AAAI 2024

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

Jiayi Chen[1], Aidong Zhang[2]

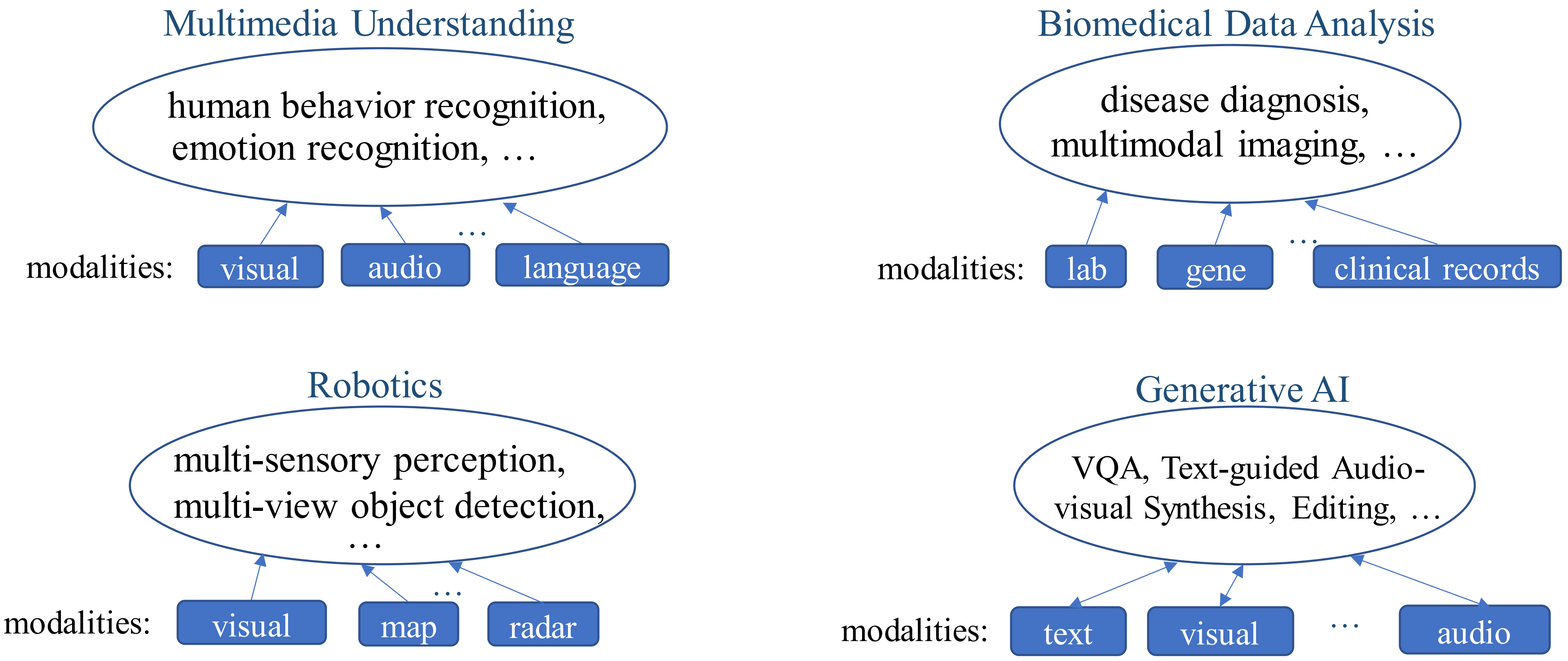
[1,2]*Department of Computer Science, University of Virginia*

[1]Contact: jc4td@virginia.edu

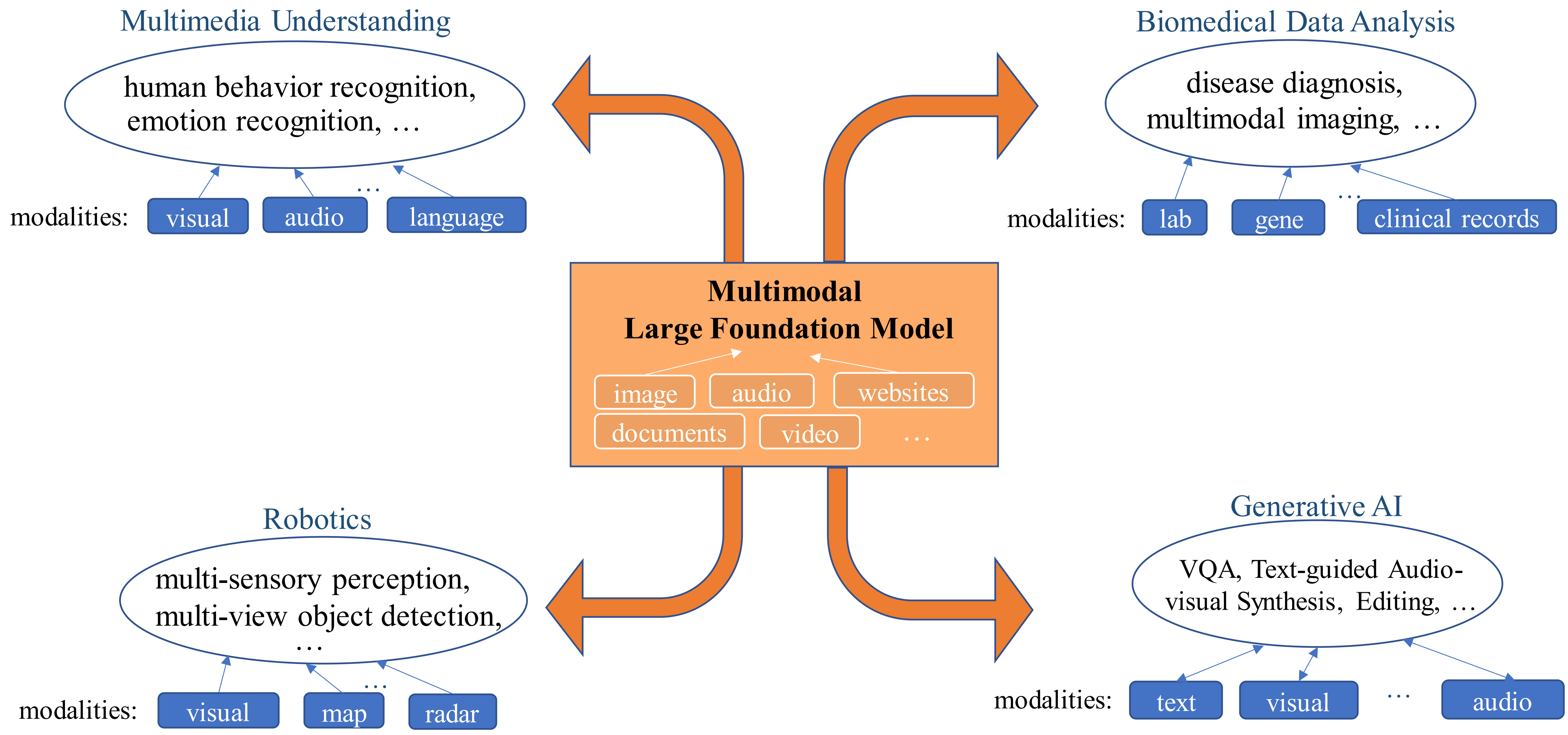
[2]Thomas M. Linville Professor

Multimodal Artificial Intelligence

- **Real-world Use Cases**:
  - *Discriminative tasks:* Multimodal/cross-modal Understanding, Alignment, Multimodal fusion, …
  - *Generative tasks:* Cross-modal guided data synthesis, Video captioning, grounding, …

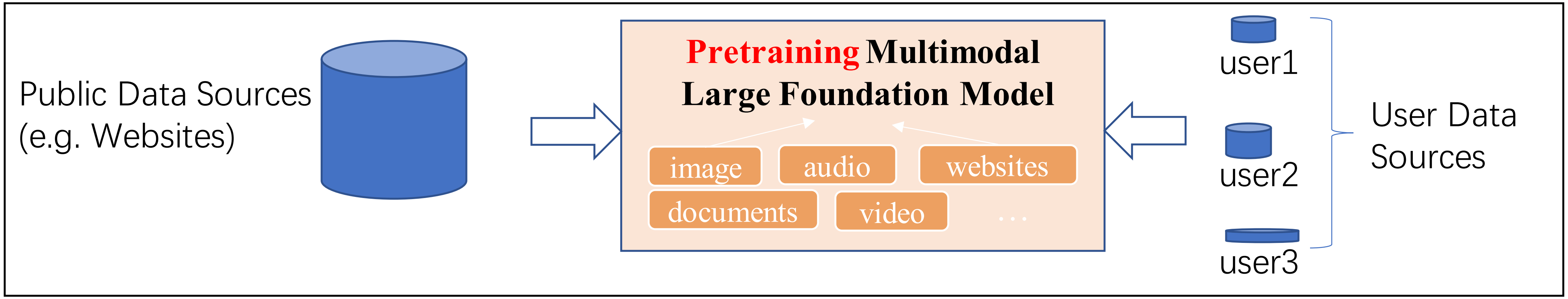


Artificial General Intelligence (AGI)



