EXTENDIBLE MULTI-STAGE FEDERATED MULTI-MODAL LEARNING WITH MISSING MODALITIES

Anonymous authors
Paper under double-blind review

ABSTRACT

In federated multi-modal learning, communicating the entire model every round increases communication cost and, more importantly, limits how easily the framework can be extended, even though handling missing modalities, optimizing the downstream task, and, in some settings, training the encoders are each essential parts of the pipeline. Rather than a single jointly trained model, this paper proposes a multi-stage federated learning view in which the component that handles missing modalities is separated from the downstream task: the earliest rounds are devoted to optimizing the missing-modality handler alone, with only that component communicated between server and clients, while the remaining rounds are devoted to optimizing the downstream task alone, with only that component communicated instead. When pretrained encoders are not available, this view extends naturally to a third stage in which the initial rounds are devoted to training the encoders themselves. Separately, this paper proposes the missing-modality handler itself, which reconstructs each missing modality from whichever modalities are present. The resulting framework is supported by both theoretical and experimental results. Being task-agnostic, the handler applies directly to regression, classification, and multi-label classification alike, and its F1-micro performance exceeds that of state-of-the-art baselines in the majority of evaluated cases. Code is available at `https://github.com/for-reviewers-arz/FArms_wArms/tree/main`.

1 INTRODUCTION

Multimodal learning benefits from combining complementary signals across data sources, but real-world deployments routinely suffer missing modalities from sensor failure, privacy limits, or incomplete collection Ma et al. (2022; 2021); Dai et al. (2025). Centralized approaches address this via generative reconstruction Liu et al. (2024); Zhao et al. (2021); Wu et al. (2021), adaptive prompting Lee et al. (2023), or expert mixtures Li et al. (2025), while federated multimodal learning (MFL) extends these ideas to privacy-preserving distributed settings through prototypes Bao et al. (2023), reconfigurable representations Nguyen et al. (2025), and expert routing Yun et al. (2024); Orzikulova et al. (2024). Most MFL methods, however, communicate the entire model every round, coupling the missing-modality handler to the downstream task and limiting extendibility, and rely on indirect alignment rather than explicit per-sample reconstruction.

The proposed FArms addresses this through two separable choices: (i) a multi-stage federated schedule that separates the missing-modality handler from the downstream task, communicating only one block per round, with a natural extension to a third stage for encoder training; and (ii) a handler built from **sequential simulating arms**, which reconstruct each absent modality from an already-completed prefix under a fixed ordering, trained label-free with a case-aware ℓ_2 loss. The result is a task-agnostic, extendible framework at reduced communication cost. Our contributions:

- A multi-stage federated schedule separating the missing-modality handler from the task head, reducing communication cost and extending naturally to additional stages.
- A missing-modality handler using sequential simulating arms, requiring only $3K - 4$ trainable modules independent of the number of availability patterns, trained without labels.
- Demonstrated task-agnostic applicability across multi-label classification (MM-IMDb), regression (CMU-MOSI), and classification (Sleep-EDF).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

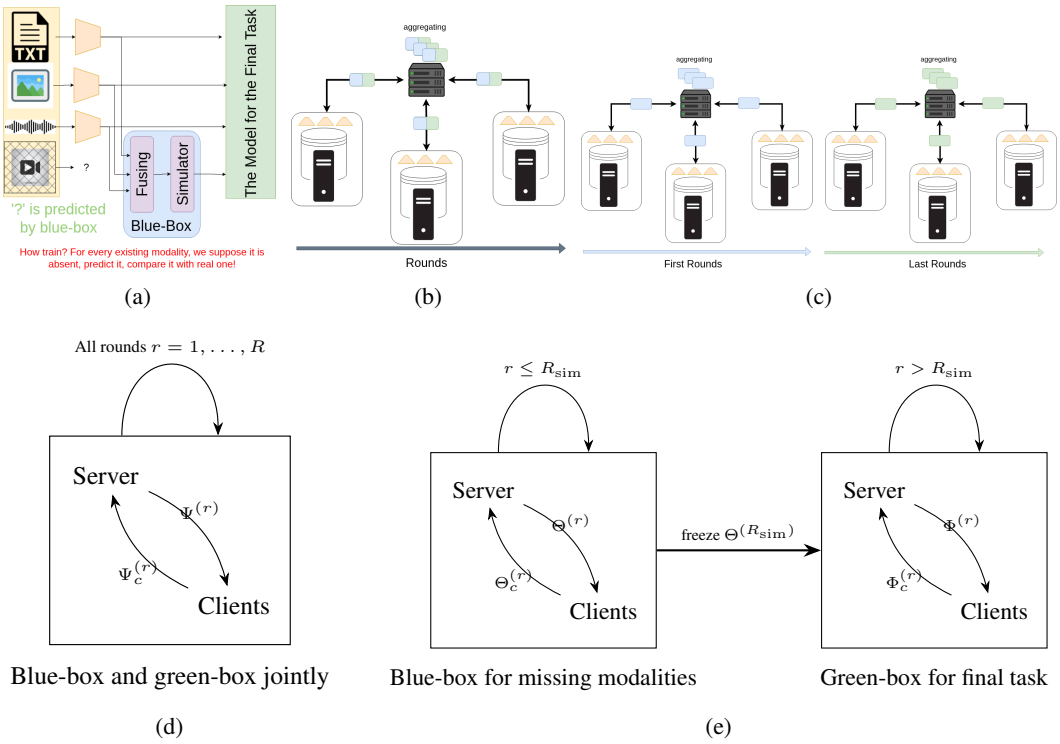


Figure 1: (a) The missing-modality handler (blue box, parameters Θ) is separated from the final task (green box, parameters Φ). (b) In ordinary federated optimization both boxes are updated and communicated in every round. (c) In the proposed schedule the blue box is trained in the initial rounds and the green box in the remaining ones. (d) Communication pattern of ordinary federated learning: the full parameter set $\Psi = \Theta \cup \Phi$ is exchanged in all R rounds. (e) Communication pattern of the proposed two-stage schedule: only Θ is exchanged while $r \le R_{\text{sim}}$, and only Φ afterwards.

- A comparative study against strong centralized and federated baselines — ShaSpec Wang et al. (2023), M^3Care Zhang et al. (2022), PmcmFL Bao et al. (2023), FLEX-MoE Zhang et al. (2025), and PEPSY Nguyen et al. (2025) — under heavy, heterogeneous missing-modality regimes.

So, section 2 introduces the multi-stage federated view. Section 3 presents the sequential simulating-arms construction. Section 4 gives the accompanying theoretical results. Section 5 reports experiments and analysis, Section 6 reviews related work, and Section 7 concludes.

2 Multi-Stage Federated Multi-Modal Learning

We consider N clients indexed by $c \in \{1, \ldots, N\}$, where client c holds a private dataset $\mathcal{D}_c$ of size $n_c = |\mathcal{D}_c|$ that is never shared. Each sample in $\mathcal{D}_c$ carries its own available-modality set $\mathcal{A} \subseteq \{1, \ldots, K\}$, and the distribution of $\mathcal{A}$ may differ across clients.

The model is partitioned into two disjoint parameter blocks, illustrated in Figure 1a. The *blue box* Θ comprises the simulating arms and their prefix-fusion modules, which map the observed embeddings $\{\mathbf{M}'_k\}_{k \in \mathcal{A}}$ to a complete set $\{\mathbf{C}_k\}_{k=1}^K$ as specified in Section 3. The *green box* Φ comprises the final fusion module and the task head, which map $\{\mathbf{C}_k\}_{k=1}^K$ to a prediction. The frozen encoders E_k are held locally and are never transmitted. Writing $\Psi = \Theta \cup \Phi$ for the full parameter set, we have $\Theta \cap \Phi = \emptyset$, and the two blocks are driven by two different objectives: Θ by the label-free reconstruction loss $\mathcal{L}_{\text{sim}}(\Theta; \mathcal{D}_c)$ of Section 3, and Φ by the task loss $\mathcal{L}_{\text{task}}(\Phi; \Theta, \mathcal{D}_c)$, which depends on Θ only through the completed representations it receives.

