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

Jiayi Chen and 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 ovover 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.

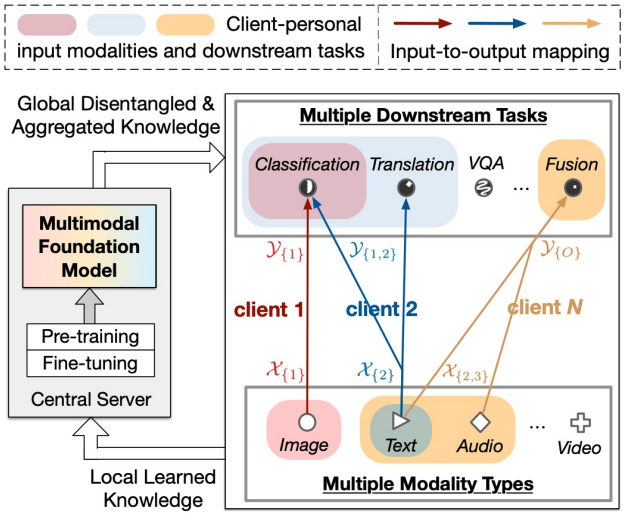


Figure 1: Modality-task Agnostic Federated Learning (AFL) for Privacy-preserving AGI. Clients learn personal models for their specific modalities and tasks using local data.

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 $q_i : \mathcal{X} \rightarrow \mathcal{Y}_i$ 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

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

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 $q_i(\widetilde{\boldsymbol{y}}|\widetilde{\boldsymbol{x}}) : \mathcal{X}_{\mathcal{I}_i} \rightarrow \mathcal{Y}_{\mathcal{O}_i}$ 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.

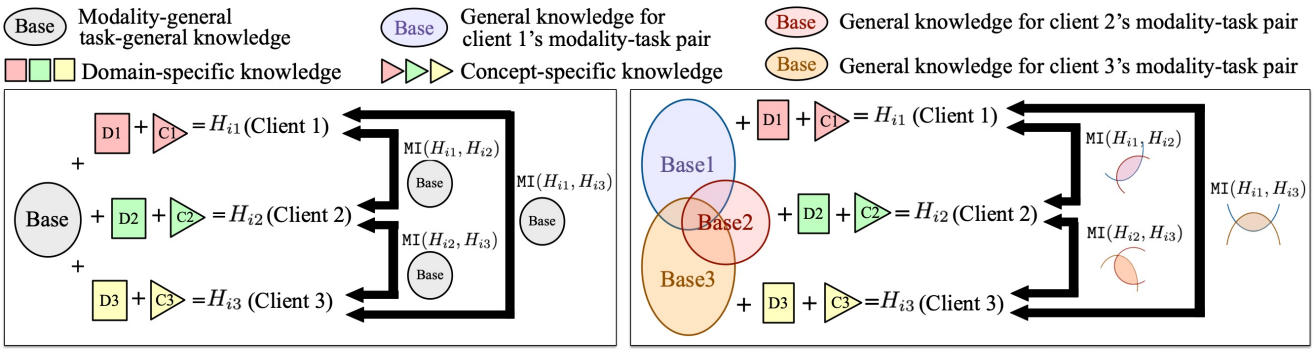
Asymmetrical Knowledge Transfer in AFL

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

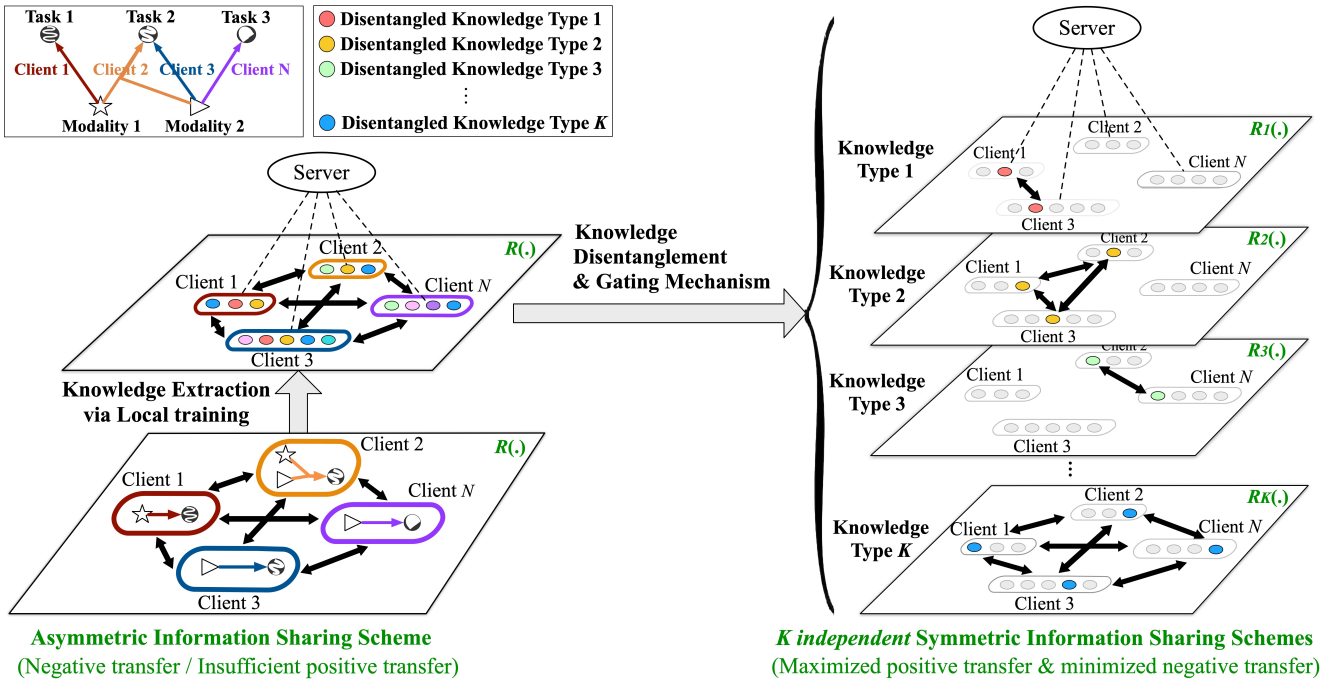


Figure 3: Overview of the proposed DisentAFL.

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

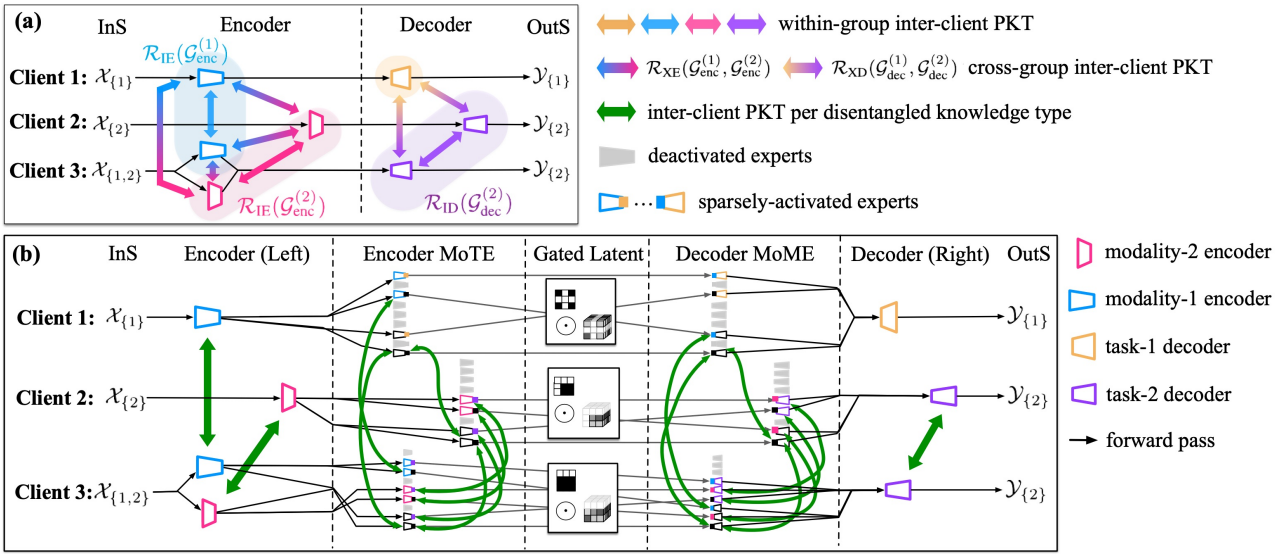


Figure 4: The two-stage Knowledge Disentanglement and Gating mechanism in DisentAFL. (a) The intermediate asymmetrical information sharing schemes after coarse-grained disentanglement; (b) The final symmetrical information sharing schemes after fine-grained disentanglement.

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.

Appendix

Brief Summary

As for the *information sharing scheme* among clients in FL, selecting accurate part of knowledge to transfer–i.e., *imposing all the aligned knowledge to be jointly learned (maximizing positive transfer) and avoiding all conflicting knowledge to be transferred (minimizing negative transfer)*, is crucial for both convergence and overall performance. However, real-world FL scenarios, such as the *newly introduced AFL* (cross-modal and cross-task FL) scenario with MTDC (**M**odality, **T**ask, **D**omain, and **C**oncept shifts) client heterogeneity, has *asymmetric knowledge relationships*–i.e., the common knowledge type(s) between each pair of clients are largely diversified. Consequently, the traditional FL methods tend to perform insufficient positive transfer, or, not be able to fully avoid negative transfer (**motivation**). We highlight the proposed DisentAFL to addresses this problem with the following points:

- DisentAFL is a new FL paradigm that can be used for improving traditional heterogeneous FL scenarios, and, in particular, for solving the newly introduced AFL scenario with more complex MTDC heterogeneity (**application**).
- DisentAFL is one of first attempts that seek to *explicitly* transfer sufficient positive knowledge while excluding negative knowledge, for better information sharing in FL (**contribution 1**).
- DisentAFL achieves the above goal by empowering the knowledge transfer strategy the ability of being aware of the true pairwise mutual knowledge type(s), through a *Knowledge Disentanglement and Gating* mechanism (**contribution 2**). In details, DisentAFL leverages a two-stage process to disentangle the *mixed knowledge* into *individual knowledge types*, where each unique knowledge type is captured by an *expert* network. In this way, we explicitly decompose the overall *asymmetric* information sharing scheme (with mixed knowledge) into K *symmetric* information sharing schemes (with disentangled knowledge types) $\mathcal{R}(\boldsymbol{w}_{1:N}) = \sum_{k=1}^{K} \mathcal{R}_k(\{\boldsymbol{w}_i^{(k)} | \forall i \in C_k \subseteq [N]\})$, where each $\mathcal{R}_k$ focuses on one knowledge type k. An overview of DisentAFL is shown in Figure 3.

