Fed-MoSeC: A Federated Learning Framework for Cross-Modal Semantic Communication Systems in Mobile Networks

Yushi Wang, Zhengxin Yu, Guhan Zheng, *Member, IEEE*, Haris Pervaiz, *Member, IEEE*, Aryan Kaushik, *Member, IEEE*, Weizhi Meng, *Senior Member, IEEE*, and Shunqing Zhang, *Senior Member, IEEE*

***Abstract*—Semantic communication systems in mobile networks necessitate real-time collaborative updates of semantic encoders as a result of node mobility that induces semantic extraction drift. However, diversity within modal transmission content and heterogeneous encoder architectures, along with challenges such as imbalanced training requirements and pseudo-label noise, limit the effectiveness of general collaborative update approaches. In this paper, we propose Fed-MoSeC, a novel federated learning framework for updating cross-modal semantic encoders. Our framework trains only a newly designed graph neural network-based adapter while freezing heterogeneous encoders for various modalities, converting heterogeneous cross-modal updates into a homogeneous aggregation task, and significantly reducing communication overhead. By combining confidence-based filtering with similarity-matrix distillation, the novel integrated Pseudo-label Noise Counteracting Component (PNCC) is designed to be robust to noisy data. The Training Optimization Component (TOC) based on a bi-level Cournot–Stackelberg game theoretical algorithm achieves near-optimal Subgame-Perfect Nash Equilibrium (SPNE) to incentivize across nodes and maximize update nodes' utilities with different training levels. Experimental results highlight the advantages of Fed-MoSeC over existing potential application algorithms, reducing communication by 95–97% and improving RSUM by 18% at 70% pseudo-label noise.**

***Index Terms*—Semantic communications, federated learning, cross-modal, mobile networks**

I. Introduction

With the rapid development of the Internet of Things (IoT) and edge computing, mobile devices are introducing numerous emerging data-intensive applications, such as mobile augmented reality, mobile 3D reconstruction and virtual reality. These applications transform mobile devices into highly perceptive agents that rely on massive data streams generated and processed by various devices, *e.g.*, camera visual inputs, LiDAR point clouds, auditory data, radar, and textual traffic reports [1]. Although various modal data exchange enables high service quality and new capabilities, it also significantly increases data transmission loads on Mobile Networks (MNs) with constrained spectrum resources.

Semantic Communication (SC) has been considered as a promising communication paradigm to enhance spectral efficiency and reduce transmission load in MNs. Fig. 1 illustrates the workflow of SC. Unlike the traditional communication paradigm that transmits raw or lightly compressed data bits, SC extracts and transmits only the essential meaning or intent relevant to the task [2]. At the transmitter, a Deep Learning (DL)-based semantic encoder maps input data to semantic symbols and aligns them with a Semantic Knowledge Base (SKB) for efficient encoding. At the receiver, the semantic decoder uses the shared SKB to reconstruct the intended information from the received semantic symbols [3]. DL-based semantic encoders are typically task-oriented, which makes them vulnerable to node mobility induced shifts in environment and input distribution. Such shifts degrade semantic extraction accuracy, disrupt alignment with the SKB, and impair symbol reconstruction at the receiver. Consequently, practical deployment of the SC system faces a fundamental challenge: the semantic encoder of nodes in MNs requires real-time updates to evolve semantic extraction capabilities for sending accurate known symbols in conjunction with SKB.

To ensure update accuracy and encoder environmental adaptability, collaborative learning through joint updates across multiple nodes becomes essential [4]. However, direct data sharing is often infeasible due to privacy concerns and communication constraints, which motivates a learning mechanism that enables collaborative SC encoder updates and continual model adaptation without raw data exchange. Federated Learning (FL) can naturally support collaborative updating of the semantic encoder [5] [6]. It enables joint updates of semantic encoders on participating nodes while preserving data privacy by training encoder models locally and aggregating model parameters at the edge server.

A. Challenges

Directly applying conventional FL methods to dynamic MNs presents several distinct challenges that arise from heterogeneous node mobility. Nodes differ in movement patterns:

This work was supported by the National Natural Science Foundation of China under Grant 62571307, the National Key Research and Development Program of China under Grants 2022YFB2902304, the Science and Technology Commission Foundation of Shanghai under Grants 24DP1500500, and the EPSRC and DSIT through the Communications Hub for Empowering Distributed Cloud Computing Applications and Research (CHEDDAR) [grant numbers EP/X040518/1 and EP/Y037421/1].

Yushi Wang, Zhengxin Yu, and Weizhi Meng are with the School of Computing and Communications, Lancaster University, LA1 4YW, United Kingdom. (e-mail: y.wang216@lancaster.ac.uk; z.yu8@lancaster.ac.uk; w.meng3@lancaster.ac.uk).

Guhan Zheng and Shunqing Zhang are with the School of Communication and Information Engineering, Shanghai University, 200444, China. (email: gzheng@shu.edu.cn; shunqing@shu.edu.cn).

Haris Pervaiz is with the School of Computer Science and Electronic Engineering, University of Essex, CO4 3SQ, United Kingdom. (email: haris.pervaiz@essex.ac.uk).

Aryan Kaushik is with the RakFort, Ireland, and IIITD, India. (email: a.kaushik@ieee.org).

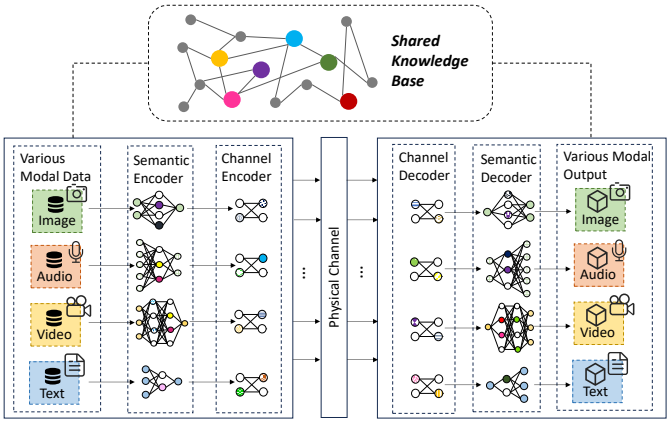


Fig. 1. Illustration of SC workflow. The transmitter employs heterogeneous, various modal task-specific semantic encoders to extract semantic representations with the assistance of the SKB, and heterogeneous semantic decoders reconstruct the original task semantics.

Roadside Sensors (RSs) and infrastructure units are stationary and act as stable nodes, whereas Autonomous Vehicles (AVs) traverse cells rapidly and act as mobile nodes. This mobility diversity, coupled with diverse computational budgets and latency requirements, leads to model heterogeneity. During initial training, different usage scenarios produce distinct encoder architectures. For example, RSs train small encoder models with fast transmissions to fit their limited computational and storage capacity, whereas AVs train large models with higher transmission accuracy to guarantee driving security. Despite these architectural differences, encoders should be jointly updated for the same shared type of tasks to maintain the benefits of SC and thus reduce network spectrum resource consumption. Traditional FL approaches, however, assume uniform model structures across nodes and become ineffective for SC scenarios in MNs.

The challenges are further compounded by the need for SC in MNs to transmit data in multiple modalities. Newly joining nodes (*i.e.,* mobile nodes) often need to simultaneously update multiple modalities' task-oriented encoder models. For instance, an AV that joins the network may require concurrent updates of speech, text, and image semantic encoders. Each modality requires separate FL sessions with its own network participant sets, resulting in multiple parallel FL processes that dramatically increase computational overhead and communication costs. This resource burden becomes especially severe in the case of joint accounting for model heterogeneity within each modality group.

A critical bottleneck in this process is the lack of accurate training labels in mobility scenarios, which are required to compute loss values during model updates. Although pseudo-labeling methods [7] can generate labels assisted by SKB, these labels often contain important information but different errors/noise due to the movement. Each erroneous label degrades learning accuracy, biases gradient estimates, and correlations across modalities allow pseudo-labels from one modality to propagate noise to others, resulting in interdependent noise amplification [8]. As noise accumulates across data streams and modalities, semantic encoders gradually lose accuracy in extracting and interpreting meaning. In case performance falls below service threshold, nodes may abandon SC altogether, reverting to traditional bit-based transmission. This fallback mechanism substantially increases network latency and load, potentially compromising the stringent security and timing requirements of MNs.

Moreover, as heterogeneous movement patterns of nodes, the MNs also include stable nodes with well-trained models adapted to their specific operational environments. These models are trained with high-quality labels, making them relatively accurate and robust. However, they still need to incorporate data from mobile nodes and to receive periodic updates to maintain accuracy in evolving conditions. This situation creates an imbalance in FL update requirements. Mobile nodes with substantial noise labels need extensive model alterations, while stable nodes with numerous accurate labels only require minor refinements. Applying the same FL process to both groups is therefore highly suboptimal, since it subjects stable nodes with high-quality labels to low-quality updates while failling to allocate sufficient attention to the models on mobile nodes. Therefore, it is necessary to design robust semantic encoder update strategies to mitigate the vastly heterogeneous movement patterns of nodes in MNs, considering unique and interconnected operational factors. Such factors include model heterogeneity across nodes, diversity modalities encoder concurrent updating, pseudo-label noise propagation, and imbalanced training requirements.

B. Related Works

Several existing FL-based approaches show promise for SC in MNs, but critical gaps remain. FL-based methods for model heterogeneity synchronize differing encoder architectures [9], [10], but they operate in single-modality scenarios and require separate encoder synchronization, which creates bandwidth bottlenecks that scale with sensor diversity. Multimodal FL frameworks [11], [12] enable cross-modal learning but overlooked model heterogeneity and label noise. To mitigate label noise, pseudo-label denoising techniques [8], [13] offered partial solutions. They assume static neighbor graphs or rely on costly confidence voting, both unrealistic in mobile networks with rapidly changing topology. The resulting label noise amplifies rapidly due to context drift in mobile environments. Moreover, many existing schemes assumed uniform FL training requirements across nodes, ignoring the disparate needs of stable versus mobile nodes. Mobility-aware schedulers [14], [15] reduced stragglers by dropping low quality updates, but they inadvertently excluded mobile nodes from semantic-layer training. It forces them back to inefficient bit-level transmission and negates SC's core benefits. Therefore, these limitations in handling model heterogeneity, multiple modalities, different training requirements, and pseudo-labeling noise prevent current methods from fully addressing SC challenges in MNs.

C. Motivation and Contributions

To bridge the abovementioned gap, we propose Fed-MoSeC, a lightweight FL framework for SC in MNs. Fed-MoSeC

converts heterogeneous updates of multiple modality-specific encoders into homogeneous updates of a shared cross-modal semantic encoder adapter. It enables nodes to perform federated aggregation and thus collaboratively train encoder models. Stable nodes and mobile nodes also allow mutual assistance for enhanced both parties training efficiency considering label noise via dynamically assuming teacher or student roles. Notably, Fed-MoSeC also incorporates a novel Pseudo-label Noise Counteracting Component (PNCC) and a proprietary Training Optimization Component (TOC). PNCC mitigates the adverse effects of cross-modal pseudo-labels noise, improving training accuracy. Moreover, the TOC establishes a mutually beneficial training paradigm by adapting the teaching-study process to the mobile-aware utilities of participating nodes.

The main contributions are summarized as follows:

- We propose a novel SC updating framework, Fed-MoSeC, tailored for multiple modalities semantic encoder updates between heterogeneous nodes in MNs. The envisioned framework utilizes the designed GNN model to transform the heterogeneous semantic encoder updating into a homogeneous cross-modal updating, for various modal tasks. FL then aggregates this homogeneous distilled knowledge, creating a bridge for mutual learning between established stable nodes and newly joining mobile nodes, allowing both to benefit and enhance their respective multi-modal semantic encoders.
- We develop PNCC, an integrated mechanism tailored for the Fed-MoSeC framework and its adapter to enable robust federated cross-modal learning. PNCC combines a reliability-weighted knowledge transfer with adaptive pseudo-label filtering, leveraging a multi-dimensional evaluation scheme to quantify node reliability. The key dimensions include diagonal alignment strength, discrimination ratio, cross-modal consistency, and intra-modal similarity. To further enhance generalization under privacy constraints, PNCC incorporates a similarity matrix distillation approach, which mitigates the impact of noisy pseudo-labels through adaptive threshold-based filtering that dynamically adjusts to local data distributions. By integrating the dual perspectives of data reliability and model diversity, PNCC empowers Fed-MoSeC to function as a robust and efficient federated cross-modal learning ecosystem.
- We investigate the training motivation of both stable nodes and mobile nodes in Fed-MoSeC, and propose a TOC to maximize the training benefits of nodes via incentivizing stable nodes and ensuring that Fed-MoSeC operates properly. Our component takes a comprehensive stance, jointly considering velocity, energy cost, noise-aware distillation process, and federated aggregation strategies. Facing the utility maximization objective formulated as a multi-node coupled intricate mixed integer nonlinear programming (MINLP) problem, we relax the discrete variables using the binary variable relaxation method. A bi-level Cournot-Stackelberg game theoretical algorithm is then presented to approximate the subgame-perfect Nash equilibrium (SPNE) via multistage backward induction.