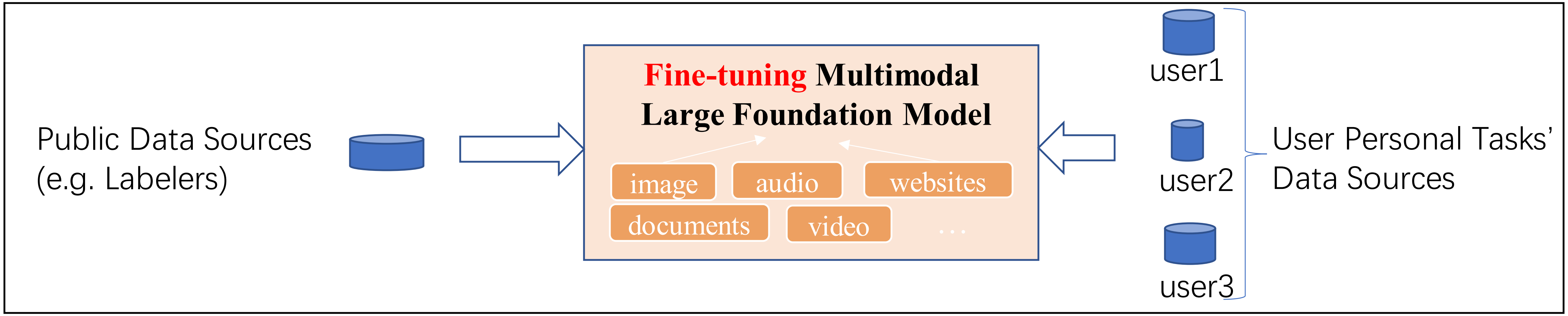
User-AGI Interactions

➢ Pretraining Stage



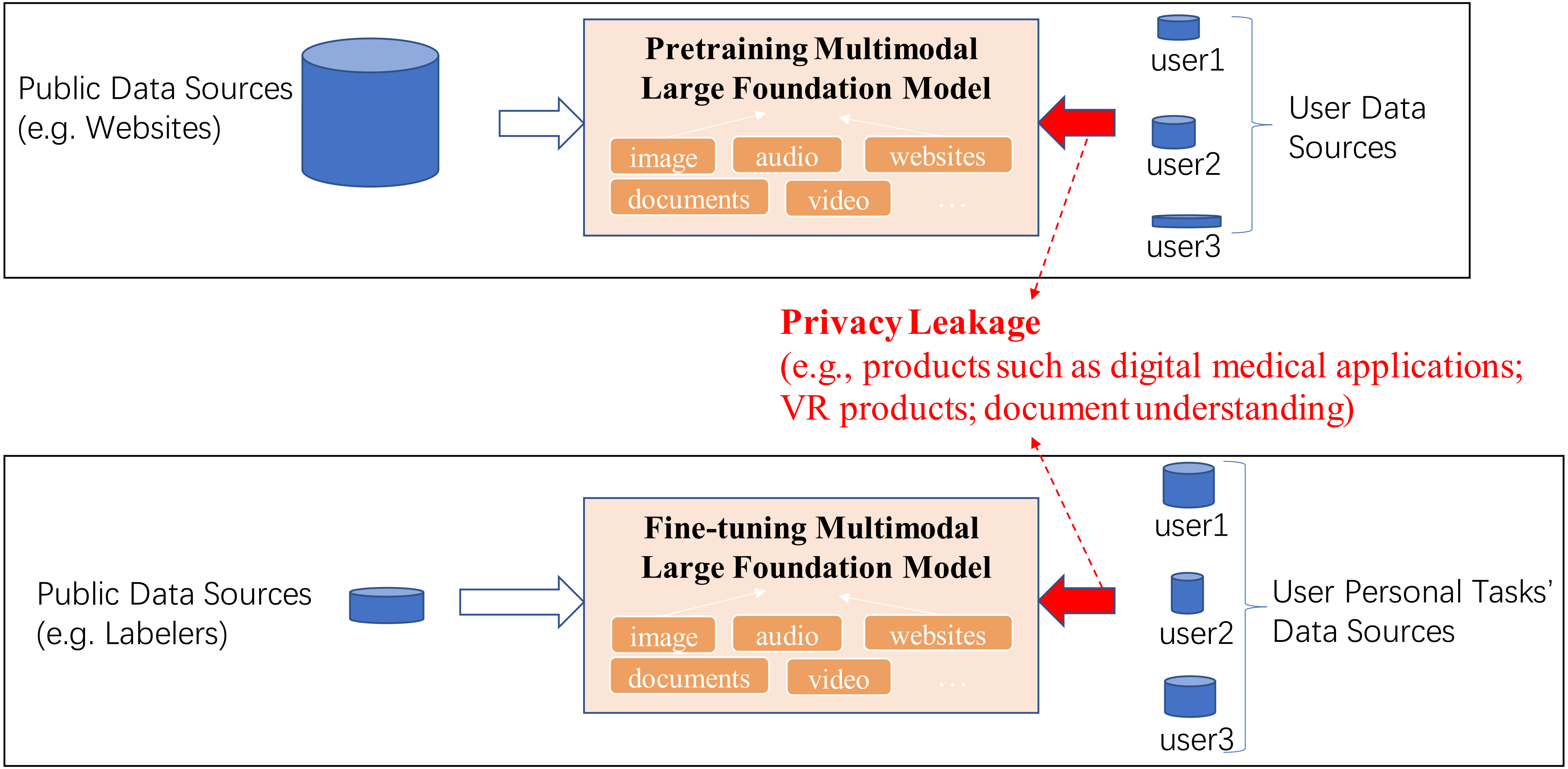
➢ Fine-tuning Stage

- *Training*: RLHF, meta-tuning, …
- *Instruction/Data:* Prompts, RAG, …
- *Model modification:* PEFT (Adaptors, Prefix Tuning, …)

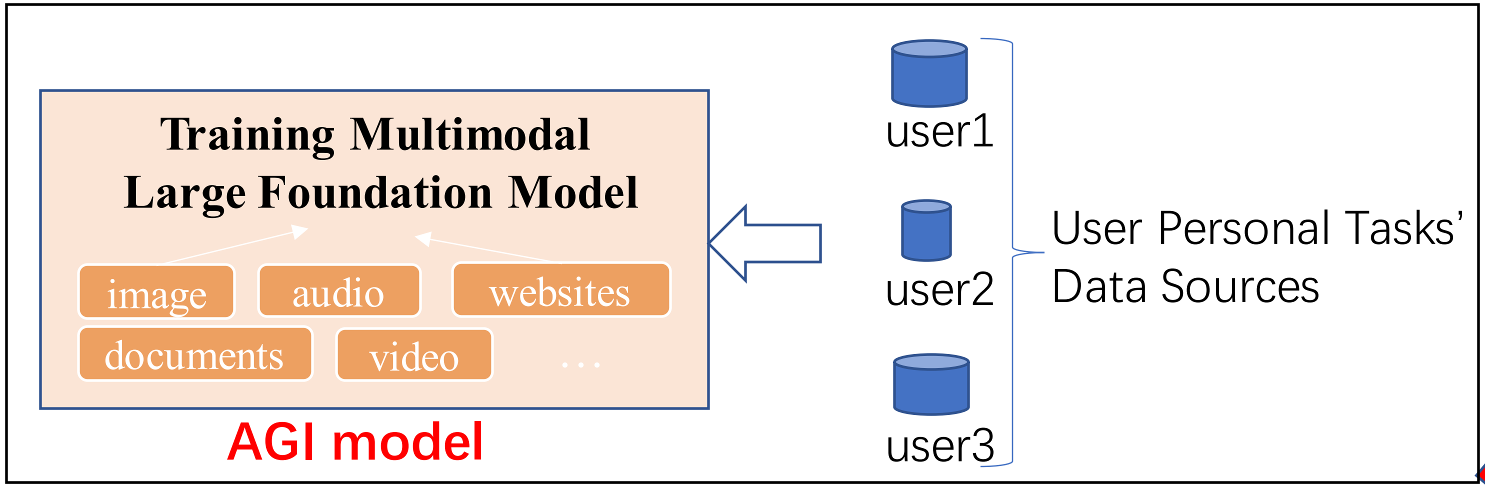


Privacy Issue in User-AGI Interactions

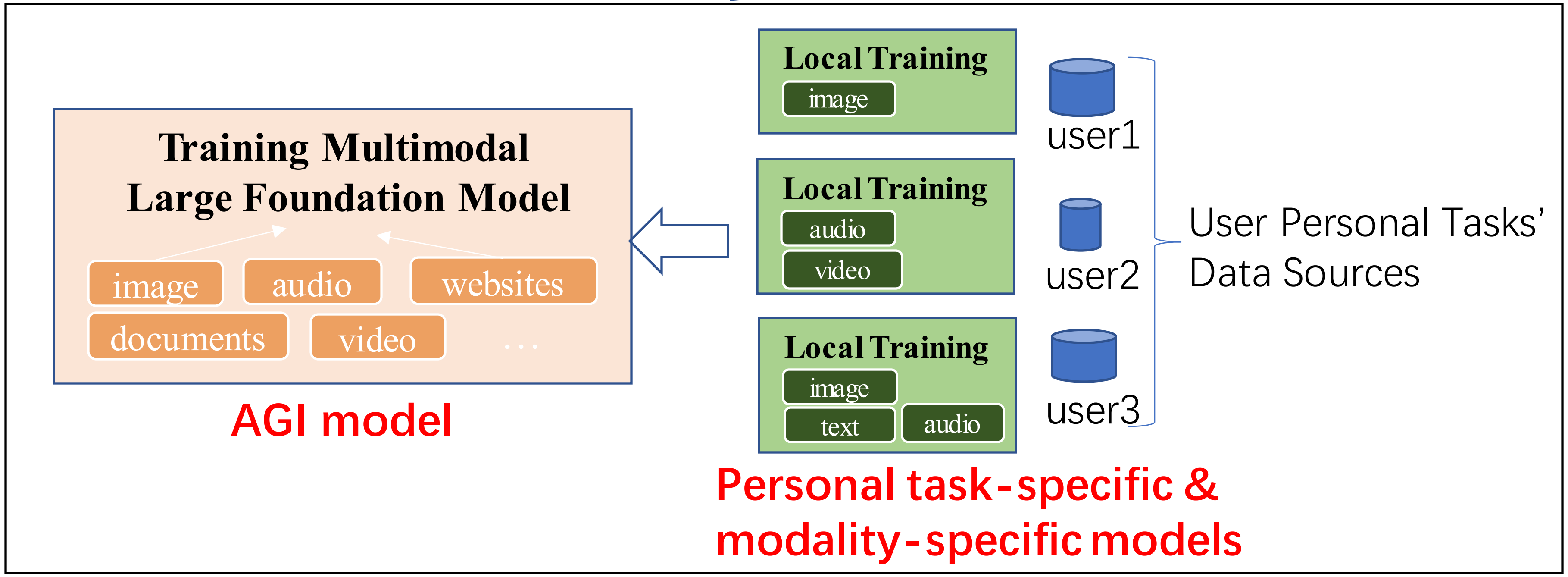
- **Problem?**



Privacy-preserving AGI via User Collaboration



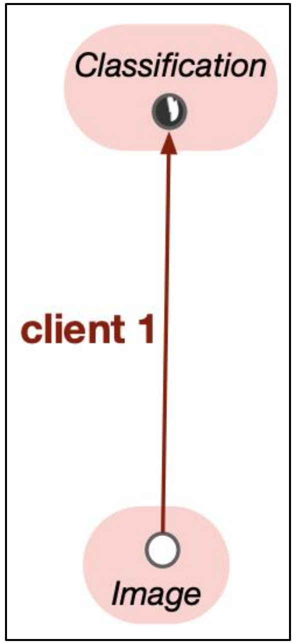
Can we use Personalized Federated Learning (PFL) [1] to achieve
- **user privacy of AGI while handling**
- **diverse modalities and task categories?**

