On Disentanglement of Asymmetrical Knowledge Transfer for Modality-Task Agnostic Federated Learning

Jiayi Chen, Aidong Zhang

University of Virginia
jc4td@virginia.edu, aidong@virginia.edu

Abstract

There has been growing concern regarding data privacy during the development and deployment of Multimodal Foundation Models for Artificial General Intelligence (AGI), while Federated Learning (FL) allows multiple clients to collaboratively train models in a privacy-preserving manner. This paper formulates and studies **Modality-task Agnostic Federated Learning** (**AFL**) to pave the way toward privacy-preserving AGI. A unique property of AFL is the asymmetrical knowledge relationships among clients due to modality gaps, task gaps, and domain shifts between clients. This raises a challenge in learning an optimal inter-client information-sharing scheme that maximizes positive transfer and minimizes negative transfer for AFL. However, prior FL methods, mostly focusing on symmetrical knowledge transfer, tend to exhibit insufficient positive transfer and fail to fully avoid negative transfer during inter-client collaboration. To address this issue, we propose **DisentAFL**, which leverages a two-stage Knowledge Disentanglement and Gating mechanism to explicitly decompose the original asymmetrical inter-client information-sharing scheme into several independent symmetrical inter-client information-sharing schemes, each of which corresponds to certain semantic knowledge type learned from the local tasks. Experimental results demonstrate the superiority of our method on AFL than baselines.

Introduction

Artificial General Intelligence (AGI) aims to build foundation models that emulate human-like intelligence on a variety of cognitive tasks, across diverse modality types and domains (Bubeck et al. 2023). Yet recently, there has been a growing concern regarding data privacy of AGI models, in both pre-training and fine-tuning phases (Xu et al. 2023a). For example, the massive multimodal data for pre-training and the user-specific data for downstream task fine tuning might include sensitive or personal information, thus centralizing these data is not possible. Meanwhile, Federated Learning (FL) (Zhang et al. 2021) techniques allow multiple clients to collaboratively train models in a privacy-preserving manner. In this paper, we attempt to leverage FL to achieve better data privacy for AGI.

However, simply applying existing FL techniques in training or fine-tuning a large foundation model is impractical.

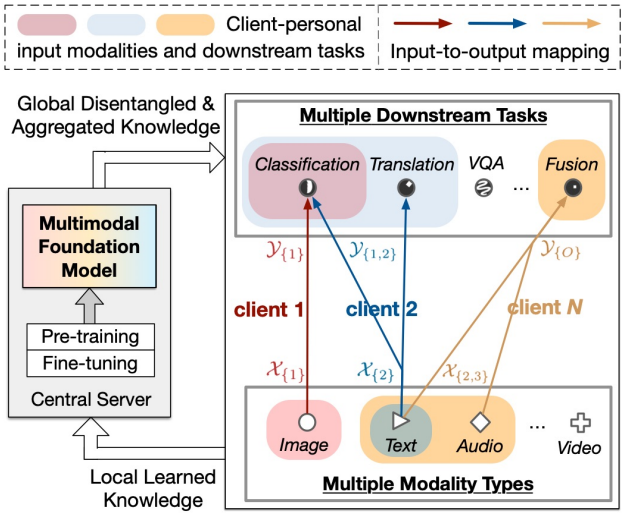


Figure 1: Modality-task Agnostic Federated Learning (AFL) for Privacy-preserving AGI. Clients learn personal models for their specific modalities and tasks using local data.

Due to computing resource limitations, clients cannot afford to train or fine-tune with the entire multimodal foundation model with billions of parameters. In addition, in the real world, each client focuses only on its specific modality types and tasks, making numerous parameters redundant for individual users. Given these facts, we explore **Modality-task Agnostic Federated Learning** (**AFL**), where each client independently trains a personalized model on its *own* modalities and tasks, while periodically collaborating with each other to aggregate knowledge onto a central server housing the large foundation model, as illustrated in Figure 1.

AFL is still an under-explored research direction in FL community. Because of the inconsistency of input modality types and downstream task types between clients, client heterogeneity in AFL is complex–there are *simultaneous* Modality gaps, Task gaps, Domain shifts, and Concept drifts (MTDC) among clients. Such an MTDC client heterogeneity imposes a unique **property** of AFL–the **Asymmetrical Knowledge Relationships** (**AKR**) among clients, meaning that the mutual knowledge between each pair of clients are greatly diversified. This raises a crucial **challenge** in learn-

ing an optimal inter-client information sharing scheme (i.e. maximizing positive transfer and minimizing negative transfer) for AFL–it would be difficult to efficiently and automatically identify correct transferable knowledge for each pair of clients through client-server interactions. Existing FL works (Jeong and Hwang 2022; Chen and Zhang 2022a) mainly address the symmetrical knowledge transfer between clients, which struggle to perform sufficient positive transfer and cannot fully avoid negative transfer during the inter-client collaboration under an AKR situation.

To overcome the abovementioned challenge in AFL and achieve an optimal inter-client information sharing scheme that maximizes positive transfer and minimizes negative transfer, we propose a novel knowledge disentanglement-based federated learning framework, namely **DisentAFL**. The key idea of DisentAFL is to explicitly disentangle the original *asymmetrical* inter-client information sharing scheme into several independent *symmetrical* inter-client information sharing schemes, each of which corresponds to certain semantic knowledge type learned from the local tasks. In details, DisentAFL empowers the server-client communication to be aware of the true pairwise mutual knowledge type(s) through a two-stage **Knowledge Disentanglement and Gating** mechanism. The stage one leverages *coarse-grained group-wise disentanglement* to reduce the original asymmetrical problem into several *intermediate* asymmetrical subproblems, and the stage two leverages *fine-grained knowledge-type disentanglement* that further decomposes each of the asymmetrical subproblems into several independent symmetric information sharing schemes. Our contributions are summarized as follows.

- We systematically study and formulate the problem of modality-task agnostic federated learning (**AFL**). To the best of our knowledge, this is one of the early attempts paving the way towards privacy-preserving AGI for multimodal tasks. Also, AFL has the potential to extend multimodal intelligence capabilities beyond traditional FL
- We propose **DisentAFL** to address the complex asymmetrical inter-client knowledge relationships of AFL. Technically, DisentAFL is one of first FL methods that explicitly leverage the fine-grained disentanglement of inter-client relationships to achieve sufficient positive knowledge while excluding negative knowledge.
- We evaluate DisentAFL on six AFL simulations, with at most 4 modalities and 4 downstream tasks. The empirical results demonstrate the effectiveness of our method.

Related Works

Artificial General Intelligence (AGI). AGI aims to attain Foundation Models that emulate human-like intelligence on a variety of cognitive *tasks* across diverse *modalities* (Bubeck et al. 2023). Multimodal Large Language Models (MLLMs) *pretrained* on large-scale multimodal data have emerged as a pivotal paradigm for AGI (Sanderson 2023; Wu et al. 2023; Yu et al. 2023). Pretrained MLLMs could quickly *adapt* to various multimodal downstream tasks through few-shot fine-tuning or zero-shot inference, catering to both deterministic tasks (e.g. multimodal fusion) (Chen and Zhang 2020, 2022b, 2021; Chen et al. 2023; Wang et al. 2022) and generative tasks (e.g. cross-modal video generation) (Chen and Zhang 2023; Seo et al. 2022b). To enhance the success of AGI, many Multimodal Interaction Modeling techniques have been incorporated into MLLMs and have played important roles (Wu et al. 2023; Li et al. 2023), including *model design* (e.g., inter-modal interaction architecture), *training* algorithms (e.g., co-training of different modalities), and *task adaptation* mechanisms (e.g., hypernetworks, soft prompting, and the prompt design of input structures that combines multiple modalities).