The rest of this paper is organized as follows. Related works are reviewed in Section II. We then present the considered system model in Section III. In Section IV, the design details of the Fed-MoSeC framework are presented with the PNCC and the TOC. Section V illustrates the simulation results to demonstrate the effectiveness of our Fed-MoSeC. Finally, we conclude this paper in Section VI.

II. Related Works

The FL-based methods have emerged as a promising approach for collaborative semantic encoder updating while preserving data privacy in various mobile scenarios. Zheng et al. [10] proposed a privacy-preserving personalized FL framework for SC in vehicular networks, enabling collaborative encoder training across vehicular participants. Similarly, hierarchical FL-powered SC for UAV swarm cooperation was proposed in [16], organizing UAVs into clusters for efficient model training. In [17], PSFed for updating SC codecs in satellite-borne edge cloud networks was introduced. Nevertheless, these works are conventional FL aggregation algorithms such as FedAvg [18] or FedProx [19] for non-independent and identically distributed (non-IID) data, but applicable only to homogeneous encoders' architecture.

The recent efforts of FL also have the potential for SC encoder updating in MNs under the heterogeneous or different modal encoder. Lin et al. [9] investigated ensemble distillation for robust model fusion across different network structures. Tan et al. [20] used federated prototype learning to manage clients through shared representations. Moreover, Xiong et al. [11] proposed MMFed considers multimodal federated aggregation. Yu et al. [12] also presented CreamFL, which relaxes the homogeneity assumption by allowing heterogeneous client models and modalities via representation-level ensemble knowledge transfer on public data with global–local cross-modal contrastive aggregation. These approaches, however, rely on the availability of accurately labeled data and significantly increase communication overhead, posing a substantial challenge in mobile scenarios.

In mobile scenarios, the lack of accurate training labels necessitates pseudo-labeling noise-aware approaches. Sohn et al. [21] proposed FixMatch, which combines high-confidence pseudo-labeling with consistency regularization on strongly augmented data. Furthermore, Dai et al. [8] and Ding et al. [22] proposed graph-based methods to prevent error propagation. The GRAND [14] and PTA [13] were also designed to enhance model robustness through random feature propagation or an adaptive weighting strategy. Nonetheless, these methods are fundamentally unsuitable for dynamic mobile FL. They predominantly assume a static network topology, where node relationships are fixed. This assumption breaks down in mobile environments where changing locations and connections make neighborhood-based consistency checks unreliable. Moreover, they are generally not designed for FL, lacking mechanisms to handle data privacy and communication constraints. They also require a synchronized learning process, which is impractical for mobile nodes that may join or leave the training process at

TABLE I
SUMMARY OF NOTATIONS

Sym.	Description	Sym.	Description	Sym.	Description
$\mathcal{S}, \mathcal{M}$	stable/mobile node sets	K	total number of nodes	T	total FL rounds
$\mathcal{U}$	modality set	μ, ν, ρ	modality indices	$\mathcal{P}$	paired modality set
$\mathbf{x}_m(t)$	position of node m	$v_m(t), u_m(t)$	speed/direction of node m	t	continuous time index
$\mathcal{D}_k$	local dataset of node k	$\mathcal{D}_{k,\text{filtered}}$	filtered dataset after PNCC	B_s	training batch size
$\mathbf{F}^{(\mu)}$	modality-μ feature matrix	n_μ	modal dimension	d	unified embedding dimension
$N=\sum_\mu n_\mu$	total feature components	$\mathbf{S}$	unified similarity matrix	$\mathbf{H}_{\mu\nu}$	cross-modal similarity
$\mathbf{h}_i^{(l)}$	node representation at layer l	$W_q^{(l)}, W_k^{(l)}, W_v^{(l)}$	query/key/value projections	L	number of GNN layers
$\mathbf{z}^{(\mu)}$	pooled representation of μ	λ_{att}	attention scaling parameter	$\boldsymbol{\theta}$	GNN adapter parameters
$\boldsymbol{\theta}^t_{\text{global}}$	global model at round t	$\boldsymbol{\theta}^t_k$	local model of node k	η_l	local learning rate
E	local epochs	$g_k(\theta;\xi)$	stochastic gradient	$\xi_k^{t,e}$	sampling randomness
$\mathcal{L}_k(\theta)$	local objective of node k	$\mathcal{L}(\theta)$	global objective	q_k	ideal data-proportional weight
σ^2	stoch. gradient variance bound	G	stoch. gradient norm bound	$\mathcal{F}_t$	Filtration
Ω	agg. weight variance bound	w_k^t	PNCC aggregation weight	Δ^t	PNCC aggregation deviation
$\mathcal{L}_{\text{inf}}$	global loss lower bound	ρ_k	pseudo-label noise rate		
c_i	sample confidence score	Δ	confidence gap	$d_i^{\mu\nu}$	diagonal similarity of sample i
$\bar{s}_i^{\text{off}}$	mean off-diagonal similarity	$r_i^{\mu\nu}$	discrimination ratio	$\mu_{c,k}, \sigma_{c,k}$	mean/std of of confidence
$\tau_{k,\text{adaptive}}$	adaptive filtering threshold	γ	threshold scaling factor	Q_k, Div_k	quality/diversity factors
CONF_k	node confidence score	ϵ	numerical stability constant	K_{eff}	effective node count $1/\sum_k w_k^2$
$\mathcal{L}^{\mu\nu}_{\text{triplet}}$	cross-modal triplet loss	$\mathcal{L}_{\text{sim}}$	similarity regularization loss	$\mathcal{L}_{\text{distill}}$	distillation loss
λ_{reg}	regularization weight	λ_{distill}	distillation loss weight	ζ	triplet loss margin
x, y	pre-train/collab rounds	$X_{\max}$	maximum total rounds	$\mathbf{R}$	total incentive payment pool
U_{s_i}, U_{m_j}	stable/mobile node utilities	A_i, B_j	base benefit coefficients	λ_i, μ_j	diminishing return rates
ϕ_{s_i}, η_{m_j}	stable weight / mobile efficiency	w_{s_i}, α_{m_j}	stable/mobile contribution wts.	ρ_b	pre-train/collab balance coeff.
δ_j	participation indicator	t_j	mobile node j staying time	t_γ	time cost per FL round
ω_{model}	model size in bits	B_k	allocated bandwidth	p_k	transmission power
h_k	channel gain	σ_0^2	channel noise power	τ_k^{comm}	communication delay
e_k^{comm}	communication energy cost	E_k^{comp}	computation energy cost	ζ_k, f_k	capacitance / CPU freq.
C_k	total CPU cycles	κ, d_u	FLOPs per sample / cycles		

different times. These limitations need a denoising approach specifically for FL in dynamic mobile networks.

In addition, the dynamic nature of MNs creates a sharp imbalance in FL training requirements between different types of nodes. Wu et al. [15] proposed the SAFA protocol, using asynchronous FL to solve heterogeneous training accuracy diversity between nodes. Sun et al. [23] developed SC-AFL to manage asynchronicity by controlling model accuracy, *e.g.* , staleness, to balance efficiency and accuracy. However, these methods are designed for asynchronous FL with model training differences rather than label noise differences between nodes. Furthermore, once a user's update delay exceeds a predefined threshold, their local model is considered deprecated. This strategy is particularly impractical for SC systems. The excluded node must revert to conventional bit-level transmission, sharply increasing network latency and load. Hence, a clear gap remains for an FL framework tailored to the distinct and imbalanced requirements of stable and mobile nodes.

Therefore, an FL framework is still urgently needed for SC in MNs to jointly consider the challenges of different modal, model heterogeneity, the absence of reliable labels, and training collaboration with disparate training requirements.

III. System Model

In this section, we describe the considered MN, followed by introducing the collaborative updating framework used for SC and the corresponding computing and communication model.

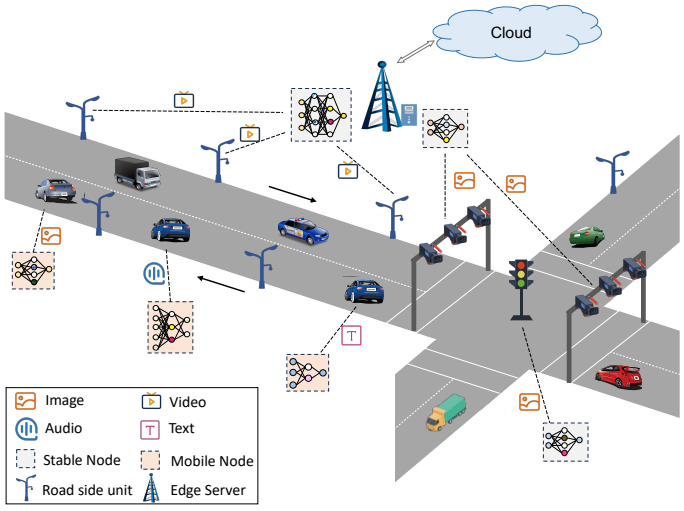


Fig. 2. Multiple modalities communication in MNs with heterogeneous stable and mobile nodes.

A. System Description

We consider an MN that employs the SC system, comprising of multiple client nodes and an edge server, as shown in Fig. 2. The transmitter of nodes in this network utilizes the semantic encoder to extract the essential meaning from the information to be transmitted (*e.g.*, images, audio, and text). This extracted semantic information is then encoded via the SKB to generate concise symbols for efficient transmission. At the receiver, a corresponding semantic decoder utilizes the shared SKB to

reconstruct the original information.

Furthermore, according to the velocities, nodes in MNs using SCs can be composed of two types:

- **Stable Nodes** ($\forall s \in \mathcal{S}$): Stable infrastructure units (*e.g.*, based stations and roadside units) with fixed positions ($v_S = 0$). These nodes maintain well-trained semantic codecs and high-quality local datasets tailored for this MN's scenario, ensuring reliable semantic encoding/decoding. They, however, operating in relatively stable environments, still require periodic encoder updates to adapt to subtle local distribution shifts during semantic extraction.
- **Mobile Nodes** ($\forall m \in \mathcal{M}$): Devices of mobile entities (*e.g.*, pedestrians and vehicles) with time-varying positions. Their mobility exposes them to changing environments, which degrade semantic extraction accuracy and disrupt SKB alignment. This necessitates real-time collaborative updating mechanisms for their semantic encoders to maintain communication robustness in different MNs.

The local dataset $\mathcal{D}_k$ at each node $k \in \mathcal{S} \cup \mathcal{M}$ is drawn from a node-specific distribution $P_k(\mathbf{x}, y)$, where $P_i \neq P_j$ for $i \neq j$, reflecting the non-IID nature of data across heterogeneous nodes. The Semantic Knowledge Base remains fixed during federated updates, i.e., $\mathrm{SKB}^{(t)} = \mathrm{SKB}^{(0)}, \forall t \in \{1, \ldots, T\}$, as SKB synchronization operates at a coarser timescale than encoder updates.

To model the dynamic behavior of mobile nodes in MNs, we employ the general definition of vehicle motion trajectory [24] as the trajectory of a mobile node m, which can be expressed by

$$x_m(t) = x_m(0) + \int_0^t v_m(\tau) u_m(\tau) d\tau, \tag{1}$$

where $x_m(t)$ is the position of the node at time t, $x_m(0)$ is its initial position, $v_m(\tau)$ is its speed at a given moment τ, and $u_m(\tau)$ represents its direction of movement.

B. Fl-based updating

The FL is a general and effective approach for collaborative updating of the semantic encoder in MNs [18]. It enables multiple nodes in MNs to collaboratively train an encoder without exchanging their local data. Specifically, each node independently computes an update to the current global model based on its local data and then communicates this update to a central server. The central server aggregates the updates from all local nodes to construct an improved global model, which is then sent back to the nodes for the subsequent training round [25]. This approach inherently preserves data privacy because the raw training data never leaves the local node.

In the considered MNs, stable nodes and mobile nodes can act as the student node, and the edge acts as the central server. The FL-based updating target can be denoted by

$$\mathcal{L}_{\mathrm{FL}}(\theta) = \sum_{k \in \mathcal{S} \cup \mathcal{M}} w_k \mathcal{L}_k(\theta), \tag{2}$$

where w_k reflects the relative importance of node k, typically proportional to local data volume. Because all local models are identical, FedAvg achieves this through weighted parameter averaging: $\theta_{\text{global}} = \sum_k w_k \theta_k$.

This straightforward aggregation of model parameters is only possible because each local model θ_k has the same size, structure, and parameter ordering. This requirement, however, poses a significant challenge in heterogeneous scenarios.

C. Training Overhead

During the FL-based updating, each participating node $k \in \mathcal{S} \cup \mathcal{M}$ incurs both computational overhead for local training and communication costs when uploading model parameters to the aggregator.