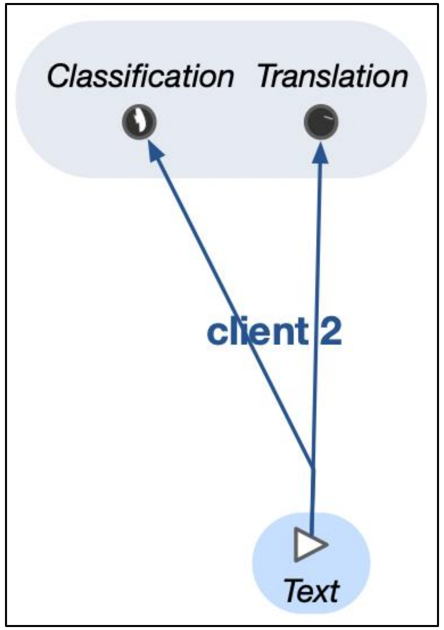


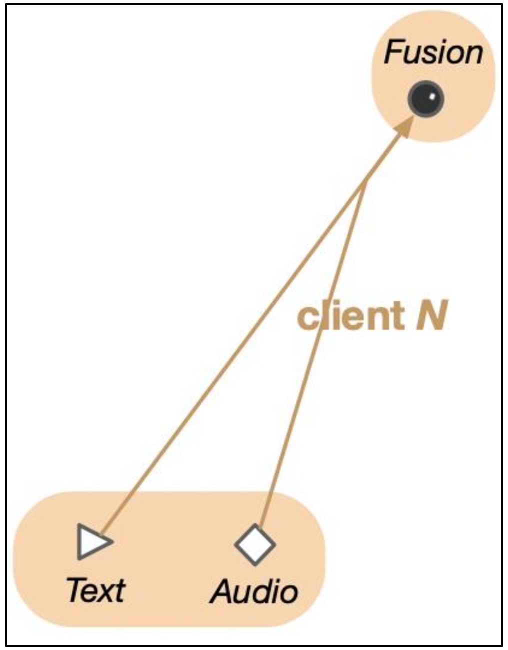
[1] Tan, Alysa Ziying, et al. "Towards personalized federated learning." *IEEE Transactions on Neural Networks and Learning Systems* (2022).