Personalized Federated Learning (PFL). In PFL, multiple clients/users train their *personal* models while periodically *collaborating* with each other's learned knowledge without directly exchanging their local data. The personalization in PFL is typically achieved by fine-tuning (T Dinh, Tran, and Nguyen 2020), meta-learning (Finn, Abbeel, and Levine 2017; Zheng and Zhang 2022; Zheng et al. 2023), mixture methods (Guo et al. 2021), hypernetworks (Shamsian et al. 2021), or multi-task learning (Smith et al. 2017). Another line of PFL consider the personalization of *neural architectures*, including approaches based on collaborative knowledge distillation (Jiang, Shan, and Zhang 2020; Ahmad and Aral 2022) and personally masked supernetwork (Shi et al. 2021; Kim et al. 2023; Dai et al. 2022). While our work adopts the masked super-network methods, we address an under-explored asymmetrical information sharing problem using disentanglement. Another research direction, Multimodal PFL, considers the personalization of *input modalities*, allowing different clients/users to train from different multimodal combinations (McMahan et al. 2018; Chen and Zhang 2022a; Xiong et al. 2022; Che et al. 2023). While these methods assume an embedding space where knowledge is symmetrically shared across clients, our approach considers the asymmetry of knowledge transfer.

Privacy-preserving Federated AGI. In AGI, there has been a growing concern regarding data privacy during the pre-training and fine-tuning phases of Multimodal Foundation Models. For example, the massive multimodal corpora for pre-training might include *sensitive or personal information*, thus centralizing these data is not possible; also, commercial competition tend to isolate users' feedbacks, hindering direct collaboration and knowledge sharing for downstream task fine tuning. Recent works have shown that text-only LLMs can be trained/tuned with Federated Learning for protecting users' privacy (Hilmkil et al. 2021; Zhang et al. 2023; Fowl et al. 2022; Xu et al. 2023b,a; Ait-Mlouk et al. 2023). However, there has been limited discussions on privacy-preserving AGI focusing on multimodal scenarios.

Disentanglement for Knowledge Transfer. Disentanglement, initially studied in deep generative models (Mathieu et al. 2019; Tran, Yin, and Liu 2017), has been recently utilized in multimodal representation decoupling (Hazarika, Zimmermann, and Poria 2020), cross-modal and cross-domain transfer learning (Gonzalez-Garcia, Van De Weijer, and Bengio 2018), and multimodal knowledge distillation (Li, Wang, and Cui 2023) to enhance knowledge transfer ef-

fectiveness. In Federated Learning community, recent works (Yang et al. 2023; Bercea et al. 2022; Jeong and Hwang 2022; Ye et al. 2023) show disentanglement helps to achieve better interpretability and privacy protection, as well as perform better the global-local knowledge tradeoff. Different from them, our work employs *finer*-grained disentanglement to *purify* the positive knowledge transfer among clients.

Problem Formulation

In traditional PFL, there are N clients and each client $i = 1, 2, ..., N$ aims to solve a local ***i.i.d*** learning problem $\mathcal{D}_i := (\mathcal{X}, \mathcal{Y}_i, p_i(\boldsymbol{x}), q_i(\boldsymbol{y}|\boldsymbol{x}))$ with a globally-shared input space $\mathcal{X}$, a local label space $\mathcal{Y}_i$ within a global task $\mathcal{T}$, a personal input distribution $p_i(\boldsymbol{x})$, where $\boldsymbol{x} \in \mathcal{X}$, and a ground-truth mapping function $q_i : \mathcal{X} \to \mathcal{Y}_i$ that predicts a conditional output distribution $q_i(\boldsymbol{y}|\boldsymbol{x})$, where $\boldsymbol{x} \in \mathcal{X}$ and $\boldsymbol{y} \in \mathcal{Y}_i$.

Different from traditional PFL, **Modality-task Agnostic Federated Learning (AFL)** incorporates diversified input modality types (e.g. image, text, video, audio, tabular) and diversified categories of downstream tasks (e.g. classification, fusion, translation, representation learning) into the learning systems. AFL can be widely applied to many real-world scenarios, such as Privacy-preserving AGI, Artificial Internet of Things (AIoT), and Learning-at-home (Wu et al. 2020). Formally, assuming a total of M types of modalities and O types of downstream tasks over the N clients. Let $\mathcal{X}^{(m)}$ denote the raw input space associated to the m-th modality type and $\mathcal{Y}^{(o)}$ the label space for the o-th task. Typically, **each client i does not learn all the modalities and all the tasks**; instead, it has its own input modality types $\mathcal{I}_i \subseteq [M]$ and target task types $\mathcal{O}_i \subseteq [O]$. Each client $i = 1, 2, ..., N$ aims to learn a personal mapping function

$$\widehat{q}_i(\cdot; \boldsymbol{\omega}_i) : \mathcal{X}_{\mathcal{I}_i} \to \mathcal{Y}_{\mathcal{O}_i} \tag{1}$$

from a client-specific structured/joint input space $\mathcal{X}_{\mathcal{I}_i} := \texttt{Join}(\mathcal{X}^{(m)}|\forall m \in \mathcal{I}_i)$ to the *client-specific* label spaces of each local tasks $\mathcal{Y}_{\mathcal{O}_i} := \{\mathcal{Y}^{(o)}|\forall o \in \mathcal{O}_i\}$, where $\boldsymbol{\omega}_i \in \mathbb{R}^{d_i^{\text{param}}}$ denotes trainable weights. Then, the local problem is formulated as $\mathcal{D}_i := (\mathcal{X}_{\mathcal{I}_i}, \mathcal{Y}_{\mathcal{O}_i}, p_i(\widetilde{\boldsymbol{x}}), q_i(\widetilde{\boldsymbol{y}}|\widetilde{\boldsymbol{x}}))$, where $\widetilde{\boldsymbol{x}} \in \mathcal{X}_{\mathcal{I}_i}$ and $q_i(\widetilde{\boldsymbol{y}}|\widetilde{\boldsymbol{x}}) : \mathcal{X}_{\mathcal{I}_i} \to \mathcal{Y}_{\mathcal{O}_i}$ is the ground-truth conditional output distribution with $\widetilde{\boldsymbol{y}} = \{\boldsymbol{y}^{(o)}\}_{o \in \mathcal{O}_i} \in \mathcal{Y}_{\mathcal{O}_i}$. Figure 1 shows an illustration of the AFL problem setting.