Broader Impact

This research makes substantial advances in our modern life and can be used in the following **applications**: **(1)** ***Learning-at-home and Federated Artificial General Intelligence.*** Artificial General Intelligence (AGI) aims to train a general-purpose AI model to serve broader capabilities of intelligence, by seeking better generality over different modalities (image, text, video, audio, tabular, etc.) and different downstream tasks (i.e., classification, generation, ranking, QA, etc.). Considering the data privacy issue during achieving the centralized training of AGI, thinking about switching to decentralized, learning-at-home, and federated AGI, turns to be reasonable and more efficient. Learning-at-home and federated-AGI involves multiple users/workers jointly training a *general-purpose* AI model, where individual users/workers learn their own *customized* tasks (focusing the specific modality and task) anywhere at anytime, using their private data. Despite the importance, the field of Federated-AGI is still under-explored. Our work introduce the novel AFL to facilitate typical FL systems with unlimited modalities and categories of downstream tasks, and therefore, our work can be viewed as a step toward Federated-AGI. **(2)** ***Artificial Internet of Things (AIoT).*** Internet of Things have widely penetrated in different aspects of modern life and many intelligent IoT services and applications are emerging. AIoT applications often deploy different types of smart sensors or devices that generate data from different modalities (e.g., sensory, visual, and audio). For example, in one smart home, activities of a person can be recorded by body sensors in a smartwatch worn by the person, and also by a video camera in the room at the same time. Meanwhile, for smart homes with different device setups, some of them may have multimodal local data (i.e., multimodal clients) while the others may have unimodal local data (i.e., unimodal clients). With huge amounts of smart devices connected together in IoT, we are able to get access to massive user data to yield insights, train task-specified machine learning models and ultimately provide high-quality smart services and products. Despite of rare existing work, Federated AIoT will make substantial advances in all aspects of our modern life, including Internet of Medical Things (IoMT), Internet of Augmented Reality Things (IoART), intelligent transportation infrastructure, etc. **(3)** ***Collaborative Learning.*** Traditionally, it is hard to let different users, who focus on extremely diversified modalities and tasks, collaborate with each other due to their knowledge gap. With the help of the proposed method for AFL, different users with diversified modalities and tasks are able to explore any potential collaboration opportunity that can help each other to learn local tasks more efficiently and accurately.

Technically, the proposed DisentAFL is a new FL framework that can improve traditional heterogeneous FL scenarios, and, in particular, can achieve state-of-the-art results on the newly introduced AFL problem that have more complex client heterogeneity. ***Moreover***, DisentAFL is one of first attempts that seek to explicitly transfer sufficient positive knowledge while excluding negative knowledge, for better information sharing in FL. Our experiments have demonstrated the importance of understanding and supervising such positive and negative transfer behaviors during FL, which has been neglected by prior work. ***Finally***, we propose the first knowledge disentanglement and gating method for improving FL, which provides some explainability to the FL-learned models. We believe all these ideas may provide some helpful insights to future research in FL.

Global Knowledge Type Design

We aim to find the largest knowledge components that can sufficiently describe the global asymmetric PKT (Pairwise Knowledge Transfer) problem as the combination of several symmetric PKT problems.

The four sub-problems after the 1st-stage group-wise disentanglement give us the following inspirations: (1) First, to reduce the asymmetrical knowledge relationships of the subproblem-1 (TD heterogeneity), we investigate how a single modality serves different downstream tasks. Then, the

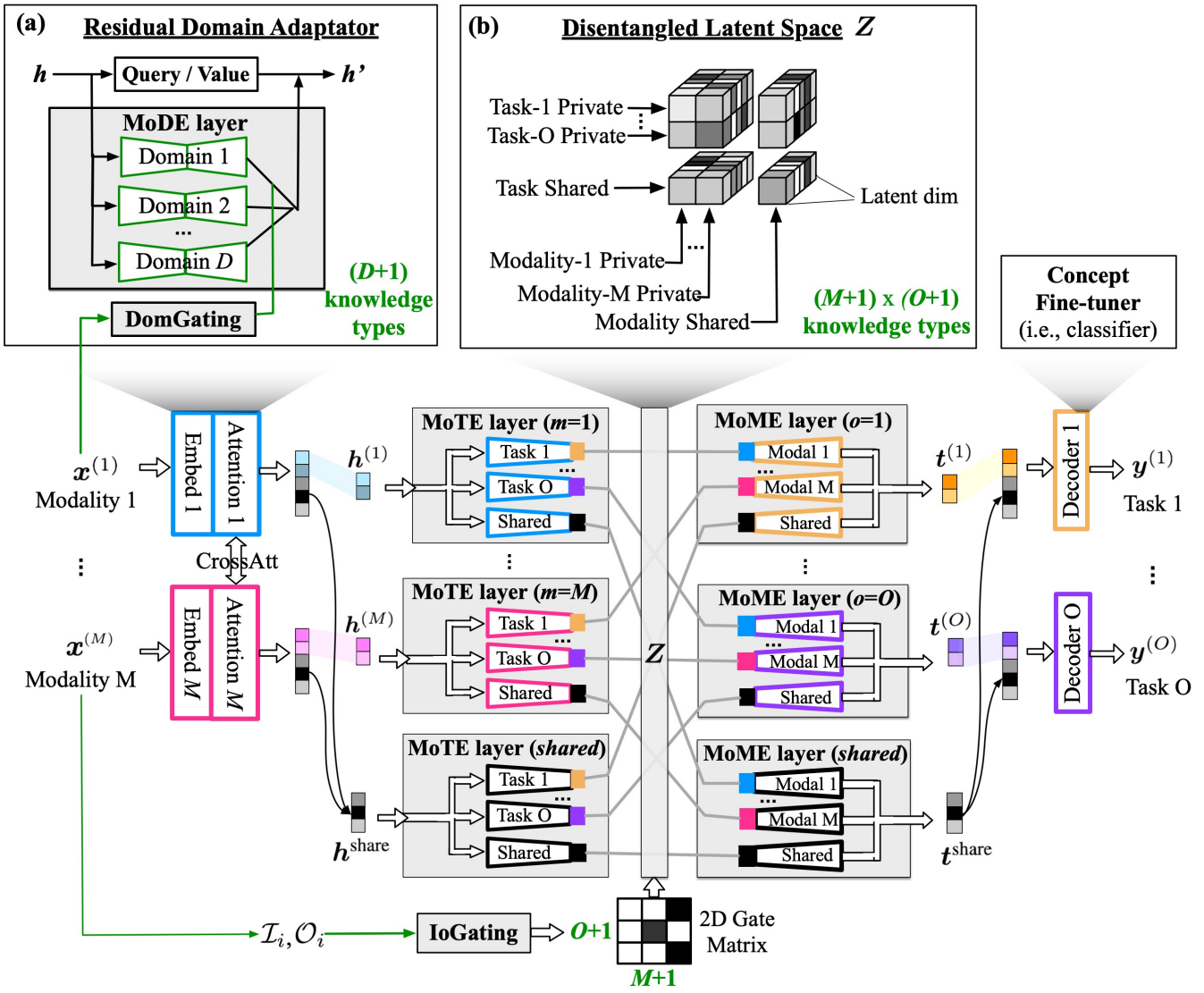


Figure 5: The super-network $\boldsymbol{w}^{\text{sup}}$ at the server, associated with the proposed DisentAFL method.

problem is that, how to disentangle the modality-specific knowledge into several knowledge types, each of which serves a different task? And, how to disentangle the domain-specific knowledge from domain-agnostic knowledge? (2) To reduce the asymmetrical knowledge relationships of the subproblem-2 (MC heterogeneity), we investigate how multiple modalities serve the same task. We need to disentangle the task-specific knowledge into several modality types, each of which receive information from a different modality, as well as to disentangle the concept-specific knowledge between clients? (3) To reduce the asymmetrical knowledge relationships of the subproblem-3 (MT heterogeneity), it is necessary to *first learn the common knowledge shared by different modalities* and then *disentangle the modality-shared knowledge into several knowledge types, each of which serves a different task*. (4) To reduce the asymmetrical knowledge relationships of the subproblem-4 (MT heterogeneity), it is necessary to *first learn the common knowledge shared by different task* and then *disentangle the task-shared knowledge into several knowledge types, each of which decodes a different modality*.

Inspired by them, we assume $K = M(D+1) + O(N+1) + (M+1)(O+1)$ fine-grained knowledge types over N clients can sufficiently describe the global asymmetric PKT problem as the combination of several symmetric PKT problems. For **concept shift** (C), we simply use fine-tuning for the last layer of the decoder. The last layer each decoder is considered as one knowledge type. We assume the early layers except the last layer of each decoder $o \in [O]$ is globally shareable by those clients that have task o. The last layer of each decoder is unique for each client. Therefore, a total of $O(N+1)$ concept knowledge over N clients will be needed. For **domain shift** (D), each domain should contain a domain-specific knowledge type and, in addition,

there is a single domain-agnostic based knowledge type, which results in $M(D+1)$ knowledge types related to domain shift of each modality. **Modality and Task shift** (MT) are correlated and more complex. for disentangling modality and task-related knowledge relationship asymmetric, we consider three levels of knowledge types as follows. First, for each modality, it can serve different tasks in different clients, and therefore, disentangling clients implies disentangling the **task-wise modality-specific knowledge**. Likewise, for each task, it can receive the information from different modalities in different clients, and therefore, disentangling clients implies disentangling the task-specific knowledge into **modality-wise task-specific knowledge**. More importantly, sub-problem-3/4 seeks any common knowledge shared by modalities and task to bridge the gap between clients that have neither similar modality nor similar tasks. Without shareable knowledge or without private knowledge types, the positive transfer is not maximized or the negative transfer can happen (see Appendix Figure 7). Therefore, it is straightforward to learn an additional **modality-shared knowledge type** and **task-shared knowledge type**. Over all the N clients with a total of O tasks, there would be $O+1$ independent types of knowledge: the **task-shared knowledge** and **task-private knowledge** per downstream task type. In summary, we need $(M+1)(O+1)$ knowledge types for disentangling MT heterogeneity.

Super-net at Server

In order to learn the K global fine-grained knowledge types mentioned above, we design a *wide and deep* super-network $\boldsymbol{w}^{\text{sup}}$ stored at the central server, whose overview is shown in Figure 5. The mission of $\boldsymbol{w}^{\text{sup}}$ is to accommodate K knowledge types. To this end, we design $\boldsymbol{w}^{\text{sup}}$ as follows. Basically, $\boldsymbol{w}^{\text{sup}}$ is a Multi-oath Multi-head Transformer-like architecture, consisting of M input channels for all seen modalities over clients and O output channels for all seen tasks over clients, where the interaction between modalities can be captured by cross-attention mechanism. On the basis of that, we introduce several modules and operations to learn fine-grained knowledge types from the model to further disentangling the asymmetric $\mathcal{R}_{\text{IE}}(\cdot)$, $\mathcal{R}_{\text{ID}}(\cdot)$, $\mathcal{R}_{\text{XE}}(\cdot,\cdot)$ and $\mathcal{R}_{\text{XD}}(\cdot,\cdot)$. Specifically, we adopt the idea of Mixture of Experts and replace several layers in the Transformer with Mixture of Domain Experts (**MoDE**) layers, Mixture of Task Experts (**MoTE**) layers, or Mixture of Modality Experts (**MoME**) layers. The main ideas are as follows.