The computational cost can be denoted by

$$E_k^{\text{comp}} = P_k^{\text{comp}} \cdot T_k^{\text{comp}}, \tag{3}$$

where P_k^{comp} is the computational power and T_k^{comp} is the computational delay. The power P_k^{comp} is proportional to the cube of the frequency, we have

$$P_k^{\text{comp}} = \zeta_k f_k^3, \tag{4}$$

where ζ_k is a hardware-dependent effective capacitance coefficient and f_k is CPU frequency. Further, the computational delay T_k^{comp} can be expressed as:

$$T_k^{\text{comp}} = \frac{C_k}{f_k}, \tag{5}$$

where $C_k = |\mathcal{D}_k| \cdot \kappa \cdot E \cdot d_u$. Here, $|\mathcal{D}_k|$ is the number of data samples in the local dataset on node k, E is the number of local training epochs per round, κ is the number of operations per data sample for one forward and backward pass, and d_u is the number of CPU cycles required per operation.

The energy consumption E_k^{comp} is thus can be expressed as:

$$E_k^{\text{comp}} = P_k^{\text{comp}} \cdot T_k^{\text{comp}} = (\zeta_k f_k^3) . \left(\frac{C_k}{f_k} \right) = \zeta_k C_k f_k^2. \tag{6}$$

Similarly, communication overhead is communication power multiplied by transmission time. We have the transmission delay τ_k^{comm} as:

$$\tau_k^{\text{comm}} = \frac{\omega_{\text{model}}}{B_k \log_2(1 + \frac{p_k h_k}{\sigma_0^2})}, \tag{7}$$

where B_k, p_k, h_k and σ_0 are the bandwidth, transmission power, channel gain and noise power, respectively [26]. Therefore, the energy cost e_k^{comm} during updating can be expressed as:

$$e_k^{\text{comm}} = p_k \cdot \tau_k^{\text{comm}}. \tag{8}$$

IV. The Proposed Fed-MoSeC

In this section, the proposed SC updating framework, Fed-MoSeC is presented, which tailored for multiple modalities semantic encoder updates between heterogeneous nodes in MNs. We first introduce the GNN-based adapter design to mitigate encoder heterogeneity. We then describe the novel federated aggregation process for both mobile and stable nodes. Finally, we detail the integrated PNCC and TOC components.

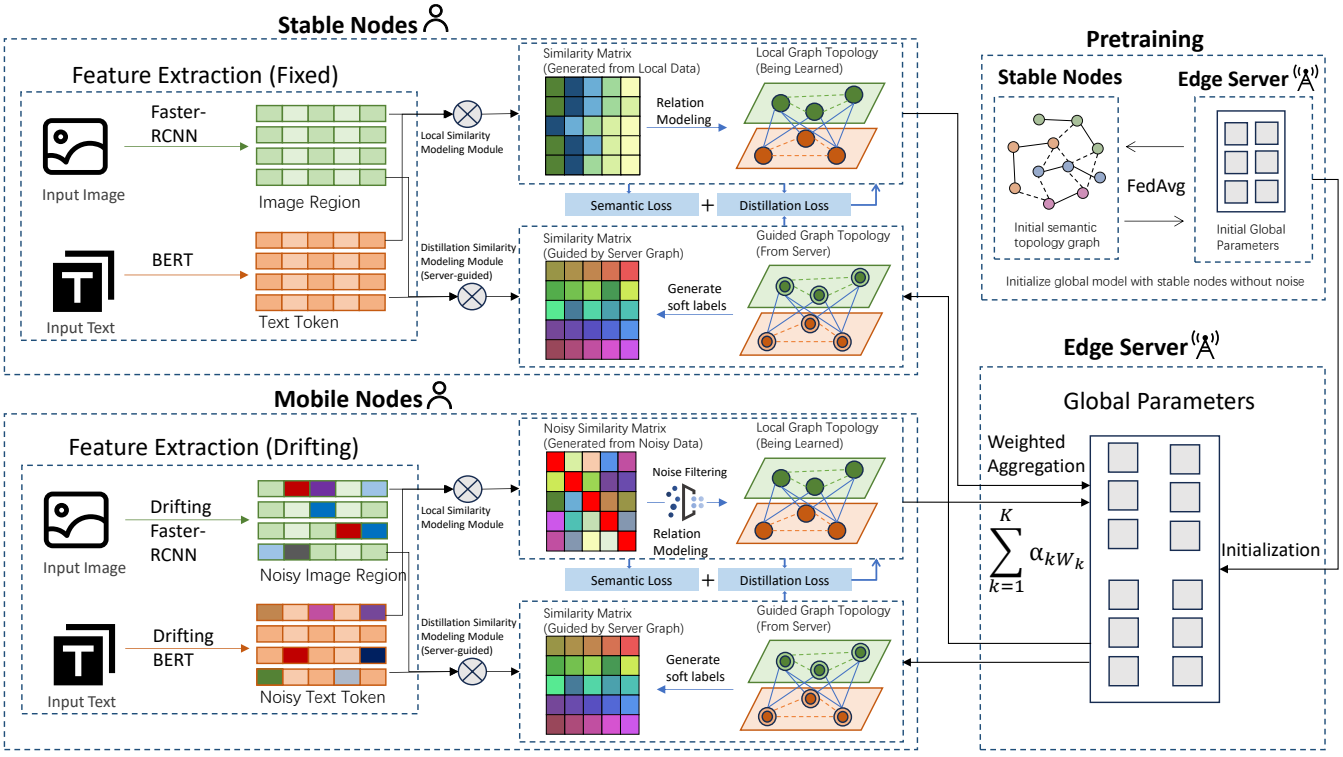


Fig. 3. Overview of Fed-MoSeC framework. Schematic of collaborative training process for stable and mobile nodes.

A. Framework Description

The Fed-MoSeC framework operates via a robust two-stage training process to ensure stability and adaptability, as illustrated in Fig. 3.

The first stage is pre-training and Initialization with stable nodes. This process begins with an early training phase involving only the stable nodes, which possess reliable, high-quality data. We first design a novel unified adapter to process heterogeneous encoder outputs from multimodal inputs, where only the adapter is trained and shared while keeping the original encoders frozen to maintain their pretrained representations. In this stage, stable nodes first train their local GNN adapters, after which they aggregate these adapters into a global model based on the decision from TOC. This strategy effectively leverages the collective knowledge from stable nodes to create a strong foundational adapter that serves as a reliable basis for subsequent system-wide adaptation.

The second stage is collaborative updating. After initialization, the mobile nodes are introduced into the network, and the framework enters its main collaborative training loop. The PNCC is employed in this stage for handling the data noise and environmental shifts associated with mobile nodes. Each round of FL commences with the server broadcasting the current global encoder parameters $\boldsymbol{\theta}_{\text{global}}$ and each node k then performs a sequence of local operations. It first filters its local data to generate a high-quality subset $\mathcal{D}_{k,\text{filtered}}$; it then extracts the target similarity matrix $\mathbf{S}_{\text{target}}^{(k)}$ by applying the received global parameters to this data. Subsequently, each node executes local training by optimizing the joint objective $\mathcal{L}_{\text{local}}$, which incorporates both task-specific learning and similarity-based distillation. Upon completion of local training, each node assesses its contribution by computing the confidence score CONF_k, which balances data quality with model diversity. Finally, each node transmits its updated parameters $\boldsymbol{\theta}_k$ and its confidence score to the server. The round concludes with the server performing confidence-weighted aggregation to produce the next iteration of the global model, thereby balancing contributions from local nodes with reliable data against those providing diverse knowledge patterns.

The procedure of the Fed-MoSeC is demonstrated in Algorithm 1.

B. GNN-based Adapter Design

To address heterogeneity across modalities and encoder architectures, Fed-MoSeC employs a novel GNN-based adapter. This adapter enables the transformation of heterogeneous features into a unified, homogeneous representation. Therefore, the SC encoders for different modality inputs retain their previously trained states frozen during federated training, requiring only the lightweight adapters to be federated learned.

On each node in the same MN, less precise features are extracted from multiple modalities using frozen pre-trained encoders. This is because the variation of the environment makes it difficult for frozen semantic encoders to extract accurate features. We consider $|\mathcal{U}|$ modalities, where modality $\mu \in \mathcal{U}$ is represented by feature matrix $\mathbf{F}^{(\mu)} \in \mathbb{R}^{n_\mu \times d}$ with rows $\{\mathbf{f}_1^{(\mu)}, \ldots, \mathbf{f}_{n_\mu}^{(\mu)}\}$. Here, d denotes the unified feature dimension, and n_μ represents the number of components (e.g., regions, tokens, frames) in modality μ.

Algorithm 1 Fed-MoSeC

Require: Stable node features $\{\{\mathbf{F}_s^{(\mu)}\}_{\mu\in\mathcal{U}} \mid s \in \mathcal{S}\}$, Mobile node features $\{\{\mathbf{F}_m^{(\mu)}\}_{\mu\in\mathcal{U}} \mid m \in \mathcal{M}\}$, Maximum rounds $X_{\max}$

Initialize: GNN adapters, global model $\boldsymbol{\theta}_{global}^0$
$(x^*, y^*, R^*) \leftarrow$ TOC($\mathcal{S}, \mathcal{M}, X_{max}$)
for all stable node $s \in \mathcal{S}$ **do**
  $R_s \leftarrow R^* \cdot w_s / \sum_{i\in\mathcal{S}} w_i$
end for
for all mobile node $m \in \mathcal{M}$ **do**
  $R_m \leftarrow R^* \cdot \alpha_m / \sum_{j\in\mathcal{M}} \alpha_j$
end for
for $r = 1$ to x^* **do**
  for all stable node $s \in \mathcal{S}$ **do**
    $\boldsymbol{\theta}_s^r \leftarrow$ GNN-AdapterUpdate($\boldsymbol{\theta}_s^{r-1}$, $\{\mathbf{F}_s^{(\mu)}\}_\mu$)
  end for
  $\boldsymbol{\theta}_{global}^r \leftarrow$ FedAvg($\{\boldsymbol{\theta}_s^r\}$)
end for
for $r = x^* + 1$ to $x^* + y^*$ **do**
  Server broadcast $\boldsymbol{\theta}_{global}^{r-1}$ to all nodes
  for all stable node $s \in \mathcal{S}$ (receives R_s) **do**
    $\mathbf{S}_{target}^{(s)} \leftarrow$ Similarity($\{\mathbf{F}_s^{(\mu)}\}_\mu, \boldsymbol{\theta}_{global}^{r-1}$)
    $\boldsymbol{\theta}_s^r \leftarrow$ GNN-AdapterUpdate($\{\mathbf{F}_s^{(\mu)}\}_\mu, \mathbf{S}_{target}^{(s)}$)
    CONF$_s \leftarrow$ ComputeConfidence($\{\mathbf{F}_s^{(\mu)}\}_\mu, \boldsymbol{\theta}_s^r$)
  end for
  for all mobile node $m \in \mathcal{M}$ (pays R_m) **do**
    $\{\mathbf{F}_{m,\text{filtered}}^{(\mu)}\}_\mu, \mathbf{S}_{target}^{(m)} \leftarrow$ PNCC($\{\mathbf{F}_m^{(\mu)}\}_\mu, \boldsymbol{\theta}_{global}^{r-1}$)
    $\boldsymbol{\theta}_m^r \leftarrow$ GNN-AdapterUpdate($\{\mathbf{F}_{m,\text{filtered}}^{(\mu)}\}_\mu, \mathbf{S}_{target}^{(m)}$)
    CONF$_m \leftarrow$ ComputeConfidence($\{\mathbf{F}_{m,\text{filtered}}^{(\mu)}\}_\mu, \boldsymbol{\theta}_m^r$)
  end for
  $\boldsymbol{\theta}_{global}^r \leftarrow \sum_{k\in\mathcal{S}\cup\mathcal{M}} \frac{\text{CONF}_k}{\sum_j \text{CONF}_j} \boldsymbol{\theta}_k^r$
end for
return $\boldsymbol{\theta}_{global}$

This adapter learns graph-node representations through iterative message passing over a unified multimodal graph. Here, each graph node corresponds to a modality-specific feature unit, such as an image region, a text token, or an audio segment. This process is based on our proposed unified similarity matrix $\mathbf{S} \in \mathbb{R}^{N\times N}$, where $N = \sum_{\mu\in\mathcal{U}} n_\mu$ is the total number of graph nodes across all modalities. The matrix is designed to jointly represent intra-modal and cross-modal relationships across all modalities and has a block structure denoted by

$$\mathbf{S} = \begin{bmatrix} \mathbf{S}_{11} & \mathbf{S}_{12} & \cdots & \mathbf{S}_{1|\mathcal{U}|} \\ \mathbf{S}_{21} & \mathbf{S}_{22} & \cdots & \mathbf{S}_{2|\mathcal{U}|} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{S}_{|\mathcal{U}|1} & \mathbf{S}_{|\mathcal{U}|2} & \cdots & \mathbf{S}_{|\mathcal{U}||\mathcal{U}|} \end{bmatrix}, \tag{9}$$

where diagonal blocks $\mathbf{S}_{\mu\mu} \in \mathbb{R}^{n_\mu\times n_\mu}$ capture intra-modal similarities within modality μ, and off-diagonal blocks $\mathbf{S}_{\mu\nu} \in \mathbb{R}^{n_\mu\times n_\nu}$ ($\mu \neq \nu$) represent cross-modal similarities between modalities μ and ν.

