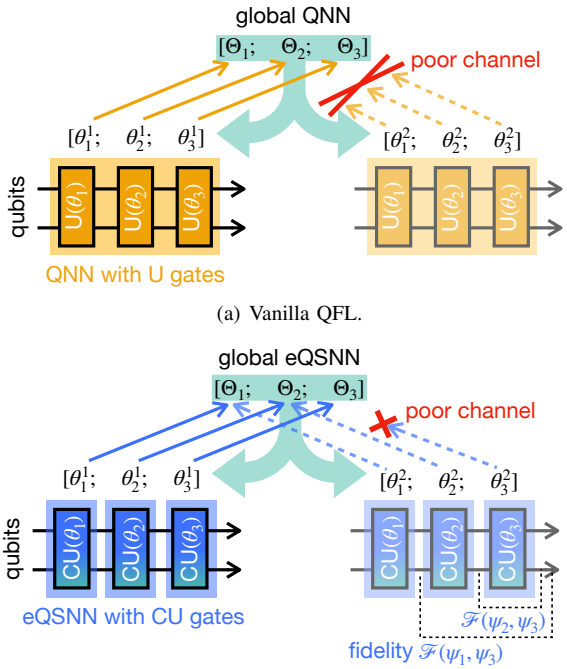
This article has been accepted for publication in IEEE Internet of Things Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2025.3540103

Entanglement-Controlled Quantum Federated Learning

Soohyun Park, *Member, IEEE*, Hyunsoo Lee, Soyi Jung, *Member, IEEE*, Jihong Park, *Senior Member, IEEE*, Mehdi Bennis, *Fellow, IEEE*, and Joongheon Kim, *Senior Member, IEEE*

***Abstract*—According to the advances in quantum computing and distributed learning, quantum federated learning (QFL) has recently become an emerging field of study. In QFL, each quantum computer or device locally trains its quantum neural network (QNN) with trainable gates, and communicates only these gate parameters over classical channels, without costly quantum communications. To successfully opeate QFL under various and dynamic channel conditions in Internet of Things (IoT) environments, this paper develops a novel depth-controllable architecture of entangled slimmable quantum neural networks (eSQNNs), and thus, proposes an entangled slimmable QFL (eSQFL) that communicates the superposition-coded parameters of eSQNNs. Even though the proposed eSQNN-based eSQFL is superior, training the depth-controllable eSQNN architecture is challenging due to high entanglement entropy and inter-depth interference. Therefore, the proposed method in this paper mitigates the interference using entanglement controlled universal (CU) gates and an inplace fidelity distillation (IPFD) regularizer penalizing inter-depth quantum state differences, respectively. Furthermore, the proposed method optimizes the superposition coding power allocation by deriving and minimizing the convergence bound of eSQFL. The novelty of this work is evaluated via extensive simulations in terms of prediction accuracy, fidelity, and entropy compared to Vanilla QFL as well as under different channel conditions and various data distributions.**

***Index Terms*—Quantum Machine Learning, Quantum Federated Learning, Superposition Coding, Slimmable Neural Network.**



(a) Vanilla QFL.

(b) eSQFL.

Fig. 1. A schematic illustration of (a) Vanilla quantum federated learning (QFL) and (b) the proposed *entangled slimmable quantum FL (eSQFL)* with 2 devices each of which has the *entangled slimmable quantum neural network (eSQNN)* having 3 depth layers.

I. Introduction

Recent advances in quantum computing hardware and algorithms have recently lead to the emergence of quantum machine learning [2], [3], [4], [5], [6]. As opposed to classical computation at a linear scale in bits, quantum computing can perform calculations at an exponential scale in qubits [7]. The main enablers are the stochastic nature and the entanglement phenomenon of qubits, allowing one to make each qubit represent superimposed multiple states and to simultaneously control multiple qubits, respectively [8]. Consequently, even in the current era of noisy intermediate scale quantum (NISQ) [9], *i.e.*, with 50 to a few hundred qubits, quantum machine learning has achieved linear or sublinear complexity in various applications, as compared with the polynomial complexity of classical machine learning [10].

As analogous to neural networks, the parameterized quantum circuit (PQC), also known as a quantum neural network (QNN), has become a de facto standard quantum machine learning architecture [10], [11]. In a PQC, qubits flow through the gates associated with trainable classical parameters, during which the states of the qubits can be adjusted. For various applications ranging from image classification [12] to reinforcement learning [13], with a much smaller number of trainable parameters, PQC training has achieved the prediction accuracy on par with

The parts of this research were presented at IEEE Conference on Computer Communications (INFOCOM), London, United Kingdom, May 2022 [1].

This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2024-RS-2024-00436887) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation); and also by IITP grant funded by MSIT (RS-2024-00439803, SW Star Lab) for Quantum AI Empowered Second-Life Platform Technology. *(Corresponding authors: Soyi Jung, Joongheon Kim)*

S. Park is with the Division of Computer Science, Sookmyung Women's University, Seoul 04310, Korea (e-mail: soohyun.park@sookmyung.ac.kr).

H. Lee and J. Kim are with the Department of Electrical and Computer Engineering, Korea University, Seoul 02841, Republic of Korea (e-mails: {hyunsoo,joongheon}@korea.ac.kr).

S. Jung is with the Department of Electrical and Computer Engineering, Ajou University, Suwon 16499, Republic of Korea (e-mail: sjung@ajou.ac.kr).

J. Park is with the Singapore University of Technology and Design (SUTD), Singapore (e-mail: jihong_park@sutd.edu.sg).

M. Bennis is with the Centre for Wireless Communications, University of Oulu, Oulu 90014, Finland (e-mail: mehdi.bennis@oulu.fi).

the neural network.

Focusing on the parameter efficiency of PQCs, by integrating federated learning (FL) [14], [15], [16], [17], [18] into standalone quantum machine learning, quantum FL (QFL) has recently attracted attention [19], [20]. Without communicating qubits via costly quantum communications, QFL enables distributed quantum machine learning in Internet of Things (IoT) environments at scale by communicating the PQC's trainable parameters via classical communications, even over wireless channels [21].

A. Algorithm Design Concept

Motivated by this trend in QFL, the overarching goal of this article is to develop a communication-efficient QFL framework that can cope with heterogeneous and time-varying channel conditions and computing resources. To this end, we first revisit slimmable FL (SFL) in classical machine learning [1], wherein each device has a width-controllable local model, known as a *slimmable neural network (SNN)* [22], [23], and communicates its superposition-coded local model with different width levels, enabling multi-level local information exchanges depending on channel conditions. Inspired from this, as visualized in Fig. 1, we propose an *entangled slimmable quantum FL (eSQFL)* framework with *entangled slimmable QNNs (eSQNNs)*, which is a non-trivial extension from SFL with SNNs to their quantum versions as summarized later.

Unlike multi-width SNNs, the eSQNN is a multi-depth PQC wherein more depth levels incur higher *von Neumann entanglement entropy* on average. Unfortunately, the PQC trainability is often challenged by the problem of vanishing all gradients, known as the barren plateaus [24], which is exacerbated under higher entanglement entropy [25]. Meanwhile, too low entanglement may negate the benefit of quantum machine learning. To resolve this issue for an unknown target degree of entanglement, our proposed eSQNN entangles different qubits using the *controlled universal (CU) quantum gates* [26] such that the degree of entanglement is trainable.

Next, simultaneous local training of the multiple eSQNN depths may induce inter-depth interference, hindering convergence. In classical machine learning, SFL avoids its similar inter-width interference issue by adding the inplace knowledge distillation (IPKD) regularizer that penalizes the output difference from any smaller width to the largest width level [22]. Since the IPKD uses the Kullback-Leibler (KL) divergence, it becomes less accurate (or even diverging) for the larger differences. Alternatively, leveraging the *Uhlmann's fidelity* function in quantum information theory [27], we propose a novel *inplace fidelity distillation (IPFD)* regularizer that is bounded within 0 and 1 while accurately measuring the quantum state difference even between the smallest and the largest levels. This IPFD regularizer plays a big role in handling multi-depth by regularizing parameters between different layers of QNN.

Finally, for communication efficiency, the eSQNN parameters in multiple depths are superposition-coded and transmitted with a different transmit power allocation to each depth. Like SFL, the transmit power is optimized by deriving and minimizing the convergence bound of the eSQFL. Nevertheless, the convergence analysis is completely different since the gradient in PQC training is measured in a quantum computing way using the *parameter shift rule* [28].

Not only by analysis but also by extensive simulations, we corroborate that the proposed eSQFL with eSQNNs achieves convergence while each different width level can be trained to be a separate model with reasonable accuracy and fidelity, under various channel conditions as well as independent and identically distributed (IID) or non-IID data distributions. Note that unlike eSQFL, Vanilla quantum FL having fixed local PQC architectures cannot cope with different channel conditions [19], [29]. A recent work [30] also considers a slimmable architecture in the context of QFL. However, it does not theoretically guarantee convergence, and its specific architecture (*i.e.*, angle/pole parameters) only allows two-level superposition coding, as opposed to generalized multi-level architectures using CU gates in eSQFL.

B. Contributions

The major contributions of the work in this paper can be summarized as follows.

- First of all, a multi-depth QNN architecture with CU gates, *i.e.*, eSQNN, is proposed to enable superposition-coded transmissions while avoiding barren plateaus. We measure von Neumann entropy between inter-quantum states of different depths. Indeed, CU gates increase the trainability when designing multi-depth QNN.
- In addition, a local eSQNN training algorithm with a fidelity-inspired regularizer, *i.e.*, IPFD, is proposed in order to mitigate inter-depth interference. In eSQNN training, the proposed IPFD in this paper shows the crucial role.
- Moreover, with eSQNNs and IPFD, a novel quantum FL framework, *i.e.*, eSQFL, is proposed, and its convergence bound is theoretically derived.
- Lastly, based on the derived convergence bound, transmit power allocation in superposition coding is optimized. In addition, we corroborate that the derived convergence bound helps eSQFL achieve high accuracy.

C. Organization

The rest of this paper is organized as follows. Sec. II presents the related work to the proposed quantum federated learning. Sec. III introduces the eSQNN architecture and its local training with IPFD regularizer. Sec. IV describes superposition coding, successive decoding, and the proposed eSQFL framework. Sec. V provides the convergence analysis on eSQFL and its insight. Sec. VI presents the numerical experimental results to corroborate eSQFL empirically. Lastly, Sec. VII concludes this paper. Notice that the notations in this paper are in Tab. III.

II. Preliminaries

A. Quantum Machine Learning Basics

Basic Quantum Gates. A qubit is a quantum computing unit where the quantum state is represented with two bases $|0\rangle$, and $|1\rangle$ in Bloch sphere [31]. Consider the q qubits system,

in which the quantum state defined in Hilbert space $\boldsymbol{\psi} \in \mathbb{C}^{2^q}$ can be expressed as follows,

$$|\boldsymbol{\psi}\rangle = \Lambda_1|0\cdots0\rangle + \cdots + \Lambda_{2^q}|1\cdots1\rangle \tag{1}$$

where $\sum_{i=1}^{2^q} \Lambda_i^2 = 1$. A classical data $\boldsymbol{x}$ is encoded as a quantum state with the rotation gates $R_{\mathrm{x}}(\boldsymbol{x})$, $R_{\mathrm{y}}(\boldsymbol{x})$, and $R_{\mathrm{z}}(\boldsymbol{x})$, where the rotation of $(\boldsymbol{x})$ occurs in the direction of x-, y-, and z-axes in Bloch sphere, respectively. Moreover, qubits are entangled with *controlled-NOT* gates (CNOT) [32]. CNOT gates act on two qubits to entangle them by using the first qubit as the control qubit and performing *XOR* operation on the second qubit. These basic quantum gates configure the QNNs.

