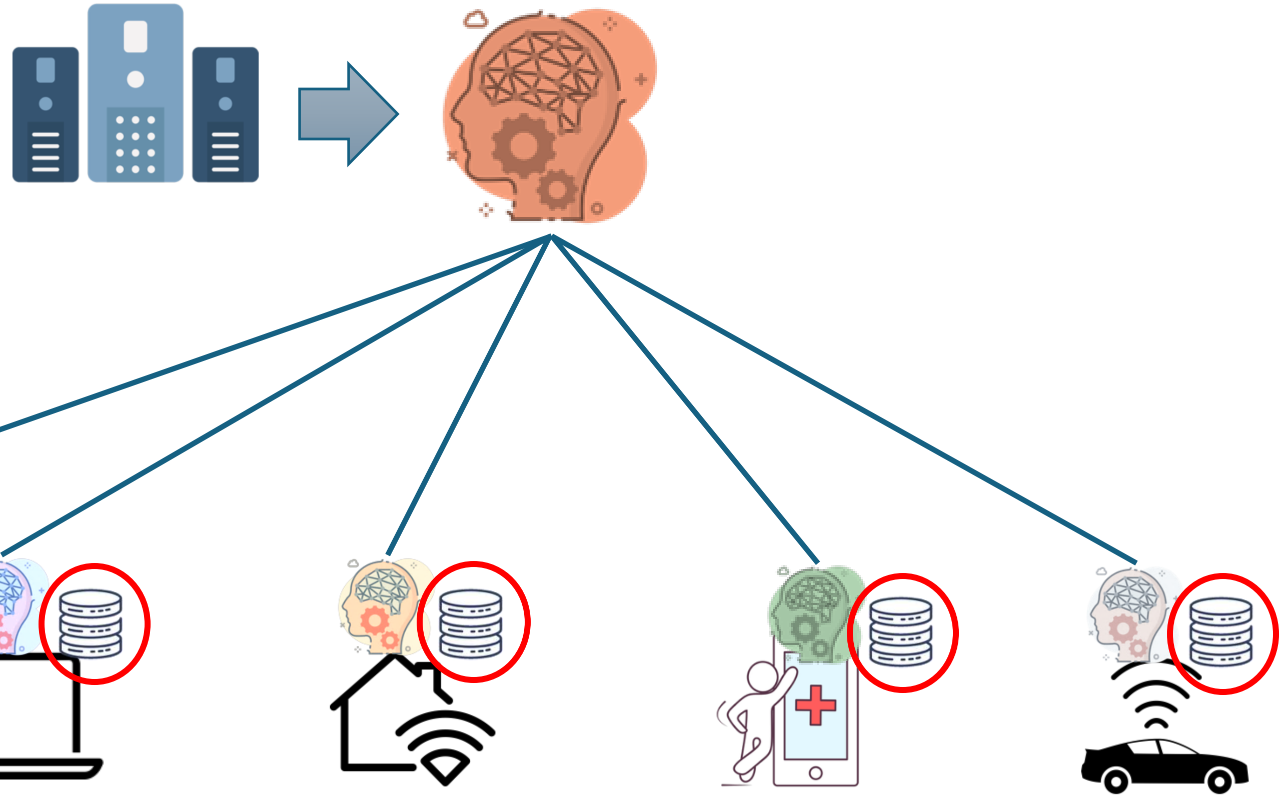
Federated Prompt-Tuning with Heterogeneous and Incomplete Multimodal Client Data

Thu Hang Phung[1], Duong M. Nguyen[1], Thanh Trung Huynh[2], Quoc Viet Hung Nguyen[3], Trong Nghia Hoang[4], Phi Le Nguyen[1]

[1]Institute for AI Innovation and Societal Impact, Hanoi University of Science and Technology, [2]EPFL ,
[3]Griffith University, [4]Washington State University

Federated Learning (FL)