- We propose the **MoDE** layer (denoted as $\boldsymbol{\phi}_{\text{mode}}$) to capture D **domain-specific** knowledge types. It consists of D parallel expert models, where each expert model (denoted as $\texttt{MoDE}_d(\cdot;\boldsymbol{\phi}^d_{\text{mode}})$) in MoDE stands for the knowledge type for a specific domain d. MoDE acts as a residual connection attached to an original model block, which we treat as the **domain-agnostic** knowledge (as show in Figure 5(a)). In practice, we apply MoDE layers to the query and value linear layers.
- Suppose $\boldsymbol{\phi}^{(m)}_{\text{left}}$ denotes a combination of modality-m's embedding layers, attention layers, and MoDE layers (as shown as the colored-contour rectangles on the left of Figure 5). Let $\mathcal{H}^{(m)}$ denote the feature space learned by $\boldsymbol{\phi}^{(m)}_{\text{left}}$ (after necessary cross-attention multimodal interaction). $\mathcal{H}^{(m)}$ contains the modality-m specific information. It is straightforward different modalities' output space $\mathcal{H}^{(m)} \neq \text{H}^{(m')}$ contains complementary and heterogeneous information that describe different aspects/views of any object. However, in order to maximize positive transfer between different modalities that may contain common knowledge, especially in the asymmetric $\mathcal{R}_{\text{XE}}(\cdot)$ (cross-modal multi-task) problem, the modality-shared information should be learned as well. Therefore, we follow (Lee and Pavlovic 2021) and split the modality-m's output feature space as $\mathcal{H}^{(m)}$ into two types of information: $[\boldsymbol{h}^{\text{share}}||\boldsymbol{h}^{(m)}] = \boldsymbol{h}'^{(m)} \in \mathcal{H}^{(m)}$, where $||$ denotes concatenation operation, $\boldsymbol{h}^{(m)} \in \mathbb{R}^{d^{\text{modal}}_m}$ represents the **modality-private knowledge** that is not shared with the other modalities and $\boldsymbol{h}^{\text{share}} \in \mathbb{R}^{d^{\text{modal}}_{\text{share}}}$ represents the **modality-shared knowledge** type.
- Let $\mathcal{T}^{(o)}$ denote the input feature space of the task o's decoder network (denoted as $\boldsymbol{\theta}^{(o)}_{\text{right}}$). $\mathcal{T}^{(o)}$ contains the information that specifically used for prediction or decision making for the downstream task o. It is straightforward different tasks' input spaces $\mathcal{T}^{(o)} \neq \mathcal{T}^{(o')}$ contains diversified information; for example, the information for classification should be different from that used for segmentation. However, in order to maximize positive transfer between different tasks that may contain common knowledge, especially in the asymmetric $\mathcal{R}_{\text{XD}}(\cdot)$ (multi-modal cross-task) problem, it is beneficial to leverage any task-shared information as well. Therefore, we assume $\mathcal{T}^{(o)}$ is a fused space by combining a task-private and a task-shared feature space, $[\boldsymbol{t}^{\text{share}}||\boldsymbol{t}^{(o)}] = \boldsymbol{t}'^{(o)} \in \mathcal{T}^{(o)}$. where $\boldsymbol{t}^{(o)} \in \mathbb{R}^{d^{\text{task}}_o}$ represents the **task-private knowledge** that is not shared with the other tasks and $\boldsymbol{t}^{\text{share}} \in \mathbb{R}^{d^{\text{task}}_{\text{share}}}$ represents the **task-shared knowledge**.
- The asymmetrical problems–the single-modality *multi-task* $\mathcal{R}_{\text{IE}}$ and the *cross-modality multi-task* $\mathcal{R}_{\text{XE}}$, both seek how one modality-level knowledge type (i.e., modality-private or modality-shared) can serve diverse tasks. The reason that may cause asymmetrical knowledge relationships in $\mathcal{R}_{\text{IE}}$ or $\mathcal{R}_{\text{XE}}$ is the downstream task identity–those clients that share more downstream tasks should transfer more knowledge then other pairs with less/no common downstream task. To this end, we propose the **MoTE** layer (denoted as $\boldsymbol{\phi}_{\text{mote}}$) to capture $(O+1)$ **task-related knowledge type for each types of modality-private/shared knowledge**. In the server's super-network, there are a total $(M+1)$ MoTE layers: $\boldsymbol{\phi}^{(1)}_{\text{mote}}, \boldsymbol{\phi}^{(2)}_{\text{mote}}, ..., \boldsymbol{\phi}^{(M)}_{\text{mote}}, \boldsymbol{\phi}^{(\text{share})}_{\text{mote}}$, where $\boldsymbol{\phi}^{(m)}_{\text{mote}}$ denote the MoTE layer for modality-m's private knowledge and $\boldsymbol{\phi}^{(\text{share})}_{\text{mote}}$ denote the MoTE layer for modality-shared knowledge. The MoTE layer of each modality m contains $(O+1)$ expert models $\boldsymbol{\phi}^{(m)}_{\text{mote}} = \{\boldsymbol{\phi}^{(m)1}_{\text{mote}}, \boldsymbol{\phi}^{(m)2}_{\text{mote}}, ..., \boldsymbol{\phi}^{(m)O}_{\text{mote}}, \boldsymbol{\phi}^{(m)\text{share}}_{\text{mote}}\}$. As shown in Figure 5(b), the output of all MoTE layers

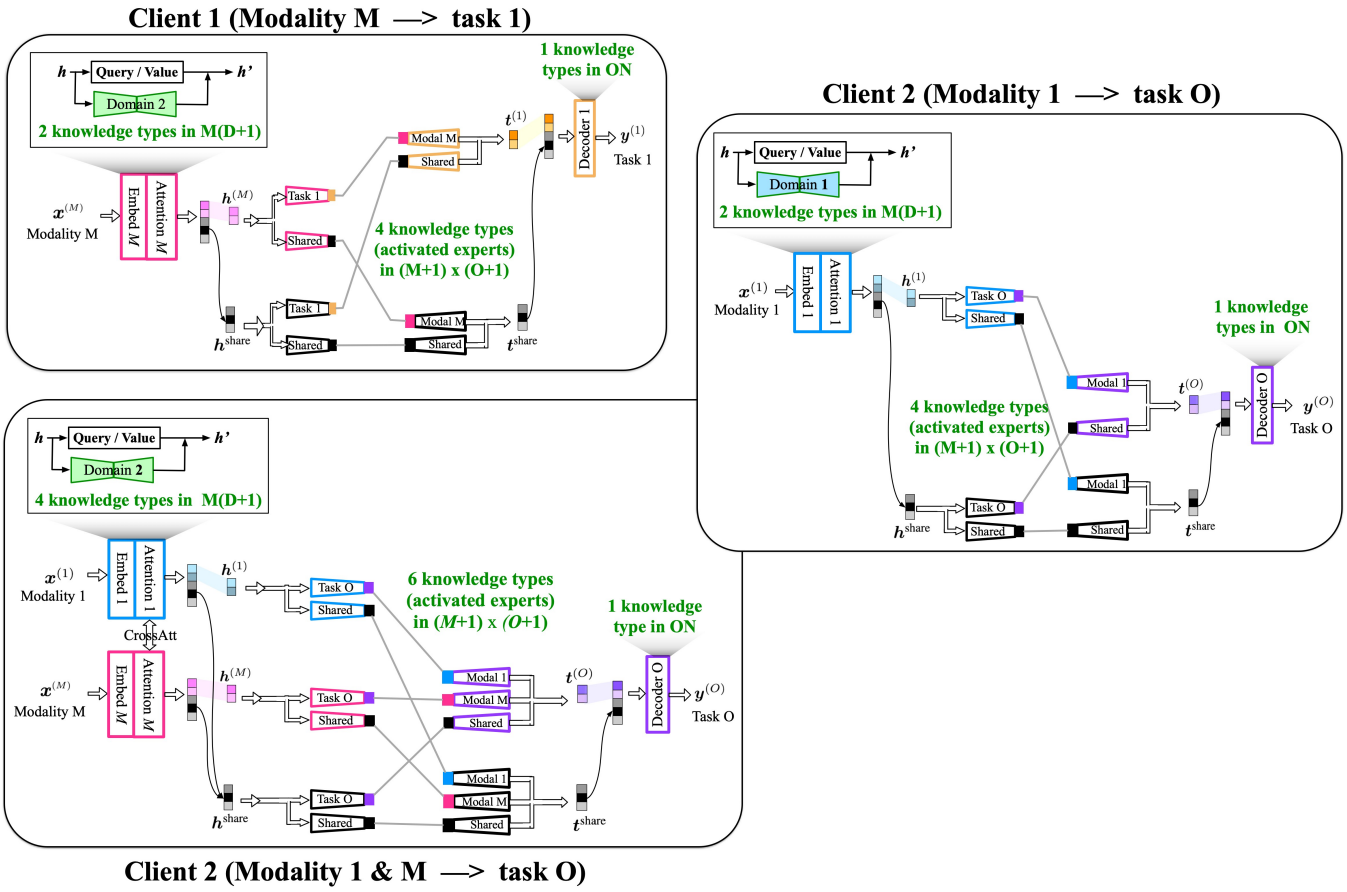


Figure 6: The routed network structures at three example clients.

of the server network can be represented as a tensor $\boldsymbol{Z} \in \mathbb{R}^{(M+1)\times(O+1)\times d^{\text{latent}}}$ consists of $(M+1)(O+1)$ features from each of the expert models, where d^{latent} is the feature dimensional of each knowledge type.

- The asymmetrical problems–the *multi-modality* single-task $\mathcal{R}_{\text{ID}}$ and the *multi-modality cross-task* $\mathcal{R}_{\text{XD}}$, both seek how one task-level knowledge type (i.e., task-private or task-shared) can be learned from diverse tasks. The reason that may cause asymmetrical knowledge relationships in $\mathcal{R}_{\text{ID}}$ or $\mathcal{R}_{\text{XD}}$ is the incoming modality identity–those clients that share more modality types should transfer more knowledge then other pairs with less/no common modalities types. To this end, We propose the **MoME** layer (denoted as $\boldsymbol{\theta}_{\text{mome}}$) to capture $(M+1)$ **modality-related knowledge types for each type of task-private/shared knowledge**. Similar to MoTE, we denote the MoME layer for task-o's private knowledge as $\boldsymbol{\theta}^{(o)}_{\text{mome}} = \{\boldsymbol{\theta}^{(o)1}_{\text{mome}}, \boldsymbol{\theta}^{(o)2}_{\text{mome}}, ..., \boldsymbol{\theta}^{(o)M}_{\text{mome}}, \boldsymbol{\theta}^{(o)\text{share}}_{\text{mome}}\}$. The order of inputs feed to all the MoME layers of the server network is the transposed $\boldsymbol{Z}^{\top} \in \mathbb{R}^{(O+1)\times(M+1)\times d^{\text{latent}}}$.

Sparsely-gated Client Network Routing

Each client's local network $\boldsymbol{w}_i$ is a sparsely-activated version of $\boldsymbol{w}^{\text{sup}}$. Formally, each client has two auxiliary ***gating functions***: **(1)** $\texttt{IoGate}(\cdot)$ takes the input samples or the 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 any client with any modality-task pair learn the task-shared and modality shared knowledge, which bridge the gap between a pair of clients with $\mathcal{I}_i \cap \mathcal{I}_{i'} = \emptyset \wedge \mathcal{O}_i \cap \mathcal{O}_{i'} = \emptyset$. **(2)** $\texttt{DomGate}(\cdot; \psi)$ takes the input samples and produces a D-dimensional binary vector $\boldsymbol{g}_i \in \{0,1\}^D$, where $D <= N$ denotes the pre-defined number of domains over clients and ψ is the parameters of the function.

The binary outputs of the two gating functions $\boldsymbol{S}_i, \boldsymbol{g}_i$ are used to ***route*** each client's network through the supernetwork $\boldsymbol{w}_i = \texttt{ROUTE}(\boldsymbol{S}_i, \boldsymbol{g}_i; \boldsymbol{w}^{\text{sup}})$. The details of module routing are as follows.