Problem Setting

➢ **Local Training:** Totally personal, no AGI benefit

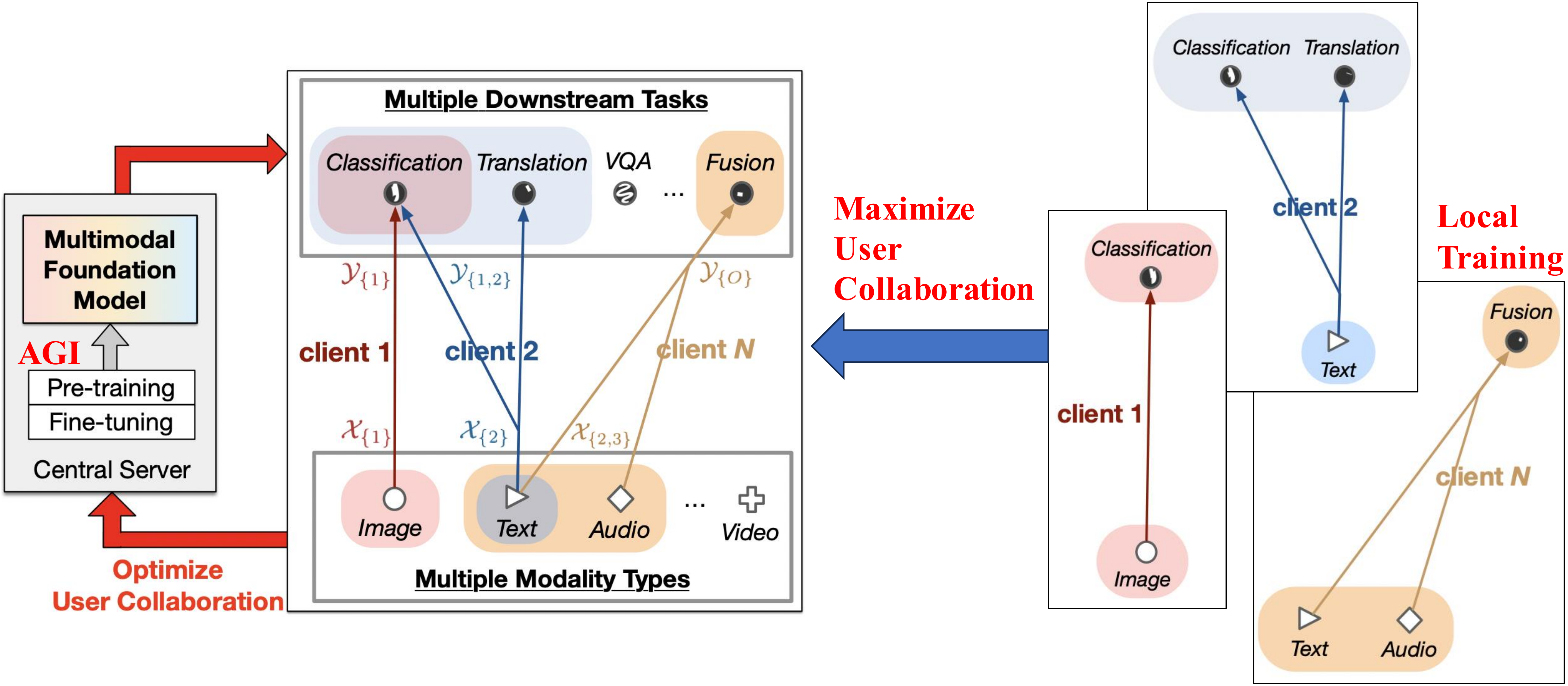




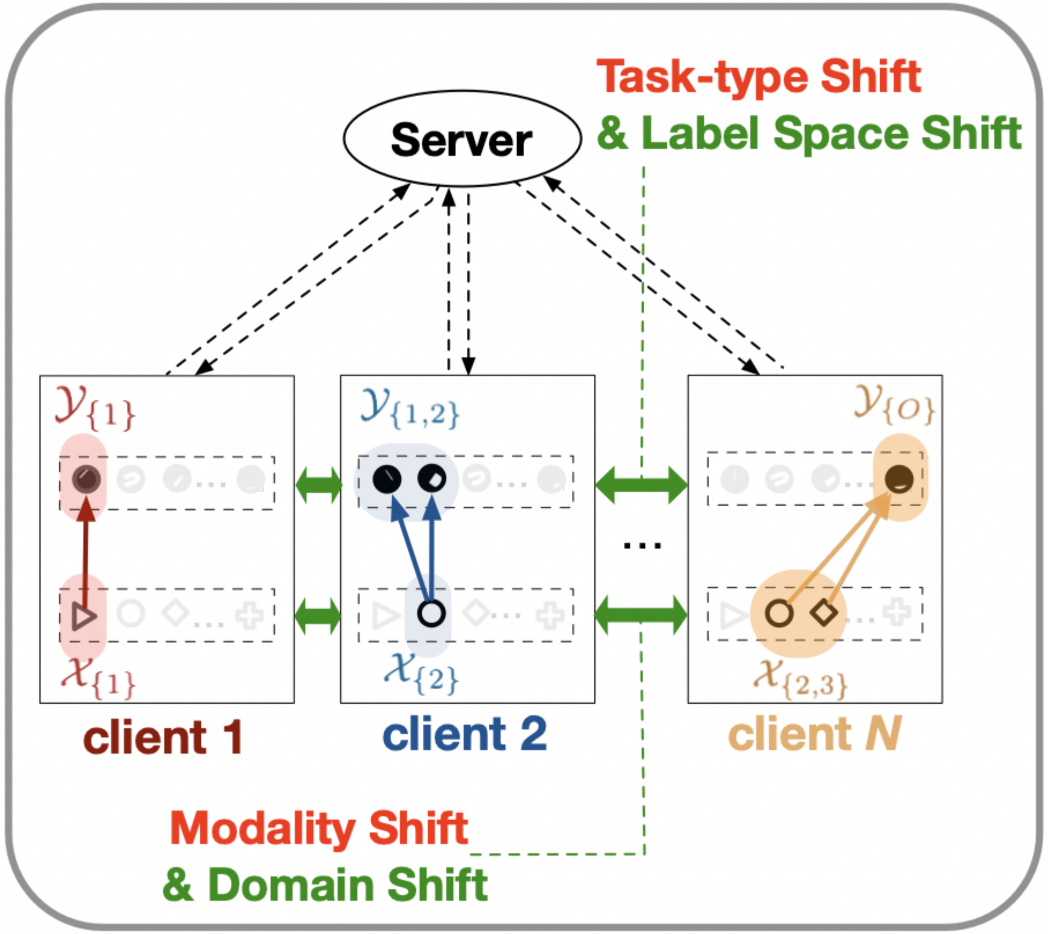


Problem Setting

➢ **Our Setting:** leverage user collaboration to learn an AGI model



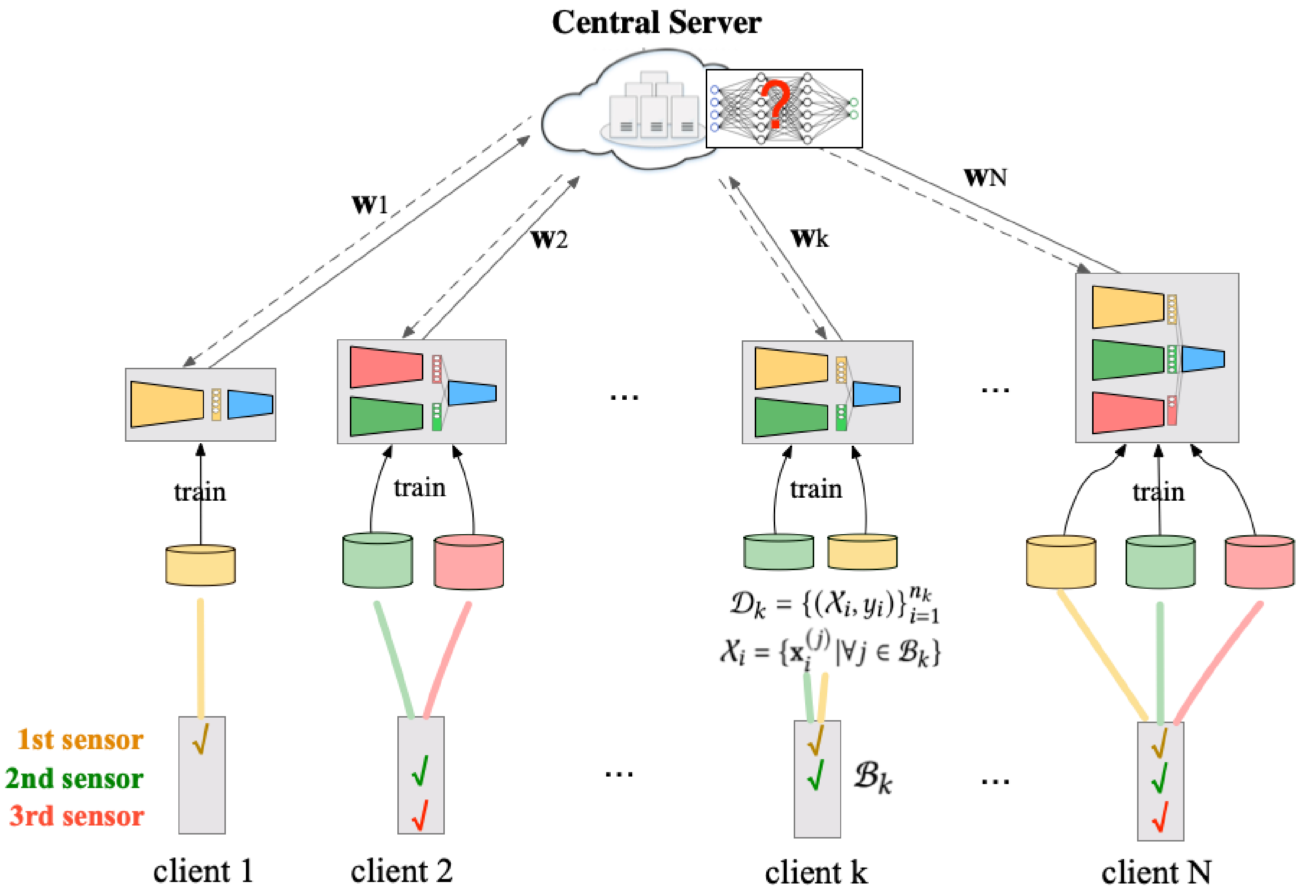
Problem Setting



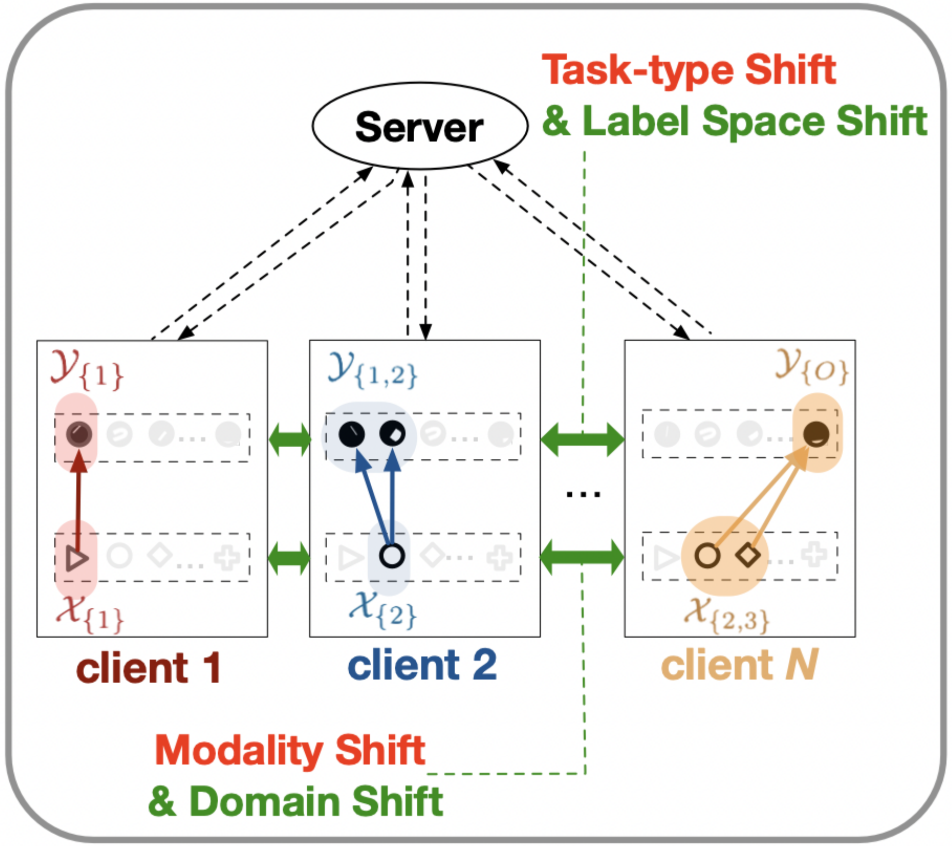
4 Heterogeneity Patterns:

- **Domain shift**
- **Concept shift**
- **Modality gap** (image, text, audio, video) across different domains
- **Task type difference** (object classification, image captioning, audio generation, emotion recognition)

Unique Challenges compared to Existing Solutions



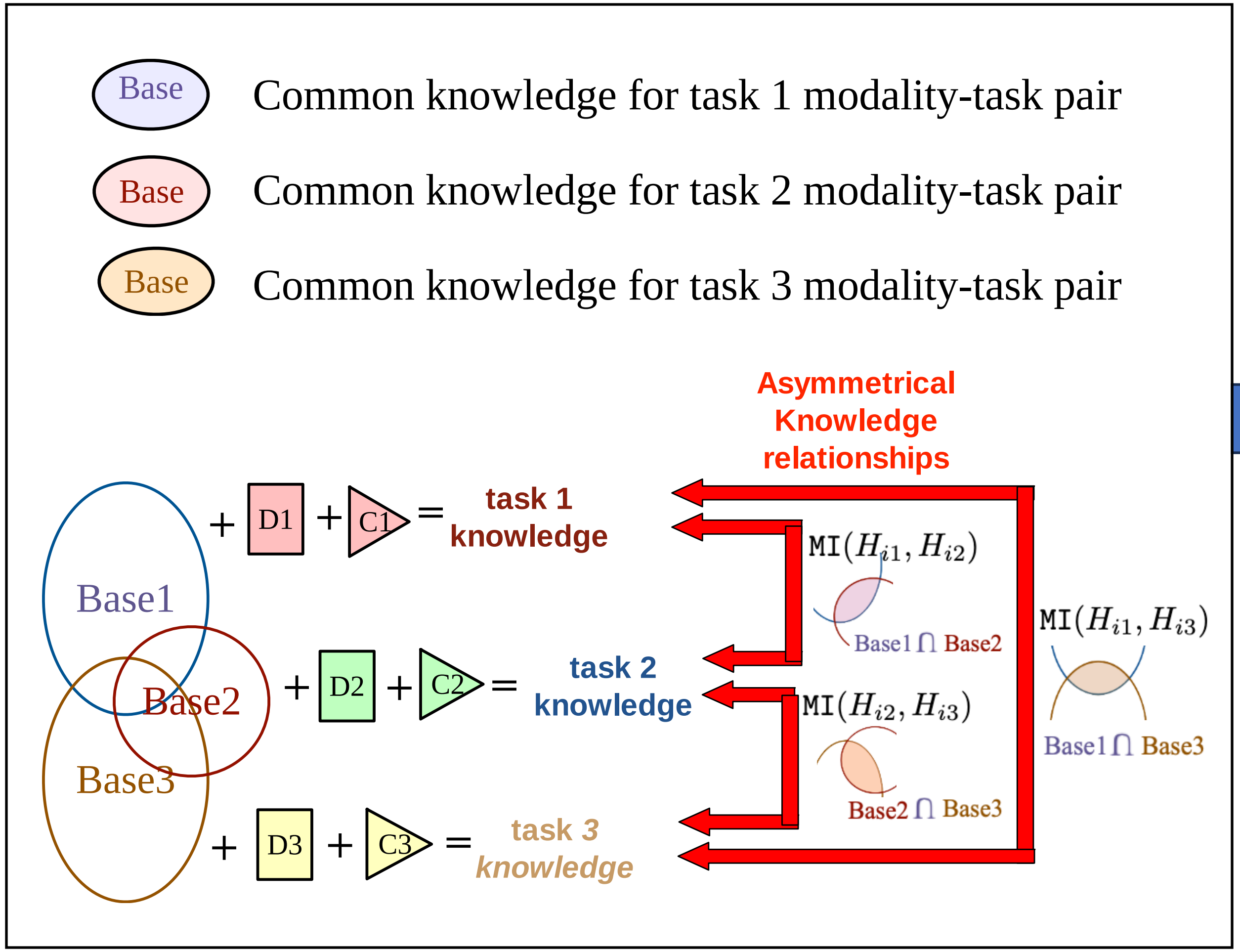
Existing Solutions: Multimodal FL via Latent Space Alignment



Challenge: Suboptimal solution with large modality gap & task gap

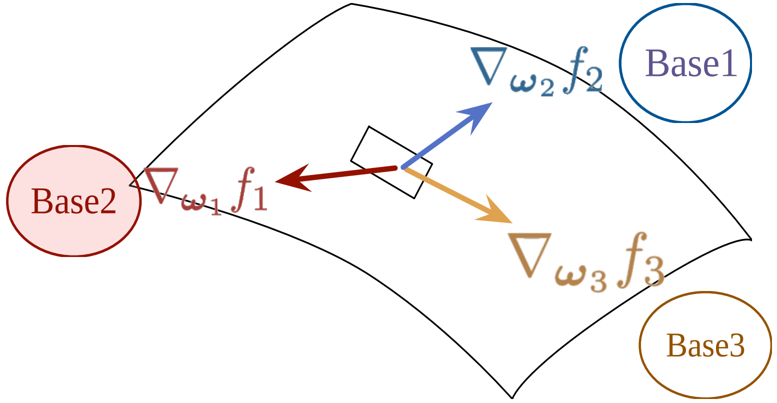
A Closer Look: Knowledge Unalignment between Users

