

Decentralized Federated Learning for Heterogeneous Multi-Task Semantic Communication

Lin Yin, Tiejun Lv, Senior Member, IEEE, Weicai Li, Xi Yu, and Xiaoyu He

Abstract—Collaborative training in distributed semantic communication (DSC) networks typically relies on decentralized federated learning (DFL). However, pushing topology-agnostic aggregation into heterogeneous, multi-task environments creates a fundamental bottleneck: it drives negative transfer and over-consensus bias (OCB). This paper introduces a personalized DSC framework that cuts off this cross-task interference. At the node level, a policy-driven multi-path routing mechanism separates task-specific features from shared representations to preserve local fidelity. Across the network, we deploy a “communication-while-aggregation” protocol. It calibrates a column-stochastic consensus matrix using task affinities. This limits the system to absorbing complementary knowledge while actively blocking mismatched parameter updates. To bound the convergence, we derive a unified Lyapunov drift analysis. We reveal a strict U-shaped trade-off: deeper topological mixing reduces variance but amplifies structural OCB. Resolving this tension yields a closed-form expression for the optimal aggregation depth. We evaluate the proposed framework on NYU-v2, where the results reveal a clear trade-off between insufficient aggregation and excessive topological mixing. At the analytically derived optimal aggregation depth, our method achieves a 4.77% global relative improvement over the no-aggregation baseline and outperforms decentralized FedAvg, FedAMP, and heuristic max aggregation. We further evaluate the framework on Taskonomy and imperfect wireless links to examine the effects of network-size variation and wireless-link reliability.

Index Terms—Decentralized federated learning, heterogeneous multi-task learning, similarity-aware aggregation, distributed semantic communication, over-consensus bias.

I. Introduction

Modern wireless networks increasingly execute data-intensive inference tasks directly at the edge. To support this shift, semantic communication (SC) bypassing traditional bit-level transmission has emerged as a practical solution to bandwidth bottlenecks. By extracting and transmitting only task-relevant representations, SC drastically cuts communication overhead [1]. Yet, most existing SC architectures assume single-task environments [2], [3].

Manuscript received 01 July 2025; revised 07 February 2026; accepted 13 August 2026. This paper was supported in part by the National Natural Science Foundation of China under No. 62271068. (corresponding author: Tiejun Lv.)

L. Yin, T. Lv, X. Yu, and X. He are with the School of Information and Communication Engineering, Beijing University of Posts and Telecommunications (BUPT), Beijing 100876, China (e-mail: {yinlin, lvtiejun, yusy, xiaoyuhhh}@bupt.edu.cn).

W. Li is with the Center for Target Cognition Information Processing Science and Technology, and the Key Laboratory of Modern Measurement and Control Technology, Ministry of Education, both at Beijing Information Science and Technology University, Beijing, China (e-mail: liweicai@bistu.edu.cn).

Real-world edge applications are rarely this simple; they require concurrent execution of diverse, heterogeneous tasks, necessitating multi-task learning (MTL) [4], [5]. Centralized MTL orchestration dictates heavy communication costs, strict latency bounds, and severe privacy risks. Decentralized federated learning (DFL) deployed over peer-to-peer meshes avoids this central bottleneck entirely. It forces collaboration directly to the edge [6].

A. Related Work and Motivations

Distributed optimization and task-oriented representations must merge to support edge intelligence. Initial SC research relied heavily on single-modal deep joint source-channel coding (DJSCC) [2], [7]. The field now targets multi-task SC (MT-SC) designs. Recent architectures execute multiple downstream tasks through graph attention mechanisms [8], scalable feature ranking [9], and shared codebooks [1]. Asynchronous multi-task semantic communication provides another route to extract task-independent knowledge [10]. A critical blind spot remains: almost all these MT-SC systems presuppose a central server. They strip away the realities of decentralized data distributions and the dynamic topological limits inherent to wireless edge networks.

DFL protocols take the opposite approach. They cut out the orchestrator via peer-to-peer gossip averaging. The presence of non-independent and identically distributed (non-IID) data severely hinders the convergence of global consensus, thereby necessitating local adaptation through personalized federated learning (PFL). To mitigate the performance degradation induced by such data heterogeneity, recent literature explores clustered subgraphs [11], model and gradient decoupling [12], robust decentralized personalization [13] and over-the-air computation [14]. Security- and resource-aware decentralized learning has also been studied in edge systems, where asynchronous reinforcement federated learning and task offloading are used to support secure and privacy-preserving operation [15]. This line of work mainly concerns security, privacy, and offloading decisions, whereas our setting is shaped by semantic mismatch among heterogeneous task portfolios. In heterogeneous multi-task learning, task compatibility can influence whether shared training is beneficial or leads to interference. Studies on visual task transferability, task grouping, and gradient conflict provide related evidence for using task compatibility to regulate knowledge sharing [16]–[18]. Similarity-guided collaboration has also been used in PFL, where aggregation gives

larger influence to related clients rather than enforcing a single averaged model [19]. These studies motivate our use of task-level semantic similarity as a criterion for guiding peer-to-peer aggregation in decentralized multi-task distributed SC (DSC) networks.

Mixing mismatched tasks over a decentralized topology guarantees severe negative transfer, client drift, and over-consensus bias (OCB) [20]. Standard PFL fixes like clustering or meta-learning [21] break down in SC environments. They demand massive communication payloads. Adaptive aggregation schemes [22], [23] routinely force clients to broadcast dense, high-dimensional neural parameters. The edge physical layer simply cannot support this bandwidth overhead.

Recent literature has attempted to narrow this communication bottleneck through optimizations at the physical and network layers. For example, [24] and [25] study decentralized federated learning over imperfect channels to mitigate transmission impairments. In the semantic communication domain, recent works investigate robust semantic digital-to-analog conversion for discrete quantization [26] and develop empirical performance models for resource allocation [27]. In addition, [28] incorporates semantic encoding into a hierarchical federated architecture to reduce satellite communication overhead. Despite these advances, existing frameworks remain insufficient for multi-task edge environments. Many rely on semi-centralized multi-tier architectures [28], while others treat the underlying neural model as a monolithic object during communication and aggregation [26], [29]. As a result, they cannot explicitly separate task-specific semantic components from the shared representation space [30].

Current designs largely decouple communication-efficient semantic compression from personalized decentralized training. This limitation suggests that robust multi-task edge intelligence requires personalization to be incorporated directly into the semantic compression process. In particular, neighborhood updates should be selectively screened and reweighted so that mismatched semantic information is suppressed while local task fidelity is preserved.

B. Contributions and Organization

Solving this structural disconnect requires moving personalization from the federated averaging step into the semantic representation process. To achieve this, we propose a multi-path semantic encoder with a learnable feature routing policy. Rather than relying on a monolithic representation, the proposed encoder balances shared semantic extraction with task-specific feature preservation. At the network level, these learned routing policies support a similarity-aware decentralized aggregation scheme, where each client evaluates semantic alignment with its neighbors based on task-level semantic similarity and adjusts the aggregation weights accordingly. The resulting framework treats the semantic transceiver as the interface for exchanging task-relevant features, while the

main optimization is placed on routing-policy learning and similarity-aware neighbor aggregation. As a result, mismatched updates are suppressed, whereas semantically aligned updates are assigned greater influence.

In summary, our primary contributions are threefold:

- A distributed, personalized learning framework for heterogeneous semantic communication: instead of relying on monolithic full-model exchange, collaborative modeling is organized around the multi-path encoder that extracts task-relevant semantic representations. This design reduces the communication burden associated with dense parameter sharing while retaining task-specific modeling capacity at local clients.
- A similarity-aware aggregation protocol designed specifically to counter OCB and negative transfer: we utilize the task-specific policy vectors to calibrate the topological weights during training. Consequently, the network blocks mismatched updates and forces compatible peers into tighter cooperation.
- A rigorous convergence proof for the “communication-while-aggregation” protocol: using a unified Lyapunov drift analysis, we uncover a fundamental U-shaped tension; i.e., the variance reduction gained from topology mixing is in direct conflict with structural OCB. We resolve this tension by deriving a closed-form solution for the optimal aggregation depth J^* , establishing the exact point where the steady-state error floor hits its absolute minimum.

Empirical evaluations on the NYU-v2 benchmark validate the existence of these theoretical boundaries. If we restrict the network to a shallow mixing regime ($J = 1$), the similarity-aware routing still manages a 2.11% global relative improvement. It also acts as a safeguard, preventing the highly sensitive depth estimation task from collapsing entirely. Allowing the system to operate at the analytically derived optimal depth unlocks the full potential of the framework, yielding a 4.77% overall relative gain. Looking at individual tasks, Semantic segmentation jumps by 8.04% and depth estimation improves by 3.31%. These margins confirm a critical design hypothesis: explicitly blocking topology-agnostic mixing is mandatory to survive cross-task interference at the edge. We further include Taskonomy experiments and communication-quality simulations to examine the effects of network-size variation and imperfect wireless links.

The remainder of the paper proceeds as follows. Section II sets up the system model. Section III formulates the decentralized training protocol alongside the core optimization objectives. The theoretical convergence analysis, culminating in the derivation of the optimal aggregation depth, is presented in Section IV. Section V discusses the experimental findings, and Section VI draws the final conclusions.

II. System Model

We detail the architecture underpinning our DSC framework. The network functions as a decentralized mesh.

TABLE 1: Summary of Key Notations

Notation	Definition
$i, \mathcal{N}$	Client index, and set of clients in the network.
$t, \mathcal{T}$	Task index, and set of global semantic tasks.
$\mathcal{T}_i$	Personalized task set of client i .
$n, \mathcal{N}_i$	One-hop neighbor index, and set of neighbors of client i .
$\mathbf{u}_t$	Learnable binary policy vector for task t .
$s_{t,t'}$	Semantic similarity between task t and task t' .
$S_{i,n}$	Overall semantic alignment between client i and its neighbor n .
$\mathcal{L}_{i,t}$	Task-level local loss of client i on task t .
$\mathcal{L}_i$	Client-level local loss of client i .
$\mathcal{L}$	Global network-wide loss.
r, R	Communication round index and total number of communication rounds.
e, E	Training epoch index and total epochs in one round.
j, J	Aggregation depth index and total number of aggregation steps.
$\psi_{i,r,j}$	Model parameters of client i after the j -th aggregation step in round r .
$w_{i,n}, \mathbf{W}$	Aggregation weight between client i and neighbor n , and global aggregation weight matrix.
$E_{\text{cons},r,J}$	Round-wise consensus error after J aggregation steps in communication round r .
$\Phi_{r,J}$	Composite Lyapunov potential function.
Ω_*^2	Over-consensus bias.
J^*	Optimal aggregation depth.

Clients bypass central coordination entirely, relying instead on direct neighborhood exchanges to jointly train their local multi-path encoders and task-specific decoders.

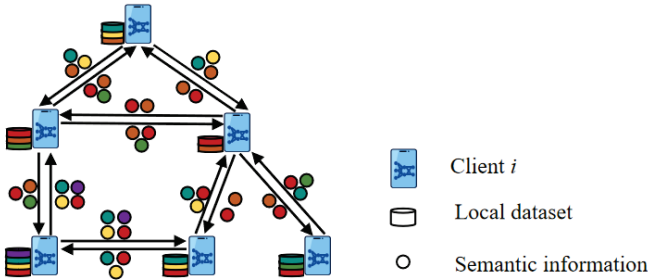


Fig. 1. The proposed distributed semantic communication framework. The distinct colors indicate different tasks and semantic information.

A. Distributed Semantic Communication Model

The DSC network comprises a set of clients $\mathcal{N} = \{1, \dots, N\}$ forming a mesh topology, and a global semantic task set $\mathcal{T} = \{1, \dots, T\}$. Each client $i \in \mathcal{N}$ is assigned a personalized semantic task portfolio $\mathcal{T}_i \subseteq \mathcal{T}$ and possesses an independent local dataset $\mathcal{D}_i = \cup_{t \in \mathcal{T}_i} \mathcal{D}_{i,t}$, where $\mathcal{D}_{i,t} = \{(x, y_t)\}$ contains the raw input x and the ground truth labels y_t specific to task t . Client i communicates exclusively with its one-hop neighbor set $\mathcal{N}_i$. The one-hop neighbor set $\mathcal{N}_i$ is specified by a distance-constrained

communication graph. Each edge represents a feasible peer-to-peer link between two clients. Information from non-neighboring clients is not obtained through direct long-range broadcasting, but through repeated one-hop aggregation over the connected graph. This setting keeps the decentralized training process local while allowing model information to propagate across the network over multiple aggregation steps.

Rather than reconstructing raw bit streams, the localized SC model extracts and transmits only task-relevant features. This design explicitly bypasses conventional bandwidth bottlenecks. As Fig. 1 illustrates, the network assigns a heterogeneous multi-task portfolio to each client operating on a strictly local dataset. Clients completely avoid transmitting raw data or dense model weights. They locally encode task-specific semantic representations and exchange them directly across peer-to-peer mesh links, eliminating the need for central coordination.

Consider a raw input image $x_i \in \mathbb{R}^{C_{\text{in}} \times W_{\text{in}} \times H_{\text{in}}}$ at client i . The joint encoder $f_t^{\text{enc}}(\cdot)$ parameterized by $\theta_{i,t}^E$ generates the task-relevant latent representation $X_{i,t}$ for task t as

$$X_{i,t} = f_t^{\text{enc}}(x_i; \theta_{i,t}^E), \quad X_{i,t} \in \mathbb{R}^{C_{\text{out}} \times W_{\text{out}} \times H_{\text{out}}}. \quad (1)$$

Here, C_{in} , W_{in} , and H_{in} denote the channel number, width, and height of the input image, respectively, while C_{out} , W_{out} , and H_{out} denote those of the latent semantic feature.

The latent tensor $X_{i,t}$ is mapped into a sequence of complex channel symbols $\mathbf{x}_{i,t} \in \mathbb{C}^S$, where S denotes the available symbol budget of the edge link. The symbols are transmitted from client i to its one-hop neighbor n over the decentralized communication graph. During neighborhood exchange, transmissions follow a time division multiple access (TDMA)-based link schedule. To construct the schedule, we build a conflict graph over the feasible one-hop links: two links are treated as conflicting if they share a transmitter or receiver, or if one transmission may interfere with the receiver of the other link under the adopted interference range. A graph-coloring rule then assigns conflicting links to different time slots, following coloring-based link scheduling for TDMA transmissions in wireless networks [31]. Under this scheduled one-hop exchange, an active link is modeled with the desired signal, channel fading, and additive white Gaussian noise (AWGN), while conflicting links are deactivated in the same time slot.