Same type of data in clients

Types of Multimodal Dataset

Text-only Dataset
All samples are texts

Complete Dataset
All samples are complete

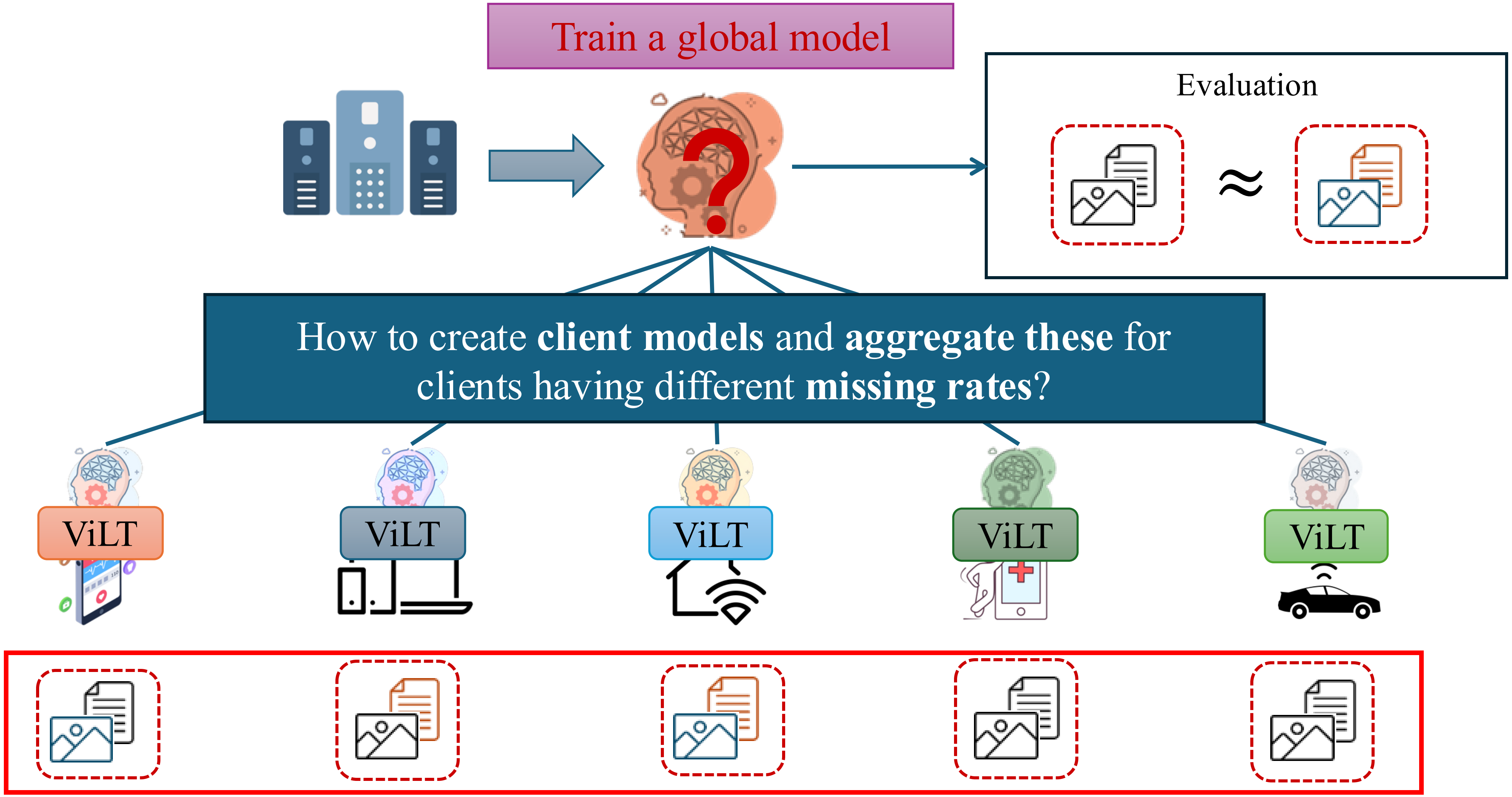
Miss-image Dataset
Some samples are complete
The rest are text-only

Image-only Dataset
All samples are images

Miss-both Dataset
Some samples are complete
The rest are image-only and text-only

Miss-text Dataset
Some samples are complete
The rest are image-only

Modality Missing in FL – ViLT[1]



1. Kim, W., Son, B. and Kim, I. Vilt: Vision-and-language transformer without convolution or region supervision. ICML 2021

Related Works

FL methods

- Solve modality missing but **not utilizing pretrained models** in clients
- Address **either** inter-heterogeneity or intra-heterogeneity

Apply Centralized method [1] to FL

Lead to inter-client heterogeneity (local prompts not aligned when aggregates in the server)