For intra-modal similarities, we introduce k-nearest neighbor constrained Gaussian kernels to maintain computational efficiency while preserving local neighborhood structures. For modality μ, the intra-modal similarity is computed as:

$$(\mathbf{S}_{\mu\mu})_{ij} = \begin{cases} \exp\left(-\dfrac{\|\mathbf{f}_i^{(\mu)} - \mathbf{f}_j^{(\mu)}\|_2^2}{\sigma_b^2}\right), & j \in \mathcal{N}_k(\mathbf{f}_i^{(\mu)}), \\ 0, & \text{otherwise}, \end{cases} \tag{10}$$

where σ_b denotes the bandwidth parameter and $\mathcal{N}_k(\mathbf{f}_i^{(\mu)})$ represents the k-nearest neighbors of feature $\mathbf{f}_i^{(\mu)}$ within the same modality.

For cross-modal similarities, we distinguish between two types of modality pairs. For directly paired modalities, *i.e.*, modalities with ground-truth correspondences in training data, we calculate the raw similarity matrix $\mathbf{H}_{\mu\nu} = (\mathbf{F}^{(\mu)})^T\mathbf{F}^{(\nu)}$, and then obtain the cross-modal similarity through learned transformations combining top-k selection and mean pooling operations:

$$\mathbf{S}_{\mu\nu} = \text{MLP}(\text{TopK}(\mathbf{H}_{\mu\nu})) + \text{MeanPool}(\mathbf{H}_{\mu\nu}). \tag{11}$$

For indirectly related modality pairs, *i.e.*, modalities without direct supervision, such as text-speech when only image-text and image-speech pairs exist, we bridge the semantic domains of disjoint modalities by aligning them with a shared intermediary modality. Specifically, if modality ρ serves as a bridge connecting modalities μ and ν, we have:

$$\mathbf{S}_{\mu\nu}^{\text{bridge}} = \mathbf{S}_{\mu\rho} \cdot \mathbf{S}_{\rho\nu}. \tag{12}$$

By leveraging a common modal anchor as a bridge, we integrate disparate modalities into a unified embedding space, enabling cross-modal retrieval even without directly supervised data. The symmetry constraint ensures $\mathbf{S}_{\mu\nu} = \mathbf{S}_{\nu\mu}^T$ for all modality pairs.

Using the constructed similarity matrix, we build a heterogeneous graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where nodes $\mathcal{V} = \{\mathbf{f}_1^{(1)}, \ldots, \mathbf{f}_{n_1}^{(1)}, \ldots, \mathbf{f}_1^{(|\mathcal{U}|)}, \ldots, \mathbf{f}_{n_{|\mathcal{U}|}}^{(|\mathcal{U}|)}\}$ include features from all modalities. The similarity matrix $\mathbf{S}$ is then incorporated into a scaled attention mechanism for message passing, formulated as follows:

$$\mathbf{h}_i^{(l+1)} = \sum_j \text{softmax}\left(\frac{(\mathbf{W}_q^{(l)}\mathbf{h}_i^{(l)})^T(\mathbf{W}_k^{(l)}\mathbf{h}_j^{(l)})}{\sqrt{d}} + \lambda_{att}\mathbf{S}_{ij}\right) \times \mathbf{W}_v^{(l)}\mathbf{h}_j^{(l)}, \tag{13}$$

where $\mathbf{W}_q^{(l)}$, $\mathbf{W}_k^{(l)}$, and $\mathbf{W}_v^{(l)}$ are learnable query, key, and value projection matrices respectively, λ_{att} is a scaling parameter that balances learned attention scores with similarity-based weights. $\mathbf{h}_i^{(0)}$ corresponds to the initial node features $\mathbf{f}_i^{(\mu)}$ for the node from modality μ.

After L layers of feature aggregation, global representations for each modality are obtained through pooling. For modality μ, we compute:

$$\mathbf{z}^{(\mu)} = \text{Pool}(\{\mathbf{h}_i^{(L)}\}_{i\in\mathcal{I}_\mu}), \tag{14}$$

where $\mathcal{I}_\mu$ denotes the index set of nodes belonging to modality μ.

The training objective combines triplet alignment across all paired modalities and similarity regularization to ensure effective cross-modal learning:

$$\mathcal{L} = \sum_{(\mu,\nu)\in\mathcal{P}} \mathcal{L}_{\text{triplet}}^{\mu\nu} + \mathcal{L}_{\text{sim}}, \tag{15}$$

where $\mathcal{P}$ denotes the set of directly paired modality combinations available in the training data. For each paired modality combination (μ, ν), the triplet loss promotes cross-modal alignment by maximizing similarity between paired samples while minimizing similarity with negative samples:

$$\begin{aligned}\mathcal{L}_{\text{triplet}}^{\mu\nu} = \frac{1}{B_s}\sum_{i=1}^{B_s}[\zeta - \cos(\mathbf{z}_i^{(\mu)}, \mathbf{z}_i^{(\nu)}) + \cos(\mathbf{z}_i^{(\mu)}, \mathbf{z}_i^{(\nu)-})]_+ \\ + [\zeta - \cos(\mathbf{z}_i^{(\mu)}, \mathbf{z}_i^{(\nu)}) + \cos(\mathbf{z}_i^{(\mu)-}, \mathbf{z}_i^{(\nu)})]_+,\end{aligned} \tag{16}$$

where B_s denotes the batch size, $\cos(\cdot,\cdot)$ represents cosine similarity, ζ is the margin parameter, $[\cdot]_+ = \max(\cdot, 0)$, $(\mathbf{z}_i^{(\mu)}, \mathbf{z}_i^{(\nu)})$ denotes a positive pair from the same training instance, while $(\mathbf{z}_i^{(\mu)}, \mathbf{z}_i^{(\nu)-})$ and $(\mathbf{z}_i^{(\mu)-}, \mathbf{z}_i^{(\nu)})$ denote hardest non-matching cross-modal pairs selected from the same mini-batch. The similarity regularization term enforces matrix consistency by penalizing asymmetry in cross-modal similarities:

$$\mathcal{L}_{\text{sim}} = \sum_{\mu=1}^{|\mathcal{U}|-1}\sum_{\nu=\mu+1}^{|\mathcal{U}|} \|\mathbf{S}_{\mu\nu} - \mathbf{S}_{\nu\mu}^T\|_F^2, \tag{17}$$

where $\|\cdot\|_F$ denotes the Frobenius norm.

C. Fed-MoSeC: PNCC

A challenge in mobile FL is the conflict between data diversity and data quality. While mobile nodes joining the network bring valuable data from diverse environments, this data is often corrupted by significant pseudo-label noise. Directly aggregating noisy updates from mobile nodes can degrade the global model, potentially negating the benefit of their diverse data. In contrast, stable nodes possess well-trained encoder adapters and higher-quality local data, providing a source of reliable knowledge for the system. This necessitates a mechanism to leverage the diversity of mobile nodes while protecting the integrity of the global model from noise.

To address this, we propose PNCC, which designs the local updating and global aggregation processes to create a robust teacher-student knowledge distillation. In this process, the 'teacher' is the global model, which is initially aggregated from the well-trained stable nodes. Subsequently, all participating nodes, including both the noisy mobile nodes and the stable nodes, act as 'students' that distill knowledge from this teacher. This framework enables students to effectively learn from the teacher model, while the final aggregation process intelligently weighs each node's contribution based on both its data quality and model diversity. The PNCC operates through the following integrated steps.

1) Local Nodes Side Confidence-Based Filtering: To enhance local data quality, each node k utilizes the cross-modal similarity matrix $\mathbf{S}_{k,\mu\nu}$, generated by its local model $\boldsymbol{\theta}_k$, to assess the reliability of its training samples. For each cross-modal pair (i, i) in a batch of size B, the diagonal similarity is defined as:

$$d_i^{\mu\nu} = (\mathbf{S}_{k,\mu\nu})_{ii}. \tag{18}$$

The discriminative strength of this match is quantified by the ratio r_i, which contrasts the diagonal similarity against the average off-diagonal similarity:

$$r_i^{\mu\nu} = \frac{d_i^{\mu\nu}}{\bar{s}_i^{\text{off}}}, \quad \text{where} \quad \bar{s}_i^{\text{off}} = \frac{1}{B-1}\sum_{j\neq i}(\mathbf{S}_{k,\mu\nu})_{ij}. \tag{19}$$

A composite confidence score c_i is then computed for each sample:

$$c_i = d_i^{\mu\nu} + r_i^{\mu\nu}. \tag{20}$$

Given the data heterogeneity inherent in FL, we employ an adaptive threshold $\tau_{k,\text{adaptive}}$ derived from the batch-wise confidence distribution:

$$\tau_{k,\text{adaptive}} = \mu_{c,k} + \gamma \cdot \sigma_{c,k}, \tag{21}$$

where $\mu_{c,k}$ and $\sigma_{c,k}$ are the mean and standard deviation of the confidence scores, and γ is a sensitivity parameter. The filtered training set $\mathcal{D}_{k,\text{filtered}}$ is thus formed by retaining samples whose confidence exceeds this threshold:

$$\mathcal{D}_{k,\text{filtered}} = \{i : c_i \geq \tau_{k,\text{adaptive}},\ i \in \{1, 2, \ldots, B\}\}. \tag{22}$$

2) Parameter-Driven Similarity Matrix Extraction: The target similarity matrix $\mathbf{S}_{\text{target}}^{(k)}$ for knowledge distillation is extracted by applying the global model parameters $\boldsymbol{\theta}_{\text{global}}$ to the node's local filtered data, $\{\mathbf{F}_k^{(\mu)}\}_{\mu\in\mathcal{U}}$. This process is deterministic:

$$\mathbf{S}_{\text{target}}^{(k)} = \text{Similarity}(\{\mathbf{F}_k^{(\mu)}\}_{\mu\in\mathcal{U}}; \boldsymbol{\theta}_{\text{global}}). \tag{23}$$

where the Similarity$(\cdot)$ function constructs the matrix blocks following the framework defined in Section III-C, yielding:

$$\mathbf{S}_{\text{target}}^{(k)} = \left[\mathbf{S}_{\mu\nu}^{(\text{global})}\right]_{\mu,\nu\in\mathcal{U}} \in \mathbb{R}^{N_k\times N_k}, \tag{24}$$

where $N_k = \sum_{\mu\in\mathcal{U}} n_{\mu,k}$ is the total number of components across all modalities at node k. This matrix encodes the teacher global model's learned relational patterns and serves as the objective for local model alignment.

3) Local Similarity-Based Distillation: Each student node k initializes its local model with the global parameters, $\boldsymbol{\theta}_k^{(0)} = \boldsymbol{\theta}_{\text{global}}$, and performs joint optimization. The local objective combines a task-specific loss with a similarity distillation loss:

$$\mathcal{L}_{\text{local}} = \mathcal{L}_{\text{task}} + \lambda_{\text{distill}}\mathcal{L}_{\text{distill}}. \tag{25}$$

The task loss, applied to the filtered dataset $\mathcal{D}_{k,\text{filtered}}$, is composed of contrastive and similarity regularization objectives as:

$$\mathcal{L}_{\text{task}} = \mathcal{L}_{triplet}(\mathcal{D}_{k,\text{filtered}}) + \mathcal{L}_{\text{sim}}(\mathcal{D}_{k,\text{filtered}}). \tag{26}$$

The distillation loss enforces alignment between the local model's similarity matrix $\mathbf{S}_k^{(t)}$ and the target matrix $\mathbf{S}_{\text{target}}^{(k)}$ at each local training iteration t:

$$\mathcal{L}_{\text{distill}} = \left\|\mathbf{S}_k^{(t)} - \mathbf{S}_{\text{target}}^{(k)}\right\|_F^2, \tag{27}$$

where $\mathbf{S}_k^{(t)} = \text{Similarity}\left(\{\mathbf{F}_k^{(\mu)}\}_{\mu\in\mathcal{U}}; \boldsymbol{\theta}_k^{(t)}\right)$. This loss acts as a regularizer, penalizing deviations from the teacher model's relational patterns during local adaptation.