- MoDE layers are gated by the one-hot vector $\boldsymbol{g}_i$. Suppose the input feature is $\boldsymbol{h}$. The output of any MoDE layer is $\Delta\boldsymbol{h} = \sum_{d=1}^{D} \boldsymbol{g}_{i,d} \texttt{MoDE}_d(\boldsymbol{h}; \boldsymbol{\phi}^d_{\text{mode},i})$, which will be added to the original model's output. Since $\boldsymbol{g}_i$ is one hot,

only one expert is activated and receives gradients from backward propagation. That is, only one **domain's specific knowledge** is learned at each client. The learned $\boldsymbol{g}_i$ is treated as a guess of the domain identity of client i.

- Modality encoders and task-specific decoders are gated by $\mathcal{I}_i$ and $\mathcal{O}_i$, respectively. The client i's network contains encoders $\{\boldsymbol{\phi}^{(m)}_{\text{left}}\}_{m\in\mathcal{I}_i}$ and decoders $\{\boldsymbol{\theta}^{(o)}_{\text{right}}\}_{o\in\mathcal{O}_i}$. Accordingly, the client learns $|\mathcal{I}_i|$ **modality-private** knowledge types, one **modality-shared** knowledge type, $|\mathcal{O}_i|$ **task-private** knowledge types, and one **task-shared** knowledge type.
- MoTE and MoME layers are routed by the gate matrix $\boldsymbol{S}_i$. Given a sample, the client model learns $(|\mathcal{I}_i|+1)\times(|\mathcal{O}_i|+1)$ **modality-task interactive** knowledge types, which are represented in a sparse tensor $\boldsymbol{Z}\in\mathbb{R}^{(M+1)\times(O+1)\times d^{\text{latent}}}$, where only $(|\mathcal{I}_i|+1)\times(|\mathcal{O}_i|+1)\times d^{\text{latent}}$ values in this tensor is activated and requires back-propagation. Examples are shown in Figure 4(c). For example, given a modality-private feature $\boldsymbol{h}^{(m)}$, the outputs of an MoTE layer is $|\mathcal{O}_i|+1$ features; each expert $o\in\mathcal{O}_i$ outputs $\boldsymbol{Z}_{o,m,\cdot}=\texttt{MoTE}^o_m(\boldsymbol{h}^{(m)};\boldsymbol{\phi}^{(m)o}_{\text{mote},i})$. After $\boldsymbol{Z}$ is learned, MoME layers reconstruct the task-private/shared knowledge from $\boldsymbol{Z}$. Specifically, the task o's private and task-shared knowledge are decoded as $\boldsymbol{t}^{(o)}=\sum_{m=1}^{M+1}\boldsymbol{S}_{i,o,m}\texttt{MoME}^m_o(\boldsymbol{Z}_{o,m,\cdot};\boldsymbol{\theta}^{(o)m}_{\text{mome},i})$ and $\boldsymbol{t}^{\text{share}}=\sum_{m=1}^{M+1}\boldsymbol{S}_{i,o,m}\texttt{MoME}^m_{\text{share}}(\boldsymbol{Z}_{o,m,\cdot};\boldsymbol{\theta}^{(\text{share})m}_{\text{mome},i})$.

Through the sparse routing process, the number of knowledge types that can be disentangled from client i's model is $K_i=2|\mathcal{I}_i|+2|\mathcal{O}_i|+(|\mathcal{I}_i|+1)(|\mathcal{O}_i|+1)$. The collection of client-specific knowledge types is a subset of globally shareable knowledge types. Figure 6 shows three example clients and their client-specific knowledge types.

In practice, $\boldsymbol{S}_i,\boldsymbol{g}_i$ is very sparse, i.e., $K_i << K$. Hence the client network $\boldsymbol{w}_i$ is much thinner than $\boldsymbol{w}^{\text{sup}}$ and does not exceed the local memory constraint. Also, the number of activated MoTE and MoME experts depends on $\mathcal{I}_i,\mathcal{O}_i$, thus a larger M and O over the large-scale clients have no influence on the size of the local model $\boldsymbol{w}_i$.

Decomposition of Information Sharing Scheme through Knowledge Disentanglement

We here prove that our fine-grained knowledge disentanglement can therefore decompose the *original asymmetric client relationships* into $K=M(D+1)+O(N+1)+(M+1)(O+1)$ independent *symmetric client relationships*, that is

$$\begin{aligned}&\mathcal{R}(\boldsymbol{w}_i|i\in[N])\\ =\;&\textstyle\sum_{m\in[M]}\mathcal{R}^m_{\text{IE}}(\mathcal{G}^{(m)}_{\text{enc}})+\sum_{o\in[O]}\mathcal{R}^o_{\text{ID}}(\mathcal{G}^{(o)}_{\text{dec}})\\ &\textstyle+\sum_{m,m'\in[M],m\neq m'}\mathcal{R}^{m,m'}_{\text{XE}}(\mathcal{G}^{(m)}_{\text{enc}},\mathcal{G}^{(m')}_{\text{enc}})\\ &\textstyle+\sum_{o,o'\in[O],o\neq o'}\mathcal{R}^{o,o'}_{\text{XD}}(\mathcal{G}^{(o)}_{\text{dec}},\mathcal{G}^{(o')}_{\text{dec}})\\ =\;&\textstyle\sum_{k=1}^{K}\mathcal{R}_k(\boldsymbol{w}^{(k)}_i|i\in C_k\subseteq[N]).\end{aligned}\tag{5}$$

where $\mathcal{G}^{(m)}_{\text{enc}}=\{\boldsymbol{\phi}^{(m)}_i|i\in[N],m\in\mathcal{I}_i\}$ and $\mathcal{G}^{(o)}_{\text{dec}}=\{\boldsymbol{\theta}^{(o)}_i|i\in[N],o\in\mathcal{O}_i\}$. The notation $\boldsymbol{\phi}^{(m)}_i=\{\boldsymbol{\phi}^{(m)}_{\text{BE},i},\boldsymbol{\phi}^{(m)}_{\text{mode},i},\boldsymbol{\phi}^{(m)}_{\text{mote},i},\boldsymbol{\phi}^{(\text{share})}_{\text{mote},i}\}$ consists of the modality m's encoder learned at client i consisting of embedding, attention layers, MoDE layers, and the MoTE layers for modality-m, as well as the MoTE layers for modality-shared information. The notation $\boldsymbol{\theta}^{(o)}_i=\{\boldsymbol{\theta}^{(o)}_{\text{mome},i},\boldsymbol{\theta}^{(\text{share})}_{\text{mome},i},\boldsymbol{\theta}^{(o)}_{\text{BD},i},\boldsymbol{\theta}^{(o)}_{\text{final},i}\}$ consists of the task-o MoME layers, the MoME layers for task-shared information, and the task o's decoder consisting of early layers and the final concept layers, which are learned at client i. The mixed knowledge captured in $\boldsymbol{w}_i=\{\boldsymbol{\phi}^{(m)}_i|m\in\mathcal{I}_i\}\cup\{\boldsymbol{\theta}^{(o)}_i|o\in\mathcal{O}_i\}$.

Knowledge Split We define $(M+O+MO)$ subsets of clients: (1) For each modality type m, we define a subset of clients that contain the modality m, that is, $C^m_{\text{modal}}=\{i|i\in[N],m\in\mathcal{I}_i\}$; (2) For each downstream task type o, we define a subset of clients that contain the task o, that is, $C^o_{\text{task}}=\{i|i\in[N],o\in\mathcal{O}_i\}$; and, (3) For each modality-task pair (m,o), we define a subset of clients that share both the modality m and the task o, that is, $C^{m,o}_{\text{pair}}=\{i|i\in[N],m\in\mathcal{I}_i\wedge o\in\mathcal{O}_i\}$.

Each asymmetric term $\mathcal{R}^m_{\text{IE}}(\mathcal{G}^{(m)}_{\text{enc}})$ can be split into $K^m_{\text{IE}}=(D+1)+2(O+1)$ symmetric terms

$$\begin{aligned}&\mathcal{R}^m_{\text{IE}}(\mathcal{G}^{(m)}_{\text{enc}})\\ =\;&\mathcal{R}^m_{\text{IE}}\left(\boldsymbol{\phi}^{(m)}_i|i\in[N],m\in\mathcal{I}_i\right)\\ =\;&\mathcal{R}^m_{\text{BE}}\left(\boldsymbol{\phi}^{(m)}_{\text{BE},i}|i\in[N],m\in\mathcal{I}_i\right)\\ &+\textstyle\sum_{d=1}^{D}\mathcal{R}^{m,d}_{\text{MoDE}}\left(\boldsymbol{\phi}^{(m)d}_{\text{mode},i}|i\in[N],g_{i,d}=1\right)\\ &+\mathcal{R}^{m,\text{share}}_{\text{MoTE}}\left(\boldsymbol{\phi}^{(m)\text{share}}_{\text{mote},i}|i\in[N],m\in\mathcal{I}_i\right)\\ &+\textstyle\sum_{o=1}^{O}\mathcal{R}^{m,o}_{\text{MoTE}}\left(\boldsymbol{\phi}^{(m)o}_{\text{mote},i}|i\in[N],m\in\mathcal{I}_i\wedge o\in\mathcal{O}_i\right)\\ &+\mathcal{R}^{\text{share,share}}_{\text{MoTE}}\left(\boldsymbol{\phi}^{(\text{share})\text{share}}_{\text{mote},i}|i\in[N],m\in\mathcal{I}_i\right)\\ &+\textstyle\sum_{o=1}^{O}\mathcal{R}^{\text{share},o}_{\text{MoTE}}\left(\boldsymbol{\phi}^{(\text{share})o}_{\text{mote},i}|i\in[N],m\in\mathcal{I}_i\wedge o\in\mathcal{O}_i\right)\\ =\;&\mathcal{R}^m_{\text{BE}}\left(\boldsymbol{\phi}^{(m)}_{\text{BE},i}|i\in C^m_{\text{modal}}\right)\\ &+\textstyle\sum_{d=1}^{D}\mathcal{R}^{m,d}_{\text{MoDE}}\left(\boldsymbol{\phi}^{(m)d}_{\text{mode},i}|i\in C^d_{\text{MoDE}}\right)\\ &+\mathcal{R}^{m,\text{share}}_{\text{MoTE}}\left(\boldsymbol{\phi}^{(m)\text{share}}_{\text{mote},i}|i\in C^m_{\text{modal}}\right)\\ &+\textstyle\sum_{o=1}^{O}\mathcal{R}^{m,o}_{\text{MoTE}}\left(\boldsymbol{\phi}^{(m)o}_{\text{mote},i}|i\in C^{m,o}_{\text{pair}}\right)\\ &+\mathcal{R}^{\text{share,share}}_{\text{MoTE}}\left(\boldsymbol{\phi}^{(\text{share})\text{share}}_{\text{mote},i}|i\in C^m_{\text{modal}}\right)\\ &+\textstyle\sum_{o=1}^{O}\mathcal{R}^{\text{share},o}_{\text{MoTE}}\left(\boldsymbol{\phi}^{(\text{share})o}_{\text{mote},i}|i\in C^{m,o}_{\text{pair}}\right),\end{aligned}\tag{6}$$

where $\mathcal{R}^m_{\text{BE}}(\cdot)$ denotes the information sharing over the ***modality-m domain-agnostic*** knowledge type of a subset of clients that contain the modality m, that is, C^m_{modal}. $\mathcal{R}^d_{\text{MoDE}}(\cdot)$ denotes the information sharing over the ***domain-d-specific*** knowledge across a subset of clients that belong to domain d, that is, $C^d_{\text{MoDE}}=\{i|i\in[N],g_{i,d}=1\}$. $\mathcal{R}^{m,\text{share}}_{\text{MoTE}}(\cdot)$ denotes the information sharing over the ***modality-m-private and task-shared*** knowledge across a subset of clients that contain the modality m, that is, C^m_{modal}. $\mathcal{R}^{m,o}_{\text{MoTE}}(\cdot)$ denotes the information sharing over the ***modality-m-private and task-o-private*** knowledge across a subset of clients that share both the modality m and task o, that is, $C^{m,o}_{\text{pair}}$. $\mathcal{R}^{\text{share,share}}_{\text{MoTE}}(\cdot)$ de-

notes the information sharing over the ***modality-shared and task-shared*** knowledge across the group C^m_{modal}. $\mathcal{R}^{\text{share},o}_{\text{MoTE}}(\cdot)$ denotes the information sharing over the ***modality-shared and task-o-private*** knowledge across $C^{m,o}_{\text{pair}}$.

