Adaptive Hyper-graph Aggregation for Modality-Agnostic Federated Learning

Supplementary Material

APPENDIX. In this appendix we introduce the hypergraph neural networks preliminary in Appendix A. The Modular Architecture for Local Model is illustrated in Appendix B. The datasets and segmentation methods used for our experiments are presented in Appendix C. Additional results of the experiment are given in Appendix D. More experimental details are in Appendix E.

A. Hypergraph Preliminary

Hypergraph Neural Networks. Consider an attributed hypergraph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ and $\mathcal{E}$ denote the node and hyperedge sets, respectively. A hyperedge e is defined as a subset of $\mathcal{V}$, $e = \{v_1^{(e)}, \ldots, v_{|e|}^{(e)}\}$, that can connect more than two nodes. The node attribute matrix $X = [\ldots, x_v, \ldots]^T$ belongs to $\mathbb{R}^{K\times N}$, where x_v encapsulates the latent features of node v.

In hypergraph convolutional networks, the adjacency matrix $H \in \mathbb{R}^{|V|\times|E|}$ represents node-to-hyperedge connections, where H_{ij} indicates node v_i's membership in hyperedge e_j. A hyperedge convolutional layer is formulated as:

$$g(X, W, \Theta) : X^{t+1} = \sigma\left(D_v^{-\frac{1}{2}} H W D_e^{-1} H^T D_v^{-\frac{1}{2}} X^t \Theta^t\right) \tag{13}$$

where W is initialized as the identity matrix to assign equal importance to all hyperedges. The feature transformation filter Θ operates on the nodes, and σ denotes a nonlinear activation function. D_e and D_v are the diagonal degree matrices of hyperedges and vertices, respectively.

B. Modular Architecture for Local Model

As shown in Fig.2, we modular each local model into three modules as follows:

- ▷ The **Modality-Specific module**(ψ_c^i) is tailored to capture the unique characteristics of each data modality.
- ▷ The **Modality-Shared module**(ψ_d^i) is specifically designed to learn features common to all modalities. Its learning process is guided and constrained by alignment with the **Global Modality-Shared Prototype** (P_{global}), ensuring consistency and shared knowledge across the network.
- ▷ The **Personalized Interaction module** (ϕ_p^i) module functions as an attention mechanism, guiding the integration of modality-shared features with modality-specific representations to enhance multimodal complementarity.

Modality-Specific module and Modality-Shared module. The Modality-Specific module specializes in intra-modality feature extraction, allowing it to isolate features peculiar to each modality by enabling exclusive interactions among clients sharing the same modality. In contrast, the Modality-Shared module is tailored to detect and analyze complex cross-correlations across clients from different modalities. This cross-modal interaction empowers the module to identify and extract higher-order features that are common across modalities, thereby enriching the feature representation with shared information that cuts across modality-specific boundaries. Therefore, a divergence loss is introduced to produce distinct feature representations between the Modality-Specific and Modality-Shared modules. Inspired by domain separation

Algorithm 1 FHNN

Input: Communication rounds T, number of client K, local datasets $\{D_i\}_{i=1}^K$, Common dataset D_{com}, learning rate η_h, locat step E, Hypergraph model w_h^0, Prototype Enhancer model w_p^0, server model $\{\psi_c^0, \psi_d^0, \phi_p^0\}$

Server broadcasts $\{\psi_c^0, \psi_d^0, \phi_p^0\}$ to all clients
for t=0 to $T-1$ **do**
  for $Client\ i = 1, 2...K$ in parallel **do**
    ClientUpdate$(\{\psi_c^{i,t}, \psi_d^{i,t}, \phi_p^{i,t}\}, D_i, D_{com})$
  end for
  Server executes:
  // Global Consensus Prototype Enhancer
  Calculated to the aggregated prototype p_{agg} according to Eq.(4-6)
  Update w_p^{t+1} according to $\mathcal{L}_{CE} + \lambda \sum_m \mathcal{L}_{proto}^m$
  Computing the Global Consensus Prototype P_{global}^{t+1} on the new model w_p^{t+1}
  // Multimodal Hypergraph Aggregation
  Initialize the generation of the hyperedge $\mathcal{E}$ and Nodes Attributes $\{X_c, X_d\}$
  Update $h_v^{(t+1)}$ according to Eq.(9-10)
  Aggregate local models $\{\psi_c^{i,t+1}, \psi_d^{i,t+1}, \phi_p^{i,t}\}_{i=1}^K$
  Compute $\mathcal{L}_{hnn} = \frac{1}{K}\sum_{i=1}^K (1 - \delta^{acc_i - 1})$
  Update $w_h^{t+1} \leftarrow w_h^t - \eta_h \nabla_{w_h^t} \mathcal{L}_{hnn}$
  Server sends $\{\psi_c^{i,t+1}, \psi_d^{i,t+1}\}_{i=1}^K$ and P_{global}^{t+1}
end for
ClientUpdate$(\{\psi_c^{i,t}, \psi_d^{i,t}, \phi_p^{i,t}\}, D_i, D_{com})$:
for $e = 1$ to E **do**
  Compute $\mathcal{L}(D_i) = \mathcal{L}_{fc} + \lambda_1 \mathcal{L}_{dif}$, and updating local models with private dataset D_k
  Compute $\mathcal{L}'(D_{com}) = \mathcal{L}_{fc} + \lambda_1 \mathcal{L}_{div} + \lambda_2 \mathcal{L}_{NCE}$, and fine-tuning local models with public dataset
  Upload the local model to the server
end for