For a scheduled transmission over link (i, n) , the received signal vector at client n is modeled as

$$\mathbf{y}_{n,i,t} = \mathbf{H}_{n,i} \mathbf{x}_{i,t} + \mathbf{z}_{n,i,t}, \quad (2)$$

where $\mathbf{H}_{n,i} \in \mathbb{C}^{S \times S}$ represents the equivalent channel matrix of link (i, n) , and $\mathbf{z}_{n,i,t} \sim \mathcal{CN}(\mathbf{0}, \sigma_c^2 \mathbf{I})$ is the AWGN vector.

Assuming that client n has channel state information for the scheduled link, a zero-forcing equalizer estimates the transmitted symbols as

$$\hat{\mathbf{x}}_{n,i,t} = (\mathbf{H}_{n,i}^H \mathbf{H}_{n,i})^{-1} \mathbf{H}_{n,i}^H \mathbf{y}_{n,i,t} = \mathbf{x}_{i,t} + \tilde{\mathbf{z}}_{n,i,t}, \quad (3)$$

where $\mathbf{H}_{n,i}^H$ denotes the Hermitian transpose, and $\tilde{\mathbf{z}}_{n,i,t}$ is the amplified noise after equalization.

The encoder and decoder for task t are trained as a paired semantic transceiver at client i . After training, the corresponding decoder parameters $\theta_{i,t}^D$ are made available to the intended receiving neighbor for semantic reconstruction. Client n reshapes $\hat{\mathbf{x}}_{n,i,t}$ into the feature tensor $\hat{X}_{n,i,t}$ and reconstructs the task prediction as

$$\hat{y}_{n,i,t} = f_t^{\text{dec}}(\hat{X}_{n,i,t}; \theta_{i,t}^D), \quad (4)$$

where $\theta_{i,t}^D$ denotes the task-specific decoder trained at client i and used by client n to decode the received semantic feature.

The scheduled one-hop link is further associated with packet-level reliability. For link (i, n) , we use the packet loss probability $p_{n,i}^{\text{loss}}$ to characterize the reliability of packet delivery. This probability reflects the link-level signal-to-noise ratio (SNR) condition and is determined by the nominal SNR setting, link distance, and packet length [24]. A packet-level outage or decoding failure is counted as an unsuccessful packet delivery.

Let $K_{n,i}^{\text{pkt}}$ denote the number of packets transmitted over link (i, n) , which is determined by the message size and the packet length. When a packet is lost, a retransmission is triggered until successful delivery. Under this latency model, packet loss does not alter the decoded semantic feature after successful reception, but it increases the communication time. If ℓ_{pkt} denotes the duration of one packet transmission, the expected link transmission latency is

$$\ell_{n,i}^{\text{link}} = \frac{K_{n,i}^{\text{pkt}} \ell_{\text{pkt}}}{1 - p_{n,i}^{\text{loss}}}. \quad (5)$$

This packet-level abstraction allows link reliability to affect decentralized aggregation through retransmission latency, without introducing a detailed physical-layer modulation, coding, or outage analysis.

B. Design of Multi-Task Joint Source-Channel Coding

1) Policy-Driven Multi-Path Encoder: To enable heterogeneous multi-task extraction without instantiating independent backbones, our joint encoder employs a policy-driven feature routing architecture [32]. The shared backbone comprises K sequential computational blocks. We define a learnable binary policy vector $\mathbf{u}_t = [u_{t,1}, \dots, u_{t,K}]^T \in \{0, 1\}^K$ for each task t , which dictates the execution path.

For the t -th task of client i , the latent representation evolves recursively through the K blocks according to

$$\alpha_{i,t,k} = u_{t,k} \cdot \mathcal{F}_{i,k}(\alpha_{i,t,k-1}) + \alpha_{i,t,k-1}, \quad (6)$$

for all $k \in \{1, \dots, K\}$, where $\alpha_{i,t,0} = x$ is the initial input, and $\mathcal{F}_{i,k}(\cdot)$ represents the transfer function of the k -th shared block parameterized by the local client. The indicator $u_{t,k}$ dynamically controls whether the k -th block is activated or bypassed. The final output $\alpha_{i,t,K}$ corresponds to the extracted semantic feature X_t .

This mechanism ensures that the active encoder parameters $\theta_{i,t}^E$ are a functionally restricted subset of the global parameter space $\mathcal{F}_i = \{\mathcal{F}_{i,k}\}_{k=1}^K$, governed by the task-specific policy $\mathbf{u}_t$. By disentangling the execution paths, the network prevents gradient interference and structural OCB among disjoint tasks. As illustrated in Fig. 2, a policy-driven multi-path encoder adaptively extracts task-specific semantic representations by routing the input through shared computational blocks. Subsequently, independent task-specific decoders reconstruct the final predictions for their corresponding tasks.

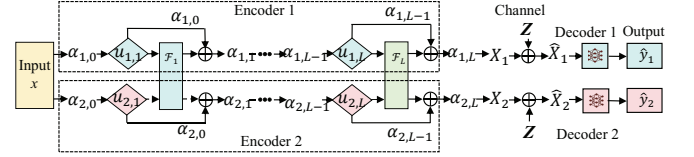


Fig. 2. Schematic of the policy-driven multi-path semantic encoding and task-specific decoding architecture for heterogeneous multi-task execution.

The routing policy also provides a compact descriptor of task-specific use of the shared encoder. Since $u_{t,k}$ records whether task t activates the k -th shared block, two tasks with more overlapping active blocks tend to update more overlapping parts of the shared backbone during local training. This overlap serves as a lightweight indicator of compatibility in the shared representation space. Consequently, we quantify inter-task affinity by defining the semantic similarity between any two tasks $t \in \mathcal{T}_i$ and $t' \in \mathcal{T}_n$ as the cosine similarity of their policy vectors expressed as

$$s_{t,t'} = \frac{\mathbf{u}_t^\top \mathbf{u}_{t'}}{\|\mathbf{u}_t\| \|\mathbf{u}_{t'}\|}. \quad (7)$$

Under this binary policy representation, the cosine score can be understood as a normalized measure of active-path overlap. Its numerator counts the shared activated blocks between two tasks, while the normalization avoids favoring tasks merely because they activate more blocks. Thus, the metric provides a lightweight indicator of task compatibility within the proposed multi-path encoder, and its interpretation is based on how tasks select and share encoder blocks.

And the overall semantic alignment between client i and its neighbor n is evaluated using symmetric best matching as

$$S_{i,n} = \frac{1}{2|\mathcal{T}_i|} \sum_{t \in \mathcal{T}_i} \max_{t' \in \mathcal{T}_n} s_{t,t'} + \frac{1}{2|\mathcal{T}_n|} \sum_{t' \in \mathcal{T}_n} \max_{t \in \mathcal{T}_i} s_{t,t'}, \quad (8)$$

where $S_{i,n}$ guides decentralized aggregation weights, prioritizing clients with overlapping pathways while filtering out those with disjoint task portfolios.

2) Task-Specific Decoders: While the encoder employs a shared routing backbone, each task t maintains an independent, lightweight DeepLab-style dilated convolutional decoder. The decoder exploits multi-scale contextual features via parallel branches, which are linearly fused along

the channel dimension. Because the parameters $\theta_{i,t}^D$ are strictly task-specific, they are decoupled from the neighbor aggregation process, ensuring high local task fidelity.

C. Personalized and Global Optimization Objectives

For client i , the empirical loss on task t is evaluated over its corresponding local dataset $\mathcal{D}_{i,t}$, as given by

$$\mathcal{L}_{i,t}(\lambda_{i,t}, \mathcal{D}_{i,t}, \theta_{i,t}) = \frac{1}{|\mathcal{D}_{i,t}|} \sum_{(x,y_t) \in \mathcal{D}_{i,t}} \lambda_{i,t} L_t(\hat{y}_{i,t}, y_t), \quad (9)$$

where $\lambda_{i,t}$ denotes the task-specific weighting coefficient that governs the relative importance of task t , $L_t(\cdot, \cdot)$ is the task-specific criterion (e.g., cross-entropy for segmentation), and $\theta_{i,t}$ encompasses both the localized encoder pathway and the task-specific decoder parameters. While all parameters in $\theta_{i,t}$ are jointly updated during local gradient descent, only the shared encoder routing backbone participates in the decentralized network-wide consensus protocol.

Let $\theta_i = \mathcal{F}_i$ denote the entirety of the aggregatable encoder backbone on client i . By treating the localized decoders as conditionally optimized internal variables, the aggregate local objective for client i is formulated with respect to the shared backbone as

$$\mathcal{L}_i(\Lambda_i, \mathcal{D}_i, \theta_i) = \sum_{t \in \mathcal{T}_i} \mathcal{L}_{i,t}(\lambda_{i,t}, \mathcal{D}_{i,t}, \theta_{i,t}), \quad (10)$$

where $\Lambda_i = \{\lambda_{i,t} | t \in \mathcal{T}_i\}$ collects the predefined importance weights for the local task portfolio.

Building upon the local objectives, the generalized global loss function over the entire mesh network is defined as the uniform average of all client-specific losses given by

$$\mathcal{L}(\Lambda, \mathcal{D}, \theta) = \frac{1}{N} \sum_{i \in \mathcal{N}} \mathcal{L}_i(\Lambda_i, \mathcal{D}_i, \theta_i), \quad (11)$$

where $\Lambda = \{\Lambda_i | i \in \mathcal{N}\}$, $\mathcal{D} = \{\mathcal{D}_i | i \in \mathcal{N}\}$, and $\theta = \{\theta_i | i \in \mathcal{N}\}$ aggregates the shared parameters across all clients.

The DFL system optimizes two interdependent objectives. First, the localized personalization objective seeks the optimal parameters θ_i^* that minimize client i 's empirical risk formulated as

$$\theta_i^* = \arg \min_{\theta_i} \mathcal{L}_i(\Lambda_i, \mathcal{D}_i, \theta_i), \quad \forall i \in \mathcal{N}. \quad (12)$$

Secondly, the network-level consensus objective seeks the global minimizer θ^* formulated as

$$\theta^* = \arg \min_{\theta} \mathcal{L}(\Lambda, \mathcal{D}, \theta). \quad (13)$$

The topological tension between the local minimizers θ_i^* and the global minimizer θ^* is regulated by the within-round aggregation depth J , a phenomenon we analytically characterize in the subsequent convergence analysis.

III. Decentralized Training for the DSC Model

Every client within the fully decentralized DSC network independently tackles a personalized empirical risk minimization problem. Consequently, the sharing of knowledge

across the network is dictated exclusively by peer-to-peer communications. Fig. 3 breaks down our DFL training paradigm. We enforce an alternating sequence for every communication round. Local data drives the initial gradient-based semantic updates. Immediately after, the client pulls in peer parameters, fusing them through dynamic, similarity-aware weighting.

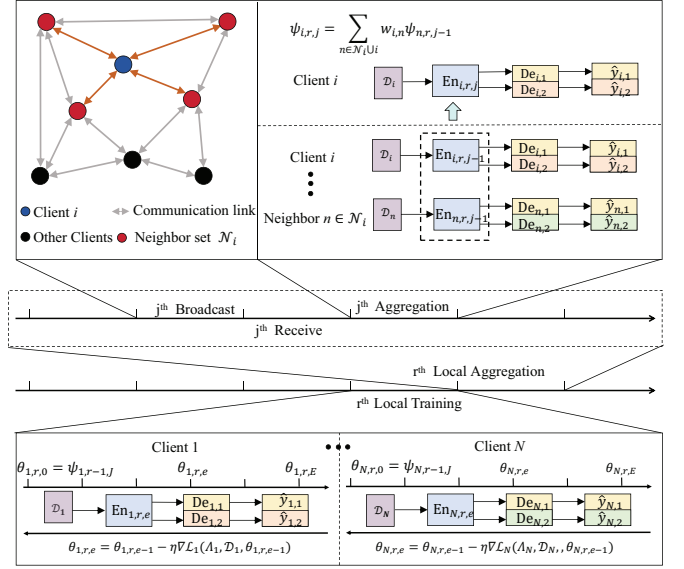


Fig. 3. Workflow of the proposed “communication-while-aggregation” protocol for the distributed personalized SC system.

A. Local Model Training

In the DSC network, clients perform parallel training by iteratively updating aggregatable model parameters θ_i via stochastic gradient descent (SGD).

Each client i performs R communication rounds during the training process. In the r -th communication round, each client i executes E local epochs based on its local dataset $\mathcal{D}_i$. The local training starts with its locally aggregated model from the $(r-1)$ -th communication round. During local training, the model at the e -th epoch, denoted by $\theta_{i,r,e}$, is updated via

$$\theta_{i,r,e} = \theta_{i,r,e-1} - \eta \nabla \mathcal{L}_i(\Lambda_i, \mathcal{D}_i, \theta_{i,r,e-1}), \quad (14)$$

where η is the learning rate, and $\nabla(\cdot)$ denotes the stochastic gradient. Consequently, the output parameter of the local training phase in round r after E epochs is given by

$$\theta_{i,r,E} = \theta_{i,r,0} - \eta \sum_{e=0}^{E-1} \nabla \mathcal{L}_i(\Lambda_i, \mathcal{D}_i, \theta_{i,r,e}). \quad (15)$$

To quantify the structural divergence inherent in heterogeneous decentralized training, we introduce an idealized reference trajectory. Assuming client i evaluates gradients at the instantaneous network average $\bar{\theta}_{r,e} = \frac{1}{N} \sum_{i \in \mathcal{N}} \theta_{i,r,e}$

rather than its local state, the resulting synchronized reference state $\hat{\theta}_{i,r,E}$ is formulated as

$$\hat{\theta}_{i,r,E} = \theta_{i,r,0} - \eta \sum_{e=0}^{E-1} \nabla \mathcal{L}_i(\Lambda_i, \mathcal{D}_i, \bar{\theta}_{r,e}). \quad (16)$$

This reference model establishes a rigorous mathematical anchor to bound the consensus error and characterize the aggregation trade-offs in the subsequent theoretical analysis.