Quantum Neural Network. The structure of a QNN is tripartite: the state encoder, PQC, and the measurement layer [33], [34]. In the forward propagation, classical input data $\boldsymbol{x}$ needs to be first encoded with the state encoder via basic rotation gates, which is a unitary operation and denoted as $U(\boldsymbol{x})$. Then, the encoded quantum state is processed through the PQC $U(\boldsymbol{\theta})$, a multi-layered set of CNOT gates and rotation gates associated with trainable parameters $\boldsymbol{\theta}$. The quantum state $\boldsymbol{\psi}$ can be expressed as,

$$|\boldsymbol{\psi}_{\boldsymbol{\theta}}\rangle = U(\boldsymbol{\theta}) \cdot |\boldsymbol{\psi}_0\rangle = U(\boldsymbol{\theta}) \cdot U(\boldsymbol{x})|0\rangle. \tag{2}$$

The output of the PQC is the entangled quantum state that can be measured after applying a projection matrix $M \in \mathcal{M} \equiv \{M_1, \cdots, M_c, \cdots, M_C\}$ onto the reference z-axis. The measured output $\langle V\rangle_{\boldsymbol{\theta}} \in [-1, 1]^C$ is called an *observable*, where C denotes the output dimension. The operation of QNN corresponding to c-th observable is as follows,

$$\langle V_c\rangle_{\boldsymbol{\theta}} = \langle 0|U^\dagger(\boldsymbol{x})U^\dagger(\boldsymbol{\theta})M_cU(\boldsymbol{\theta})U(\boldsymbol{x})|0\rangle = \langle\boldsymbol{\psi}|M_c|\boldsymbol{\psi}\rangle \tag{3}$$

where $(\cdot)^\dagger$ denotes the complex conjugate operator. Using the observable, a given loss function is calculated. Unlike classical NNs having visible activations in their hidden layers, the quantum states within QNNs are not measurable; otherwise, the quantum states collapse [31]. This does not allows quantum machine learning to compute the loss gradients via the chain rule, i.e., backpropagations. Alternatively, quantum machine learning evaluates the gradients using the zero-th order method called the parameter shift rule [28] (see Appendix A).

B. Related Work

Classical Federated Learning. Federated learning (FL) is a distributed learning architecture made up of a server, local devices, and a global model [16]. The conventional FL workflow is as follows [35], [36], [37]. The server transmits the global model to all the local devices via one-hop direct communications, and each device produces local parameters by training the received global model. Then, these parameters are sent back to the central server via one-hop direct communications, where all the data is aggregated to update the global model. Finally, the updated global model is transmitted to the local devices again for another iteration. Due to this mechanism, FL allows a large number of devices to learn a global model simultaneously without transmitting any data, ensuring data privacy as well. Considering the recent increase in the number and computational power of edge devices, FL is an extremely useful tool for reducing computational overhead and protecting data security which is both emerging challenges in the field of machine learning [38]. Within this architecture, various techniques with differing methods of aggregating data and training the global model exist, *e.g.*, FedAvg [39], FedBN [40]. The convergence analysis of FL algorithms is especially challenging because of the data heterogeneity in FL, which forces researchers to rely on copious numbers of assumptions. Consequently, gaps in the understanding of FL analysis occur. Over the years, many major works have attempted to better understand FL by removing assumptions and exploring new techniques [41], [42], [39], [43], [44]. Even now, research on FL convergence in various aspects is still being carried out (*i.e.*, non-convex, convergence bounds). For example, [45] has successfully proposed an analysis of local stochastic gradient descent (SGD) using only arbitrarily heterogeneous data while also using weaker assumptions than previous works. On the other hand, convergence analysis of most QFL algorithms has not been fully developed yet. This paper aims to further discuss the convergence analysis of QFL via the analysis of eSQFL with the characteristic of quantum computing. The convergence analysis on a dynamic QFL is elaborated in Sec. IV.

Classical Slimmable Federated Learning. SlimFL is a framework that executes FL by using *slimmable neural networks (SNNs)* with *SC* and *SD* [1]. The architectural properties of SNN reduce memory costs of SlimFL [22] while with rigorous communication and computational efficiencies. SC is a process of compressing two different data signals into one signal. As the signals are encoded, different power levels are assigned to each data signal which is used to decide the priority of signals during SD. Additionally, the SNN is composed of the left-hand (LH) and the right-hand (RH) sides. The LH side is occupied by the high priority signal, while the lower priority signal goes to the RH side. After SC is finished, the encoded message will be uploaded to the server, which then undergoes SD. Assuming that the state of the communication channel is good, the LH of the SNN will be decoded first, followed by the RH signal. However, if the communication channel is not stable enough, only LH will be decoded, resulting in a small model. Finally, if the communication channel is completely unstable, no signal will be obtained. This flexible characteristic of SNN allows SlimFL to be adaptable to dynamic communication environments, making it suitable for practical applications.

The proposed algorithm in this paper is superior to SlimFL [1] as follows. First of all, the proposed algorithm in this paper fundamentally based on quantum neural network (QNN) utilization in order to realize quantum supremacy [46], [47]. For this objective, the concept of SNN is used for the re-design of QNN for depth-controllable operations. Furthermore, the IPFD method is re-designed under the consideration of the unique characteristics of QNN, which outperforms classical IPKD. Lastly, this paper provides comprehensive convergence analysis for the QNN-based eSQFL using quantum gradient descent, which is different from the convergence analysis of SlimFL [1].

Quantum Federated Learning. In this section, the concept

This article has been accepted for publication in IEEE Internet of Things Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2025.3540103

of QFL is elaborated in depth. QFL is implementing FL via quantum computation by replacing all the NNs with QNNs. Chen *et al.*, [19] is the first to propose a hybrid quantum-classical QFL architecture where the local devices are replaced with quantum devices, unlike FL models. After receiving global model parameters, the quantum computers carry out quantum machine learning using QNNs. Then, the output of each device is aggregated to update the global model before repeating the process. Nouhaila *et al.*, [48] propose QFL as a framework for training quantum machine learning models across distributed networks, addressing data privacy concerns inherent in traditional machine learning. The proposed federated quantum neural network (FedQNN) framework combines quantum machine learning with classical federated learning principles, enabling secure and collaborative learning without direct data sharing. Song *et al.*, [49] introduce a QFL framework tailored for classical clients (CC), termed CC-QFL, which enables collaborative training of a quantum machine learning model without requiring clients to possess quantum computing capabilities. By utilizing the shadow tomography technique, the server constructs a classical representation of the quantum machine learning model, allowing clients to compute local gradients using their data. Qiao *et al.*, [50] provide a comprehensive survey on QFL, highlighting its potential to revolutionize machine learning by combining the strengths of classical FL with the speed and parallelism of quantum computing. It addresses the gap in current research by exploring how quantum machine learning can bridge traditional FL and QFL, and reviews key topics such as heterogeneity, privacy, and security in FL. The survey also examines how quantum computing can enhance FL's computational efficiency and discusses future directions for optimizing QFL in distributed machine learning. In Chehimi *et al.*, [29], a purely QFL framework is proposed. Similar to [19], this model is composed of a server and multiple quantum devices. However, instead of converting classic data into quantum states, the local devices generate quantum data by labeling qubits as excited or not excited according to the degree of rotation on the Bloch sphere. Both [19], [29] use FedAvg to aggregate data and execute training. As seen from the two examples above, a QFL and FL share an identical system structure, but QFL leverages quantum machine learning instead of machine learning in order to exploit the advantages of quantum computing. For this work, Vanilla QFL is referring to a purely quantum version of [19]. In addition, quantum application of SFL is studied to improve the communication opportunities [30]. SQFL utilized trainable measurement parameters to configure two messages which contain both the trainable measurement parameters and PQC parameters, respectively. However, in this work, multiple layer architectures and local training algorithms are proposed, which are not present in [30].

III. Architecture and Training of eSQNNs

In this section, we describe the architecture of eSQNN and its local training algorithm. To elaborate on this, by slightly modifying the depth-fixed architecture of Vanilla QNN [10], we first prepare its depth-controllable counterpart without controlling the level of entanglement, dubbed Vanilla SQNN, followed by introducing the proposed eSQNN controlling both the depth and the level of entanglement.

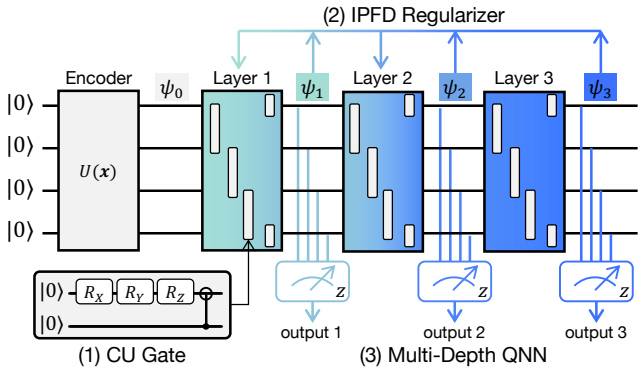


Fig. 2. Illustration of eSQNN: (1) we use *CU gates* in eSQNN, (2) the *fidelity regularizer* is applied sub-model layers (*e.g.*, *Layer 1 and 2*) by receiving quantum states, (3) our proposed eSQNN is based on the *multi-depth* QNN.

Architecture of Vanilla SQNN. Suppose that Vanilla SQNN consists of L layers, and produces the desired outputs at any layer $l \in [1, L]$. In this paper, the number of sub-models must be larger than 1 (*i.e.*, $L \geq 2$). When the l-th sub-model is used, it means that the l-th model will be configured from the encoding layer to the l-th layer. For an arbitrary $k \in \mathbb{N}[1, K]$ and $l \in \mathbb{N}[1, L]$, the model parameters of k-th local device and the l-th layer is denoted as $\boldsymbol{\theta}^k \odot \sum_{l'=1}^{l} \Xi_{l'}$. Note that $\Xi_{l'}$ is a binary mask which eliminates all trainable parameters except parameters of the l'-th layer. The operation $\odot$ denotes an element-wise product. However, it is difficult to make desirable results at any random layer because the vanilla SQNN is vulnerable to the barren plateau problem [24], [51]. The barren plateau is a bad local optimum which hinders convergence. It is known that more entanglement's degree introduces worse the barren plateau problem [25]. The operations in Vanilla SQNN are as follows: 1) rotate quantum state $|\psi\rangle$ with rotation gates, 2) entangle qubits, and 3) repeat the first and second steps. We predict that the operations mentioned above will increase the degree of entanglement.

