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.2025.3639892

CSQoS: Continual Sparse QoS Measurement for Edge Clouds with GNN-based Variational Bayesian

Heng Zhang, *Member, IEEE,* Liping Yi, *Member, IEEE,* Xiaofei Wang, *Member, IEEE,* and Wenyu Wang

***Abstract*—Edge computing, an emerging paradigm, utilizes decentralized edge nodes to offer low-latency, high-quality network services. Quality of Service (QoS) is a crucial metric to measure network service quality, and network resource scheduling relies on QoS measurement results. However, current QoS measurement methods often measure all QoS data among edge nodes, these dense measurement approaches introduce significant costs. Besides, edge nodes may adopt varied network access manners, these factors cause fluctuations in QoS between edge nodes. But existing QoS measurement works often focus on measuring exact QoS values while merely considering QoS fluctuations, resulting in unreliable measured QoS data. In addition, existing QoS measurement methods often can not support online QoS measurement, leading to stale offline QoS data affecting network resource scheduling. To tackle the two issues, we propose a novel Continual Sparse QoS range measurement method (CSQoS) with four innovative designs: (1) To reduce measurement costs, we propose to measure QoS by only sampling partial QoS data and using them to impute unmeasured QoS data. To achieve sparse QoS imputation, we propose a novel variational Bayesian model (BayGNN) with an edge-enhanced Graph Neural Network (GNN) as the encoder for feature extraction and a Multilayer Perceptron (MLP) as the decoder to predict unmeasured QoS data. (2) To assess QoS data ranges, we design the proposed BayGNN model to produce uncertainty simultaneously. (3) To fulfill reliable online QoS predictions, we incorporate continual learning and residual connections in BayGNN. Experimental results on 2 real-world datasets demonstrate that CSQoS has minimal QoS imputation error with the lowest measurement costs, reducing 17.6% RMSE and 20% sampling costs.**

***Index Terms*—Edge Computing, Edge Clouds, Sparse QoS Measurement, Uncertainty Estimation, Variational Bayesian, GNN, Continual Learning, Residual Connection**

I. Introduction

The rapid proliferation of artificial intelligence (AI) applications—such as Large Language Models (LLMs)—has led to a dramatic surge in the demand for computing resources [1], [2], [3]. As an emerging paradigm, edge computing [4], [5] (e.g., distributed edge clouds) fully exploits decentralized computing resources to deliver low-latency and high-quality network services. Furthermore, modern large-scale Internet of Things (IoT) applications—such as industrial sensing, real-time video analytics at the edge, and autonomous device coordination—offload sensing, inference, storage, and actuation workloads to geographically distributed edge or cloud nodes. These applications are highly sensitive to latency and bandwidth constraints, and thus rely on predictable end-to-end Quality of Service (QoS) among devices, proximate edge nodes, and backend services.

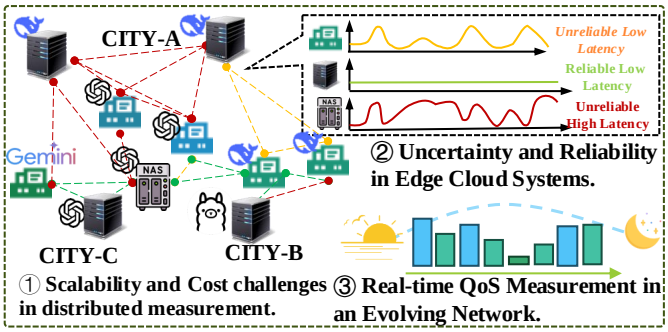


Fig. 1. Illustration for Challenges in the QoS Measurement of Edge Clouds.

The network service quality of edge clouds depends on dense-deployed edge nodes that bring edge servers closer to users in physical distance. Prior work [6] has measured that only if the distances between servers and users are within 30km, network service latency is slightly affected by network fluctuations. However, achieving such latency benefits requires much infrastructure investment for dense deployments. The economic viability of edge clouds necessitates cost-efficient deployment through crowdsourced devices. For example, network service providers can use personal PCs or phones that are willing to provide network services for nearby users as edge nodes[7], [8], [9].

