Computer Science and Information Systems 00(0):0000–0000 https://doi.org/10.2298/CSIS260626049H

Paper Type: Research Paper

Scalable Structured Variational Sparse Tensor Representation Learning for Distributed Multimodal Data Streams

Lianming Hao[1] and Fengxu Zhu[1]

College of Mathematics and Computer, Jilin Normal University,
Siping 136000, China
haolianming@jlnu.edu.cn (corresponding author); mqyzfx@163.com

Abstract. Representation learning for distributed multimodal data streams is an important task in intelligent healthcare, mobile sensing, human activity recognition, and edge-assisted monitoring. Its goal is to learn stable, scalable, and uncertainty-aware representations from continuously arriving multimodal data distributed across multiple clients. However, existing federated learning and multimodal fusion methods often rely on complete modality availability and a shared input structure across clients, making them less suitable for scenarios with client-level modality inconsistency, asynchronous modality arrival, and highly uncertain observations. In addition, many existing approaches mainly depend on parameter averaging, feature reconstruction, or missing-modality generation, which makes it difficult to capture high-order cross-modal dependencies and posterior uncertainty in distributed environments. To address these challenges, this paper proposes a Scalable Structured Variational Sparse Tensor framework, abbreviated as SVST. Specifically, SVST uses sparse low-rank tensor factors to directly generate the natural parameters of a structured variational posterior, enabling high-order multimodal interactions and uncertainty correlations to be modeled during posterior construction. Meanwhile, each client uploads only compact natural-parameter increments and sparse factor updates, and the server performs aggregation in the natural-parameter space. This design avoids raw-data transmission, full neural-parameter averaging, and dense global tensor construction. Furthermore, SVST introduces a missing-modality-aware online ELBO update, delay-aware weighting, and uncertainty-confidence regulation to adapt to incomplete and asynchronous multimodal streams. Experiments on distributed multimodal datasets, including MIMIC, CMU-MOSEI, and OPPORTUNITY, show that SVST achieves overall superior or comparable performance compared with existing federated learning, multimodal fusion, and multimodal variational inference baselines in terms of predictive performance, calibration error, communication cost, and client-level stability. Further robustness analysis indicates that, under increasing modality missingness, stronger asynchronous delay, and non-IID client partitions, the proposed SVST effectively reduces performance degradation and improves prediction reliability and model robustness. These results demonstrate that SVST provides a practical and robust solution for representation learning in distributed multimodal data stream tasks.

Keywords: Distributed Multimodal Data Streams, Structured Variational Inference, Sparse Tensor Factorization, Natural-parameter Aggregation, Missing-modality Learning.

1. Introduction

Distributed multimodal data streams have become common in real-world intelligent systems, including clinical decision support, mobile sensing, human activity recognition, affective computing, and edge-assisted monitoring. In these scenarios, data are continuously generated by multiple clients, each observing different combinations of images, text, audio, physiological signals, event logs, or wearable sensor streams. Compared with centralized multimodal learning, distributed multimodal stream learning faces additional challenges: raw data often cannot be shared because of privacy, bandwidth, or institutional constraints, while modality availability, sampling frequency, local distribution, and communication delay may vary across clients. Therefore, a key challenge is learning scalable and uncertainty-aware representations from incomplete, asynchronous, and distributed observations [5].

Existing federated learning methods provide a foundation for distributed model training. Parameter averaging, proximal regularization, client-drift correction, and normalized aggregation support global training without centralizing raw data. However, these methods generally assume a shared input structure and model interface. In distributed multimodal streams, one client may receive text and sensor data, another may contain images and event records, and a third may observe only a subset of modalities. Consequently, direct neural-parameter averaging may combine updates derived from different modality subsets and uncertainty levels, reducing aggregation reliability [17,34].

Multimodal fusion methods model interactions among heterogeneous modalities. Tensor fusion networks represent high-order interactions, low-rank fusion reduces dense tensor costs, and multimodal Transformers capture dependencies among unaligned sequences. However, most are developed for centralized or relatively complete modality settings and focus on combining available features rather than aggregating uncertain, client-specific posterior information. Tensor-based fusion is also commonly used for feature interaction or compression rather than posterior uncertainty parameterization or defining the statistics exchanged across clients [25].

Recent multimodal federated learning studies address heterogeneous modality availability through reconfigurable representations, contribution-aware aggregation, reconstruction, distillation, or routing. These methods establish modality incongruity as a central federated learning problem. Nevertheless, reconstruction-oriented strategies may introduce bias when missing modalities are highly uncertain or client distributions differ substantially. Moreover, many approaches aggregate model parameters or learned features rather than variational posterior statistics, so uncertainty from client-specific modality subsets is represented only indirectly during server aggregation [22,18].

Probabilistic multimodal representation learning offers another approach to uncertainty under incomplete observations. Multimodal variational autoencoders and product-of-experts models infer latent variables from available modalities and estimate posterior means and variances. However, mean-field or simple expert-combination assumptions may not capture high-order modality dependencies. Tensor decomposition compactly represents such interactions but is usually applied as post-hoc factorization or feature fusion. Its use for structured posterior parameterization remains insufficiently explored in distributed multimodal streams, particularly when posterior statistics must be aligned across heterogeneous clients [29].

Asynchronous modality arrival creates an additional challenge. Clinical records, images, textual notes, audio, visual streams, and wearable sensors may arrive at different times or sampling rates. Although sequence and online learning methods handle temporal misalignment or continuous updating, distributed multimodal systems must also determine how delayed modalities affect posterior inference, how incomplete local objectives are updated, and how unreliable client updates are prevented from dominating aggregation [26].

To address these issues, this paper proposes a Scalable Structured Variational Sparse Tensor framework, abbreviated as SVST. Instead of treating tensor decomposition as post-processing, SVST uses sparse low-rank tensor factors to generate the natural parameters of a structured variational posterior. This allows high-order multimodal interactions and uncertainty correlations to be modeled during posterior construction without forming a dense global interaction tensor.

SVST contains three coupled mechanisms. First, a sparse tensor-parameterized structured posterior models multimodal uncertainty under incomplete observations. Second, distributed natural-parameter aggregation allows clients to upload compact posterior increments and sparse-factor updates rather than raw data, full tensors, or complete neural parameters. Third, missing-modality-aware online inference uses observed modalities, delay-aware weights, and uncertainty confidence to regulate asynchronous or unreliable observations. The novelty therefore lies not in merely combining tensor factorization, variational inference, and federated learning, but in establishing a common posterior-statistics pathway. Sparse tensor factors construct the posterior, natural parameters form the client–server communication object, and missingness, delay, and uncertainty regulate the same update process. Unlike conventional tensor fusion, SVST models both latent representation and posterior uncertainty; unlike centralized multimodal variational models, it communicates aligned posterior statistics; and unlike reconstruction-, distillation-, routing-, or parameter-averaging-based multimodal federated methods, it performs observed-modality inference and reliability-aware aggregation in a shared natural-parameter space. Figure 1 illustrates this workflow.

SVST is evaluated on clinical, language, and sensor-based multimodal datasets. The evaluation covers predictive performance, uncertainty calibration, communication cost, client-level stability, missing-modality degradation, and robustness under bounded asynchronous delay. Repeated runs, confidence intervals, and ablation studies assess component contributions and stability under increasing missingness and asynchronous updates.

The main contributions are as follows. First, this paper formulates distributed multimodal stream representation learning as structured variational inference under missingness, asynchrony, client heterogeneity, and communication constraints. Second, it proposes a posterior-construction mechanism in which sparse low-rank factors directly generate structured posterior natural parameters rather than deterministic fused features. Third, it develops a distributed inference mechanism that communicates natural-parameter increments and compact factors instead of complete model parameters, reconstructed modalities, or distilled features. Fourth, it introduces a unified posterior-update procedure in which observed-modality evidence, delay, and uncertainty jointly regulate local inference and global aggregation, and evaluates it across multiple datasets and baseline categories.

2. Related Works

This section reviews studies related to representation learning for distributed multimodal data streams. The central problem is how to learn stable and scalable multimodal representations when missing modalities, asynchronous arrival, client heterogeneity, and communication constraints coexist. Existing studies can be grouped into five categories: traditional federated optimization, multimodal fusion and cross-modal representation learning, multimodal federated learning with missing modalities, probabilistic representation learning based on tensor decomposition and variational inference, and online inference for asynchronous multimodal streams. These categories form a natural progression: federated optimization addresses distributed training, multimodal fusion models cross-modal interaction, missing-modality federated learning considers client-level modality inconsistency, tensor and variational methods provide structural and probabilistic modeling tools, and online inference handles dynamic data arrival.

2.1. Traditional Federated Optimization Methods

Traditional federated learning methods mainly address distributed model training when raw data cannot be centrally shared. FedAvg performs weighted averaging of client model parameters at the server. FedProx introduces proximal regularization into the local objective to reduce update bias caused by non-IID data and system heterogeneity. SCAFFOLD uses control variates to correct client drift, while FedNova normalizes local updates to mitigate inconsistencies caused by different numbers of local training steps. These methods form important foundations for federated learning and serve as traditional optimization baselines in this paper [14,10,28].

However, these methods generally assume that clients share similar input structures. This assumption does not always hold in distributed multimodal streams, where clients may observe different modality combinations and modalities may be unavailable because of sampling differences, communication delays, or sensor states. The problem is therefore not only how to average model parameters, but also how to aggregate posterior information generated from heterogeneous modality subsets and uncertainty levels. Improving parameter aggregation alone is insufficient without explicitly modeling cross-modal interactions and observation uncertainty.

2.2. Multimodal Fusion and Cross-Modal Representation Learning Methods

Multimodal representation learning aims to learn unified representations from heterogeneous data sources. Early methods commonly use feature concatenation, decision-level fusion, or attention mechanisms to map text, image, audio, sensor, or clinical-event modalities into a shared space. Tensor Fusion Network explicitly models high-order multimodal interactions through tensor outer products. Low-rank Multimodal Fusion reduces the cost of high-order fusion through low-rank tensor decomposition. Multimodal Transformer uses cross-modal attention to handle unaligned multimodal sequences and has been widely applied to language, audio, and visual sequence modeling [35,16,24].

These methods demonstrate that multimodal learning should not rely only on simple feature concatenation. Tensor-based fusion captures high-order interactions, while cross-modal attention supports dynamic relation modeling. TFN, LMF, and MulT therefore provide important technical baselines for this paper.

Nevertheless, most multimodal fusion methods assume centralized or relatively complete modality access. Even when they address sequence misalignment, they generally focus on temporal differences within the same sample rather than modality-structure differences across clients. Moreover, tensor decomposition and attention are typically used to generate deterministic fused representations, rather than to parameterize a structured posterior or define the statistics communicated in distributed inference. SVST differs by using sparse tensor factors to generate posterior natural parameters, so high-order interactions influence both latent representation and uncertainty rather than only feature fusion.

2.3. Multimodal Federated Learning with Missing Modalities

Recent studies have combined missing-modality learning with federated learning. Incongruent MMFL focuses on medical vision-language tasks in which clients have different modality combinations. PEPSY studies reconfigurable representations under missing modalities. FedMobile models client knowledge contributions under incomplete mobile or sensor modalities. FedMIR represents approaches that mitigate unavailable modalities through imputation, routing, or reconstruction [15,30,20].

These methods are closely related to this study because they recognize that multimodal federated learning is affected not only by label distribution, sample size, and system heterogeneity, but also by structural differences in modality availability [19,33]. Ordinary parameter averaging is often insufficient when clients cannot provide the same modalities. Incongruent MMFL, PEPSY, FedMobile, and FedMIR therefore serve as direct baselines for evaluating SVST under missing modalities and client heterogeneity.

However, several limitations remain. First, many methods treat missing modalities primarily as a feature-recovery problem and rely on generation, reconstruction, distillation, or routing. These strategies may introduce bias when missing information is highly uncertain or client distributions differ substantially. Second, existing methods generally aggregate model parameters, features, or distillation signals rather than variational posterior statistics. Consequently, the uncertainty associated with client-specific modality subsets is represented only indirectly during server aggregation. Third, many methods focus on static missingness or round-level modality inconsistency, while asynchronous arrival and online posterior updating receive less attention. In contrast, SVST does not reconstruct unavailable raw modalities; it constructs the local objective from observed modalities and regulates posterior updates using completeness, delay, and uncertainty information.

2.4. Probabilistic Representation Methods based on Tensor Decomposition and Variational Inference

Tensor decomposition provides an effective tool for modeling high-order multimodal interactions. Through CP decomposition, Tucker decomposition, or low-rank approximation, high-dimensional interactions can be represented using compact factors, reducing the cost of explicit tensor outer products. Low-rank tensor fusion shows that major cross-modal patterns can often be captured without constructing a full dense tensor [21,7,23].

Variational inference, in contrast, emphasizes uncertainty modeling. MVAE, PoE-VAE, and later multimodal VAE methods construct latent-variable distributions through multiple modality encoders. They can infer representations from available modalities and

output posterior means and variances, making them suitable for incomplete and noisy observations. MVAE/PoE-VAE therefore provides an important probabilistic baseline.

Despite these advantages, tensor decomposition and variational inference are usually studied separately. Tensor decomposition is commonly used for feature fusion, dimensionality reduction, or interaction compression, rather than for generating posterior natural parameters. Variational methods express uncertainty, but mean-field assumptions or simple expert combinations may not capture complex high-order correlations. This limitation becomes more evident when different clients infer from different modality subsets. SVST links these two directions by allowing sparse low-rank factors to directly parameterize a structured posterior, thereby embedding high-order interaction structure into both posterior location and uncertainty. It further extends this posterior representation to distributed aggregation through natural-parameter increments.

2.5. Online Inference Methods for Asynchronous Multimodal Streams

Distributed multimodal streams also involve asynchronous modality arrival. Clinical events, medical images, text records, language, audio, visual signals, and wearable sensors may be generated at different times or sampling rates. Existing unaligned sequence models can handle certain temporal differences, but they are usually designed for centralized modeling rather than federated settings where missing modalities and communication constraints coexist [13].

Online learning and streaming inference methods update model parameters or latent representations as new data arrive. In multimodal federated scenarios, however, client updates may originate from different time windows, modality subsets, and uncertainty levels. Without reliability regulation, delayed or uncertain observations may exert excessive influence on global statistics.

In summary, traditional federated optimization pays limited attention to modality-structure heterogeneity; multimodal fusion usually produces deterministic representations under centralized access; missing-modality multimodal federated learning often relies on reconstruction, routing, distillation, or parameter averaging; and multimodal variational inference generally remains centralized and uses relatively simple posterior structures. The fundamental distinction of SVST is therefore not the simultaneous use of tensor factorization, variational inference, federated learning, and online updating. Instead, these elements operate on one shared statistical object: sparse tensor factors generate structured posterior natural parameters, natural-parameter increments constitute the client–server communication unit, and missingness, delay, and uncertainty regulate the same posterior update and aggregation process.

Accordingly, SVST differs from tensor-fusion methods by using tensor factors for posterior construction rather than deterministic feature fusion; it differs from multimodal variational methods by communicating structured posterior statistics across heterogeneous clients; and it differs from missing-modality multimodal federated methods by avoiding raw-modality reconstruction and full-parameter averaging. During distributed training, clients transmit compact natural-parameter and sparse-factor increments, while online inference uses observed-modality evidence and reliability-aware weighting under bounded missingness and delay.

3. Used Methods

This section proposes a Scalable Structured Variational Sparse Tensor Framework for robust representation learning in distributed multimodal data streams. The framework is designed to address incomplete modality observations, asynchronous modality arrival, client-level heterogeneity, and communication constraints in distributed learning environments. Section 3.1 first formulates the research problem, including distributed multimodal streaming inputs, learning outputs, task objectives, optimization targets, and uncertainty modeling. Section 3.2 introduces the overall architecture of the proposed framework, which consists of a modality-specific encoding layer, a sparse tensor-parameterized structured variational posterior module, a distributed natural-parameter aggregation module, and an online robust inference module, with efficient information interaction achieved through low-rank sparse tensor factors and variational natural parameters. Section 3.3 further details the sparse tensor-parameterized structured variational posterior, which is designed based on structured variational inference and low-rank sparse tensor factorization theory to enhance uncertainty-aware multimodal representation learning. Section 3.4 introduces the distributed natural-parameter aggregation mechanism, which aggregates client-side variational statistics in the natural-parameter space to avoid centralized full-tensor construction and reduce communication cost. Section 3.5 presents the online robust inference mechanism for missing modalities and asynchronous streams, where observed-subset likelihood masking, delay-aware weighting, and online ELBO updates are jointly used to stabilize representation learning under bounded incomplete observations. Section 3.6 describes the training and optimization strategy, including local variational updates, sparse tensor factor optimization, natural-parameter communication, server-side aggregation, and dynamic online updating. Section 3.7 provides theoretical analysis of communication complexity, update stability, and uncertainty-estimation variance. Section 3.8 summarizes the complete training and inference workflow in pseudocode [1].