Each asymmetric term $\mathcal{R}^o_{\text{ID}}(\mathcal{G}^{(o)}_{\text{dec}})$ can be split into $K^o_{\text{ID}} = (N+1)+2(M+1)$ symmetric terms:

$$\begin{aligned}
&\mathcal{R}^o_{\text{ID}}(\mathcal{G}^{(o)}_{\text{dec}}) \\
=\ &\mathcal{R}^o_{\text{ID}}\left(\boldsymbol{\theta}^{(o)}_i | i\in[N], o\in\mathcal{O}_i\right) \\
=\ &\mathcal{R}^o_{\text{BD}}\left(\boldsymbol{\theta}^{(o)}_{\text{BD},i} | i\in[N], o\in\mathcal{O}_i\right) \\
&+\textstyle\sum_{i=1}^N \mathcal{R}^{o,i}_{\text{concept}}\left(\boldsymbol{\theta}^{(o)}_{\text{final},i} | \{i\} \text{ if } o\in\mathcal{O}_i \text{ else } \emptyset\right) \\
&+\mathcal{R}^{\text{share},o}_{\text{MoME}}\left(\boldsymbol{\theta}^{(o)\text{share}}_{\text{mome},i} | i\in[N], o\in\mathcal{O}_i\right) \\
&+\textstyle\sum_{m=1}^M \mathcal{R}^{m,o}_{\text{MoME}}\left(\boldsymbol{\theta}^{(o)m}_{\text{mome},i} | i\in[N], m\in\mathcal{I}_i\wedge o\in\mathcal{O}_i\right) \\
&+\mathcal{R}^{\text{share,share}}_{\text{MoME}}\left(\boldsymbol{\theta}^{(\text{share})\text{share}}_{\text{mome},i} | i\in[N], o\in\mathcal{O}_i\right) \\
&+\textstyle\sum_{m=1}^M \mathcal{R}^{m,\text{share}}_{\text{MoME}}\left(\boldsymbol{\theta}^{(\text{share})m}_{\text{mome},i} | i\in[N], m\in\mathcal{I}_i\wedge o\in\mathcal{O}_i\right) \\
=\ &\mathcal{R}^o_{\text{BD}}\left(\boldsymbol{\theta}^{(o)}_{\text{BD},i} | i\in C^o_{\text{task}}\right) \\
&+\textstyle\sum_{i=1}^N \mathcal{R}^{o,i}_{\text{concept}}\left(\boldsymbol{\theta}^{(o)}_{\text{final},i} | \{i\} \text{ if } o\in\mathcal{O}_i \text{ else } \emptyset\}\right) \\
&+\mathcal{R}^{\text{share},o}_{\text{MoME}}\left(\boldsymbol{\theta}^{(o)\text{share}}_{\text{mome},i} | i\in C^o_{\text{task}}\right) \\
&+\textstyle\sum_{m=1}^M \mathcal{R}^{o,m}_{\text{MoME}}\left(\boldsymbol{\theta}^{(o)m}_{\text{mome},i} | i\in C^{m,o}_{\text{pair}}\right) \\
&+\mathcal{R}^{\text{share,share}}_{\text{MoME}}\left(\boldsymbol{\theta}^{(\text{share})\text{share}}_{\text{mome},i} | i\in C^o_{\text{task}}\right) \\
&+\textstyle\sum_{m=1}^M \mathcal{R}^{\text{share},m}_{\text{MoME}}\left(\boldsymbol{\theta}^{(\text{share})m}_{\text{mome},i} | i\in C^{m,o}_{\text{pair}}\right),
\end{aligned} \tag{7}$$

where $\mathcal{R}^o_{\text{BD}}(\cdot)$ denotes the information sharing over the ***task-o label-agnostic*** knowledge type of a subset of clients that contain the task o, that is, C^o_{task}. $\mathcal{R}^{o,i}_{\text{concept}}(\cdot)$ denotes the **task-o's concept that is specific on the client** i. $\mathcal{R}^{\text{share},o}_{\text{MoME}}(\cdot)$ denotes the information sharing over the ***task-o-private and modality-shared*** knowledge across a subset of clients that contain the task o, that is, C^o_{task}. $\mathcal{R}^{o,m}_{\text{MoME}}(\cdot)$ denotes the information sharing over the ***task-o-private and modality-m-private*** knowledge across a subset of clients that share both the modality m and task o, that is, $C^{m,o}_{\text{pair}}$. $\mathcal{R}^{\text{share,share}}_{\text{MoME}}(\cdot)$ denotes the information sharing over the ***task-shared and modality-shared*** knowledge across the group. $\mathcal{R}^{\text{share},m}_{\text{MoME}}(\cdot)$ denotes the information sharing over the ***task-shared and modality-m-private*** knowledge across a subset of clients that share both the modality m and task o, that is, $C^{m,o}_{\text{pair}}$.

Each cross-group asymmetric term $\mathcal{R}^{m,m'}_{\text{XE}}(\mathcal{G}^{(m)}_{\text{enc}}, \mathcal{G}^{(m')}_{\text{enc}})$ with $m\neq m'$ can be split into $K^{m,m'}_{\text{XE}} = O+1$ symmetric terms

$$\begin{aligned}
&\mathcal{R}^{m,m'}_{\text{XE}}(\mathcal{G}^{(m)}_{\text{enc}}, \mathcal{G}^{(m')}_{\text{enc}}) \\
=\ &\mathcal{R}^{m,m'}_{\text{XE}}\left(\{\boldsymbol{\phi}^{(m)}_i | i\in[N], m\in\mathcal{I}_i\}, \{\boldsymbol{\phi}^{(m')}_i | i\in[N], m'\in\mathcal{I}_i\}\right) \\
=\ &\mathcal{R}^{\text{share,share}}_{\text{MoTE}}\left(\boldsymbol{\phi}^{(\text{share})\text{share}}_{\text{mote},i}, \boldsymbol{\phi}^{(\text{share})\text{share}}_{\text{mote},i'} | \forall(i,i')\in C^m_{\text{modal}}\times C^{m'}_{\text{modal}}\right) \\
&+\textstyle\sum_{o=1}^O \mathcal{R}^{\text{share},o}_{\text{MoTE}}\left(\boldsymbol{\phi}^{(\text{share})o}_{\text{mote},i}, \boldsymbol{\phi}^{(\text{share})o}_{\text{mote},i'} | \forall(i,i')\in C^{m,o}_{\text{pair}}\times C^{m',o}_{\text{pair}}\right)
\end{aligned} \tag{8}$$

where $\mathcal{R}^{\text{share,share}}_{\text{MoTE}}(\cdot)$ denotes the information sharing over the ***task-shared and modality-shared*** knowledge across **all** pairs of clients from separate modality groups, and $\mathcal{R}^{\text{share},o}_{\text{MoTE}}(\cdot)$ denotes the information sharing over the ***task-o-private and modality-shared*** knowledge across **all** pairs of clients from separate modality groups as well as sharing the task o.

Each cross-group asymmetric term $\mathcal{R}^{o,o'}_{\text{XD}}(\mathcal{G}^{(o)}_{\text{dec}}, \mathcal{G}^{(o')}_{\text{dec}})$ with $o\neq o'$ can be split into $K^{o,o'}_{\text{XD}} = M+1$ symmetric schemes

$$\begin{aligned}
&\mathcal{R}^{o,o'}_{\text{XD}}(\mathcal{G}^{(o)}_{\text{dec}}, \mathcal{G}^{(o')}_{\text{dec}}) \\
=\ &\mathcal{R}^{o,o'}_{\text{XD}}\left(\{\boldsymbol{\theta}^{(o)}_i | i\in[N], o\in\mathcal{O}_i\}, \{\boldsymbol{\theta}^{(o')}_i | i\in[N], o'\in\mathcal{O}_i\}\right) \\
=\ &\mathcal{R}^{\text{share,share}}_{\text{MoME}}\left(\boldsymbol{\theta}^{(\text{share})\text{share}}_{\text{mome},i}, \boldsymbol{\theta}^{(\text{share})\text{share}}_{\text{mome},i'} | \forall(i,i')\in C^o_{\text{task}}\times C^{o'}_{\text{task}}\right) \\
&+\textstyle\sum_{m=1}^M \mathcal{R}^{\text{share},m}_{\text{MoME}}\left(\boldsymbol{\theta}^{(\text{share})m}_{\text{mome},i}, \boldsymbol{\theta}^{(\text{share})m}_{\text{mome},i'} | \forall(i,i')\in C^{m,o}_{\text{pair}}\times C^{m,o'}_{\text{pair}}\right)
\end{aligned} \tag{9}$$

where $\mathcal{R}^{\text{share,share}}_{\text{MoME}}(\cdot)$ denotes the information sharing over the ***task-shared and modality-shared*** knowledge across **all** pairs of clients from separate task groups, and $\mathcal{R}^{\text{share},m}_{\text{MoME}}(\cdot)$ denotes the information sharing over the ***task-shared and modality-m-private*** knowledge across **all** pairs of clients from separate task groups as well as sharing the same input modality m.

Knowledge Independence Considering Eq.(6) for each of M modalities, the terms $\mathcal{R}^{\text{share},1}_{\text{MoME}}, \mathcal{R}^{\text{share},2}_{\text{MoME}}, ..., \mathcal{R}^{\text{share},O}_{\text{MoME}}$, $\mathcal{R}^{m,1}_{\text{MoME}}, \mathcal{R}^{m,2}_{\text{MoME}}, ..., \mathcal{R}^{m,O}_{\text{MoME}}, \mathcal{R}^{\text{share,share}}_{\text{MoME}}$, and $\mathcal{R}^{m,\text{share}}_{\text{MoME}}$ are encouraged to be independent with each other through the auxiliary orthogonal loss over the latent space $\mathcal{Z}$. When adding up Eq.(6) over all modalities, the modality-specific terms in different modalities $\mathcal{R}^{1,o}_{\text{MoME}}, \mathcal{R}^{2,o}_{\text{MoME}}, ..., \mathcal{R}^{M,o}_{\text{MoME}}$ are independent due to the disentanglement of the latent space; however, the $(O+1)$ modality-shared terms in different modalities, $\mathcal{R}^{\text{share},o}_{\text{MoME}}$ and $\mathcal{R}^{\text{share,share}}_{\text{MoME}}$ are not independent and can be combined. The independence between domain-agnostic and domain-specific knowledge types can be done as first-order meta learning, where domain-agnostic parameters are treated as the meta-parameters and domain-specific parameters are treated as the inner-loop tunable parameters. Overall, by adding up Eq.(6) over all modalities, we end up with K_{IE} independent symmetric information sharing terms $K_{\text{IE}} = (O+1) + \sum_{m=1}^M [K^m_{\text{IE}} - (O+1)] = M(D+1) + (M+1)(O+1)$. Likewise, by adding up Eq.(7) over all task types, we end up with K_{ID} independent symmetric information sharing terms $K_{\text{ID}} = (M+1) + \sum_{o=1}^O [K^o_{\text{ID}} - (M+1)] = O(N+1) + (O+1)(M+1)$.