Architecture of eSQNN. eSQNN is proposed to cope with the problem of Vanilla SQNN architecture. Fig. 2 shows the illustration of eSQNN. eSQNN is mainly composed of CU gates. The operations of the CU gate in two qubits are written as $\begin{bmatrix} I & 0 \\ 0 & U \end{bmatrix}$, where U is expressed as $U = \begin{bmatrix} u_{00} & u_{01} \\ u_{10} & u_{11} \end{bmatrix}$. Note that U is an unitary matrix, *i.e.*, $U^\dagger U = I$. We focus on the architectural advantage of CU gates because CU gates can adjust the direction of entanglement, disentanglement, or rotation while training. We describe the advantages of eSQNN and the barren plateau phenomenon next.

To this end, at first we consider the *von Neumann entanglement entropy*, a metric for measuring the degree of quantum entanglement of bipartite subsystems in an entire system [52]. For instance, consider two subsystems, *e.g.*, l and l'-th model configuration, where $l > l'$. According to a two-copy test from [53], we can compare the different quantum states $|\psi_l\rangle$

and $|\psi_{l'}\rangle$ by using additional qubits. Then, we can measure the entanglement entropy by following the statement below. Suppose a quantum state that exists in l and l'-th depth is represented as $\boldsymbol{\psi}_{l',l} \in \mathbb{C}^{2^{2q}}$. Its pure state is obtained by $\rho_{l',l} \triangleq |\boldsymbol{\psi}_{l',l}\rangle\langle\boldsymbol{\psi}_{l',l}|$. Finally, the entanglement entropy is calculated as follows,

$$S_l(\rho_{l',l}) = -\operatorname{Tr}_l(\rho_{l',l}\log\rho_{l',l}) \tag{4}$$

where $\operatorname{Tr}_l(\cdot)$ stands for partial trace over the l-th layer. As discussed in many studies, avoiding the barren plateaus requires the reduction of the entanglement entropy [25].

In addition, amplitude encoding is used as a method to encode the classical state in our proposed algorithms and systems. This approach represents the classical state as the probability amplitude of a quantum state, allowing the information to be expressed through quantum superposition. If $\sum_{l'=0}^{l-1} S(\rho_{l',l}) \geq S_{l,th}$, the barren plateau becomes severe and training of l-th model fails. For this, we observe the entanglement entropy between the encoding state and the layer of eSQNN. In order to ensure all model configurations are trained, we define a metric as,

$$\mathbb{1}_{\text{train}} = \prod_{l=1}^{L} \mathbb{1}\left(\sum_{l'=0}^{l-1} S(\rho_{l',l}) < S_{l,th}\right) \tag{5}$$

where $\mathbb{1}(\cdot)$ stands for an indicator function. In order to verify whether the metric works correctly, we provide the following two cases. If all model configurations are satisfied $\sum_{l'=0}^{l-1} S(\rho_{l',l}) < S_{l,th})$, then $\mathbb{1}_{\text{train}} = 1$, which means it avoids the barren plateau. On the other hand, suppose that $\exists l$ that satisfies $\sum_{l'=0}^{l-1} S(\rho_{l',l}) \geq S_{l,th}$, we have $\mathbb{1}_{\text{train}} = 0$ which it means it suffers from barren plateau. We conjecture that eSQNN is more robust to the barren plateau than Vanilla SQNN because the event $\mathbb{1}_{\text{train}} = 1$ frequently occurs in eSQNN. The eSQNN addresses this issue by using CU gates, which allow entanglement level control and adjustment. This design has potential to help for maintaining lower entanglement entropy across the layers, thus avoiding the conditions that typically lead to barren plateaus. More details are in Sec. VI-B.

eSQNN Local Training. This section presents the eSQNN local training algorithm. In general, classic SNNs use the IPKD regularizer $\mathcal{L}_{KD}$ to transfer knowledge from a large model to a small model [22], which can be expressed as,

$$\mathcal{L}_{KD} = D_{KL}(p(\boldsymbol{y}_{t,e}^{k,L}) \| p(\boldsymbol{y}_{t,e}^{k,l})) \tag{6}$$

where D_{KL} is the KL divergence. IPKD is ill-suited when the difference between the outputs of two models becomes large, where the KL divergence may even diverge. Alternatively, we propose the IPFD regularizer $\mathcal{L}_{FD}$, inspired by the Uhlmann's fidelity function [54] in quantum information theory, measuring the similarity between two quantum states [54]. Precisely, the fidelity of the quantum states in the L-th and l-th model configurations is defined as follows,

$$\mathcal{F}(\boldsymbol{\psi}_L, \boldsymbol{\psi}_l) = |\langle\boldsymbol{\psi}_L|\boldsymbol{\psi}_l\rangle|^2. \tag{7}$$

In (7), if $\mathcal{F}(\boldsymbol{\psi}_L, \boldsymbol{\psi}_l) \approx 1$, $\boldsymbol{\psi}_l$ is similar to $\boldsymbol{\psi}_L$, which means the logits of l-th model are almost same as the logits of L-th model. On the other hand, the opposite condition $\mathcal{F}(\boldsymbol{\psi}_L, \boldsymbol{\psi}_l) \approx 0$ means the l-th model does not follow the L-th model.

Algorithm 1: Local-eSQNN Training

```
Initialization. local-QNN parameters, θ;
for e = {1, 2, ..., E} do
    for (x, y) ∈ D do
        Get logits with L-th model;
        Calculate loss with labels and accumulate loss;
        for l = {1, 2, ..., L − 1} do
            Get logits with l-th model;
            Calculate loss gradient with parameter-shift rule;
        θ^k_{t,e+1} ← θ^k_{t,e} − η_t ∇_{θ^k_{t,e}} L^{k,l}_{t,e};
```

Consequently, in a classification task, the local training of an eSQNN with the IPFD regularizer is described as follows. The parameters $(\boldsymbol{x}, \boldsymbol{y})$ are denoted as data and label, respectively. The predicted label $\boldsymbol{y} = \{y_c\}_{c=1}^{C}$ is an one-hot encoded vector wherein the element y_c becomes unity for a true label and otherwise 0, *i.e.*, $y_{c'} = 0, \forall c' \neq c$. Hereafter, we describe the local training for the parameters of local device k in the t-th communication round and e-th local training iteration. The logits of class and its prediction of l-th model are denoted as,

$$y_{t,e}^{k,l,c} = \exp(a\langle V_c\rangle_{\boldsymbol{\theta}_{t,e}^k \odot \sum_{l'=1}^{l} \Xi_{l'}}), \tag{8}$$

$$p(y_{t,e}^{k,l,c}|\boldsymbol{x}) = \frac{y_{t,e}^{k,l,c}}{\sum_{c=1}^{C} y_{t,e}^{k,l,c}}, \tag{9}$$

where a represents the observable hyperparameter. Additionally, the cross-entropy loss and the fidelity regularization are as,

$$\mathcal{L}_{CE} = -\sum_{c=1}^{C}[y_c \log(p(y_{t,e}^{k,l,c})|\boldsymbol{x})], \tag{10}$$

$$\mathcal{L}_{FD} = 1 - \mathcal{F}(\boldsymbol{\psi}_{t,e,\boldsymbol{x}}^{k,L}, \boldsymbol{\psi}_{t,e,\boldsymbol{x}}^{k,l}). \tag{11}$$

The loss function is given as,

$$\mathcal{L}_{t,e}^{k,l} = \frac{1}{D}\sum_{(\boldsymbol{x},\boldsymbol{y})\in\zeta^k}[\lambda\mathcal{L}_{CE} + (1-\lambda)\mathcal{L}_{FD}] \tag{12}$$

where D and λ stand for the batch size and the balanced parameter of fidelity regularization, respectively. The gradient of (12) can be calculated with parameter shift rule [28]. Algorithm 1 summarizes the local training process before one communication round. After training with Algorithm 1, the gradient to be transmitted to the server can be as,

$$g_t^k = \sum_{e=1}^{E}\sum_{l=1}^{L} \nabla_{\boldsymbol{\theta}_{t,e}^k} \mathcal{L}_{t,e}^{k,l} \tag{13}$$

where η_t denotes the learning rate at communication round t.

IV. Entangled Slimmable Quantum Federated Learning

A. Superposition Coding & Successive Decoding

The successful reception of a wireless signal is mainly affected by the signal-to-interference-plus-noise ratio (SINR) [55]. At a receiver, SINR can be expressed as,

$$\gamma = \chi d^{-\beta} P/(\sigma^2 + P^I) \tag{14}$$

where P, P^I, d, and σ^2 denote the transmission interference, reception interference, a transmitter-receiver distance, and noise powers, respectively. In addition, $\beta \geq 2$ is a path loss exponent and χ is small-scale fading parameter (*i.e.*, Rayleigh fading). Following the Shannon's capacity formula with a Gaussian codebook, the received throughput R with the bandwidth W is $R = W\log_2(1+\gamma)$ (bits/sec). When the transmitter encodes raw data with a code rate u, its receiver successfully decodes the encoded data if $R > u$. Finally, the decoding success probability can be given as follows,

$$\Pr(R \geq u) = \Pr\left(\frac{\chi d^{-\beta} P}{\sigma^2 + P^I} \geq u'\right) \tag{15}$$

where $u' = 2^{\frac{u}{W}} - 1$. Consider transmitting L messages from a transmitter to a receiver simultaneously. Before transmission, these messages are SC-encoded [56], and the whole transmission power budget P is allotted to the l-th message, with $P_l = \nu_l P$ transmission power for $l \in [1, L]$. Note that ν_l is an allocation variable such that $\nu_l > u' \sum_{l'=l+1}^{L} \nu_{l'}$, $\sum_{l=1}^{L} \nu_l = 1$, and $\forall \nu_l \geq 0$.

The SC-encoded message is meant to be sequentially decoded at the receiver by first decoding the strongest signal, then canceling out the decoded signal, and finally decoding the next strongest signal, *i.e.*, SD, also known as successive interference cancellation [57], [58]. The small-scale fading parameter χ under Rayleigh fading follows an exponential distribution, *i.e.*, $\chi \sim \exp(1)$. Assuming $l' > l$, the receiver may gradually decode the l-th message while experiencing the remaining messages as interference P_l^I, *i.e.*,

$$P_l^I = \chi d^{-\beta} P \sum_{l'=l+1}^{L} \nu_{l'}, \tag{16}$$

for $l \leq L-1$. However, $P_L^I = 0$ as there is no interference for the last message. Assume that R_l represents the throughput of the l-th message. Then, the distribution of R_l is given as,

$$\Pr(R_l \geq u) = \Pr\Big(\chi \geq \frac{1/\bar{\gamma}}{\nu_l/u' - \sum_{l'=1+1}^{L} \nu_l'}\Big) \tag{17}$$

where $\bar{\gamma} = \frac{P d^{-\beta}}{\sigma^2}$ denotes the averaged signal-to-noise ratio (SNR). By using this result, the l-th message's decoding success probability p_l can be expressed as follows,

$$p_l = \Pr(R_1 \geq u, \cdots, R_l \geq u), \tag{18}$$

$$= \Pr\Big(\chi \geq \max\big(\frac{1/\bar{\gamma}}{\frac{\nu_l}{u'} - \sum_{l'=2}^{L} \nu_{l'}}, \cdots, \frac{1/\bar{\gamma}}{\frac{\nu_l}{u'} - \sum_{l'=l+1} \nu_{l'}}\big)\Big). \tag{19}$$