3.1. Problem Formulation

The paper considers a distributed multimodal streaming scenario with K clients and M modalities. Each client continuously receives local multimodal observations over time, while different clients may have different modality availability, data distributions, sampling rates, and communication conditions. Let the client set be defined as $K = \{1, 2, \ldots, K\}$ and the modality set as $M = \{1, 2, \ldots, M\}$. At time step t, client k receives a multimodal streaming sample represented as:

$$x_{k,t} = \{x_{k,t}^{(m)} \mid m \in M\} \tag{1}$$

Here, $x_{k,t}^{(m)}$ denotes the raw observation of modality m at client k and time t. Due to missing sensors, incomplete data transmission, privacy restrictions, or asynchronous acquisition, not all modalities are always observable. The observed modality subset is denoted as:

$$\Omega_{k,t} = \{m \in M \mid x_{k,t}^{(m)} \text{ isobserved}\} \tag{2}$$

The corresponding binary missing-modality mask is defined as:

$$r_{k,t}^{(m)} = I(m \in \Omega_{k,t}) \tag{3}$$

where $r_{k,t}^{(m)} = 1$ indicates that modality m is available at client k and time t, while $r_{k,t}^{(m)} = 0$ indicates that the modality is missing. To model asynchronous arrival, each modality is associated with a timestamp $\tau_{k,t}^{(m)}$. The delay of modality m relative to the current inference time t is defined as:

$$\Delta_{k,t}^{(m)} = t - \tau_{k,t}^{(m)} \tag{4}$$

The paper focuses on bounded missingness and bounded asynchronous delay, where the missing rate and delay satisfy:

$$E_{k,t}\left[1 - r_{k,t}^{(m)}\right] \leq \epsilon_m, \qquad 0 \leq \Delta_{k,t}^{(m)} \leq \Delta_{\max} \tag{5}$$

This assumption reflects common distributed sensing scenarios where modality absence and communication delay occur frequently but do not correspond to permanent sensor failure or indefinitely disconnected clients.

Each raw modality is first mapped into a latent representation through a modality-specific encoder:

$$h_{k,t}^{(m)} = E_{\phi_m}\left(x_{k,t}^{(m)}\right) \tag{6}$$

where E_{ϕ_m} denotes the encoder of modality m with parameters ϕ_m. The proposed framework does not claim to solve complete cross-modal alignment from raw observations. Instead, it assumes that modality-specific encoders provide initial latent features, and the main objective is to learn a robust distributed multimodal representation over these latent features. The latent multimodal representation for client k at time t is denoted as $z_{k,t}$.

The learning objective is to infer a posterior distribution over latent representations under incomplete and asynchronous multimodal streams. Let θ denote generative parameters and ψ denote variational parameters. The ideal posterior $p_\theta(z_{k,t} \mid x_{k,t})$ is generally intractable, so the paper introduces a structured variational posterior $q_\psi(z_{k,t} \mid x_{k,t}^{\Omega})$, where $x_{k,t}^{\Omega}$ denotes the observed modality subset. The variational learning objective is formulated through an online evidence lower bound:

$$\begin{aligned} L_{k,t} = {} & E_{q_\psi(z_{k,t} \mid x_{k,t}^{\Omega})}\left[\log p_\theta(x_{k,t}^{\Omega} \mid z_{k,t})\right] \\ & - \mathrm{KL}\left(q_\psi(z_{k,t} \mid x_{k,t}^{\Omega}) \| p_\theta(z_{k,t})\right) \end{aligned} \tag{7}$$

The global objective over distributed clients and streaming time steps is:

$$\max_{\theta,\psi,A} \sum_{t=1}^{T} \sum_{k=1}^{K} w_{k,t} L_{k,t} - \lambda_s \|A\|_1 - \lambda_f R_{\text{factor}} \tag{8}$$

where A denotes sparse tensor factors, $w_{k,t}$ is the reliability weight of client k at time t, λ_s controls sparsity, and λ_f controls factor consistency. This formulation integrates representation learning, uncertainty estimation, sparse tensor regularization, and distributed online inference into a unified optimization objective.

3.2. Overall Architecture of the Proposed SVST Framework

The proposed method is named the Structured Variational Sparse Tensor Framework, abbreviated as SVST. The architecture is designed around a central methodological principle: distributed multimodal uncertainty should not be represented by point estimates or post-hoc tensor compression, but by a structured variational posterior whose natural parameters are directly generated from sparse low-rank tensor factors. This design allows the framework to capture high-order multimodal interactions while keeping communication and memory costs scalable [11].

The overall architecture contains four tightly coupled components. The first component is the modality-specific encoding layer, which transforms heterogeneous raw modalities into latent features. For an observed modality m, the encoded representation is obtained as:

$$h_{k,t}^{(m)} = E_{\phi_m}(x_{k,t}^{(m)}) \tag{9}$$

For missing modalities, no raw-modality generation is performed. Instead, the inference objective is constructed only over the observed modality subset $\Omega_{k,t}$, and unavailable likelihood terms are excluded through the modality mask. This operation is likelihood masking rather than probabilistic marginalization over missing raw-modality values.

The second component is the sparse tensor-parameterized structured variational posterior module. This module constructs the variational posterior by mapping sparse low-rank tensor factors into posterior natural parameters. Rather than applying tensor decomposition after representation learning, the tensor factors participate directly in posterior construction. This is the core difference between the proposed framework and conventional tensor-factorization-based dimensionality reduction.

The third component is the distributed natural-parameter aggregation module. Each client computes local posterior statistics and uploads compact natural-parameter increments and sparse tensor factor updates. The server aggregates these statistics in the natural-parameter space rather than averaging complete neural parameters or reconstructing a full global tensor. This mechanism is compatible with the additive structure of exponential-family variational inference and reduces the communication burden.

The fourth component is the online robust inference module. This module updates the local variational posterior under missing modalities and asynchronous arrival. Delayed modalities are not simply discarded; their contributions are weighted by delay-aware reliability terms. The online ELBO is updated incrementally, and uncertainty estimates guide the magnitude of local factor updates.

Therefore, the framework forms a coherent algorithmic chain: sparse tensor factors define the structure of the variational posterior; the posterior natural parameters provide compact statistics for distributed aggregation; and the aggregated statistics support online inference under incomplete and asynchronous observations.

3.3. Sparse Tensor-Parameterized Structured Variational Posterior

A central limitation of conventional variational representation learning is the frequent use of mean-field assumptions. A standard mean-field posterior can be written as:

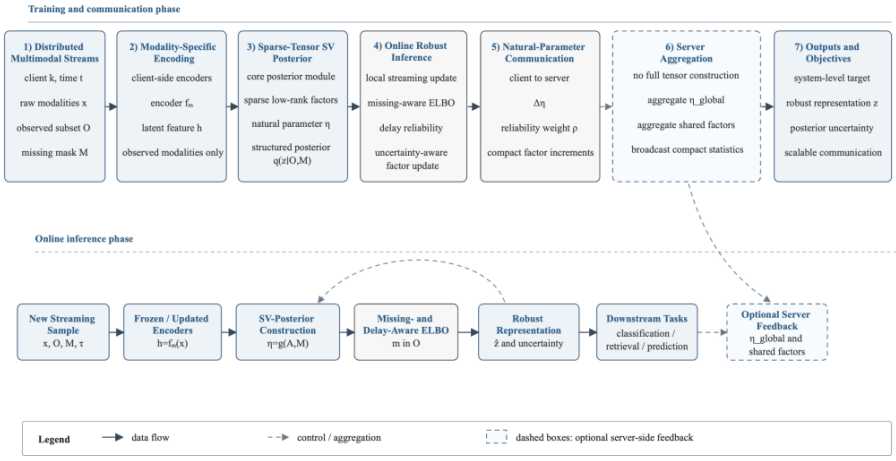


Fig. 1. Overall architecture of the proposed SVST framework for distributed multimodal data streams.

$$q_\psi^{\mathrm{MF}}(\mathbf{z}_{k,t}) = \prod_{j=1}^{d_z} q_\psi(z_{k,t,j}). \tag{10}$$

Although computationally convenient, this form ignores dependencies among latent dimensions and cannot explicitly capture high-order multimodal interactions. Such dependencies are important because modality relations may vary across clients, time steps, and missing-observation patterns.

To address this limitation, the proposed framework defines a structured variational posterior:

$$q_\psi\left(\mathbf{z}_{k,t} \mid \mathbf{x}_{k,t}^{\Omega}, \mathbf{r}_{k,t}\right) = \exp\left\{\langle \boldsymbol{\eta}_{k,t}, T(\mathbf{z}_{k,t})\rangle - A(\boldsymbol{\eta}_{k,t})\right\}, \tag{11}$$

where $\boldsymbol{\eta}_{k,t}$ denotes the natural parameter, $T(\mathbf{z}_{k,t})$ denotes sufficient statistics, and $A(\boldsymbol{\eta}_{k,t})$ is the log-partition function. The key design is that $\boldsymbol{\eta}_{k,t}$ is generated by sparse low-rank tensor factors rather than an unconstrained dense parameterization. The modality mask $\mathbf{r}_{k,t} \in \{0,1\}^M$ is included explicitly to distinguish unavailable modalities from observed near-zero features.

Let the high-order multimodal interaction tensor be denoted as $\mathcal{W}$. Direct construction is expensive because its size grows multiplicatively with modality dimensions. It is therefore parameterized through a sparse CP-style factorization:

$$\mathcal{W} \approx \sum_{r=1}^{R} \lambda_r \mathbf{a}_r^{(1)} \circ \mathbf{a}_r^{(2)} \circ \cdots \circ \mathbf{a}_r^{(M)}, \tag{12}$$

where R is the tensor rank, λ_r is the rank weight, $\mathbf{a}_r^{(m)} \in \mathbb{R}^{d_h}$ is the factor vector for modality m, and $\circ$ denotes the outer product. The sparse factor set is:

$$\mathcal{A} = \left\{ \lambda_r, \mathbf{a}_r^{(m)} \right\}_{r=1,m=1}^{R,M}. \tag{13}$$

For each modality, the masked feature is $\widetilde{\mathbf{h}}_{k,t}^{(m)} = r_{k,t}^{(m)} \mathbf{h}_{k,t}^{(m)}$. The rank-$R$ sparse interaction vector $\mathbf{s}_{k,t} = [s_{k,t,1}, \ldots, s_{k,t,R}]^\top$ is computed as

$$s_{k,t,r} = \lambda_r \prod_{m=1}^{M} \left[r_{k,t}^{(m)} \langle \mathbf{a}_r^{(m)}, \mathbf{h}_{k,t}^{(m)} \rangle + 1 - r_{k,t}^{(m)} \right].$$

The term $1 - r_{k,t}^{(m)}$ provides a neutral value for missing modalities, preventing one unavailable modality from nullifying the full interaction. The mapping input is

$$\mathbf{v}_{k,t} = \left[\widetilde{\mathbf{h}}_{k,t}^{(1)}; \ldots; \widetilde{\mathbf{h}}_{k,t}^{(M)}; \mathbf{r}_{k,t}; \mathbf{s}_{k,t} \right] \in \mathbb{R}^{Md_h + M + R}.$$

The natural-parameter mapping is defined as

$$\left(\boldsymbol{\eta}_{1,k,t}, \boldsymbol{\delta}_{k,t}, \mathbf{U}_{k,t} \right) = g_\omega(\mathbf{v}_{k,t}), \tag{14}$$

where g_ω is a two-layer multilayer perceptron with 256 and 128 hidden units, ReLU activation, and dropout of 0.1. Its three output heads generate

$$\boldsymbol{\eta}_{1,k,t} \in \mathbb{R}^{d_z}, \quad \boldsymbol{\delta}_{k,t} \in \mathbb{R}^{d_z}, \quad \mathbf{U}_{k,t} \in \mathbb{R}^{d_z \times R}.$$

The default implementation uses $d_h = d_z = 128$ and $R = 16$, giving an input dimension of $128M + M + 16$.

Figure 2 illustrates the sparse tensor-parameterized posterior. Observed features and masks produce the rank-R interaction vector, while the mapping network outputs the first natural parameter, diagonal precision coefficients, and a low-rank precision factor.

The posterior is implemented as a low-rank-plus-diagonal precision Gaussian:

$$q_\psi \left(\mathbf{z}_{k,t} \mid \mathbf{x}_{k,t}^{\Omega}, \mathbf{r}_{k,t} \right) = \mathcal{N} \left(\boldsymbol{\mu}_{k,t}, \boldsymbol{\Lambda}_{k,t}^{-1} \right),$$

where

$$\boldsymbol{\Lambda}_{k,t} = \operatorname{diag} \left[\operatorname{softplus}(\boldsymbol{\delta}_{k,t}) + \varepsilon \right] + \mathbf{U}_{k,t} \mathbf{U}_{k,t}^\top.$$

The Gaussian natural parameters are:

$$\boldsymbol{\eta}_{k,t} = \left(\boldsymbol{\eta}_{1,k,t}, \boldsymbol{\eta}_{2,k,t} \right) = \left(\boldsymbol{\Lambda}_{k,t} \boldsymbol{\mu}_{k,t}, -\frac{1}{2} \boldsymbol{\Lambda}_{k,t} \right). \tag{15}$$

The posterior mean is recovered by solving $\boldsymbol{\Lambda}_{k,t} \boldsymbol{\mu}_{k,t} = \boldsymbol{\eta}_{1,k,t}$.

This posterior is neither mean-field nor unrestricted full-covariance. The positive diagonal term models dimension-specific uncertainty, while $\mathbf{U}_{k,t} \mathbf{U}_{k,t}^\top$ c aptures low-rank latent dependencies. Positive definiteness follows from the softplus-transformed diagonal, $\varepsilon = 10^{-6}$, and the positive-semidefinite low-rank term. Linear solves are used instead of explicit matrix inversion, and the structure requires $O(d_z R + d_z)$ rather than $O(d_z^2)$ precision parameters.

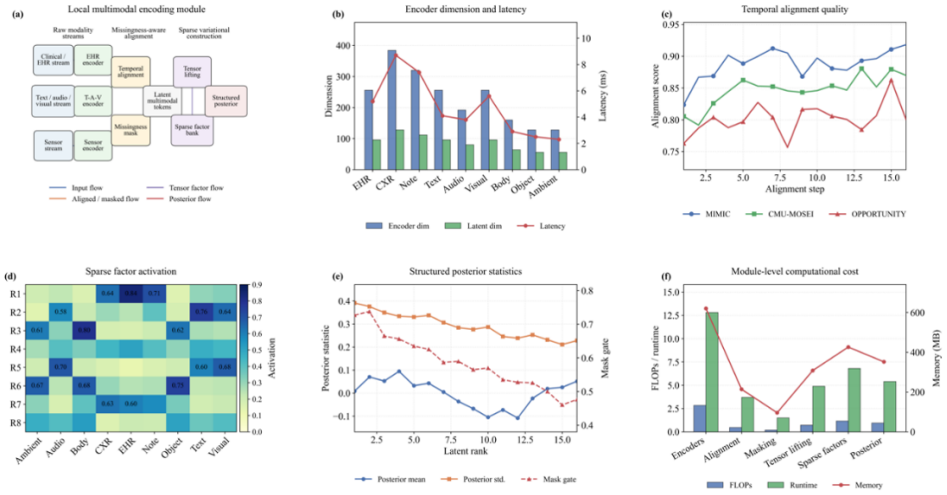


Fig. 2. Sparse tensor-parameterized structured variational posterior module.

Sparsity is imposed through an l_1-regularized factor penalty:

$$\mathcal{R}_{\text{sparse}} = \sum_{r=1}^{R} \sum_{m=1}^{M} \|\mathbf{a}_r^{(m)}\|_1. \tag{16}$$

This penalty selects informative cross-modal interactions when modalities are noisy, redundant, or frequently missing. The posterior learning objective becomes:

$$\mathcal{L}_{k,t}^{\text{SV}} = \mathcal{L}_{k,t} - \lambda_s \mathcal{R}_{\text{sparse}}. \tag{17}$$

This design provides two benefits. First, the structured posterior captures uncertainty correlations ignored by mean-field approximations. Second, sparse low-rank factors avoid dense high-order tensors, supporting scalable distributed posterior construction.