True Knowledge Alignment & Conflict

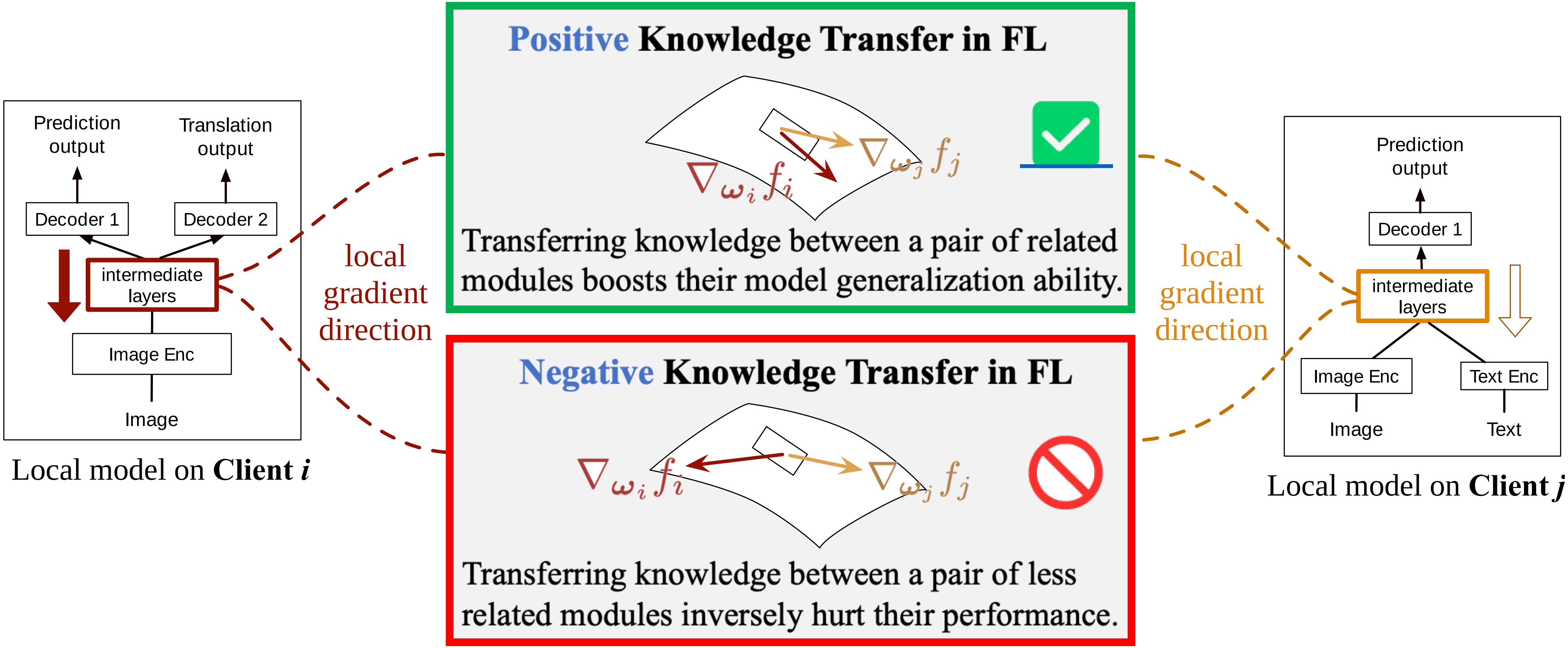


Our Observation:

If assuming an aligned latent space, gradients at the latent layers



Inspiration



[1] Riemer, Matthew, et al. "Learning to Learn without Forgetting by Maximizing Transfer and Minimizing Interference." *ICLR*. 2019.

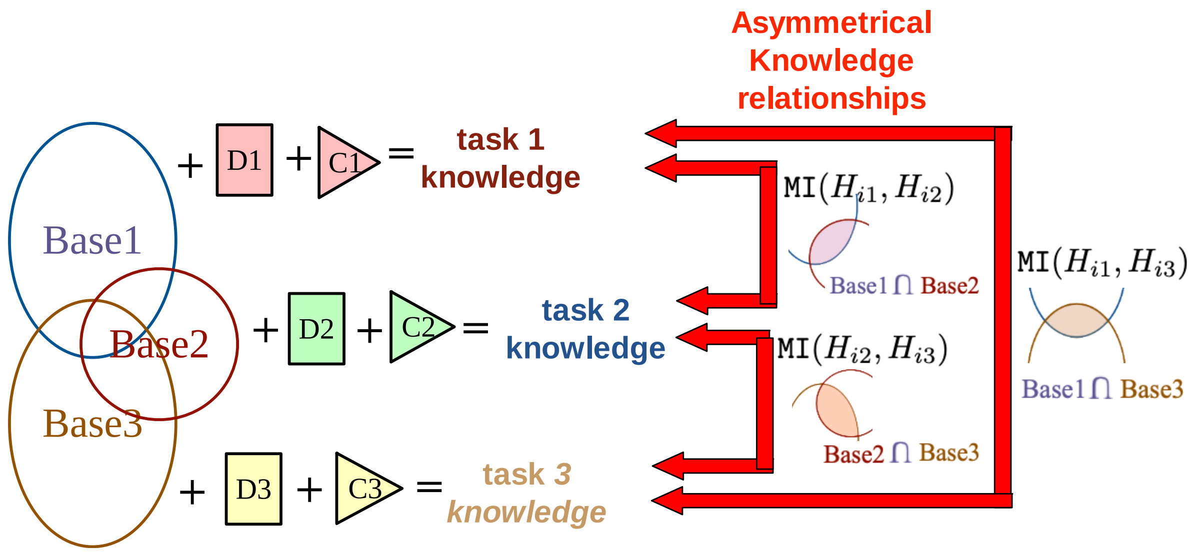
A Closer Look: Knowledge Unalignment between Users

True Knowledge Alignment & Conflict

Base — Common knowledge for task 1 modality-task pair

Base — Common knowledge for task 2 modality-task pair

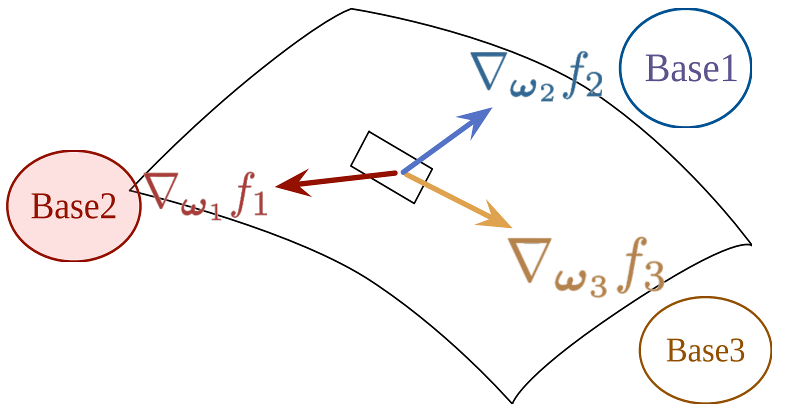
Base — Common knowledge for task 3 modality-task pair



Our Observation:

If assuming an aligned latent space,

gradients at the latent layers

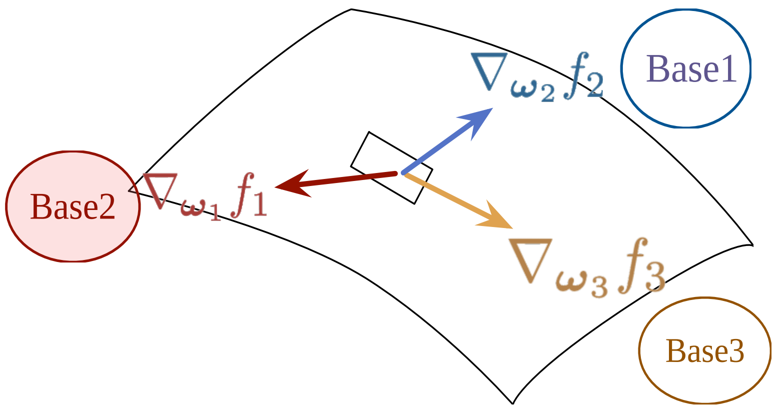


Consequence:

- **Insufficient Positive Knowledge Transfer**
- **Unwanted Negative Knowledge Transfer**

Main Idea

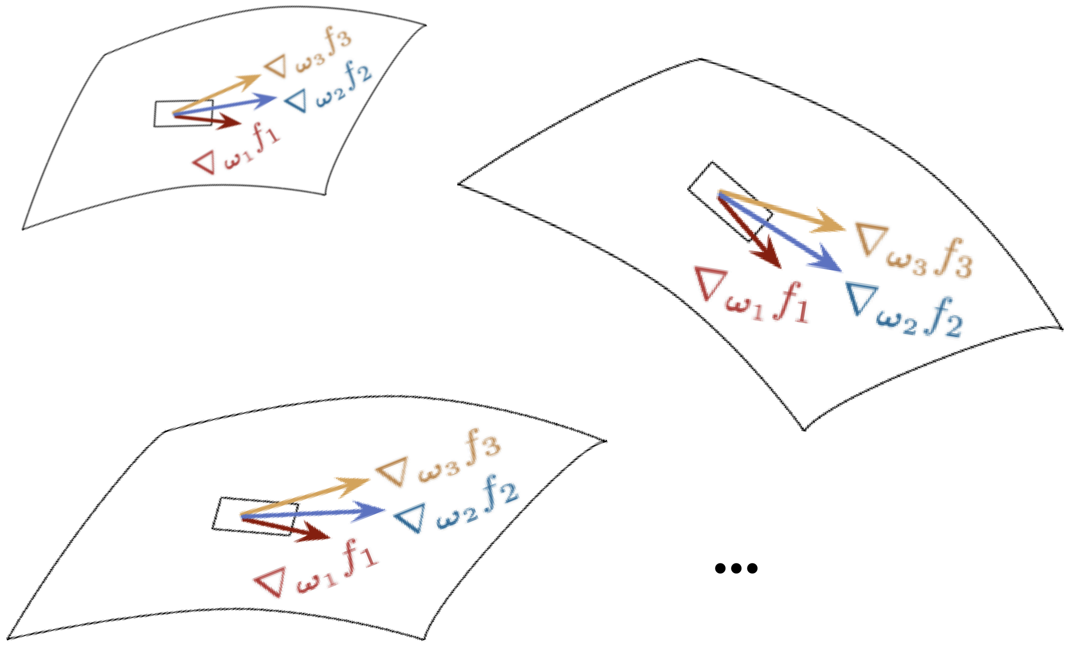
One aligned latent space



- **Insufficient Positive Knowledge Transfer**
- **Unwanted Negative Knowledge Transfer**