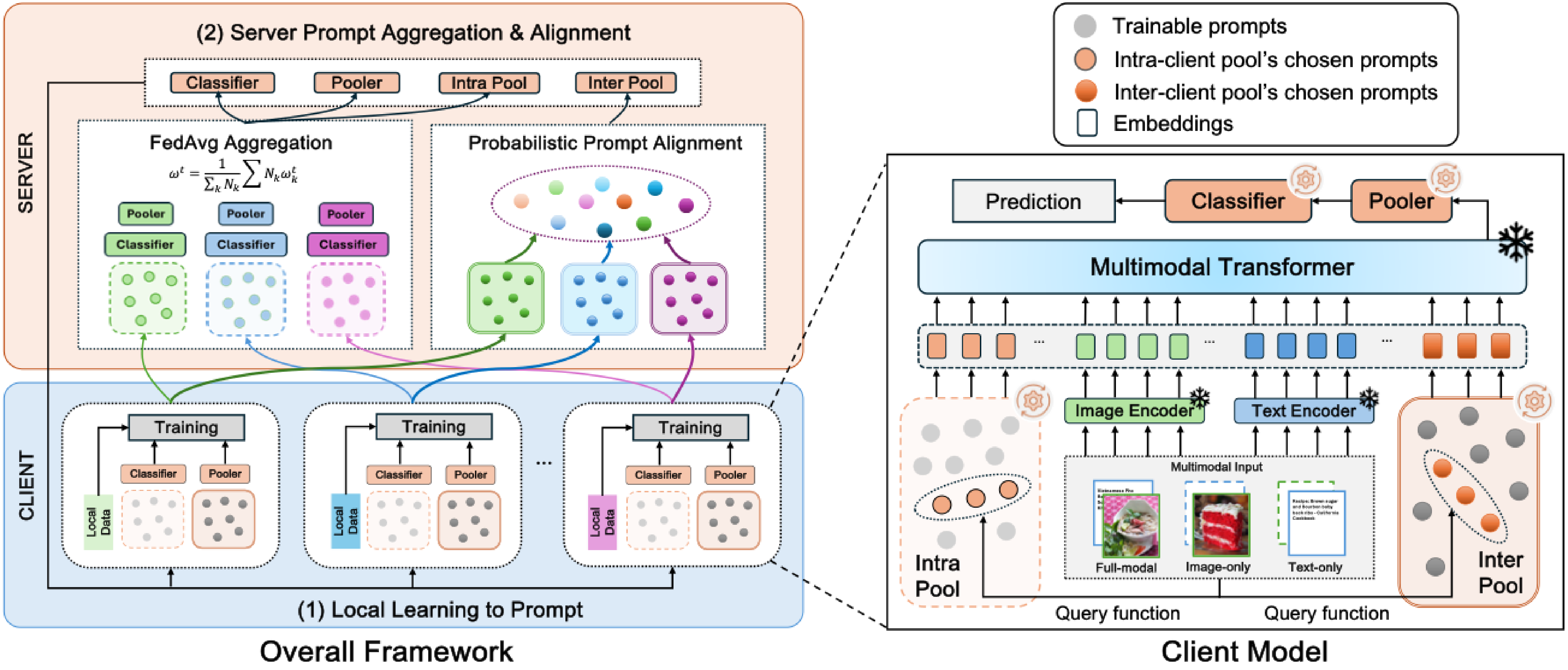
FED-PRIME

✓ State **a new problem** of modality missing in federated prompt-tuning

✓ Propose a method to aggregate 2 types of prompt to **solve 2 problems**: Address both **inter- and intra-client heterogeneities**

1. Lee, Y.L., Tsai, Y.H., Chiu, W.C. and Lee, C.Y., 2023. Multimodal Prompting with Missing Modalities for Visual Recognition. CVPR 2023.