3.4. Distributed Natural-Parameter Aggregation for Multimodal Streams

The second core mechanism is distributed natural-parameter aggregation. Conventional federated learning usually aggregates model parameters through weighted averaging. However, direct parameter averaging is not always suitable for structured variational inference because different clients may have different modality subsets, different uncertainty levels, and different posterior statistics. The proposed framework instead performs aggregation in the natural-parameter space [6]. This aggregation assumes a shared latent space, posterior family, and natural-parameter coordinate system across clients. All modality-specific encoders project observations into the same d_z-dimensional space, and all clients use the low-rank-plus-diagonal Gaussian posterior defined in Section 3.3. Under these conditions, local natural-parameter increments are dimensionally and semantically compatible.

At client k, the local variational update produces a natural-parameter increment:

$$\Delta\eta_{k,t} = \eta_{k,t}^{\text{local}} - \eta_{t-1}^{\text{global}} \tag{18}$$

The reliability of this update depends on sample size, modality completeness, delay, and posterior uncertainty. The paper defines the client reliability weight as:

$$w_{k,t} = \frac{n_{k,t} c_{k,t} d_{k,t} u_{k,t}}{\sum_{j=1}^{K} n_{j,t} c_{j,t} d_{j,t} u_{j,t}} \tag{19}$$

where $n_{k,t}$ is the effective local sample size, $c_{k,t}$ measures modality completeness, $d_{k,t}$ denotes delay reliability, and $u_{k,t}$ denotes uncertainty confidence. A possible completeness score is:

$$c_{k,t} = \frac{1}{M} \sum_{m=1}^{M} r_{k,t}^{(m)} \tag{20}$$

The delay reliability can be defined as:

$$d_{k,t} = \exp(-\beta_{\Delta} \bar{\Delta}_{k,t}) \tag{21}$$

where $\bar{\Delta}_{k,t}$ is the average modality delay and β_{Δ} controls delay sensitivity. The uncertainty confidence can be defined as:

$$u_{k,t} = \exp(-\beta_u \operatorname{Tr}(\Sigma_{k,t})) \tag{22}$$

The server updates the global natural parameter as:

$$\eta_t^{\text{global}} = (1 - \rho_t)\eta_{t-1}^{\text{global}} + \rho_t \sum_{k=1}^{K} w_{k,t} \Delta\eta_{k,t} \tag{23}$$

Here, ρ_t is an online aggregation step size. This formulation differs from FedAvg because it aggregates posterior statistical contributions rather than full neural parameters. Under the shared latent space, posterior family, and natural-parameter coordinate system, the additive structure of exponential-family natural parameters supports the aggregation of compatible local increments. However, this is a reliability-weighted approximate procedure rather than exact Bayesian fusion. Non-IID data, missing modalities, and local optimization errors may still produce heterogeneous posterior estimates. Since natural parameters in exponential-family distributions correspond to additive sufficient statistics, the aggregation is better aligned with variational inference.

The sparse tensor factors are also aggregated, but only through their compact low-rank representation:

$$A_t^{\text{global}} = (1 - \rho_t) A_{t-1}^{\text{global}} + \rho_t \sum_{k=1}^{K} w_{k,t} A_{k,t} \tag{24}$$

Because CP factors are ambiguous under permutation, sign, and scale transformations, local factors are stabilized before aggregation. Each round, clients initialize from the same broadcast global factors to preserve rank ordering. After local optimization, each factor is normalized to unit l_2 norm, with its scale absorbed into λ_r, and its sign is aligned

with the corresponding global factor using a nonnegative inner product. The component-wise consistency term in Equation (32) further limits rank switching and local drift. Thus, Equation (24) aggregates normalized, reference-aligned factors rather than arbitrary CP decompositions. This alignment procedure improves cross-client factor correspondence but does not establish a formal identifiability guarantee for the CP decomposition.

Although clients transmit only natural-parameter increments and compact sparse-factor updates rather than raw data, this mechanism does not provide formal privacy guarantees. The transmitted statistics may still reveal information about local observations through gradient inversion, membership inference, or related statistical attacks.

Figure 3 provides an overview of the distributed natural-parameter aggregation process. Local clients first construct posterior statistics and sparse tensor updates, which are weighted according to modality completeness, delay reliability, and uncertainty confidence before being aggregated on the server. The figure also summarizes the resulting aggregation consistency, communication efficiency, and stability characteristics of the proposed framework.

The server never constructs the full tensor W. If a full interaction tensor has complexity $O(\prod_{m=1}^{M} d_m)$, the sparse low-rank representation only requires:

$$O(R \sum_{m=1}^{M} d_m) \tag{25}$$

This reduction is essential for scalable distributed multimodal representation learning. It also makes the framework experimentally verifiable through communication-cost comparisons, rank-sensitivity analysis, and client-scaling experiments.

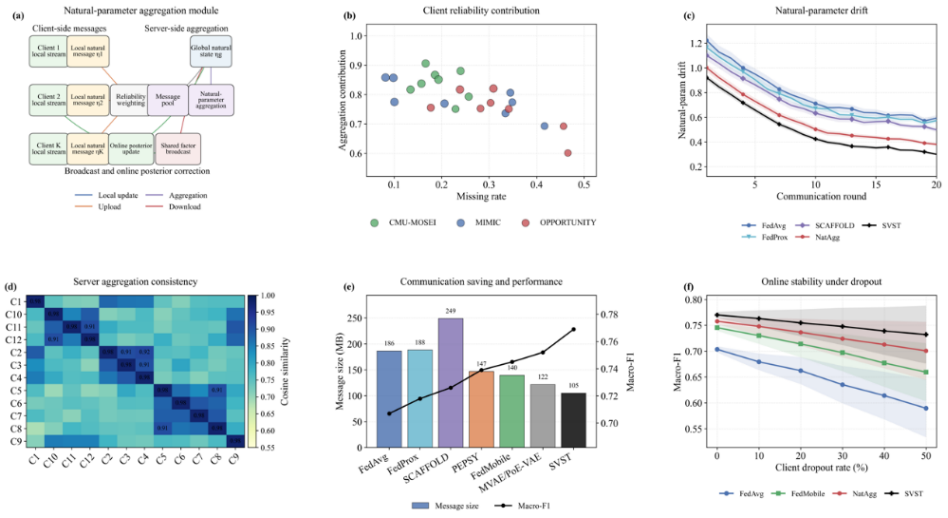


Fig. 3. Distributed natural-parameter aggregation module.

3.5. Online Robust Inference under Missing Modalities and Asynchronous Streams

The third core mechanism addresses online inference when multimodal streams are incomplete and asynchronous. The proposed method does not synthesize missing raw modalities. Instead, unavailable likelihood terms are masked, and the variational posterior is conditioned only on observed modalities. Thus, missing-modality handling in this paper denotes observed-subset posterior inference rather than raw-modality generation or probabilistic marginalization over missing inputs.

Given the observed modality subset $\Omega_{k,t}$, the reconstruction term is defined as:

$$\log p_\theta(x_{k,t}^{\Omega} \mid z_{k,t}) = \sum_{m \in \Omega_{k,t}} \log p_{\theta_m}(x_{k,t}^{(m)} \mid z_{k,t}) \tag{26}$$

Equivalently, it can be implemented using the binary mask $M_{k,t}^{(m)}$:

$$\log p_\theta\left(\mathbf{x}_{k,t}^{\Omega} \mid \mathbf{z}_{k,t}\right) = \sum_{m=1}^{M} M_{k,t}^{(m)} \log p_{\theta_m}\left(\mathbf{x}_{k,t}^{(m)} \mid \mathbf{z}_{k,t}\right),$$

where $M_{k,t}^{(m)} = 1$ for observed modalities and 0 otherwise. The mask removes unavailable likelihood factors while remaining an explicit posterior input.

$$\begin{aligned} L_{k,t}^{\Omega} = E_{q_\psi(z_{k,t} \mid x_{k,t}^{\Omega})} & \left[\sum_{m \in \Omega_{k,t}} \log p_{\theta_m}(x_{k,t}^{(m)} \mid z_{k,t}) \right] \\ & - \mathrm{KL}\left(q_\psi(z_{k,t} \mid x_{k,t}^{\Omega}) \| p_\theta(z_{k,t})\right) \end{aligned} \tag{27}$$

This objective avoids reconstructing unavailable modalities and therefore reduces the bias that may arise from generated missing inputs.

For asynchronous observations, each observed modality receives a delay-aware weight:

$$\alpha_{k,t}^{(m)} = \frac{r_{k,t}^{(m)} \exp(-\gamma \Delta_{k,t}^{(m)})}{\sum_{j \in \Omega_{k,t}} \exp(-\gamma \Delta_{k,t}^{(j)})} \tag{28}$$

Here, $r_{k,t}^{(m)} \in (0, 1]$ denotes modality reliability, $\Delta_{k,t}^{(m)}$ is the arrival delay, and $\gamma > 0$ controls delay decay. Exponential decay is adopted because the relevance of an asynchronous observation is expected to decrease continuously with its delay rather than disappear at a predefined cutoff. The exponential function is positive, monotonic, and smooth, allowing moderately delayed modalities to retain a limited contribution while progressively suppressing highly stale observations. Compared with a linear decay function, it avoids an abrupt zero-weight boundary and requires only one decay parameter. Moreover, the temporal memory of the weighting function can be interpreted through its half-life, $\ln 2/\gamma$.

The delay-aware ELBO is then defined as:

$$R_{k,t}^{\mathrm{async}} = \sum_{m \in \Omega_{k,t}} \alpha_{k,t}^{(m)} \log p_{\theta_m}(x_{k,t}^{(m)} \mid z_{k,t}) \tag{29}$$

The notation $\bar{\mathcal{L}}_{k,t}$ distinguishes the smoothed objective from the instantaneous ELBO and avoids a self-referential definition.

The online ELBO is updated by exponential smoothing:

$$\widehat{L}_{k,t} = (1 - \zeta_t)\widehat{L}_{k,t-1} + \zeta_t L^{\Omega}_{k,t} \tag{30}$$

The local sparse tensor factors are updated through uncertainty-aware gradients:

$$A_{k,t+1} = \mathrm{prox}_{\alpha_t \lambda_s}\left(A_{k,t} + \alpha_t u_{k,t} \nabla_A \widehat{L}_{k,t}\right) \tag{31}$$

Here, $\mathrm{prox}_{\alpha_t \lambda_s}$ denotes the proximal operator for sparse regularization. The confidence score $u_{k,t} \in (0, 1]$, derived from posterior uncertainty, scales the update magnitude. Higher uncertainty produces a smaller value and suppresses unstable sparse-factor updates.

This module assumes bounded missingness and delay and does not cover permanent modality absence, continuous sensor failure, or indefinite disconnection. When these assumptions are violated, performance may degrade because posterior updates receive insufficient evidence and uncertainty estimates become less reliable. The system can retain partial functionality by using available modalities and suppressing unreliable updates. However, permanent loss or complete disconnection requires fault detection, modality replacement, or adaptive client reconfiguration, which remain beyond the current scope.

3.6. Training and Optimization Strategy

The training strategy integrates stochastic variational inference, sparse tensor factor optimization, and distributed natural-parameter communication. Training proceeds through alternating local and global updates. At each communication round, each client receives the latest global natural parameter and shared sparse tensor factors. The client then processes its local streaming mini-batch, constructs the observed modality subset, computes the structured variational posterior, and updates local natural parameters and tensor factors.

The overall optimization objective is:

$$\begin{aligned} J(\theta, \psi, A) = &- \sum_{t=1}^{T} \sum_{k=1}^{K} w_{k,t} L^{\Omega}_{k,t} + \lambda_s R_{\mathrm{sparse}} \\ &+ \lambda_f \| A_{k,t} - A^{\mathrm{global}}_t \|_2^2 \end{aligned} \tag{32}$$

Here, $\bar{\mathcal{L}}_{k,t}$ is the smoothed delay-aware ELBO defined in Section 3.5, and $w_{k,t}$ is the normalized client reliability weight. The first term maximizes the observed-modality and delay-aware online ELBO, the second encourages sparse cross-modal interaction selection, and the third is applied component-wise after factor normalization and sign alignment, keeping each local rank component close to its corresponding global reference while allowing client-specific adaptation. Together with shared initialization, it reduces permutation drift during local optimization.

Because the posterior uses a low-rank-plus-diagonal precision structure, latent sampling is performed through precision-based reparameterization. Let $\boldsymbol{\Lambda}_{k,t} = \mathbf{C}_{k,t}\mathbf{C}_{k,t}^{\top}$,

where $\mathbf{C}_{k,t}$ is the Cholesky factor of the positive-definite precision matrix. The latent variable is sampled as:

$$z_{k,t} = \mu_{k,t} + \sigma_{k,t} \odot \epsilon, \qquad \epsilon \sim N(0, I) \tag{33}$$

The posterior mean is recovered by solving $\boldsymbol{\Lambda}_{k,t}\boldsymbol{\mu}_{k,t} = \boldsymbol{\eta}_{1,k,t}$. Triangular linear solves are used for mean recovery and sampling, avoiding explicit inversion of $\boldsymbol{\Lambda}_{k,t}$ while preserving differentiability.

The encoder, decoder, and natural-parameter mapping parameters are updated by stochastic gradients, whereas sparse factors use the proximal-gradient rule in Equation (31). The diagonal precision coefficients and low-rank factors generated by g_ω are jointly optimized through backpropagation. Periodic global aggregation further reduces communication overhead.

During inference, a client only requires the observed modality subset, current global natural parameter, and shared sparse tensor factors. The framework outputs the posterior mean as the robust latent representation and the posterior covariance as uncertainty estimation:

$$\widehat{z}_{k,t} = E_{q_\psi(z_{k,t}|x^{\Omega}_{k,t})}[z_{k,t}], \qquad \widehat{U}_{k,t} = \mathrm{Tr}(\Sigma_{k,t}) \tag{34}$$

The representation $\widehat{z}_{k,t}$ can be used for downstream classification, retrieval, clustering, prediction, or anomaly detection, while $\widehat{U}_{k,t}$ supports uncertainty-aware rejection or confidence calibration.

3.7. Theoretical Analysis

The theoretical analysis focuses on properties that are directly supported by the proposed design: communication complexity, online update stability, and uncertainty-estimation variance. The paper avoids claiming global optimality because the complete objective involves non-convex neural encoders, sparse tensor factors, and asynchronous online updates.

First, the communication complexity is reduced by transmitting sparse low-rank factors and natural parameters instead of full tensors. A full multimodal interaction tensor requires:

$$C_{\text{full}} = O(\prod_{m=1}^{M} d_m) \tag{35}$$

By contrast, the proposed factorized representation requires:

$$C_{\text{SVST}} = O(R \sum_{m=1}^{M} d_m + d_\eta) \tag{36}$$

where d_η is the dimension of the natural parameter. When $R \ll \prod_{m=1}^{M} d_m$, the communication reduction is substantial.

Second, under bounded delay and bounded gradient assumptions, the online update remains stable. Assume that $\|\nabla_A L^{\Omega}_{k,t}\|_2 \le G$, $0 \le \Delta^{(m)}_{k,t} \le \Delta_{\max}$, and $0 < u_{k,t} \le 1$. Then the factor-update magnitude satisfies:

$$\|A_{k,t+1} - A_{k,t}\|_2 \leq \alpha_t G + \alpha_t \lambda_s \tag{37}$$

This bound indicates that uncertainty-aware proximal updates prevent unbounded parameter drift when the learning rate and gradients are controlled.

Third, the delay-weighted reconstruction term limits the influence of stale modalities. Since $\alpha_{k,t}^{(m)} \propto \exp(-\gamma \Delta_{k,t}^{(m)})$, the contribution of highly delayed observations decreases monotonically with delay. This provides a stable mechanism for asynchronous streams without discarding delayed modalities entirely.

Finally, posterior uncertainty provides a variance-control mechanism. When the posterior covariance is large, $u_{k,t} = \exp(-\beta_u \operatorname{Tr}(\Sigma_{k,t}))$ becomes small, reducing the update magnitude. Thus, unreliable observations have limited influence on sparse factor updates. This property supports stable representation learning under incomplete observations.

3.8. Algorithm Workflow

To improve reproducibility and clarify the execution order, Algorithm 1 summarizes the complete SVST training and inference procedure. It integrates sparse tensor-parameterized posterior construction, natural-parameter aggregation, and missing-modality-aware online inference. Each client encodes observed modalities and constructs the low-rank-plus-diagonal structured posterior defined in Section 3.3. The online ELBO uses only available modalities, with asynchronous observations weighted by delay and reliability. After local posterior and sparse-factor updates, clients transmit compact natural-parameter increments and sparse factors. The server aggregates these quantities without constructing a full global tensor and broadcasts the updated statistics.