B. Local Model Aggregation

We adopt a ‘‘communication-while-aggregation’’ protocol as our decentralized aggregation scheme. This peer-to-peer gossip protocol is scalable, fault-tolerant, and inherently bypasses the central bottleneck. During the aggregation phase of each communication round, each client transmits its task-specific policy vectors together with the current shared encoder parameters to its one-hop neighbors. The received shared encoder parameters are then integrated over J consecutive peer-to-peer aggregation steps according to

$$\psi_{i,r,j} = \sum_{n \in \mathcal{N}_i \cup \{i\}} w_{i,n} \psi_{n,r,j-1}, \quad (17)$$

where $\psi_{i,r,0} = \theta_{i,r,E}$, and $w_{i,n}$ is the aggregation weight denoting the fraction of model information transmitted from neighbor n to client i .

To prioritize semantically aligned updates while preserving the exact global parameter average during communication, the raw similarities $S_{i,n}$ defined in Eq. (8) are converted into normalized aggregation coefficients. Specifically, we apply a column-wise softmax operation over the outgoing edges of each sender n , expressed as

$$w_{i,n} = \frac{\exp(S_{i,n}/\tau)}{\sum_{m \in \mathcal{N}_n \cup \{n\}} \exp(S_{m,n}/\tau)}, \quad (18)$$

where $\tau > 0$ is a temperature parameter, and the self-similarity $S_{n,n}$ is defined as the maximum possible affinity metric. This guarantees that $\sum_{i \in \mathcal{N}_n \cup \{n\}} w_{i,n} = 1$ for any client n . In this protocol, cross-client knowledge enters a local model through neighborhood aggregation. Following the usual interpretation of negative transfer [33], aggregation-induced negative transfer refers to the case where received neighbor updates are poorly aligned with a client’s local task requirements and degrade its local task performance. This degradation is induced by peer-to-peer model mixing under heterogeneous task portfolios, rather than by multi-task learning itself. The similarity-aware weights in Eq. (18) are designed to mitigate the influence of such incompatible updates during aggregation.

This successive aggregation model allows each client to implicitly incorporate multi-hop information without establishing global connections. Let $\Psi_{r,j} = [\psi_{1,r,j}, \dots, \psi_{N,r,j}]^\top \in \mathbb{R}^{N \times M}$ denote the collective parameter matrix. We define the global aggregation weight matrix as $\mathbf{W} = \{w_{i,n}\}_{i,n=1}^N \in \mathbb{R}^{N \times N}$, where $w_{i,n} = 0$ for any $i \notin \mathcal{N}_n \cup \{n\}$. By construction via Eq. (18), $\mathbf{W}$ is a strictly column-stochastic matrix, satisfying $\mathbf{1}_N^\top \mathbf{W} = \mathbf{1}_N^\top$ with $\mathbf{1}_N = [1, \dots, 1]^\top$. This column-stochastic property

ensures that the global arithmetic mean of the network parameters is perfectly preserved across any aggregation step j , i.e., $\frac{1}{N} \sum_{i=1}^N \psi_{i,r,j} = \frac{1}{N} \sum_{i=1}^N \psi_{i,r,j-1}$.

Assuming a fixed and strongly connected topology, the iterative aggregation obeys

$$\Psi_{r,j} = \mathbf{W} \Psi_{r,j-1} = (\mathbf{W})^j \Psi_{r,0}. \quad (19)$$

By the Perron-Frobenius theorem, the irreducible column-stochastic matrix $\mathbf{W}$ possesses a unique positive right stationary eigenvector $\mathbf{v} \in \mathbb{R}^N$ satisfying $\mathbf{W}\mathbf{v} = \mathbf{v}$, where $v_i > 0$ and $\mathbf{1}_N^\top \mathbf{v} = 1$. Let $\rho = \max\{|\rho_W| : \rho_W \in \text{spec}(\mathbf{W}), \rho_W \neq 1\} < 1$ denote the subdominant spectral radius governing the mixing rate. The powers of $\mathbf{W}$ geometrically converge to a rank-one consensus projector expressed as

$$\lim_{j \rightarrow \infty, j \in \mathbb{N}_+} (\mathbf{W})^j = \mathbf{v} \mathbf{1}_N^\top. \quad (20)$$

Under the ‘‘communication-while-aggregation’’ protocol, executing J aggregations traversing multi-hop paths exponentially suppresses the topology-dependent variance. However, in heterogeneous multi-task settings, an unbounded J aggressively pulls personalized models toward a weighted global state $\mathbf{v} \mathbf{1}_N^\top \Psi_{r,0}$, amplifying the structural OCB. This structural tension necessitates the existence of an optimal, finite J^* .

Algorithm 1 Communication-while-Aggregation for Personalized DFL

Input: $\mathcal{N}, \mathcal{D}, \theta_0, \mathcal{T}, \Lambda, R, E, J, \eta$.

Output: Trained personalized local models θ_R .

- 1: Clients initialize shared backbone parameters θ_0 .
 - 2: for each communication round $r = 1, \dots, R$ do
 - 3: for each client $i \in \mathcal{N}$ in parallel do
 - 4: Perform parallel local training to obtain $\theta_{i,r,E}$ by executing E epochs of SGD on $\mathcal{D}_i$ via Eq. (14).
 - 5: Initialize the local aggregation buffer with state, $\psi_{i,r,0} \leftarrow \theta_{i,r,E}$.
 - 6: for each aggregation step $j = 1, \dots, J$ do
 - 7: Broadcast current local state $\psi_{i,r,j-1}$ and synchronize with all semantically aligned one-hop neighbors $n \in \mathcal{N}_i$.
 - 8: Compute the updated mixing state $\psi_{i,r,j}$ by performing similarity-aware neighborhood aggregation via Eq. (17).
 - 9: end for
 - 10: Set the local model $\theta_{i,r+1,0} \leftarrow \psi_{i,r,J}$ for the subsequent training round.
 - 11: end for
 - 12: end for
-

IV. Theoretical Formulation and Convergence Analysis

We study a DSC network under the ‘‘communication-while-aggregation’’ protocol. To simultaneously reflect global learning and client-level personalization, we introduce a round-wise composite objective that couples the loss at the round mean iterate with a consensus-error term measuring post-aggregation disagreement. Based on this

objective, we establish convergence guarantees in the DSC setting and further characterize how the aggregation depth J affects the steady-state performance.

A. Composite Objective for Global Generalization and Personalization

After establishing the client-wise personalized objective in Eq. (12), evaluating whether decentralized training reaches a satisfactory operating point requires jointly accounting for global generalization and client-level personalization. In distributed SC with heterogeneous task portfolios, the minimizer of the global objective in Eq. (13) differs from the collection of client-specific minimizers. As a result, optimizing the global loss alone may still leave many clients far from their preferred personalized solutions.

To capture this global-personalization discrepancy within each communication round, we consider the model parameters after J aggregations in round r . Let $\psi_{i,r,J}$ denote the model parameter of client i , and let $\bar{\psi}_{r,J} = \frac{1}{N} \sum_{i \in \mathcal{N}} \psi_{i,r,J}$ be their network average. Because the aggregation weight matrix $\mathbf{W}$ is strictly column-stochastic, the intra-round aggregation rigorously preserves the global arithmetic mean according to

$$\bar{\psi}_{r,J} = \bar{\theta}_{r,E}, \quad (21)$$

which implies that the network average after J topological mixings exactly coincides with the global mean state immediately after the local training phase.

Building on this structural property, the exact global loss parameterized by the dispersed local models can be algebraically decomposed into the loss at the global mean and a personalization-induced variation term, formulated as

$$\begin{aligned} \mathcal{L}(\Lambda, \mathcal{D}, \Psi_{r,J}) &= \mathcal{L}(\Lambda, \mathcal{D}, \bar{\psi}_{r,J}) \\ &+ \frac{1}{N} \sum_{i \in \mathcal{N}} \left[\mathcal{L}_i(\Lambda_i, \mathcal{D}_i, \psi_{i,r,J}) - \mathcal{L}_i(\Lambda_i, \mathcal{D}_i, \bar{\psi}_{r,J}) \right], \end{aligned} \quad (22)$$

where $\Psi_{r,J} = \{\psi_{i,r,J} \mid i \in \mathcal{N}\}$. The first term reflects the generalized optimization progress, while the second term captures the exact loss deviation induced by task heterogeneity.

To analytically upper-bound this exact deviation, we invoke the L_{eff} -smoothness property (formally defined in Assumption 1). For any client i , the smoothness inequality yields

$$\begin{aligned} \mathcal{L}_i(\Lambda_i, \mathcal{D}_i, \psi_{i,r,J}) &\leq \mathcal{L}_i(\Lambda_i, \mathcal{D}_i, \bar{\psi}_{r,J}) + \frac{L_{\text{eff}}}{2} \|\psi_{i,r,J} - \bar{\psi}_{r,J}\|^2 \\ &+ \langle \nabla \mathcal{L}_i(\Lambda_i, \mathcal{D}_i, \bar{\psi}_{r,J}), \psi_{i,r,J} - \bar{\psi}_{r,J} \rangle. \end{aligned} \quad (23)$$

When averaging Eq. (23) over all clients, the residual inner product term does not trivially vanish due to the non-IID data distributions. However, by applying the Cauchy-Schwarz and Young's inequalities, this inner product can be strictly decoupled and bounded by a combination of the squared gradient norms and the squared parameter dispersion.

This mathematical decoupling reveals that the exact loss deviation in Eq. (22) is dominated by the network consensus variance. We define this round-wise consensus error as

$$E_{\text{cons},r,J} \triangleq \frac{1}{N} \sum_{i \in \mathcal{N}} \|\psi_{i,r,J} - \bar{\psi}_{r,J}\|^2, \quad (24)$$

where $E_{\text{cons},r,J}$ serves as a tractable metric for geometric disagreement across clients.

Because the exact physical loss $\mathcal{L}(\Lambda, \mathcal{D}, \Psi_{r,J})$ is difficult to optimize directly in a decentralized topology, the aforementioned bounding technique allows us to construct a surrogate upper bound. We formulate the following global- and personalization-oriented composite Lyapunov potential function $\Phi_{r,J}$ given by

$$\Phi_{r,J} = \mathcal{L}(\Lambda, \mathcal{D}, \bar{\theta}_{r,E}) + \alpha E_{\text{cons},r,J}, \quad (25)$$

where $\alpha > 0$ is a composite penalty weight incorporating the Lipschitz constant L_{eff} and Young's inequality scaling factors. Our convergence analysis focuses on the Lyapunov potential $\Phi_{r,J}$, whose minimization ensures descent of the exact loss: it drives the network mean toward a global stationary point while J constrains personalized models within a bounded consensus neighborhood.

B. Convergence Analysis

We now formally analyze the convergence behavior of the proposed DFL system under heterogeneous multi-task settings. To simplify the subsequent derivations, we introduce the following shorthand notations for the task-wise and client-wise objective functions: $\mathcal{L}_{i,t}(\theta_{i,t}) \triangleq \mathcal{L}_{i,t}(\lambda_{i,t}, \mathcal{D}_{i,t}, \theta_{i,t})$ and $\mathcal{L}_i(\theta_i) \triangleq \mathcal{L}_i(\Lambda_i, \mathcal{D}_i, \theta_i)$.

Definition 1 (Statistical Heterogeneity). At a common reference parameter vector θ , let $\nabla \mathcal{L}_i(\theta)$ denote the exact gradient of the local objective of client i . Define the mean gradient over all clients as

$$\bar{g}(\theta) \triangleq \frac{1}{N} \sum_{i \in \mathcal{N}} \nabla \mathcal{L}_i(\theta). \quad (26)$$

The statistical heterogeneity at θ is quantified by the gradient dispersion formulated as

$$\Gamma(\theta) \triangleq \frac{1}{N} \sum_{i \in \mathcal{N}} \|\nabla \mathcal{L}_i(\theta) - \bar{g}(\theta)\|^2. \quad (27)$$

Definition 2 (Structural Heterogeneity and OCB). Let θ_i^* be a minimizer of the local objective for client i , and define the global average of these client-specific optima as

$$\bar{\theta}^* \triangleq \frac{1}{N} \sum_{i \in \mathcal{N}} \theta_i^*. \quad (28)$$

The structural heterogeneity is measured by the geometric dispersion of these client optima given by

$$\Omega_*^2 \triangleq \frac{1}{N} \sum_{i \in \mathcal{N}} \|\theta_i^* - \bar{\theta}^*\|^2. \quad (29)$$

We refer to Ω_*^2 as the OCB. When $\Omega_*^2 > 0$, a single consensus model cannot simultaneously match all

client optima. From the aggregation perspective, Ω_*^2 measures the dispersion of client-specific personalized optima around a common consensus point. When Ω_*^2 is small, the personalized optima are relatively close, and additional aggregation is less likely to introduce a large structural bias. A larger Ω_*^2 reflects stronger disagreement among local optima. In that case, deeper mixing can reduce consensus variance, but it may also move local models away from their preferred personalized solutions. This persistent deviation corresponds to the OCB term in our convergence analysis.

Assumption 1 (Lipschitz Smoothness). For client i and local task t , the task-specific loss function is $L_{i,t}$ -Lipschitz smooth. Specifically, for any two model parameters $\theta_{i,t}$ and $\theta'_{i,t}$, it holds that

$$\|\nabla \mathcal{L}_{i,t}(\theta_{i,t}) - \nabla \mathcal{L}_{i,t}(\theta'_{i,t})\| \leq L_{i,t} \|\theta_{i,t} - \theta'_{i,t}\|. \quad (30)$$

The effective smoothness constant across all clients and tasks is defined as $L_{\text{eff}} \triangleq \max_i \sum_{t \in \mathcal{T}_i} L_{i,t}$.

Assumption 2 (Polyak-Łojasiewicz (PL) Condition). The aggregate local objective $\mathcal{L}_i$ for each client i , as well as the global objective $\mathcal{L}$, satisfy the PL condition [34]. Specifically, there exists an effective PL constant $\mu_{\text{eff}} > 0$ such that for any parameter θ

$$2\mu_{\text{eff}}(\mathcal{L}(\theta) - \mathcal{L}^*) \leq \|\nabla \mathcal{L}(\theta)\|^2, \quad (31)$$

where $\mathcal{L}^*$ denotes the optimal value of the corresponding loss function. This condition is used as a regularity assumption for the convergence argument, rather than as a global description of the full neural objective.