Fed-Prime Overview



Local Training Objectives

Given input embedding after concatenated with selected inter-client prompts and intra-client prompts

$$Input_{augmented} = F_e(x) \circ w_p^{inter} \circ w_p^{intra}$$

Task loss for client t given m data points, F_c and F_p are classifier (updatable) and ViLT frozen encoder; w_c is the classifier weight, and $z_{t,s}$ is sample's label

$$L_t(w) = \frac{1}{m}\sum_{t=1}^{m} l\left(F_c\left(F_p\left(Input_{augmented}\right); w_c\right), z_{t,s}\right)$$

Prompt relevant loss (contrastive)

$$R_t = -\frac{1}{m}\sum_{s=1}^{m}\left[S_{pos} - S_{neg}\right]$$

$$S_{pos} = \sum_{i \in selected} \log\left(\sigma\left(q.k(p_i)\right)\right); S_{neg} = \sum_{i \in unselected} \log\left(\sigma\left(-q.k(p_i)\right)\right)$$

Final loss

$$L_{total} = L_t + \lambda R_t$$

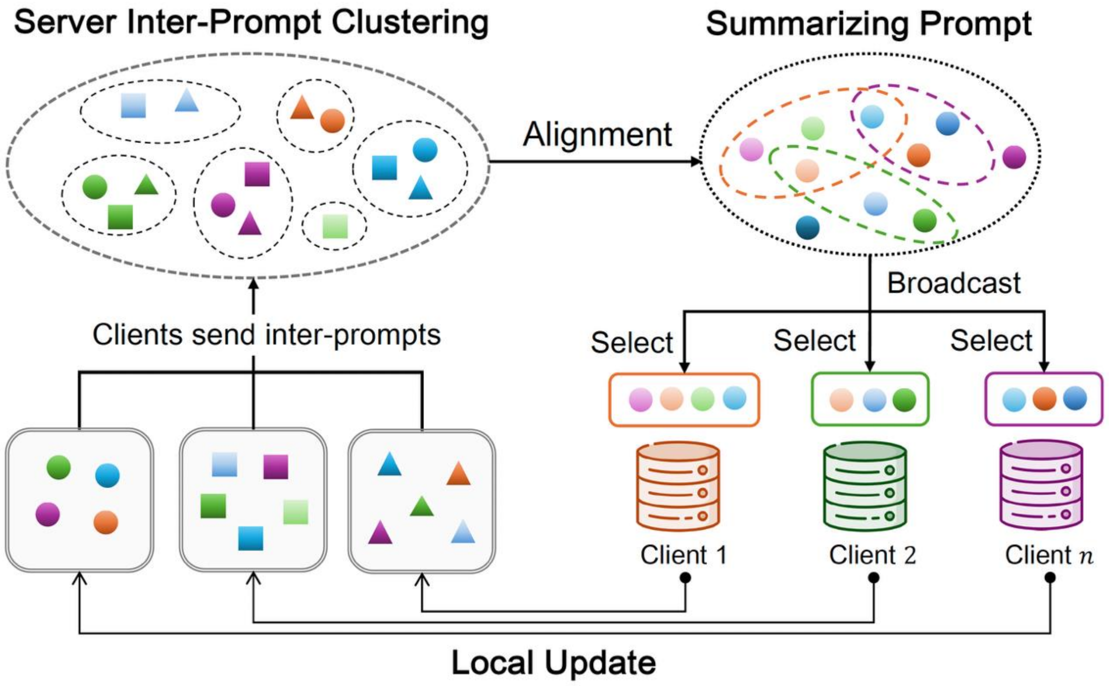
Inter-client Prompt Alignment - Server

Motivation

- Prompt positions can encode different meanings due to **client heterogeneity** (e.g., missing data, different tasks).
- Simply averaging prompts **breaks semantic alignment** and hurts performance.
- We need to **group similar prompts** across clients before aggregation.

How It Works:

1. Receive inter-prompts from all clients.
2. Group prompts by semantic similarity (clustering).
3. Alignment: Create a global (**summarizing prompt pool**) using the **clusters' centroids**
4. Drop empty or unused clusters.
5. Broadcast to client to be the **next inter-client prompt pool**