This separation admits a schedule that ordinary federated optimization does not. In the baseline protocol of Figures 1b and 1d, every round r transmits the entire Ψ in both directions, and the local objective is the weighted sum $\alpha\mathcal{L}_{\text{sim}} + \beta\mathcal{L}_{\text{task}}$ with $\alpha, \beta > 0$. In the proposed protocol of Figures 1c

and 1e, the total budget of R rounds is split at a round $R_{\text{sim}} < R$, and exactly one block is optimized and communicated on either side of the split.

Formally, the server initializes $\Theta^{(0)}$ and $\Phi^{(0)}$ and, at each round $r = 1, \dots, R$, samples a client subset $\mathcal{C}^{(r)} \subseteq \{1, \dots, N\}$ with $|\mathcal{C}^{(r)}| = \lceil qN \rceil$ for a participation rate $q \in (0, 1]$. Let $\Omega^{(r)}$ denote the block that is active in round r,

$$\Omega^{(r)} = \begin{cases} \Theta, & r \le R_{\text{sim}} \quad \text{(stage 1)}, \\ \Phi, & r > R_{\text{sim}} \quad \text{(stage 2)}, \end{cases} \qquad \mathcal{L}^{(r)} = \begin{cases} \mathcal{L}_{\text{sim}}, & r \le R_{\text{sim}}, \\ \mathcal{L}_{\text{task}}, & r > R_{\text{sim}}. \end{cases} \tag{1}$$

The server broadcasts only $\Omega^{(r-1)}$ to the selected clients. Each $c \in \mathcal{C}^{(r)}$ performs P local epochs of stochastic gradient descent on $\mathcal{L}^{(r)}$ with respect to Ω alone, holding the inactive block fixed at its current global value, and returns the resulting $\Omega_c^{(r)}$. The server aggregates by dataset size,

$$\Omega^{(r)} = \sum_{c \in \mathcal{C}^{(r)}} \frac{n_c}{\sum_{c' \in \mathcal{C}^{(r)}} n_{c'}} \, \Omega_c^{(r)}, \tag{2}$$

and the inactive block is carried forward unchanged. At the transition $r = R_{\text{sim}} + 1$ the arms are frozen globally at $\Theta^{(R_{\text{sim}})}$, so that throughout stage 2 every client completes its samples with an identical imputation map and the only remaining source of client disagreement is the task block itself. Raw data and raw embeddings never leave a client in either stage, and the frozen encoders guarantee that no modality-specific input representation is exposed.

3 Missing-Modality Imputation with Sequential Simulating Arms

Let $\mathcal{M} = \{\mathbf{M}_1, \dots, \mathbf{M}_K\}$ be a set of K modalities with input spaces $\mathcal{X}_1, \dots, \mathcal{X}_K$, and let $\mathcal{A} \subseteq \{1, \dots, K\}$, $\mathcal{A} \neq \emptyset$, denote the index set of the modalities that are *available* for a given sample; $\mathcal{A}$ may vary arbitrarily from sample to sample. Imputation is performed in representation space: each modality is equipped with an encoder $E_k : \mathcal{X}_k \to \mathbb{R}^{d_k}$ producing $\mathbf{M}'_k = E_k(\mathbf{m}_k)$, evaluated only for $k \in \mathcal{A}$. In our instantiation the encoders are pretrained and frozen, so that $\{\mathbf{M}'_k\}_{k \in \mathcal{A}}$ is a fixed function of the observed inputs, but nothing below depends on this choice. The module we describe takes $\{\mathbf{M}'_k\}_{k \in \mathcal{A}}$ and returns a complete set $\{\mathbf{C}_k\}_{k=1}^K$ with $\mathbf{C}_k \in \mathbb{R}^{d_k}$ for every k, which any downstream consumer may use as if no modality were absent. It is defined and trained without reference to any downstream objective.

We fix once and for all a permutation π of the modality indices, drawn at random, and relabel the modalities according to it, referring to them in the ordered sequence $\mathbf{M}_1, \mathbf{M}_2, \dots, \mathbf{M}_K$. The ordering is global: it is identical for every sample, at training and at inference. Completion then proceeds strictly left to right along this sequence, and the defining property of the construction is that position k is reconstructed only from the *already-completed* prefix $\mathbf{C}_1, \dots, \mathbf{C}_{k-1}$, whose entries are real embeddings where the corresponding modality was observed and simulated embeddings where it was not. Since every prefix is complete by construction, a single family of modules suffices for all $2^K - 1$ possible availability patterns, and the number of trainable modules grows only linearly in K.

The trainable components are the following. First, $K - 1$ *anchor simulators* $S_{j \to 1} : \mathbb{R}^{d_j} \to \mathbb{R}^{d_1}$ for $j = 2, \dots, K$, used when the anchor modality $\mathbf{M}_1$ is itself missing. Second, $K - 2$ *prefix-fusion* modules $f_{1:\ell}$ for $\ell = 2, \dots, K - 1$, where $f_{1:\ell}$ maps the first ℓ completed embeddings to a joint representation $\mathbf{F}_{1:\ell} \in \mathbb{R}^{d_f}$ of fixed width d_f; any permutation-tolerant fusion mechanism may be used here, and we employ multi-head self-attention (MHSA). Third, $K - 1$ *prefix-to-target simulators* $S_{1:\ell \to (\ell+1)}$ for $\ell = 1, \dots, K - 1$, where for $\ell = 1$ the module $S_{1 \to 2} : \mathbb{R}^{d_1} \to \mathbb{R}^{d_2}$ is an ordinary pairwise simulator (no fusion is needed for a prefix of length one), and for $\ell \ge 2$ the module $S_{1:\ell \to (\ell+1)} : \mathbb{R}^{d_f} \to \mathbb{R}^{d_{\ell+1}}$ maps the fused prefix to the embedding space of the next modality. All simulators are lightweight MLPs. We write Θ for the union of these parameters. The total count is $3K - 4$ modules, independent of the number of availability patterns, so the module set scales linearly with K rather than with the number of modality subsets.

Given a sample with available set $\mathcal{A}$, the completed sequence is built as follows. The anchor is set to $\mathbf{C}_1 = \mathbf{M}'_1$ if $1 \in \mathcal{A}$, and otherwise to $\mathbf{C}_1 = S_{j^* \to 1}(\mathbf{M}'_{j^*})$, where $j^* \in \mathcal{A}$ is the available modality whose anchor simulator currently reconstructs modality 1 best. Concretely, for each j we maintain a running score ρ_j, an exponential moving average of the observed reconstruction error

$\|\mathbf{M}'_1 - S_{j\to 1}(\mathbf{M}'_j)\|^2$ over recent training batches in which both modality 1 and modality j were present, and select $j^* = \arg\min_{j\in\mathcal{A}} \rho_j$. Thereafter, for each position $k = 2, \ldots, K$ in turn, we form the simulated embedding

$$\hat{\mathbf{M}}'_k = \begin{cases} S_{1\to 2}(\mathbf{C}_1), & k = 2, \\ S_{1:(k-1)\to k}\big(f_{1:(k-1)}(\mathbf{C}_1, \ldots, \mathbf{C}_{k-1})\big), & k \geq 3, \end{cases} \tag{3}$$

and set $\mathbf{C}_k = \mathbf{M}'_k$ if $k \in \mathcal{A}$ and $\mathbf{C}_k = \hat{\mathbf{M}}'_k$ otherwise. Two points are worth making explicit. First, $\hat{\mathbf{M}}'_k$ is computed at every position, irrespective of whether modality k is available; availability decides only which of the two candidates is carried forward into the prefix used by positions $k+1, \ldots, K$. Second, when modality k is available the real embedding always takes precedence, so an observed modality is never replaced by its own estimate, and the estimate $\hat{\mathbf{M}}'_k$ is instead retained as a supervision signal, as described next. By the time position k is reached, $k-2$ prefix fusions have been applied, and the prefix feeding $f_{1:(k-1)}$ is complete regardless of $\mathcal{A}$.