Disentangle

Multiple disentangled latent subspaces

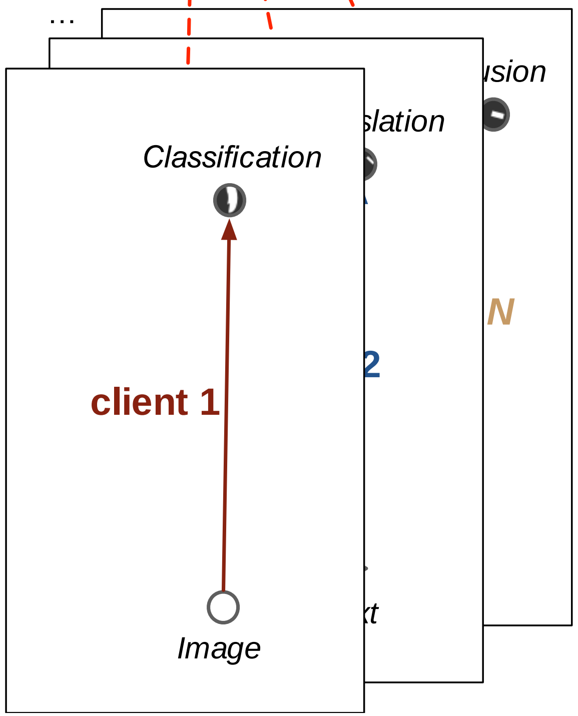


- **maximized Positive Knowledge Transfer**
- **minimized Negative Knowledge Transfer**

Global Objective

$$\min_{\omega_1,\omega_2,\ldots,\omega_N} \left[\frac{1}{N} \sum_{i=1}^{N} \boxed{f_i(\omega_i)} \right] + \mathcal{R}(\omega_1, \omega_2, \ldots, \omega_N)$$

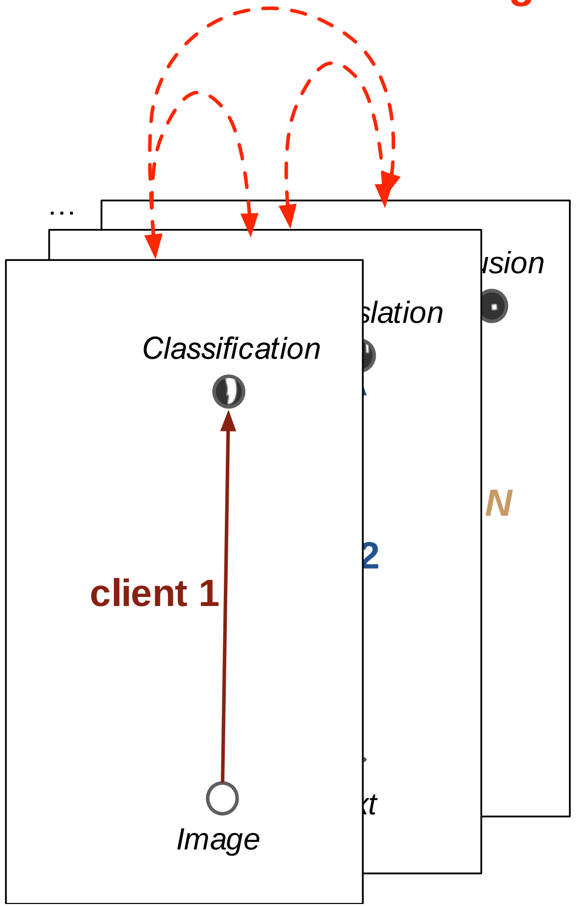
local objectives: $\min_{\omega_i} f_i(\omega_i)$



Global Objective

$$\min_{\omega_1,\omega_2,\ldots,\omega_N} \left[\frac{1}{N} \sum_{i=1}^{N} f_i(\omega_i) \right] + \boxed{\mathcal{R}(\omega_1, \omega_2, \ldots, \omega_N)}$$

privacy preserving
knowledge sharing scheme



Global Objective

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

original knowledge sharing scheme

$$\sum_{k=1}^{K} \boxed{\mathcal{R}_k(\{\boldsymbol{w}_i^{(k)} | \forall i \in C_k\})}$$

fine-grained knowledge sharing schemes

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

Disentangle Knowledge Sharing Schemes (Details)

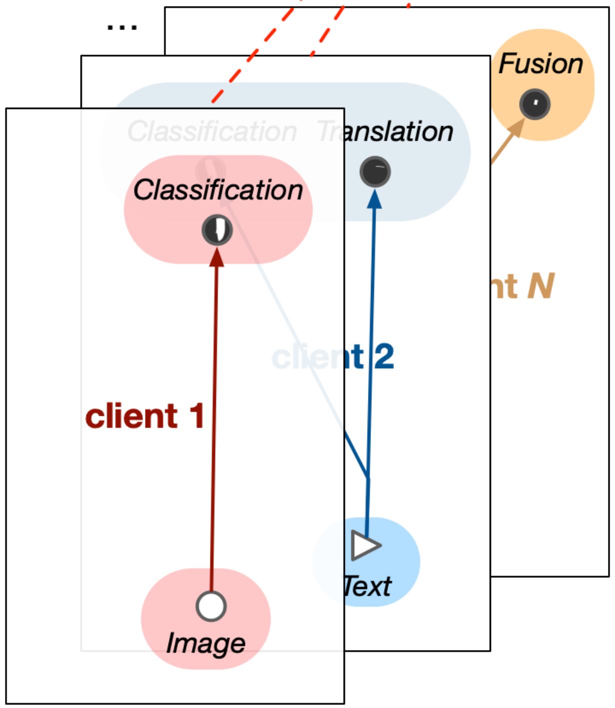
How to Solve Global Objective?

Research Question: How to **maximize positive transfer** while **minimizing negative transfer**?

$$\min_{\omega_1,\omega_2,\ldots,\omega_N} \left[\frac{1}{N}\sum_{i=1}^{N} \boxed{f_i(\omega_i)}\right] + \boxed{\sum_{k=1}^{K}\mathcal{R}_k(\{w_i^{(k)}|\forall i\in C_k\})}$$