networks, we adopt a soft subspace orthogonality constraint for this purpose:

$$\mathcal{L}_{div}(D_i) = \sum_{j}^{|D_i|} ||(h_c^i)^T h_d^i||_F^2$$

where $h_c^i \in \mathbb{R}^{|h|}$ is the output features of the Modality-Specific module,$h_c^i = \psi_c^i(x_j^i; \Theta_c^i)$. Similarly, $h_d^i \in \mathbb{R}^{|h|}$ is the output features of the Modality-Shared module, $h_d^i = \psi_d^i(x_j^i; \Theta_d^i)$. $||\cdot||_F$ is Frobenius norm.

While the divergence loss promotes distinct representation spaces, it does not ensure the consistency of Modality-Shared outputs across different modalities within a shared latent space. To rectify this, we introduce the Global Consensus Prototype (detailed in Section 3.4), normalizing the Modality-Shared module's outputs via a contrastive loss, specifically the InfoNCE loss, to augment intra-class compactness and inter-class separability in the shared prototype space:

$$\mathcal{L}_{NCE}(D_i, P_{global}) = -\sum_{j}^{|D_i|} \log \frac{exp(h_d^i \cdot p_{y_j}/\tau)}{\sum_{c=1}^{C} exp(h_d^i \cdot p_c/\tau)}$$

where C denotes the total number of classes. y_j is the ground-truth, and τ is a temperature coefficient.

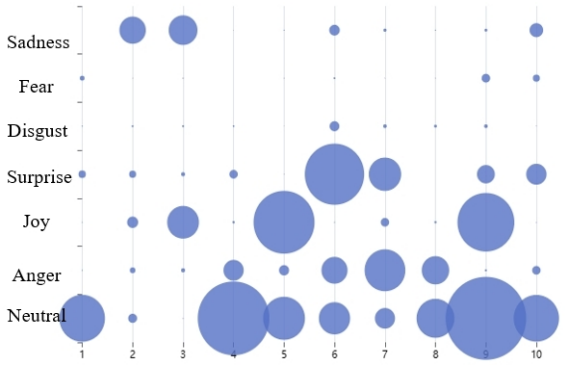
During local training, to avoid the misguidance that might result from directly applying server-derived prototypes which do not account for individual client characteristics, we omit the $\mathcal{L}_{NCE}$ loss. Instead, we harness the public dataset to capture cross-client generic knowledge by jointly computing the $\mathcal{L}_{NCE}$ loss between local data and prototypes, subsequently fine-tuning the local models for enhanced performance.

Personalized Interaction module. Leveraging multimodal attention mechanisms, the Personalized Interaction module is crafted to adeptly fuse cross-modal information. This fusion entails the effective blending of disparate modality-specific and modality-shared features, culminating in a rich, integrated representation. The module's training harnesses a cross-entropy loss function, defined as:

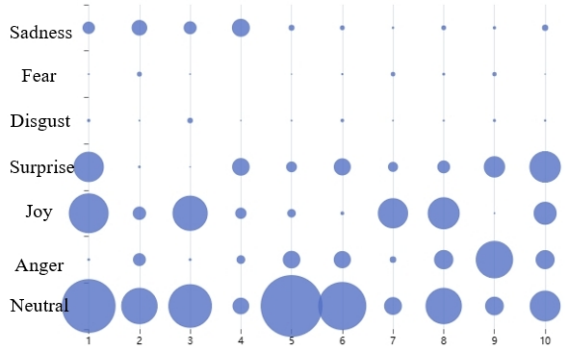
$$\mathcal{L}_{fc}(D_i) = -\sum_{j}^{|D_i|} y_j log[\phi_p^i(h_c^i, h_d^i; \Theta_c^i)]$$

where y_j is the ground-truth. ϕ_p^i is the modality fusion attention module of i-th client and Θ_c^i is the parameter.

In the end, the training phase for the local client is divided into two main parts: the client uses its own private dataset $\{D_i\}$ to update the model; the public dataset D_{com} is used to fine-tune the updated model. The only difference between these two parts is the loss function: $\mathcal{L}(D_i) = \mathcal{L}_{fc} + \lambda_1 \mathcal{L}_{div}$ and $\mathcal{L}'(D_{com}) = \mathcal{L}_{fc} + \lambda_1 \mathcal{L}_{div} + \lambda_2 \mathcal{L}_{NCE}$. We summarize the optimization steps of HAMFL in Algorithm 1.