4) Diversity-Aware Confidence Computation: To guide the aggregation process, each node computes a confidence score reflecting its contribution. This score, CONF_k, linearly interpolates a quality factor Q_k and a diversity factor Div_k:

$$\text{CONF}_k = Q_k + Div_k. \tag{28}$$

The quality factor Q_k is the fraction of samples retained after filtering, directly measuring data cleanliness:

$$Q_k = \frac{|\mathcal{D}_{k,\text{filtered}}|}{|\mathcal{D}_{k,\text{total}}|}. \tag{29}$$

The diversity factor Div_k quantifies the model's normalized divergence from the global model after local training. It is defined as:

$$Div_k = \frac{\sigma_k}{\sigma_k + \epsilon}, \quad \text{where} \quad \sigma_k = \frac{\left\|\mathbf{S}_k^{\text{final}} - \mathbf{S}_{\text{target}}^{(k)}\right\|_F}{\left\|\mathbf{S}_{\text{target}}^{(k)}\right\|_F}, \tag{30}$$

where $\mathbf{S}_k^{\text{final}}$ is the similarity matrix from the node's fully trained local model and ϵ is a small constant for numerical stability. A higher Div_k value signifies a greater contribution of unique local knowledge.

5) Confidence-Weighted Federated Aggregation: The server performs confidence-weighted aggregation of updated parameters. Upon receiving the updated parameters $\{\boldsymbol{\theta}_k\}$ and confidence scores $\{\text{CONF}_k\}$ from all K nodes, the server updates the global model as follows:

$$\boldsymbol{\theta}_{\text{global}}^{(t+1)} = \sum_{k=1}^{K} w_k \boldsymbol{\theta}_k^{(t)}, \tag{31}$$

where the aggregation weights $\{w_k\}$ are derived by normalizing the confidence scores:

$$w_k = \frac{\text{CONF}_k}{\sum_{j=1}^{K} \text{CONF}_j}. \tag{32}$$

D. Fed-MoSeC: TOC

We observe that nodes in the system exhibit significant heterogeneity in terms of computing and communication capabilities, as well as varying mobility levels. During the update process, different nodes experience diverse utility gains and incur varying training costs, leading to inconsistent willingness to participate in the training. Especially, in the PNCC, stable nodes in teacher-student knowledge distillation require several additional updating rounds to assist mobile nodes in suppressing pseudo-label noise. Paradoxically, stable nodes only need minor updates, but require more training rounds than mobile nodes, which require major updates. It is necessary for mobile nodes to incentivize stable users to update the encoder collaboratively. We hence propose TOC to be integrated into Fed-MoSeC to ensure the smooth operation of the Fed-MoSeC. First, we investigate the utilities of stable nodes and mobile nodes.

1) Utility Function of Stable Nodes: Generally, the utility of updating can be denoted by the updating benefit minus the cost. The utility for the stable node i can hence be expressed as:

$$\begin{aligned} U_{s_i}(x, y, \mathbf{R}) = \phi_{s_i} A_i \left(1 - e^{-\lambda_i f_i(x,y)}\right) + R_i \\ -(E^{\text{comp}} + e^{\text{comm}})(x + y), \end{aligned} \tag{33}$$

where $\phi_{s_i} A_i \left(1 - e^{-\lambda_i f_i(x,y)}\right)$ is the updating benefit; R_i is the incentive reward from mobile nodes to incentive updating participation; $(E^{\text{comp}} + e^{\text{comm}})(x + y)$ is the updating cost. Specifically, x and y are the decision variables representing the number of pre-training and collaboration rounds, respectively. The term R_i denotes the individual reward for stable node i, distributed from the total incentive payment pool $\mathbf{R}$ paid by mobile nodes. Furthermore, ϕ_{s_i} is the weight parameter of stable node i; A_i is the base benefit coefficient; λ_i is the learning sensitivity parameter; and E^{comp} and e^{comm} are the computational and communication energy costs, respectively.

Notably, we use the negative exponential function $\phi_{s_i} A_i \left(1 - e^{-\lambda_i f_i(x,y)}\right)$ to capture the benefit, with a theoretical upper bound of $\phi_{s_i} A_i$. This choice is motivated by the typical convergence dynamics in FL. The accuracy of the aggregated global model is known to exhibit a trend of rapid initial improvement that slows as it approaches a performance plateau, a behavior that is well-captured by the saturating nature of the negative exponential function.

Here, $f_i(x, y)$ is the accuracy auxiliary function according to epoch x and y, we have

$$f_i(x, y) = x \sum_{k\in\mathcal{S}} w_{s_k} + \rho_b \cdot y \cdot \sum_{j\in\mathcal{M}} \alpha_{m_j}, \tag{34}$$

where $x \sum_{k\in\mathcal{S}} w_{s_k}$ represents the contribution from pre-training updates, scaled by the collective weight of all stable nodes w_{s_k}; $\rho_b \cdot y \cdot \sum_{j\in\mathcal{M}} \alpha_{m_j}$ represents the contribution from cross-node collaboration, scaled by the weights of all mobile nodes α_{m_j}; ρ_b is a coefficient to balance the two types of contributions.

For the stable node, the updating objective is to maximize its utility function. Thus, the optimization problem for the stable nodes is

P1:

$$\max_{x,y} \quad U_{s_i}(x, y, \mathbf{R}^*(x, y)) \tag{35a}$$

$$\text{s.t.} \quad x \in \mathbb{Z}_+, \quad y \in \mathbb{Z}_+, \tag{35b}$$

$$x + y \le X^{\max}, \tag{35c}$$

$$y \le \min_{j\in\mathcal{M}} \left\{\frac{t_j}{t_\gamma}\right\}, \tag{35d}$$

$$\sum_{i\in\mathcal{S}} R_i = \sum_{j\in\mathcal{M}} R_j. \tag{35e}$$

The constraint (35b) shows that the decision variables, pre-training rounds x and collaboration rounds y, must be non-negative integers. The constraint (35c) bounds the total workload by the stable node's maximum resource capacity. The constraint (35d) ensures that the collaboration rounds y do not exceed the capacity of the most constrained mobile node, where t_j is the node's staying time and t_γ is the time

per round. The constraint (35e) is a budget balance equation, ensuring the total payments made by mobile nodes equal the total rewards received by stable nodes.

2) Utility Function of Mobile Nodes: Similarly, the utility for mobile node j can be denoted as its updating benefit minus the incentive payment and its own energy cost. The utility can hence be expressed as:

$$U_{m_j}(R_j; x, y) = t_j \eta_{m_j} B_j \left(1 - e^{-\mu_j g_j(x,y)}\right) - R_j - (E^{\text{comp}} + e^{\text{comm}})y, \tag{36}$$

where $t_j \eta_{m_j} B_j \left(1 - e^{-\mu_j g_j(x,y)}\right)$ is the updating benefit, which is also formulated via the negative exponential function, with a theoretical upper bound of $t_j \eta_{m_j} B_j$. Here, t_j is the time of mobile node staying in this system j, η_{m_j} is its learning efficiency, B_j is the base benefit coefficient, μ_j is the learning sensitivity parameter, and $g_j(x, y)$ is the accuracy auxiliary function. Unlike stable nodes, the accuracy auxiliary function $g_j(x, y)$ is different from $f_i(x, y)$, as mobile nodes are students. We have

$$g_j(x, y, R_j) = g_j(x, y) \cdot q_j(R_j), \tag{37}$$

where $g_j(x, y) = \beta_1 \sum_{i \in \mathcal{S}} w_{s_i} x + \beta_2 y$ represents the base learning contribution from pre-training and collaboration, with β_1 and β_2 as weight coefficients. The quality scaling factor $q_j(R_j)$ captures the effect of incentive payment on learning quality and is defined as:

$$q_j(R_j) = 1 + \nu_j \ln\left(1 + \frac{R_j}{c_j}\right), \tag{38}$$

where ν_j is a payment sensitivity parameter and c_j is a normalization constant. R_j is the incentive payment provided to stable nodes for their participation, and $(E^{\text{comp}} + e^{\text{comm}})y$ is the mobile node's energy cost.

The optimization problem for the mobile nodes is also to maximize their utilities, which can be denoted by

P2:

$$\max_{R_j} \quad U_{m_j}(R_j; x, y), \tag{39a}$$

$$\text{s.t.} \quad R_j \geq R_j^{\min} \cdot \delta_j, \tag{39b}$$

$$R_j \leq R_j^{\max} \cdot \delta_j, \tag{39c}$$

$$\delta_j \in \{0, 1\}, \tag{39d}$$

where δ_j is the participation indicator. The constraints (38b), (38c), and (38d) jointly define the payment rules that if a node participates ($\delta_j = 1$), its payment R_j is bounded between a minimum threshold $R_j^{\min}$ and a maximum cap $R_j^{\max}$. If the node does not participate ($\delta_j = 0$), these constraints enforce that the payment is zero, thus ensuring only participants make a payment.

3) bi-level Cournot-Stackelberg game theoretical algorithm: The optimization problems for the stable (P1) and mobile (P2) nodes are coupled. The solution for P1 requires the mobile nodes' decisions on payment $\mathbf{R}$ and collaboration y, while the solution for P2 requires the stable nodes' decision on pre-training x. This structure forms a complex, coupled MINLP problem. Hence, we propose a bi-level game theoretical algorithm based on the Cournot-Stackelberg game.

We first relax the integer constraints by allowing the stable nodes' decision variables, x and y, to take any non-negative real value, effectively changing their domain from $\mathbb{Z}_+$ to $\mathbb{R}_+$ and transforming the MINLP into a continuous non-linear problem. The utility functions are continuous and differentiable, and the strategy sets are compact, non-empty, and convex. Therefore, a Subgame-Perfect Nash Equilibrium (SPNE) that satisfies both inter-group and intra-group objectives is guaranteed to exist.

Our proposed algorithm decomposes the problem into two main stages, *i.e.,* an inter-group bargaining game to set the total payment, followed by an intra-group Stackelberg game to determine the effort levels.

In the first stage, we model the interaction between the set of stable nodes $\mathcal{S}$ and the set of mobile nodes $\mathcal{M}$ as a Cournot game to decide the terms of their cooperation. For any given total incentive payment $\mathbf{R}$ that could be transferred from $\mathcal{M}$ to $\mathcal{S}$, the system coordinates the effort levels (x, y) through a Stackelberg subgame, which results in a unique Subgame-Perfect Nash Equilibrium (SPNE). Once this total payment $\mathbf{R}$ is determined, it is allocated within each group. The stable nodes receive individual rewards R_i, and the mobile nodes provide individual payments R_j, with the allocation in both groups being proportional to the members' respective contribution weights.

Given the agreed-upon total payment $\mathbf{R}^*$, the two groups play a Stackelberg game to determine the optimal effort levels. This leader-follower game is solved using backward induction.

The mobile nodes, acting as a collective follower, observe the leader's pre-training effort x. Their goal is to choose a common collaboration level y and individual payments R_j to maximize their joint utility, knowing they must pay $\mathbf{R}^*$ in total. With payments fixed, the problem reduces to finding a common y that maximizes the sum of utilities. As the joint utility function $\sum U_{m_j}$ is concave with respect to y, the optimal common response $y^*(x)$ is found by solving the collective first-order condition $\frac{\partial}{\partial y}\left(\sum_{j \in \mathcal{M}} U_{m_j}\right) = 0$

$$\sum_{j \in \mathcal{M}} \left(t_j \eta_{m_j} B_j \mu_j \beta_2 e^{-\mu_j g_j(x,y)}\right) - \sum_{j \in \mathcal{M}} (E^{\text{comp}} + e^{\text{comm}}) = 0. \tag{40}$$

The total payment $\mathbf{R}^*$ is then sourced from the individual mobile nodes' payments.

The stable nodes, acting as the leader, anticipate the followers' response $y^*(x)$. They choose the optimal pre-training effort x^* to maximize their utility, knowing they will receive a total reward of $\mathbf{R}^*$. Their optimization problem is:

$$\max_{x} \quad U_{s_i}(x, y^*(x), \mathbf{R}^*) \tag{41}$$

We employ a numerical gradient ascent algorithm to solve this optimization. The algorithm is initialized with a feasible starting point x^0, a learning rate α, and a convergence tolerance ϵ. At each step k, we approximate the gradient numerically:

$$\frac{dU_{s_i}}{dx} \approx \frac{U_{s_i}(x^k + h) - U_{s_i}(x^k)}{h}, \tag{42}$$

where h is a small step size. Note that each function evaluation $U_{s_i}(x)$ necessitates first solving for the followers' optimal

common response $y^*(x)$ from the collective first-order condition, and then determining the individual payments $R_j^*(x)$. Stable node then updates its strategy:

$$x^{k+1} = x^k + \alpha \cdot \frac{dU_{s_i}}{dx}(x^k). \tag{43}$$

The individual rewards R_i are distributed from the total $\mathbf{R}^*$ proportionally to the stable nodes' weights.

Finally, we recover an integer solution from the continuous optimum (x^*, y^*). We explore the set of all integer points $(\tilde{x}, \tilde{y})$ in neighborhood of the continuous solution, defined as $\mathcal{C} = \{(x, y) \in \mathbb{Z}_+^{|\mathcal{S}|+1} : \|(x, y) - (x^*, y^*)\|_\infty \leq 1\}$.