fine-grained knowledge sharing schemes

local objectives



② latent space disentanglement

purification

① Mixture of Expert w/ Routing Mechanism

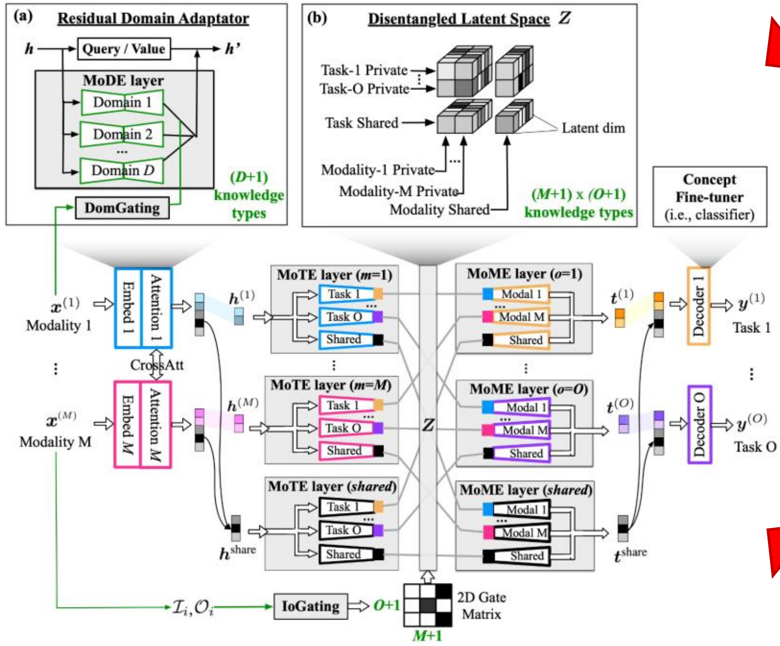
communication

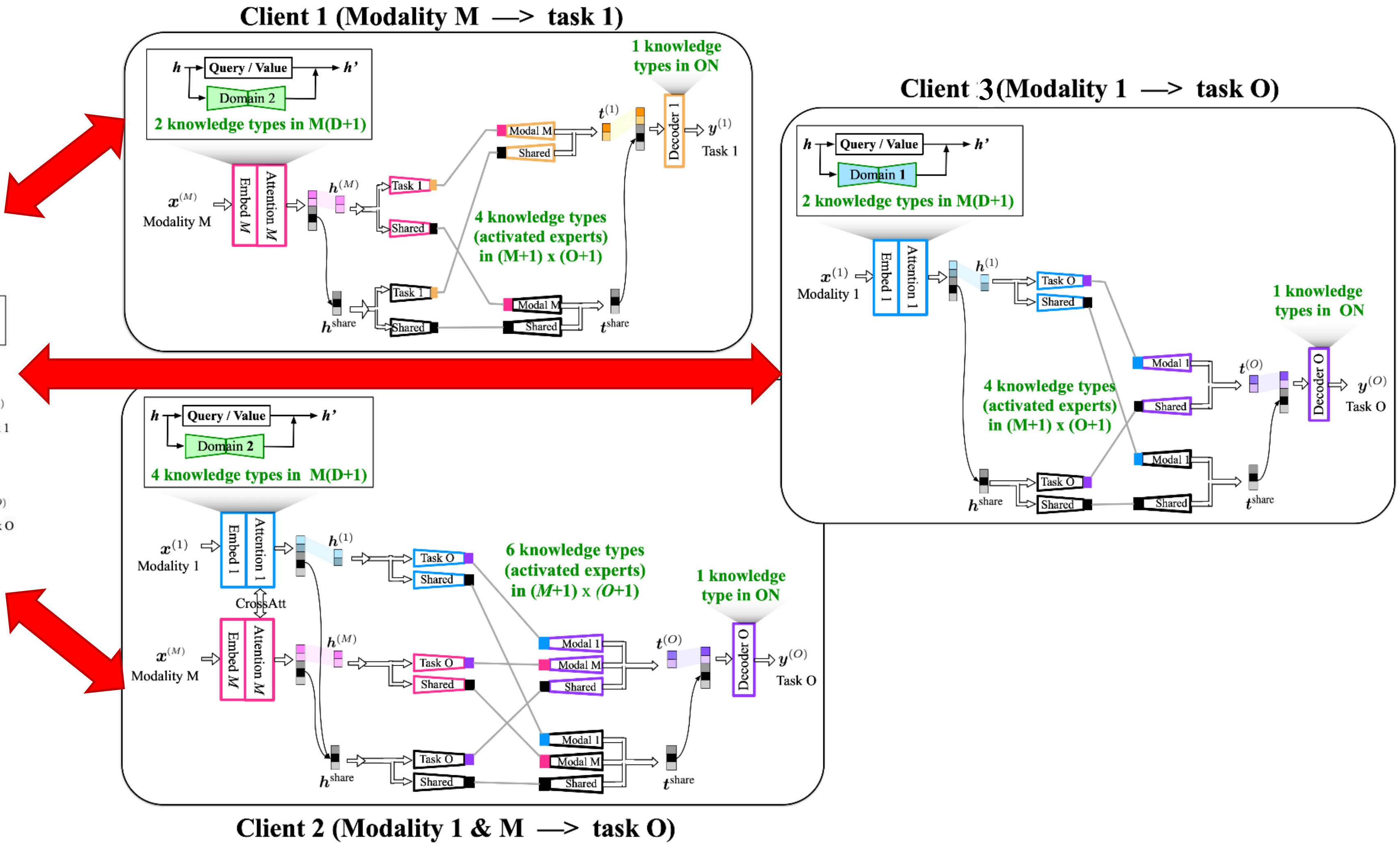
Communication

- **Model architectures based on Mixture of Experts**

  - **Client Personal Models (MoE)**

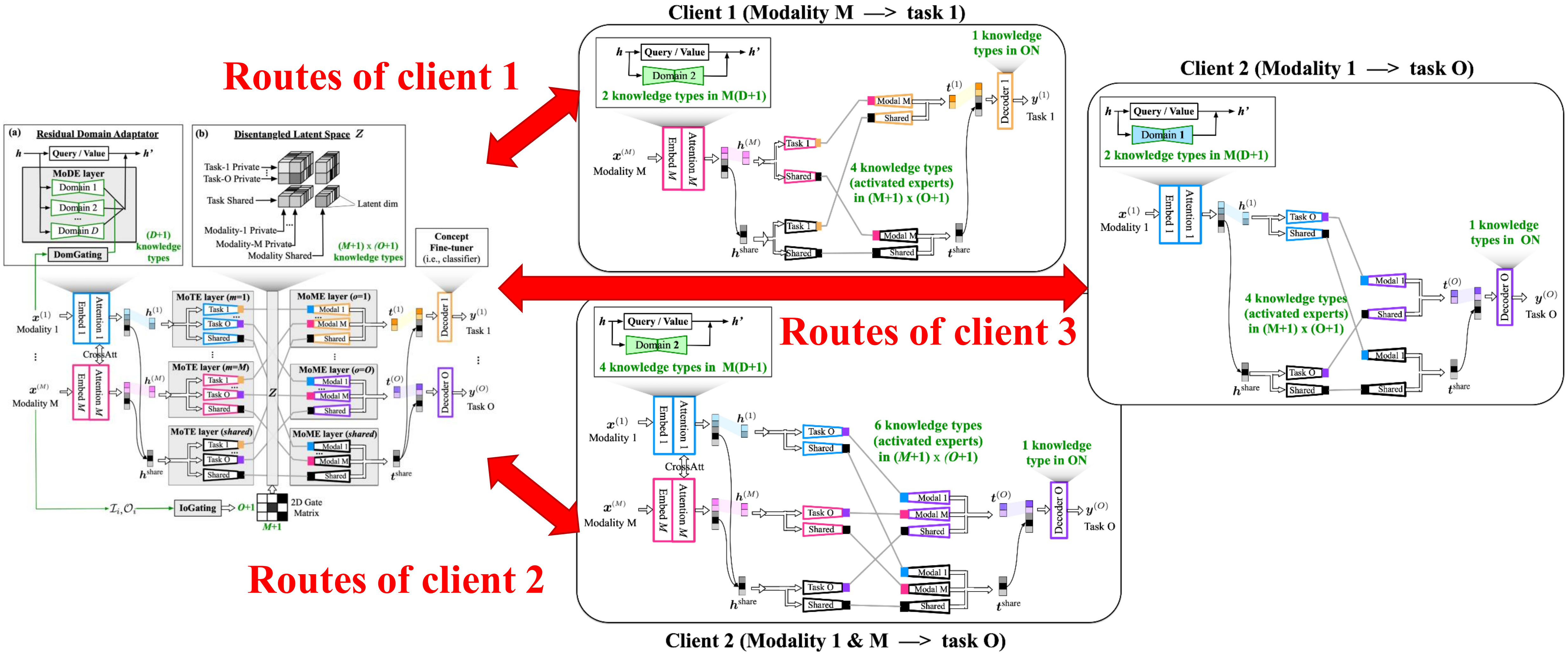
Supernetwork (MoE) on Server





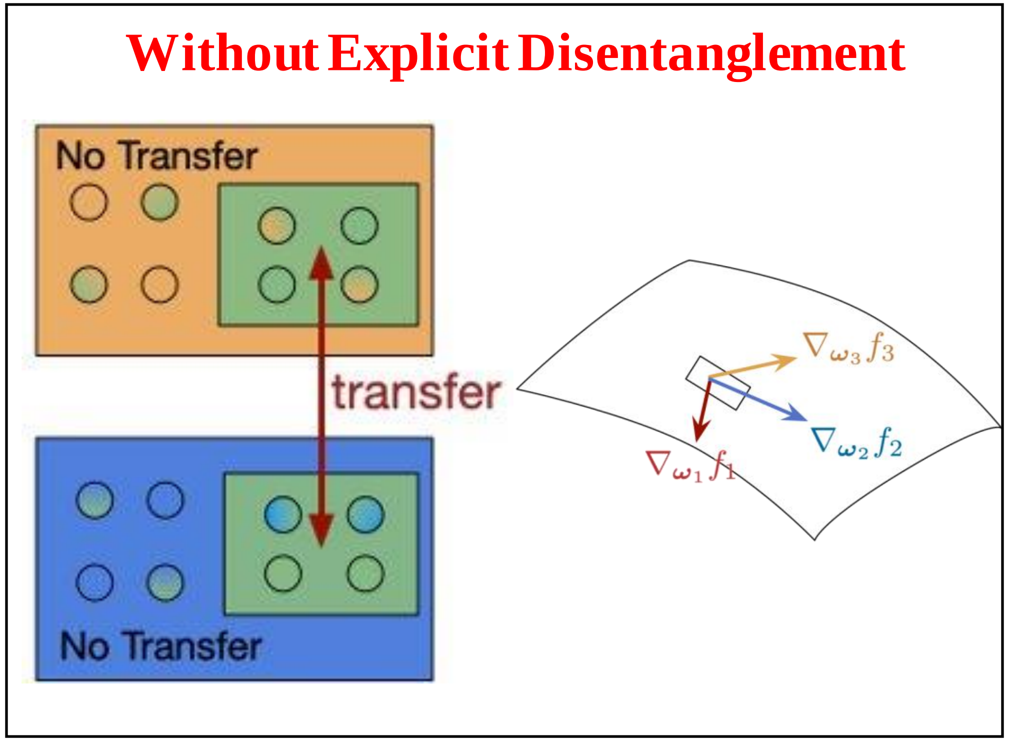
Communication

- **Automatic routing during communication**

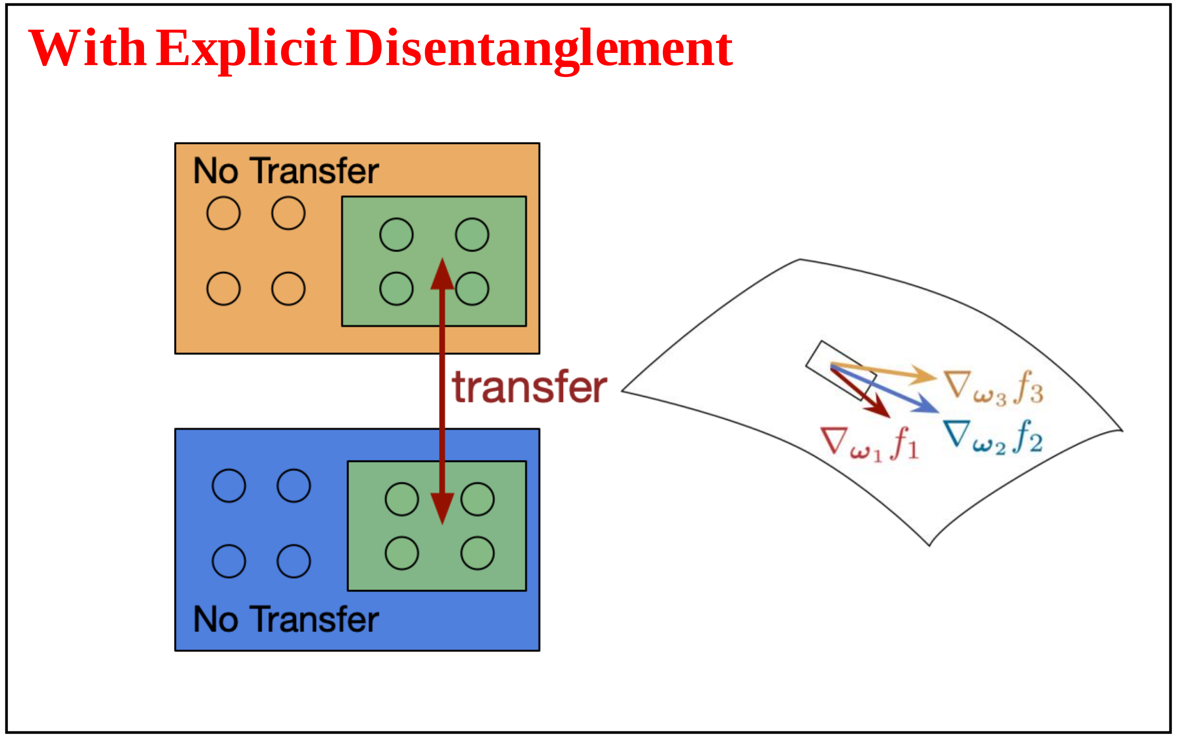


Purification

- **Multiple disentangled latent subspaces**



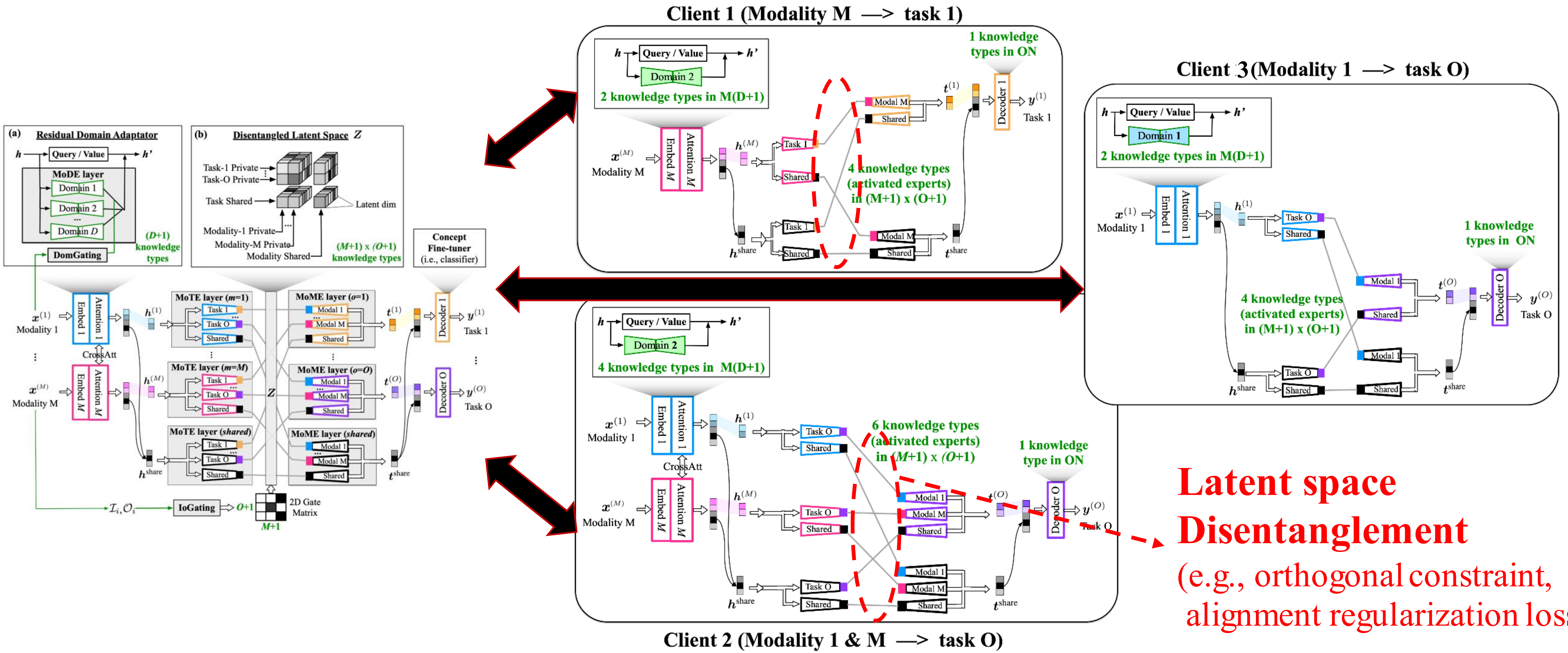
v.s



A more purified knowledge split is beneficial to produce more aligned gradients.

Purification

- **Leverage explicit disentanglement losses to enhance purification**



Latent space Disentanglement

(e.g., orthogonal constraint, alignment regularization loss)

Datasets/Simulations

- #clients (<50), #modalities (<5), #downstream tasks (<5)
- model size: 4 self/cross-attention layers, 3 heads

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)}

“Multimedia Understanding”

$$
\begin{aligned}
&3\rho \times (\mathcal{X}_{\{\text{video}\}} \rightarrow \mathcal{Y}_{\{\text{sentiment}\}})\\
&3\rho \times (\mathcal{X}_{\{\text{audio}\}} \rightarrow \mathcal{Y}_{\{\text{sentiment}\}})\\
&3\rho \times (\mathcal{X}_{\{\text{text}\}} \rightarrow \mathcal{Y}_{\{\text{sentiment}\}})\\
&3\rho \times (\mathcal{X}_{\{\text{video, audio}\}} \rightarrow \mathcal{Y}_{\{\text{sentiment}\}})\\
&3\rho \times (\mathcal{X}_{\{\text{audio, text}\}} \rightarrow \mathcal{Y}_{\{\text{sentiment}\}})\\
&3\rho \times (\mathcal{X}_{\{\text{video, text}\}} \rightarrow \mathcal{Y}_{\{\text{sentiment, emotions}\}})\\
&3\rho \times (\mathcal{X}_{\{\text{text}\}} \rightarrow \mathcal{Y}_{\{\text{emotions}\}})\\
&3\rho \times (\mathcal{X}_{\{\text{video, audio}\}} \rightarrow \mathcal{Y}_{\{\text{emotions}\}})\\
&3\rho \times (\mathcal{X}_{\{\text{audio, text}\}} \rightarrow \mathcal{Y}_{\{\text{sentiment, emotions}\}})\\
&3\rho \times (\mathcal{X}_{\{\text{audio}\}} \rightarrow \mathcal{Y}_{\{\text{cls\_vehicle}\}})\\
&3\rho \times (\mathcal{X}_{\{\text{seismic}\}} \rightarrow \mathcal{Y}_{\{\text{cls\_vehicle}\}})\\
&10(1-\rho) \times (\mathcal{X}_{\{\text{audio, seismic}\}} \rightarrow \mathcal{Y}_{\{\text{cls\_vehicle}\}})\\
&10(1-\rho) \times (\mathcal{X}_{\{\text{video, text, audio}\}} \rightarrow \mathcal{Y}_{\{\text{emotions}\}})
\end{aligned}
$$

Client inputs　Client outputs

Cross-modal Generation & Understanding

$$
\begin{aligned}
&3 \times (\mathcal{X}_{\{\text{image}\}} \rightarrow \mathcal{Y}_{\{\text{cls\_digits}\}})\\
&3 \times (\mathcal{X}_{\{\text{audio}\}} \rightarrow \mathcal{Y}_{\{\text{cls\_digits}\}})\\
&3 \times (\mathcal{X}_{\{\text{image, audio}\}} \rightarrow \mathcal{Y}_{\{\text{cls\_digits}\}})\\
&3 \times (\mathcal{X}_{\{\text{audio}\}} \rightarrow \mathcal{Y}_{\{\text{cls\_digits, gen\_image}\}})\\
&3 \times (\mathcal{X}_{\{\text{image}\}} \rightarrow \mathcal{Y}_{\{\text{gen\_audio}\}})\\
&3 \times (\mathcal{X}_{\{\text{image, audio}\}} \rightarrow \mathcal{Y}_{\{\text{gen\_image, gen\_audio}\}})\\