Remark 1 (Role of PL condition). The PL condition is introduced as a regularity condition for the convergence argument, rather than as a claim that the full neural objective is globally well behaved over the entire parameter space. PL-type inequalities are commonly used to establish linear convergence for non-convex objectives without requiring convexity [34], and related analyses have also discussed such behaviors in over-parameterized nonlinear models [35]. In our analysis, this condition is only invoked when converting the Lyapunov drift relation into the linear objective-gap recursion in Theorem 1. The preceding drift bound does not rely on this conversion; without the PL step, it can still be interpreted as a stationarity-oriented result in terms of the gradient norm and consensus error.

Assumption 3 (Unbiased Stochastic Gradients and Bounded Variance). For each client i , the stochastic gradient $g_{i,r,e}$ at local step e of round r is an unbiased estimator of the true local gradient, satisfying

$$\mathbb{E}[g_{i,r,e} \mid \theta_{i,r,e}] = \nabla \mathcal{L}_i(\theta_{i,r,e}). \quad (32)$$

Furthermore, its conditional variance is uniformly bounded as

$$\mathbb{E}[\|g_{i,r,e} - \nabla \mathcal{L}_i(\theta_{i,r,e})\|^2 \mid \theta_{i,r,e}] \leq \sigma_i^2, \quad (33)$$

where $\sigma_i^2 \geq 0$ represents the local gradient variance bound

for client i , and $\sigma^2 = \frac{1}{N} \sum_{i \in \mathcal{N}} \sigma_i^2$ denotes the network-wide average gradient noise level.

Assumption 4 (Stepsize). The per-round learning rate is chosen as $\eta = \kappa \cdot \frac{1-\rho^{2J}}{L_{\text{eff}} E}$, where $0 < \kappa \leq \kappa_{\text{max}} = \min\{\frac{1}{4}, \frac{\mu_{\text{eff}}}{4L_{\text{eff}}}\}$.

Assumption 5 (Bounded Statistical Heterogeneity). The statistical heterogeneity measure defined in Eq. (27) is uniformly bounded. There exists such a constant $\hat{\Gamma} \geq 0$ that $\Gamma(\theta) \leq \hat{\Gamma}$ for all θ .

Remark 2 (Statistical and structural heterogeneity). The quantity $\Gamma(\theta)$ measures the dispersion of local gradients evaluated at a common parameter θ , and therefore characterizes the statistical mismatch among client objectives during local optimization. Similar bounded dissimilarity or bounded gradient-dissimilarity conditions are widely used in heterogeneous FL analyses to make use of non-IID data analytically tractable [36], [37]. In the proposed DSC setting, this gradient-level mismatch arises from both heterogeneous data distributions and heterogeneous task portfolios. Since the clients optimize over a shared semantic task space through a common multi-path encoder, $\hat{\Gamma}$ provides an effective upper bound on the gradient dispersion along the considered training trajectory. This statistical heterogeneity is distinct from the structural heterogeneity measured by Ω_*^2 , which describes the dispersion of client-specific optima and gives rise to the OCB term. Consequently, $\hat{\Gamma}$ affects the noise- and variance-related constants in the convergence bound, whereas Ω_*^2 determines the bias induced by excessive aggregation depth.

The coupled dynamics of local training and topology-aware aggregation govern the convergence trajectory through two competing mechanisms. On one hand, local SGD minimizes the objective but inevitably injects structural disagreement due to data heterogeneity and gradient noise. On the other hand, the iterative gossip protocol contracts this disagreement, though overly aggressive mixing may trigger an irreducible structural bias. To rigorously decouple these effects, we first establish a cross-round recursion for the consensus error.

Lemma 1 (Cross-Round Recursion of Consensus Error). Under Assumptions 1–4, there exist universal constants $a_1 > 0$, $a_2 > 0$, and $c_{\text{OCB}} > 0$ such that for all $r \geq 0$ the consensus error is bounded by

$$E_{\text{cons},r+1,J} \leq \rho^{2J} \left(1 + a_1 \eta^2 E^2 L_{\text{eff}}^2 \right) E_{\text{cons},r,J} + \rho^{2J} a_2 \eta^2 E^2 (\hat{\Gamma} + \sigma^2) + c_{\text{OCB}} (1 - \rho^{2J}) \Omega_*^2. \quad (34)$$

Proof: See Appendix A. ■

In the recursion bound presented above, the first term on the right-hand side (RHS) of Eq. (34) characterizes the expansion of disagreement due to E epochs, followed by the contraction induced by J gossip steps. The second term on the RHS captures the divergence injected by gradient noise and statistical heterogeneity, while the

final term mathematically isolates the irreducible OCB. This explicit term-by-term separation forms the quantitative foundation for analyzing how the aggregation depth J shapes network coherence. By substituting this disagreement evolution into a smoothness-based descent inequality, we derive a unified Lyapunov drift bound.

Lemma 2 (Per-Round Lyapunov Drift Bound). We define the expected per-round drift as $\Delta\Phi_r \triangleq \mathbb{E}[\Phi_{r+1,J} - \Phi_{r,J}]$. Under Assumptions 1–4, there exist universal constants $c_{\Phi,1}, c_{\Phi,3} \in (0,1)$ and $c_{\Phi,2}, C_{\mathcal{L}} > 0$ such that if the personalization weight α satisfies the admissible condition given by

$$\alpha \left((1 - \rho^{2J}) - \rho^{2J} a_1 \eta^2 E^2 L_{\text{eff}}^2 \right) \geq (2C_{\mathcal{L}} + c_{\Phi,3}) \eta E L_{\text{eff}}^2, \quad (35)$$

then the drift is strictly bounded by

$$\begin{aligned} \Delta\Phi_r \leq & -c_{\Phi,1} \eta E \mathbb{E} \|\nabla \mathcal{L}(\bar{\theta}_{r,E})\|^2 - c_{\Phi,3} (1 - \rho^{2J}) \mathbb{E}[E_{\text{cons},r,J}] \\ & + c_{\Phi,2} \eta E (\hat{\Gamma} + \sigma^2) + \alpha c_{\text{OCB}} (1 - \rho^{2J}) \Omega_*^2. \end{aligned} \quad (36)$$

Proof: See Appendix B. ■

Lemma 2 successfully couples global optimization progress with consensus error contraction. By invoking the PL condition on this drift, we can formally extract the linear convergence rate and the exact J -dependent steady-state error floor.

Theorem 1 (Linear Convergence and Trade-off Characterization). Under Assumptions 1–5, and for any $\alpha > 0$ satisfying Eq. (35), the composite objective $\Phi_{r,J}$ converges linearly to a steady-state error floor, as given by

$$\begin{aligned} \mathbb{E}[\Phi_{r,J} - \mathcal{L}^*] \leq & (1 - \kappa_{\text{lin}})^r (\Phi_{0,J} - \mathcal{L}^*) + C_{\text{base}} \\ & + \frac{C_{\Gamma}}{1 - \rho^{2J}} + \alpha C_{\Omega} (1 - \rho^{2J}) \Omega_*^2, \end{aligned} \quad (37)$$

where $\kappa_{\text{lin}} = c_{\Phi,1} \mu_{\text{eff}} \eta E$ defines the linear contraction rate. The geometric bounding constants evaluate to $C_{\text{base}} = \frac{\eta c_{\Phi,2}}{c_{\Phi,1} \mu_{\text{eff}}} (\hat{\Gamma} + \sigma^2)$, $C_{\Gamma} = \frac{\eta^2 C_{\text{var}}}{c_{\Phi,1} \mu_{\text{eff}}} (\hat{\Gamma} + \sigma^2)$, and $C_{\Omega} = \frac{c_{\text{OCB}}}{c_{\Phi,1} \mu_{\text{eff}}}$. All three terms remain strictly positive. Structurally, they are entirely decoupled from both the communication round index r and the aggregation depth J .

Proof: See Appendix C. ■

Theorem 1 exposes a mechanical conflict driven by the aggregation depth J . Pushing communication deeper suppresses the topology-dependent consensus variance. However, it simultaneously inflates the structural OCB. This U-shaped error floor forces a critical design choice: resource-constrained edge deployments must lock into an optimal depth J^* to prevent model divergence.

C. Optimal Aggregation Depth via Upper Bound Minimization

To resolve this structural tension, we must isolate the terms actively controlled by J inside the theoretical upper bound. High task heterogeneity violently scatters the local optima across the parameter space. This exact scattering dictates the severe OCB penalty during deep mixing. We therefore extract J as our primary decision variable.

System hardware and bandwidth budgets inherently restrict J to a discrete feasible domain:

$$J \in \{1, 2, \dots, J_{\text{max}}\}, \quad (38)$$

with $J_{\text{max}} \in \mathbb{N}_+$ acting as the hard communication ceiling. Next, we discard the intrinsic SGD noise C_{base} from the error bound. It remains entirely invariant to the mixing steps. By defining $A \triangleq C_{\Gamma}$ to capture the topological variance scale and $B \triangleq C_{\Omega} \Omega_*^2$ to quantify the structural bias, we condense the J -dependent dynamics into a simplified objective function:

$$\text{Bound}(J) \triangleq \frac{A}{1 - \rho^{2J}} + \alpha B (1 - \rho^{2J}). \quad (39)$$

Locating the optimal operating point J^* thus reduces to a constrained discrete minimization problem:

$$J^* \in \arg \min_{J \in \{1, \dots, J_{\text{max}}\}} \text{Bound}(J), \quad (40)$$

which must strictly satisfy the admissibility limits established earlier in Lemma 2.

Eq. (39) translates our U-shaped trade-off into pure algebra. Expanding J pushes the effective mixing gap $(1 - \rho^{2J})$ toward unity. The variance fraction $A/(1 - \rho^{2J})$ decays rapidly, but the structural bias term $\alpha B (1 - \rho^{2J})$ grows right alongside it. In highly heterogeneous environments where $\Omega_*^2 \gg 0$, increasing the gossip depth helps at first by crushing gradient noise. However, performance soon collapses once OCB takes over the error floor. The opposite happens in strictly homogeneous settings where $\Omega_*^2 \rightarrow 0$. Here, the bias penalty disappears entirely, meaning the network always benefits from a larger J .

Theorem 2 (Optimal Aggregation Depth J^*). Let us relax the steady-state objective in Eq. (39) to the continuous domain. If structural heterogeneity is strong enough to satisfy the interior condition $\alpha B > A$, the unique global minimizer is given by

$$J^* = \frac{1}{-2 \ln \rho} \ln \left(\frac{1}{1 - \sqrt{\frac{A}{\alpha B}}} \right). \quad (41)$$

Otherwise, if the network is highly homogeneous such that $\alpha B \leq A$, the function $\text{Bound}(J)$ decreases monotonically. In this scenario, the theoretical optimal aggregation depth has no upper bound.

Proof: See Appendix D. ■

Theorem 2 provides a direct configuration rule for system deployment. When $\alpha B > A$, Eq. (41) pinpoints the exact boundary where variance reduction perfectly balances structural bias. Because practical communication depths require integer values, the actual deployed J is obtained by testing the two nearest integers to the continuous root J^* . The final value is strictly capped by the hardware threshold J_{max} . This explicit configuration maps the theoretical minimum directly to real-world edge bandwidth constraints.

V. Experiments

This section evaluates the similarity-aware aggregation protocol against standard peer-to-peer baselines. The primary goal is to verify the multi-task performance gains within heavily personalized client networks. We sweep the aggregation depth J to observe the convergence behavior. These empirical observations directly validate the theoretical optimal mixing depth derived in Section IV-C. Finally, we examine whether the observed gains remain stable under network-size variation and how imperfect wireless links affect packet loss and communication latency.

A. Experimental Setup

Datasets and task settings. We conduct experiments on both NYU-v2 and Taskonomy. The NYU-v2 benchmark is used as the main testbed, where we evaluate the effectiveness of the proposed similarity-aware aggregation, examine the convergence behavior, and study the influence of the aggregation depth J . NYU-v2 contains 795 training and 654 test images with dense annotations for semantic segmentation (Seg), depth estimation (Dep), and surface normal prediction (Sn). Unless otherwise specified, the simulated DSC network on NYU-v2 consists of $N = 10$ clients.

To examine whether the observed gains remain stable beyond this setting, we also conduct supplementary experiments on Taskonomy. This dataset introduces a different multi-task composition, including Seg, Sn, key-point detection (KP), and edge detection (Edge). In the Taskonomy experiments, we vary the number of clients over $N \in [2, 30]$.

Communication setup. We construct the decentralized topology from randomly generated client locations in a two-dimensional area. Rather than assuming a fully connected mesh, we use a random geometric graph to reflect distance-constrained peer-to-peer communication. Two clients are connected when their distance falls within the communication radius. For each topology realization, the radius is gradually increased until the graph becomes connected, and the resulting adjacency defines the one-hop neighbor set $\mathcal{N}_i$ used for decentralized aggregation.

For wireless access, we adopt the TDMA-based coloring schedule described in the system model. Feasible one-hop links are assigned to time slots such that conflicting transmissions are separated. This setting provides a controlled abstraction of local wireless contention, while keeping the comparison focused on how different aggregation rules exploit the same neighborhood structure.

We further incorporate packet-level reliability into the communication evaluation. For communication-volume accounting, both the shared encoder parameters and the routing-policy entries are represented using 32-bit floating-point values. Each transmitted message is divided into fixed-length packets, and the packet number on each link is computed from the message size. The packet loss probability of a valid one-hop link depends on the nominal regional SNR, the inter-client distance, and the packet

length. For latency evaluation, links assigned to the same TDMA color slot are treated as parallel transmissions, and the slot duration is determined by the maximum link latency among the active links in that slot. The latency of one aggregation step is obtained by summing the durations of all color slots, while the total communication latency is accumulated over the required communication rounds and aggregation steps. In the communication-quality experiments, both packet loss probability and communication latency are averaged over random topology realizations.

Implementation Details. The local models utilize a DeepLab-ResNet-34 backbone. For a given client i , the backbone splits into $|\mathcal{T}_i|$ task-specific decoders. We train the network weights using standard SGD and update the routing policy parameters via the Adam optimizer. Both optimizers are initialized with a learning rate of 1×10^{-4} . The training process spans $R = 60$ global communication rounds. During each round, clients complete $E = 10$ local epochs before triggering the gossip aggregation. The aggregation depth J is evaluated over the interval $[1, 15]$.