For all resulting feasible candidates, the leader's utility $U_{s_i}(\tilde{x}, \tilde{y}, \mathbf{R}^*(\tilde{x}, \tilde{y}))$ is evaluated, a process which requires computing the mobile nodes' corresponding optimal responses. The feasible integer point that yields the maximal utility is selected as the final solution:

$$(x^{\text{int}}, y^{\text{int}}) = \arg \max_{(\tilde{x}, \tilde{y}) \in \mathcal{C}} U_{s_i}(\tilde{x}, \tilde{y}, \mathbf{R}^*(\tilde{x}, \tilde{y})). \tag{44}$$

The resulting integer profile together with $\mathbf{R}$ constitutes a Subgame-Perfect Nash Equilibrium (SPNE) of the Cournot-bargaining game: $\{(x^{\text{int}}, y^{\text{int}}), \mathbf{R}^*(x^{\text{int}}, y^{\text{int}})\}$. Within each set, the side payment R is shared proportionally to the members' weights, ensuring intra-set Pareto efficiency.

V. Performance Results and Analysis

In this section, we present the simulation settings and illustrate the superior performance of the proposed Fed-MoSeC.

A. Experimental Setting

To evaluate our proposed framework Fed-MoSeC, we emulate a realistic MN with 10 nodes and one edge server. The experiments were executed in a Lancaster University HEC GPU cluster using NVIDIA V100 (32 GB) and L40 (48 GB) GPU nodes through the gpu-short / gpu-medium queues.

We evaluate both SC performance and FL convergence efficiency by the RSUM metric. RSUM is calculated as the sum of R@1, R@5, and R@10 for both image-to-text and text-to-image retrieval. This metric, R@K (Recall@K), confirms whether the correct target is found within the top-K retrieved items. The total sum therefore provides a comprehensive assessment of encoder retrieval capability.

B. Ablation Study on Cross-Modal GNN Framework

This section validates the proposed cross-modal GNN adapter through ablation studies from multiple perspectives, including the visual and textual feature extractors, the use of fine-tuning versus frozen encoders, and different graph modeling strategies. Specifically, we compare seven variants: single-modality baselines (CNN-only; BERT-only), cross-modal baselines without our adapter (CNN+GNN; BERT+GNN), partial integration variants that use our adapter while still fine-tuning the encoders (CNN+Fed-MoSeC; BERT+Fed-MoSeC), and our Fed-MoSeC, which freezes the pre-trained encoders and only updates the lightweight adapter.

We evaluate on the Flickr30K dataset [27], which is designed for cross-modal SC tasks. This dataset contains 31,000 images with 158,000 descriptive captions, following the standard split of 29,000 images for training, 1,000 for validation, and 1,000 for testing. Each image is paired with five descriptive captions, enabling cross-modal learning for SC in mobile networks.

The network architecture extracts features from visual and textual modalities. For visual feature extraction, we employ a Faster R-CNN with ResNet-101 [28] backbone, extracting 36 region features of dimension 2048. Textual features are extracted using BERT-base-uncased [29], producing token-level features of dimension 768. Both modalities are aligned through linear projection layers to a unified dimension $d = 1024$. The batch size $B_s = 128$, GNN layer $L = 1$, the hyper-parameters as $k = 10$, $\sigma_b = 1.0$, Top-$K = 10$, $\lambda_{att} = 1.5$ for attention scaling, and margin of triplet loss $\zeta = 0.2$. We use max pooling for global feature aggregation and the Adam optimizer with a learning rate $1e-4$.

Our experiments evaluate robustness in two scenarios: a stable node and a mobile node. Stable node evaluation tests performance using clean, high-quality data. For mobile node evaluation, we add 30%-50% random outliers to simulate conditions faced by new mobile nodes experiencing data distribution shifts.

Figure 4(a) shows the performance from the stable node perspective. Our Fed-MoSeC framework, alongside the CNN+Fed-MoSeC and BERT+Fed-MoSeC variants, reaches approximately 475 RSUM after convergence. This indicates that our approach maintains SC quality while offering significant computational and communication advantages. In particular, all three federated approaches outperform their corresponding GNN counterparts (CNN+GNN: 462.0, BERT+GNN: 462.9), which validates the effectiveness of our graph-structured framework over conventional fusion strategies. In addition, the comparison of results shows a clear advantage for cross-modal processing, as all cross-modal approaches significantly outperform the single-modality baselines (CNN: 451.5, BERT: 456.6). This result confirms that joint visual-textual semantic learning provides benefits for mobile network applications with diverse data modalities.

From the mobile node perspective, as shown in Figure 4(b), the robustness of our FL approach is more apparent under noisy conditions. While all methods experience performance degradation, Fed-MoSeC maintains a higher resilience with approximately 440 RSUM. The performance gap to other methods widens in the presence of noise, with traditional approaches dropping to the 380-420 range.

Our Fed-MoSeC framework achieves these competitive results with significantly lower computational and communication overhead than encoder-updating approaches. As detailed in Table II, communication cost is quantified by the total size of transmitted parameters, assuming each is a standard 32-bit floating-point number (4 bytes). Unlike methods that require transmitting large pre-trained model parameters (*e.g.*, $\sim 240MB$ for CNN or $\sim 440MB$ for BERT), our approach only communicates the lightweight GNN adapter weights. This design reduces communication costs by over an order of magnitude to just $\sim 12MB$ per round, while maintaining a comparable computational complexity of $O(N^2)$ and

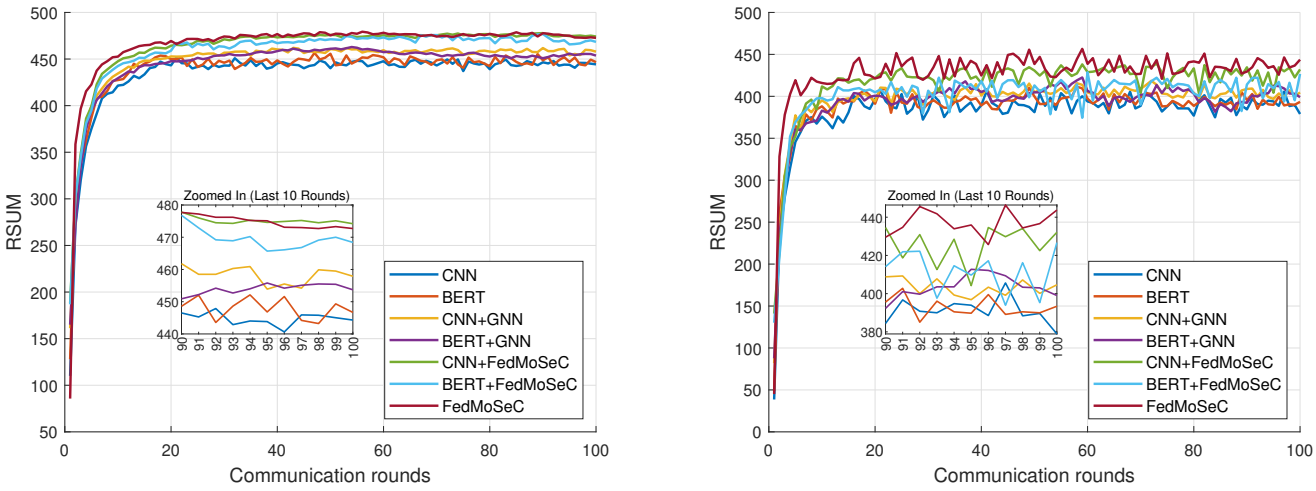


Fig. 4. Performance comparison of cross-modal SC approaches under different node conditions.

achieving similar semantic accuracy. Such efficiency makes our framework particularly suitable for resource-constrained mobile environments where bandwidth and energy conservation are critical.

TABLE II
COMPUTATIONAL AND COMMUNICATION OVERHEAD COMPARISON

Method	Communication Cost (MB/round)	Computational Complexity
CNN	∼240	O(N)
BERT	∼440	$O(N^2)$
CNN+GNN	∼248	$O(N^2)$
BERT+GNN	∼448	$O(N^2)$
CNN+Fed-MoSeC	∼252	$O(N^2)$
BERT+Fed-MoSeC	∼452	$O(N^2)$
Fed-MoSeC	∼12	$O(N^2)$

C. Game Theoretic Analysis

To validate the efficacy of our proposed TOC, we conduct a game-theoretic simulation. This experiment is designed to show the optimal resource allocation strategy for FL by analyzing the relationship between system utility and the number of training rounds.

Figure 5 shows the relationship between system utility $U_{sys}(x, y)$ and the number of pre-training rounds conducted exclusively by stable nodes x, the number of subsequent collaborative training rounds involving both stable and mobile nodes y. In this experiment, we configure a network with 5 stable nodes and 5 mobile nodes, where the computational and communication costs per round are set to 0.8 and 0.4 units, respectively. The analysis of this utility surface reveals that the optimal equilibrium is achieved at $x^* = 5.219$ and $y^* = 102.754$. This theoretical finding guides our selection of hyperparameters for the main FL experiments. Consequently, we set the number of pre-training rounds to 5 and the number of collaborative rounds to 100, thereby ensuring that the system operates in a state of near-maximal efficiency.

Figure 6 details the resulting incentive distribution at this optimal strategy, showing the reward R_i allocated to each stable node i and the corresponding incentive payment R_j from each mobile node j. These allocations are budget-balanced, satisfying $\sum R_i = \sum R_j$. The variations in individual rewards and payments reflect the simulated heterogeneity of the nodes. For instance, the local data quality of stable nodes were randomized between 0.5 and 0.7, mobile nodes participation times were randomized between 30 and 70 units. This resulting differentiated allocation thus validates the TOC's ability to establish a fair, contribution-based reward mechanism for nodes with varying characteristics.

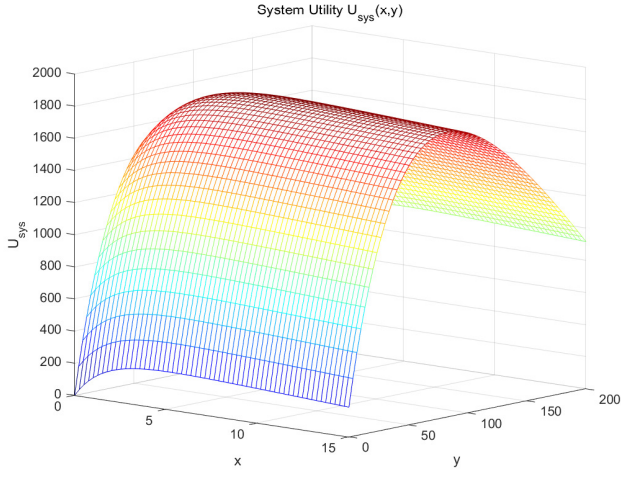
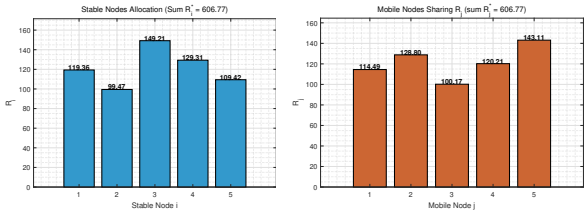


Fig. 5. System utility as a function of pre-training rounds x and collaborative training rounds y.

D. Convergence Performance Under Heterogeneous Conditions

This section conducted experiments to evaluate our proposed Fed-MoSeC Flickr30K dataset [27]. Since heterogeneity



(a) Reward allocation R_i (b) Incentive payment R_j

Fig. 6. Incentive distribution at game theoretic equilibrium.

and modal diversity of SC encoders, most of the methods cannot be directly applied in our considered scenarios, we assume the encoder models are homomorphic to facilitate comparison of potential solutions. We select FedAvg [18] and FedProx [19] as baselines, and GNN-based GCFL+ [30], FedSage+ [31], and FedGTA [32] as application schemes. For graph-structured data, GCFL+ [30] uses personalized aggregation for graph-level tasks, FedSage+ [31] mitigates missing-neighbor information induced by distributed subgraphs, and FedGTA [32] achieves topology-aware aggregation across subgraphs. Regarding the evaluation of our PNCC, we address pseudo-label noise at a different point of intervention compared to other graph-based paradigms. For example, NRGNN [8] refines the graph's topological structure via edge prediction and PTA [13] improves the label propagation process. PNCC instead operates on the intermediate relational representation, namely the similarity matrix. Since these methods act at different levels of the pipeline, a direct numerical comparison is not meaningful. We therefore evaluate FedMoSeC framework against the selected aggregation schemes, measuring convergence speed and final performance under varied node compositions and noise levels.

Figure 7 presents the performance of these aggregation methods under different experimental conditions. We evaluated three stable-to-mobile node ratios (7:3, 5:5, and 3:7) combined with four noise levels (10%, 30%, 50%, and 70%) applied to mobile node data. This design simulates realistic mobile network scenarios where new nodes have varying data quality. Regarding the impact of node composition, Fed-MoSeC's performance and convergence speed improve as the proportion of mobile nodes increases. In the 7:3 stable-to-mobile configuration, our method converges quickly and maintains a stable performance around 470 RSUM. When the ratio shifts to 5:5 and then to 3:7, Fed-MoSeC reaches stable performance in 20 communication rounds, while baseline methods require 60-80 rounds and performance decreases significantly.