Training the arms uses exactly this sweep and requires no labels. Whenever a position $k \geq 2$ is available, the corresponding $\hat{\mathbf{M}}'_k$ from Equation (3) admits the ground-truth target $\mathbf{M}'_k$, and contributes the squared error $\|\mathbf{M}'_k - \hat{\mathbf{M}}'_k\|^2$; positions that are genuinely missing contribute nothing, as no reference exists for them. Because the prefix $\mathbf{C}_1, \ldots, \mathbf{C}_{k-1}$ never contains $\mathbf{M}_k$ itself, this term is well defined without any masking or re-running of the procedure: one forward sweep simultaneously completes the sample and generates every available supervision term. Separately, the anchor simulators are supervised whenever $1 \in \mathcal{A}$, each available $j \in \mathcal{A} \setminus \{1\}$ contributing $\|\mathbf{M}'_1 - S_{j\to 1}(\mathbf{M}'_j)\|^2$; these same values update the running scores ρ_j that govern the choice of j^*. Collecting both parts and normalizing by the number of supervised terms so that samples with larger $\mathcal{A}$ do not dominate the gradient, the objective for a sample is

$$\mathcal{L}_{\text{sim}} = \frac{1}{|\mathcal{A}\setminus\{1\}|} \sum_{k\in\mathcal{A},\, k\geq 2} \big\|\mathbf{M}'_k - \hat{\mathbf{M}}'_k\big\|^2 \;+\; \frac{\mathbf{1}_{1\in\mathcal{A}}}{|\mathcal{A}\setminus\{1\}|} \sum_{j\in\mathcal{A}\setminus\{1\}} \big\|\mathbf{M}'_1 - S_{j\to 1}(\mathbf{M}'_j)\big\|^2, \tag{4}$$

with $\mathcal{L}_{\text{sim}} = 0$ when $\mathcal{A} = \{1\}$ or, more generally, whenever $|\mathcal{A}| = 1$, since a single observed modality yields no usable target. The parameters Θ are updated by stochastic gradient descent on $\mathcal{L}_{\text{sim}}$ averaged over the minibatch; frozen encoders receive no gradient. Note that a term at position k backpropagates through the whole completed prefix, so that simulated entries $\mathbf{C}_j$ with $j < k$ and the fusions that produced them are updated as well; this is what allows the limited set of arms to be trained jointly rather than position by position. Detaching $\mathbf{C}_j$ from the graph is a valid alternative that shortens the chain at the cost of this coupling.

Since $\mathcal{L}_{\text{sim}}$ involves no targets beyond the modality embeddings themselves, the module can be fitted on unlabeled or partially labeled data and, once converged, frozen and reused as a fixed preprocessing stage. At inference, a sample with available set $\mathcal{A}$ is encoded, swept once through Equation (3), and emitted as $\{\mathbf{C}_k\}_{k=1}^K$; the cost is at most $K-2$ fusion passes and $K-1$ MLP evaluations, independent of which modalities happen to be absent.

4 Theoretical Notes

This section collects the properties that follow from the two design choices of the previous sections: the sequential, prefix-based construction of the simulating arms (Section 3) and the block-alternating federated schedule (Section 2). Throughout, K is the number of modalities, d the width of the embedding space in which imputation is performed, N the number of clients, q the participation rate, R the total number of communication rounds and R_{sim} the round at which the schedule switches blocks. We write $|\Theta|$ and $|\Phi|$ for the parameter counts of the simulation block and of the task block.

4.1 Cost of the Sequential Sweep

Proposition 1 (Pattern-independent module set)**.** *The construction of Section 3 uses $K-1$ anchor simulators, $K-2$ prefix-fusion modules and $K-1$ prefix-to-target simulators, that is $3K-4$ trainable modules in total. This count is independent of the availability set $\mathcal{A}$ and, in particular, does not grow with the number $2^K - 1$ of non-empty availability patterns.*

The reason is that position k is never reconstructed from the raw availability pattern but only from the completed prefix $\mathbf{C}_1, \ldots, \mathbf{C}_{k-1}$, which is complete by construction. A pattern-indexed design would instead require one module per (source subset, target) pair, hence a count exponential in K; a pairwise design would require a count quadratic in K.

Proposition 2 (Inference cost is independent of how much is missing)**.** *For a sample with availability set $\mathcal{A}$, one forward sweep performs at most $K-1$ simulator evaluations and at most $K-2$ prefix fusions, regardless of $|\mathcal{A}|$ and of which positions are absent. The added cost on top of the frozen encoders is*

$$O\big((K-1)d^2\big) \;+\; O\Big(\textstyle\sum_{\ell=2}^{K-1} \ell^2 d\Big) \;=\; O\big(Kd^2 + K^3 d\big).$$

Each simulator is a small MLP acting on d-dimensional embeddings, costing $O(d^2)$ per call, and each prefix fusion is a self-attention block over at most K tokens, costing $O(\ell^2 d)$ at position ℓ. The availability pattern selects, at every position, which of the two candidates is carried forward; it does not change the number of positions visited. Consequently there is no combinatorial blow-up as modalities are removed, and the worst case (one observed modality) costs the same as the best case (none missing). Since the frozen encoders act on raw inputs whose length is typically far larger than both K and d, they dominate total inference compute, and the sweep remains a small additive term as K grows.

4.2 Communication of the Block-Alternating Schedule

Proposition 3 (Payload)**.** *Under the schedule of Equation* (1)*, the uplink volume over R rounds is*

$$\mathcal{B}_{\text{ours}} = \lceil qN \rceil \big(R_{\text{sim}}|\Theta| + (R - R_{\text{sim}})|\Phi|\big), \qquad \mathcal{B}_{\text{joint}} = \lceil qN \rceil R\big(|\Theta| + |\Phi|\big),$$

with the downlink volume reduced identically. The ratio $\mathcal{B}_{\text{ours}}/\mathcal{B}_{\text{joint}}$ is a convex combination of $|\Theta|/(|\Theta| + |\Phi|)$ and $|\Phi|/(|\Theta| + |\Phi|)$ and is therefore strictly below one for every $0 < R_{\text{sim}} < R$; when the two blocks are of comparable size the payload per participating client and round is halved, irrespective of R_{sim}.

Two further reductions are structural rather than scheduled. The frozen encoders E_k are held locally and never enter the exchange, so the payload is governed by the lightweight simulators, fusions and head rather than by the backbone; and the simulation block is not re-transmitted after R_{sim}, so its contribution to the total budget is bounded by R_{sim} and not by R.

Remark 1 (Label-free first stage)**.** *The objective of Equation* (4) *is defined purely on modality embeddings and requires no targets. Stage 1 can therefore be run on unlabeled or partially labeled local data that the joint objective cannot use, and the resulting arms can be frozen and reused as a fixed preprocessing map for a different downstream task without retraining.*

Remark 2 (A single global imputation map during stage 2)**.** *At the transition the arms are frozen globally at $\Theta^{(R_{\text{sim}})}$. For $r > R_{\text{sim}}$, every client therefore completes its samples with the same deterministic map, and the composition of frozen encoders with frozen arms is a fixed feature extractor shared by the whole federation. Stage 2 is then an ordinary federated optimization problem over Φ on a stationary feature space, so the aggregation rule of Equation* (2) *may be replaced by any other federated aggregator and its assumptions apply unchanged. Under joint optimization the feature map moves with every aggregation, and the disagreement between clients has two coupled sources instead of one.*

4.3 Extendibility

The schedule is defined by a partition of the trainable parameters into blocks driven by different objectives, together with an ordering of those blocks over the round budget; nothing in Equation (1) is specific to two blocks. Writing $\Xi_1, \ldots, \Xi_S$ for a partition into S blocks with objectives $\mathcal{L}_1, \ldots, \mathcal{L}_S$ and switch points $0 < R_1 < \cdots < R_{S-1} < R$, the active block in round r is Ξ_s for $R_{s-1} < r \le R_s$, and the uplink volume becomes $\lceil qN \rceil \sum_{s=1}^{S} (R_s - R_{s-1})|\Xi_s|$, again a convex combination of the individual block sizes and therefore never larger than the joint protocol. The instance we use when pretrained encoders are unavailable or inadmissible takes $S = 3$: the encoders E_k are themselves learned federatedly, on local unimodal data only, during the first interval; they are frozen; the arms are trained on the label-free objective during the second interval and frozen in turn; and only the

fusion and head are exchanged in the last interval. Privacy properties are unaffected, since raw data and raw embeddings leave no client in any stage, and neither the simulating-arms construction nor the objective of Equation (4) is modified. Together with Proposition 1, this gives two orthogonal directions of growth — more modalities and more stages — each of which costs only a linear increase in the number of exchanged blocks.