Evaluation Metrics. We track the global loss convergence alongside the final task-specific performance. Each downstream task relies on its standard evaluation metrics. Seg is evaluated via mean intersection over union (mIoU) and pixel accuracy (PAcc). Sn performance is measured using mean and median angular errors, plus the accuracy thresholds at 11.25° , 22.5° , and 30° . For Dep, we calculate absolute (Abs) and relative (Rel) errors, along with the threshold percentage $\delta \in \{1.25, 1.25^2, 1.25^3\}$. For KP and Edge, we calculate the L_1 loss, where lower values indicate better performance.

We establish a strict no-aggregation baseline to isolate the collaborative gains. Let $b_{t,q}$ denote this baseline for metric q on task t . If client i achieves an experimental measurement $m_{i,t,q}$, its relative improvement is calculated as $\Delta_{i,t,q} = (m_{i,t,q} - b_{t,q})/b_{t,q} \times 100\%$. The evaluation metrics scale in opposite directions; higher is better for accuracy, while lower is better for errors. A directional multiplier $I_{t,q} \in \{-1, 1\}$ is therefore introduced to unify the metric orientation. The final overall improvement is computed through the following hierarchical aggregation:

$$\Delta_{i,t} = \frac{1}{|Q_t|} \sum_{q \in Q_t} I_{t,q} \Delta_{i,t,q}, \quad (42)$$

$$\Delta_t = \frac{1}{|\mathcal{N}_t|} \sum_{i \in \mathcal{N}_t} \Delta_{i,t}, \quad (43)$$

$$\Delta_{\text{all}} = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \Delta_t, \quad (44)$$

Here, Q_t is the specific metric set for task t . The subset of clients assigned to task t is denoted by $\mathcal{N}_t$, and $\mathcal{T}$ represents the comprehensive set of all tasks. The relative improvement scores also reflect the effect of negative transfer during aggregation. Since the non-aggregation model is used as the no-transfer reference, $\Delta_{i,t} < 0$ indicates that aggregation degrades task t on client i . Here, Δ_t measures the average transfer effect for a specific task; Δ_{all} reports the network-level average across all tasks. A positive Δ_{all} means that aggregation is beneficial on

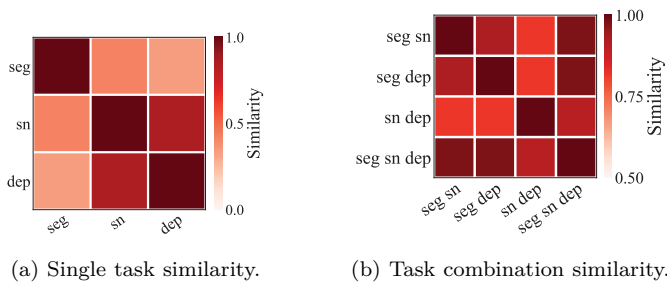


Fig. 4. Task Correlation Heatmap.

average, although some individual clients or tasks may suffer from negative transfer.

B. Results and Analysis

Most existing PFL frameworks assume single-task setups or rely on central orchestration. They fail in decentralized MT-SC environments. We benchmark our framework against four standard peer-to-peer alternatives to highlight the necessity of task-aware routing:

- 1) No-aggregation: The strict performance lower bound. Clients train in total isolation.
- 2) Decentralized FedAvg [38]: A topology-agnostic baseline that weights all neighbors uniformly, ignoring task differences entirely.
- 3) Heuristic max aggregation: A naive task-aware approach. Rather than evaluating the full portfolio, it defines peer weights using only the single highest task similarity found between two clients.
- 4) FedAMP [19]: A PFL method based on attentive message passing. It measures the similarity between client models and gives higher weights to more related clients during aggregation.

Table 2 reports the relative gains against the isolated baseline for both shallow ($J = 1$) and deep ($J = 5$) mixing configurations. To contextualize these results, we must reference the task affinities in Fig. 4. The matrix confirms a strong semantic correlation linking Sn with Dep. The stronger correlation between Sn and Dep is consistent with prior visual task-transfer studies [16]. Seg sits apart, showing weak affinity to both. This specific structural asymmetry dictates the severity of the OCB penalty. As the network mixes deeper, aggregation rules that do not explicitly model task-path compatibility are more likely to introduce negative transfer. Our similarity-aware design inherently blocks this, distributing collaborative gains evenly without triggering sudden performance collapses.

Shallow Aggregation. Under limited mixing, Decentralized FedAvg delivers only a marginal overall improvement of 1.17%. Uniform neighbor weighting indiscriminately injects mismatched semantic features, which severely dilutes the potential collaborative gain. The greedy max strategy improves Seg by 2.42% and Sn by 1.99%, but does so at the direct expense of Dep, whose performance drops by 0.12%. This result confirms that Dep is particularly sensitive to representation drift. Absorbing parameters

from an uncorrelated task such as Seg actively degrades its geometric structure. FedAMP improves the overall gain to 1.77%, showing that model-level attentive weighting can reduce part of the mismatch among clients. Its gain is still lower than that of our method, since the attention weights are not explicitly tied to task-path compatibility. In contrast, our proposed method avoids this failure mode and achieves a more balanced overall improvement of 2.11%.

Deep Aggregation and Cross-Task Interference. Increasing the communication depth to $J = 5$ amplifies both the variance reduction effect and the underlying structural bias. FedAvg reaches only a modest overall gain of 2.46%. Meanwhile, the max strategy induces severe cross-task interference: it pushes Seg to a peak improvement of 10.24%, but completely destabilizes Dep, whose performance falls by 1.24%. This sharp degradation highlights a critical flaw: propagating incompatible Seg features into Dep models over multiple communication hops destroys their geometric consistency. FedAMP outperforms these two baselines, reaching a 4.03% overall improvement. Its attentive weighting gives related clients more influence and therefore avoids the sharp Dep degradation observed in the max strategy. Nevertheless, the improvement remains below that of the proposed method, especially on Sn and Dep. Our routing scheme instead exploits the known Sn–Dep synergy to filter out these harmful updates. As a result, both tasks remain protected from negative transfer, with Sn improving by 2.95% and Dep by 3.31%. Although the Seg gain of our method, 8.04%, is slightly below that of the greedy baseline, it avoids the over-consensus trap and creates a substantially more stable collaborative environment, ultimately achieving the highest global improvement of 4.77%.

Convergence Behavior Under Aggregation. Fig. 5 tracks the training metrics over 60 communication rounds under a deep mixing regime. The primary goal is to confirm optimization stability despite extreme task heterogeneity across the mesh network.

Empirical Validation of the J -dependent Trade-off. Theorem 2 predicts a U-shaped relationship between performance and aggregation depth. To examine this analytical prediction, Fig. 6 plots the relative improvement versus J across four different task portfolios.

A clear non-monotonic trend is observed across all four subplots. The relative improvement initially increases with J , reaches an empirical optimum, and then declines. At small J , communication is limited, and the network operates in a variance-dominated regime in which local updates cannot absorb sufficient external knowledge. Increasing the propagation depth helps reduce this variance by enabling broader neighborhood information exchange. However, overly deep aggregation pushes the system into an OCB-dominated regime, where accumulated heterogeneous parameters begin to act as structural interference. This empirical transition closely aligns with the behavior predicted by Eq. (39).

The location of the optimal depth J^* is highly sensitive

TABLE 2: Comparison of Different Aggregation Methods.

Method	Seg			Sn							Dep					Δ_{all} (%)
	mIoU	PAcc	Δ_{Seg}	Mean $\downarrow$	Median $\downarrow$	11.25°	22.5°	30°	Δ_{Sn}	Abs $\downarrow$	Rel $\downarrow$	1.25	1.25 ²	1.25 ³	Δ_{Dep}	
No aggregation	28.46	59.21	-	16.54	13.05	44.11	73.12	83.14	-	0.565	0.223	62.38	88.72	96.71	-	-
FedAvg($J=1$)	28.89	59.72	1.19	16.29	12.83	44.80	73.62	83.66	1.21	0.557	0.219	63.30	89.16	97.14	1.12	1.17
max($J=1$)	29.31	60.31	2.42	16.30	12.66	45.36	73.78	84.60	1.99	0.570	0.220	61.44	88.81	97.06	-0.12	1.43
FedAMP($J=1$)	29.14	59.77	1.67	16.23	12.74	45.16	73.80	84.67	1.88	0.553	0.215	63.42	89.48	97.32	1.77	1.77
Similarity($J=1$)	29.26	59.92	2.01	16.22	12.67	45.33	73.89	84.84	2.14	0.551	0.212	63.49	89.67	97.37	2.19	2.11
FedAvg($J=5$)	29.37	60.87	3.00	16.18	12.68	45.25	74.50	84.85	2.31	0.551	0.217	63.94	90.22	97.60	2.06	2.46
max($J=5$)	31.87	64.24	10.24	16.16	12.77	45.06	74.02	84.90	1.99	0.572	0.226	61.63	87.66	95.55	-1.24	3.66
FedAMP($J=5$)	31.21	61.74	6.97	16.18	12.57	45.48	74.34	84.91	2.55	0.546	0.213	63.82	90.26	97.63	2.57	4.03
Similarity($J=5$)	31.43	62.55	8.04	16.15	12.51	45.85	74.69	84.93	2.95	0.541	0.209	64.26	90.49	97.68	3.31	4.77

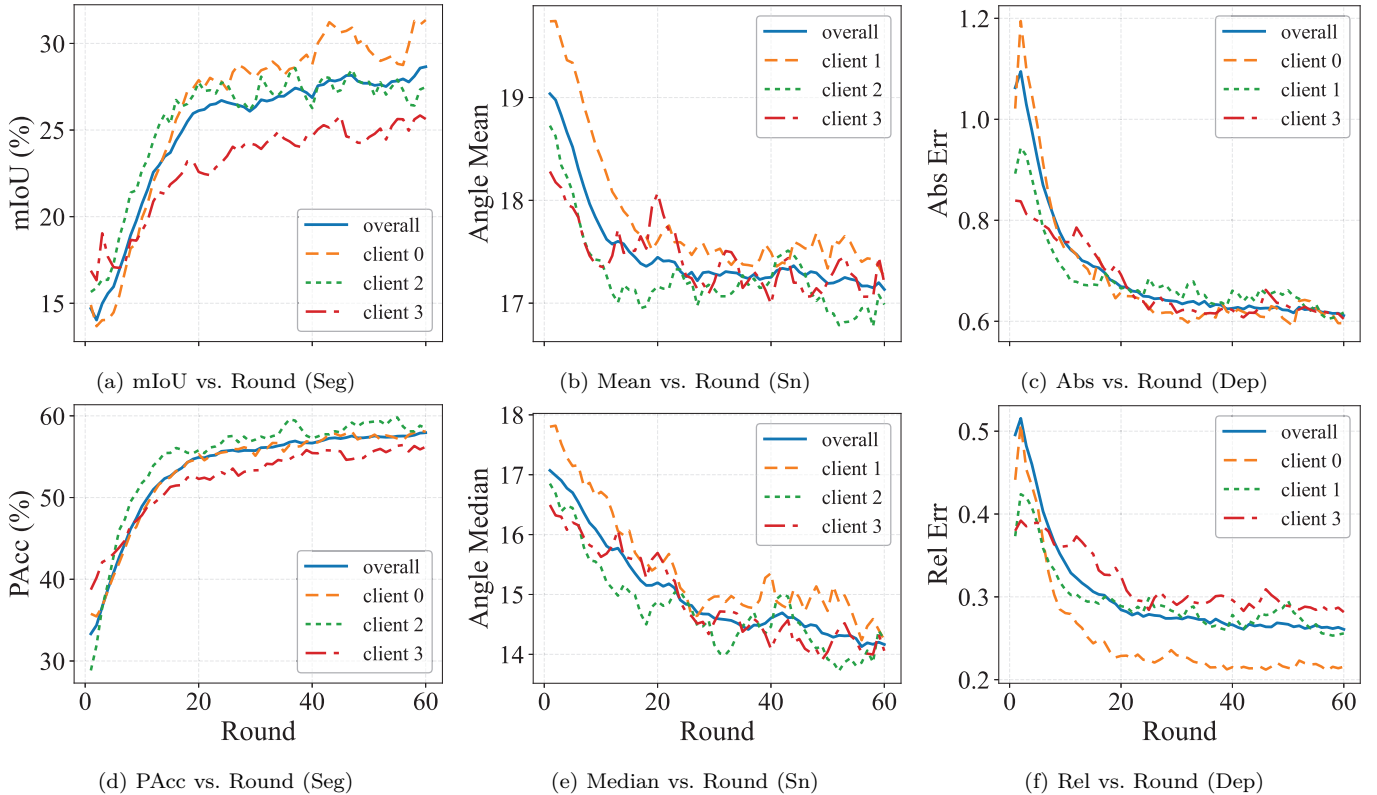


Fig. 5. Convergence behavior of different tasks (Seg, Sn, Dep) over communication rounds.

to the underlying task composition. This observation is consistent with the structural heterogeneity term Ω_*^2 in the derived upper bound. For the high-affinity “sn dep” portfolio (Fig. 6d), collaborative gains appear immediately at $J = 1$, and the post-peak degradation remains relatively mild, indicating a stronger tolerance to deeper parameter mixing. By contrast, introducing tasks with lower affinity or higher dimensionality, such as “seg sn dep” or “seg sn”, substantially changes this pattern. The optimal peak shifts leftward, and the subsequent performance drop becomes much sharper. The task composition therefore limits the effective communication depth by determining the point at which heterogeneity-induced interference outweighs the marginal benefit of additional neighborhood averaging.

Client-level results further show that local dataset size affects the maximum achievable collaborative gain. Within the “sn dep” group, the client with the richest local data consistently attains the largest peak improvement.

A larger local dataset helps anchor the optimization trajectory and makes the model more robust to heterogeneous external updates. In contrast, data-scarce clients obtain smaller peak gains and exhibit faster performance degradation as J increases, indicating greater vulnerability to externally induced OCB.