(a) $\alpha = 0.2$



(b) $\alpha = 1$

Figure 7. Illustration of MELD Non-IID data distributions over 10 clients with $\alpha = 0.2$ and $\alpha = 1$. The x-axes represents the client IDs. The y-axes represents the emotion labels. The dot sizes represent the number of data.

C. Datasets

In this section we introduce the datasets and data partition method used in this paper.

EPIC-Kitchens. The dataset is a large-scale egocentric video dataset collected from daily kitchen activities. It contains over 100 hours of videos collected from 37 home cooks, with over 90K action instances across 45 kitchen environments. In our experiments, we are using two modalities of this dataset to train and test. We only use data segments from EPIC-Kitchens that contain video and audio modalities, which contains 89K action recognition segments. The test set we use the official test set division for testing.

UCF-101. The UCF101 dataset is a popular action recognition dataset collected from YouTube videos. It consists of 13,320 video clips distributed across 101 human action categories, ranging from daily life to sports activities. We take the test1.txt of the action recognition task in the dataset file to divide the test set and the rest of the data as the training set to divide the client. The duration of the videos ranges from several seconds to over 20 seconds.

MELD. Multimodal EmotionLines Dataset (MELD) has

Methods	Times(seconds)		
	EPIC-Kitchens	UCF-101	MELD
SingleSet	0.21±0.11	0.08±0.05	0.17±0.09
FedAVG	0.39±0.24	0.19±0.15	0.25±0.13
FedProto	0.57±0.36	0.45±0.27	0.62±0.45
FedLAW	1.12±0.81	0.57±0.51	1.56±0.37
pFedGraph	2.79±1.22	1.44±0.52	2.14±1.08
FedMSplit	5.37±2.25	3.49±1.17	4.82±2.24
FDARN	6.36±2.78	3.61±1.64	5.93±2.69
CreamFL	5.49±2.83	2.27±1.31	5.18±3.82
Ours	7.29±3.53	4.39±1.92	6.72±2.75

Table 3. Average one training run time for clients of different methods.

been created by enhancing and extending EmotionLines dataset. MELD contains the same dialogue instances available in EmotionLines, but it also encompasses audio and visual modality along with text. MELD has more than 1400 dialogues and 13000 utterances from Friends TV series. Multiple speakers participated in the dialogues. Each utterance in a dialogue has been labeled by any of these seven emotions – Anger, Disgust, Sadness, Joy, Neutral, Surprise and Fear.

Data Partition. Each client is allocated a proportion of the samples of each label according to Dirichlet distribution. In detail, we sample the data by simulating $p_j \sim Dir(\alpha)$ and allocate a portion of $p_{j,i}$ of the samples in class j to client i. Here α controls the degree of skewness. Note that when using this partitioning strategy, the training data of each client may have majority classes, minority classes, or even some missing classes, which is more practical in real-world applications. See Fig.7 for the detailed two Non-IID ($\alpha = 0.2, \alpha = 1$) data partitions on MELD datasets.

D. Additional Experimental Results

Acquisition Time. In Tab.3, we report the average training times of our method compared with others. Due to the introduction of hypergraph neural networks and prototype learning networks on the server side, HAMFL incurs increased training time. However, compared with other FL algorithms, the computational overhead of HAMFL is acceptable, especially on UCF-101.

Ablation Studies. Here, we conduct an ablation study on three datasets to demonstrate the effectiveness of HAMFL's the Global Consensus Prototype Enhancer (GCPE) and the Multimodal Hypergraph Aggregation (MHA), as well as the Modality Speculative Domain (MSD) and Distributional Speculative Domain (DSD) in the hypergraph structure. The quantitative analysis results are shown in Tab.4. In the first row, we remove the GCPH, which corresponds to the removal of the NCE loss function

Method	EPIC-Kitchens	UCF-101	MELD
ω/o *GCPE*	40.2	71.5	52.3
ω/o *MSD*	39.1	68.9	51.9
ω/o *DSD*	40.6	71.4	52.6
ω/o *MHA*	38.6	69.2	51.4
Ours	**41.6**	**73.3**	**54.1**

Table 4. Ablation studies on three datasets.

for client training. In the second and third rows, we remove MSD and DSD, respectively. In the fourth row, we remove MHA and use Fedavg to aggregate the global model.

Performance drops on all three datasets after removing either GCPE or MHA, confirming their importance. The larger decrease caused by removing MHA suggests its greater effectiveness in multimodal client model aggregation. Additionally, removing MSD and DSD also reduces performance, indicating the validity of these domain designs. The more significant decline with MSD removal across datasets points to challenges in cross-modality model aggregation and potential issues with model convergence.

E. More Experimental Details

For visual data, we use MobileNetV2 as a feature extraction network to extract potential features. For text data, we use MobileBERT to extract representations from textual data. For audio data, we extract audio representations using the current widely used Wav2Vec 2.0 speech recognition model. Optical flow data obtains potential feature representations by a pre-training I3D model. The code will be made available publicly at : github.com/MM-Fed/HAMFL.