5 Experimental Results

Datasets and missing-modality regimes. We evaluate on **MM-IMDb** ($K = 2$: image and text), a multi-label movie genre classification benchmark over 27 genres in which a title may carry several genres at once (Section B), and on **CMU-MOSI** ($K = 3$: text, audio and video), whose segments carry a continuous sentiment score in $[-3, +3]$ and which we treat as a regression task. Scalability to a larger modality count is assessed on **Sleep-EDF** ($K = 5$ physiological channels, five sleep stages). For MM-IMDb we use two families of patterns: α-image-β-text, where $\alpha\%$ of samples keep the image and $\beta\%$ keep the text and at least one of α, β equals 100; and complex-γ-α-β, where $\gamma\%$ of samples have both modalities, $\alpha\%$ only the image and $\beta\%$ only the text. For CMU-MOSI we use X-text-Y-audio-Z-video, and the complex pattern complex-TAV-T-A-V-TA-TV-AV whose entries give the percentage of samples retaining all three modalities, text only, audio only, video only, text with audio, text with video, and audio with video respectively.

Protocol. MM-IMDb and CMU-MOSI are partitioned non-IID over 10 clients, of which 50% participate in each round; Sleep-EDF is partitioned over 30 clients with a participation rate of 10%. All methods — ShaSpec Wang et al. (2023), M^3Care Zhang et al. (2022), PmcmFL Bao et al. (2023), FLEX-MoE Zhang et al. (2025) and PEPSY Nguyen et al. (2025), of which the first two are centralized methods and the last three federated ones — are run on the *same* frozen encoders (CLIP for images and video frames, BERT for text, WavLM for audio), the same client partitions and the same missing patterns, so that the reported margins reflect the multi-modal adaptation and the aggregation mechanism rather than the backbone. Every method is given the same budget of $R = 10$ communication rounds with $P = 10$ local epochs per selected client; FArms splits this budget at $R_{\text{sim}} = 2$ and reaches its reported accuracy by round four, after which further rounds leave it unchanged (Figure 4), while the baselines are always allowed the full budget. Optimization uses Adam with learning rate 5×10^{-5} in both stages. Results on MM-IMDb and CMU-MOSI are averaged over ten independent runs and on Sleep-EDF over three; the last rows of Tables 1a and 1b name the winner under a paired t-test and under a Wilcoxon signed-rank test (WCOX) at the 5% level, with *N/A* where no difference is significant. Remaining hyper-parameters are listed in Section C.

Table 1: F-micro and F-macro obtained by the algorithms on MM-IMDb and CMU-MOSI.

(a) F-micro

	MM-IMDb						CMU-MOSI							
Method	100-image-100-text	100-image-20-text	20-image-100-text	complete-10-40-45	complete-20-40-40	complete-30-35-35	100-text-100-audio-100-video	100-text-100-audio-20-video	100-text-20-audio-100-video	100-text-20-audio-20-video	20-text-100-audio-20-video	20-text-20-audio-100-video	20-text-20-audio-20-video	complete-20-20-20-10-10-10
ShaSpec	0.49	0.39	0.45	0.39	0.38	0.43	0.35	0.33	0.32	0.30	0.28	0.25	0.27	0.29
M^3Care	0.46	0.39	0.43	0.43	0.37	0.45	0.36	0.34	0.33	0.31	0.29	0.26	0.28	0.30
PmcmFL	0.66	0.51	0.48	0.50	0.51	0.54	0.38	0.36	0.35	0.33	0.31	0.28	0.30	0.32
FLEX-MoE	0.11	0.09	0.10	0.08	0.09	0.10	0.61	0.54	0.51	0.51	0.49	0.43	0.49	0.51
PEPSY	0.13	0.11	0.12	0.10	0.11	0.12	0.74	0.72	0.71	0.62	0.64	0.55	0.55	0.55
FArms	**0.94**	**0.93**	**0.93**	**0.93**	**0.93**	**0.93**	**0.82**	**0.81**	**0.82**	**0.81**	**0.67**	**0.57**	**0.67**	**0.71**
t-test	FArms wins on all columns													
WCOX	FArms wins on all columns													

(b) F-macro

	MM-IMDb						CMU-MOSI							
Method	100-image-100-text	100-image-20-text	20-image-100-text	complete-10-40-45	complete-20-40-40	complete-30-35-35	100-text-100-audio-100-video	100-text-100-audio-20-video	100-text-20-audio-100-video	100-text-20-audio-20-video	20-text-100-audio-20-video	20-text-20-audio-100-video	20-text-20-audio-20-video	complete-20-20-20-10-10-10
ShaSpec	0.50	0.38	0.44	0.39	0.38	0.43	0.33	0.31	0.30	0.28	0.26	0.23	0.25	0.27
M^3Care	0.45	0.38	0.43	0.42	0.36	0.44	0.34	0.32	0.31	0.29	0.27	0.24	0.26	0.28
PmcmFL	**0.66**	0.49	0.45	0.48	0.49	0.54	0.36	0.34	0.33	0.31	0.29	0.26	0.28	0.30
FLEX-MoE	0.08	0.06	0.07	0.05	0.06	0.07	0.57	0.51	0.48	0.48	0.46	0.40	0.46	0.48
PEPSY	0.09	0.07	0.08	0.06	0.07	0.08	0.70	0.68	0.67	0.59	0.61	0.52	0.52	0.52
FArms	0.64	**0.57**	**0.60**	**0.57**	**0.57**	**0.58**	**0.82**	**0.81**	**0.82**	**0.81**	**0.67**	**0.57**	**0.67**	**0.71**
t-test	PmcmFL/FArms						FArms wins on all columns							
WCOX	PmcmFL/FArms						FArms wins on all columns							

Overall comparison. Table 1a shows FArms attains the best F-micro in every configuration of both benchmarks, staying between 0.93 and 0.94 on MM-IMDb even as four samples in five lose a modality, while the baselines degrade steadily. Table 1b follows the same trend, except on the fully available MM-IMDb column, where PmcmFL is slightly ahead. The micro/macro gap on MM-IMDb reflects the benchmark's long-tailed genre distribution rather than any method (Section B), and Figure 3 shows the arms' simulated embeddings falling inside the manifold of ground-truth embeddings rather than forming a separate cluster.

Why FLEX-MoE and PEPSY behave differently on the two benchmarks. Please see Section E

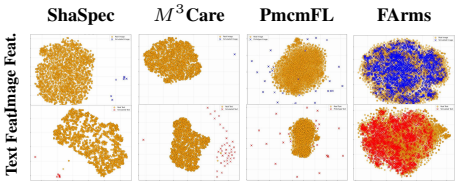


Figure 3: Visualisation of real vs simulated feature embeddings for missing modalities across different methods on MM-IMDb.

Table 2: Two-stage versus joint optimization on CMU-MOSI.

	100-20-20	20-100-20	20-20-100	complex
F-micro (Two-stage)	0.811	0.584	**0.670**	**0.741**
F-micro (Joint)	**0.820**	**0.607**	0.645	0.732
Payload/client/round: Two-stage(**MB**)	**29.48**	**29.48**	**29.48**	**29.48**
Payload/client/round: Joint(**MB**)	58.95	58.95	58.95	58.95

Two-stage versus joint optimization. Table 2 isolates the schedule while holding the architecture, partition and round budget fixed. Accuracy is close between the two protocols, with each ahead in two of the four configurations and no gap exceeding a few points, while the payload differs by a factor of two throughout, as Proposition 3 predicts for blocks of comparable size. The schedule thus buys a halved per-round payload, a label-free first stage, and a stationary feature space in stage two, at no systematic cost in accuracy.