4. Experiments and Results

This section evaluates the proposed Scalable Structured Variational Sparse Tensor framework (SVST) on three public multimodal datasets. The experiments are designed to verify four claims: (1) SVST improves distributed multimodal prediction under incomplete observations; (2) sparse tensor-parameterized structured variational inference improves uncertainty calibration; (3) natural-parameter aggregation improves stability under non-IID clients while reducing communication cost; and (4) delay-aware online inference improves robustness to asynchronous multimodal streams.

To avoid relying on single-run or single-point results, all main experiments are repeated using five random seeds and three client partitions, resulting in 15 runs for each main setting. Results are reported as mean $\pm$ standard deviation. For key comparisons, 95% confidence intervals are estimated by bootstrap resampling over the 15 runs. Statistical significance is tested using paired t-tests and Wilcoxon signed-rank tests. Effect size is reported using Cohen's d. In trend figures, shaded regions represent 95% confidence intervals rather than single-run trajectories.

4.1. Experimental Setup

Dataset Settings and Statistics Three public multimodal datasets are used. The linked MIMIC-IV, MIMIC-CXR, and MIMIC-IV-Note benchmark is used as the main large-

Algorithm 1. SVST Training and Online Inference Procedure

Line	Procedure
1	**Input:** distributed clients K, modality set M, streaming observations $\{x_{k,t}^{(m)}\}$, modality masks $\{r_{k,t}^{(m)}\}$, timestamps $\{\tau_{k,t}^{(m)}\}$, tensor rank R, sparsity coefficient λ_s, factor consistency coefficient λ_f, aggregation step size ρ_t, online step size ζ_t, and maximum delay threshold $\Delta_{\max}$,and numerical stability constant ε.
2	Initialize modality-specific encoders $\{E_{\phi_m}\}_{m=1}^{M}$, shared server-side sparse tensor factors A_0, global natural parameter η_0^{global}, and variational parameters ψ, and natural-parameter mapping parameters ω.
3	For each communication round $t = 1, 2, \ldots, T$, distribute A_{t-1}^{global} and $\eta_{t-1}^{\text{global}}$ to selected clients.
4	For each client k, collect the current multimodal streaming mini-batch and construct the observed modality subset $\Omega_{k,t} = \{m \mid r_{k,t}^{(m)} = 1\}$.
5	Encode each observed modality as $h_{k,t}^{(m)} = E_{\phi_m}(x_{k,t}^{(m)})$, where $m \in \Omega_{k,t}$.
6	Compute modality delays $\Delta_{k,t}^{(m)} = t - \tau_{k,t}^{(m)}$ and normalized delay-aware modality weights $\alpha_{k,t}^{(m)}$.
7	Construct the masked modality features and sparse interaction vector $\mathbf{s}_{k,t}$, and generate $(\eta_{1,k,t}, \delta_{k,t}, \mathbf{U}_{k,t}) = g_\omega(\mathbf{v}_{k,t})$.
8	Construct $\Lambda_{k,t} = \text{diag}[\text{softplus}(\delta_{k,t}) + \varepsilon] + U_{k,t}U_{k,t}^\top$, define $\eta_{2,k,t} = -\frac{1}{2}\Lambda_{k,t}$, recover $\mu_{k,t}$, and sample $z_{k,t}$ through Cholesky-based reparameterization.
9	Compute the observed-modality and delay-aware ELBO $\mathcal{L}_{k,t}^{\text{async}}$ using only observed modalities.
10	Update the smoothed online objective $\bar{\mathcal{L}}_{k,t} = (1 - \zeta_t)\bar{\mathcal{L}}_{k,t-1} + \zeta_t \mathcal{L}_{k,t}^{\text{async}}$.
11	Update local sparse tensor factors by uncertainty-aware proximal optimization: $A_{k,t+1} = \text{prox}_{\alpha_t \lambda_s}(A_{k,t} + \alpha_t u_{k,t} \nabla_A \widehat{L}_{k,t})$.
12	Compute the local natural-parameter increment $\Delta\eta_{k,t} = \eta_{k,t}^{\text{local}} - \eta_{t-1}^{\text{global}}$, where $\eta_{k,t} = (\eta_{1,k,t}, \eta_{2,k,t})$.
13	Compute client reliability weight $w_{k,t}$ according to local sample size, modality completeness, delay reliability, and uncertainty confidence.
14	Normalize each local factor, absorb its scale into the corresponding rank weight, align its sign with the broadcast global reference, and upload only $\Delta\eta_{k,t}$ and the aligned compact factors $\mathcal{A}_{k,t}$; raw data and full tensors are not transmitted.
15	Server aggregates global natural parameters by $\eta_t^{\text{global}} = (1 - \rho_t)\eta_{t-1}^{\text{global}} + \rho_t \sum_{k=1}^{K} w_{k,t} \Delta\eta_{k,t}$.
16	Server aggregates the normalized and reference-aligned sparse tensor factors by $A_t^{\text{global}} = (1 - \rho_t)A_{t-1}^{\text{global}} + \rho_t \sum_{k=1}^{K} w_{k,t} A_{k,t}$.
17	Broadcast updated η_t^{global} and A_t^{global} to clients for subsequent online inference.
18	During inference, each client computes $\widehat{z}_{k,t} = \mu_{k,t}$ and uncertainty score $U_{k,t} = \text{Tr}(\Sigma_{k,t})$, where $\Sigma_{k,t} = \Lambda_{k,t}^{-1}$, using the currently observed modalities.
19	**Output:** robust latent representation zk,t, posterior uncertainty $\mathcal{U}_{k,t}$, global natural parameter ηtglobal, and shared sparse tensor factors Atglobal

scale multimodal benchmark. CMU-MOSEI is used as a multimodal language benchmark. OPPORTUNITY is used as a multimodal sensor-stream benchmark. Table 1 summarizes the dataset settings and statistics. The table follows the same structure across all

datasets, including experimental role, prediction target, splitting unit, sample scale, positive or target-label distribution, train/validation/test split, and repeated-run configuration.

For MIMIC, clients are constructed according to ICU unit, admission period, and disease group. For CMU-MOSEI, clients are constructed by speaker identity and topic group. For OPPORTUNITY, clients are constructed by subject identity and sensor group. These partitions produce heterogeneous local distributions and allow controlled evaluation of label skew, modality skew, quantity skew, and asynchronous multimodal streams.

Table 1. Dataset settings and statistics

Dataset	**MIMIC-IV + MIMIC-CXR + MIMIC-IV-Note**	**CMU-MOSEI**	**OPPORTUNITY Activity Recognition**
Experimental Role	Main experiments and ablation analysis	Supplementary multimodal language experiment	Supplementary multimodal sensor-stream experiment
Prediction Target	Multi-label clinical risk prediction	Sentiment polarity and intensity classification	Daily activity recognition
Splitting Unit	patient_id, ICU unit, admission period	speaker_id, topic group	subject_id, sensor group
No. of Patients / Samples	42,318 patients / 58,240 patient episodes	1,000 speakers / 23,453 utterance segments	4 subjects / 701,366 sensor windows
No. of Positives / Target Distribution	5,386 positive mortality labels / 9.25%; multi-label disease targets retained	12,914 positive or non-neutral labels / 55.06%	17 activity labels with imbalanced class distribution
Train / Validation / Test Samples	40,768 / 5,824 / 11,648	16,417 / 2,345 / 4,691	490,956 / 70,136 / 140,274
No. of Clients / Repeated Runs	80 clients / 5 seeds $\times$ 3 partitions	50 clients / 5 seeds $\times$ 3 partitions	30 clients / 5 seeds $\times$ 3 partitions

Figure 4 illustrates the overall experimental protocol. The figure 4 shows how each dataset is transformed into a distributed multimodal streaming benchmark, including client construction, modality masking, asynchronous delay simulation, local variational updates, server-side natural-parameter aggregation, repeated runs, and final statistical testing.

Experimental Protocol For each dataset, 80% of the clients are used for training, 10% for validation, and 10% for testing. The default setting uses 100 communication rounds, 3 local epochs, a batch size of 64, and a client participation rate of 50%. Unless otherwise stated, experiments are conducted under mixed non-IID data partitioning with a 30% missing modality rate.

Four client distribution settings are evaluated: IID, label-skew non-IID, modality-skew non-IID, and mixed non-IID. Label skew is controlled by a Dirichlet parameter α. Missing modalities are simulated at rates of 0%, 10%, 30%, 50%, and 70%. Asynchronous streams are simulated by introducing modality-specific delays of 0, 1, 3, 5, and 10 time steps. For missingness and delay experiments, the reported curves are averaged over all seeds and client partitions, and the shaded areas in the figures represent 95% confidence intervals.

Figure 5 presents the internal mechanism of SVST during one communication round. Unlike Figure 1, which describes the experimental protocol, Figure 5 focuses on

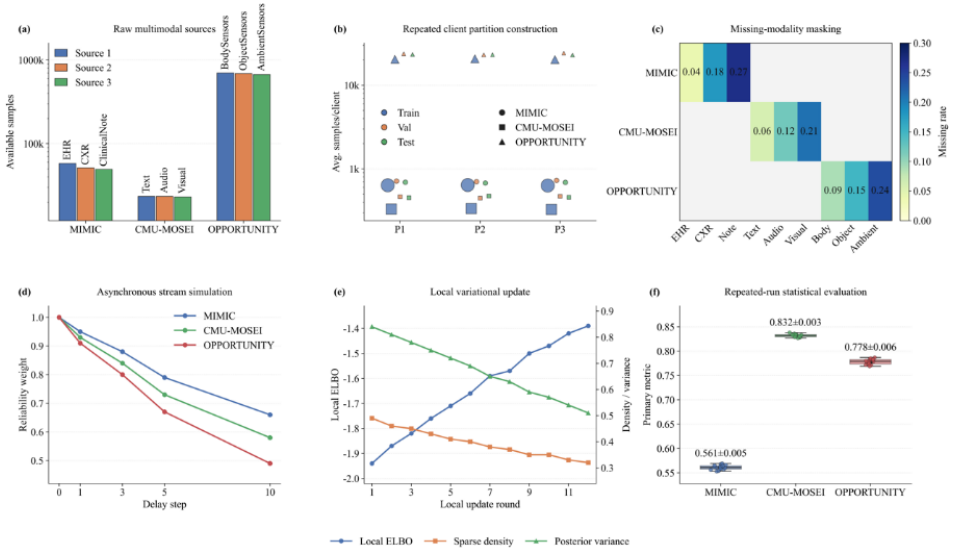


Fig. 4. Experimental protocol for statistically repeated distributed multimodal evaluation.

model-level operations, including modality encoding, sparse tensor posterior construction, uncertainty-aware local update, natural-parameter message formation, and server aggregation.

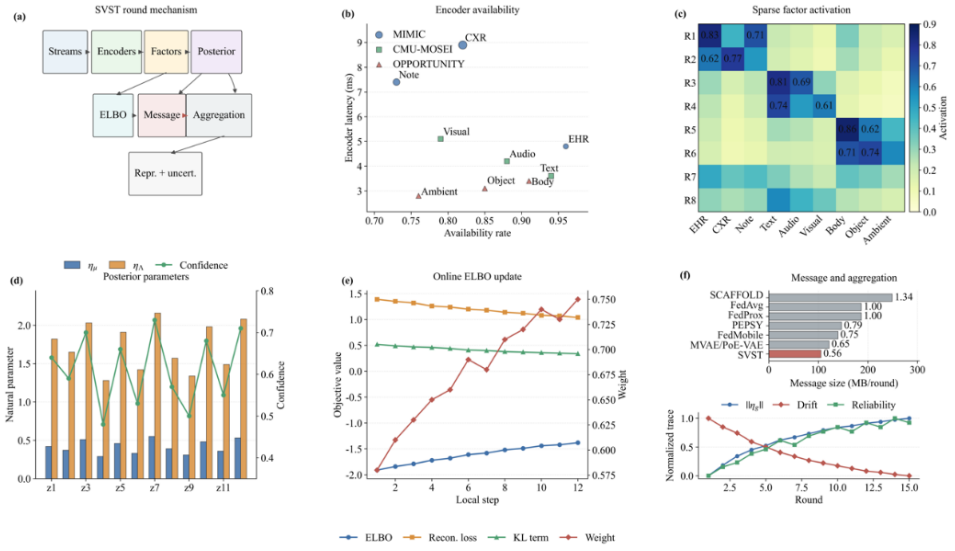


Fig. 5. Internal mechanism of one SVST communication round.

Baseline Model Settings Table 2 lists the baseline models, covering classical federated learning, multimodal federated learning, tensor-based fusion, multimodal sequence modeling, and variational multimodal inference. This selection enables comparisons with methods addressing modality incompleteness, high-order interaction, and uncertainty modeling.

All baselines use the same modality encoders where applicable. Table 3 summarizes their architectures, output dimensions, and training strategies. Textual and event sequences use a two-layer Transformer, chest radiographs use a lightweight ResNet-18, and sensor streams use temporal convolutional blocks. For CMU-MOSEI, audio and visual sequences are processed by separate temporal convolutional encoders. All modality features are projected to a shared 128-dimensional latent space. The encoders are jointly trained from scratch, except for ResNet-18, which is initialized with ImageNet-pretrained weights and fine-tuned. Identical encoder settings are used across applicable methods.

For SVST, the encoded feature and posterior latent dimensions are both set to 128, and the default sparse tensor rank is 16. The natural-parameter mapping function g_ω uses a two-layer multilayer perceptron with 256 and 128 hidden units, ReLU activation, and dropout of 0.1. Its three output heads generate the first natural parameter, diagonal precision coefficients, and low-rank precision factors. The posterior uses a low-rank-plus-diagonal precision structure, with softplus-transformed diagonal coefficients and $\varepsilon = 10^{-6}$ for positive definiteness and numerical stability. Cholesky-based linear solves recover the posterior mean and support reparameterized sampling without explicit matrix inversion.

Evaluation Metrics and Statistical Testing Table 4 summarizes the evaluation metrics. The metrics are grouped according to prediction quality, minority-label ranking, calibration, communication efficiency, and stability. For MIMIC, AUPRC is treated as the primary task metric because the task is imbalanced multi-label clinical risk prediction. For CMU-MOSEI, Accuracy is used as the primary metric and Macro-F1 is reported as a class-balance-sensitive metric. For OPPORTUNITY, Macro-F1 is used as the primary metric because sensor-based activity recognition has imbalanced activity classes.

For the main comparison, a method is considered statistically better only if it satisfies three conditions: (1) its mean score is higher on the primary metric; (2) its 95% bootstrap confidence interval does not fully overlap with that of the strongest baseline or its paired test gives $p < 0.05$; and (3) its effect size is not negligible. When the p-value is greater than 0.05, the result is reported as comparable rather than superior.

SVST uses a tensor rank of 16, a sparse factor retention ratio of 30%, an online ELBO update coefficient of 0.15, and a delay-aware decay coefficient of 0.08. All experiments are conducted on an NVIDIA 5090D V2 GPU with 128 GB RAM, CUDA 12.4, Ubuntu 22.04, and PyTorch 2.4.