In Eq.(8), the terms $\mathcal{R}^{\text{share,share}}_{\text{MoME}}$ and $\mathcal{R}^{\text{share},o}_{\text{MoME}}, o = 1, 2, ..., O$ are shared over the $M(M-1)/2$ pairs of modalities. By adding up all pairs, we end up with K_{XE} independent symmetric information sharing terms $K_{\text{XE}} = \sum_{m,m'\in[M],m\neq m'} K^{m,m'}_{\text{XE}}/(M(M-1)/2) = O+1$. Likewise, by adding up Eq.(9) over all task types, we end up with K_{XD} independent symmetric information sharing terms $K_{\text{XD}} = \sum_{o,o'\in[O],o\neq o'} K^{o,o'}_{\text{XD}}/(O(O-1)/2) = M+1$.

Moreover, Eq.(6-9) have the following correlations: the term $\mathcal{R}^{\text{share,share}}_{\text{MoTE}}$ is related to $\mathcal{R}^{\text{share,share}}_{\text{MoME}}$, the term $\mathcal{R}^{\text{share},o}_{\text{MoTE}}$ is

related to $\mathcal{R}^{\text{share},o}_{\text{MoME}}$, the term $\mathcal{R}^{m,\text{share}}_{\text{MoTE}}$ is related to $\mathcal{R}^{m,\text{share}}_{\text{MoME}}$, and the term $\mathcal{R}^{m,o}_{\text{MoTE}}$ is related to $\mathcal{R}^{m,o}_{\text{MoME}}$.

Finally, we end up with the following independent terms

$$\begin{aligned} K = \;& K_{\text{IE}} + K_{\text{ID}} + K_{\text{XE}} + K_{\text{XD}} - 3(M+1)(O+1) \\ =\;& M(D+1) + O(N+1) + (M+1)(O+1) \end{aligned} \tag{10}$$

$$\begin{aligned} & \mathcal{R}(\boldsymbol{w}_i | i \in [N]) \\ =\;& \textstyle\sum_{m\in[M]} \mathcal{R}^m_{\text{BE}}\left(\boldsymbol{\phi}^{(m)}_{\text{BE},i} | i \in C^m_{\text{modal}}\right) \\ & + \textstyle\sum_{o\in[O]} \mathcal{R}^o_{\text{BD}}\left(\boldsymbol{\theta}^{(o)}_{\text{BD},i} | i \in C^o_{\text{task}}\right) \\ & + \textstyle\sum_{m\in[M]} \sum_{d=1}^{D} \mathcal{R}^{m,d}_{\text{MoDE}}\left(\boldsymbol{\phi}^{(m)d}_{\text{mode},i} | i \in C^d_{\text{MoDE}}\right) \\ & + \textstyle\sum_{o\in[O]} \sum_{i=1}^{N} \mathcal{R}^{o,i}_{\text{concept}}\left(\boldsymbol{\theta}^{(o)}_{\text{final},i} | \{i\}\}\right) \\ & + \mathcal{R}^{\text{share,share}}_{\text{MoTE \&MoTE}}\left(\{\boldsymbol{\phi}^{(\text{share})\text{share}}_{\text{mote},i}, \boldsymbol{\theta}^{(\text{share})\text{share}}_{\text{mome},i}\} | i \in [N]\right) \\ & + \textstyle\sum_{m\in[M]} \left[\mathcal{R}^{m,\text{share}}_{\text{MoTE \&MoTE}}\left(\{\boldsymbol{\phi}^{(m)\text{share}}_{\text{mote},i}, \boldsymbol{\theta}^{(\text{share})m}_{\text{mote},i}\} | i \in C^m_{\text{modal}}\right)\right] \\ & + \textstyle\sum_{m\in[M]} \sum_{o\in[O]} \left[\mathcal{R}^{m,o}_{\text{MoTE \&MoTE}}\left(\{\boldsymbol{\phi}^{(m)o}_{\text{mote},i}, \boldsymbol{\theta}^{(o)m}_{\text{mome},i}\} | i \in C^{m,o}_{\text{pair}}\right)\right] \\ & + \textstyle\sum_{o\in[O]} \left[\mathcal{R}^{m,\text{share}}_{\text{MoTE \&MoTE}}\left(\{\boldsymbol{\phi}^{(\text{share})o}_{\text{mote},i}, \boldsymbol{\theta}^{(o)\text{share}}_{\text{mome},i}\} | i \in C^o_{\text{task}}\right)\right] \end{aligned} \tag{11}$$

Case Study ($M = O = 2$). For simplicity, we have assumed a single shared-knowledge space can shared by all pairs of modalities. If there is only two modalities and two tasks, the single-shared knowledge assumption is true, and we have $K = M \cdot K_{\text{IE}} + O \cdot K_{\text{ID}} + M(M-1)K_{\text{XE}}/2 + O(O-1)K_{\text{XD}}/2 = M \cdot K_{\text{IE}} + O \cdot K_{\text{ID}} + K_{\text{XE}} + K_{\text{XD}}$.

Algorithm and Pseudocode of DisentAFL

Auxiliary Losses. (1) First, we incorporate an auxiliary loss added to local objective–the *alignment regularization loss* between the shared feature learned by each modality, $f^{\text{align}}_i(\boldsymbol{\phi}^{(m)}_{\text{left},i}, m \in \mathcal{I}_i) := \sum_{m,m'\in\mathcal{I}_i} ||\boldsymbol{h}'^{(m)}_{:d^{\text{modal}}_{\text{share}}} - \boldsymbol{h}'^{(m')}_{:d^{\text{modal}}_{\text{share}}}||^2_2$. (2) Second, in order to explicitly enforce the latent space to be as disentangled as designed above, we propose the *orthogonal regularization loss* between each knowledge types $f^{\text{orth}}_i(\boldsymbol{\phi}^{(m)}_{\text{left},i}, \boldsymbol{\phi}^{(m)o}_{\text{mote},i} | m \in \mathcal{I}_i, o \in \mathcal{O}_i) := \sum_{(m,o),(m',o')\in\mathcal{I}_i\times\mathcal{O}_i} \boldsymbol{Z}^\top_{o,m,\cdot} \boldsymbol{Z}_{o',m',\cdot}$. Without an accurate disentangled latent space, there would be false positive knowledge transfer and false negative knowledge transfer. Only an accurate disentangled space together with the designed parameter sharing strategy will encourage positive transfer and avoid negative transfer.

Client Sampling for Balanced Load of Experts. Following (Chen and Zhang 2022a), for training stability, we make sure each expert model (each knowledge type) is loaded and trained in a balanced manner over clients. That is, over the training rounds the number of clients by which each expert is activated should be balanced. We select clients having larger local loss (i.e., exploitation) as well as having blocks that were less frequently seen before (i.e., exploration). To balance the exploration-exploitation trade-off in the multimodal client selection problem, we employ Multi-Armed Bandit (MAB) algorithms for the problem of client selection in Cross-modal Cross-task FL. Regarding the local loss of individual clients are non-stationary during training, we make use of the discounted MAB algo-

Algorithm 1: Proposed DisentAFL

Input: Total number of clients N; total number of modalities M; total number of tasks O; total communication rounds T; each client's dataset $\mathcal{D}_i$, sensor sets $\mathcal{I}_i$, and task types $\mathcal{O}_i$, $\forall i \in [N]$.
Initialization: Randomly initialize super-network at server $\boldsymbol{w}^{\text{sup}}_0$ and gating function $\boldsymbol{\psi}_0$; compute and fix each client's MoME/MoTE gate matrix $\boldsymbol{S}_i = \texttt{IoGate}(\mathcal{I}_i, \mathcal{O}_i)$; each client's memory buffer storing validation accuracy during policy search $\mathcal{B}_i = \{\}$.
for round $t = 0$ to $T-1$ **do**
  Sampling a subset of clients $V_t \subset [N]$ with balanced load of MoME/MoTE experts.
  // local SGD independently
  for client $i \in V_t$ in parallel **do**
    // Expert activation
    Download the current policy $\boldsymbol{\psi}_{i,t} \leftarrow \boldsymbol{\psi}_t$.
    Sample an MoDE expert with ϵ-greedy: $\widetilde{\boldsymbol{g}}_{i,t} = \texttt{DomGate}(\mathcal{D}_i; \boldsymbol{\psi}_{i,t}) + \epsilon$.
    Query the current server model $\boldsymbol{w}^{\text{sup}}_t$ using $\widetilde{\boldsymbol{g}}_{i,t}$ and $\boldsymbol{S}_{i,t}$.
    Receive the sparsely-gated model from server $\boldsymbol{w}_{i,t} \leftarrow \texttt{ROUTE}(\boldsymbol{S}_{i,t}, \widetilde{\boldsymbol{g}}_{i,t}; \boldsymbol{w}^{\text{sup}}_t)$
    // Local Training
    $\widetilde{\boldsymbol{w}}_{i,t,0} = \boldsymbol{w}_{i,t}$
    for each local update step $\tau = 0$ to $U-1$ **do**
      Sample a batch of data $\mathcal{D}^b_i$ from local training dataset $\mathcal{D}_i$
      Local update

$$\widetilde{\boldsymbol{w}}_{i,t,\tau+1} = \widetilde{\boldsymbol{w}}_{i,t,\tau} - \gamma \nabla_{\widetilde{\boldsymbol{w}}_{i,t,\tau}} f_i(\widetilde{\boldsymbol{w}}_{i,t,\tau}; \mathcal{D}^b_i)$$

    end for
    $\boldsymbol{w}_{i,t+\frac{1}{2}} = \widetilde{\boldsymbol{w}}_{i,t,U}$
    Evaluate the current sampled expert ACC^d_i using local validation dataset and $\boldsymbol{w}_{i,t+\frac{1}{2}}$
    Add $(\widetilde{\boldsymbol{g}}_{i,t}, \text{ACC}^d_i)$ to memory buffer $\mathcal{B}_i$
    Update the policy $\boldsymbol{\psi}_{t+\frac{1}{2}} = \boldsymbol{\psi}_t - \alpha \nabla_{\boldsymbol{\psi}_t} Q$ using $\mathcal{B}_i$ such that the best expert is produced.
    Compute the best MoDE expert $\boldsymbol{g}_{i,t+1} = \texttt{DomGate}(\mathcal{D}_i; \boldsymbol{\psi}_{i,t+\frac{1}{2}})$.
    Upload $\boldsymbol{w}_{i,t+\frac{1}{2}}, \boldsymbol{\psi}_{t+\frac{1}{2}}, \boldsymbol{g}_{i,t+1}$ to server.
  end for
  // Knowledge aggregation
  Update $\boldsymbol{\psi}_{t+1} = \frac{1}{|V_t|} \sum_{i\in C_t} \boldsymbol{\psi}_{t+\frac{1}{2}}$
  for knowledge type $k = 1$ to K in parallel **do**
    Aggregate knowledge type k's corresponding parameters, $\boldsymbol{w}^{\text{sup}(k)}_{t+1}$, from $\boldsymbol{w}^{(k)}_{i,t+\frac{1}{2}}$, $i \in C_k$ as each term in Eq.(11).
  end for
end for
Output: Super-network at server $\boldsymbol{w}^{\text{sup}}_T$ and gating network $\boldsymbol{\psi}_T$; each client's personalized model $\boldsymbol{w}_{i,T} = \texttt{ROUTE}(\boldsymbol{S}_i, \boldsymbol{g}_{i,T}; \boldsymbol{w}^{\text{sup}}_T)$, where MoDE gate $\boldsymbol{g}_{i,T} = \texttt{DomGate}(\mathcal{D}_i; \boldsymbol{\psi}_T)$.