Regression Results CMU-MOSI is natively a regression task, and our task head already emits a continuous value, with Table 3 reporting the corresponding MSE and MAE. We additionally report F-micro and F-macro in Tables 1a and 1b, derived from these same continuous outputs, since the baselines are classification-based and not designed to produce regression outputs, making the F-scores the only common ground for comparison. Please also see the sign rule of Section A.

Table 3: Regression metrics of FArms on the CMU-MOSI

Configuration	**MSE** ↓	**MAE** ↓
100-text-100-audio-100-video	1.26	0.91
100-text-100-audio-20-video	1.14	0.84
100-text-20-audio-100-video	1.13	0.84
100-text-20-audio-20-video	1.10	0.84
20-text-100-audio-100-video	2.01	1.22
20-text-100-audio-20-video	2.39	1.35
20-text-20-audio-100-video	2.12	1.24
complex-20-20-20-10-10-10-10	1.72	1.07

Table 4: F-micro of FArms on CMU-MOSI as the number of clients grows, at a fixed participation rate of 50% and a fixed budget of $R = 10$ rounds.

Clients	**C1**	**C2**
10	0.81	0.60
15	0.78	0.60
20	0.78	0.47
25	0.78	0.47
30	0.78	0.51

C1: 100-text-20-audio-20-video
C2: 20-text-100-audio-20-video.

Scaling the Client Pool Table 4 grows the CMU-MOSI client pool from ten to thirty at a fixed participation rate and non-IID partition. F-micro is flat as the pool triples in the text-dominated regime, but in the audio-dominated regime it holds only to fifteen clients before dipping and partially recovering at thirty. This asymmetry reflects supervision density, not the schedule: when text is absent for most samples, the pairs available to train the anchor simulators are already scarce, and spreading them over more non-IID shards leaves individual clients under-supervised. A longer first stage, rather than a larger overall budget, is the natural remedy.

Communication and computation compared to the baselines. Table 5 compares the per-round exchange of FArms with the federated baselines, in the notation of Section 4. FArms transmits one of two lightweight blocks and never the encoders, with a simple weighted-average server step. PEPSY additionally exchanges per-client embedding controls and solves a clustering problem superlinear in the number of clients and controls; PmcmFL exchanges the full model plus per-class prototypes, so its overhead grows with the label space; FLEX-MoE exchanges a shared extractor plus assigned experts and solves a per-round assignment program over the client–expert matrix. Client-side, FArms's cost is the linear sweep of Proposition 2, free of the quadratic terms prototype banks or expert routing introduce.

Rounds. Figure 4 tracks F-micro as the budget grows from five to twenty rounds in the severe CMU-MOSI regimes. The curves are flat within a point or two beyond the fourth round, so the budget of ten rounds granted to every method in Tables 1a and 1b is not a limitation for FArms; the comparison is, if anything, conservative towards it, since the baselines use the full budget while FArms does not need it. Combined with Proposition 3, this gives the two multiplicative savings that the framework claims: fewer rounds, and a smaller payload within each of them.

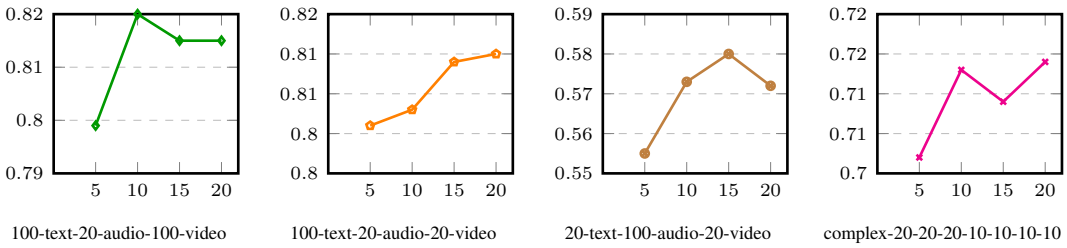


Figure 4: F-Micro changes with number of rounds across different data configurations in federated setting.

What the arms contribute. Table 6 replaces the arms with three fillers (zero, mean, available-only masked fusion) sharing the same encoders, fusion and head, and further ablates the prefix fusions and the case-aware loss of Equation (4) against joint optimization at two weightings. Results are reported across all three regimes for the reader to compare directly.

More modalities. Table 7 moves from $K = 2$ and $K = 3$ to $K = 5$ on Sleep-EDF, where the client pool is larger and only one client in ten participates per round. FArms is the significant winner in all four IID regimes and in one of the four non-IID ones, and is statistically indistinguishable from PEPSY in the remaining three; no configuration is won by another method. What the table supports is therefore that the construction transfers to a five-modality, signal-only setting without modification to the objective or to the schedule, and that its advantage narrows rather than disappears as the data move away from the vision–language regime. Note also that accuracy is essentially flat in p_m for FArms, which is what Proposition 2 leads one to expect: increasing the number of absent modalities changes which branch is taken at each position of the sweep, not how much work or how many parameters are involved.

Table 5: Communication and computational characteristics per communication round. N is the number of clients, q the participation rate, $|\Theta|$ and $|\Phi|$ the two FArms blocks, $|\psi|$ and $|\theta_e|$ the shared extractor and one expert of FLEX-MoE with $k_c \le E$ experts per client, $|C|$ the number of classes, d_p the width of an embedding control or prototype, p the number of controls a PEPSY client transmits, and M the number of inner iterations of its server-side clustering.

Method	**Payload per client and round**	**Exchanged / trainable components**	**Server-side step and key cost factor**
FArms (ours)	$O(\lvert\Theta\rvert)$ for $r \le R_{\text{sim}}$, $O(\lvert\Phi\rvert)$ afterwards; total $\lceil qN \rceil \big(R_{\text{sim}}\lvert\Theta\rvert + (R - R_{\text{sim}})\lvert\Phi\rvert\big)$	Lightweight MLP simulators and small MHSA fusions, plus the task head; frozen encoders are never transmitted	Weighted averaging, $O(N\lvert\Omega\rvert)$; only one block is active per round
PEPSY Nguyen et al. (2025)	$O(d + p\, d_p)$	Full model plus client-side embedding controls and reconfiguration modules	Non-parametric clustering of the transmitted controls, $O(Nd + MN^2p^2d_p^2)$; alignment of per-client missing-pattern profiles
FLEX-MoE Zhang et al. (2025)	$O(\lvert\psi\rvert + k_c\lvert\theta_e\rvert)$	Shared feature extractor and the assigned expert subset; gating kept locally	Linear program over the $N \times E$ client–expert matrix in every round, plus usage-weighted expert averaging
PmcmFL Bao et al. (2023)	$O(d + 3\lvert C\rvert\, d_p)$	Full model plus image, text and fusion prototype libraries	FedAvg plus prototype-library construction; overhead grows with the number of classes

Table 6: Component ablation on CMU-MOSI under three regimes. *No-sim controls* keep the frozen encoders, the final fusion and the head, and differ only in how absent positions are filled. *Joint* rows use the single objective $\alpha\mathcal{L}_{\text{sim}} + \beta\mathcal{L}_{\text{task}}$ with all parameters updated every round.