Algorithm 2: eSQFL

1 **Notation.** $\boldsymbol{\theta}_t^k$: k-th device's parameters, Θ_t: parameters of global eSQNN, X_l: set of l-th subdivided gradient;
2 **Initialization.** $X_l \leftarrow \emptyset, \forall l \in [1, L]$;
3 **for** $k = \{1, \ldots, K\}$ **do**
4 Sample $\chi^k \sim \exp(1)$;
5 **for** $l = \{1, 2, \ldots, L\}$ **do**
6 **if** $\chi_k \geq u_l$ **then**
7 $X_l \leftarrow X_l \cup k$;
8 **if** $\prod_{l=1}^{L} \mathbb{1}(X_l = \emptyset) \neq 0$ **then**
9 $\Theta_{t+1} \leftarrow \Theta_t - \eta_t \sum_{l=1}^{L} \frac{1}{|X_l|} \sum_{k \in X_l} g_t^k \odot \Xi_l$;
10 **else**
11 Skip aggregation;
12 **for** $k = \{1, \cdots, K\}$ **do**
13 $\boldsymbol{\theta}_{t+1,1}^k \leftarrow \Theta_{t+1}$;

B. eSQFL Operations

This section describes the operations of eSQFL. Algorithm 2 shows the eSQFL algorithm. First of all, local devices are trained with Algorithm 1. The power allocation is conducted to configure SC-encoded model parameters, *i.e.*, $\boldsymbol{\nu} = \{\nu_l\}_{l=1}^L$ for the gradient $\{g_t^k \odot \Xi_l\}_{l=1}^L$ of the subdivided model configuration. After that, the local devices transmit their SC-encoded model parameters to the server. The server decodes the devices' SC-encoded model parameters with SD. If the server receives at least one local gradient for every model configuration, the server aggregates; otherwise, no aggregation occurs. In the aggregation of sub-divided model configuration, FedAvg is utilized [39]. The updates of eSQFL will be explained later.

V. Convergence Analysis

A. Setup

In order to analyze the convergence rate of eSQFL, the following assumptions are considered. Firstly, the local-side decoding is always successful (Algorithm 2, lines 12–13) because the server-side transmission power is higher than the uplink power. Secondly, K is assumed to be big enough such that $|X_l| \approx K p_l$, for all l. During the t-th communication round, the server builds the global model which can be expressed as follows,

$$\Theta_{t+1} \leftarrow \Theta_t - \eta_t \underbrace{\sum_{l=1}^{L} \frac{1}{K p_l} \sum_{k \in X_l} g_t^k \odot \Xi_l}_{:= f_t}. \tag{20}$$

The objective function of the global model and the local objective functions are denoted as F and $\{F^k\}$ respectively. The bar notation $\bar{\cdot}$ is used for the averaged value over $\{\zeta_t^k\}$, and the superscript * is used to indicate the optimum. For mathematical amenability, we consider the following assumptions on F and $\{F^k\}$, as used in [59].

Assumption 1 (**β-Smoothness**). *If F and $\{F^k\}$ are β-smooth,*

$$F^k(\boldsymbol{\theta}_v) \leq F^k(\boldsymbol{\theta}_w) + (\boldsymbol{\theta}_v - \boldsymbol{\theta}_w)^T \nabla F^k(\boldsymbol{\theta}_w) + \frac{\beta}{2}\|\boldsymbol{\theta}_v - \boldsymbol{\theta}_w\|^2, \quad (21)$$

for all $v, w > 0$.

Assumption 2 (**μ-Strong Convexity**). *If F and $\{F^k\}$ are μ-strong convex,*

$$F^k(\boldsymbol{\theta}_v) \geq F^k(\boldsymbol{\theta}_w) + (\boldsymbol{\theta}_v - \boldsymbol{\theta}_w)^T \nabla F^k(\boldsymbol{\theta}_w) + \frac{\mu}{2}\|\boldsymbol{\theta}_v - \boldsymbol{\theta}_w\|^2, \quad (22)$$

for all $v, w > 0$.

Assumption 3 (**Bounded Local Gradient Variance**). *For all device $k \in \mathbb{N}[1, K]$ and its local data $\zeta^k \in \mathbf{Z}$, the difference between the local gradient $F^k(\boldsymbol{\theta}^k; \zeta^k)$ and $\bar{F}^k(\boldsymbol{\theta}; \mathbf{Z})$ is bounded, i.e.,*

$$\mathbb{E}[\|\nabla_{\boldsymbol{\theta}} F^k(\boldsymbol{\theta}^k, \zeta_t^k) - \nabla_{\boldsymbol{\theta}} \bar{F}^k(\boldsymbol{\theta}^k; \mathbf{Z})\|^2] \leq \sigma_k^2. \quad (23)$$

According to [45], the metric for the non-IIDness of $\mathbf{Z}$ is given as follows,

$$\delta = \frac{1}{K}\sum_{k=1}^{K} \sigma_k^2. \quad (24)$$

B. Convergence Analysis

In classical machine learning, the convergence of FedAvg has been analyzed by assuming bounded local gradients in [59]. Without such an unrealistic assumption, the convergence bound of SFL has been derived in [1]. In quantum machine learning, local gradients can be shown to be inherently bounded thanks to the bounded fidelity and the parameter shift rule computing quantum gradients [28]. Hence, rather than adopting the methods in [1], we first derive the local gradient bound, and then derive the convergence bound of eSQFL by following the steps [59]. The detailed proofs are deferred to Appendix, and only the results are presented as elaborated next.

Lemma 1 (**Bounded Local Gradient**). *For $t \geq 1$ and $\eta_t \leq \eta_{t+1}$, it follows that*

$$\mathbb{E}[\|g_t^k\|^2] \leq EL(2 + (a-2)\lambda)^2. \quad (25)$$

Lemma 2 (**Bounded Global Gradient**). *For $t \geq 1$, the global gradient has bound as,*

$$\mathbb{E}[\|f_t\|^2] \leq EL^2(2 + (a-2)\lambda)^2 \sum_{l=1}^{L} \frac{1}{p_l^2}. \quad (26)$$

Lemma 3 (**Bounded Global Gradient Variance**). *Under Assumption 3, the variance of the global gradient f_t is bounded within $\mathbf{Z}$, which is given as,*

$$\mathbb{E}\|f_t - \bar{f}_t\|^2 \leq L\delta \sum_{l=1}^{L} \frac{1}{p_l^2}. \quad (27)$$

Note that Lemmas 2 and 3 are different, in the sense that Lemma 2 focuses on the actual gradient, whereas Lemma 3 is related to data distributions. The convergence analysis utilizes Lemmas 1–3 and eSQFL convergence can be proven by [59].

Theorem 1 (**eSQFL Convergence**). *Under Assumptions 1 and 3 with the learning rate $\eta_t = \frac{2}{\mu t + 2\beta - \mu}$, we obtain*

$$\mathbb{E}[F(\theta_t)] - F^* \leq \frac{\beta}{\mu} \cdot \frac{\mu\beta\Delta_1 + 2B}{\mu t + 2\beta - \mu}, \quad (28)$$

where $$\Delta_t \triangleq \mathbb{E}\|\Theta_t - \Theta^*\|^2, \quad (29)$$

$$B = (EL^2(2 + (a-2)\lambda)^2 + L\delta)\sum_{l=1}^{L} \frac{1}{p_l^2}. \quad (30)$$

Hence, $\lim_{t\to\infty} \mathbb{E}[F(\theta_t)] = F^*$.

Theorem 1 exhibits several insights of eSQFL as follows.

1) *Failure under extremely poor channels:* Consider an extremely poor channel condition, where the server cannot receive $[l, L]$-th model configurations, *i.e.*, $p_{l'} \simeq 0, \forall l' \in [l, L]$. In this case, the RHS of (28) diverges.
2) *Importance of successful reception:* The optimal gap of eSQFL becomes smaller by increasing the communication opportunities. Consider a perfect channel condition, where the RHS of (28) is minimized. By optimizing the SC transmission, the optimality gap is reduced which is referred to Proposition 1 and Corollary 1.
3) *Other important metrics:* The optimality gap is affected by the local iterations per communication round E, balance factor λ, and the number of layers L.

Proposition 1 (**Optimal SC Power Allocation**). *The transmission power allocation $\boldsymbol{\nu}^*$ minimizing the optimality gap is given as,*

$$\boldsymbol{\nu}^* = \arg\min_{\boldsymbol{\nu}} \left(\sum_{l=1}^{L} \exp\left(-\frac{2/\bar{\gamma}}{\nu_l/u' - \sum_{l'=l+1}^{L} \nu_{l'}} \right) \right) \quad (31)$$

where $L \geq 2$, $\nu_l > u' \sum_{l'=l+1}^{L} \nu_{l'}$ for $\forall l \in [1, L)$, and $\sum_{l=1}^{L} \nu_l = 1$.

Proof. Substituting the term p_l into Theorem 1, the optimality gap is minimized by optimizing the power allocation. □

Corollary 1 (**Low SNR, $L = 2$**). *For $L = 2$, $\bar{\gamma} \to 0$, and $u' \geq (1 + \sqrt{5})/2 \approx 1.618$, the optimal power allocation is as,*

$$(\nu_1^*, \nu_2^*) = \left(-\frac{\sqrt{u'+1} - u'^2 + 1}{u'^2 + u'}, 1 + \frac{\sqrt{u'+1} - u'^2 + 1}{u'^2 + u'} \right). \quad (32)$$

Proof. Since $\exp(-x) = 1 - x$ for $x \to 0$, the RHS of (31) becomes $2 + \frac{2/\bar{\gamma}}{\nu_1/u' - (1-\nu_1)} + \frac{2/\bar{\gamma}}{(1-\nu_1)/u'}$, which is piece-wise convex. Applying the first-order necessary condition (FONC) with respect to ν_1 completes the proof. □

VI. Experiments

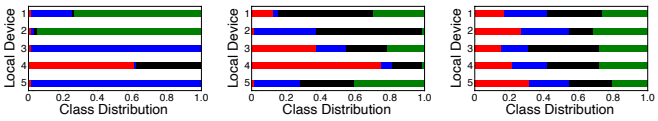
A. Experimental Design

To corroborate the main analysis and hypothesis of this paper, the experiments are designed as follows:

This article has been accepted for publication in IEEE Internet of Things Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2025.3540103

TABLE I
LIST OF SIMULATION PARAMETERS.

Description	Value
Number of devices (N)	10
Local iterations per communication round (E)	10
Epoch (T)	100
Optimizer	SGD
Learning rate (η_1),	0.01
Decaying rate	0.001
Observable hyperparameter (a)	2
Number of qubits	4
Number of parameters in eSQFL & Vanilla QFL	36
Number of data per device	128
Batch size (D)	32

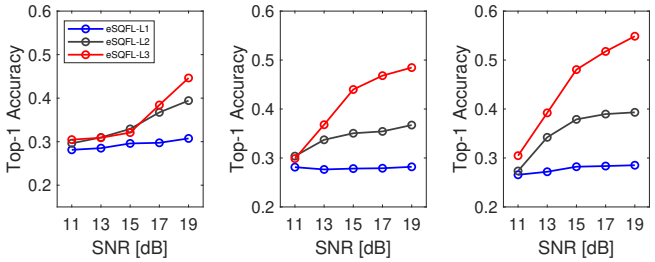


(a) $\alpha = 0.1$. (b) $\alpha = 1.0$. (c) $\alpha = 10$.