rithms. The clients are viewed as *arms* in the MAB problem. The discounted cumulative local loss of each client is $L_i(t) = \sum_{t'=1}^{t} \gamma^{t-t'} f_i(t')$; the discounted number of times each client has been selected over the previous rounds is $I_i(t) = \sum_{t'=1}^{t} \gamma^{t-t'} 1_{i \in C_{t'}}$; and, the discounted number of times each expert k has been sampled over previous rounds is, $P_i(t) = \sum_{t'=1}^{t} \gamma^{t-t'} 1_{(m,o,d) \in \mathcal{I}_i \times \mathcal{O}_i \times [D] \; \forall i \in C_{t'}}$. Here, $0 \leq \gamma \leq 1$ is the discount rate. Then, we define the estimated UCB reward of client k up to round t as $A_i(t) = L_i(t)/I_i(t) + U_i(t)$ where $U_i(t) = \sqrt{\sum_{r=1}^{t} \gamma^{t-r} / (I_i(t) + \sum_{(m,o,d) \in \mathcal{I}_i \times \mathcal{O}_i \times [D]} P_i(t))}$ is the exploration term for client i. At communication round t, we select the top C clients with largest discounted UCB rewards. The first term enforces selecting clients with estimated larger local loss (exploitation). However, if certain client has not been selected recently, or any type of model block of the client has not been selected recently, $U_i(t)$ will get larger. This forces the server to select them regardless of their local loss values (exploration).

Reproducibility

Datasets & Simulation

Since we introduced a new AFL problem, there is no existing dataset that can directly be used to evaluate our method. We constructed **six** simulated AFL federated datasets using *seven* existing multimodal or multitask datasets. The statistics of AFL simulations are summarized in Table **??**.

Source Datasets We select *seven* multimodal/multitask source datasets, whose statistics is summarized in Table **??**:

Aircraft and ***CIFAR-100*** (Finn, Abbeel, and Levine 2017) are two image classification datasets. Aircraft contains 10,200 images of aircraft with 100 images for each of 102 different aircraft classes. CIFAR-100 contains 60,000 images of real-world objects with 600 images for each of 100 classes. We resize images to 84 × 84.

Vehicle Sensor (Duarte and Hu 2004) for classifying vehicles driving by a segment of road. It contains 23 instances and each instance has 1,000 samples. Each sample is described by 50 acoustic and 50 seismic features and we predict between AAV-type and DW-type vehicles.

ModelNet40 (Wu et al. 2015) dataset for *multi-view 3D object recognition* tasks. It contains 12,311 3D shapes covering 40 common categories, including airplane, bathtub, bed, bookshelf, chair, cone, cup, and so on. Each 3D CAD object has $M = 2$ modalities as two views of its shapes. (Feng et al. 2019).

CMU-MOSEI from MultiBench (Liang et al. 2021) is a large dataset of sentence-level sentiment analysis and emotion recognition in real-world online videos with more than 65 hours of annotated video from more than 1,000 speakers and 250 topics. Each video is annotated for sentiment as well as the presence of *9 discrete emotions* (angry, excited, fear, sad, surprised, frustrated, happy, disappointed, and neutral) as well as *continuous emotions* (valence, arousal, and dominance).

AV-MNIST from MultiBench (Liang et al. 2021) is a multimodal dataset created by paring audio of a human reading digits from the FSDD dataset with written digits in the MNIST dataset with tasked to predict the digit into one of 10 classes (0-9).

Multi-FMNIST (Multi-(Fashion + MNIST)) (Liang et al. 2021) was created by randomly pick a pair of images from the original MNIST and FashionMNIST datasets, respectively, and then combine these two images into a new one by putting the digit on the top-left corner and the fashion object on the bottom-right corner. For each image, we therefore have two objectives to classify the item on the top-left (task 1) and to classify the item on the bottom-right (task 2).

AFL Simulation Setup From the above original datasets, we created *six* AFL federated datasets as follows.

- ***MERGE-AC*** simulates a *single-modal, single-task, and multi-domain* federated learning scenario with concept shifts. It focuses on only the image classification task, where clients' data has discrepant input and output distributions. This simulation is used to evaluate the performance of the mixture of domain expert (MoDE) module.
- ***ModelNet-xM*** and ***Vehicle-xM*** both simulate the *cross-modal, single-task, and single-domain* federated learning scenario, with and without concept shifts, respectively. Both of them focus on the classification task from diversified input modality types, where clients' data has discrepant input spaces. These simulations are used to evaluate the mixture of modality expert (MoME) module.
- ***MERGE-VM*** simulates an AFL scenario with the simultaneous 3 patterns of heterogeneity (MTC)–a *cross-modal, cross-task, and single-domain* federated learning scenario with concept shift, which is the entire. Clients' data has discrepant input and output spaces and distributions. Since the two source datasets has no common modality type and no common task task, this is a single-domain simulation ($D = 1$). This simulation is used to evaluate the entire DisentAFL framework with $D = 1$, to see whether any modality-shared and task-shared knowledge can help to improve both of ModelNet-xM and Vehicle-xM.
- ***MERGE-MM*** simulates an AFL scenario with the simultaneous 4 patterns of heterogeneity (MTDC)–a *cross-modal, cross-task, and cross-domain* federated learning scenario with concept shift, which is the entire. Clients' data has discrepant input and output spaces and distributions. Since the two source datasets both have the audio modality, this is a multi-domain simulation. This simulation is used to evaluate the entire DisentAFL framework.
- ***MERGE-FA*** simulates an AFL scenario with the simultaneous 4 patterns of heterogeneity (MTDC)–a *cross-modal, cross-task, and cross-domain* federated learning scenario with concept shift, which is the entire. Clients' data has discrepant input and output spaces and distributions. This simulation is used to evaluate the entire DisentAFL framework. The two input modalities are image or audio. The four downstream tasks include classifying the item on the top-left, classifying the item on the bottom-right, generating the digit image where the item is on the middle, and generating the audio signal

Dataset	# Samples	Modalities	Tasks
Aircraft	10,200	{Image}	{*Classification* (102 aircraft classes)}
CIFAR-100	60,000	{Image}	{*Classification* (100 object classes)}
Vehicle Sensor	23,000	{Audio, Seismic}	{*Classification* (2 vehicle types)}
ModelNet40	12,300	{View1, View2}	{*Classification* (40 3d objects)}
CMU-MOSEI	22,777	{Audio, Text, Video}	{*Classification* (9 sentiments), *Regression* (3 emotions) }
Multi-FMNIST	70,000	{Image}	{*Classification Task 1* (10 digits), *Classification Task 2* (10 objects)}
AV-MNIST	70,000	{Image, Acoustic}	{*Generation Task1* (image), *Generation Task2* (audio), *Classification* (10 digits)}

Table 3: Statistics of 7 Source Datasets.

of the digits. The images containing one item and those containing two items are considered as two different domains.

Others. For MERGE-VM or MERGE-MM, since there are more than three modalities, we use ρ to control the percentage of clients that have less input modalities. ρ can be treated as a control signal for the degree of inter-client discrepancy. The target class categories at each client are different. For simulation of ***concept shifts***, we assume the number of target classes in each task at each client is fixed to 5. For each client's each task, we randomly sample the 5 classes from the corresponding complete categories in the original dataset. Each client has equal ***number of samples***. The local dataset at each client $\mathcal{D}_i$ are ***split*** into 60% training, 10% validation, and 10% testing samples.

Baseline Reproducibility

We implemented six baseline methods: Local, FedAvg, Cross-FedAvg, Align-FedAvg, Cross-PFL, and Align-PFL. Implementation details are as follows.

- **Local**: We *separately train local models that have different encoder-decoder architectures, without any information sharing among clients*–no positive transfer and no negative transfer ($\mathcal{R}(\cdot)$=0). Each client i's model consists of the embedding and attention layers for each of $|\mathcal{I}_i|$ input modalities, which is similar to DisentAFL, and $|\mathcal{O}_i|$ decoders. The input dimension of every task's decoder is $768(|\mathcal{I}_i|+1)/(|\mathcal{O}_i|+1)$. (1) First, similar to DisentAFL, the 768-dimensional *encoded inputs at each position* are split into two parts: a 384-dimensional modality-private feature and a 384-dimensional modality-shared feature. Over the all the positions of each modality, we compute the average of the 384-dimensional modality-private features; then, over all positions of all modalities, we compute the average of the 384-dimensional modality-shared features. Finally, we have $(|\mathcal{I}_i|+1)$ separate 384-dimensional features. (2) We concatenate these features and get a $384(|\mathcal{I}_i|+1)$-dimensional vector $\boldsymbol{z}$. (3) Then, similar to DisentAFL, the $384(|\mathcal{I}_i|+1)$-dimensional vector is split into $(|\mathcal{O}_i|+1)$ parts: a 384-dimensional task-shared feature and a 384-dimensional task-private feature for each of the $|\mathcal{O}_i|$ tasks. (4) For every task's decoder, we fuse the task-shared features and the corresponding task-private features and obtain a $768(|\mathcal{I}_i|+1)/(|\mathcal{O}_i|+1)$-dimensional vector, which is used as the input of the decoder. (5) Every decoder takes a $768(|\mathcal{I}_i|+1)/(|\mathcal{O}_i|+1)$-dimensional feature as the input. The classification task's decoder architecture consists of two linear layers (with the hidden dimension 128) followed by a softmax layer.
- **FedAvg** (McMahan et al. 2018): Client models have the same structures as in "Local". Clients are split into several disjoint groups of clients, such that the clients within each group share the same modality-task pair and clients from different groups have different modality-task pairs. For each group, there exists one single group-specific model shared by all the client models within the group. We *separately train each group of clients using FedAvg, without any information sharing between different groups*–i.e., exclusive training.
- **Align-FedAvg**: Similar to the previous baseline, we apply the (McMahan et al. 2018) method to every group of clients, where clients having the same modality-task pair can fully share their learned knowledge with each other. In addition to FedAvg, Align-FedAvg encourages *information sharing between different modality-task pairs through latent alignment*, which assumes there exists an unified latent space shared by all modalities as well as shared by all downstream tasks. To achieve this, the modal architectures of different groups of clients are forced to learn the same latent space: once we obtain the 768-dimensional encoded inputs at each position, we take the average of them over all positions and obtain a 768-dimensional latent vector $\boldsymbol{z}$, and then, $\boldsymbol{z}$ will be the input of all downstream decoders. The inter-group information sharing can be done by simply (1) directly taking the average of encoder weights if two groups both learn this information; (2) directly taking the average of decoder weights if two groups both learn this information; (3) if two groups neither have common modality nor common task, using Knowledge Distillation to perform alignment on the full latent space (Seo et al. 2022a). The illustration of this baseline is shown in Figure 7(b). This baseline maximizes the positive transfer; however, it also incorrectly transfers any conflicting knowledge–i.e.,