	100-text-20-audio-20-video				**20-text-100-audio-20-video**				**complex-20-20-20-10-10-10-10**			
Variant	MSE ↓	MAE ↓	F-mi ↑	F-ma ↑	MSE ↓	MAE ↓	F-mi ↑	F-ma ↑	MSE ↓	MAE ↓	F-mi ↑	F-ma ↑
FArms (full, two-stage)	1.1000	0.8400	**0.8100**	**0.8100**	2.3900	1.3500	**0.5700**	**0.5700**	1.7200	1.0700	**0.7100**	**0.7100**
No-sim control: zero imputation	1.1209	0.8597	0.7955	0.7945	2.3568	1.3001	0.5227	0.4921	1.8455	1.0791	0.6568	0.6496
No-sim control: mean imputation	1.2916	0.8756	0.7591	0.7569	2.6274	1.3971	0.4409	0.3060	**1.6533**	**1.0220**	0.7091	0.7086
No-sim control: available-only (masked fusion)	1.2016	**0.8208**	0.7727	0.7715	**2.3017**	**1.2751**	0.5227	0.4844	1.9711	1.0993	0.6864	0.6863
Arms without prefix fusion (pairwise only)	**1.0518**	0.8213	0.7909	0.7869	2.3363	1.3206	0.5886	0.5213	1.8319	1.0788	0.6409	0.6293
Case-aware simulation loss off	1.0894	0.8487	0.7886	0.7885	2.4473	1.3597	0.5795	0.4444	2.4134	1.1824	0.6977	0.6977
Joint, simulation-heavy (α=2.0, β=1.0)	1.1990	0.8349	0.8045	0.8003	2.4496	1.3597	0.4545	0.3359	1.9801	1.1553	0.5977	0.5970
Joint, task-heavy (α=0.5, β=1.0)	1.1599	0.8342	0.8023	0.7989	2.4465	1.3602	0.5591	0.3586	1.8990	1.0911	0.6545	0.6519

Table 7: Accuracy on Sleep-EDF ($K = 5$), with 30 clients and a participation rate of 10%. Eighty percent of the samples are affected by missingness and p_m is the fraction of modalities absent in those samples. Averages over three runs; the last row names the winner under a paired t-test.

	IID				non-IID			
Method	p_m=0.2	p_m=0.4	p_m=0.6	p_m=0.8	p_m=0.2	p_m=0.4	p_m=0.6	p_m=0.8
PmcmFL	0.32	0.34	0.32	0.32	0.33	0.33	0.32	0.32
FLEX-MoE	0.36	0.34	0.34	0.34	0.33	0.34	0.32	0.32
PEPSY	0.39	0.27	0.27	0.23	0.43	0.43	**0.42**	**0.42**
FArms	**0.44**	**0.44**	**0.43**	**0.43**	**0.44**	**0.45**	0.41	0.41
Winner	FArms	FArms	FArms	FArms	N/A	FArms	N/A	N/A

6 Related Work

Missing-modality learning has been studied extensively in centralized settings, through adaptive fusion and missing-aware prompting Ma et al. (2022); Lee et al. (2023); generative reconstruction via imagination, translation, shared-private modeling, diffusion, or prototype decomposition Zhao et al. (2021); Liu et al. (2024); Wu et al. (2021); Dai et al. (2025); Jin et al. (2023); and improved representation learning through text-enhanced or hybrid cross-modal features, self-supervised or interactive representations, dynamic attention, mixtures of experts, and constraint-based robustness Chen et al. (2025); Fu et al. (2024); Yu et al. (2021); Huang et al. (2023); Lin & Hu (2023a); Li et al. (2025); Lin & Hu (2023b); Zhou et al. (2020); Mirza et al. (2023). Cross-modality consistency and shared-specific feature learning further reduce representation gaps under heavy missingness You et al. (2016); Liu et al. (2022); Wang et al. (2023), and Bayesian meta-learning enables flexible handling of severe missingness at both train and test time Ma et al. (2021).

Federated multi-modal learning (MFL) extends these ideas to distributed, privacy-preserving settings, via foundation-model integration, meta-learning, prototype-based approaches, and contrastive alignment Che et al. (2024); Singh et al. (2023); Zeng et al. (2024); Le et al. (2025); Bao et al. (2023), and, more recently, reconfigurable representations for heterogeneous missing patterns Nguyen et al. (2025), retrieval-based cross-modal augmentation Poudel et al. (2024), federated feature-imputation networks Poudel et al. (2025), mixture-of-experts for arbitrary modality combinations Yun et al. (2024), and quality-aware aggregation for time-series healthcare sensing Orzikulova et al. (2024).

These approaches largely rely on indirect alignment (contrastive losses, expert routing Yun et al. (2024), reconfigurable embeddings Nguyen et al. (2025)), prototype matching Le et al. (2024); Zhu et al. (2025), or quality-aware aggregation and distillation Orzikulova et al. (2024) rather than explicit per-sample reconstruction supervision across all availability regimes, leaving them more sensitive to extreme missing rates or task shifts.

7 Conclusion

We separated the missing-modality handler from the downstream task in federated multi-modal learning, turning what is usually a single jointly trained model into a schedule of blocks with distinct objectives, only one of which is communicated in a given round. The handler itself, a sequence of lightweight simulating arms operating over a fixed modality ordering, reconstructs any missing modality from an already-completed prefix, requiring only $3K - 4$ trainable modules regardless of the number of availability patterns, and is trained without labels. Theoretically, this construction gives a pattern-independent inference cost, a communication payload that is a convex combination of the two blocks' sizes rather than their sum, and a natural extension to additional stages such as federated encoder training. Empirically, across multi-label classification, regression, and a five-modality signal setting, FArms matches or exceeds state-of-the-art federated and centralized baselines in the large majority of configurations, does so within a fraction of the allotted communication rounds, and remains stable as the severity and heterogeneity of missing modalities increase. Ablations confirm that both the prefix-fusion mechanism and the case-aware training objective, rather than the schedule alone, are responsible for this stability. Taken together, the framework offers a task-agnostic, extendible alternative to jointly trained multi-modal federated models, at reduced communication cost and no systematic loss in accuracy.

8 AI USE STATEMENT

We used AI to aid or polish writing, and for retrieval and discovery.

REFERENCES

Guangyin Bao, Qi Zhang, Duoqian Miao, Zixuan Gong, Liang Hu, Ke Liu, Yang Liu, and Chongyang Shi. Multimodal federated learning with missing modality via prototype mask and contrast. *arXiv preprint arXiv:2312.13508*, 2023.

Liwei Che, Jiaqi Wang, Xinyue Liu, and Fenglong Ma. Leveraging foundation models for multimodal federated learning with incomplete modality. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 401–417. Springer, 2024.

Jingwen Chen, Shuxiang Song, Yumei Tan, and Haiying Xia. Temsa: Text enhanced modal representation learning for multimodal sentiment analysis. *Computer Vision and Image Understanding*, 258:104391, 2025.

Ruiting Dai, Chenxi Li, Yandong Yan, Lisi Mo, Ke Qin, and Tao He. Unbiased missing-modality multimodal learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 24507–24517, 2025.

Yanping Fu, Zhiyuan Zhang, Ruidi Yang, and Cuiyou Yao. Hybrid cross-modal interaction learning for multimodal sentiment analysis. *Neurocomputing*, 571:127201, 2024.

Jian Huang, Yanli Ji, Yang Yang, and Heng Tao Shen. Cross-modality representation interactive learning for multimodal sentiment analysis. In *Proceedings of the 31st ACM international conference on multimedia*, pp. 426–434, 2023.

Tao Jin, Xize Cheng, Linjun Li, Wang Lin, Ye Wang, and Zhou Zhao. Rethinking missing modality learning from a decoding perspective. In *Proceedings of the 31st ACM international conference on multimedia*, pp. 4431–4439, 2023.

Huy Q Le, Chu Myaet Thwal, Yu Qiao, Ye Lin Tun, Minh NH Nguyen, Eui-Nam Huh, and Choong Seon Hong. Cross-modal prototype based multimodal federated learning under severely missing modality. *Information Fusion*, pp. 103219, 2025.

Huy Q Le et al. Cross-modal prototype based multimodal federated learning under severely missing modality. *arXiv preprint arXiv:2401.13898*, 2024.

Yi-Lun Lee, Yi-Hsuan Tsai, Wei-Chen Chiu, and Chen-Yu Lee. Multimodal prompting with missing modalities for visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14943–14952, 2023.

Sijie Li, Chen Chen, and Jungong Han. Simmlm: A simple framework for multi-modal learning with missing modality. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 24068–24077, 2025.