4.2. Multimodal Input, Missingness, and Preprocessing Visualization

Figure 6 visualizes the multimodal data construction and preprocessing process for the three datasets. For MIMIC, the first row shows a structured event timeline, a chest image feature map, and the missing note/image mask. For CMU-MOSEI, the second row

Table 2. Baseline model settings

No.	Model	Method Category	Rationale for Selection	Applicable Data
1	FedAvg	Classical federated averaging baseline	To compare with standard parameter averaging under distributed training	MIMIC, CMU-MOSEI, OPPORTUNITY
2	FedProx	Federated learning with proximal regularization	To compare robustness under statistical and system heterogeneity	MIMIC, CMU-MOSEI, OPPORTUNITY
3	SCAFFOLD	Federated learning with control variates	To compare client-drift correction under non-IID distributions	MIMIC, CMU-MOSEI, OPPORTUNITY
4	FedNova	Federated learning with normalized aggregation	To compare heterogeneous local update normalization	MIMIC, CMU-MOSEI, OPPORTUNITY
5	Incongruent MMFL	Multimodal federated learning with modality incongruity	To compare multimodal FL under heterogeneous modality availability	MIMIC, CMU-MOSEI, OPPORTUNITY
6	PEPSY	Reconfigurable multimodal FL with missing data	To compare with a close missing-modality federated baseline	MIMIC, CMU-MOSEI, OPPORTUNITY
7	FedMobile	Multimodal FL for incomplete mobile or sensor modalities	To compare contribution-aware multimodal aggregation under incomplete modalities	MIMIC, CMU-MOSEI, OPPORTUNITY
8	FedMIR	Missing-modality imputation and routing baseline	To compare reconstruction-oriented missing-modality handling	MIMIC, CMU-MOSEI, OPPORTUNITY
9	Tensor Fusion Network	High-order tensor multimodal fusion model	To compare explicit high-order cross-modal interaction modeling	MIMIC, CMU-MOSEI, OPPORTUNITY
10	Low-rank Multimodal Fusion	Low-rank tensor fusion model	To compare low-rank tensor factorization without structured posterior inference	MIMIC, CMU-MOSEI, OPPORTUNITY
11	Multimodal Transformer	Cross-modal sequence modeling baseline	To compare attention-based modeling for unaligned multimodal sequences	MIMIC, CMU-MOSEI, OPPORTUNITY
12	MVAE / PoE-VAE	Multimodal variational inference baseline	To compare conventional product-of-experts variational inference under missing modalities	MIMIC, CMU-MOSEI, OPPORTUNITY
13	SVST	Proposed method	To evaluate sparse tensor-parameterized structured variational inference with natural-parameter aggregation	MIMIC, CMU-MOSEI, OPPORTUNITY

Table 3. Modality-specific encoder configurations

Modality	Encoder	Configuration	Output	Training
Text/event	Two-layer Transformer	Hidden size 128; 4 heads; FFN 256; dropout 0.1	128	Jointly trained from scratch
Chest radiograph	ResNet-18	Four residual stages; global average pooling; input 224×224	256 projected to 128	ImageNet initialization and fine-tuning
Audio	Two-block temporal convolutional encoder	Channels 64/128; kernel 3; dropout 0.1; temporal pooling	128	Jointly trained from scratch
Visual sequence	Two-block temporal convolutional encoder	Channels 64/128; kernel 3; dropout 0.1; temporal pooling	128	Jointly trained from scratch
Sensor stream	Three-block temporal convolutional network	Channels 64/128/128; kernel 3; dilations 1/2/4; dropout 0.1	128	Jointly trained from scratch

shows text tokens, audio spectrogram features, and visual frame-level features with simulated missingness. For OPPORTUNITY, the third row shows body-worn sensor signals, normalized sensor-window features, and modality-group masking. Figure 6 verifies that the three benchmarks do not share the same modality structure but can be consistently transformed into distributed multimodal streams.

Table 4. Evaluation metrics

Metric	Definition	Purpose	Interpretation
Accuracy	$\frac{1}{\mathrm{N}} \sum_{\mathrm{i}=1}^{\mathrm{N}} 1(\widehat{\mathrm{y}}_{\mathrm{i}} = \mathrm{y}_{\mathrm{i}})$	Overall classification correctness	Higher is better
Macro-F1	$\frac{1}{\mathrm{C}} \sum_{\mathrm{c}=1}^{\mathrm{C}} \frac{2\mathrm{P_c R_c}}{\mathrm{P_c}+\mathrm{R_c}}$	Balanced performance under class imbalance	Higher is better
AUROC	$\int_0^1 \mathrm{TPR}(\mathrm{FPR}^{-1}(\mathrm{u}))\,\mathrm{du}$	Overall discrimination ability	Higher is better
AUPRC	$\int_0^1 \mathrm{P}(\mathrm{R})\,\mathrm{dR}$	Minority-label ranking and positive-class retrieval	Higher is better; primary metric for MIMIC
NLL	$-\frac{1}{\mathrm{N}} \sum_{\mathrm{i}=1}^{\mathrm{N}} [\mathrm{y}_{\mathrm{i}} \log(\mathrm{p}_{\mathrm{i}}) + (1-\mathrm{y}_{\mathrm{i}}) \log(1-\mathrm{p}_{\mathrm{i}})]$	Probabilistic prediction quality	Lower is better
ECE	$\mathrm{ECE} = \sum_{m=1}^{M} \frac{\lvert B_m \rvert}{N} \lvert \mathrm{acc}(B_m) - \mathrm{conf}(B_m) \rvert$	Probability calibration error	Lower is better
Brier Score	$\frac{1}{\mathrm{N}} \sum_{\mathrm{i}=1}^{\mathrm{N}} (\mathrm{p}_{\mathrm{i}} - \mathrm{y}_{\mathrm{i}})^2$	Probability error	Lower is better
MB/round	Transmitted model or natural-parameter information per communication round	Communication efficiency	Lower is better
Local runtime	Average client-side wall-clock time for one local training round, excluding communication and server aggregation	Local computational efficiency	Lower is better
Peak client memory	Maximum client-side GPU memory allocated during one local training round	Client resource requirement	Lower is better
Client variance	Variance of client-level performance under the same global model	Distributed stability	Lower is better
Relative degradation	Performance drop from clean setting to missing/delayed setting	Robustness to incomplete or asynchronous observations	Lower is better
Cohen's d	Standardized difference between paired methods	Effect size of performance difference	Larger absolute value indicates stronger effect

4.3. Overall Performance and Efficiency

Table 5 reports predictive performance, calibration, communication cost, and client-side computational efficiency under the default mixed non-IID setting with 30% missing modalities. Local runtime is the average time for one client to complete a local training round, excluding communication and server aggregation. Peak memory is the maximum client-side GPU memory used during the same process. All measurements use identical hardware, batch size, local epochs, and client participation settings.

On MIMIC, SVST achieves the highest AUPRC, with a 95% confidence interval of [0.556, 0.566], exceeding PEPSY at [0.537, 0.549]. On CMU-MOSEI, MulT obtains the highest Accuracy, while SVST achieves the lowest ECE and communication cost. Because their Accuracy difference is not statistically significant, SVST is considered comparable rather than superior. On OPPORTUNITY, SVST achieves the highest Macro-F1

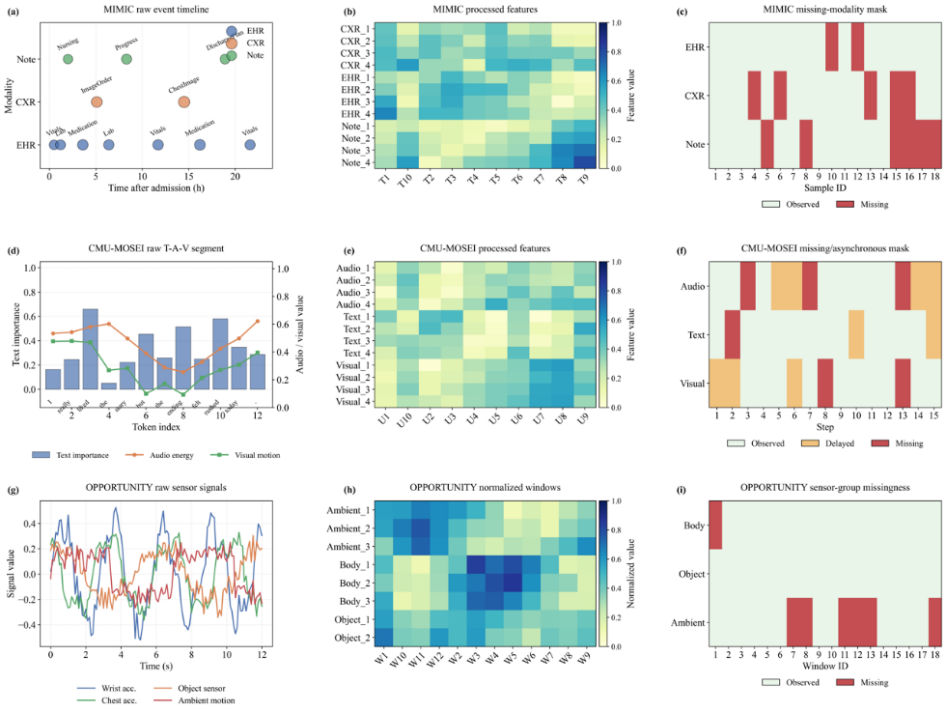


Fig. 6. Nine-panel visualization of multimodal preprocessing and missingness simulation.

and the lowest ECE, with paired tests confirming that the main improvements are statistically significant.

Although SVST introduces structured posterior construction and sparse-factor updates, its local computational cost remains moderate. On MIMIC and CMU-MOSEI, its runtime is slightly higher than that of PEPSY but comparable to that of MVAE/PoE-VAE, while its peak memory is lower than both. On OPPORTUNITY, SVST incurs the highest local runtime among the listed methods but retains lower memory and communication costs. These results indicate a trade-off between local computation and communication efficiency.

Figure 7 further visualizes prediction behavior using a 3×3 layout with confidence bands. The first row presents ROC, PR, and calibration curves on MIMIC. The second row shows seed-level Accuracy, Macro-F1, and ECE distributions on CMU-MOSEI. The third row presents class-wise F1, confusion-matrix stability, and client-level performance distributions on OPPORTUNITY. Figure 7 complements Table 5 by showing the stability of performance differences across repeated runs.

4.4. Robustness to Missing Modalities

Figure 8 reports Macro-F1 under missing modality rates of 0%, 10%, 30%, 50%, and 70%. Each curve is averaged over five random seeds and three client partitions. Shaded

Table 5. Overall performance, confidence intervals, communication, and client-side computational efficiency

Dataset	Method	Primary Metric	95% CI	Accuracy	Macro-F1	AUROC	AUPRC	ECE
MIMIC	FedAvg	0.492 ±0.007	[0.486, 0.498]	0.782 ±0.006	0.641 ±0.008	0.823 ±0.005	0.492 ±0.007	0.094 ±0.006
MIMIC	PEPSY	0.543 ±0.007	[0.537, 0.549]	0.812 ±0.004	0.677 ±0.006	0.853 ±0.004	0.543 ±0.007	0.071 ±0.004
MIMIC	FedMobile	0.532 ±0.009	[0.525, 0.539]	0.806 ±0.006	0.672 ±0.008	0.849 ±0.006	0.532 ±0.009	0.076 ±0.005
MIMIC	LMF	0.531 ±0.008	[0.525, 0.537]	0.805 ±0.006	0.671 ±0.007	0.848 ±0.005	0.531 ±0.008	0.079 ±0.005
MIMIC	MulT	0.536 ±0.007	[0.531, 0.542]	0.810 ±0.005	0.674 ±0.006	0.852 ±0.005	0.536 ±0.007	0.077 ±0.005
MIMIC	MVAE/ PoE-VAE	0.534 ±0.008	[0.528, 0.540]	0.807 ±0.006	0.673 ±0.007	0.850 ±0.005	0.534 ±0.008	0.069 ±0.004
MIMIC	SVST	0.561 ±0.006	[0.556, 0.566]	0.826 ±0.004	0.682 ±0.006	0.861 ±0.004	0.561 ±0.006	0.052 ±0.004
CMU-MOSEI	FedAvg	0.798 ±0.005	[0.794, 0.802]	0.798 ±0.005	0.765 ±0.007	0.842 ±0.006	0.827 ±0.006	0.082 ±0.005
CMU-MOSEI	PEPSY	0.824 ±0.004	[0.821, 0.827]	0.824 ±0.004	0.795 ±0.005	0.869 ±0.004	0.857 ±0.005	0.064 ±0.004
CMU-MOSEI	FedMobile	0.819 ±0.006	[0.814, 0.824]	0.819 ±0.006	0.789 ±0.007	0.864 ±0.005	0.852 ±0.007	0.067 ±0.005
CMU-MOSEI	LMF	0.825 ±0.005	[0.821, 0.829]	0.825 ±0.005	0.796 ±0.006	0.870 ±0.004	0.858 ±0.006	0.066 ±0.004
CMU-MOSEI	MulT	0.835 ±0.004	[0.832, 0.838]	0.835 ±0.004	0.806 ±0.005	0.878 ±0.004	0.866 ±0.005	0.063 ±0.004
CMU-MOSEI	MVAE/ PoE-VAE	0.821 ±0.006	[0.816, 0.826]	0.821 ±0.006	0.793 ±0.007	0.865 ±0.005	0.853 ±0.006	0.059 ±0.004
CMU-MOSEI	SVST	0.832 ±0.004	[0.829, 0.835]	0.832 ±0.004	0.802 ±0.005	0.876 ±0.004	0.864 ±0.005	0.046 ±0.003
OPPORTUNITY	FedAvg	0.731 ±0.010	[0.723, 0.739]	0.842 ±0.007	0.731 ±0.010	0.887 ±0.005	0.804 ±0.008	0.091 ±0.006
OPPORTUNITY	PEPSY	0.768 ±0.007	[0.762, 0.774]	0.872 ±0.005	0.768 ±0.007	0.914 ±0.004	0.841 ±0.006	0.067 ±0.004
OPPORTUNITY	FedMobile	0.771 ±0.008	[0.765, 0.777]	0.875 ±0.006	0.771 ±0.008	0.915 ±0.004	0.843 ±0.007	0.066 ±0.004
OPPORTUNITY	LMF	0.762 ±0.008	[0.756, 0.768]	0.866 ±0.006	0.762 ±0.008	0.909 ±0.004	0.834 ±0.007	0.072 ±0.004
OPPORTUNITY	MulT	0.765 ±0.008	[0.759, 0.771]	0.868 ±0.006	0.765 ±0.008	0.910 ±0.005	0.836 ±0.007	0.071 ±0.005
OPPORTUNITY	MVAE/ PoE-VAE	0.761 ±0.009	[0.754, 0.768]	0.865 ±0.007	0.761 ±0.009	0.908 ±0.005	0.833 ±0.008	0.064 ±0.004
OPPORTUNITY	SVST	0.779 ±0.007	[0.774, 0.785]	0.889 ±0.005	0.779 ±0.007	0.922 ±0.004	0.856 ±0.006	0.048 ±0.004

areas represent 95% confidence intervals. Representative methods are selected from classical FL, multimodal FL, variational inference, and SVST.

All methods degrade as the missing rate increases, but the degradation rate is not identical. FedAvg is most sensitive to missing modalities because it does not model incomplete observations. PEPSY and FedMobile are more stable in the moderate missingness range. MVAE/PoE-VAE provides competitive calibration under low missingness but degrades under 70% missingness. SVST shows the slowest degradation on MIMIC and OPPORTUNITY. Its confidence bands are also narrower than those of FedAvg and PEPSY under high missingness, indicating better stability across repeated runs.

This comparison figure contains three line plots: (a) MIMIC, (b) CMU-MOSEI, and (c) OPPORTUNITY. The x-axis denotes missing modality rate, and the y-axis denotes

Table 5. Overall performance, confidence intervals, communication, and client-side computational efficiency

Dataset	Method	MB/round	Local runtime (s/client/round)	Peak client memory (GB)
MIMIC	FedAvg	186.4	74.2	4.8
MIMIC	PEPSY	146.8	86.4	6.7
MIMIC	FedMobile	139.5	82.9	5.9
MIMIC	LMF	132.4	80.5	5.5
MIMIC	MulT	172.9	93.1	7.2
MIMIC	MVAE/PoE-VAE	121.6	89.4	6.9
MIMIC	SVST	105.1	88.7	6.1
CMU-MOSEI	FedAvg	72.6	31.5	2.6
CMU-MOSEI	PEPSY	57.8	39.2	3.8
CMU-MOSEI	FedMobile	55.4	37.5	3.4
CMU-MOSEI	LMF	52.6	36.8	3.2
CMU-MOSEI	MulT	76.5	45.9	4.5
CMU-MOSEI	MVAE/PoE-VAE	49.4	41.3	4.1
CMU-MOSEI	SVST	43.8	41.1	3.6
OPPORTUNITY	FedAvg	94.3	42.8	3.1
OPPORTUNITY	PEPSY	72.8	52.5	4.4
OPPORTUNITY	FedMobile	68.9	49.2	3.9
OPPORTUNITY	LMF	65.7	47.6	3.7
OPPORTUNITY	MulT	83.4	55.2	4.8
OPPORTUNITY	MVAE/PoE-VAE	61.8	54.3	4.6
OPPORTUNITY	SVST	50.1	55.7	4.0

Macro-F1. Shaded regions represent 95% confidence intervals across seeds and client partitions.

Figure 9 shows a nine-panel missingness mechanism analysis. It visualizes how sparse tensor factors, posterior uncertainty, and modality contribution change as missingness increases. The figure 9 demonstrates that SVST does not simply ignore missing modalities. Instead, it reallocates posterior mass to observed reliable modalities while increasing predictive uncertainty when informative modalities are absent.