&3 \times (\mathcal{X}_{\{\text{image}\}} \rightarrow \mathcal{Y}_{\{\text{cls\_digits, cls\_objects}\}})\\
&3 \times (\mathcal{X}_{\{\text{image}\}} \rightarrow \mathcal{Y}_{\{\text{cls\_objects, gen\_audio}\}})\\
&3 \times (\mathcal{X}_{\{\text{image}\}} \rightarrow \mathcal{Y}_{\{\text{cls\_digits, cls\_objects, gen\_image}\}})
\end{aligned}
$$

Client inputs　Client outputs

Some Experimental Results

- Server model size: 22M
- Average of client model sizes: 63% (Quantization: 15%)

Methods	Average Testing Accuracy on Classification Tasks	
Local	88.23 ± 0.72	70.23 ± 0.93
FedAvg	84.63 ± 0.02	74.12 ± 0.93
Multi-FedAvg	84.82 ± 0.29	69.65 ± 0.73
FedMSplit	87.37 ± 0.03	73.25 ± 0.31
Ours	**96.38 ± 0.41**	**75.96 ± 0.83**

$$
\begin{aligned}
&3\rho \times (\mathcal{X}_{\{\text{video}\}} \rightarrow \mathcal{Y}_{\{\text{sentiment}\}})\\
&3\rho \times (\mathcal{X}_{\{\text{audio}\}} \rightarrow \mathcal{Y}_{\{\text{sentiment}\}})\\
&3\rho \times (\mathcal{X}_{\{\text{text}\}} \rightarrow \mathcal{Y}_{\{\text{sentiment}\}})\\
&3\rho \times (\mathcal{X}_{\{\text{video, audio}\}} \rightarrow \mathcal{Y}_{\{\text{sentiment}\}})\\
&3\rho \times (\mathcal{X}_{\{\text{audio, text}\}} \rightarrow \mathcal{Y}_{\{\text{sentiment}\}})\\
&3\rho \times (\mathcal{X}_{\{\text{video, text}\}} \rightarrow \mathcal{Y}_{\{\text{sentiment, emotions}\}})\\
&3\rho \times (\mathcal{X}_{\{\text{text}\}} \rightarrow \mathcal{Y}_{\{\text{emotions}\}})\\
&3\rho \times (\mathcal{X}_{\{\text{video, audio}\}} \rightarrow \mathcal{Y}_{\{\text{emotions}\}})\\
&3\rho \times (\mathcal{X}_{\{\text{audio, text}\}} \rightarrow \mathcal{Y}_{\{\text{sentiment, emotions}\}})\\
&3\rho \times (\mathcal{X}_{\{\text{audio}\}} \rightarrow \mathcal{Y}_{\{\text{cls\_vehicle}\}})\\
&3\rho \times (\mathcal{X}_{\{\text{seismic}\}} \rightarrow \mathcal{Y}_{\{\text{cls\_vehicle}\}})\\
&10(1-\rho) \times (\mathcal{X}_{\{\text{audio, seismic}\}} \rightarrow \mathcal{Y}_{\{\text{cls\_vehicle}\}})\\
&10(1-\rho) \times (\mathcal{X}_{\{\text{video, text, audio}\}} \rightarrow \mathcal{Y}_{\{\text{emotions}\}})
\end{aligned}
$$

"Multimedia Understanding"

$$
\begin{aligned}
&3 \times (\mathcal{X}_{\{\text{image}\}} \rightarrow \mathcal{Y}_{\{\text{cls\_digits}\}})\\
&3 \times (\mathcal{X}_{\{\text{audio}\}} \rightarrow \mathcal{Y}_{\{\text{cls\_digits}\}})\\
&3 \times (\mathcal{X}_{\{\text{image, audio}\}} \rightarrow \mathcal{Y}_{\{\text{cls\_digits}\}})\\
&3 \times (\mathcal{X}_{\{\text{audio}\}} \rightarrow \mathcal{Y}_{\{\text{cls\_digits, gen\_image}\}})\\
&3 \times (\mathcal{X}_{\{\text{image}\}} \rightarrow \mathcal{Y}_{\{\text{gen\_audio}\}})\\
&3 \times (\mathcal{X}_{\{\text{image, audio}\}} \rightarrow \mathcal{Y}_{\{\text{gen\_image, gen\_audio}\}})\\
&3 \times (\mathcal{X}_{\{\text{image}\}} \rightarrow \mathcal{Y}_{\{\text{cls\_digits, cls\_objects}\}})\\
&3 \times (\mathcal{X}_{\{\text{image}\}} \rightarrow \mathcal{Y}_{\{\text{cls\_objects, gen\_audio}\}})\\
&3 \times (\mathcal{X}_{\{\text{image}\}} \rightarrow \mathcal{Y}_{\{\text{cls\_digits, cls\_objects, gen\_image}\}})
\end{aligned}
$$

Cross-modal Generation & Understanding

Thank You!
Q & A