Experiment Results

Datasets		UPMC Food-101					MM-IMDB				
Train	Method	Test (∼ Train)	Test (Miss Both)	Test (Full Modal)	Test (Text only)	Test (Image only)	Test (∼ Train)	Test (Miss Both)	Test (Full Modal)	Test (Text only)	Test (Image only)
Miss Text	FEDAVG-P	15.71 ± 2.27	14.90 ± 1.57	21.56 ± 7.81	16.91 ± 0.69	15.36 ± 0.43	22.42 ± 2.27	21.89 ± 1.34	30.78 ± 1.65	18.40 ± 0.06	14.53 ± 4.56
	FEDMSPLIT-P	15.62 ± 1.51	17.50 ± 1.97	25.27 ± 8.22	18.78 ± 0.64	17.50 ± 1.97	21.02 ± 1.89	19.97 ± 0.74	24.39 ± 6.08	14.38 ± 2.76	18.09 ± 6.13
	FED-INTER	54.82 ± 19.01	48.87 ± 24.64	59.17 ± 27.06	35.13 ± 26.78	56.59 ± 15.12	18.25 ± 3.50	16.95 ± 3.57	18.67 ± 7.63	15.03 ± 4.66	18.01 ± 1.95
	FED-INTRA	61.71 ± 17.22	48.09 ± 19.12	62.06 ± 26.98	22.51 ± 5.92	62.64 ± 11.83	13.38 ± 1.73	12.77 ± 0.85	12.55 ± 1.67	11.31 ± 0.38	14.33 ± 1.80
	FED-PRIME	**78.88 ± 0.90**	**80.38 ± 0.65**	**92.12 ± 0.40**	**73.01 ± 4.25**	**76.83 ± 1.22**	**31.92 ± 0.20**	**31.48 ± 0.30**	**37.67 ± 0.04**	**30.60 ± 1.41**	**30.69 ± 1.41**
	Improv. (%)	**27.82 ↑**	**64.48 ↑**	**48.44 ↑**	**107.83 ↑**	**22.65 ↑**	**42.37 ↑**	**43.81 ↑**	**22.35 ↑**	**66.30 ↑**	**69.65 ↑**
Miss Image	FEDAVG-P	17.35 ± 4.77	15.12 ± 1.48	16.84 ± 2.37	18.12 ± 6.49	14.81 ± 0.24	27.69 ± 5.97	22.55 ± 3.06	31.94 ± 0.98	23.76 ± 11.72	12.29 ± 0.47
	FEDMSPLIT-P	74.16 ± 10.56	48.88 ± 10.26	45.64 ± 32.43	88.65 ± 2.17	14.81 ± 0.90	19.11 ± 11.33	16.61 ± 7.22	18.19 ± 12.55	18.12 ± 12.30	12.81 ± 1.25
	FED-INTER	77.96 ± 11.62	64.62 ± 10.22	82.08 ± 7.75	77.69 ± 12.35	37.56 ± 6.49	18.79 ± 5.23	17.93 ± 3.60	20.56 ± 3.56	17.67 ± 6.79	15.47 ± 2.51
	FED-INTRA	22.84 ± 3.52	20.13 ± 1.72	23.48 ± 1.86	24.46 ± 2.99	16.66 ± 1.32	15.75 ± 4.34	14.06 ± 3.16	15.68 ± 4.77	14.53 ± 3.65	11.71 ± 0.42
	FED-PRIME	**90.55 ± 0.22**	**79.12 ± 0.49**	**92.89 ± 0.21**	**90.18 ± 0.29**	**54.14 ± 2.50**	**36.08 ± 0.35**	**31.35 ± 0.61**	**38.49 ± 0.56**	**36.91 ± 0.59**	**18.15 ± 0.66**
	Improv. (%)	**16.15 ↑**	**22.44 ↑**	**13.17 ↑**	**1.73 ↑**	**44.14 ↑**	**30.30 ↑**	**39.02 ↑**	**20.51 ↑**	**55.35 ↑**	**17.32 ↑**
Miss Both	FEDAVG-P	14.57 ± 1.50	-	17.17 ± 4.37	16.40 ± 4.05	13.24 ± 0.32	26.45 ± 2.63	-	33.03 ± 2.56	24.12 ± 11.30	20.21 ± 1.98
	FEDMSPLIT-P	49.15 ± 24.76	-	64.78 ± 36.62	64.62 ± 36.51	21.49 ± 7.19	24.25 ± 5.02	-	26.05 ± 11.17	26.02 ± 9.64	19.79 ± 6.20
	FED-INTER	56.32 ± 21.77	-	69.57 ± 19.41	45.15 ± 34.09	59.30 ± 10.84	26.53 ± 0.90	-	31.97 ± 2.22	29.69 ± 2.21	21.63 ± 0.77
	FED-INTRA	49.28 ± 32.87	-	56.70 ± 37.90	43.24 ± 34.19	49.85 ± 25.44	11.90 ± 0.37	-	12.47 ± 0.45	11.46 ± 0.33	12.83 ± 0.92
	FED-PRIME	**84.44 ± 2.65**	-	**93.64 ± 0.58**	**87.95 ± 0.91**	**72.41 ± 3.88**	**32.01 ± 2.51**	-	**38.68 ± 0.65**	**31.00 ± 2.97**	**26.01 ± 0.12**
	Improv. (%)	**49.93 ↑**	-	**34.60 ↑**	**36.10 ↑**	**22.11 ↑**	**20.66 ↑**	-	**17.11 ↑**	**4.41 ↑**	**20.25 ↑**