The **local objective** of client i minimizes multiple losses $\min_{\boldsymbol{\omega}_i} f_i(\boldsymbol{\omega}_i) := \mathbb{E}_{(\widetilde{\boldsymbol{x}}, \widetilde{\boldsymbol{y}}) \sim \mathcal{D}_i} \frac{1}{|\mathcal{O}_i|} \sum_{o \in \mathcal{O}_i} \mathcal{L}^{(o)} \left(\boldsymbol{y}^{(o)}, \widehat{q}_i(\widetilde{\boldsymbol{x}}; \boldsymbol{\omega}_i)_o \right)$ where $\mathcal{L}^{(o)}$ is the loss function for the type-o task. Then, following PFL (Chen and Zhang 2022a; T Dinh, Tran, and Nguyen 2020), the **global objective** of AFL is formulated as

$$\min_{\boldsymbol{\omega}_1, \boldsymbol{\omega}_2, ..., \boldsymbol{\omega}_N} \left[\frac{1}{N} \sum_{i=1}^{N} f_i(\boldsymbol{\omega}_i) \right] + \mathcal{R}(\boldsymbol{\omega}_1, \boldsymbol{\omega}_2, ..., \boldsymbol{\omega}_N), \tag{2}$$

where the regularizer $\mathcal{R}(\cdot)$ indicates the information sharing scheme (i.e. knowledge transfer) among clients, which is encouraged to transfer beneficial knowledge among clients to boost each local model's performance.

Client Heterogeneity in AFL. Since clients in AFL do not necessarily have the same input modalities or downstream tasks, there could be simultaneous 4 heterogeneity patterns between clients: **M**odality gap, **T**ask gap, **D**omain shift, and **C**oncept drift (**MTDC**). (1) **M** (***modality gap***): the clients vary in their *input spaces* due to their input modality divergence, that is, $\mathcal{X}_{\mathcal{I}_i} \neq \mathcal{X}_{\mathcal{I}_{i'}}$ when $\mathcal{I}_i \neq \mathcal{I}_{i'}$. For example, a vehicle may use its onboard camera to capture videos to predict traffics, while another vehicle may use both video and RADAR signals to predict traffics. (2) **T** (***task gap***): clients vary in their *output spaces* $\mathcal{Y}_{\mathcal{O}_i} \neq \mathcal{Y}_{\mathcal{O}_{i'}}$ since they target at different downstream tasks $\mathcal{O}_i \neq \mathcal{O}_{i'}$. For example, while a client may focus on image classification, the other client may focus on image segmentation. (3) **D** (***domain shift***): slightly different from traditional FL's definition on domain shift, AFL considers the joint distribution shift, meaning that the multimodal interaction behaviors can vary between clients. (4) **C** (***concept drift***): clients vary in their conditional *output distribution*, or label space.

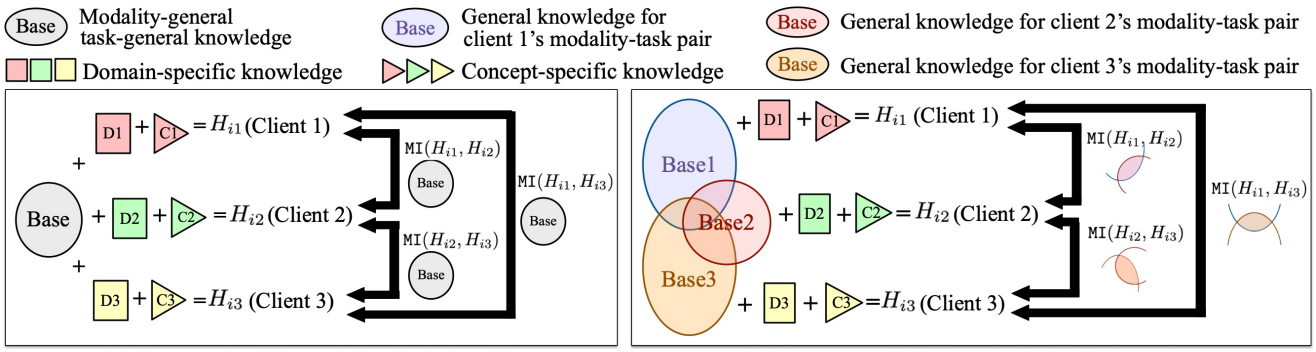
Asymmetrical Knowledge Transfer in AFL

We begin with discussing the key challenges in solving the AFL's global objective (Eq.(2)) due to MTDC heterogeneity.

Definition 1 **(Positive & Negative Knowledge Transfer).** Positive Transfer (**PT**) is defined as the information sharing behavior between a *pair* of clients that will lead to the improvement of each other models. Negative Transfer (**NT**), on the other hand, is a phenomenon when sharing parameters between two local models results in poorer results than solving individual tasks (or, *unlearning*).

Rethinking Information Sharing in Federated Learning. The information sharing scheme $\mathcal{R}(\boldsymbol{w}_{1:N})$ in FL is essentially to find an inter-client Pairwise Knowledge Transfer (**PKT**) mechanism that can lead to the improvement of each client model. For any pair of clients, there exists both mutual common knowledge and conflicting knowledge between them–if $\nabla_\psi f_i(\psi) \nabla_\psi f_{i'}(\psi) > 0$, we say the knowledge representation ψ at client i and client i' aligns/matches with each other; on the other hand, if $\nabla_\psi f_i(\psi) \nabla_\psi f_{i'}(\psi) < 0$, the knowledge ψ at client i and client i' conflicts. As in (Wu, Zhang, and Ré 2020), the transfer behavior of conflicting knowledge will result in Negative Transfer; and, the un-transfer of common knowledge will result in *insufficient* Positive Transfer. Both need to be avoided for better performance. Therefore, the optimal $\mathcal{R}(\boldsymbol{w}_{1:N})$ relies on a PKT mechanism that can ***maximize positive transfer*** *and* ***minimize negative transfer*** between each pair of clients–that is, all the true aligned knowledge is encouraged to be transferred and all the true conflicting knowledge should be excluded during transfer.

Definition 2 **(Symmetrical & Asymmetrical Knowledge Relationships).** Suppose H_i denotes the knowledge learned by the client i and $\texttt{MI}(H_i, H_{i'})$ denotes the *true* mutual/common knowledge between by a pair of clients (i, i'). We say the knowledge relationships over N clients is ***symmetrical*** if the mutual information (common knowledge) between each pair of clients are the same $\texttt{MI}(H_{i1}, H_{i2}) =$



(a) Symmetrical Knowledge Relationships (b) Asymmetrical Knowledge Relationships

Figure 2: Comparison between symmetrical and asymmetrical inter-client knowledge relationships. In (a), the three example clients share the same modality-task pair. In (b), the three example clients have different modality-task pairs.

$\mathtt{MI}(H_{i2}, H_{i3}) = \mathtt{MI}(H_{i1}, H_{i3}), \forall i_1, i_2, i_3 \in [N]$. On the other hand, we say the knowledge relationships over N clients is ***asymmetrical*** if $\mathtt{MI}(H_{i1}, H_{i2}) \neq \mathtt{MI}(H_{i2}, H_{i3}) \neq \mathtt{MI}(H_{i1}, H_{i3}), \exists i_1, i_2, i_3 \in [N]$. Figure 2 shows an comparison between the two scenarios.