Impact of Network Size. We next use the Taskonomy setting to examine whether the average learning gain changes as the decentralized network becomes larger. As shown in Fig. 8, the proposed method achieves the highest Δ_{all} under all tested network sizes. When N increases from 10 to 20, the curves of all methods remain relatively flat. Throughout this range, the changes in Δ_{all} are 0.09%, 0.10%, 0.13%, and 0.14% for FedAvg, max, FedAMP, and the proposed similarity-aware aggregation, respectively. This indicates that, within the examined range, enlarging the decentralized network does not substantially change the average learning gain. In this case,

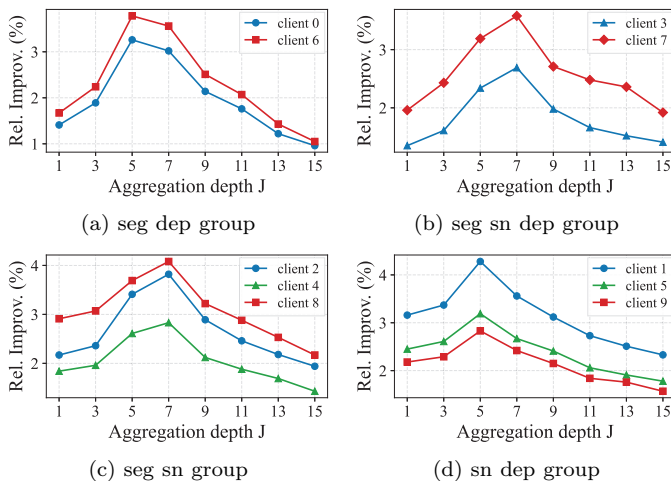


Fig. 6. Relative improvement versus aggregation depth J for different task groups.

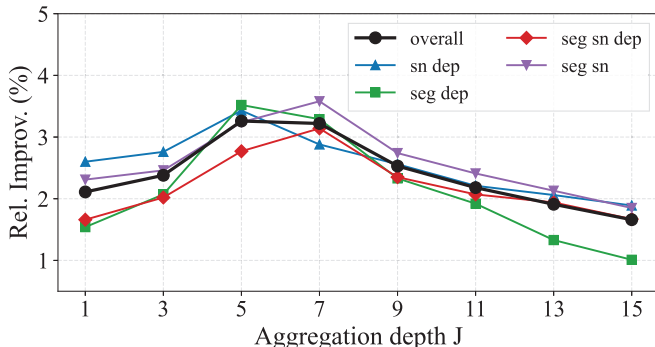


Fig. 7. Overall relative improvement versus aggregation depth J .

increasing N expands the global network scale, while the effective aggregation behavior is still governed by local neighborhood composition and the way neighbor updates are weighted. This variation is partly attributable to the proposed policy-driven multi-path encoder, which preserves task-dependent execution paths while allowing shared encoder blocks to learn reusable representations across tasks. Under this architecture, exchanging encoder parameters can bring collaborative gains to all aggregation rules, while the proposed similarity-aware aggregation benefits more by assigning larger weights to neighbors with more compatible routing-policy patterns.

These results suggest that the proposed method is not sensitive to network-size variation. Since the average local communication density is fixed, the main performance difference comes from whether the aggregation rule can select and weight compatible neighbor updates, rather than from the number of clients. This observation is consistent with the NYU-v2 results, where explicitly modeling task compatibility leads to stronger overall performance under heterogeneous task portfolios.

Communication Cost and Impact of Regional Communication Quality. Using ResNet-18 as the shared backbone, FedAvg and FedAMP transmit approximately 44.71 MB

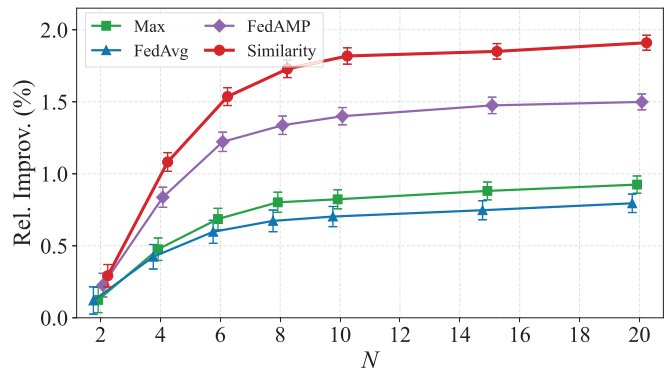


Fig. 8. Overall relative improvement with varying network size. The curves show the mean over task-assignment realizations, and the error bars indicate the standard deviation.

of model parameters in each client-to-neighbor exchange. The heuristic max and the proposed similarity-aware method additionally transmit the routing-policy vectors, while their total communication volume remains approximately 44.71 MB. For a client assigned with three tasks, these policy vectors introduce only 96 bytes of additional payload, resulting in a marginal communication-latency overhead. Thus, the proposed method achieves improved task performance under essentially the same communication budget, with the gain primarily arising from the designed aggregation mechanism rather than a larger transmitted model.

We next examine the effect of regional channel quality on packet loss and communication latency. The nominal regional SNR is varied to control the area-level channel condition, while different network sizes lead to different client densities and one-hop link distances. We set $R = 30$ and $J = 1$, and count one packet transmission as one packet-time unit.

As shown in Fig. 9(a), the average packet loss probability decreases as the regional SNR increases. The difference among network sizes mainly comes from the generated spatial topology. In a denser network, many one-hop links become shorter, so the packet loss probability can drop more quickly when the channel condition improves.

Fig. 9(b) shows how this link reliability affects communication latency. When the regional SNR is low, retransmission dominates the latency, and larger backbones suffer more because each model exchange contains more packets. As the SNR reaches a moderate range, denser networks may benefit from shorter and more reliable links, which can offset part of the additional scheduling cost. In the high-SNR region, packet loss is already low for all network sizes, and the remaining latency difference is mainly determined by the number of scheduled links and the model size.

VI. Conclusion

This paper investigated how structural heterogeneity affects decentralized multi-task optimization in DFL systems. Our analysis revealed a U-shaped trade-off in

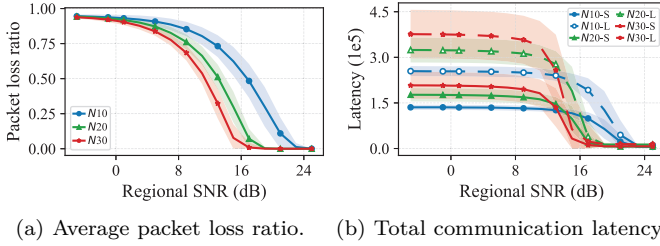


Fig. 9. Effect of regional communication quality on packet loss probability and communication latency. Here, $N10$, $N20$, and $N30$ denote networks with 10, 20, and 30 clients, respectively. S and L denote the small and large backbone settings, corresponding to ResNet-18 and ResNet-34. The curves show the mean over random topology realizations, and the shaded regions indicate the 10th-90th percentile range.

the steady-state error: increasing the aggregation depth reduces topology-dependent variance, but also amplifies OCB. To address this issue, we proposed a similarity-aware routing strategy that selectively filters structurally mismatched updates, enabling more reliable and effective decentralized collaboration.

Experiments on the NYU-v2 dataset support the theoretical analysis. When operated at the analytically derived optimal depth, the proposed framework mitigates negative transfer and consistently outperforms decentralized FedAvg, FedAMP and heuristic baselines, achieving a 4.77% global relative improvement. Results on Taskonomy further show that the proposed aggregation strategy is only mildly affected by moderate changes in network size.

Future work will extend the framework to more realistic edge settings, including time-varying wireless channels, dynamic mesh topologies, and adaptive online schemes for aggregation depth control.

Appendix

A. Proof of Lemma 1

In this proof, we derive an explicit cross-round recursion for the consensus error $E_{\text{cons},r+1,J}$ in terms of $E_{\text{cons},r,J}$ by combining a local-training disagreement expansion bound and a local-aggregation disagreement bound.

a) Local Training: Let $\theta_{i,r,e}$ be the local model of client i at step e of round r . We define the pre-aggregation consensus error at the end of local training as

$$E_{\text{cons},r,0} \triangleq \frac{1}{N} \sum_{i \in \mathcal{N}} \|\theta_{i,r,e} - \bar{\theta}_{r,e}\|^2, \quad (45)$$

where $\bar{\theta}_{r,e} \triangleq \frac{1}{N} \sum_{i \in \mathcal{N}} \theta_{i,r,e}$. Note that at the start of the round, $e = 0$, the error is $E_{\text{cons},r-1,J}$.

We first bound the expected drift of the local updates. The update rule is $\theta_{i,r,e+1} = \theta_{i,r,e} - \eta g_{i,r,e}$. The mean model evolves as $\bar{\theta}_{r,e+1} = \bar{\theta}_{r,e} - \eta \bar{g}_{r,e}$, where $\bar{g}_{r,e} \triangleq \frac{1}{N} \sum_i g_{i,r,e}$. The deviation from the mean evolves as

$$\theta_{i,r,e+1} - \bar{\theta}_{r,e+1} = (\theta_{i,r,e} - \bar{\theta}_{r,e}) - \eta(g_{i,r,e} - \bar{g}_{r,e}). \quad (46)$$

Squaring the norm and taking expectations yields the expansion given by

$$\mathbb{E}\|\theta_{i,r,e+1} - \bar{\theta}_{r,e+1}\|^2 = \mathbb{E}\|\theta_{i,r,e} - \bar{\theta}_{r,e}\|^2 + \eta^2 \mathbb{E}\|g_{i,r,e} - \bar{g}_{r,e}\|^2 - 2\eta \mathbb{E}\langle \theta_{i,r,e} - \bar{\theta}_{r,e}, g_{i,r,e} - \bar{g}_{r,e} \rangle. \quad (47)$$

Using Young's inequality $\|a - b\|^2 \leq (1 + \beta)\|a\|^2 + (1 + \frac{1}{\beta})\|b\|^2$, we simplify the analysis by bounding the expected deviation as

$$\mathbb{E}\|\theta_{i,r,e+1} - \bar{\theta}_{r,e+1}\|^2 \leq (1 + \eta^2 E L_{\text{eff}}^2) \mathbb{E}\|\theta_{i,r,e} - \bar{\theta}_{r,e}\|^2 + \left(1 + \frac{1}{\eta^2 E L_{\text{eff}}^2}\right) \eta^2 \mathbb{E}\|g_{i,r,e} - \bar{g}_{r,e}\|^2. \quad (48)$$

To handle the gradient variance term, we decompose it using the exact mean gradient $\nabla \bar{\mathcal{L}}(\theta_{r,e}) \triangleq \frac{1}{N} \sum_{j \in \mathcal{N}} \nabla \mathcal{L}_j(\theta_{j,r,e})$ expressed as

$$g_{i,r,e} - \bar{g}_{r,e} = (\nabla \mathcal{L}_i(\theta_{i,r,e}) - \nabla \bar{\mathcal{L}}(\theta_{r,e})) + g_{i,r,e} - \nabla \mathcal{L}_i(\theta_{i,r,e}) - \frac{1}{N} \sum_{j \in \mathcal{N}} (g_{j,r,e} - \nabla \mathcal{L}_j(\theta_{j,r,e})). \quad (49)$$

Applying Jensen's inequality and Assumption 3, we bound the expected squared norm of the stochastic noise by $4(\sigma_i^2 + \sigma^2)$. Similarly, the heterogeneity drift is bounded by averaging over all clients $i \in \mathcal{N}$ via the same inequality,

$$\begin{aligned} \frac{1}{N} \sum_{i \in \mathcal{N}} \|\nabla \mathcal{L}_i(\theta_{i,r,e}) - \nabla \bar{\mathcal{L}}(\theta_{r,e})\|^2 &\leq \frac{3}{N} \sum_{i \in \mathcal{N}} \|\nabla \mathcal{L}_i(\theta_{i,r,e}) - \nabla \mathcal{L}_i(\bar{\theta}_{r,e})\|^2 \\ &+ 3 \|\nabla \bar{\mathcal{L}}(\bar{\theta}_{r,e}) - \nabla \bar{\mathcal{L}}(\theta_{r,e})\|^2 + \frac{3}{N} \sum_{i \in \mathcal{N}} \|\nabla \mathcal{L}_i(\bar{\theta}_{r,e}) - \nabla \bar{\mathcal{L}}(\bar{\theta}_{r,e})\|^2 \\ &\leq 6L_{\text{eff}}^2 E_{\text{cons},r,e} + 3\hat{\Gamma}. \end{aligned} \quad (50)$$

Substituting these bounds back into (48), recursively applying it over $e = 0, \dots, E-1$, and defining appropriate bounding constants a_1 and a_2 leads to

$$E_{\text{cons},r,0} \leq a_2 \eta^2 E^2 (\hat{\Gamma} + \sigma^2) + (1 + a_1 \eta^2 E^2 L_{\text{eff}}^2) E_{\text{cons},r-1,J}. \quad (51)$$

This bound quantifies how statistical heterogeneity and gradient noise introduce disagreement into the system prior to the aggregation step.

b) Local Aggregation: Next, we determine how the disagreement contracts after J aggregation steps. Let $\Psi_{r,0} \in \mathbb{R}^{N \times M}$ be the matrix of local models before aggregation. The aggregation proceeds as $\Psi_{r,J} = \mathbf{W}^J \Psi_{r,0}$. Since $\mathbf{W}$ is strictly column-stochastic, it perfectly preserves the arithmetic mean, yielding

$$\bar{\psi}_{r,J} = \frac{1}{N} \mathbf{1}^\top \Psi_{r,J} = \frac{1}{N} \mathbf{1}^\top \mathbf{W}^J \Psi_{r,0} = \bar{\psi}_{r,0}. \quad (52)$$

The consensus error is formulated as $E_{\text{cons},r,J} = \frac{1}{N} \|(\mathbf{I} - \mathbf{M})\Psi_{r,J}\|_F^2$, where $\mathbf{M} \triangleq \frac{1}{N} \mathbf{1} \mathbf{1}^\top$ denotes the mean-centering projection matrix. To capture the impact of dispersed client optima $\Theta^* = [\theta_1^*, \dots, \theta_N^*]^\top$, we decompose the error relative to the centered optima matrix as

$$(\mathbf{I} - \mathbf{M})\Psi_{r,J} = (\mathbf{I} - \mathbf{M})\mathbf{W}^J \Theta^* + (\mathbf{I} - \mathbf{M})\mathbf{W}^J (\Psi_{r,0} - \Theta^*). \quad (53)$$