Fig. 3. Class distributions with different Dirichlet concentration α.

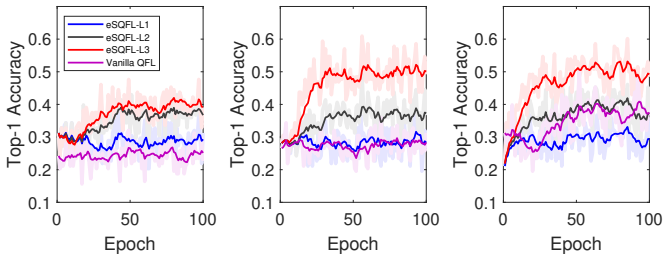
- From Sec. V-A, the derived convergence bound is highly affected by the decoding success probability and non-IIDness. To corroborate these results numerically, we compare the top-1 accuracy of eSQFL in various channel conditions and degrees of non-IIDness with Vanilla QFL (referred to Fig. 1(a)).
- We investigate the advantage of CU gates that compose eSQNN by designing an experiment which measures entanglement entropy and top-1 accuracy of eSQNN and standard QNNs under the same conditions. Then, the two metrics are compared to demonstrate the advantage of CU gates.
- The increased effectiveness of local training with IPFD compared to IPKD is proven. IPFD trains the local models by regularizing the fidelity of two quantum states. In contrast, IPKD trains local models by ensuring that the small model follows the large model via its prediction. The benchmark scheme comparing the fidelity and top-1 accuracy of IPFD and IPKD is designed.
- According to Proposition 1 and Corollary 1, the convergence bound is minimized by optimal transmission power allocation. To corroborate this, we compare the optimal power allocation scheme to its random power allocation counterpart.
- Finally, we conduct experiments by controlling various variables and assess their various impact on the performance.

For the experiment, eSQFL and Vanilla QFL are implemented via the TorchQuantum [60] library. eSQFL is the proposed model which leverages eSQNN. This specific QNN consists of three sub-models named 'L1', 'L2', and 'L3'. In contrast, Vanilla QFL uses a standard QNN which is made up of basic quantum gates [10], and does not consider SC and SD [19]. Despite the difference in structure, both eSQNN and standard QNN use equivalent number of parameters. Moreover, we conduct ablation studies on our eSQNN by comparing it with Vanilla SQNN, a depth-controllable yet entanglement-fixed



(a) $\alpha = 0.1$. (b) $\alpha = 1$. (c) $\alpha = 10$.

Fig. 4. Comparison of top-1 accuracy under various avg. SNR [dB] and α.



(a) $\alpha = 0.1$. (b) $\alpha = 1$. (c) $\alpha = 10$.

Fig. 5. Comparison of top-1 accuracy under various α ($\bar{\gamma} = 17\,\text{dB}$).

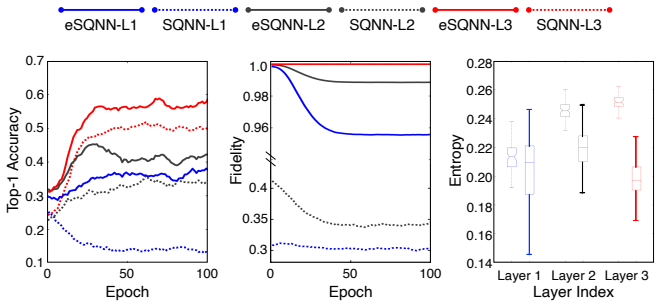
QNN. Since the performance of QFL suffers under a system with a large number of qubits, many QFL works use a simple dataset [19]. In this paper, the MNIST dataset is transformed into a simpler form: the dimension of MNIST data is reduced to 4×4 by inter-area interpolation, and only four classes are used (*i.e.*, 0, 1, 2 and 3) [61]. The four classes are represented with red, blue, black, and green respectively. In addition, Dirichlet distribution is used to investigate non-IIDness of data [62]. Fig. 3 shows the data distribution with the different values of the Dirichlet concentration ratio α. Data with high Dirichlet concentration ratio (*i.e.,* $\alpha = 10$) is IID while data with low Dirichlet concentration ratio (*i.e.,* $\alpha = 0.1$) is non-IID.

To compare IPFD and IPKD, we initialize the parameter of eSQNN identically. The simulation parameters used in these numerical experiments are summarized in Table I. Four qubits are used for each local device during the encoding process. The number of qubits used in the local model should match the number used in the global model.

B. Numerical Results

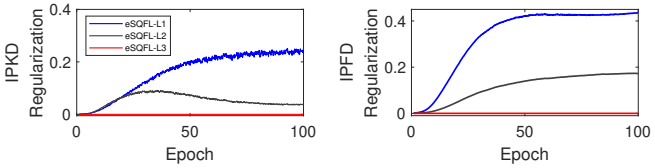
Numerical Results and Convergence Analysis. According to Theorem 1, the convergence bound decreases if the decoding success probability increases. Fig. 4 shows the performance of eSQFL under various channel conditions obtained through various σ^2. As $\bar{\gamma}$ increases from 11 dB to 19 dB, the decoding success probability and top-1 accuracy of the eSQFL with all layers increase. The small models, *i.e.*, eSQFL-L2 and eSQFL-L1, also show improvement in performance along with eSQFL-L3. Especially, eSQFL-L2 shows significant improvement in top-1 accuracy from 28% to 39%. Fig. 5 shows the top-1 accuracy and convergence of eSQFL and comparison models. When $\bar{\gamma} = 17\,\text{dB}$, the sub-models in eSQFL (*i.e.,* eSQFL-L2,

This article has been accepted for publication in IEEE Internet of Things Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2025.3540103



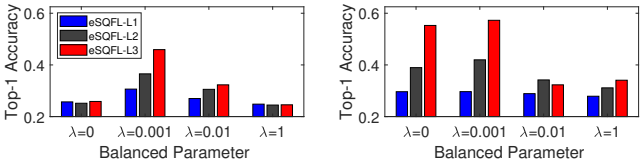
(a) Top-1 accuracy. (b) Fidelity. (c) Entropy.

Fig. 6. Model architectural difference (eSQNN vs. Vanilla SQNN).



(a) IPKD Regularization. (b) IPFD Regularization.

Fig. 7. Comparison of IPFD training algorithm under non-IID and IID.



(a) $\alpha = 0.1$. (b) $\alpha = 10$.

Fig. 8. Comparison of fidelity training algorithm under non-IID and IID.

eSQFL-L3) achieve higher accuracy than Vanilla QFL. The final standard deviations of eSQFL under $\bar{\gamma} = 17$ dB are 0.041, 0.051, and 0.066 for eSQFL-L1, eSQFL-L2, and eSQFL-L3, respectively.

According to Theorem 1, the data distribution affects the convergence bound of eSQFL. With non-IID data, the convergence bound is widened. As shown in Fig. 5, we test various Dirichlet concentration, *i.e.*, $\alpha = \{0.1, 1, 10\}$. The overall performance of all comparison models decreases as α decreases. However, eSQFL shows robustness under non-IID data distribution. Vanilla QFL shows low top-1 accuracy under $\alpha = 1$ and $\alpha = 0.1$. In contrast, eSQFL maintains the top-1 accuracy of 52% and 41% under $\alpha = 1.0$ and $\alpha = 0.1$ respectively. From the results in Fig. 4 and Fig. 5, eSQFL is robust under various channel conditions and non-IID data distribution.

Structural Advantage of eSQNN. In this subsection, we investigate the general performance, fidelity, and entanglement entropy of eSQNN. We conduct ablation studies corresponding to model architecture (*i.e.*, eSQNN and Vanilla SQNN). Fig. 6 shows the experiment results. As shown in Fig. 6(a), eSQNN shows better top-1 accuracy than Vanilla eSQNN. Especially, eSQNN-L3 achieves a 20% performance improvement over Vanilla SQNN. When eSQNN is used, the quantum state (*i.e.*, knowledge) is successfully distilled to the small model as shown in Fig. 6(b). In contrast, Vanilla SQNN fails to distill the knowledge to its sub-model. To understand why its model is successfully trained, we calculate the entanglement entropy (referred to Sec. III). Fig. 6(c) exhibits the von Neumann entanglement entropy of each layer of eSQNN and Vanilla SQNN. The entanglement entropy of eSQNN is less than Vanilla SQNN for all layers. It means that the event of exceeding the entropy threshold aforementioned in Sec. VI, *i.e.*, $\mathbb{1}_{\text{train}} = 0$, rarely occurs compared to Vanilla SQNN. This underscores that eSQNN is more robust to barren plateaus than Vanilla SQNN.

TABLE II
TOP-1 ACCURACY COMPARISON WITH ($\boldsymbol{\nu}^*$) AND WITHOUT OPTIMIZATION.

Condition	$L=3$			$L=2$	
	$l=1$	$l=2$	$l=3$	$l=1$	$l=2$
with optimization ($\boldsymbol{\nu}^*$)	**29.6**	**40.1**	**55.8**	**33.4**	**50.7**
w.o. optimization	29.5	39.3	50.7	33.0	48.1

Effectiveness of IPFD. To investigate the effectiveness of IPFD used in eSQNN local training, we compare the results of local training using IPFD regularizer to IPKD regularizer. Fig. 7 (a)/(b) show the learning curve of $\mathcal{L}_{FD}$ and $\mathcal{L}_{KL}$, respectively. The learning curve of IPFD starts at $\mathcal{L}_{FD} = 0$ due to the fidelity $\mathcal{F}(\boldsymbol{\psi}_l, \boldsymbol{\psi}_L) \approx 1$. As eSQNN is trained, the fidelity decreases and converges to 0.955 for L1 and 0.987 for L2. In the learning curve of $\mathcal{L}_{KD}$, the curve has a tendency to decrease and converge. However, the fluctuation of IPKD regularization is larger than IPFD, especially in eSQFL-L1. This is because the KL divergence becomes unstable when the difference between the two distributions is large, *i.e.*, the overlapping area between the distributions is small. Then, if there is no overlapping area, it diverges. In contrast, the aforementioned phenomena does not occur in IPFD regularization because IPFD regularizer is bounded from 0 to 1. Therefore, IPFD regularization provides more stable noise to its eSQNN than IPKD regularization.

Impact on Optimal Power Allocation. We verify the proofs of Proposition 1, and Corollary 1. When $L = 3$, we calculate the power allocation variable as $\boldsymbol{\nu}^* = \{0.8909, 0.0989, 0.0102\}$ by non-convex optimization. When $L = 2$, we obtain $\boldsymbol{\nu}^* = \{0.8969, 0.1059\}$, where the comparison of power allocation is set to $\boldsymbol{\nu} = \{0.9170, 0.0820, 0.001\}$ for $L = 3$, and $\boldsymbol{\nu} = \{0.8333, 0.1667\}$ for $L = 2$. The final accuracy is Tab. II. Compared to the eSQFL with $\boldsymbol{\nu}$, the eSQFL with $\boldsymbol{\nu}^*$ achieves 10.1% higher top-1 accuracy when $L = 3$ and 5.41% higher top-1 accuracy when $L = 2$. Thus, we corroborate that the optimal power allocation minimizes the convergence bound.