(*) Improv. shows the relative performance improvement between our proposal and the second-best. (in percentage).

Prompting Analysis

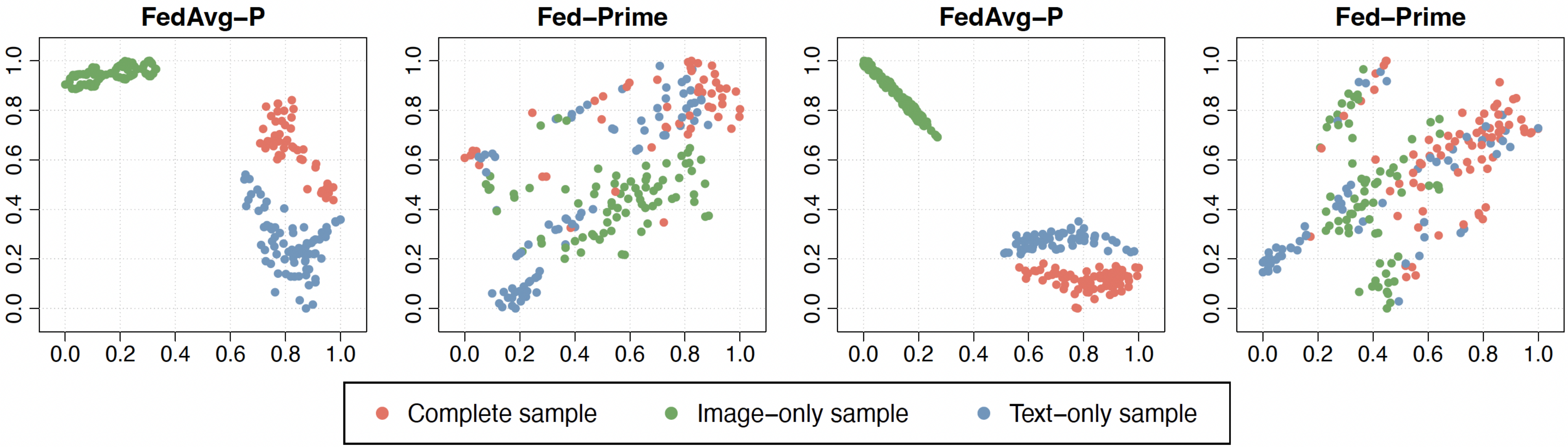


Figure 4.6: t-SNE plots of embeddings prior to classification on MM-IMDB under the **Miss Both** training scenario for Client #4 (left) and Client #14 (right), with two subfigures per client.

Label-skewed NonIID

Dirichlet $\alpha = 0.1$ vs FEDPROX-P[1]

Train	Method	Test (∼ Train)	Test (Miss Both)	Test (Full Modal)	Test (Text only)	Test (Image only)
Miss Text	FEDPROX-P	67.42	64.34	77.29	56.83	68.24
	FED-PRIME	**71.20**	**71.15**	**85.08**	**63.91**	**69.56**
Miss Image	FEDPROX-P	82.56	71.26	85.24	83.05	45.31
	FED-PRIME	**87.38**	**75.59**	**89.25**	**87.05**	**48.47**
Miss Both	FEDPROX-P	75.75	-	89.36	83.98	69.61
	FED-PRIME	**79.98**	-	**91.00**	**86.70**	**70.38**

1. Li, T., Sahu, A.K., Zaheer, M., Sanjabi, M., Talwalkar, A. and Smith, V., 2020. Federated optimization in heterogeneous networks. Proceedings of Machine learning and systems.

Thank you