Applying the generalized triangle inequality $\|A + B\|_F^2 \leq (1 + \gamma)\|A\|_F^2 + (1 + \frac{1}{\gamma})\|B\|_F^2$ for any $\gamma > 0$ yields

$$E_{\text{cons},r,J} \leq \frac{1+\gamma}{N}\|(\mathbf{I} - \mathbf{M})\mathbf{W}^J(\Psi_{r,0} - \Theta^*)\|_F^2 + \frac{1+1/\gamma}{N}\|(\mathbf{I} - \mathbf{M})\mathbf{W}^J\Theta^*\|_F^2. \quad (54)$$

The stochastic matrix contracts the disagreement component orthogonal to the average-consensus subspace at a geometric rate ρ , i.e.,

$$\|(\mathbf{I} - \mathbf{M})\mathbf{W}^J(\mathbf{I} - \mathbf{M})X\|_F \leq \rho^J\|(\mathbf{I} - \mathbf{M})X\|_F. \quad (55)$$

Since $\mathbf{W}$ is column-stochastic but not necessarily doubly stochastic, applying $\mathbf{W}^J$ to a matrix with a nonzero mean component may additionally induce an imbalance residual $(\mathbf{I} - \mathbf{M})\mathbf{W}^J\mathbf{M}X$. In particular, when the local optima Θ^* are dispersed across clients, $\mathbf{W}^J\Theta^*$ may exhibit a persistent structural shift. While $\mathbf{W}^J$ suppresses network disagreement [39], it may simultaneously displace the iterates away from their respective empirical minimizers. Leveraging spectral decay, this OCB floor scales with the structural dispersion Ω_*^2 as

$$\Delta_{\text{OCB}}(J) \leq c_{\text{OCB}}(1 - \rho^{2J})\Omega_*^2. \quad (56)$$

The finalized consensus error bound after J mixing steps is thus given by

$$E_{\text{cons},r,J} \leq \rho^{2J}E_{\text{cons},r,0} + c_{\text{OCB}}(1 - \rho^{2J})\Omega_*^2. \quad (57)$$

This characterization reveals a dual effect: the first term suppresses initial disagreement and gradient noise, while the second introduces an irreducible bias floor dictated by the decentralized network's structural heterogeneity.

Substituting the local expansion bound (51) into the gossip contraction bound (57), we obtain

$$E_{\text{cons},r+1,J} \leq \rho^{2J}(1 + a_1\eta^2E^2L_{\text{eff}}^2)E_{\text{cons},r,J} + \rho^{2J}a_2\eta^2E^2(\hat{\Gamma} + \sigma^2) + c_{\text{OCB}}(1 - \rho^{2J})\Omega_*^2. \quad (58)$$

This explicitly separates the expansion, contraction, and bias terms, concluding the proof.

B. Proof of Lemma 2

We analyze the per-round drift $\Delta\Phi_r \triangleq \mathbb{E}[\Phi_{r+1,J} - \Phi_{r,J}]$ of the Lyapunov function $\Phi_{r,J} = \mathcal{L}(\bar{\theta}_{r,E}) + \alpha E_{\text{cons},r,J}$. Applying L_{eff} -smoothness to $\mathcal{L}$, the expected change over E local steps is given by

$$\mathbb{E}[\mathcal{L}(\bar{\theta}_{r+1,E}) - \mathcal{L}(\bar{\theta}_{r,E})] \leq \frac{L_{\text{eff}}}{2}\mathbb{E}\|\bar{\theta}_{r+1,E} - \bar{\theta}_{r,E}\|^2 + \mathbb{E}\langle \nabla \mathcal{L}(\bar{\theta}_{r,E}), \bar{\theta}_{r+1,E} - \bar{\theta}_{r,E} \rangle. \quad (59)$$

Standard bounding for decentralized settings yields the descent in terms of gradient norm, variance, and model disagreement as

$$\mathbb{E}[\mathcal{L}(\bar{\theta}_{r+1,E}) - \mathcal{L}(\bar{\theta}_{r,E})] \leq -c_{\Phi,1}\eta E\mathbb{E}\|\nabla \mathcal{L}(\bar{\theta}_{r,E})\|^2 + \frac{C_{\mathcal{L}}\eta L_{\text{eff}}^2}{N}\sum_{e,i}\mathbb{E}\|\theta_{i,r+1,e} - \bar{\theta}_{r,E}\|^2 + C_{\mathcal{L}}\eta E(\sigma^2 + \hat{\Gamma}), \quad (60)$$

where $c_{\Phi,1}, C_{\mathcal{L}} > 0$. Decoupling the summation term via the triangle inequality gives

$$\mathbb{E}\|\theta_{i,r+1,e} - \bar{\theta}_{r,E}\|^2 \leq 2\mathbb{E}\|\theta_{i,r+1,0} - \bar{\theta}_{r,E}\|^2 + 2\mathbb{E}\|\theta_{i,r+1,e} - \theta_{i,r+1,0}\|^2. \quad (61)$$

The first term evaluates to $2\mathbb{E}E_{\text{cons},r,J}$ since $\theta_{i,r+1,0} \equiv \psi_{i,r,J}$. For the second, unfolding local updates and applying Jensen's inequality yields

$$\mathbb{E}\|\theta_{i,r+1,e} - \theta_{i,r+1,0}\|^2 \leq \eta^2e \sum_{k=0}^{e-1} \mathbb{E}\|g_{i,r+1,k}\|^2. \quad (62)$$

Using Lemma 1, the average local drift is bounded as

$$\frac{2}{EN} \sum_{e,i} \mathbb{E}\|\theta_{i,r+1,e} - \theta_{i,r+1,0}\|^2 \leq 2a_2\eta^2E^2(\hat{\Gamma} + \sigma^2). \quad (63)$$

Combining these components strictly bounds the expected disagreement in (60) by $2\mathbb{E}E_{\text{cons},r,J} + 2a_2\eta^2E^2(\hat{\Gamma} + \sigma^2)$. Substituting this bound and the consensus recursion from Lemma 1 into the total drift definition yields

$$\begin{aligned} \mathbb{E}[\Delta\Phi_r] &= \mathbb{E}[\mathcal{L}(\bar{\theta}_{r+1,E}) - \mathcal{L}(\bar{\theta}_{r,E})] + \alpha\mathbb{E}[E_{\text{cons},r+1,J} - E_{\text{cons},r,J}] \\ &\leq -c_{\Phi,1}\eta E\mathbb{E}\|\nabla \mathcal{L}(\bar{\theta}_{r,E})\|^2 + 2C_{\mathcal{L}}\eta EL_{\text{eff}}^2\mathbb{E}E_{\text{cons},r,J} \\ &\quad + C_{\mathcal{L}}\eta E(\sigma^2 + \hat{\Gamma}) + 2C_{\mathcal{L}}a_2\eta^3E^3L_{\text{eff}}^2(\hat{\Gamma} + \sigma^2) \\ &\quad + \alpha(\rho^{2J}(1 + a_1\eta^2E^2L_{\text{eff}}^2) - 1)\mathbb{E}E_{\text{cons},r,J} \\ &\quad + \alpha\rho^{2J}a_2\eta^2E^2(\hat{\Gamma} + \sigma^2) + \alpha c_{\text{OCB}}(1 - \rho^{2J})\Omega_*^2. \end{aligned} \quad (64)$$

Defining $c_{\Phi,2} \triangleq C_{\mathcal{L}} + 2C_{\mathcal{L}}a_2\eta^2E^2L_{\text{eff}}^2 + \alpha\rho^{2J}a_2\eta E$ consolidates noise and heterogeneity. To ensure monotonicity, the coefficient of $\mathbb{E}E_{\text{cons},r,J}$ must be strictly negative, which requires

$$\alpha(1 - \rho^{2J}) - \rho^{2J}a_1\eta^2E^2L_{\text{eff}}^2 - 2C_{\mathcal{L}}\eta EL_{\text{eff}}^2 \geq c_{\Phi,3}(1 - \rho^{2J}). \quad (65)$$

Defining $c_{\Phi,3} \triangleq \alpha(1 - \frac{\rho^{2J}a_1\eta^2E^2L_{\text{eff}}^2}{1 - \rho^{2J}}) - \frac{2C_{\mathcal{L}}\eta EL_{\text{eff}}^2}{1 - \rho^{2J}}$ and choosing a sufficiently large α to ensure $c_{\Phi,3} \in (0, 1)$, the per-round drift is effectively bounded as

$$\mathbb{E}[\Delta\Phi_r] \leq -c_{\Phi,1}\eta E\mathbb{E}\|\nabla \mathcal{L}(\bar{\theta}_{r,E})\|^2 - c_{\Phi,3}(1 - \rho^{2J})\mathbb{E}E_{\text{cons},r,J} + c_{\Phi,2}\eta E(\hat{\Gamma} + \sigma^2) + \alpha c_{\text{OCB}}(1 - \rho^{2J})\Omega_*^2. \quad (66)$$

This completes the proof.

C. Proof of Theorem 1

Building on Lemma 2, we establish linear convergence to a steady-state error floor. To explicitly characterize the trade-off, we decouple the fundamental gradient noise from the consensus-induced noise, yielding the refined expected per-round drift formulated as

$$\begin{aligned} \Delta\Phi_r &\leq -c_{\Phi,1}\eta E\mathbb{E}\|\nabla \mathcal{L}(\bar{\theta}_{r,E})\|^2 - c_{\Phi,3}(1 - \rho^{2J})\mathbb{E}E_{\text{cons},r,J} \\ &\quad + \eta^2E(c_{\text{base}} + \alpha\rho^{2J}c_{\text{cons}})(\hat{\Gamma} + \sigma^2) + \alpha c_{\text{OCB}}(1 - \rho^{2J})\Omega_*^2, \end{aligned} \quad (67)$$

where c_{base} characterizes the intrinsic SGD noise, and c_{cons} captures the residual disagreement noise proportional to the spectral contraction ρ^{2J} .

Applying the PL condition from Assumption 2 to the mean parameter $\bar{\theta}_{r,E}$ yields

$$\begin{aligned} \Delta\Phi_r \leq & -2c_{\Phi,1}\mu_{\text{eff}}\eta E\mathbb{E}[\mathcal{L}(\bar{\theta}_{r,E}) - \mathcal{L}^*] - c_{\Phi,3}(1-\rho^{2J})\mathbb{E}E_{\text{cons},r,J} \\ & + \eta^2 E(c_{\text{base}} + \alpha\rho^{2J}c_{\text{cons}})(\hat{\Gamma} + \sigma^2) + \alpha c_{\text{OCB}}(1-\rho^{2J})\Omega_*^2. \end{aligned} \quad (68)$$

We define the linear contraction factor for the composite Lyapunov function as

$$\kappa_{\text{lin}} \triangleq c_{\Phi,1}\mu_{\text{eff}}\eta E. \quad (69)$$

By enforcing the admissible condition in (35), we ensure that the consensus error contracts at least as fast as the global objective, satisfying $\kappa_{\text{lin}}\alpha \leq c_{\Phi,3}(1-\rho^{2J})$. Furthermore, since the expected loss gap is strictly non-negative, we can safely bound $-2\kappa_{\text{lin}} \leq -\kappa_{\text{lin}}$. Consequently, the negative descent terms in (68) can be merged and bounded by the composite function as

$$-2\kappa_{\text{lin}}\mathbb{E}[\mathcal{L}(\bar{\theta}_{r,E}) - \mathcal{L}^*] - \kappa_{\text{lin}}\alpha\mathbb{E}E_{\text{cons},r,J} \leq -\kappa_{\text{lin}}\mathbb{E}[\Phi_{r,J} - \mathcal{L}^*]. \quad (70)$$

Substituting (70) back into (68) and rearranging the terms using $\Delta\Phi_r = \mathbb{E}[\Phi_{r+1,J}] - \mathbb{E}[\Phi_{r,J}]$, we establish the linear recurrence relation formulated as

$$\begin{aligned} \mathbb{E}[\Phi_{r+1,J} - \mathcal{L}^*] & \leq (1 - \kappa_{\text{lin}})\mathbb{E}[\Phi_{r,J} - \mathcal{L}^*] \\ & + \eta^2 E(c_{\text{base}} + \alpha\rho^{2J}c_{\text{cons}})(\hat{\Gamma} + \sigma^2) + \alpha c_{\text{OCB}}(1-\rho^{2J})\Omega_*^2. \end{aligned} \quad (71)$$

Let the sum of the last two additive terms be denoted as Total Noise. Iterating this recursion from round 0 to r yields the standard linear convergence bound given by

$$\mathbb{E}[\Phi_{r,J} - \mathcal{L}^*] \leq (1 - \kappa_{\text{lin}})^r (\Phi_{0,J} - \mathcal{L}^*) + \frac{\text{Total Noise}}{\kappa_{\text{lin}}}. \quad (72)$$

To explicitly characterize the trade-off induced by the gossip depth J , we decouple the steady-state error floor (i.e., the second term in (72)) into its fundamental statistical components. By expanding κ_{lin} and algebraically factoring out $(1-\rho^{2J})$ in the denominator of the consensus variance term to highlight the spectral gap dependence, the total steady-state error is strictly bounded by three distinct components expressed as

$$\begin{aligned} \text{Error Floor} = & \frac{\eta c_{\text{base}}}{c_{\Phi,1}\mu_{\text{eff}}}(\hat{\Gamma} + \sigma^2) + \frac{\eta^2 c_{\text{var}}}{c_{\Phi,1}\mu_{\text{eff}}(1-\rho^{2J})}(\hat{\Gamma} + \sigma^2) \\ & + \frac{\alpha c_{\text{OCB}}(1-\rho^{2J})}{c_{\Phi,1}\mu_{\text{eff}}}\Omega_*^2. \end{aligned} \quad (73)$$

By defining the bounding constants as $C_{\text{base}} = \frac{\eta c_{\text{base}}}{c_{\Phi,1}\mu_{\text{eff}}}(\hat{\Gamma} + \sigma^2)$, $C_{\Gamma} = \frac{\eta^2 c_{\text{var}}}{c_{\Phi,1}\mu_{\text{eff}}}(\hat{\Gamma} + \sigma^2)$, and $C_{\Omega} = \frac{\alpha c_{\text{OCB}}}{c_{\Phi,1}\mu_{\text{eff}}}$, we recover the exact theoretical upper bound presented in Theorem 1. This structural form mathematically isolates the dual impact of J , establishing the foundation for optimizing the communication depth. This completes the proof.