Impact on Balanced Parameter. The balanced parameter λ is an important parameter in eSQNN local training. Fig. 8 shows the top-1 accuracy according to λ in various data distributions (*i.e.*, $\alpha = 0.1$ and $\alpha = 10$). With finitely adjusted IPFD parameter ($\lambda^* = 0.01$), eSQNN shows the highest top-1 accuracy under non-IID data distribution (*i.e.*, $\alpha = 0.1$). In addition, by not using IPFD ($\lambda = 0$) or only using IPFD ($\lambda = 1$), eSQNN fails to classify the mini-MNIST dataset. Under IID data distribution (*i.e.*, $\alpha = 10$), eSQNN with $\lambda^* = 0.01$ outperforms eSQNN with only using label training

($\lambda = 0$) about 1.03%. From the result, we recommend utilizing eSQNN with $\lambda^* = 0.001$ for robust performance in both IID and non-IID data distribution.

VII. Conclusions and Future Work

In this paper, we developed a depth-adjustable QNN architecture, and proposed a novel QFL framework based on wireless communications, termed eQSNN and eSQFL, respectively. To control the level of entanglement and reduce its entropy, we applied CU gates to the eSQNN architecture. To mitigate the inter-depth interference inspired from the fidelity in quantum information theory, we introduced a novel IPFD regularizer. Finally, to cope with various channel conditions, we applied SC across multiple depths and optimized the SC power allocation by deriving and minimizing the convergence bound of eSQFL. In conclusion, we were able to propose a QFL model that shows stable performance despite the NISQ limitation and variable channel conditions. Additionally, the fidelity regularizer was also designed. This novel method decreases the error rate of QNN in a way that is exclusively suitable with QC, instead of depending on classical optimization methods.

Our considering future research directions are as follows. First of all, it is worthy to evaluate the realistic efficacy of the proposed algorithm based on the theoretical analysis in this paper. Furthermore, it is required to find various applications and use cases for our proposed eSQFL.

Acknowledgement

The authors thank Hankyul Baek for initiating this research and his constructive discussions.

References

[1] H. Baek, W. J. Yun, Y. Kwak, S. Jung, M. Ji, M. Bennis, J. Park, and J. Kim, "Joint superposition coding and training for federated learning over multi-width neural networks," in *Proc. IEEE Conference on Computer Communications (INFOCOM)*, May 2022.

[2] W. J. Yun, J. P. Kim, S. Jung, J.-H. Kim, and J. Kim, "Quantum multiagent actor–critic neural networks for Internet-connected multirobot coordination in smart factory management," *IEEE Internet Things J.*, vol. 10, no. 11, pp. 9942–9952, Jun. 2023.

[3] F. Arute, K. Arya, R. Babbush, D. Bacon, J. C. Bardin, R. Barends, R. Biswas, S. Boixo, F. G. Brandao, D. A. Buell *et al.*, "Quantum supremacy using a programmable superconducting processor," *Nature*, vol. 574, no. 7779, pp. 505–510, 2019.

[4] Silvirianti, B. Narottama, and S. Y. Shin, "Layerwise quantum deep reinforcement learning for joint optimization of UAV trajectory and resource allocation," *IEEE Internet Things J.*, vol. 11, no. 1, pp. 430–443, 2024. [Online]. Available: https://doi.org/10.1109/JIOT.2023.3285968

[5] C. Park, W. Yun, J. Kim *et al.*, "Quantum multiagent actor-critic networks for cooperative mobile access in multi-UAV systems," *IEEE Internet Things J.*, vol. 10, no. 22, pp. 20 033–20 048, Nov. 2023.

[6] W. J. Yun, Y. Kwak, J. P. Kim, H. Cho, S. Jung, J. Park, and J. Kim, "Quantum multi-agent reinforcement learning via variational quantum circuit design," in *Proc. IEEE International Conference on Distributed Computing Systems (ICDCS)*, Bologna, Italy, July 2022.

[7] P. W. Shor, "Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer," *SIAM Journal on Computing*, vol. 26, no. 5, pp. 1484–1509, October 1997.

[8] S. Park, J. P. Kim, C. Park, S. Jung, and J. Kim, "Quantum multi-agent reinforcement learning for autonomous mobility cooperation," *IEEE Communications Magazine*, vol. 62, no. 6, pp. 106–112, 2024.

[9] J. Preskill, "Quantum computing in the NISQ era and beyond," *Quantum*, vol. 2, p. 79, August 2018.

[10] S. Y.-C. Chen, C.-H. H. Yang, J. Qi, P.-Y. Chen, X. Ma, and H.-S. Goan, "Variational quantum circuits for deep reinforcement learning," *IEEE Access*, vol. 8, pp. 141 007–141 024, 2020.

[11] "Quantum distributed deep learning architectures: Models, discussions, and applications," *ICT Express*, 2022.

[12] V. Havlíček, A. D. Córcoles, K. Temme, A. W. Harrow, A. Kandala, J. M. Chow, and J. M. Gambetta, "Supervised learning with quantum-enhanced feature spaces," *Nature*, vol. 567, no. 7747, pp. 209–212, 2019.

[13] O. Lockwood and M. Si, "Reinforcement learning with quantum variational circuit," in *Proc. AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, vol. 16, no. 1, 2020, pp. 245–251.

[14] N. P. Patel, R. Parekh, S. A. Amin, R. Gupta, S. Tanwar, N. Kumar, R. Iqbal, and R. Sharma, "LEAF: A federated learning-aware privacy-preserving framework for healthcare ecosystem," *IEEE Trans. Netw. Serv. Manag.*, vol. 21, no. 1, pp. 1129–1141, 2024. [Online]. Available: https://doi.org/10.1109/TNSM.2023.3287393

[15] H. Sedjelmaci and A. Boualouache, "When two-layer federated learning and mean-field game meet 5G and beyond security: Cooperative defense systems for 5G and beyond network slicing," *IEEE Trans. Netw. Serv. Manag.*, vol. 21, no. 1, pp. 1178–1189, 2024. [Online]. Available: https://doi.org/10.1109/TNSM.2023.3294568

[16] J. Park, S. Samarakoon, A. Elgabli, J. Kim, M. Bennis, S.-L. Kim, and M. Debbah, "Communication-efficient and distributed learning over wireless networks: Principles and applications," *Proceedings of the IEEE*, vol. 109, no. 5, pp. 796–819, May 2021.

[17] S. Niknam, H. S. Dhillon, and J. H. Reed, "Federated learning for wireless communications: Motivation, opportunities, and challenges," *IEEE Communications Magazine*, vol. 58, no. 6, pp. 46–51, 2020.

[18] D. Kwon, J. Jeon, S. Park, J. Kim, and S. Cho, "Multiagent DDPG-based deep learning for smart ocean federated learning IoT networks," *IEEE Internet of Things Journal*, vol. 7, no. 10, pp. 9895–9903, 2020.

[19] S. Y.-C. Chen and S. Yoo, "Federated quantum machine learning," *Entropy*, vol. 23, no. 4, p. 460, 2021.

[20] H. Zhou, K. Lv, L. Huang, and X. Ma, "Quantum network: Security assessment and key management," *IEEE/ACM Transactions on Networking*, vol. 30, no. 3, pp. 1328–1339, 2022.

[21] R. Pujahari and A. Tanwar, "Quantum federated learning for wireless communications," in *Federated Learning for IoT Applications*. Springer, 2022, pp. 215–230.

[22] J. Yu and T. S. Huang, "Universally slimmable networks and improved training techniques," in *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, Korea, October 2019, pp. 1803–1811.

[23] D. Kim, J. Kim, J. Kwon, and T.-H. Kim, "Depth-controllable very deep super-resolution network," in *Proc. IEEE International Joint Conference on Neural Networks (IJCNN)*, Budapest, Hungary, July 2019, pp. 1–8.

[24] J. R. McClean, S. Boixo, V. N. Smelyanskiy, R. Babbush, and H. Neven, "Barren plateaus in quantum neural network training landscapes," *Nature Communications*, vol. 9, no. 1, pp. 1–6, 2018.

[25] S. H. Sack, R. A. Medina, A. A. Michailidis, R. Kueng, and M. Serbyn, "Avoiding barren plateaus using classical shadows," *PRX Quantum*, vol. 3, no. 2, June 2022.

[26] T. Sleator and H. Weinfurter, "Realizable universal quantum logic gates," *Physical Review Letters*, vol. 74, no. 20, p. 4087, 1995.

[27] M. M. Wilde, *Quantum information theory*. Cambridge University Press, 2013.

[28] K. Mitarai, M. Negoro, M. Kitagawa, and K. Fujii, "Quantum circuit learning," *Physical Review A*, vol. 98, no. 3, p. 032309, 2018.

[29] M. Chehimi and W. Saad, "Quantum federated learning with quantum data," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, May 2022, pp. 8617–8621.

[30] W. J. Yun, J. P. Kim, S. Jung, J. Park, M. Bennis, and J. Kim, "Slimmable quantum federated learning," in *Proc. of ICML Workshop on Dynamic Neural Networks*, Baltimore, MD, USA, July 2022.

[31] D. Bouwmeester and A. Zeilinger, "The physics of quantum information: basic concepts," in *the Physics of Quantum Information*, 2000, pp. 1–14.

[32] C. P. Williams, S. H. Clearwater *et al.*, *Explorations in quantum computing*. Springer, 1998.

[33] N. Killoran, T. R. Bromley, J. M. Arrazola, M. Schuld, N. Quesada, and S. Lloyd, "Continuous-variable quantum neural networks," *Physical Review Research*, vol. 1, no. 3, p. 033063, 2019.

[34] O. Simeone, "An introduction to quantum machine learning for engineers," *Foundations and Trends® in Signal Processing*, vol. 16, no. 1-2, pp. 1–223, 2022.

[35] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," *IEEE Signal Processing Magazine*, vol. 37, no. 3, pp. 50–60, May 2020.

[36] W. Y. B. Lim, N. C. Luong, D. T. Hoang, Y. Jiao, Y.-C. Liang, Q. Yang, D. Niyato, and C. Miao, "Federated learning in mobile edge networks: A comprehensive survey," *IEEE Communications Surveys & Tutorials*, vol. 22, no. 3, pp. 2031–2063, Thirdquarter 2020.

[37] K. Wei, J. Li, M. Ding, C. Ma, H. H. Yang, F. Farokhi, S. Jin, T. Q. S. Quek, and H. Vincent Poor, "Federated learning with differential privacy: Algorithms and performance analysis," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 3454–3469, April 2020.

[38] N. H. Tran, W. Bao, A. Y. Zomaya, M. N. H. Nguyen, and C. S. Hong, "Federated learning over wireless networks: Optimization model design and analysis," in *Proc. IEEE Conference on Computer Communications (INFOCOM)*, Paris, France, 2019, pp. 1387–1395.

[39] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, Ft. Lauderdale, FL, USA, April 2017, pp. 1273–1282.

[40] X. Li, M. Jiang, X. Zhang, M. Kamp, and Q. Dou, "FedBN: Federated learning on non-iid features via local batch normalization," in *Proc. International Conference on Learning Representations (ICLR)*, 2021.

[41] L. Mangasarian, "Parallel gradient distribution in unconstrained optimization," *SIAM Journal on Control and Optimization*, vol. 33, no. 6, pp. 1916–1925, 1995.