Challenge of Optimizing Information Sharing in AFL. Existing FL algorithms mainly address the ***symmetrical*** knowledge relationships. For example, Non-IID PFL (Jeong and Hwang 2022; Guo et al. 2021) with a universal domain shift and concept shift can be a *symmetrical* case (Figure 2(a)) since there exists global common knowledge $\mathtt{MI}_{\mathrm{g}} = \mathtt{MI}(H_i, H_{i'})$ shared by all pairs of clients $i, i' \in [N]$ and all the other learned knowledge is considered as personalized knowledge. However, due to the complexity of MTDC heterogeneity, **AFL** has more complex ***asymmetrical*** knowledge relationships among clients, as illustrated in Figure 2(b). Using the 4 clients in Figure 4 as an example, the common knowledge between client 1 and client 3 includes the modality-1's encoding function, which is however not shareable between client 1 and client 2 since the modality 1 is not learned at client 2. Unfortunately, the asymmetrical knowledge relationships in AFL brings difficulties in optimizing the information sharing scheme $\mathcal{R}$–it is hard to efficiently and adaptively identify transferable knowledge for each pair of clients through client-server interactions. Existing FL methods may result in negative transfer or insufficient positive transfer under the asymmetry of AFL.

Given such *complex* and *unknown* user-to-user knowledge sharing in AFL, it is desirable to explicitly maximize positive transfer and minimize negative transfer for the optimization of $\mathcal{R}$. Ideally, for any pair of clients (i, i'), an optimal PKT mechanism should perform the transfer to approximate the true mutual knowledge $\mathtt{MI}(H_i, H_{i'})$.

Proposed DisentAFL

In order to achieve an efficient and optimal PKT mechanism that maximizes positive transfer and minimizes negative transfer with the *asymmetrical* knowledge relationships of AFL, we propose **DisentAFL**, whose overview is shown in Figure 3. The key idea is to *disentangle* the asymmetrical information sharing scheme on the original knowledge space into K independent *symmetrical* information sharing schemes on each of the disentangled knowledge subspaces

$$\mathcal{R}(\boldsymbol{w}_1, \boldsymbol{w}_2, ..., \boldsymbol{w}_N) = \sum_{k=1}^{K} \mathcal{R}_k(\{\boldsymbol{w}_i^{(k)} | \forall i \in C_k\}) \quad (3)$$

such that each $\mathcal{R}_k(\cdot)$ is a *symmetric* information sharing scheme among a subset of clients $C_k \subseteq [N]$, where $\boldsymbol{w}_i^{(k)}$ is the disentangled knowledge type k extracted from $\boldsymbol{w}_i$.

Specifically, to find an optimal inter-client communication solution for Eq.(3), we propose a Knowledge Disentanglement and Gating (KDG) mechanism, which consists of two stages: **coarse-grained** group-wise disentanglement and **fine-grained** knowledge-type disentanglement. The two-stage KDG mechanism is shown in Figure 4.

Stage One: Coarse-grained Disentanglement

Group-wise disentanglement reduces the *original asymmetrical problem* with complex MTDC heterogeneity into several *intermediate asymmetrical subproblems* with less complex client diversity. **First,** we separate the encoding and decoding related knowledge such that the clients sharing the same modality or downstream task could share the corresponding encoder or decoder parameters/representations. For example, a client aiming at image classification task using the ViT (Liu et al. 2023) encoder and a MLP classification head, might share the image encoder with an image-text classification client that uses a Multimodal Transformer backbone (Xu, Zhu, and Clifton 2022). We rewrite the local parameters of i-th client as $\boldsymbol{w}_i = \{\boldsymbol{\phi}_i^{(m)}\}_{m \in \mathcal{I}_i} \cup \{\boldsymbol{\theta}_i^{(o)}\}_{o \in \mathcal{O}_i}$, where $\boldsymbol{\phi}_i^{(m)}$ denotes the modality m's encoder and $\boldsymbol{\theta}_i^{(o)}$ denotes the decoder for the type-o downstream task. We define two types of knowledge groups: (1) encoding-knowledge groups $\mathcal{G}_{\mathrm{enc}}^{(m)} = \{\boldsymbol{\phi}_i^{(m)} | \forall i \in [N] \text{ if } m \in \mathcal{I}_i\}$, where each group is a collection of encoders from those clients having the modality m within their inputs; and (2) decoding-knowledge groups $\mathcal{G}_{\mathrm{dec}}^{(o)} = \{\boldsymbol{\theta}_i^{(o)} | \forall i \in [N] \text{ if } o \in \mathcal{O}_i\}$, where each group is a collection of decoders from those clients having the target downstream task type o. **Then,** we can rewrite

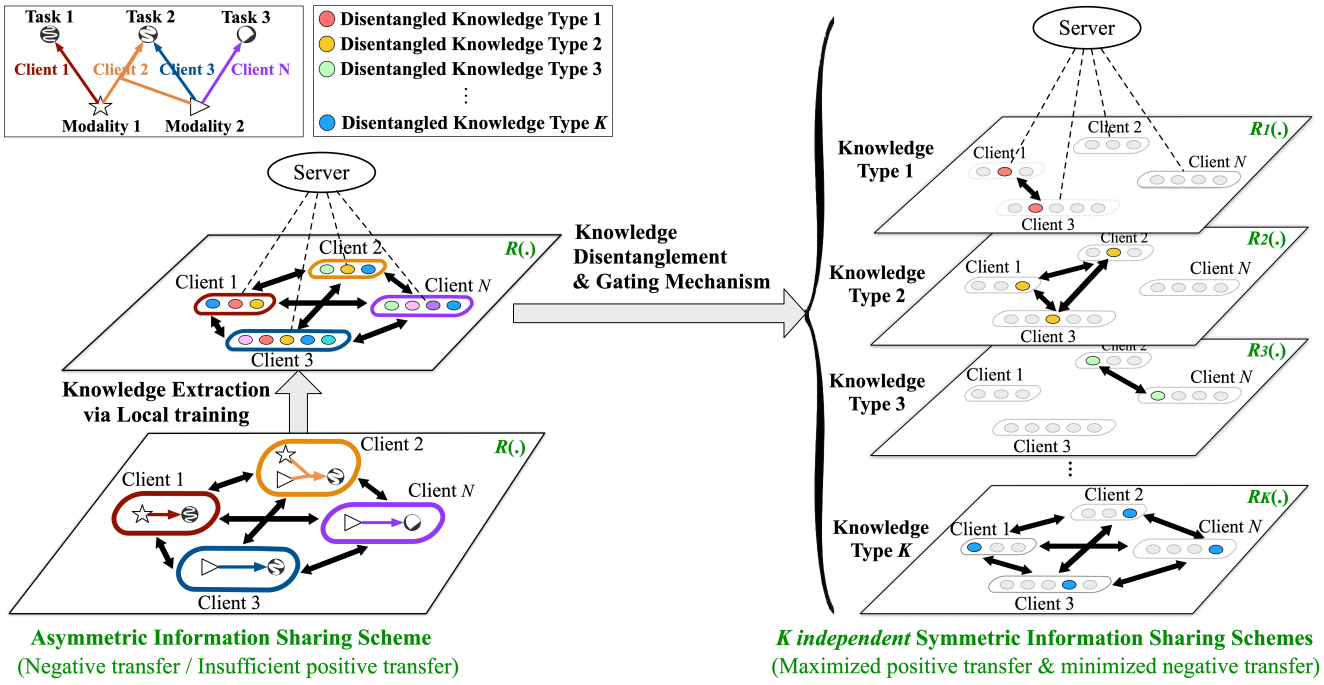


Figure 3: Overview of the proposed DisentAFL.