Ronghao Lin and Haifeng Hu. Dynamically shifting multimodal representations via hybrid-modal attention for multimodal sentiment analysis. *IEEE Transactions on Multimedia*, 26:2740–2755, 2023a.

Ronghao Lin and Haifeng Hu. Missmodal: Increasing robustness to missing modality in multimodal sentiment analysis. *Transactions of the Association for Computational Linguistics*, 11:1686–1702, 2023b.

Huan Liu, Ke Li, Jianping Fan, Caixia Yan, Tao Qin, and Qinghua Zheng. Social image–text sentiment classification with cross-modal consistency and knowledge distillation. *IEEE Transactions on Affective Computing*, 14(4):3332–3344, 2022.

Zhizhong Liu, Bin Zhou, Dianhui Chu, Yuhang Sun, and Lingqiang Meng. Modality translation-based multimodal sentiment analysis under uncertain missing modalities. *Information Fusion*, 101: 101973, 2024.

Mengmeng Ma, Jian Ren, Long Zhao, Sergey Tulyakov, Cathy Wu, and Xi Peng. Smil: Multimodal learning with severely missing modality. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 2302–2310, 2021.

Mengmeng Ma, Jian Ren, Long Zhao, Davide Testuggine, and Xi Peng. Are multimodal transformers robust to missing modality? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18177–18186, 2022.

M Jehanzeb Mirza, Leonid Karlinsky, Wei Lin, Horst Possegger, Rogerio Feris, and Horst Bischof. Tap: Targeted prompting for task adaptive generation of textual training instances for visual classification. *arXiv preprint arXiv:2309.06809*, 2023.

Manh Duong Nguyen, Trong Nghia Hoang, Thanh Trung Huynh, Quoc Viet Hung Nguyen, and Phi Le Nguyen. Learning reconfigurable representations for multimodal federated learning with missing data. *arXiv preprint arXiv:2510.22880*, 2025.

Adiba Orzikulova, Jaehyun Kwak, Jaemin Shin, and Sung-Ju Lee. Federated learning for time-series healthcare sensing with incomplete modalities. *arXiv preprint arXiv:2405.11828*, 2024.

Pranav Poudel, Prashant Shrestha, Sanskar Amgain, Yash Raj Shrestha, Prashnna Gyawali, and Binod Bhattarai. Car-mfl: Cross-modal augmentation by retrieval for multimodal federated learning with missing modalities. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 102–112. Springer, 2024.

Pranav Poudel, Aavash Chhetri, Prashnna Gyawali, Georgios Leontidis, and Binod Bhattarai. Multimodal federated learning with missing modalities through feature imputation network. In *Annual Conference on Medical Image Understanding and Analysis*, pp. 289–299. Springer, 2025.

Apoorva Singh, Siddarth Chandrasekar, Sriparna Saha, and Tanmay Sen. Federated meta-learning for emotion and sentiment aware multi-modal complaint identification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 16091–16103, 2023.

Hu Wang, Yuanhong Chen, Congbo Ma, Jodie Avery, Louise Hull, and Gustavo Carneiro. Multimodal learning with missing modality via shared-specific feature modelling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15878–15887, 2023.

Yang Wu, Zijie Lin, Yanyan Zhao, Bing Qin, and Li-Nan Zhu. A text-centered shared-private framework via cross-modal prediction for multimodal sentiment analysis. In *Findings of the association for computational linguistics: ACL-IJCNLP 2021*, pp. 4730–4738, 2021.

Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang. Cross-modality consistent regression for joint visual-textual sentiment analysis of social multimedia. In *Proceedings of the Ninth ACM international conference on Web search and data mining*, pp. 13–22, 2016.

Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 10790–10797, 2021.

Sukwon Yun, Inyoung Choi, Jie Peng, Yangfan Wu, Jingxuan Bao, Qiyiwen Zhang, Jiayi Xin, Qi Long, and Tianlong Chen. Flex-moe: Modeling arbitrary modality combination via the flexible mixture-of-experts. In *Advances in Neural Information Processing Systems*, volume 37, pp. 98782–98805, 2024.

Huimin Zeng, Zhenrui Yue, and Dong Wang. Open-vocabulary federated learning with multimodal prototyping. *arXiv preprint arXiv:2404.01232*, 2024.

Boyang Zhang, Xiaobing Chen, Songyang Zhang, Shuai Zhang, Xiangwei Zhou, Jian Zhang, and Mingxuan Sun. Flex-moe: Federated mixture-of-experts with load-balanced expert assignment for edge computing. *arXiv preprint arXiv:2512.23070*, 2025.

Chaohe Zhang, Xu Chu, Liantao Ma, Yinghao Zhu, Yasha Wang, Jiangtao Wang, and Junfeng Zhao. M3care: Learning with missing modalities in multi-modal healthcare data. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 2418–2428, 2022.

Jinming Zhao, Ruichen Li, and Qin Jin. Missing modality imagination network for emotion recognition with uncertain missing modalities. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 2608–2618, 2021.

Tao Zhou, Jiuxin Cao, Xuelin Zhu, Bo Liu, and Shancang Li. Visual-textual sentiment analysis enhanced by hierarchical cross-modality interaction. *IEEE Systems Journal*, 15(3):4303–4314, 2020.

Sa Zhu, Dayan Wu, Chenming Wu, Pengwen Dai, and Bo Li. Endogenous recovery via within-modality prototypes for incomplete multimodal hashing. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI)*, 2025.

APPENDIX

A COMPUTATION OF F-MICRO AND F-MACRO

MM-IMDb: multi-label protocol. MM-IMDb assigns each title a subset of $C = 27$ genres, so the task is multi-label rather than single-label and the softmax cross-entropy used for single-label classification does not apply. Let $y_{i,c} \in \{0, 1\}$ be the ground-truth indicator of genre c for sample i, let $\hat{y}_{i,c}$ be the corresponding output logit and let

$$p_{i,c} \;=\; \sigma(\hat{y}_{i,c}) \;=\; \frac{1}{1 + e^{-\hat{y}_{i,c}}}$$

be its sigmoid probability. The task loss is the binary cross-entropy applied independently per genre,

$$\mathcal{L}_{\text{task}} = -\frac{1}{NC} \sum_{i=1}^{N} \sum_{c=1}^{C} \Big[\, y_{i,c} \log p_{i,c} + (1 - y_{i,c}) \log \big(1 - p_{i,c}\big) \Big]. \tag{5}$$

Predictions are obtained by thresholding each genre independently, $\tilde{y}_{i,c} = \mathbf{1}[\, p_{i,c} > 0.5 \,]$, so a sample may be assigned any number of genres, including none. The per-genre confusion counts are

$$TP_c = \sum_{i=1}^{N} \mathbf{1}[y_{i,c}{=}1,\ \tilde{y}_{i,c}{=}1], \quad FP_c = \sum_{i=1}^{N} \mathbf{1}[y_{i,c}{=}0,\ \tilde{y}_{i,c}{=}1], \quad FN_c = \sum_{i=1}^{N} \mathbf{1}[y_{i,c}{=}1,\ \tilde{y}_{i,c}{=}0].$$

The micro-averaged score pools these counts over all genres before forming a single F1,

$$F1_{\text{micro}} = \frac{2 \sum_{c=1}^{C} TP_c}{2 \sum_{c=1}^{C} TP_c + \sum_{c=1}^{C} FP_c + \sum_{c=1}^{C} FN_c}, \tag{6}$$

whereas the macro-averaged score forms one F1 per genre and averages them with equal weight,

$$F1_c = \frac{2\,TP_c}{2\,TP_c + FP_c + FN_c}, \qquad F1_{\text{macro}} = \frac{1}{C} \sum_{c=1}^{C} F1_c. \tag{7}$$

Every label position of every sample is counted, and genres that never appear in a prediction contribute $F1_c = 0$ to Equation (7) while contributing only their FN_c to Equation (6). This is the entire source of the gap between the two metrics reported in Tables 1a and 1b: Equation (6) is dominated by the genres that carry most of the label mass, while Equation (7) is dominated by the tail. Section B quantifies the tail.

