When a Modality is Missing: A Cross-Modal Recovery for Federated Multimodal Models

Md. Akil Raihan Iftee
Department of Computer Science and Engineering
Khulna University of Engineering & Technology, Bangladesh
iftee1807002@stud.kuet.ac.bd

I. Abstract

Multimodal Federated Learning (MFL) enables collaborative training of models across decentralized clients, each possessing local and private data with multiple modalities. However, in real-world scenarios, clients often suffer from missing modalities due to data collection constraints, privacy concerns, or device limitations. Existing solutions typically struggle with robust performance when one or more modalities are absent, especially under heterogeneous data availability across clients. In this work, we propose a novel framework that disentangles semantic components from multimodal features to facilitate missing modality recovery in a federated setting. Specifically, our approach introduces image feature decomposition into text-aligned and purely visual subspaces, a cross-modal mapper for reconstructing missing text embeddings from images, and a diffusion-based image generator for recovering missing image modalities from text. Furthermore, we incorporate a visual-linguistic memory retrieval mechanism to leverage high-confidence embeddings from previous training steps for zero-shot modality approximation during inference. Our method preserves data privacy, requires minimal communication overhead, and adapts flexibly to different modality configurations.

II. Introduction

Federated Learning (FL) is a decentralized paradigm that enables collaborative model training across multiple clients without exchanging raw data. Its extension to the multimodal domain—Multimodal Federated Learning (MFL)—has opened new frontiers in privacy-preserving learning across diverse modalities like text, image, and audio. However, a key challenge in MFL is the *missing modality* problem, where some clients may lack one or more modalities due to hardware limitations, privacy restrictions, or domain-specific constraints.

Recent studies have explored various techniques to address this issue. Some propose prototype masking and contrastive alignment to deal with incomplete modalities across clients [1]. Others develop cross-model reconstruction networks that adaptively map between modalities [2], or design frameworks tailored to healthcare scenarios with incomplete multimodal data [3]. Benchmarking efforts like FedMultimodal [4] and model-agnostic approaches such as contrastive representation ensembles [5] further highlight the importance of robust modality-incomplete learning.

Despite these advances, several limitations remain. First, most approaches treat modality alignment globally, ignoring the fine-grained semantic relationships within features. Second, they often fail when no parallel modality data is available, limiting their flexibility. Third, they lack mechanisms to recover rich modality-specific features in a privacy-preserving and locally computable way.

To address these gaps, we propose a novel disentangled and generative framework for MFL with missing modalities. Our contributions include:

- A disentangled representation learning approach that splits image embeddings into text-aligned and visual components using orthogonality constraints.
- A cross-modal mapper that reconstructs missing text embeddings directly from disentangled image features, enabling robust adaptation when text is unavailable.
- A diffusion-based image generator that synthesizes image features from text descriptions, supporting clients without image modalities.
- A visual-linguistic memory retrieval (VLMR) mechanism that enables zero-shot modality estimation using high-confidence feature memories.
- A comprehensive federated pipeline integrating these components under strict privacy preservation and efficient communication protocols.

Our framework bridges the gap between local semantic disentanglement and global federated coordination, offering a robust, adaptive, and privacy-preserving solution to multimodal learning with missing modalities.

III. Methodology

A. Problem Setup

We consider a **Federated Learning (FL)** scenario comprising a central server and a set of N distributed

clients $\mathcal{C} = \{c_1, c_2, \ldots, c_N\}$, each possessing a private, potentially incomplete multimodal dataset $\mathcal{D}_i$. In this setup, raw data is never shared between clients and server; only model parameters and latent representations are communicated under strict privacy constraints.

Each client holds three types of elements: a text modality x_t, which may be absent; an image modality x_i, which may also be absent; and a class label y. Based on their data availability, clients are categorized into three groups: those with text only, those with image only, and those with both modalities.

The goal is to collaboratively train a global multimodal classifier that performs robustly even when clients possess only partial modalities, while preserving privacy and ensuring efficient communication. This is achieved by integrating local missing modality adaptation mechanisms with server-coordinated aggregation.

B. Client-Side Learning Pipeline

Each client implements a complete local pipeline for encoding, disentangling, recovering missing modalities, and classification. The steps are described below.

1) Multimodal Encoders: Each client independently encodes the available modalities using pretrained encoders. Given a text input x_t, the text encoder E_t produces an embedding vector $f_t \in \mathbb{R}^d$, while the image encoder E_i processes the image input x_i to produce an embedding $f_i \in \mathbb{R}^{d'}$:

$$f_t = E_t(x_t), \quad f_i = E_i(x_i) \tag{1}$$

These latent feature vectors serve as inputs for subsequent disentanglement and recovery steps.

2) Image Feature Disentanglement: To disentangle text-related semantics and purely visual cues from image embeddings, each client decomposes f_i into two orthogonal components: $f_{i\to t}$ aligned with text semantics, and $f_{i\to v}$ capturing purely visual information:

$$D_i(f_i) = \{f_{i\to t}, f_{i\to v}\}, \quad f_i = f_{i\to t} + f_{i\to v} \tag{2}$$

An orthogonality constraint is applied to ensure these two components capture independent aspects of the image:

$$\mathcal{L}_{\text{orth}} = \left\| f_{i\to t}^{\top} f_{i\to v} \right\|_2 \tag{3}$$

This disentanglement allows for more effective cross-modal mapping and fusion downstream.