the asymmetrical information sharing scheme $\mathcal{R}(\boldsymbol{w}_{1:N})$ as

$$\mathcal{R}(\boldsymbol{w}_{1:N}) = \sum_{m=1}^{M} \mathcal{R}_{\text{IE}}(\mathcal{G}_{\text{enc}}^{(m)}) + \sum_{o=1}^{O} \mathcal{R}_{\text{ID}}(\mathcal{G}_{\text{dec}}^{(o)}) + \sum_{m,m'=1}^{M} \mathcal{R}_{\text{XE}}(\mathcal{G}_{\text{enc}}^{(m)}, \mathcal{G}_{\text{enc}}^{(m')}) + \sum_{o,o'=1}^{O} \mathcal{R}_{\text{XD}}(\mathcal{G}_{\text{dec}}^{(o)}, \mathcal{G}_{\text{dec}}^{(o')}). \tag{4}$$

The $\mathcal{R}(\boldsymbol{w}_{1:N})$ with MTDC client heterogeneity is split into four sub-problems: (1) $\mathcal{R}_{\text{IE}}(\cdot)$ indicates the information sharing scheme within each modality-specific group $\mathcal{G}_{\text{enc}}^{(m)}$, which is an *asymmetrical* but *single-modal task-agnostic* problem with **TD** heterogeneity (no modality shift and concept shift). (2) $\mathcal{R}_{\text{ID}}(\cdot)$ indicates that within each task-specific group $\mathcal{G}_{\text{dec}}^{(o)}$, which is an *asymmetrical* but *modality-agnostic single-task* problem with **MC** heterogeneity (no task shift and domain shift). (3) $\mathcal{R}_{\text{XE}}(\cdot,\cdot)$ indicates the potential encoding-information sharing between clients having different modalities, which is an *asymmetrical* but *cross-modal task-agnostic* problem with **MT** heterogeneity (no domain shift and concept shift). (4) $\mathcal{R}_{\text{XD}}(\cdot,\cdot)$ indicates the decoding-information sharing scheme between the clients that have diversified downstream tasks, which is an *asymmetrical* but *cross-task modality-agnostic* problem with **MT** heterogeneity (no domain shift and concept shift).

Stage Two: Fine-grained Disentanglement

We further disentangle each of the above four asymmetrical sub-problems into several independent symmetric problems.

To achieve this, we first need to find the largest knowledge components that can sufficiently describe the global asymmetric PKT problem as the combination of several symmetric PKT problems. Specifically, we assume a total of $K = M(D+1)+O(N+1)+(M+1)(O+1)$ fine-grained knowledge types **globally existing** over the N clients: **(1)** $M(D+1)$ knowledge types related to domain shift of each modality; each domain $d \in [D]$ consists of a domain-specific and a domain-agnostic knowledge. **(2)** At most $O(N+1)$ knowledge types related to concept drift regarding individual fine-tuning on the decoder. **(3)** $(M+1)(O+1)$ knowledge types related to modality and task gaps, including the task-specific and task-shared knowledge per modality; the modality-specific and modality-shared knowledge per task type; and the knowledge shared by all tasks and all modalities, such as the commonsense cognition.

Supernetwork on Server. The central server hosts a multimodal multitask large model, which serves as a supernetwork $\boldsymbol{w}^{\text{sup}}$ that can accommodate the K global fine-grained knowledge types mentioned above. The neural architecture of $\boldsymbol{w}^{\text{sup}}$ can be any popular foundation models. In our experiments, we use a Multi-input Multi-head Transformer, consisting of M input channels and O output channels, respectively, for all the seen modalities and task types over clients. Within the network, we design several Mixture of Domain Experts (**MoDE**) layers to capture D domain-specific knowledge types and one additional domain-agnostic type. Each of the D parallel expert models in MoDE stands for the knowledge type for a specific domain d. MoDE layers act as residual connections attached to an original model block and are located around the query and value layers. In addition, to bridge the knowledge gap between modalities and tasks in each subproblem, we employ Mixture of Task Experts (**MoTE**) and Mixture of Modality Experts (**MoME**) to capture $(M+1)(O+1)$ **modality-task interactive** knowledge

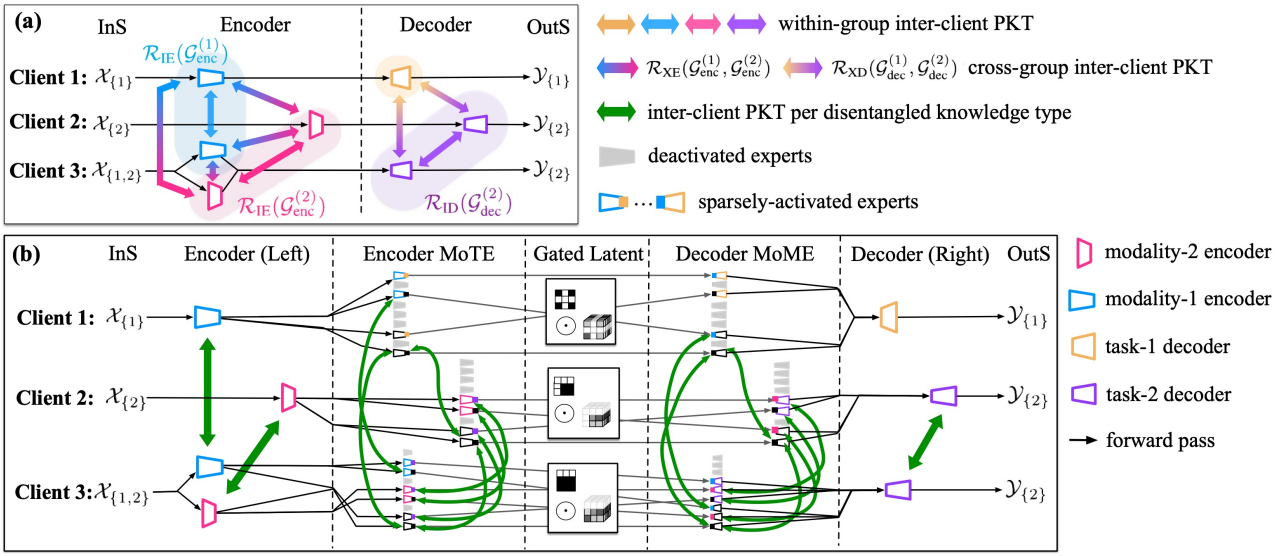


Figure 4: The two-stage Knowledge Disentanglement and Gating mechanism in DisentAFL. (a) The intermediate asymmetrical information sharing schemes after coarse-grained disentanglement; (b) The final symmetrical information sharing schemes after fine-grained disentanglement.

types. Each modality m's encoded representation is split as $\boldsymbol{h}'^{(m)} = [\boldsymbol{h}^{\text{share}}||\boldsymbol{h}^{(m)}]$, where $||$ denotes concatenation operation, $\boldsymbol{h}^{(m)}$ represents the modality-private and $\boldsymbol{h}^{\text{share}} \in \mathbb{R}^d$ the modality-shared knowledge. Likewise, the pre-decoding representation of each task o consists of task-specific and task-shared information, $\boldsymbol{t}'^{(o)} = [\boldsymbol{t}^{\text{share}}||\boldsymbol{t}^{(o)}]$. The modality-task interactive representation between a pair of MoTE and MoME is a tensor cube $\boldsymbol{Z} \in \mathbb{R}^{(M+1)\times(O+1)\times F}$ featuring the $(M+1)(O+1)$ knowledge types. The detailed architecture of $\boldsymbol{w}^{\text{sup}}$ is provided in Figure 5 in Appendix.