CMU-MOSI: sign-based binary protocol. On CMU-MOSI the head emits a single continuous value in $[-3, +3]$ and is trained with the squared error, as reported in Table 3. The F-scores in Tables 1a and 1b are computed on the same continuous outputs by the standard binary convention for this benchmark: a prediction is mapped to the positive class when its value exceeds zero and to the negative class otherwise, and the ground truth is mapped the same way, so that a prediction is counted correct exactly when it has the same sign as the reference score. Substituting $C = 2$ in Equations (6) and (7) then gives the reported values. Because each sample receives exactly one of two

labels, $\sum_c FP_c = \sum_c FN_c$ and Equation (6) reduces to plain accuracy; Equation (7) coincides with it whenever the two classes are predicted with equal quality, which is why the CMU-MOSI columns of Tables 1a and 1b agree for FArms while the baseline rows, whose errors are concentrated on one of the two classes, do not. As the baselines in Tables 1a and 1b are themselves classification-based, they are evaluated under this same sign-based binary protocol, so the reported F-scores are directly comparable across all methods rather than being converted from a regression output only for FArms.

B THE MM-IMDB DATASET

MM-IMDb pairs a movie poster with a plot summary, giving $K = 2$ modalities, and annotates each of its $25{,}959$ titles with the genres it belongs to, drawn from a fixed vocabulary of 27. Annotation is multi-label in the strict sense: a title typically carries more than one genre, the genres are not mutually exclusive, and there is no canonical primary genre, which is why the protocol of Section A treats the 27 decisions independently rather than as a choice among alternatives. Images are resized to 224×224 and normalized with the usual ImageNet statistics, and both modalities are embedded with frozen CLIP encoders producing 512-dimensional vectors.

The genre distribution is severely long-tailed, and this shapes every metric computed on the dataset. The most frequent genre, *Drama*, is present in $13{,}967$ of the $25{,}959$ titles, that is in 53.8% of them, while the rarest genre occurs once; the imbalance ratio between the two extremes is therefore roughly $13{,}967\times$, and the mean imbalance ratio across genres is 922.7, far above the value of 1.5 conventionally used to flag a multi-label dataset as imbalanced. Four genres together account for 71 samples across the training, validation and test splits combined, which leaves them effectively inestimable for any classifier: a handful of positive examples cannot support a reliable per-genre F1, yet Equation (7) weights each of them exactly as heavily as *Drama*.

The consequence is that a large gap between micro- and macro-averaged F1 is a property of the label distribution and not a symptom of a particular model. A frequency-only predictor that ignores the input entirely and always emits the k most frequent genres reaches a micro-averaged F1 of at most 0.39 while its macro-averaged F1 stays below 0.08 for every k. The separation observed in Tables 1a and 1b is thus structural, and the two metrics should be read as answering different questions: Equation (6) asks how many of the genre assignments in the corpus are recovered, and Equation (7) asks how uniformly that recovery is spread over a vocabulary in which most entries are nearly empty.

Missing modalities are simulated on top of this dataset with the two families described in Section 5. Because $K = 2$, an incomplete sample has exactly one modality, so the sweep of Section 3 reduces to a single simulator call and no prefix fusion; MM-IMDb therefore exercises the construction in its smallest non-trivial instance, and CMU-MOSI and Sleep-EDF exercise the fusion path.

C EXPERIMENTAL SETUP

Federated protocol. All federated quantities are fixed once and used identically by every method. MM-IMDb and CMU-MOSI are partitioned over $N = 10$ clients and Sleep-EDF over $N = 30$; the partition is non-IID and is produced with the Flower API using a Dirichlet allocation over labels. The participation rate is $q = 0.5$ on MM-IMDb and CMU-MOSI and $q = 0.1$ on Sleep-EDF, so five clients and three clients respectively are selected per round. Each selected client runs $P = 10$ local epochs before returning its update, and the server aggregates by dataset size according to Equation (2). The total budget is $R = 10$ communication rounds for every method, and FArms splits it at $R_{\text{sim}} = 2$. Both stages use separate Adam optimizers with learning rate 5×10^{-5}. Neither raw data nor raw embeddings leave a client at any point, and the frozen encoders are pre-loaded identically on every client and never transmitted.

Encoders. MM-IMDb uses frozen CLIP ViT-B/32 encoders for both the poster and the plot summary, producing 512-dimensional embeddings. CMU-MOSI uses `bert-base-uncased` for text, `microsoft/wavlm-large` for audio and the `openai/clip-vit-base-patch32` vision encoder for video frames. Sleep-EDF uses a convolutional encoder per channel, trained federatedly in the first stage of the three-stage instance of Section 4.3 and frozen thereafter, as

no pretrained encoder spans these signal modalities. Every baseline consumes exactly the same embeddings.

Trainable modules. Simulators are two-layer MLPs with hidden width 256 and ReLU activations; prefix fusions and the final fusion are multi-head self-attention blocks over a 256-dimensional shared space; the task head is a three-layer MLP with ReLU and dropout 0.1, ending in 27 logits on MM-IMDb, a single value on CMU-MOSI and five logits on Sleep-EDF. The module count follows Proposition 1.

Reporting. Results on MM-IMDb and CMU-MOSI are averaged over ten independent runs and on Sleep-EDF over three. Significance is assessed with a paired t-test and a Wilcoxon signed-rank test at the 5% level; *N/A* in a winner row means no significant difference. The ablations of Tables 2, 4 and 6 are single runs per cell, which is why their values deviate slightly from the averages in Table 1a.

D Comparison with a Model Designed Specifically for CMU-MOSI

The comparison below is centralized on both sides and is therefore reported separately from the federated results of Section 5. We denote by wArms the centralized instance of the framework, that is the construction of Section 3 trained with a single client and hence without Equation (1), and compare it with MissModal Lin & Hu (2023b), which is designed specifically for CMU-MOSI, trained on complete data and evaluated with modalities removed at inference time. Table 8 averages ten independent runs per cell.

Table 8: F-micro on CMU-MOSI in the centralized setting, averaged over ten independent runs. Significance is assessed with a paired t-test; *N/A* indicates no significant difference.

Configuration	MissModal	wArms	Winner
100-text-100-audio-100-video	0.80	**0.82**	wArms
100-text-100-audio-20-video	0.80	0.81	N/A
100-text-20-audio-100-video	0.80	**0.81**	wArms
100-text-20-audio-20-video	0.81	0.81	N/A
20-text-100-audio-100-video	0.67	0.67	N/A
20-text-100-audio-20-video	0.57	**0.68**	wArms
20-text-20-audio-100-video	0.68	0.66	N/A
complex-20-20-20-10-10-10-10	0.71	**0.73**	wArms

wArms wins four of the eight configurations significantly and is indistinguishable from MissModal in the remaining four, so it is never behind by a significant margin. The largest margin appears in the audio-dominated regime, where text is absent for most samples and a method trained on complete data has to cope at inference with a pattern it never saw during training. This is the distinction the construction is built around: the arms are fitted on the missing patterns themselves through the label-free objective of Equation (4), so severe and heterogeneous availability is part of the training distribution rather than a test-time perturbation.

E Why FLEX-MoE and PEPSY behave differently on the two benchmarks.

Both are strong on CMU-MOSI and weak on MM-IMDb, and the contrast is explained by the round budget rather than by the modality count. FLEX-MoE assigns experts by solving a linear program per round from accumulated client–expert fitness scores, and is trained for one hundred rounds in its original evaluation Zhang et al. (2025); PEPSY aligns client-side embedding controls through non-parametric clustering at the server and is trained for five hundred to one thousand rounds Nguyen et al. (2025). Under the common budget of ten rounds used here, the expert routing of the former and the reconfiguration signals of the latter have not stabilized, which is most visible on MM-IMDb, where the label space is large and multi-label. On CMU-MOSI, whose label space is one-dimensional and whose modality statistics are closer to the signal-oriented data these methods were designed for, both make substantially more progress within the same number of rounds. FArms is not subject to this

constraint because the arms are fitted against a label-free reconstruction target that does not depend on the global task state, and the task block is then trained on a stationary feature space (Remark 2).