In Figure 7, our method also demonstrates superior resilience to noise injected into mobile nodes, and its advantage over baselines becomes more pronounced as noise intensity increases. Under 10% noise, all methods achieve similar final performance. As noise levels rise to 30% and 50%, baseline methods show significant performance drops and training instability. In contrast, Fed-MoSeC maintains both rapid convergence and robust performance. Under severe 70% noise, simulating scenarios like disaster response, baseline methods drop to around 400 RSUM, whereas Fed-MoSeC achieves 474.0 RSUM and converges around 20 rounds, showing superior noise tolerance.

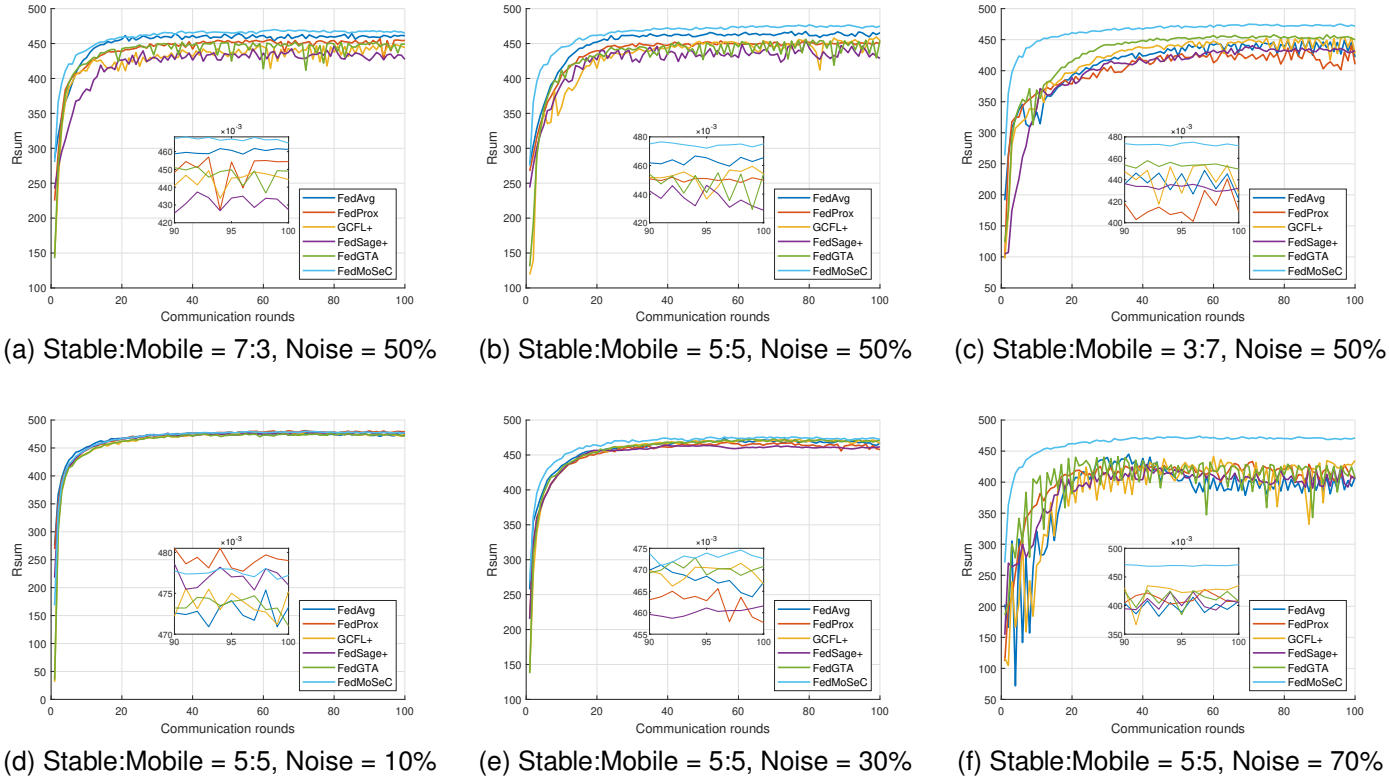
TABLE III
PERFORMANCE COMPARISON ACROSS THREE MODALITY CONVERSIONS UNDER 30% PSEUDO-LABEL NOISE. 'SUP." INDICATES THE SUPERVISED VERSION OF PARALLEL SPEECHCLIP BY REPLACING THE IMAGE ENCODER WITH TEXT ENCODER.

Method	**R@1**	**R@5**	**R@10**	**R@1**	**R@5**	**R@10**	**Rsum**
	Image → Text			Text → Image			
Fed-MoSeC	73.1	91.1	97.8	58.8	88.1	94.7	503.6
FedCor	65.7	88.8	95.1	54.0	84.8	92.0	480.4
FedMix	63.5	78.4	95.5	49.9	80.2	87.7	455.2
Trimmed Mean	62.0	81.2	88.7	47.7	78.3	84.9	442.8
Multi-Krum	61.4	80.9	88.7	46.5	78.2	84.6	440.3
Krum	59.8	79.7	86.6	45.5	77.0	83.5	432.1
	Image → Speech			Speech → Image			
Fed-MoSeC	52.9	83.8	92.7	42.9	76.5	87.2	436.0
FedCor	46.6	78.9	88.5	37.1	70.5	82.8	404.4
FedMix	43.5	76.3	87.5	34.0	68.7	80.7	390.7
Trimmed Mean	42.8	74.3	84.1	33.1	66.7	78.3	379.3
Multi-Krum	41.4	73.4	83.6	32.8	65.6	77.8	374.6
Krum	39.1	71.3	81.2	30.3	63.4	75.5	360.8
	Speech → Text (Sup.)			Text → Speech (Sup.)			
Fed-MoSeC	99.9	100.0	100.0	99.8	100.0	100.0	599.7
	Speech → Text			Text → Speech			
Fed-MoSeC	73.3	92.3	95.8	70.5	91.8	96.8	520.5
FedCor	62.6	86.6	92.5	63.6	86.1	91.9	483.3
FedMix	57.7	83.6	90.9	58.1	83.0	90.3	463.6
Trimmed Mean	55.8	79.6	85.2	57.5	79.1	85.9	443.1
Multi-Krum	54.0	78.7	85.0	56.5	79.4	84.5	438.1
Krum	52.6	77.1	83.3	54.3	76.3	82.6	426.2

The improvement is most notable under the combined challenge of a high mobile node proportion and high noise. Fed-MoSeC leverages the diversity factor in its confidence computation to extract valuable information from the noisy mobile nodes, resulting in both faster convergence and higher final accuracy. Further, it is worth noting that our framework can be performed in any circumstance.

Figure 8 evaluates our framework under non-IID data distributions. In this setup, stable nodes sample only from the first 50% of the dataset, representing specialized domains. Mobile nodes sample from the entire dataset but with 50% noise corruption, simulating their broader yet noisier data exposure. The non-IID results show the value of incorporating mobile nodes despite their data quality. Compared to a baseline of only stable nodes, including mobile nodes with 50% noise, yields both faster convergence and higher final performance, achieving round 475 RSUM versus 453.8 RSUM. This shows that the diverse perspectives from mobile nodes can compensate for data noise when managed by our confidence-weighted aggregation. Increasing the proportion of mobile nodes from 30% to 70% results in higher final performance. This finding highlights the effectiveness of our PNCC component in extracting information from noisy mobile nodes while mitigating negative effects.

These results validate that Fed-MoSeC addresses two key challenges in federated SC: 1) faster convergence through diverse mobile node data; 2) robust performance under noise via adaptive confidence weighting.



(a) Stable:Mobile = 7:3, Noise = 50% (b) Stable:Mobile = 5:5, Noise = 50% (c) Stable:Mobile = 3:7, Noise = 50%

(d) Stable:Mobile = 5:5, Noise = 10% (e) Stable:Mobile = 5:5, Noise = 30% (f) Stable:Mobile = 5:5, Noise = 70%

Fig. 7. Convergence performance comparison of federated aggregation methods under varying node compositions and noise levels.

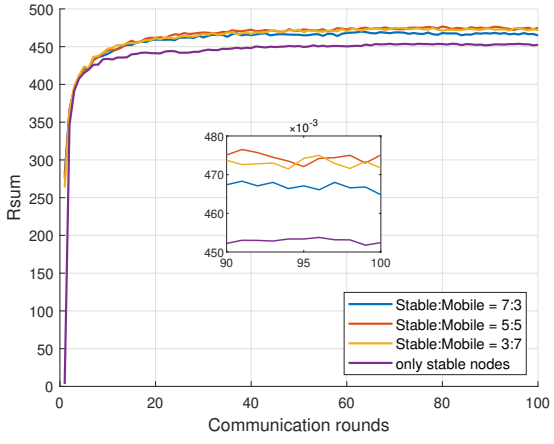


Fig. 8. Performance analysis under non-IID data distributions with different stable-to-mobile node ratios.

E. Experimental Results and Analysis on Multi-modal Scenarios

In this section, we extended the proposed framework on a new dataset to test its applicability across diverse multimodal scenarios.

1) Experimental Setup for Tri-modal: To evaluate FedMoSeC in multi-modal scenarios, we extend our experiments to a tri-modal dataset combining MSCOCO and SpokenCOCO. Because SpokenCOCO provides human-uttered spoken captions for each MSCOCO image, the combined dataset naturally forms three paired relationships: image-text (v, t), image-speech (v, a), and text-speech (t, a). In total, this dataset contains 113,287 training images, alongside 5,000 images each for validation and testing.

For visual feature extraction, we use a Faster R-CNN with a ResNet-101 backbone to extract $n_v = 36$ salient region features per image, each with a raw dimension of 2048. For the text modality, a pre-trained BERT-base-uncased model extracts token-level contextual features with a raw dimension of 768. For speech, we first apply the Montreal Forced Aligner (MFA) to determine word-level time boundaries. Within each aligned segment, a frozen pre-trained HuBERT-base model then extracts 768-dimensional vectors per word. After extracting these raw features, we use linear layers to project all of them into a unified dimension of $d = 1024$.

To jointly represent the tri-modal relationships, we expand the similarity matrix $\mathbf{S}$ of the GNN adapter into a 3×3 block structure. Unless otherwise specified, this adapter maintains the same basic configuration used in our Flickr30K experiments. During federated training, all pre-trained encoders remain frozen. We deploy the framework across a network of 10 client nodes, consisting of 5 stable nodes and 5 mobile nodes. To simulate realistic mobile network conditions, we inject 30% pseudo-label noise into the mobile node data through random cross-modal mismatching.

2) Evaluation of Bridge-Inferred Similarity: We design two training configurations to validate the bridge-inferred similarity mechanism from Eq. (12). The direct supervision configuration sets $\mathcal{P} = \{(v, t), (v, a), (t, a)\}$, utilizing all

available pairs. The bridge supervision configuration restricts $\mathcal{P}$ to $\{(v,t),(v,a)\}$, intentionally withholding direct text-speech pairs. In this setting, the cross-modal similarity between text and speech is inferred through the visual modality as a semantic anchor:

$$\mathbf{S}_{ta}^{\text{bridge}} = \mathbf{S}_{tv} \cdot \mathbf{S}_{va}. \tag{45}$$

This design simulates practical scenarios where parallel data for certain modality pairs is unavailable. It tests whether the framework can achieve effective alignment through a shared intermediary modality.

3) Performance Analysis: Fed-MoSeC achieves comparable performance on image-speech and image-text retrieval tasks. Specifically, R@10 for $v \rightarrow a$ and $a \rightarrow v$ reaches 92.7% and 87.2%, respectively. This confirms that the GNN adapter effectively integrates the speech modality into the unified semantic space. Under direct supervision, text-speech retrieval achieves near-perfect accuracy, with R@1 exceeding 99% for both $t \rightarrow a$ and $a \rightarrow t$. However, this unusually high performance stems primarily from shallow lexical mapping rather than deep semantic understanding. Because the speech signals are generated by reading text captions, direct supervision encourages simple word-to-word matching.

In contrast, under bridge supervision, text-speech retrieval still achieves an R@1 of 70.5% and an R@10 of 96.8% for $t \rightarrow a$, despite the lack of direct $t \leftrightarrow a$ pairs. This confirms that the bridge mechanism enables true semantic alignment via visual features, avoiding shallow symbol matching. Overall, transitioning from dual- to tri-modal data demonstrates the scalability of Fed-MoSeC with respect to the number of modalities $|\mathcal{U}|$. The framework achieves effective cross-modal alignment even without direct parallel data for every modality pair.