Disentanglement Losses. Disentanglement of $\boldsymbol{Z}$ is important for purifying and separating the semantics of knowledge transfer. To encourage this, we introduce auxiliary losses to the local objective. Many advanced disentanglement techniques can be applied here (Lee and Pavlovic 2021). For example, the orthogonal regularization loss $\mathcal{L}_i^{\text{orth}}(\boldsymbol{w}_i) = \sum_{(m,o),(m',o')\in\mathcal{I}_i\times\mathcal{O}_i} \boldsymbol{Z}_{o,m,\cdot}^{\top}\boldsymbol{Z}_{o',m',\cdot}$ computed from each pair of knowledge types.

Sparsely-gated Client Network. Each client i's local network $\boldsymbol{w}_i$ encapsulates only $K_i = 2|\mathcal{I}_i| + 2|\mathcal{O}_i| + (|\mathcal{I}_i|+1)(|\mathcal{O}_i|+1)$ client-personal knowledge types, therefore significantly smaller than $\boldsymbol{w}^{\text{sup}}$. The inter-client collaboration is semantically disentangled and performed by using a ***routing*** mechanism. Two gating functions is designed to achieve this. **(1)** $\texttt{IoGate}(\cdot)$ takes as input the samples or modality-task indicators $\mathcal{I}_i, \mathcal{O}_i$, and outputs a binary gate matrix $\boldsymbol{S}_i \in \{0,1\}^{(M+1)\times(O+1)}$, where each entry $\boldsymbol{S}_{i,m,o} = 1$ if $(m \in \mathcal{I}_i \wedge o \in \mathcal{O}_i) \vee (m \in \mathcal{I}_i \wedge o = O+1) \vee (m = M+1 \wedge o \in \mathcal{O}_i) \vee (m = M+1 \wedge o = O+1)$; otherwise, $\boldsymbol{S}_{i,m,o} = 0$. The $\boldsymbol{S}_{i,M+1,O+1}$ always equals to one because the commonsense knowledge is shareable between all clients, bridging the gap between any pair of clients with $\mathcal{I}_i \cap \mathcal{I}_{i'} = \emptyset \wedge \mathcal{O}_i \cap \mathcal{O}_{i'} = \emptyset$. **(2)** $\texttt{DomGate}(\cdot)$ produces a D-dimensional one-hot vector $\boldsymbol{g}_i \in \{0,1\}^D$, where $D <= N$ denotes the pre-defined number of domains over clients. The binary outputs of the two gates $\boldsymbol{S}_i, \boldsymbol{g}_i$ are used to route each client's network through the super-network $\boldsymbol{w}_i = \texttt{ROUTE}(\boldsymbol{S}_i, \boldsymbol{g}_i; \boldsymbol{w}^{\text{sup}})$. As $\boldsymbol{S}_i, \boldsymbol{g}_i$ are very sparse, i.e., $K_i << K$, the client network $\boldsymbol{w}_i$ is much thinner than $\boldsymbol{w}^{\text{sup}}$. Design details are shown in Figure 6 in Appendix.

Proof of Symmetrical PKT After Disentanglement

Due to page limit, detailed proof is provided in Appendix. We prove that the proposed two-stage disentanglement can successfully decompose the original asymmetric client relationships $\mathcal{R}(\boldsymbol{w}_{1:N})$ into $K = M(D+1)+O(N+1)+(M+1)(O+1)$ independent symmetric client relationships.

The training workflow and the pseudo-code of DisentAFL is provided in Algorithm 1 in the Supplementary Materials.

Experiments

AFL Simulation Setup

We select *seven* multimodal or multitask datasets as the source to create *six* AFL simulations. The seven **source datasets** are summarized in Table 3 in Appendix, including two image classification datasets (Finn, Abbeel, and Levine 2017), a bimodal driving dataset (Duarte and Hu 2004), a bimodal 3D object recognition dataset (Wu et al. 2015; Feng et al. 2019), a three-modal two-task multimedia emotion recognition dataset and a bimodal audio-image classification dataset (Liang et al. 2021). We then create *six* simulations from these datasets. (1) ***MERGE-AC*** simulates a basic single-modal, single-task, and *cross-domain* FL scenario with a total of 15 clients. (2) ***ModelNet-xM*** and ***Vehicle-xM*** simulate *cross-modal*, single-task, and single-domain FL scenarios with more than 20 clients. (3) ***MERGE-VM*** simulates an 4-modal 2-downstream-task AFL scenario with

Method	MERGE-AC	Vehicle-xM	ModelNet-xM	MERGE-VM	MERGE-MM
Local	77.12 ± 0.39	83.48 ± 0.89	93.01 ± 0.30	88.23 ± 0.72	70.23 ± 0.79
FedAvg (McMahan et al. 2018)	76.78 ± 0.55	73.18 ± 0.09	92.79 ± 0.12	84.63 ± 0.02	74.12 ± 0.93
Cross-FedAvg (McMahan et al. 2018)	78.16 ± 0.23	84.43 ± 0.82	91.65 ± 0.32	88.35 ± 0.20	72.41 ± 0.72
Align-FedAvg (McMahan et al. 2018)	75.30 ± 0.85	73.32 ± 0.58	89.18 ± 0.53	89.73 ± 0.68	69.65 ± 0.73
Cross-PFL (Smith et al. 2017)	81.58 ± 0.53	86.82 ± 0.38	94.20 ± 0.25	90.11 ± 0.63	75.37 ± 0.26
FedMSplit (Chen and Zhang 2022a)	78.32 ± 0.31	85.12 ± 0.03	90.79 ± 0.73	87.37 ± 0.03	73.25 ± 0.31
DisentAFL-KD	80.47 ± 0.53	87.42 ± 0.23	**96.62** ± 0.16	94.33 ± 0.32	75.17 ± 0.37
DisentAFL-Avg	**82.66** ± 0.74	**88.56** ± 0.22	96.44 ± 0.14	**96.38** ± 0.41	**75.68** ± 0.74

Table 1: Comparison of the average testing accuracy over all clients on their classification tasks.

discrepant input spaces, output spaces, and output distributions (MTC) across 50 clients. (4) ***MERGE-MM*** simulates a 4-modal 3-downstream-task AFL scenario with the 4 patterns of heterogeneity (MTDC) across 50 clients. (5) ***MERGE-FA*** simulates a 2-modal 4-downstream-task AFL scenario across 30 clients with MTDC heterogeneity. The four downstream tasks include classifying the item on the top-left, on the bottom-right, generating the digit image, and generating the audio signal of the digits. Details of our AFL simulation design are summarized in Table 4 in Appendix.