[42] A. Cotter, O. Shamir, N. Srebro, and K. Sridharan, "Better mini-batch algorithms via accelerated gradient methods," *Proc. Advances in Neural Information Processing Systems (NIPS)*, vol. 24, 2011.

[43] N. Karakoç, A. Scaglione, M. Reisslein, and R. Wu, "Federated edge network utility maximization for a multi-server system: Algorithm and convergence," *IEEE/ACM Transactions on Networking*, vol. 30, no. 5, pp. 2002–2017, 2022.

[44] C. T. Dinh, N. H. Tran, M. N. H. Nguyen, C. S. Hong, W. Bao, A. Y. Zomaya, and V. Gramoli, "Federated learning over wireless networks: Convergence analysis and resource allocation," *IEEE/ACM Transactions on Networking*, vol. 29, no. 1, pp. 398–409, February 2021.

[45] A. Khaled, K. Mishchenko, and P. Richtarik, "Tighter theory for local sgd on identical and heterogeneous data," in *Proc. International Conference on Artificial Intelligence and Statistics (AISTATS)*, vol. 108, August 2020, pp. 4519–4529.

[46] S. Park, J. P. Kim, C. Park, S. Jung, and J. Kim, "Quantum multi-agent reinforcement learning for autonomous mobility cooperation," *IEEE Communications Magazine*, vol. 62, no. 6, pp. 106–112, Jun. 2024.

[47] S. Park, J. Chung, C. Park, S. Jung, M. Choi, S. Cho, and J. Kim, "Joint quantum reinforcement learning and stabilized control for spatio-temporal coordination in metaverse," *IEEE Transactions on Mobile Computing*, pp. 1–18, 2024 (Early Access).

[48] I. Nouhaila, M. A.-Z. Khan, A. Marchisio, M. Shafique, and M. Bennai, "FedQNN: Federated learning using quantum neural networks," in *Proc. International Joint Conference on Neural Networks (IJCNN)*, Yokohama, Japan, June-July 2024.

[49] Y. Song, Y. Wu, S. Wu, D. Li, Q. Wen, S. Qin, and F. Gao, "A quantum federated learning framework for classical clients," *Science China Physics, Mechanics & Astronomy*, vol. 67, no. 250311, pp. 1–10, May 2024.

[50] C. Qiao, M. Li, Y. Liu, and Z. Tian, "Transitioning from federated learning to quantum federated learning in internet of things: A comprehensive survey," *IEEE Communications Surveys & Tutorials (Early Access)*, pp. 1–42, May 2024.

[51] X. You and X. Wu, "Exponentially many local minima in quantum neural networks," in *Proc. of the International Conference on Machine Learning (ICML)*, Virtual, July 2021.

[52] D. Greenberger, K. Hentschel, and F. Weinert, *Compendium of quantum physics: concepts, experiments, history and philosophy*. Springer Science & Business Media, 2009.

[53] Y. Subaşı, L. Cincio, and P. J. Coles, "Entanglement spectroscopy with a depth-two quantum circuit," *Journal of Physics A: Mathematical and Theoretical*, vol. 52, no. 4, p. 044001, January 2019.

[54] R. Jozsa, "Fidelity for mixed quantum states," *Journal of Modern Optics*, vol. 41, no. 12, pp. 2315–2323, 1994.

[55] D. N. C. Tse and P. Viswanath, *Fundamentals of Wireless Communications*, 2005.

[56] T. Cover, "Broadcast channels," *IEEE Transactions on Information Theory*, vol. 18, no. 1, pp. 2–14, January 1972.

[57] J. Choi, "Joint rate and power allocation for NOMA with statistical CSI," *IEEE Transactions on Communications*, vol. 65, no. 10, pp. 4519–4528, October 2017.

[58] M. Choi, D. Yoon, and J. Kim, "Blind signal classification for non-orthogonal multiple access in vehicular networks," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 10, pp. 9722–9734, 2019.

TABLE III
LIST OF NOTATIONS.

Notation	Description
K	Number of local devices, $[1, \cdots, k, \cdots, K]$
L	Number of local eSQNN blocks, $[1, \cdots, l, \cdots, L]$
E	Number of local iterations, $[1, \cdots, e, \cdots, E]$
T	Number of communication rounds, $[1, \cdots, t, \cdots, T]$
$\boldsymbol{\Xi}$	Binary mask, $\boldsymbol{\Xi} = \{\Xi_1, \cdots, \Xi_l, \cdots, \Xi_L\}$
$\boldsymbol{\psi}$	Quantum state
ρ	Reduced density matrix
$S_l(\rho)$	Entanglement entropy of ρ over subsystem l
$\boldsymbol{Z}$	Whole data, $\boldsymbol{Z} = \{\zeta^1, \cdots, \zeta^k, \cdots, \zeta^K\}$
$\boldsymbol{\nu}$	Power allocation for SC, $\boldsymbol{\nu} = \{\nu_1, \cdots, \nu_l, \cdots, \nu_L\}$
α	Dirichlet concentration

[59] X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang, "On the convergence of fedavg on non-iid data," in *Proc. of the International Conference on Learning Representations (ICLR)*, Addis Ababa, Ethiopia, April 2020.

[60] H. Wang, Y. Ding, J. Gu, Z. Li, Y. Lin, D. Z. Pan, F. T. Chong, and S. Han, "Quantumnas: Noise-adaptive search for robust quantum circuits," in *The 28th IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, Seoul, Korea, April 2022, pp. 692–708.

[61] L. Deng, "The MNIST database of handwritten digit images for machine learning research," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 141–142, 2012.

[62] T. H. Hsu, H. Qi, and M. Brown, "Measuring the effects of non-identical data distribution for federated visual classification," *CoRR*, vol. abs/1909.06335, September 2019.

APPENDIX

A. Parameter Shift Rule

Parameter shift rule [28], one of the most known quantum gradient calculators, is utilized to train the model. Subsequently, the eSQNN is trained accordingly using the zeroth stochastic gradient descent algorithm, *e.g.*, quantum natural gradient. Consider that eSQNN consists of I trainable parameters, *i.e.*, $\boldsymbol{\theta}_{t,e}^k = [\theta_{t,e,1}^k, \cdots, \theta_{t,e,i}^k, \cdots, \theta_{t,e,I}^k]$. Then, the partial derivative of k-th device's c-th observable over parameter $\theta_{t,e,i}^k$ is given as follows,

$$\frac{\partial \langle V_c \rangle_{\boldsymbol{\theta}_{t,e}^k}}{\partial \theta_{t,e,i}^k} = \frac{\langle V_c \rangle_{\boldsymbol{\theta}_{t,e}^k + \varepsilon \mathbf{e}_i} - \langle V_c \rangle_{\boldsymbol{\theta}_{t,e}^k - \varepsilon \mathbf{e}_i}}{2\varepsilon} \tag{33}$$

where $\mathbf{e}_i$ denotes the i-th standard basis, and $\varepsilon \in (0, \pi/2]$. We calculate the loss gradient using (33).

B. Proof of Lemma 1

The true label is class c and the predictions and its derivative is canceled out due to the definition of cross-entropy. Then, the cross-entropy loss is simplified as follows,

$$\mathcal{L}_{CE} = -\log p(y_{t,e}^{k,l,c} | \boldsymbol{x}). \tag{34}$$

Hereafter, we denote $\hat{y}_c = y_{t,e}^{k,l,c}$, and $p(\hat{y}_c) = p(y_{t,e}^{k,l,c} | \boldsymbol{x})$. Let's denote the partial derivative of cross-entropy loss and fidelity loss as,

$$G_1 = \frac{\partial \mathcal{L}_{CE}}{\partial \theta_{t,e,i}^k} = \frac{1}{p(\hat{y}_c)} \cdot \frac{\partial p(\hat{y}_c)}{\partial \theta_{t,e,i}^k}, \tag{35}$$

$$G_2 = \frac{\partial \mathcal{F}(\boldsymbol{\psi}_{t,e,\boldsymbol{x}}^{k,L}, \boldsymbol{\psi}_{t,e,\boldsymbol{x}}^{k,l})}{\partial \theta_{t,e,i}^k}. \tag{36}$$

By the triangle inequality, the partial derivative of (12) is bounded as follows,

$$\Big|\frac{\partial \mathcal{L}_{t,e}^{k,l}}{\partial \theta_{t,e,i}^{k}}\Big| \leq \sum_{(\boldsymbol{x},\boldsymbol{y})\in\zeta^k}\Big[\frac{\lambda}{D}|G_1| + \frac{1-\lambda}{D}|G_2|\Big]. \tag{37}$$

We have

$$G_1 = a\left(\sum_{c'\geq 1, c'\neq c}^{C} \frac{\hat{y}_{c'}}{\sum_{c=1}^{C}\hat{y}_c}\right)\cdot\frac{\partial\langle V_c\rangle_{\boldsymbol{\theta}_{t,e}^k\odot\sum_{l'=1}^{l}\Xi_{l'}}}{\partial\theta_{t,e,i}^k}. \tag{38}$$

The bound of G_1 obtained as follows,

$$|G_1| \leq a\Big|\frac{\partial\langle V_{c'}\rangle_{\boldsymbol{\theta}^k\odot\Xi_l}}{\partial\theta_{t,e,i}^k}\Big| \leq a. \tag{39}$$

The former step is due to $\sum_{c'\geq 1, c'\neq c}^{C}\hat{y}_{c'} \leq \sum_{c=1}^{C}\hat{y}_c$, and the latter step is because the gradient using parameter shift rule is bounded to 1 [28]. The term G_2 and its bound are given as,

$$G_2 = 2|\langle\boldsymbol{\psi}_{t,e,\boldsymbol{x}}^{k,L}|\boldsymbol{\psi}_{t,e,\boldsymbol{x}}^{k,l}\rangle|\cdot\frac{\partial\langle\boldsymbol{\psi}_{t,e,\boldsymbol{x}}^{k,L}|\boldsymbol{\psi}_{t,e,\boldsymbol{x}}^{k,l}\rangle}{\partial\theta_{t,e,i}^k}, \tag{40}$$

$$|G_2| \leq 2\Big|\frac{\partial\langle\boldsymbol{\psi}_{t,e,\boldsymbol{x}}^{k,L}|\boldsymbol{\psi}_{t,e,\boldsymbol{x}}^{k,l}\rangle}{\partial\theta_{t,e,i}^k}\Big| \leq 2. \tag{41}$$

The former step is due to $|\langle\boldsymbol{\psi}_{t,e,\boldsymbol{x}}^{k,L}|\cdot\boldsymbol{\psi}_{t,e,\boldsymbol{x}}^{k,l}\rangle|^2 \leq 1$, and the latter step is due to parameter shift rule. Substituting the bound of G_1 and G_2 into LHS of (37), we have the bound,

$$\Big|\frac{\partial\mathcal{L}_{t,e}^{k,l}}{\partial\theta_i^k}\Big| \leq 2 + (a-2)\lambda. \tag{42}$$

Calculate LHS of (37) for all $i\in[1,I]$, the loss gradient is obtained, and its gradient is bounded as,

$$\|\nabla_{\boldsymbol{\theta}_{t,e}^k}\mathcal{L}_{t,e}^{k,l}\| \leq 2 + (a-2)\lambda. \tag{43}$$

Applying these results to g_t^k, we complete the proof.