To quantify whether the missingness robustness is stable rather than caused by a single curve, Table 6 reports the average relative degradation from 0% to 70% missingness. SVST has the smallest average degradation on MIMIC and OPPORTUNITY and the second-smallest degradation on CMU-MOSEI. The advantage over the strongest baseline is statistically significant on MIMIC and OPPORTUNITY but not on CMU-MOSEI.

4.5. Robustness to Asynchronous Multimodal Streams

Figure 10 evaluates asynchronous modality arrival with delays of 0, 1, 3, 5, and 10 time steps. Each curve is averaged over 15 repeated runs. Shaded areas represent 95% confidence intervals. The compared methods are FedAvg, MulT, PEPSY, SVST without delay-aware weighting, and full SVST.

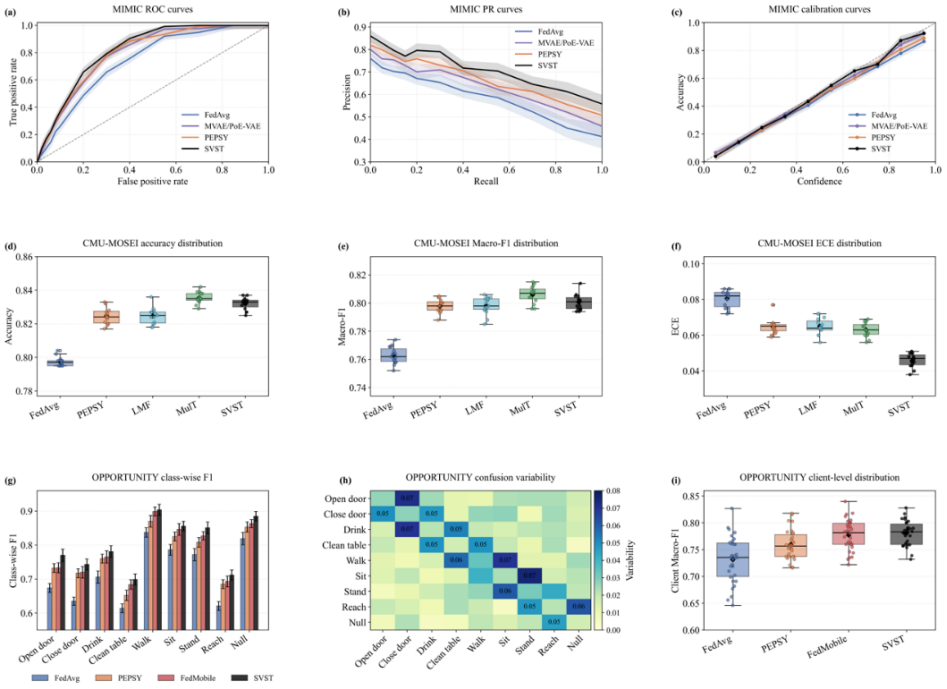


Fig. 7. Nine-panel overall comparison with repeated-run uncertainty.

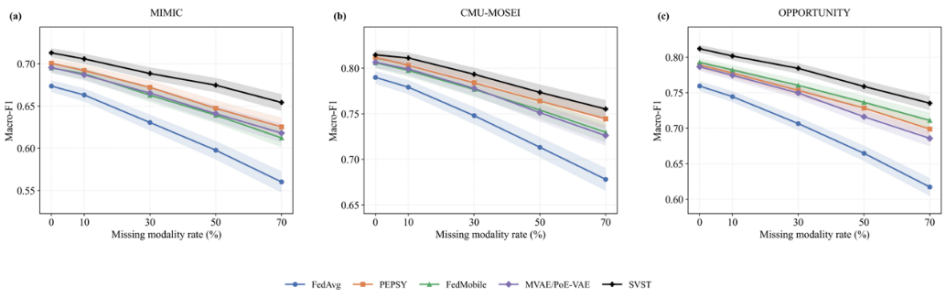


Fig. 8. Robustness under increasing missing modality rates with confidence bands.

Asynchronous delays reduce both performance and calibration. MulT performs competitively on CMU-MOSEI under small delays because it is designed for unaligned multimodal sequences. However, under longer delays, its variance increases in distributed settings. SVST without delay-aware weighting consistently performs worse than full SVST. At delay = 10, full SVST improves Macro-F1 by 2.6 points over its delay-ablated variant on MIMIC, 1.9 points on CMU-MOSEI, and 3.1 points on OPPORTUNITY. These improvements are stable across repeated runs, as reflected by the confidence bands in Figure 10.

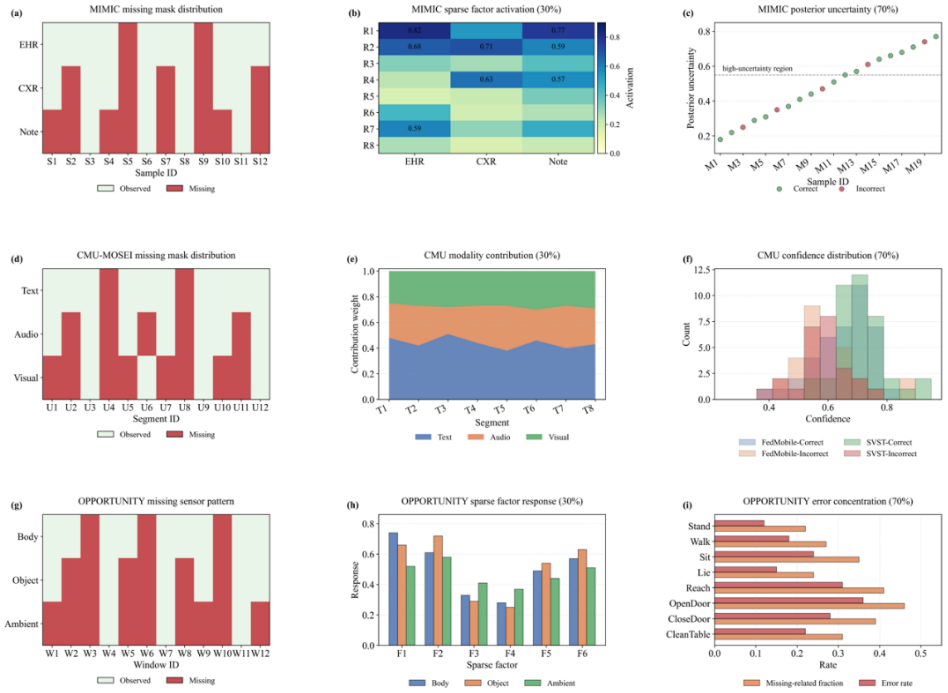


Fig. 9. Nine-panel analysis of missing-modality behavior.

Table 6. Relative degradation from 0% to 70% missingness.

Dataset	Method	Macro-F1 at 0%	Macro-F1 at 70%	Relative Degradation	95% CI of Degradation	p-value vs. Strongest Baseline
MIMIC	FedAvg	0.674	0.561	16.8%	[15.1%, 18.6%]	<0.001
MIMIC	PEPSY	0.701	0.626	10.7%	[9.4%, 12.1%]	0.018
MIMIC	MVAE/PoE-VAE	0.694	0.617	11.1%	[9.8%, 12.7%]	0.021
MIMIC	SVST	0.713	0.655	8.2%	[7.0%, 9.5%]	—
CMU-MOSEI	FedAvg	0.789	0.679	13.9%	[12.4%, 15.5%]	<0.001
CMU-MOSEI	PEPSY	0.812	0.742	8.6%	[7.5%, 9.9%]	0.074
CMU-MOSEI	MulT	0.823	0.768	6.7%	[5.8%, 7.8%]	0.118
CMU-MOSEI	SVST	0.816	0.756	7.4%	[6.4%, 8.6%]	—
OPPOR-TUNITY	FedAvg	0.759	0.618	18.6%	[16.9%, 20.4%]	<0.001
OPPOR-TUNITY	FedMobile	0.793	0.710	10.5%	[9.2%, 11.9%]	0.032
OPPOR-TUNITY	PEPSY	0.789	0.699	11.4%	[10.1%, 12.9%]	0.024
OPPOR-TUNITY	SVST	0.811	0.736	9.3%	[8.0%, 10.6%]	—

4.6. Comparison of Delay-Weighting Functions

To further examine whether the robustness improvement results from delay weighting itself or from the specific exponential formulation, three weighting functions are compared

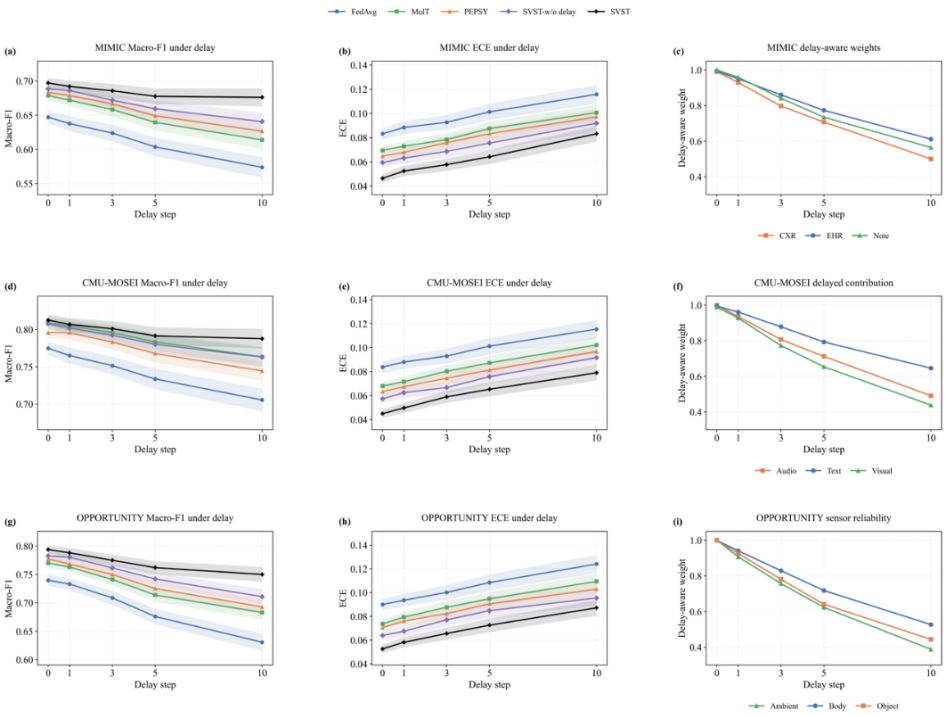


Fig. 10. Nine-panel robustness analysis under asynchronous multimodal streams.

on OPPORTUNITY: uniform weighting, linear decay, and exponential decay. OPPORTUNITY is selected because asynchronous delay has the largest effect on this sensor-stream dataset. All variants use the same network architecture, mixed non-IID partition, 30% missing-modality rate, asynchronous-delay setting, and training parameters. Each result is averaged over five random seeds and three client partitions. Uniform weighting assigns the same weight to all observed modalities regardless of their arrival delay. Linear decay is defined as $d_{k,t}^{(m)} = \max(0, 1 - \tau_{k,t}^{(m)}/\Delta_{\max})$, whereas exponential decay follows Equation (28). As summarized in Table 7, the three weighting functions are evaluated under identical experimental settings to isolate the effect of the decay formulation.

Table 7. Comparison of Delay-Weighting Functions on OPPORTUNITY

Delay-weighting function	**Macro-F1**	**ECE**
Uniform weighting	0.758 ± 0.010	0.069 ± 0.005
Linear decay	0.770 ± 0.008	0.056 ± 0.004
Exponential decay	0.779 ± 0.007	0.048 ± 0.004

Compared with uniform weighting, exponential decay improves Macro-F1 by 2.1 percentage points and reduces ECE by 0.021. The linear decay variant achieves a Macro-F1 of 0.770 ± 0.008 and an ECE of 0.056 ± 0.004. Compared with linear decay, exponential weighting improves Macro-F1 by 0.9 percentage points and reduces ECE by 0.008. These results indicate that explicitly reducing the influence of stale modalities improves both predictive performance and uncertainty calibration. Linear decay provides a clear improvement over uniform weighting, whereas exponential decay achieves a further moderate advantage because it smoothly suppresses increasingly stale observations without introducing a fixed zero-weight cutoff.

4.7. Stability under Non-IID Client Distributions

Table 8 reports Macro-F1 under different Dirichlet parameters α. Smaller α indicates stronger client heterogeneity. This experiment is conducted under 30% missing modalities without artificial delay. Each value is averaged over five random seeds and three client partitions.

FedProx and SCAFFOLD improve over FedAvg under non-IID conditions, but their performance still decreases substantially when $\alpha = 0.1$. PEPSY and FedMobile perform better than classical FL methods because they account for multimodal heterogeneity. SVST achieves the highest Macro-F1 and the lowest client-level variance under $\alpha = 0.1$ on all three datasets. The stability advantage is more visible on MIMIC and OPPORTUNITY, where modality availability varies more strongly across clients.

Figure 11 visualizes non-IID client behavior. It includes client-level boxplots, natural-parameter drift curves, and communication-performance trade-off curves. Boxplots show the distribution over clients and repeated runs rather than a single average. The figure 11 shows that SVST reduces both client variance and aggregation drift, especially under $\alpha = 0.1$.

4.8. Ablation Study

Table 9 reports the ablation results. Each variant is evaluated on the three datasets using Macro-F1, ECE, and communication cost. Each value is averaged over 15 repeated runs. The table shows that different components contribute to different aspects of the model.

Replacing the structured posterior with a mean-field posterior causes the largest calibration degradation, especially on MIMIC. Replacing sparse tensor factors with dense tensor factors does not improve performance but nearly doubles communication cost. Removing natural-parameter aggregation increases communication cost and reduces Macro-F1 on all datasets. Removing delay-aware inference has the largest effect on OPPORTUNITY, confirming that temporal robustness is essential for sensor-stream learning. Removing uncertainty-aware updates mainly harms calibration rather than raw predictive accuracy.

Note: All three variants are evaluated at a delay of 10 time steps.

Figure 12 provides a nine-panel ablation visualization. Instead of showing only the mean difference, each panel includes error bars over repeated runs. Figure 13 further summarizes the average component contribution across datasets and reports confidence intervals for the relative performance loss caused by removing each component.