D. Proof of Theorem 2

From Theorem 1, the transient term vanishes as $r \rightarrow \infty$. Ignoring the J -independent noise C_{base} , we minimize the

steady-state objective formulated as

$$\text{Bound}(J) \triangleq \frac{A}{1-\rho^{2J}} + \alpha B(1-\rho^{2J}), \quad (74)$$

where $A \triangleq C_{\Gamma}$ and $B \triangleq C_{\Omega}\Omega_*^2$. To find the stationary point, we relax J and substitute $y \triangleq 1 - \rho^{2J} \in (0, 1)$, yielding $f(y) \triangleq \frac{A}{y} + \alpha B y$. Differentiating $f(y)$ yields

$$f'(y) = \alpha B - \frac{A}{y^2}, \quad f''(y) = \frac{2A}{y^3}. \quad (75)$$

Because $A > 0$ and $y > 0$, the second derivative $f''(y) > 0$ confirms strict convexity. Setting $f'(y) = 0$ identifies the unique stationary point $y^* = \sqrt{A/(\alpha B)}$, which represents a feasible interior solution ($y^* < 1$) if $\alpha B > A$. Mapping y^* back to J^* using $1 - \rho^{2J^*} = y^*$ results in $\rho^{2J^*} = 1 - \sqrt{A/(\alpha B)}$. Solving for J^* yields the closed-form expression given by

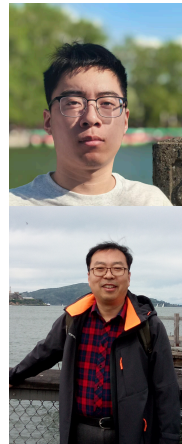
$$J^* = \frac{1}{-2\ln\rho} \ln\left(\frac{1}{1 - \sqrt{\frac{A}{\alpha B}}}\right). \quad (76)$$

Strict convexity and the monotonic relationship between y and J ensure that J^* is the unique global minimizer, mathematically confirming the U-shaped performance trend. This completes the proof.

References

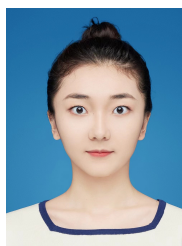
- [1] G. Zhang, Q. Hu, Z. Qin, Y. Cai, G. Yu, and X. Tao, "A unified multi-task semantic communication system for multimodal data," *IEEE Trans. Commun.*, vol. 72, no. 7, pp. 4101–4116, 2024.
- [2] E. Boursoulatz, D. Burth Kurka, and D. Gündüz, "Deep joint source-channel coding for wireless image transmission," *IEEE Trans. Cogn. Commun. Netw.*, vol. 5, no. 3, pp. 567–579, 2019.
- [3] X. Yu, W. Li, L. Yin, and T. Lv, "Adaptive dual-path framework for covert semantic communication," *IEEE Trans. Commun.*, vol. 74, pp. 8484–8499, 2026.
- [4] Y. Gao, J. Ma, M. Zhao, W. Liu, and A. L. Yuille, "NDDR-CNN: Layerwise feature fusing in multi-task CNNs by neural discriminative dimensionality reduction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019.
- [5] S. Liu, E. Johns, and A. J. Davison, "End-to-end multi-task learning with attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 1871–1880.
- [6] Z. Zhai, S. Hu, W. Ni, X. Yuan, X. Wang, and J. Wu, "Spectral-convergent decentralized machine learning: Theory and application in space networks," *IEEE Trans. Mobile Comput.*, vol. 25, no. 7, pp. 10788–10799, 2026.
- [7] S. Wang, J. Dai, Z. Liang, K. Niu, Z. Si, C. Dong, X. Qin, and P. Zhang, "Wireless deep video semantic transmission," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 1, pp. 214–229, 2023.
- [8] X. Yu, T. Lv, W. Li et al., "Multi-task semantic communication with graph attention-based feature correlation extraction," *IEEE Trans. Mobile Comput.*, vol. 24, no. 5, pp. 4371–4388, 2025.
- [9] J. Hu, F. Wang, W. Xu et al., "Scalable multi-task semantic communication system with feature importance ranking," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2023, pp. 1–5.
- [10] Z. Tian, H. Vo, C. Zhang, G. Min, and S. Yu, "An asynchronous multi-task semantic communication method," *IEEE Netw.*, vol. 38, no. 4, pp. 275–283, 2024.
- [11] K. Li, Y. Liang, P. Liò, W. Ni, F. Dressler, J. Crowcroft, and Ö. B. Akan, "User isolation poisoning on decentralized federated learning: An adversarial message-passing graph neural network approach," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 37, no. 6, pp. 2619–2633, 2026.

- [12] Z. He, Y. Li, and Z. Cai, "Personalized federated learning via gradient-fusion and gradient-decoupling for heterogeneous data," *IEEE Trans. Mobile Comput.*, vol. 25, no. 3, pp. 2956–2972, 2026.
- [13] A. Zhang, P. Zhao, W. Lu, and G. Zhang, "Personalized decentralized federated learning: A privacy-enhanced and Byzantine-resilient approach," *IEEE Trans. Comput. Social Syst.*, vol. 12, no. 5, pp. 3206–3217, 2025.
- [14] H. Feng, R. Wang, E. Liu et al., "Joint topology and beamforming optimization for decentralized federated learning," *IEEE Trans. Wireless Commun.*, vol. 25, pp. 12945–12961, 2026.
- [15] S. Nandhini, S. Poruran, and S. Devarajan, "Network security and privacy protection in cyberattacks with asynchronous reinforcement federated learning with task offloading: Decentralized real-time iteration approach," *J. Sensors*, vol. 2025, no. 1, p. 3369587, 2025.
- [16] A. R. Zamir, A. Sax, W. Shen, L. J. Guibas, J. Malik, and S. Savarese, "Taskonomy: Disentangling task transfer learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 3712–3722.
- [17] T. Standley, A. Zamir, D. Chen, L. Guibas, J. Malik, and S. Savarese, "Which tasks should be learned together in multi-task learning?" in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2020, pp. 9120–9132.
- [18] T. Yu, S. Kumar, A. Gupta, S. Levine, K. Hausman, and C. Finn, "Gradient surgery for multi-task learning," *Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 5824–5836, 2020.
- [19] Y. Huang, L. Chu, Z. Zhou, L. Wang, J. Liu, J. Pei, and Y. Zhang, "Personalized cross-silo federated learning on non-iid data," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 9, 2021, pp. 7865–7873.
- [20] A. Asheralieva and D. Niyato, "Effective UAV-aided asynchronous decentralized federated learning with distributed, adaptive and energy-aware gradient sparsification," *IEEE Internet Things J.*, vol. 12, no. 14, pp. 27461–27480, 2025.
- [21] L. You, Z. Guo, C. Yuen, C. Y.-C. Chen, Y. Zhang, and H. V. Poor, "A framework reforming personalized Internet of Things by federated meta-learning," *Nat. Commun.*, vol. 16, no. 1, 2025.
- [22] B. Xiao, F. Zhu, J. Zhang, W. Ni, and X. Wang, "Divergence-based adaptive aggregation for Byzantine robust federated learning," *ArXiv*, vol. abs/2601.06903, 2026.
- [23] Y. Lu, S. Huang, Y. Yang, S. Sirejiding, Y. Ding, and H. Lu, "FedHCA2: Towards hetero-client federated multi-task learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 5599–5609.
- [24] W. Li, T. Lv, W. Ni, J. Zhao, E. Hossain, and H. V. Poor, "Decentralized federated learning over imperfect communication channels," *IEEE Trans. Commun.*, vol. 72, no. 11, pp. 6973–6991, 2024.
- [25] —, "Route-and-aggregate decentralized federated learning under communication errors," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 36, no. 9, pp. 16675–16691, 2025.
- [26] Z. Bao, H. Liang, X. Liu et al., "sDAC—semantic digital analog converter for semantic communications," *IEEE Trans. Commun.*, vol. 73, no. 11, pp. 11061–11077, 2025.
- [27] S. Zou, Y. Zou, M. Liwang, H. Gao, W. Ni, and A. Jamalipour, "Securing semantic communications through full-lifecycle data provenance: A quantum-enhanced labeling architecture," *IEEE Netw.*, pp. 1–10, 2026, early Access.
- [28] L. X. Nguyen, S. Salman Hassan, M. Park et al., "Semspacefl: A collaborative hierarchical federated learning framework for semantic communication in 6g leo satellites," *IEEE Trans. Commun.*, vol. 74, pp. 1269–1286, 2026.
- [29] X. Zhao, Q. Cui, Z. Du et al., "Enhancing convergence, privacy and fairness for wireless personalized federated learning: Quantization-assisted min-max fair scheduling," *IEEE Trans. Mobile Comput.*, vol. 24, no. 10, pp. 9902–9918, 2025.
- [30] Y.-Q. Chen, T. Zhang, X.-L. Jiang, Q. Chen, C.-L. Gao, and W.-L. Huang, "FedBone: Towards large-scale federated multi-task learning," *J. Comput. Sci. Technol.*, vol. 39, no. 5, pp. 1040–1057, 2024.
- [31] S. Gandham, M. Dawande, and R. Prakash, "Link scheduling in wireless sensor networks: Distributed edge-coloring revisited," *J. Parallel Distrib. Comput.*, vol. 68, no. 8, pp. 1122–1134, 2008.
- [32] X. Sun, R. Panda, R. Feris, and K. Saenko, "AdaShare: Learning what to share for efficient deep multi-task learning," *ArXiv*, vol. abs/1911.12423, 2020.
- [33] Z. Wang, Z. Dai, B. Póczos, and J. Carbonell, "Characterizing and avoiding negative transfer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 11293–11302.
- [34] H. Karimi, J. Nutini, and M. Schmidt, "Linear convergence of gradient and proximal-gradient methods under the Polyak–Łojasiewicz condition," in *Proc. Mach. Learn. Knowl. Discov. Databases (ECML PKDD)*, 2016, pp. 795–811.
- [35] C. Liu, L. Zhu, and M. Belkin, "Loss landscapes and optimization in over-parameterized non-linear systems and neural networks," *Appl. Comput. Harmon. Anal.*, vol. 59, pp. 85–116, 2022.
- [36] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," *Proc. Mach. Learn. Syst.*, vol. 2, pp. 429–450, 2020.
- [37] J. Wang, R. Das, G. Joshi, S. Kale, Z. Xu, and T. Zhang, "On the unreasonable effectiveness of federated averaging with heterogeneous data," *Trans. Mach. Learn. Res.*, 2024.
- [38] S. Savazzi, M. Nicoli, and V. Rampa, "Federated learning with cooperating devices: A consensus approach for massive IoT networks," *IEEE Internet Things J.*, vol. 7, no. 5, pp. 4641–4654, 2020.
- [39] A. Hashemi, A. Acharya, R. Das, H. Vikalo, S. Sanghavi, and I. Dhillon, "On the benefits of multiple gossip steps in communication-constrained decentralized federated learning," *IEEE Trans. Parallel Distrib. Syst.*, vol. 33, no. 11, pp. 2727–2739, 2022.

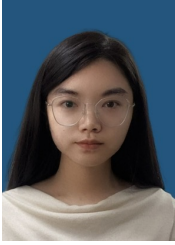


Lin Yin received the B.E. degree in applied physics from Beijing University of Posts and Telecommunications (BUPT), China, in 2023. He is pursuing his Ph.D. with the School of Information and Communication Engineering at BUPT. His research interests include personalized federated learning and semantic communication.

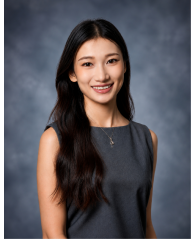
Tiejun Lv received the M.S. and Ph.D. degrees in electronic engineering from the University of Electronic Science and Technology of China (UESTC), Chengdu, China, in 1997 and 2000, respectively. From January 2001 to January 2003, he was a Post-Doctoral Fellow at Tsinghua University, Beijing, China. In 2005, he was promoted to a Full Professor at the School of Information and Communication Engineering, Beijing University of Posts and Telecommunications (BUPT). From September 2008 to March 2009, he was a Visiting Professor with the Department of Electrical Engineering, Stanford University, Stanford, CA, USA. He is currently the author of four books, one book chapter, and more than 170 published journal articles and 230 conference papers on the physical layer of wireless mobile communications. His current research interests include signal processing, communications theory, and networking. He was a recipient of the Program for New Century Excellent Talents in University Award from the Ministry of Education, China, in 2006. He received the Nature Science Award from the Ministry of Education of China for the hierarchical cooperative communication theory and technologies in 2015 and Shaanxi Higher Education Institutions Outstanding Scientific Research Achievement Award in 2025.



Weicai Li received the B.E. and Ph.D. degrees from the School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, China, in 2020 and 2025, respectively. From December 2022 to December 2023, she was a Visiting Student at the University of Technology Sydney, Australia. She is currently with the School of Information Communication Engineering, Beijing Information Science and Technology University, Beijing, China. Her research interests include integrated sensing and communications, wireless federated learning, distributed computing, and privacy preservation.



Xi Yu (Graduate Student Member, IEEE) received the B.E. degree in communication engineering from Beijing University of Posts and Telecommunications (BUPT), China, in 2020. She is pursuing her Ph.D. with the School of Information and Communication Engineering at BUPT. Her research interests include multi-task semantic communication and privacy-preserving techniques.



Xiaoyu He is currently pursuing the Ph.D. degree in Information and Communication Engineering at Beijing University of Posts and Telecommunications (BUPT), Beijing, China. From December 2025 to June 2026, she was a visiting Ph.D. student at the Singapore University of Technology and Design (SUTD), Singapore, under the supervision of Prof. Tony Q. S. Quek. Since August 2026, she has been a joint-training visiting Ph.D. student at the University of Bologna, Bologna, Italy, under the supervision of Prof. Paolo Bellavista. Her research interests include federated learning, machine unlearning, decentralized federated learning (D-FL), personalized learning, adaptive model pruning, reinforcement learning, non-cooperative game theory, multi-hop routing optimization, and resource-constrained wireless networks.