4) Comparison with Noise-Robust Federated Learning Methods: We evaluate the noise robustness of the PNCC mechanism by benchmarking Fed-MoSeC against five established noise-robust FL methods: FedCor [33], FedMix [34], Krum [35], Multi-Krum [35], and Trimmed Mean [36]. We evaluate all methods on the tri-modal SpokenCOCO dataset under identical conditions, injecting 30% pseudo-label noise into the mobile node data via random cross-modal mismatching.

As reported in Table III, Fed-MoSeC consistently achieves the highest Rsum across all three cross-modal retrieval tasks. Notably, the performance margin between Fed-MoSeC and the baselines widens as the cross-modal inference chain lengthens. Specifically, Fed-MoSeC outperforms the strongest baseline, FedCor, by 23.2 and 31.6 in Rsum on the directly supervised image-to-text and image-to-speech tasks, respectively. This gap expands to 37.2 on the bridge-inferred text-to-speech task (520.5 versus 483.3). This widening gap highlights the noise-cascading effect inherent in the bridge mechanism $\mathbf{S}_{ta}^{\text{bridge}} = \mathbf{S}_{tv} \cdot \mathbf{S}_{va}$. Perturbations in either $\mathbf{S}_{tv}$ or $\mathbf{S}_{va}$ propagate multiplicatively into the inferred text-speech similarity. Standard noise robust FL methods are not designed for noise that corrupts cross-modal relational structuress, whereas PNCC addresses this by directly filtering the cross-modal similarity representation.

Among the baselines, FedCor performs best. Its use of Gaussian Processes to model inter-client loss correlations implicitly down-weights mobile nodes with noisy updates. Krum and Multi-Krum perform worst across all tasks because they strictly exclude client updates that deviate significantly from the majority. In our scenario, mobile node updates deviate due to environmental distribution shifts and pseudo-label noise, not malicious intent. Therefore, these methods systematically discard mobile node contributions, losing valuable data diversity. The consistent underperformance of the baselines highlights a key limitation: standard noise robust FL methods are not designed for noise that corrupts cross-modal relational structures.

F. Computational Complexity Analysis of PNCC

This section quantifies the additional computation introduced by PNCC compared to a standard FL iteration. We also analyze its scaling with the number of modalities $|\mathcal{U}|$ and participating nodes K.

PNCC starts with a confidence-based filtering stage that operates on the batch-level matching matrix $\mathbf{S}_{k,\mu\nu} \in \mathbb{R}^{B\times B}$. Constructing this matrix via pairwise dot products costs $\mathcal{O}(B^2 d)$. Extracting the diagonal statistics and computing the adaptive threshold $\tau_{k,\text{adaptive}}$ require $\mathcal{O}(B^2)$ and $\mathcal{O}(B)$ operations, respectively. Because the batch size B is a small, fixed hyperparameter, this filtering stage adds negligible overhead compared to the primary backpropagation process.

After filtering, extracting the target similarity matrix becomes the dominant computational cost. Each node performs one additional forward pass of the GNN-based adapter on its filtered samples using the global model parameters $\boldsymbol{\theta}_{\text{global}}$. This generates per-sample node features $\mathbf{H} \in \mathbb{R}^{N\times d}$. It then constructs the per-sample similarity matrix $\mathbf{S}_{\text{target}}^{(k)} = \mathbf{H}\mathbf{H}^{\top} \in \mathbb{R}^{N\times N}$ at a cost of $\mathcal{O}(N^2 d)$ per sample. Next, computing the distillation loss and diversity scoring relies on element-wise operations between $\mathbf{S}_k^{(t)}$ and $\mathbf{S}_{\text{target}}^{(k)}$, which scale as $\mathcal{O}(N^2)$.

Overall, integrating PNCC only adds the cost of one adapter forward pass at $\mathcal{O}(N^2 d)$ per sample. At the system level, server-side confidence-weighted aggregation costs $\mathcal{O}(K|\boldsymbol{\theta}|)$. This complexity is identical to the standard FedAvg protocol, with the only extra communication being a single scalar CONF_k sent per node. Adding modalities primarily increases the local cost quadratically with N. As nodes execute PNCC independently, the total system computation still scales linearly with K.

VI. Conclusions

In this paper, we proposed Fed-MoSeC, an FL framework for SC in MNs. In Fed-MoSeC, the heterogeneous updates of multiple modality-specific encoders are transformed into novel homogeneous cross-modal semantic encoder adapter updates. It enables nodes to perform federated aggregation and thus collaboratively train models without sharing raw data. Notably, Fed-MoSeC also incorporates a novel PNCC and a proprietary TOC. With the proposed PNCC, Fed-MoSeC mitigates the adverse effects of cross-modal pseudo-label noise, improving training accuracy. Moreover, the TOC establishes a mutually

beneficial training paradigm by adapting the teaching-study process to the mobile-aware utilities of participating nodes. Our simulation results quantitatively validate these advantages: Fed-MoSeC reduces communication costs by over an order of magnitude to $\sim 12MB$ per round and maintains a high performance of approximately 470 RSUM even under severe 70% noise, demonstrating its efficiency, speed, and robustness.

REFERENCES

[1] G. Zhang, Q. Hu, Z. Qin, Y. Cai, G. Yu, and X. Tao, "A unified multi-task semantic communication system for multimodal data," *IEEE Transactions on Communications*, vol. 72, no. 7, pp. 4101–4116, 2024.

[2] Z. Qin, X. Tao, J. Lu, W. Tong, and G. Y. Li, "Semantic communications: Principles and challenges," *arXiv preprint arXiv:2201.01389*, 2021.

[3] Q. Lan, D. Wen, Z. Zhang, Q. Zeng, X. Chen, P. Popovski, and K. Huang, "What is semantic communication? a view on conveying meaning in the era of machine intelligence," *Journal of Communications and Information Networks*, vol. 6, no. 4, pp. 336–371, 2021.

[4] H. Xie and Z. Qin, "A lite distributed semantic communication system for internet of things," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 1, pp. 142–153, 2020.

[5] G. Shi, Y. Xiao, Y. Li, and X. Xie, "From semantic communication to semantic-aware networking: Model, architecture, and open problems," *IEEE Communications Magazine*, vol. 59, no. 8, pp. 44–50, 2021.

[6] Z. Qin, G. Y. Li, and H. Ye, "Federated learning and wireless communications," *IEEE Wireless Communications*, vol. 28, no. 5, pp. 134–140, 2021.

[7] H. Liu, B. Hu, X. Wang, C. Shi, Z. Zhang, and J. Zhou, "Confidence may cheat: Self-training on graph neural networks under distribution shift," in *Proc. of the ACM web conference 2022*, pp. 1248–1258, 2022.

[8] E. Dai, C. Aggarwal, and S. Wang, "Nrgnn: Learning a label noise resistant graph neural network on sparsely and noisily labeled graphs," in *Proc. of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, pp. 227–236, 2021.

[9] T. Lin, L. Kong, S. U. Stich, and M. Jaggi, "Ensemble distillation for robust model fusion in federated learning," *in Proc. of Advances in neural information processing systems*, vol. 33, pp. 2351–2363, 2020.

[10] G. Zheng, Q. Ni, K. Navaie, H. Pervaiz, G. Min, A. Kaushik, and C. Zarakovitis, "Mobility-aware split-federated with transfer learning for vehicular semantic communication networks," *IEEE Internet of Things Journal*, vol. 11, no. 10, pp. 17237–17248, 2024.

[11] B. Xiong, X. Yang, F. Qi, and C. Xu, "A unified framework for multimodal federated learning," *Neurocomputing*, vol. 480, pp. 110–118, 2022.

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

[13] H. Dong, J. Chen, F. Feng, X. He, S. Bi, Z. Ding, and P. Cui, "On the equivalence of decoupled graph convolution network and label propagation," in *Proc. of the Web Conference 2021*, pp. 3651–3662, 2021.

[14] W. Feng, J. Zhang, Y. Dong, Y. Han, H. Luan, Q. Xu, Q. Yang, E. Kharlamov, and J. Tang, "Graph random neural networks for semi-supervised learning on graphs," *in Proc. of Advances in neural information processing systems*, vol. 33, pp. 22092–22103, 2020.

[15] W. Wu, L. He, W. Lin, R. Mao, C. Maple, and S. Jarvis, "Safa: A semi-asynchronous protocol for fast federated learning with low overhead," *IEEE Transactions on Computers*, vol. 70, no. 5, pp. 655–668, 2020.

[16] J. Xu, H. Yao, R. Zhang, T. Mai, S. Huang, and S. Guo, "Federated learning powered semantic communication for uav swarm cooperation," *IEEE Wireless Communications*, vol. 31, no. 4, pp. 140–146, 2024.

[17] G. Zheng, Q. Ni, K. Navaie, and H. Pervaiz, "Semantic communication in satellite-borne edge cloud network for computation offloading," *IEEE Journal on Selected Areas in Communications*, vol. 42, no. 5, pp. 1145–1158, 2024.

[18] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. of Artificial intelligence and statistics*, pp. 1273–1282, PMLR, 2017.

[19] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," *Proc. of Machine learning and systems*, vol. 2, pp. 429–450, 2020.

[20] Y. Tan, G. Long, L. Liu, T. Zhou, Q. Lu, J. Jiang, and C. Zhang, "Fedproto: Federated prototype learning across heterogeneous clients," in *Proc. of the AAAI conference on artificial intelligence*, vol. 36, pp. 8432–8440, 2022.

[21] K. Sohn, D. Berthelot, N. Carlini, Z. Zhang, H. Zhang, C. A. Raffel, E. D. Cubuk, A. Kurakin, and C.-L. Li, "Fixmatch: Simplifying semi-supervised learning with consistency and confidence," *in Proc. of Advances in neural information processing systems*, vol. 33, pp. 596–608, 2020.

[22] K. Ding, X. Ma, Y. Liu, and S. Pan, "Divide and denoise: empowering simple models for robust graph semi-supervised learning against label noise," in *Proc. of ACM SIGKDD Conference on Knowledge Discovery and Data Mining (30th: 2024)*, pp. 574–584, Association for Computing Machinery, 2024.

[23] S. Sun, Z. Zhang, Q. Pan, M. Liu, Y. Wang, T. He, Y. Chen, and Z. Wu, "Staleness-controlled asynchronous federated learning: Accuracy and efficiency tradeoff," *IEEE Transactions on Mobile Computing*, vol. 23, no. 12, pp. 12621–12634, 2024.

[24] Q. Cui, X. Hu, W. Ni, X. Tao, P. Zhang, T. Chen, K.-C. Chen, and M. Haenggi, "Vehicular mobility patterns and their applications to internet-of-vehicles: A comprehensive survey," *Science China Information Sciences*, vol. 65, no. 11, p. 211301, 2022.

[25] J. Liang, Z. Yu, H. Pervaiz, G. Zheng, and N. Suri, "Game theory empowered carbon-intelligent federated multiedge caching for industrial internet of things," *IEEE Internet of Things Journal*, vol. 12, no. 17, pp. 34875–34889, 2025.

[26] G. Zheng, Q. Ni, K. Navaie, H. Pervaiz, A. Kaushik, and C. Zarakovitis, "Energy-efficient semantic communication for aerial-aided edge networks," *IEEE Transactions on Green Communications and Networking*, vol. 8, no. 4, pp. 1742–1751, 2024.

[27] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," *Transactions of the association for computational linguistics*, vol. 2, pp. 67–78, 2014.

[28] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proc. of the IEEE conference on computer vision and pattern recognition*, pp. 6077–6086, 2018.

[29] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proc. of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

[30] H. Xie, J. Ma, L. Xiong, and C. Yang, "Federated graph classification over non-iid graphs," *in Proc. of Advances in neural information processing systems*, vol. 34, pp. 18839–18852, 2021.

[31] K. Zhang, C. Yang, X. Li, L. Sun, and S. M. Yiu, "Subgraph federated learning with missing neighbor generation," *in Proc. of Advances in neural information processing systems*, vol. 34, pp. 6671–6682, 2021.

[32] X. Li, Z. Wu, W. Zhang, Y. Zhu, R.-H. Li, and G. Wang, "Fedgta: Topology-aware averaging for federated graph learning," *arXiv preprint arXiv:2401.11755*, 2024.

[33] M. Tang, X. Ning, Y. Wang, J. Sun, Y. Wang, H. Li, and Y. Chen, "Fedcor: Correlation-based active client selection strategy for heterogeneous federated learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10102–10111, 2022.

[34] T. Yoon, S. Shin, S. J. Hwang, and E. Yang, "Fedmix: Approximation of mixup under mean augmented federated learning," *arXiv preprint arXiv:2107.00233*, 2021.

[35] P. Blanchard, E. M. El Mhamdi, R. Guerraoui, and J. Stainer, "Machine learning with adversaries: Byzantine tolerant gradient descent," *Advances in neural information processing systems*, vol. 30, 2017.

[36] D. Yin, Y. Chen, R. Kannan, and P. Bartlett, "Byzantine-robust distributed learning: Towards optimal statistical rates," in *International conference on machine learning*, pp. 5650–5659, Pmlr, 2018.