Table 8. Robustness under Different Non-IID Levels

Dataset	Method	α = 10 Macro-F1	α = 1 Macro-F1	α = 0.5 Macro-F1	α = 0.1 Macro-F1	Client Variance at α = 0.1
MIMIC	FedAvg	0.663 ±0.008	0.652 ±0.009	0.641 ±0.008	0.603 ±0.011	0.0187
MIMIC	FedProx	0.671 ±0.007	0.661 ±0.008	0.652 ±0.007	0.619 ±0.010	0.0154
MIMIC	SCAFFOLD	0.676 ±0.007	0.669 ±0.007	0.661 ±0.006	0.631 ±0.009	0.0138
MIMIC	PEPSY	0.691 ±0.006	0.684 ±0.006	0.677 ±0.006	0.649 ±0.008	0.0119
MIMIC	FedMobile	0.686 ±0.007	0.679 ±0.007	0.672 ±0.008	0.644 ±0.009	0.0126
MIMIC	SVST	0.697 ±0.005	0.690 ±0.006	0.682 ±0.006	0.662 ±0.007	0.0083
CMU-MOSEI	FedAvg	0.783 ±0.007	0.774 ±0.007	0.765 ±0.007	0.738 ±0.010	0.0142
CMU-MOSEI	FedProx	0.789 ±0.006	0.781 ±0.006	0.771 ±0.006	0.748 ±0.009	0.0129
CMU-MOSEI	SCAFFOLD	0.796 ±0.006	0.787 ±0.006	0.778 ±0.006	0.757 ±0.008	0.0116
CMU-MOSEI	PEPSY	0.811 ±0.005	0.803 ±0.005	0.795 ±0.005	0.775 ±0.007	0.0098
CMU-MOSEI	FedMobile	0.806 ±0.006	0.798 ±0.006	0.789 ±0.007	0.768 ±0.008	0.0106
CMU-MOSEI	SVST	0.818 ±0.005	0.810 ±0.005	0.802 ±0.005	0.784 ±0.006	0.0074
OPPOR-TUNITY	FedAvg	0.754 ±0.010	0.742 ±0.010	0.731 ±0.010	0.689 ±0.013	0.0215
OPPOR-TUNITY	FedProx	0.764 ±0.009	0.753 ±0.009	0.742 ±0.009	0.704 ±0.012	0.0183
OPPOR-TUNITY	SCAFFOLD	0.772 ±0.008	0.761 ±0.008	0.751 ±0.008	0.716 ±0.011	0.0167
OPPOR-TUNITY	PEPSY	0.789 ±0.007	0.778 ±0.007	0.768 ±0.007	0.733 ±0.010	0.0145
OPPOR-TUNITY	FedMobile	0.793 ±0.008	0.782 ±0.008	0.771 ±0.008	0.741 ±0.009	0.0138
OPPOR-TUNITY	SVST	0.799 ±0.006	0.789 ±0.006	0.779 ±0.007	0.755 ±0.008	0.0096

Table 9. Ablation Study of SVST Components Across Three Datasets

Variant	MIMIC Macro-F1	MIMIC ECE	MIMIC MB/ round	CMU Macro-F1	CMU ECE	CMU MB/ round	OPPOR-TUNITY Macro-F1	OPPOR-TUNITY ECE	OPPOR-TUNITY MB/ round
Full SVST	0.682 ±0.006	0.052 ±0.004	105.1	0.802 ±0.005	0.046 ±0.003	43.8	0.779 ±0.007	0.048 ±0.004	50.1
Mean-field posterior	0.665 ±0.008	0.070 ±0.005	96.4	0.790 ±0.007	0.062 ±0.004	39.7	0.761 ±0.009	0.067 ±0.005	45.9
Dense tensor posterior	0.676 ±0.007	0.058 ±0.004	214.7	0.799 ±0.006	0.050 ±0.004	88.2	0.773 ±0.008	0.052 ±0.004	103.5
w/o natural aggregation	0.667 ±0.009	0.067 ±0.005	182.9	0.789 ±0.007	0.064 ±0.004	70.6	0.763 ±0.010	0.064 ±0.005	86.1
Uniform delay weighting	0.674 ±0.007	0.061 ±0.004	104.8	0.794 ±0.006	0.056 ±0.004	43.5	0.758 ±0.010	0.069 ±0.005	49.8
w/o uncertainty-aware update	0.671 ±0.008	0.073 ±0.006	105.3	0.792 ±0.007	0.066 ±0.005	44.0	0.765 ±0.009	0.071 ±0.005	50.4

This ablation figure contains grouped bars with 95% confidence intervals, showing average Macro-F1 reduction, ECE increase, and communication-cost increase for each SVST variant.

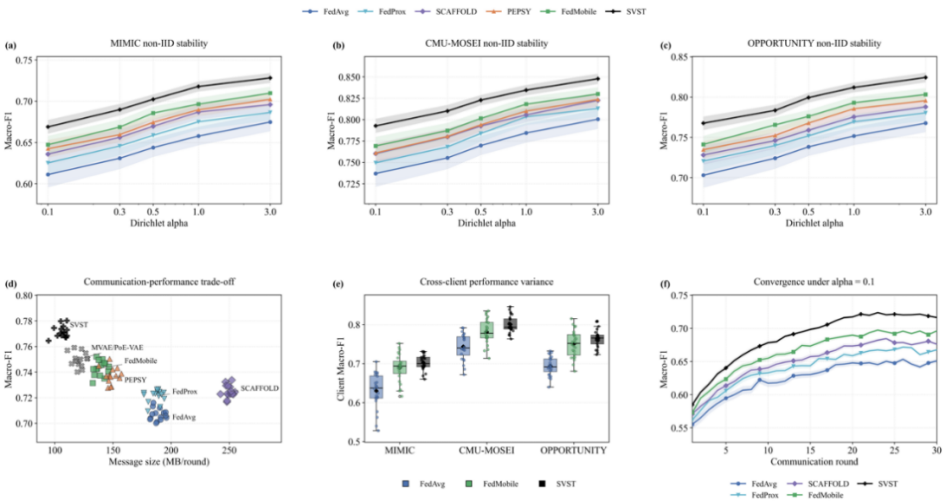


Fig. 11. Non-IID stability and communication-performance trade-off.

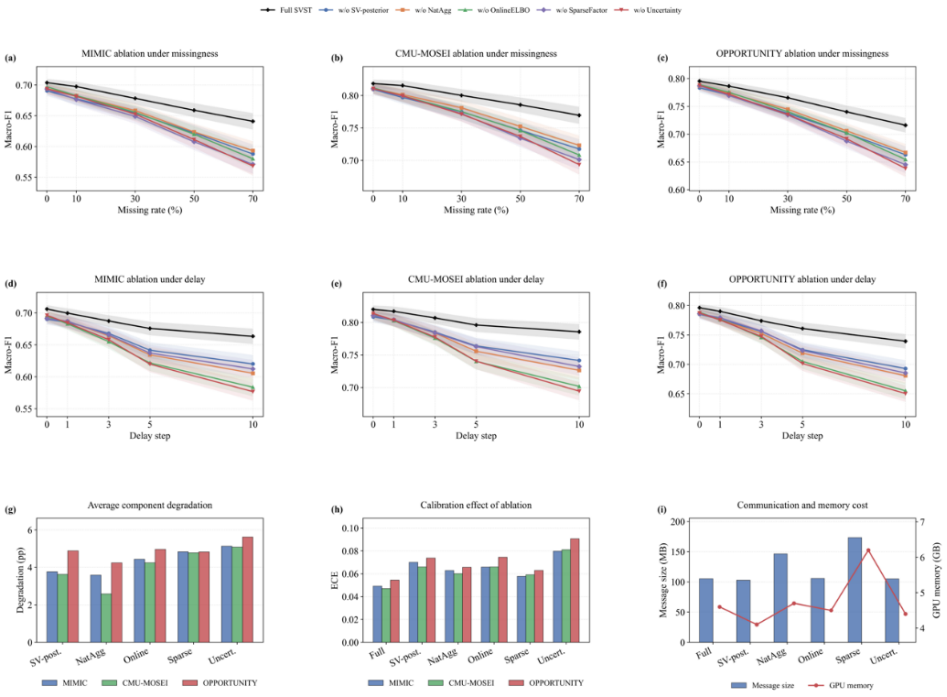


Fig. 12. Nine-panel ablation visualization with repeated-run error bars.

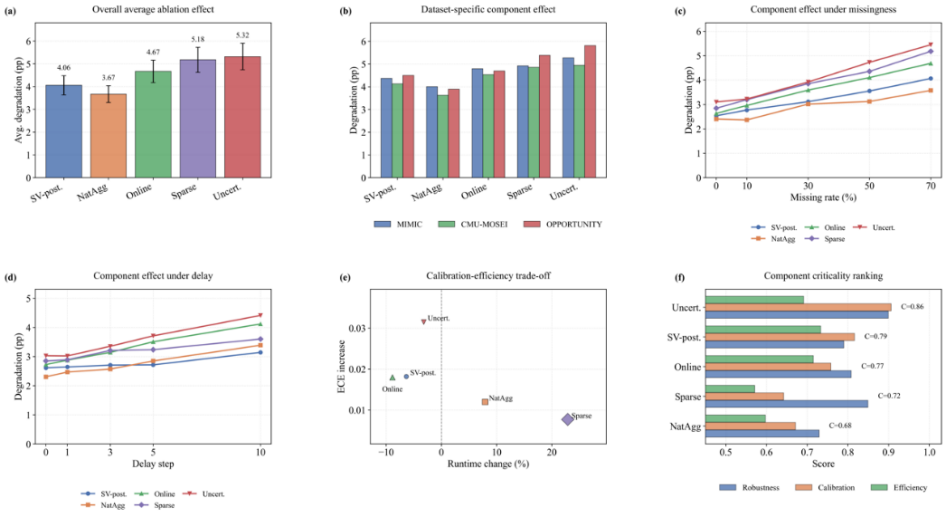


Fig. 13. Average ablation effect across datasets.

4.9. Hyperparameter Sensitivity

Figure 14 evaluates the sensitivity of SVST to tensor rank, sparse factor retention ratio, and online update coefficient. Tensor rank is varied among 4, 8, 16, 24, and 32. Sparse factor retention ratio is varied from 10% to 70%. The online update coefficient is varied from 0.05 to 0.30. Each curve is averaged across repeated runs, and shaded areas represent 95% confidence intervals.

Rank 16 provides the best trade-off between performance and communication cost. Rank 4 underfits cross-modal interactions, while rank 32 adds communication overhead with limited performance gain. A sparse factor retention ratio of 30% performs consistently well. Lower retention ratios discard weak but useful interactions, while higher ratios increase communication cost and slightly worsen calibration. The online update coefficient is stable between 0.10 and 0.20; larger values produce less stable ELBO trajectories under strong non-IID distributions.

4.10. Online Adaptation and Real-World Deployment Illustration

Figure 15 evaluates online adaptation after a distribution shift is introduced at round 50 and also illustrates the realistic deployment scenario of SVST. For MIMIC, the shift changes disease-label distribution and increases note missingness. For CMU-MOSEI, the shift changes topic distribution. For OPPORTUNITY, the shift removes one sensor group from 30% of the clients.

Static training shows a clear performance drop after the shift and recovers slowly. Periodic update improves recovery but remains unstable when the shift coincides with high missingness. SVST without online ELBO smoothing responds faster but produces larger NLL oscillations. Full SVST achieves the most stable recovery. On MIMIC, Macro-F1 drops from 0.684 to 0.651 immediately after the shift and recovers to 0.676 within 14

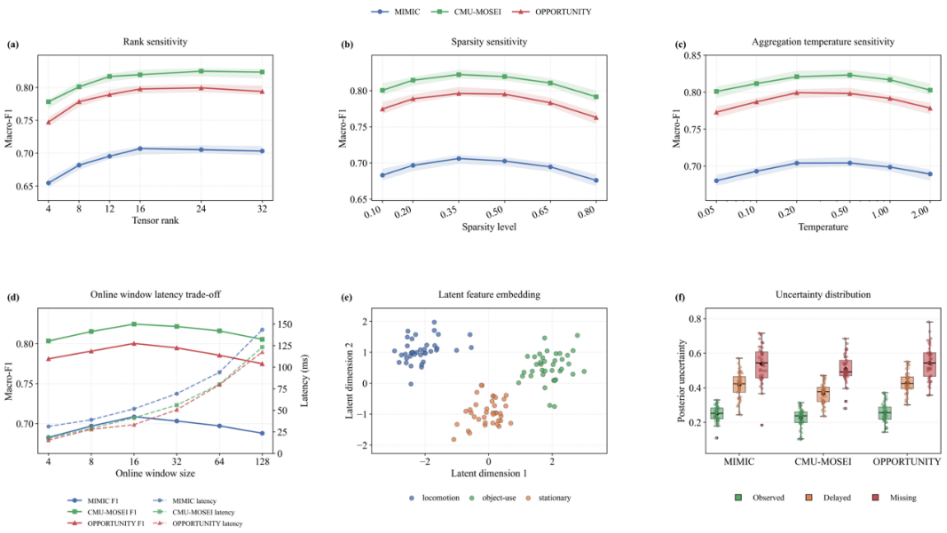


Fig. 14. Hyperparameter sensitivity and feature visualization.

rounds. PEPSY recovers to 0.661 after 21 rounds, while FedAvg recovers to 0.628 after 25 rounds. The confidence bands show that the faster recovery of SVST is stable across repeated partitions.

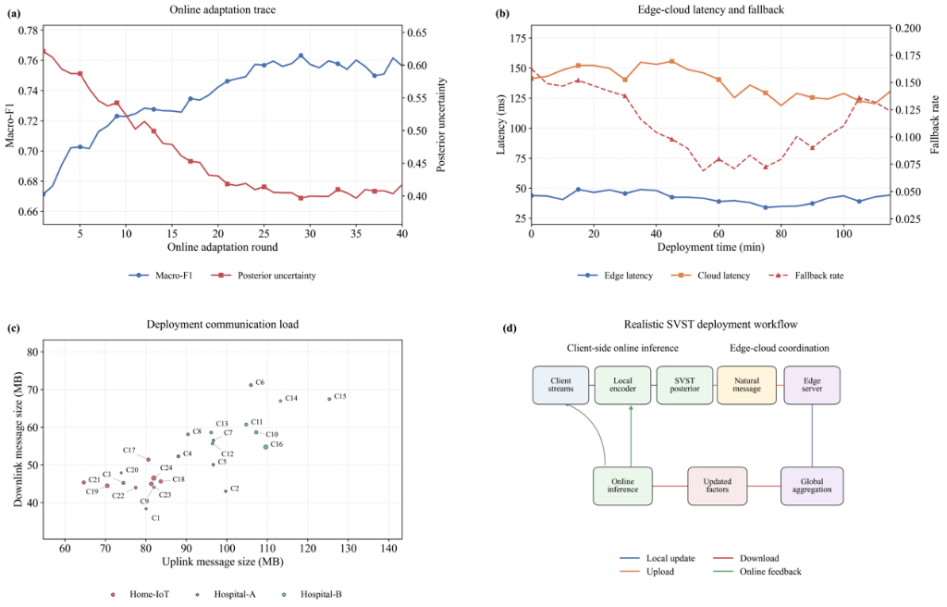


Fig. 15. Online adaptation and realistic distributed deployment scenario.

4.11. Cross-Client and Cross-Domain Generalization

Table 10 reports generalization results. For cross-client generalization, models are trained on 80% of clients and tested on unseen clients. For cross-domain generalization, CMU-MOSEI is split by topic groups, and OPPORTUNITY is evaluated using subject-held-out splits. Direct transfer between MIMIC and OPPORTUNITY is not performed because the two datasets differ substantially in task, modality, and label space.

SVST achieves the smallest performance drop in all three settings. On MIMIC unseen clients, SVST reduces the Macro-F1 drop to 2.4 points, compared with 4.7 points for FedAvg and 3.1 points for PEPSY. On OPPORTUNITY, the gap between FedMobile and SVST is small, indicating that FedMobile is a strong incomplete-modality baseline for sensor-based FL. However, SVST still achieves lower ECE and smaller degradation. The reported standard deviations indicate that the cross-client advantage is not caused by one favorable split.

Table 10. Cross-client and cross-domain generalization results

Setting	Method	Source Macro-F1	Target Macro-F1	Performance Drop	Target ECE
MIMIC seen → unseen clients	FedAvg	0.641 ±0.008	0.594 ±0.011	0.047 ±0.009	0.116 ±0.008
MIMIC seen → unseen clients	SCAFFOLD	0.661 ±0.006	0.622 ±0.010	0.039 ±0.008	0.095 ±0.007
MIMIC seen → unseen clients	PEPSY	0.677 ±0.006	0.646 ±0.008	0.031 ±0.007	0.082 ±0.006
MIMIC seen → unseen clients	MVAE/ PoE-VAE	0.673 ±0.007	0.641 ±0.009	0.032 ±0.007	0.078 ±0.006
MIMIC seen → unseen clients	SVST	0.682 ±0.006	0.658 ±0.007	0.024 ±0.006	0.061 ±0.005
CMU-MOSEI topic split	FedAvg	0.765 ±0.007	0.721 ±0.010	0.044 ±0.008	0.101 ±0.007
CMU-MOSEI topic split	MulT	0.806 ±0.005	0.774 ±0.007	0.032 ±0.006	0.077 ±0.005
CMU-MOSEI topic split	PEPSY	0.795 ±0.005	0.763 ±0.008	0.032 ±0.006	0.079 ±0.005
CMU-MOSEI topic split	MVAE/ PoE-VAE	0.793 ±0.007	0.758 ±0.009	0.035 ±0.007	0.073 ±0.005
CMU-MOSEI topic split	SVST	0.802 ±0.005	0.779 ±0.006	0.023 ±0.005	0.058 ±0.004
OPPORTUNITY subject-held-out	FedAvg	0.731 ±0.010	0.675 ±0.014	0.056 ±0.011	0.122 ±0.010
OPPORTUNITY subject-held-out	SCAFFOLD	0.751 ±0.008	0.704 ±0.012	0.047 ±0.009	0.099 ±0.008
OPPORTUNITY subject-held-out	FedMobile	0.771 ±0.008	0.735 ±0.010	0.036 ±0.008	0.082 ±0.006
OPPORTUNITY subject-held-out	PEPSY	0.768 ±0.007	0.729 ±0.011	0.039 ±0.008	0.086 ±0.007
OPPORTUNITY subject-held-out	SVST	0.779 ±0.007	0.747 ±0.009	0.032 ±0.007	0.065 ±0.005

4.12. Statistical Significance and Error Analysis

Table 11 reports paired t-tests, Wilcoxon signed-rank tests, and effect sizes over 15 repeated runs between SVST and the strongest baseline on each dataset. For MIMIC, PEPSY is selected according to AUPRC. For CMU-MOSEI, MulT is selected according to Accuracy. For OPPORTUNITY, FedMobile is selected according to Macro-F1.