C. Proof of Lemma 2

We expand the global gradient f_t as follows,

$$\|f_t\|^2 = \Big\|\sum_{l=1}^{L}\frac{1}{Kp_l}\sum_{k\in|X_l|}^{K} g_t^k\odot\Xi_l\Big\|^2 \tag{44}$$

$$\leq \frac{L}{K}\sum_{l=1}^{L}\frac{1}{p_l^2}\sum_{k=1}^{K}\|g_t^k\odot\Xi_l\|^2 \tag{45}$$

$$\leq \frac{L}{K}\sum_{k=1}^{K}\sum_{l=1}^{L}\frac{1}{p_l^2}\cdot\|g_t^k\|^2. \tag{46}$$

The first step is due to Jensen's inequality, *i.e.*,

$$\|\sum_{k=1}^{K}x_k\|^2 \leq K\sum_{k=1}^{K}\|x_k\|^2 \tag{47}$$

and the next step is due to Cauchy-Schwarz inequality, *i.e.*, $\|X\odot\Xi\|^2 \leq \|X\|^2$ [1]. Combining Lemma 1 and latter term of (46), we finalize the proof.

D. Proof of Lemma 3

According to (20) and Assumption 3, the distance between f_t and $\bar{f}_t$ is as,

$$\|f_t - \bar{f}_t\|^2 = \Big\|\sum_{l=1}^{L}\frac{1}{Kp_l}\sum_{k\in|X_l|}^{K}(g_t^k - \bar{g}_t^k)\odot\Xi_l\Big\|^2 \tag{48}$$

$$\leq \frac{L}{K}\sum_{l=1}^{L}\sum_{k=1}^{K}\frac{1}{p_l^2}. \tag{49}$$

This step is due to Jensen's inequality. With Assumption 3, we have $\mathbb{E}\|g_t^k - \bar{g}_t^k\|^2 \leq \sigma_k^2$. Combining these results finalizes the proof.

E. Completing Proof of Theorem 1

Using (20), the distance between Θ_{t+1} to the optimal is as,

$$\|\Theta_{t+1} - \Theta^*\|^2 = \|\Theta_t - \eta_t f_t - \Theta^* + \eta_t\bar{f}_t - \eta_t\bar{f}_t\|^2 \tag{50}$$

$$= \underbrace{\|\Theta_t - \eta_t\bar{f}_t - \Theta^*\|^2}_{G_3} \tag{51}$$

$$+ \underbrace{2\eta_t\langle\Theta_t - \Theta^* - \eta_t f_t, \bar{f}_t - f_t\rangle}_{G_4} \tag{52}$$

$$+ \underbrace{\eta_t^2\|f_t - \bar{f}_t\|^2}_{G_5}. \tag{53}$$

We investigate the bound of G_3 as follows,

$$G_3 = \|\Theta_t - \Theta^*\|^2 \underbrace{-2\eta_t\langle\Theta_t - \Theta^*, \bar{f}_t\rangle}_{G_6} + \eta_t^2\|\bar{f}_t\|^2. \tag{54}$$

The term $G_6/(2\eta_t)$ is bounded as,

$$\frac{G_6}{2\eta_t} \overset{(a)}{\leq} F(\Theta^*) - F(\Theta_t) - \frac{\mu}{2}\|\Theta_t - \Theta^*\|^2 \tag{55}$$

$$\overset{(b)}{\leq} -\frac{1}{2\beta}\|\bar{f}_t\|^2 - \frac{\mu}{2}\|\Theta_t - \Theta^*\|^2 \tag{56}$$

$$\overset{(c)}{\leq} -\frac{\mu}{2}\|\Theta_t - \Theta^*\|^2. \tag{57}$$

The steps (a), (b) and (c) are due to μ-strong convexity, L-smoothness, and $\|\bar{f}_t\|^2 \geq 0$, respectively. Since $\mathbb{E}[f_t] = \bar{f}_t$, $\mathbb{E}[G_5] = 0$. Combining Lemma 2, Lemma 3, and these results, we have the bound of LHS of (50). Summarizing (53) with taking expectation, and under Assumption 1 with a learning rate $\eta_t \leq \frac{1}{\beta}$, the error between the updated global model and its optimum progress as,

$$\mathbb{E}\|\Theta_{t+1} - \Theta^*\|^2 \leq (1-\eta_t\mu)\mathbb{E}\|\Theta_t - \Theta^*\|^2 + \eta_t^2\underbrace{\Big(EL^2(2-\lambda)^2 + L\delta\Big)\sum_{l=1}^{L}\frac{1}{p_l^2}}_{:=B}. \tag{58}$$

Since $\eta_t = \frac{2}{\mu t + 2L - \mu} \leq \frac{1}{L}$, applying (58), we have

$$\Delta_{t+1} \leq (1-\eta_t\mu)\Delta_t + \eta_t^2 B. \tag{59}$$

For diminishing the step-size, we focus on showing that $\Delta_t \leq \frac{v}{t+2\kappa-1}$, where $\kappa = \frac{\beta}{\mu}$ and $v = \max\{2\kappa\Delta_1, 4B/\mu^2\}$ as

elaborated next. It is trivial that $\Delta_1 \leq \frac{v}{2\kappa}$ due to the definition of v. Assuming $\Delta_t \leq \frac{v}{t+2\kappa-1}$, we have

$$\Delta_{t+1} \leq (1-\mu\eta_t)\Delta_t + \eta_t^2 B \tag{60}$$

$$\leq \left(1 - \frac{2}{t+2\kappa-1}\right)\frac{v}{t+2\kappa-1} + \frac{4B/\mu^2}{(t+2\kappa-1)^2} \tag{61}$$

$$= \frac{(t+2\kappa-2)v - (v-4B/\mu^2)}{(t+2\kappa-1)^2} \leq \frac{t+2\kappa-2}{(t+2\kappa-1)^2}v \tag{62}$$

$$\leq \frac{v}{t+2\kappa}. \tag{63}$$

For $t=1$, we obtain

$$v = \max\{2\kappa\Delta_1, \frac{4B}{\mu^2}\} \leq 2\kappa\Delta_1 + \frac{4B}{\mu^2}. \tag{64}$$

Finally, using Assumption 1, (58), the result above, we complete the proof of the theorem.

Soohyun Park has been an assistant professor at Sookmyung Women's University, Seoul, Republic of Korea, since March 2024. She was a postdoctoral scholar at the Department of Electrical and Computer Engineering, Korea University, Seoul, Republic of Korea, from September 2023 to February 2024, where she received her Ph.D. degree in electrical and computer engineering, in August 2023. She also received her B.S. degree in computer science and engineering from Chung-Ang University, Seoul, Republic of Korea, in February 2019.

She was a recipient of the Best Reviewer Award by *ICT Express* (2021) and the IEEE Vehicular Technology Society (VTS) Seoul Chapter Awards.

Hyunsoo Lee is currently pursuing the Ph.D. degree in electrical and computer engineering at Korea University, Seoul, Republic of Korea. He received a B.S. degree in electronic engineering from Soongsil University, Seoul, Republic of Korea, in 2021. His research focuses include deep learning algorithms and their applications to mobility and networking.

He was a recipient of the IEEE Vehicular Technology Society (VTS) Seoul Chapter Award in 2022.

Soyi Jung (Member, IEEE) has been an assistant professor at Ajou University, Suwon, Korea, since September 2022. Before joining Ajou University, she was an assistant professor at Hallym University, Chuncheon, Korea, from 2021 to 2022; a visiting scholar at Donald Bren School of Information and Computer Sciences, University of California, Irvine, CA, USA, from 2021 to 2022; a research professor at Korea University, Seoul, Korea, in 2021; and a researcher at Korea Testing and Research (KTR) Institute, Gwacheon, Korea, from 2015 to 2016. She received her B.S., M.S., and Ph.D. degrees in electrical and computer engineering from Ajou University, Suwon, Korea, in 2013, 2015, and 2021. Her current research interests include network optimization for autonomous vehicles communications, distributed system analysis, big-data processing platforms, and probabilistic access analysis. She was a recipient of IEEE Seoul Section Student Paper Contest Award (2018) and IEEE ICOIN Best Paper Award (2021).

Jihong Park (Senior Member, IEEE) received the B.S. and Ph.D. degrees from Yonsei University, South Korea. He is currently a Lecturer (Assistant Professor) with the School of Information Theory, Deakin University, Australia. His research interests include ultra-dense/ultra-reliable/mmWave system designs, and distributed learning/control/ledger technologies and their applications for beyond-5G/6G communication systems. He served as a Conference/Workshop Program Committee Member for IEEE GLOBECOM, ICC, and WCNC, and for NeurIPS, ICML, and IJCAI. He is an Associate Editor of Frontiers in Data Science for Communications, and a Review Editor of Frontiers in Aerial and Space Networks.

Mehdi Bennis (Fellow, IEEE) is a tenured Full Professor with the Centre for Wireless Communications, University of Oulu, Finland, an Academy of Finland Research Fellow, and the Head of the Intelligent Connectivity and Networks/Systems Group (ICON). He has published more than 200 research papers in international conferences, journals, and book chapters. His main research interests are in radio resource management, heterogeneous networks, game theory, and distributed machine learning in 5G networks and beyond.

He has been the recipient of several prestigious awards, including the 2015 Fred W. Ellersick Prize from the IEEE Communications Society, the 2016 Best Tutorial Prize from the IEEE Communications Society, the 2017 EURASIP Best Paper Award for the Journal of Wireless Communications and Networks, the All-University of Oulu Award for research, the 2019 IEEE ComSoc Radio Communications Committee Early Achievement Award, and the 2020 Clarivate Highly Cited Researcher by the Web of Science. He is an Editor of IEEE Transactions on Communications and the Specialty Chief Editor of Data Science for Communications in the Frontiers in Communications and Networks.

Joongheon Kim (M'06–SM'18) has been with Korea University, Seoul, Korea, since 2019, where he is currently an associate professor at the School of Electrical Engineering. He received the B.S. and M.S. degrees in computer science and engineering from Korea University, Seoul, Korea, in 2004 and 2006; and the Ph.D. degree in computer science from the University of Southern California (USC), Los Angeles, CA, USA, in 2014. Before joining Korea University, he was a research engineer with LG Electronics (Seoul, Korea, 2006–2009), a systems engineer with Intel Corporation (Santa Clara, CA, USA, 2013–2016), and an assistant professor with Chung-Ang University (Seoul, Korea, 2016–2019).

He serves as an editor for IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY and IEEE INTERNET OF THINGS JOURNAL. He was a recipient of Annenberg Graduate Fellowship with his Ph.D. admission from USC (2009), Intel Corporation Next Generation and Standards (NGS) Division Recognition Award (2015), IEEE SYSTEMS JOURNAL Best Paper Award (2020), IEEE ComSoc Multimedia Communications Technical Committee (MMTC) Outstanding Young Researcher Award (2020), and IEEE ComSoc MMTC Best Journal Paper Award (2021). He also received several awards from IEEE conferences including IEEE ICOIN Best Paper Award (2021), IEEE ICTC Best Paper Award (2022), and IEEE Vehicular Technology Society (VTS) Seoul Chapter Awards.