Baseline Methods

We compared DisentAFL with six baseline methods: (1) **Local**: clients separately train their models without any collaboration–neither positive transfer nor negative transfer ($\mathcal{R}(\cdot)$=0). (2) **FedAvg** (McMahan et al. 2018): clients are split into several disjoint groups such that each group share the same modality-task pair. The collaboration is within the same group of clients using FedAvg. Any information sharing between different groups is prohibited. (3) **Cross-FedAvg**, in addition to FedAvg, encourages the sharing of certain modality-to-task transmitter between different groups that have overlapping on both modalities and tasks, as illustrated in Figure 7(a) in Appendix. There is no modality-shared or task-shared representations in this baseline. (4) **Align-FedAvg**, in addition to FedAvg, encourages the sharing of certain encoders/decoders between different groups that have either overlapping modalities or overlapping tasks. The after-encoding and before-decoding representations of all modalities and task are aligned onto the same latent space. (5) **Cross-PFL** is similar to Cross-FedAvg, except that using the personalized FL method (Smith et al. 2017) to every modality-task pair group of clients. (6) **FedMSplit** (Chen and Zhang 2022a) is an Align-PFL method assuming latent space alignment, leveraging multimodal split networks to arbitrarily encourages the information sharing between different groups. Implementation details can be found in supplementary materials.

Main Results

We implemented DisentAFL using PyTorch and ran each experiment by 5 trials. The hyperparameters are listed in Appendix. Given that knowledge can be represented as features or parameters, we implemented two versions of DisentAFL: DisentAFL-KD incorporated with federated knowledge distillation (Seo et al. 2022a) for feature aggregation; DisentAFL-Avg based on gradient alignment through gradient aggregation. In Table 1, we report the results on five simulations. Table 1 (column 2) shows the results on the single-task, single-modal, and multi-domain simulation (MERGE-AC), where DisentAFL had MoDE module but the MoME and MoTE modules are removed, which demonstrates the effectiveness of the mixture of domain experts in our method. Table 1 (columns 3-4) shows the results on the cross-modal, single-task, and single-domain simulations (Vehicle-xM and ModelNet-xM), where DisentAFL had MoME module but the MoDE and MoTE modules are removed. Table 1 (columns 5-6) shows results on the cross-modal cross-task AFL simulations (MERGE-VM and MERGE-MM), where MoDE, MoME, and MoTE participated in the training and disentanglement losses were applied on the latent space.

Method	w/ Aux	w/o S1 Aux	w/o S2 Aux
DisentAFL-KD	91.51 ± 0.36	88.44 ± 0.36	87.93 ± 0.93
DisentAFL-Avg	93.30 ± 0.25	85.56 ± 0.31	86.48 ± 0.41

Table 2: Ablation Study on MERGE-FA.

Ablation Study

Table 2 compares the results on MERGE-FA with and without the auxiliary losses for knowledge type disentanglement on the latent space. Column 3 and 4 only remove the stage-one and stage-two disentanglement loss, respectively. Their performance drop indicate that the latent spaces of local models in AFL contain conflicting knowledge and shows the benefits of using disentanglement loss in DisentAFL. More ablation results are provided in the supplementary materials.

Conclusions

This paper studied the Modality-task Agnostic Federated Learning (AFL) problem, where different clients address different input modality types and downstream tasks. We discussed a unique challenge in AFL rather than traditional FL due to the asymmetrical inter-client knowledge relationships. Then, we introduced a new DisentAFL approach that can addressed this challenge via a two-stage Knowledge Disentanglement and Gating mechanism, whose main idea is to decompose the asymmetrical inter-client information sharing scheme into several independent symmetrical inter-client information sharing schemes. Experiments demonstrated our claims on AFL and effectiveness of our method.

Acknowledgments

We would like to express sincere appreciation to all the reviewers for their constructive feedbacks, which greatly improved the quality of this paper. This work is supported in part by the US National Science Foundation under grants 2217071, 2213700, 2106913, 2008208, 1955151.

References

Ahmad, S.; and Aral, A. 2022. FedCD: Personalized federated learning via collaborative distillation. In *2022 IEEE/ACM 15th International Conference on Utility and Cloud Computing (UCC)*, 189–194. IEEE.

Ait-Mlouk, A.; Alawadi, S.; Toor, S.; and Hellander, A. 2023. FedBot: Enhancing Privacy in Chatbots with Federated Learning. *arXiv preprint arXiv:2304.03228*.

Bercea, C. I.; Wiestler, B.; Rueckert, D.; and Albarqouni, S. 2022. Federated disentangled representation learning for unsupervised brain anomaly detection. *Nature Machine Intelligence*, 4(8): 685–695.

Bubeck, S.; Chandrasekaran, V.; Eldan, R.; Gehrke, J.; Horvitz, E.; Kamar, E.; Lee, P.; Lee, Y. T.; Li, Y.; Lundberg, S.; et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Che, L.; Wang, J.; Zhou, Y.; and Ma, F. 2023. Multimodal federated learning: A survey. *Sensors*, 23(15): 6986.

Chen, J.; Dai, H.; Dai, B.; Zhang, A.; and Wei, W. 2023. On Task-personalized Multimodal Few-shot Learning for Visually-rich Document Entity Retrieval. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Findings of the Association for Computational Linguistics: EMNLP 2023*, 9006–9025. Singapore: Association for Computational Linguistics.

Chen, J.; and Zhang, A. 2020. HGMF: Heterogeneous Graph-based Fusion for Multimodal Data with Incompleteness. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, 1295–1305.

Chen, J.; and Zhang, A. 2021. HetMAML: Task-heterogeneous model-agnostic meta-learning for few-shot learning across modalities. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 191–200.

Chen, J.; and Zhang, A. 2022a. FedMSplit: Correlation-adaptive federated multi-task learning across multimodal split networks. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 87–96.

Chen, J.; and Zhang, A. 2022b. Topological Transduction for Hybrid Few-Shot Learning. In *Proceedings of the ACM Web Conference 2022*, WWW '22, 3134–3142. New York, NY, USA: Association for Computing Machinery. ISBN 9781450390965.

Chen, J.; and Zhang, A. 2023. On Hierarchical Disentanglement of Interactive Behaviors for Multimodal Spatiotemporal Data with Incompleteness. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 213–225.

Dai, R.; Shen, L.; He, F.; Tian, X.; and Tao, D. 2022. DisPFL: Towards Communication-Efficient Personalized Federated Learning via Decentralized Sparse Training. In *International Conference on Machine Learning*, 4587–4604. PMLR.

Duarte, M. F.; and Hu, Y. H. 2004. Vehicle classification in distributed sensor networks. *Journal of Parallel and Distributed Computing*, 64(7): 826–838.

Feng, Y.; You, H.; Zhang, Z.; Ji, R.; and Gao, Y. 2019. Hypergraph neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 3558–3565.

Finn, C.; Abbeel, P.; and Levine, S. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, 1126–1135. PMLR.

Fowl, L. H.; Geiping, J.; Reich, S.; Wen, Y.; Czaja, W.; Goldblum, M.; and Goldstein, T. 2022. Decepticons: Corrupted Transformers Breach Privacy in Federated Learning for Language Models. In *NeurIPS ML Safety Workshop*.

Gonzalez-Garcia, A.; Van De Weijer, J.; and Bengio, Y. 2018. Image-to-image translation for cross-domain disentanglement. *Advances in neural information processing systems*, 31.

Guo, B.; Mei, Y.; Xiao, D.; and Wu, W. 2021. PFL-MoE: Personalized Federated Learning Based on Mixture of Experts. In *Web and Big Data: 5th International Joint Conference, APWeb-WAIM 2021, Guangzhou, China, August 23–25, 2021, Proceedings, Part I*, 480–486.

Hazarika, D.; Zimmermann, R.; and Poria, S. 2020. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM international conference on multimedia*, 1122–1131.