SVST significantly improves AUPRC and ECE on MIMIC. On CMU-MOSEI, the Accuracy difference between SVST and MulT is not statistically significant, but SVST significantly improves ECE. On OPPORTUNITY, SVST significantly improves Macro-F1 and ECE over FedMobile. These results show that SVST does not dominate every metric, but provides stable gains in calibration, robustness, and communication efficiency.

Table 11. Statistical Significance over 15 Repeated Runs

Dataset	Metric	Strongest Baseline	Baseline Score	SVST Score	Paired t-test p	Wilcoxon p	Cohen's d	Interpretation
MIMIC	AUPRC	PEPSY	0.543 ± 0.007	0.561 ± 0.006	0.012	0.018	2.75	significant
MIMIC	ECE	MVAE/ PoE-VAE	0.069 ± 0.004	0.052 ± 0.004	0.004	0.006	4.25	significant
CMU-MOSEI	Accuracy	MulT	0.835 ± 0.004	0.832 ± 0.004	0.184	0.211	0.75	not significant
CMU-MOSEI	ECE	MVAE/ PoE-VAE	0.059 ± 0.004	0.046 ± 0.003	0.007	0.009	3.68	significant
OPPOR-TUNITY	Macro-F1	FedMobile	0.771 ± 0.008	0.779 ± 0.007	0.031	0.038	1.06	significant
OPPOR-TUNITY	ECE	MVAE/ PoE-VAE	0.064 ± 0.004	0.048 ± 0.004	0.005	0.007	4.00	significant

Table 12. Error type analysis of SVST across repeated runs

Error Type	MIMIC Proportion	CMU-MOSEI Proportion	OPPOR-TUNITY Proportion	Main Manifestation
Severe simultaneous modality absence	$27.4\% \pm 2.1\%$	$21.8\% \pm 1.8\%$	$24.6\% \pm 2.0\%$	Multiple informative modalities are missing in the same sample
Long asynchronous delay	$14.9\% \pm 1.5\%$	$17.5\% \pm 1.6\%$	$25.1\% \pm 2.3\%$	Delayed modalities arrive after the decision window
Client distribution shift	$22.6\% \pm 2.0\%$	$15.2\% \pm 1.4\%$	$16.7\% \pm 1.7\%$	Unseen clients contain shifted labels or sensor patterns
Noisy modality dominance	$11.8\% \pm 1.2\%$	$18.6\% \pm 1.7\%$	$12.9\% \pm 1.3\%$	A noisy modality receives overly strong posterior influence
Sparse factor under-selection	$9.7\% \pm 1.0\%$	$10.4\% \pm 1.1\%$	$8.8\% \pm 0.9\%$	Weak cross-modal dependencies are removed by sparsity
Calibration failure	$8.5\% \pm 0.9\%$	$9.6\% \pm 1.0\%$	$7.4\% \pm 0.8\%$	High-confidence errors under rare labels or ambiguous states
Other errors	$5.1\% \pm 0.7\%$	$6.9\% \pm 0.8\%$	$4.5\% \pm 0.6\%$	Label noise, preprocessing errors, or rare event transitions

Table 12 summarizes the main error types of SVST. The proportions are computed over all failed predictions across repeated runs rather than from a single test split. Severe simultaneous modality absence is the most frequent error source. Long asynchronous delay is especially common on OPPORTUNITY, where delayed sensor streams may arrive after the decision window. Client distribution shift is more frequent on MIMIC because unseen clients may contain rare disease combinations or sparse note-image pairs. Sparse factor under-selection occurs when weak but useful modality interactions are pruned by the sparse tensor constraint.

5. Discussion

This paper proposes a scalable structured variational sparse tensor method for distributed multimodal data streams. Unlike methods that treat multimodal fusion, uncertainty modeling, and distributed aggregation as separate components, SVST directly parameterizes a structured variational posterior with sparse low-rank tensor factors and aggregates posterior statistics in the natural-parameter space [4,9,2]. Compared with memory-driven prompting, dual-masked VAE learning, and distillation-based missing-modality methods, SVST reduces reliance on reconstructing unavailable modalities and instead constructs the inference objective over observed modalities through likelihood masking [36,27,32]. Sparse tensor factors capture discriminative high-order interactions while limiting dependence on noisy or redundant modalities. The structured posterior models latent correlations and uncertainty, while natural-parameter aggregation enables clients to transmit compact posterior contributions without constructing a dense global tensor. Online ELBO updates and delay-aware weighting further adjust modality contributions according to observational reliability. Unavailable modality likelihood terms are omitted from the reconstruction objective rather than explicitly integrated over missing raw-modality values. The central contribution of SVST is not the parallel combination of these components, but a common posterior-statistics pathway in which sparse tensor factors construct the structured posterior, natural parameters serve as the distributed communication object, and missingness, delay, and uncertainty regulate the same posterior update and aggregation process.

Experiments across multiple multimodal datasets show that SVST generally outperforms or remains competitive with federated learning, multimodal federated learning, tensor fusion, and variational inference baselines. Under missing modalities, asynchronous streams, and non-IID client distributions, it maintains lower degradation and more reliable uncertainty estimates, consistent with studies emphasizing modality completion, staged augmentation, and reliability-aware diagnosis [31,8,37]. Compact sparse factors and natural-parameter messages also support communication-constrained applications, including clinical analysis, sensor modeling, affective computing, and cross-device perception [12,3].

Several limitations remain. The experiments rely mainly on public benchmarks and simulated distributed streams, so long-term real-world stability requires further validation. Hyperparameters such as tensor rank, sparsity strength, and delay weighting remain task dependent, and severe missingness, long disconnections, or distribution shifts may still reduce performance. The structured posterior and tensor updates also introduce additional computation on highly constrained edge devices. Although SVST avoids direct

raw-data transmission, natural-parameter increments and sparse-factor updates may still leak statistical information. The current framework therefore does not formally protect against gradient inversion, membership inference, or related reconstruction attacks. Differential privacy could be introduced through update clipping and perturbation, while secure aggregation could hide individual client contributions. Their effects on calibration, factor stability, and efficiency require further study. Future work will explore adaptive rank selection, lightweight optimization, event-driven communication, continual learning, differential privacy, secure aggregation, and long-term deployment.

6. Conclusion

This paper addresses the challenges of missing modalities, asynchronous modality arrival, client-level heterogeneity, and communication constraints in distributed multimodal data streams by proposing a scalable structured variational sparse tensor representation learning framework, namely SVST. The framework directly parameterizes the structured variational posterior with sparse low-rank tensor factors, thereby integrating high-order cross-modal interaction modeling and uncertainty estimation into a unified posterior construction process. Meanwhile, the natural-parameter aggregation mechanism transmits compact posterior statistics in distributed environments, avoiding the centralized construction of a full global tensor and reducing communication overhead. Furthermore, online ELBO updates, observed-modality likelihood masking, and delay-aware weighting jointly enhance inference stability under incomplete and asynchronous observations. Extensive experimental results show that the proposed method generally outperforms or remains competitive with existing federated learning, multimodal federated learning, tensor fusion, and variational inference methods across multiple real-world multimodal datasets and complex distributed experimental settings. SVST demonstrates stronger robustness, better generalization ability, more stable predictive performance, and more reliable uncertainty estimation. Future work will further explore adaptive tensor-rank selection, lightweight edge deployment, event-driven communication, cross-domain continual learning, differential privacy and secure aggregation for protecting communicated natural-parameter and sparse-factor updates, as well as long-term validation in real distributed multimodal systems, to further improve the practical applicability and engineering scalability of SVST.

References

1. Blei, D.M., Kucukelbir, A., McAuliffe, J.D.: Variational inference: A review for statisticians. Journal of the American Statistical Association 112(518), 859–877 (2017)
2. Chen, J., Cheng, S., Yutao, Y., Zhang, Y., Yuan, H., Peng, P., Zhong, Y.: PROMISE: Prompt-attentive hierarchical contrastive learning for robust cross-modal representation with missing modalities. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 20076–20082 (Mar 2026)
3. Chen, J., Liu, J., Liu, S., Zhang, W., Li, A., Zhu, E., Liu, X.: Sample-specific modality diagnosis and cross-modal enhancement for incomplete multimodal representations. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 20154–20162 (Mar 2026)
4. Dai, R., Li, C., Yan, Y., Mo, L., Qin, K., He, T.: Unbiased missing-modality multimodal learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 24507–24517 (2025)

5. Feng, T., Bose, D., Zhang, T., Hebbar, R., Ramakrishna, A., Gupta, R., ... & Narayanan, S.: Fedmultimodal: A benchmark for multimodal federated learning. In: Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. pp. 4035–4045 (Aug 2023)
6. Hoffman, M.D., Blei, D.M., Wang, C., Paisley, J.: Stochastic variational inference. Journal of Machine Learning Research (2013)
7. Hu, P., Qin, Y., Gou, Y., Li, Y., Yang, M., Peng, X.: Probabilistic multimodal learning with von mises-fisher distributions. In: Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence. pp. 5390–5398 (Aug 2025)
8. Huang, H., Gong, T., He, K., Wen, W., Zhang, W., Feng, M.: Recovering coherent affective patterns: Addressing modality missing in multimodal sentiment analysis. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 21957–21965 (Mar 2026)
9. Huang, W., Chen, Y., Jiang, X., Gao, C., Zhang, T., Chen, Q., Wang, Y.: Mitigating pervasive modality absence through multimodal generalization and refinement. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 26796–26804 (Apr 2025)
10. Karimireddy, S.P., Kale, S., Mohri, M., Reddi, S., Stich, S., Suresh, A.T.: Scaffold: Stochastic controlled averaging for federated learning. In: International Conference on Machine Learning. pp. 5132–5143. PMLR (Nov 2020)
11. Kolda, T.G., Bader, B.W.: Tensor decompositions and applications. SIAM Review 51(3), 455–500 (2009)
12. Kuai, Z., Zhang, H., Yang, Y., Ma, Y., Huang, M., Gui, N., Kuang, L.: Generating-filtering-ranking: A three-stage multimodal data augmentation framework under partial modality missing. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 5755–5763 (Mar 2026)
13. Lee, Y.L., Tsai, Y.H., Chiu, W.C., Lee, C.Y.: 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)
14. Li, T., Sahu, A.K., Zaheer, M., Sanjabi, M., Talwalkar, A., Smith, V.: Federated optimization in heterogeneous networks. Proceedings of Machine Learning and Systems 2, 429–450 (2020)
15. Liu, Y., Wang, C., Yuan, X.: Fedmobile: Enabling knowledge contribution-aware multi-modal federated learning with incomplete modalities. In: Proceedings of the ACM on Web Conference 2025. pp. 2775–2786 (Apr 2025)
16. Liu, Z., Shen, Y., Lakshminarasimhan, V.B., Liang, P.P., Zadeh, A.B., Morency, L.P.: Efficient low-rank multimodal fusion with modality-specific factors. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 2247–2256 (Jul 2018)
17. McMahan, B., Moore, E., Ramage, D., Hampson, S., y Arcas, B.A.: Communication-efficient learning of deep networks from decentralized data. In: Artificial Intelligence and Statistics. pp. 1273–1282. PMLR (Apr 2017)
18. Nguyen, D., Hoang, N., Huynh, T.T., Nguyen, Q.V.H., Nguyen, P.L.: Learning reconfigurable representations for multimodal federated learning with missing data. Advances in Neural Information Processing Systems 38, 28808–28845 (2026)
19. Poudel, P., Chhetri, A., Gyawali, P., Leontidis, G., Bhattarai, B.: Multimodal federated learning with missing modalities through feature imputation network. In: Annual Conference on Medical Image Understanding and Analysis. pp. 289–299. Springer Nature Switzerland, Cham (Jul 2025)
20. Poudel, P., Shrestha, P., Amgain, S., Shrestha, Y.R., Gyawali, P., Bhattarai, B.: 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 Nature Switzerland, Cham (Oct 2024)

21. Qiu, P., Zhu, W., Kumar, S., Chen, X., Yang, J., Sun, X., ... & Sotiras, A.: Multimodal variational autoencoder: A barycentric view. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 20060–20068 (Apr 2025)
22. Saha, P., Mishra, D., Wagner, F., Kamnitsas, K., Noble, J.A.: Incongruent multimodal federated learning for medical vision and language-based multi-label disease detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 28331–28339 (Apr 2025)
23. Sutter, T. M., Meng, Y., Agostini, A., Chopard, D., Fortin, N., Vogt, J. E., ... & Mandt, S.: Unity by diversity: improved representation learning for multimodal vaes. Advances in Neural Information Processing Systems 37, 74262–74297 (2024)
24. Tan, M., Ma, J., Feng, Y., Ding, J., Pan, W., Han, T., ... & Yu, Z.: FedAFD: Multimodal federated learning via adversarial fusion and distillation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3400–3409 (2026)
25. Tsai, Y.H.H., Bai, S., Liang, P.P., Kolter, J.Z., Morency, L.P., Salakhutdinov, R.: Multimodal transformer for unaligned multimodal language sequences. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 6558–6569 (Jul 2019)
26. Wang, H., Liu, X., Zhong, X., Chen, L., Liu, F., Zhang, W.: Multimodal online federated learning with modality missing in internet of things. IEEE Transactions on Mobile Computing (2025)
27. Wang, J., Guo, K., Hu, P., Peng, X., Wang, H.: Learning robust multi-view representation using dual-masked vaes. In: Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence. pp. 6298–6306 (Aug 2025)
28. Wang, J., Liu, Q., Liang, H., Joshi, G., Poor, H.V.: Tackling the objective inconsistency problem in heterogeneous federated optimization. Advances in Neural Information Processing Systems 33, 7611–7623 (2020)
29. Wu, M., Goodman, N.: Multimodal generative models for scalable weakly-supervised learning. Advances in Neural Information Processing Systems 31 (2018)
30. Xiong, H., Dai, M.: FedMIR: Multimodal federated learning with missing modality imputation and distribution-aware routing. Sensors 26(10), 2954 (2026)
31. Xu, H., Wu, H., Ke, X., Wu, J., Xu, R., Xu, J.: Mcmoe: Completing missing modalities with mixture of experts for incomplete multimodal action quality assessment. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 11241–11249 (Mar 2026)
32. Yang, Y., Qu, J., Huang, L., Dong, W.: DPMamba: Distillation prompt mamba for multimodal remote sensing image classification with missing modalities. In: Proc. 34th Int. Joint Conf. Artif. Intell. pp. 2224–2232 (Aug 2025)
33. Yu, S., Wang, J., Hussein, W., Hung, P.C.: Robust multimodal federated learning for incomplete modalities. Computer Communications 214, 234–243 (2024)
34. Yuan, L., Han, D.J., Wang, S., Upadhyay, D., Brinton, C.G.: Communication-efficient multimodal federated learning: Joint modality and client selection. IEEE Transactions on Mobile Computing (2026)
35. Zadeh, A., Chen, M., Poria, S., Cambria, E., Morency, L.P.: Tensor fusion network for multimodal sentiment analysis. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. pp. 1103–1114 (Sep 2017)
36. Zhao, Y., Xi, W., Fu, X., Zhao, J.: Enhancing multimodal model robustness under missing modalities via memory-driven prompt learning. In: Proceedings of the 34th International Joint Conference on Artificial Intelligence, IJCAI 2025. pp. 2458–2466. International Joint Conferences on Artificial Intelligence (Aug 2025)
37. Zhao, Y., Chen, C., Pang, Q. Y., Fu, Y., Li, Q., Tang, C., ... & Jin, Y.: Tackling dual-stage missing modalities in brain tumor segmentation via robust modality reconstruction and prompt-guided modality adaptation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 13314–13322 (Mar 2026)

Lianming Hao holds a Ph.D. and is an Associate Professor and Master's Supervisor at the College of Mathematics and Computer Science, Jilin Normal University, China. His research interests include mathematics education, mathematical reasoning and assessment, mathematics curriculum and instruction, and innovative teaching approaches in mathematics education.

Fengxu Zhu is currently a student at Jilin Normal University, China. His research interests mainly focus on mathematics education, mathematics teaching and learning, and the development of students' mathematical thinking and problem-solving abilities. He is particularly interested in innovative instructional approaches and the application of educational theories to mathematics classroom practice.

Received: June 26, 2026; Accepted: July 24, 2026.