3) Cross-Modality Mapping (Image to Text): In scenarios where the text modality is missing, clients reconstruct a text embedding from the text-aligned component of the image embedding. A mapping function M_{it} is trained to transform $f_{i\to t}$ into an estimated text embedding $\hat{f}_t$:

$$\hat{f}_t = M_{it}(f_{i\to t}) \tag{4}$$

The mapping is supervised by minimizing the alignment loss between the reconstructed and true text embeddings:

$$\mathcal{L}_{\text{align}} = \left\| \hat{f}_t - f_t \right\|_2^2 \tag{5}$$

This enables text-absent clients to approximate their missing text features directly from images.

4) Diffusion-Based Image Generation (Text to Image): When image modality is missing, clients employ a pretrained diffusion model $\mathcal{D}_{\text{stable}}$ to generate a synthetic image conditioned on the available text input:

$$\hat{i} = \mathcal{D}_{\text{stable}}(x_t), \quad \hat{f}_i = E_i(\hat{i}) \tag{6}$$

The generated image is encoded to obtain $\hat{f}_i$, and a loss is applied to align these generated features with real image embeddings when they are available:

$$\mathcal{L}_{\text{diff}} = \left\| \hat{f}_i - f_i \right\|_2^2 \tag{7}$$

During fine-tuning, only the cross-attention and text projection layers are updated, while other components such as the VAE and encoders remain frozen to preserve pretrained knowledge.

5) Visual-Linguistic Memory Retrieval (VLMR): Each client maintains a memory bank $\mathcal{M}_y$ of feature pairs (f_t, f_i) for each class y, storing high-confidence embeddings during training:

$$\mathcal{M}_y \leftarrow \mathcal{M}_y \cup \{(f_t, f_i)\}, \quad \text{if confidence} > \tau \tag{8}$$

At inference, when a modality is missing, the classifier predicts a label based on the available features:

$$\hat{y} = \arg\max(\text{Classifier}(f_t)) \tag{9}$$

To estimate the missing image feature, the average of the top-K closest features from memory is used:

$$\hat{f}_i = \frac{1}{K} \sum_{j=1}^{K} f_i^{(j)} \tag{10}$$

The alignment loss ensures consistency between the disentangled visual features and the retrieved estimate:

$$\mathcal{L}_{\text{mem}} = \left\| f_{i\to v} - \hat{f}_i \right\|_2^2 \tag{11}$$

6) Fusion and Classification: All available and recovered features are fused through a multilayer perceptron (MLP) to form the final multimodal feature vector:

$$f_{\text{final}} = \text{MLP}([f_t, \hat{f}_t, f_i, \hat{f}_i]) \tag{12}$$

This fused vector is passed through a classification head, and the classification loss is computed using cross-entropy:

$$\mathcal{L}_{\text{cls}} = \text{CrossEntropy}(H(f_{\text{final}}), y) \tag{13}$$

The total local loss for each client integrates all components:

$$\mathcal{L}_{\text{total}}^{(i)} = \mathcal{L}_{\text{cls}} + \lambda_1 \mathcal{L}_{\text{orth}} + \lambda_2 \mathcal{L}_{\text{align}} + \lambda_3 \mathcal{L}_{\text{diff}} + \lambda_4 \mathcal{L}_{\text{mem}} \tag{14}$$

C. Server-Side Coordination

After each communication round, the server aggregates local model updates using a weighted averaging strategy:

$$\theta_{\text{global}} = \sum_{i=1}^{N} \frac{|\mathcal{D}_i|}{|\mathcal{D}|} \theta^{(i)} \tag{15}$$

Optionally, latent feature projections can be shared for aligning the latent spaces of different clients through diffusion synchronization:

$$\{\text{Proj}(f_t), \text{Proj}(f_i)\} \tag{16}$$

D. Inference Modalities and Recovery

The modality recovery strategy depends on which modalities are available at inference time, as shown in Table I.

Modality Configuration	Used Features	Recovered Modality
Text only	f_t, $\hat{f}_i$	Diffusion or VLMR
Image only	f_i, $\hat{f}_t$	Mapper M_{it}
Both present	f_t, f_i	None required

TABLE I: Client inference strategies by available modality

E. Privacy and Communication

Throughout training, no raw modality data leaves any client. Communication involves encrypted gradients or model weights, and latent features when necessary for diffusion alignment. Optionally, differential privacy mechanisms can be applied to the projections to enhance security.

REFERENCES

[1] G. Bao, Q. Zhang, D. Miao, Z. Gong, L. Hu, K. Liu, Y. Liu, and C. Shi, "Multimodal federated learning with missing modality via prototype mask and contrast," *arXiv preprint arXiv:2312.13508*, 2023.

[2] B. Xiong, X. Yang, Y. Song, Y. Wang, and C. Xu, "Client-adaptive cross-model reconstruction network for modality-incomplete multimodal federated learning," in *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 1241–1249, 2023.

[3] Y. An, Y. Bai, Y. Liu, L. Guo, and X. Chen, "A multimodal federated learning framework for modality incomplete scenarios in healthcare," in *International Symposium on Bioinformatics Research and Applications*, pp. 245–256, Springer, 2024.

[4] T. Feng, D. Bose, T. Zhang, R. Hebbar, A. Ramakrishna, R. Gupta, M. Zhang, S. Avestimehr, and S. Narayanan, "Fedmultimodal: A benchmark for multimodal federated learning," in *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 4035–4045, 2023.

[5] Q. Yu, Y. Liu, Y. Wang, K. Xu, and J. Liu, "Multimodal federated learning via contrastive representation ensemble," *arXiv preprint arXiv:2302.08888*, 2023.