Hilmkil, A.; Callh, S.; Barbieri, M.; Sütfeld, L. R.; Zec, E. L.; and Mogren, O. 2021. Scaling federated learning for fine-tuning of large language models. In *International Conference on Applications of Natural Language to Information Systems*, 15–23. Springer.

Jeong, W.; and Hwang, S. J. 2022. Factorized-FL: Agnostic Personalized Federated Learning with Kernel Factorization & Similarity Matching. arXiv:2202.00270.

Jiang, D.; Shan, C.; and Zhang, Z. 2020. Federated learning algorithm based on knowledge distillation. In *2020 International Conference on Artificial Intelligence and Computer Engineering (ICAICE)*, 163–167. IEEE.

Kim, M.; Yu, S.; Kim, S.; and Moon, S.-M. 2023. DepthFL: Depthwise Federated Learning for Heterogeneous Clients. In *The Eleventh International Conference on Learning Representations*.

Lee, M.; and Pavlovic, V. 2021. Private-shared disentangled multimodal vae for learning of latent representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1692–1700.

Li, Y.; Quan, R.; Zhu, L.; and Yang, Y. 2023. Efficient Multimodal Fusion via Interactive Prompting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2604–2613.

Li, Y.; Wang, Y.; and Cui, Z. 2023. Decoupled Multimodal Distilling for Emotion Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6631–6640.

Liang, P. P.; Lyu, Y.; Fan, X.; Wu, Z.; Cheng, Y.; Wu, J.; Chen, L.; Wu, P.; Lee, M. A.; Zhu, Y.; et al. 2021. Multibench: Multiscale benchmarks for multimodal representation learning. *arXiv preprint arXiv:2107.07502*.

Liu, Y.; Zhang, Y.; Wang, Y.; Hou, F.; Yuan, J.; Tian, J.; Zhang, Y.; Shi, Z.; Fan, J.; and He, Z. 2023. A survey of visual transformers. *IEEE Transactions on Neural Networks and Learning Systems*.

Mathieu, E.; Rainforth, T.; Siddharth, N.; and Teh, Y. W. 2019. Disentangling disentanglement in variational autoencoders. In *International conference on machine learning*, 4402–4412. PMLR.

McMahan, H. B.; Ramage, D.; Talwar, K.; and Zhang, L. 2018. Learning Differentially Private Recurrent Language Models. In *International Conference on Learning Representations*.

Sanderson, K. 2023. GPT-4 is here: what scientists think. *Nature*, 615(7954): 773.

Seo, H.; Park, J.; Oh, S.; Bennis, M.; and Kim, S.-L. 2022a. Federated Knowledge Distillation. *Machine Learning and Wireless Communications*, 457.

Seo, P. H.; Nagrani, A.; Arnab, A.; and Schmid, C. 2022b. End-to-end generative pretraining for multimodal video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17959–17968.

Shamsian, A.; Navon, A.; Fetaya, E.; and Chechik, G. 2021. Personalized federated learning using hypernetworks. In *International Conference on Machine Learning*, 9489–9502. PMLR.

Shi, N.; Lai, F.; Kontar, R. A.; and Chowdhury, M. 2021. Fed-ensemble: Improving generalization through model ensembling in federated learning. *arXiv preprint arXiv:2107.10663*.

Smith, V.; Chiang, C.-K.; Sanjabi, M.; and Talwalkar, A. S. 2017. Federated multi-task learning. *Advances in neural information processing systems*, 30.

T Dinh, C.; Tran, N.; and Nguyen, J. 2020. Personalized federated learning with moreau envelopes. *Advances in Neural Information Processing Systems*, 33: 21394–21405.

Tran, L.; Yin, X.; and Liu, X. 2017. Disentangled representation learning gan for pose-invariant face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1415–1424.

Wang, Y.; Chen, X.; Cao, L.; Huang, W.; Sun, F.; and Wang, Y. 2022. Multimodal token fusion for vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12186–12195.

Wu, J.; Gan, W.; Chen, Z.; Wan, S.; and Yu, P. S. 2023. Multimodal large language models: A survey. *arXiv preprint arXiv:2311.13165*.

Wu, Q.; Chen, X.; Zhou, Z.; and Zhang, J. 2020. Fedhome: Cloud-edge based personalized federated learning for in-home health monitoring. *IEEE Transactions on Mobile Computing*, 21(8): 2818–2832.

Wu, S.; Zhang, H. R.; and Ré, C. 2020. Understanding and improving information transfer in multi-task learning. *arXiv preprint arXiv:2005.00944*.

Wu, Z.; Song, S.; Khosla, A.; Yu, F.; Zhang, L.; Tang, X.; and Xiao, J. 2015. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1912–1920.

Xiong, B.; Yang, X.; Qi, F.; and Xu, C. 2022. A Unified Framework for Multi-modal Federated Learning. *Neurocomputing*.

Xu, M.; Song, C.; Tian, Y.; Agrawal, N.; Granqvist, F.; van Dalen, R.; Zhang, X.; Argueta, A.; Han, S.; Deng, Y.; et al. 2023a. Training Large-Vocabulary Neural Language Models by Private Federated Learning for Resource-Constrained Devices. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.

Xu, P.; Zhu, X.; and Clifton, D. A. 2022. Multimodal learning with transformers: A survey. *arXiv preprint arXiv:2206.06488*.

Xu, Z.; Zhang, Y.; Andrew, G.; Choquette-Choo, C. A.; Kairouz, P.; McMahan, H. B.; Rosenstock, J.; and Zhang, Y. 2023b. Federated Learning of Gboard Language Models with Differential Privacy. *arXiv preprint arXiv:2305.18465*.

Yang, C.; Zhu, M.; Liu, Y.; and Yuan, Y. 2023. FedPD: Federated Open Set Recognition with Parameter Disentanglement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4882–4891.

Ye, T.; Wei, S.; Cui, J.; Chen, C.; Fu, Y.; and Gao, M. 2023. Robust Clustered Federated Learning. In *International Conference on Database Systems for Advanced Applications*, 677–692. Springer.

Yu, L.; Miao, J.; Sun, X.; Chen, J.; Hauptmann, A.; Dai, H.; and Wei, W. 2023. DocumentNet: Bridging the Data Gap in Document Pre-training. In Wang, M.; and Zitouni, I., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, 707–722. Singapore: Association for Computational Linguistics.

Zhang, C.; Xie, Y.; Bai, H.; Yu, B.; Li, W.; and Gao, Y. 2021. A survey on federated learning. *Knowledge-Based Systems*, 216: 106775.

Zhang, Z.; Yang, Y.; Dai, Y.; Wang, Q.; Yu, Y.; Qu, L.; and Xu, Z. 2023. FedPETuning: When federated learning meets the parameter-efficient tuning methods of pre-trained language models. In *Annual Meeting of the Association of Computational Linguistics 2023*, 9963–9977. Association for Computational Linguistics (ACL).

Zheng, G.; Suo, Q.; Huai, M.; and Zhang, A. 2023. Learning to Learn Task Transformations for Improved Few-Shot Classification. In *Proceedings of the 2023 SIAM International Conference on Data Mining (SDM)*, 784–792. SIAM.

Zheng, G.; and Zhang, A. 2022. Knowledge-Guided Semantics Adjustment for Improved Few-Shot Classification. In *2022 IEEE International Conference on Data Mining (ICDM)*, 1347–1352. IEEE.