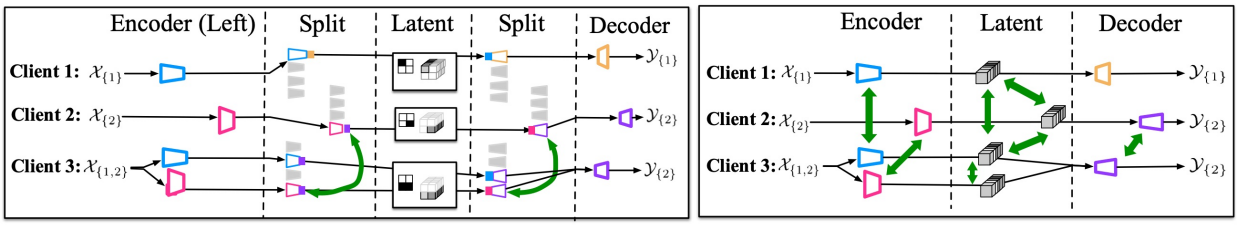
AFL Simulations	Source Data	#clients	#modalities	#tasks	Modality-task Pairs (#clients $\times(\mathcal{X}_{\mathcal{I}_i} \rightarrow \mathcal{Y}_{\mathcal{O}_i})$)
MERGE-AC	CIFAR-100 + Aircraft	15	1	1	$15 \times(\mathcal{X}_{\{\text{image}\}} \rightarrow \mathcal{Y}_{\{\text{cls\_objects}\}})$
ModelNet-xM	ModelNet40	30	2	1	$10 \times(\mathcal{X}_{\{\text{view1}\}} \rightarrow \mathcal{Y}_{\{\text{cls\_objects}\}})$ $10 \times(\mathcal{X}_{\{\text{view2}\}} \rightarrow \mathcal{Y}_{\{\text{cls\_objects}\}})$ $10 \times(\mathcal{X}_{\{\text{view1, view2}\}} \rightarrow \mathcal{Y}_{\{\text{cls\_objects}\}})$
Vehicle-xM	Vehicle Sensor	20	2	1	$7 \times(\mathcal{X}_{\{\text{audio}\}} \rightarrow \mathcal{Y}_{\{\text{cls\_vehicle}\}})$ $7 \times(\mathcal{X}_{\{\text{seismic}\}} \rightarrow \mathcal{Y}_{\{\text{cls\_vehicle}\}})$ $6 \times(\mathcal{X}_{\{\text{audio, seismic}\}} \rightarrow \mathcal{Y}_{\{\text{cls\_vehicle}\}})$
MERGE-VM	+ ModelNet40 Vehicle Sensor	50	4	2	$5\rho \times(\mathcal{X}_{\{\text{view1}\}} \rightarrow \mathcal{Y}_{\{\text{cls\_objects}\}})$ $5\rho \times(\mathcal{X}_{\{\text{view2}\}} \rightarrow \mathcal{Y}_{\{\text{cls\_objects}\}})$ $5\rho \times(\mathcal{X}_{\{\text{audio}\}} \rightarrow \mathcal{Y}_{\{\text{cls\_vehicle}\}})$ $5\rho \times(\mathcal{X}_{\{\text{seismic}\}} \rightarrow \mathcal{Y}_{\{\text{cls\_vehicle}\}})$ $5(1-\rho) \times(\mathcal{X}_{\{\text{audio, seismic}\}} \rightarrow \mathcal{Y}_{\{\text{cls\_vehicle}\}})$ $5(1-\rho) \times(\mathcal{X}_{\{\text{view1, view2}\}} \rightarrow \mathcal{Y}_{\{\text{cls\_objects}\}})$
MERGE-MM	CMU-MOSEI + Vehicle Sensor	50	4	3	$3\rho \times(\mathcal{X}_{\{\text{video}\}} \rightarrow \mathcal{Y}_{\{\text{sentiment}\}})$ $3\rho \times(\mathcal{X}_{\{\text{audio}\}} \rightarrow \mathcal{Y}_{\{\text{sentiment}\}})$ $3\rho \times(\mathcal{X}_{\{\text{text}\}} \rightarrow \mathcal{Y}_{\{\text{sentiment}\}})$ $3\rho \times(\mathcal{X}_{\{\text{video, audio}\}} \rightarrow \mathcal{Y}_{\{\text{sentiment}\}})$ $3\rho \times(\mathcal{X}_{\{\text{audio, text}\}} \rightarrow \mathcal{Y}_{\{\text{sentiment}\}})$ $3\rho \times(\mathcal{X}_{\{\text{video, text}\}} \rightarrow \mathcal{Y}_{\{\text{sentiment, emotions}\}})$ $3\rho \times(\mathcal{X}_{\{\text{text}\}} \rightarrow \mathcal{Y}_{\{\text{emotions}\}})$ $3\rho \times(\mathcal{X}_{\{\text{video, audio}\}} \rightarrow \mathcal{Y}_{\{\text{emotions}\}})$ $3\rho \times(\mathcal{X}_{\{\text{audio, text}\}} \rightarrow \mathcal{Y}_{\{\text{sentiment, emotions}\}})$ $3\rho \times(\mathcal{X}_{\{\text{audio}\}} \rightarrow \mathcal{Y}_{\{\text{cls\_vehicle}\}})$ $3\rho \times(\mathcal{X}_{\{\text{seismic}\}} \rightarrow \mathcal{Y}_{\{\text{cls\_vehicle}\}})$ $10(1-\rho) \times(\mathcal{X}_{\{\text{audio, seismic}\}} \rightarrow \mathcal{Y}_{\{\text{cls\_vehicle}\}})$ $10(1-\rho) \times(\mathcal{X}_{\{\text{video, text, audio}\}} \rightarrow \mathcal{Y}_{\{\text{emotions}\}})$
MERGE-FA	Multi-FMNIST + AV-MNIST	50	2	4	$3 \times(\mathcal{X}_{\{\text{image}\}} \rightarrow \mathcal{Y}_{\{\text{cls\_digits}\}})$ $3 \times(\mathcal{X}_{\{\text{audio}\}} \rightarrow \mathcal{Y}_{\{\text{cls\_digits}\}})$ $3 \times(\mathcal{X}_{\{\text{image, audio}\}} \rightarrow \mathcal{Y}_{\{\text{cls\_digits}\}})$ $3 \times(\mathcal{X}_{\{\text{audio}\}} \rightarrow \mathcal{Y}_{\{\text{cls\_digits, gen\_image}\}})$ $3 \times(\mathcal{X}_{\{\text{image}\}} \rightarrow \mathcal{Y}_{\{\text{gen\_audio}\}})$ $3 \times(\mathcal{X}_{\{\text{image, audio}\}} \rightarrow \mathcal{Y}_{\{\text{gen\_image, gen\_audio}\}})$ $3 \times(\mathcal{X}_{\{\text{image}\}} \rightarrow \mathcal{Y}_{\{\text{cls\_digits, cls\_objects}\}})$ $3 \times(\mathcal{X}_{\{\text{image}\}} \rightarrow \mathcal{Y}_{\{\text{cls\_objects, gen\_audio}\}})$ $3 \times(\mathcal{X}_{\{\text{image}\}} \rightarrow \mathcal{Y}_{\{\text{cls\_digits, cls\_objects, gen\_image}\}})$

Table 4: Statistics of the 6 Simulations of Cross-modal Cross-task Federated Learning (AFL).

negative transfer is also maximized.

- **Cross-FedAvg**: Again, we apply (McMahan et al. 2018) method to every group of clients as in FedAvg and Align-FedAvg, where clients having the same modality-task pair can fully share their learned knowledge with each other. Different from Align-FedAvg, Cross-FedAvg *encourages information sharing between different modality-task pairs that have* ≥ 1 *common modality and* ≥ 1 *common task*. If a client has a task A and input modality 1, it can at least share partial encoder and decoder information with another client that, at the same time, has task A and modality 1. The 768-dimensional averaged encoded inputs for each modality is equally split into O parts. Then, the part o from each modality is concatenated together to form a $768|\mathcal{I}_i|/O$-dimensional feature. For each decoder, there is an additional linear layer that maps the $768|\mathcal{I}_i|/O$-dimensional feature to the 768-dimensional input of decoder. The information sharing can be done by (1) directly taking the average of encoder weights, except the last layer; (2) directly taking the average of decoder weights, except the first layer; (3) using Knowledge Distillation on each of the $768/O$-dimensional feature segments corresponding to common modalities and common tasks. There is no knowledge sharing between tasks that have no common modality and no common task. The illustration of this method is shown in Figure 7(a). This baseline reduces negative transfer rather than Align-FedAvg; but may ignore some true positive transfer.

- **Cross-PFL** (Smith et al. 2017): Similar to Cross-FedAvg, except that using the personalized FL method to every modality-task pair group of clients.



(a) Cross-FedAvg/PFL (b) Align-FedAvg/PFL

Figure 7: Illustration of alignment-based and no-alignment based baseline methods.

- **FedMSplit/Align-PFL** (Chen and Zhang 2022a): Similar to Align-FedAvg, except that using the FedMSplit method personalized FL method (Chen and Zhang 2022a) to every modality-task pair group of clients.

The local objectives in baselines include (1) the *task-related* losses–i.e., cross-entropy loss for classification and mse loss for regression or generation, as well as (2) the knowledge distillation (KD) loss. The KD loss is only applicable in Align/Cross-FedAvg/PFL baselines.

DisentAFL Reproducibility & Hyperparameters

We implement DisentAFL using Pytorch 3 and ran all experiments on one GPU device. The **hyperparameters** are as follows:

- **Torch Seed**: 242
- **Model configuration**:
  - **Encoder output space**: in all baselines and DisentAFL, we fix $d^{\text{modal}} = 768$ as the output dimension of each modality's encoder.
  - **MoDE** is applied to each query or value linear layers. Suppose the input feature size and output feature size of a linear layer is d_{in} and d_{out}, respectively. Each MoDE expert module consists of an encoding layer (with input size d_{in} and output size 64) and an decoding layer (with input size 64 and output size d_{out}). The number of pre-defined domain expert is $D = 3$.
  - **MoTE's input and output dimension**: the 768-dimensional encoded features are split into two parts: a 384-dimensional modality-private feature ($d_m^{\text{modal}} = 384$) and a 384-dimensional modality-shared feature ($d_{\text{share}}^{\text{modal}} = 384$). For each expert model of MoTE, the input dimension is 384 and the output dimension is d^{latent}.
  - **Latent space** $\mathcal{Z}$: In all baselines and DisentAFL, we fix $d^{\text{latent}} = 128$ as the output dimension of each of task-expert models
  - **MoME's input and output dimension**: for each MoME expert, the input dimension is d^{latent} and the output dimension is 384.
  - **Decoder input dimension.** Decoders take as input the fused representation–combining the 384-dimensional task-private feature ($d_o^{\text{task}} = 384$) and the 384-dimensional task-shared feature ($d_{\text{share}}^{\text{task}} = 384$). The input dimension of each task's decoder is 768.
  - **Loss weights**: DisentAFL incorporate two auxiliary losses to the local objective to impose knowledge disentanglement to better avoid negative transfer. In all experiments, The weight of the alignment loss is fixed to 1.2. The weight of the orthogonal constraint loss was fixed to 4. For DisentAFL-KD, the weight of KD loss is 0.8.
- **Training-related hyperparameters**:
  - Local training epochs ($U = 1$); batch size (64); learning rate of model parameter ($\gamma = 0.015$); optimizer (Adam);
  - Global round ($T = 100$); client sampling ratio (0.2% per round); ϵ-greedy domain expert sampling ($\epsilon = 0.3$).