Edge clouds built on available devices with heterogeneous resources (*e.g.*, network bandwidth or computing power [10]) might cause significant fluctuations in network service quality. It's necessary to measure QoS and dynamically choose reliable edge nodes for providing network service, which enables real-time network resource scheduling [11], [12], [13], [14]. However, if an edge cloud incorporates n edge nodes, we measure the QoS data over n nodes with a $O(n^2)$ complexity, imposing substantial measurement burdens for the whole edge system.

Recently measurements [6], [15] of QoS revealed the strong

Heng Zhang, Liping Yi, and Xiaofei Wang are with the College of Intelligence and Computing, Tianjin University, Tianjin, China (e-mail: {hengzhang, lipingyi, xiaofeiwang}@tju.edu.cn).

Wenyu Wang is with PPIO Cloud Computing (Shanghai) Company, Ltd., Shanghai, China (e-mail: wayne@pplabs.org).

Xiaofei Wang is the corresponding author. This work was supported in part by Beijing-Tianjin-Hebei Basic Research Cooperation Special Project, Research on Key Technologies for Efficient Crowd Intelligence Understanding and Situation Deduction in Intelligent Connected Vehicle Environments, under Grant No. F2024201070; in part by National Natural Science Foundation of China under Grant No. U23B2049; in part by Tianjin Natural Science Foundation General Project No. 23JCYBJC00780; in part by the Tianjin Xinchuang Haihe Lab under GrantNo.22HHXCJC00002.

spatio-temporal correlation of service quality in edge clouds. Consequently, we can measure the overall service quality by partial sampling and imputing the unsampled part, thereby alleviating the measurement cost. However, real-time QoS sparse measurement and imputation methods still face the following challenges, as shown in Fig. 1:

- **Scalability and Cost Challenges.** Accurately measuring service quality in large-scale edge cloud systems presents significant challenges due to the quadratic complexity $O(n^2)$ of traditional measurement approaches. As the number of edge nodes grows, the associated measurement overhead becomes increasingly prohibitive.
- **Uncertainty and Reliability in Edge Cloud Systems.** Edge cloud systems consist of geographically dispersed and highly heterogeneous nodes with fluctuating QoS. These dynamic variations introduce uncertainty, making it extremely difficult to ensure consistent reliability.
- **Real-time QoS Measurement in an Evolving Network.** The dynamic nature of edge cloud environments requires continual and real-time QoS measurement, yet existing models struggle to adapt to evolving network conditions. Capturing rapid fluctuations in QoS while maintaining computational efficiency presents a major challenge.

Addressing these challenges is essential for maintaining high service quality and reliability in edge cloud systems. Our study focuses on developing a model that achieves efficient sampling and accurate service quality imputation while minimizing the sampling ratio. To enhance prediction accuracy, we fully exploit the spatial-temporal characteristics of edge nodes. Additionally, our model is designed to quantify uncertainty for every end-to-end path, ensuring reliable predictions while keeping training and evaluation costs minimal.

In this paper, we propose a network service quality sparse measurement framework with an impute algorithm based on GNN-based Variational Bayesian neural networks. The imputation algorithm consists of a Bayesian edge-enhanced graph neural network encoder and a multilayer perceptron decoder, utilizing variational inference for maximum ELBO (evidence lower bound) optimization [16]. The imputation can output prediction confidence intervals, enabling estimation of network fluctuations and reliability. Our contributions are listed as follows:

- We propose a framework combining uniform sampling and a variational Bayesian encoder-decoder. The encoder employs an edge-enhanced GraphSAGE with residual connections to model spatiotemporal dependencies, while an MLP decoder reconstructs missing QoS values. This framework is also integrated with continual learning and Monte Carlo uncertainty estimation to reduce measurement costs and evaluate the uncertainty of QoS, which provides sufficient information for real-time network management and scheduling.
- We propose the Bayesian variational graph auto-encoder with edge-enhanced convolution and residual connections for QoS imputation in edge cloud systems. Our framework introduces (i) a variational latent modeling scheme with reparameterization, enabling uncertainty-aware QoS prediction; (ii) an edge-enhanced GraphSAGE encoder, which explicitly incorporates QoS values on edges to capture richer relational information; (iii) dual residual paths that alleviate over-smoothing and gradient vanishing, thus improving training stability and scalability; and (iv) a continual training mechanism that reuses historical weights across time steps, accelerating convergence and enhancing generalizability. These innovations jointly ensure more accurate, robust, and uncertainty-quantified QoS reconstruction under sparse and dynamic measurement conditions.
- Experimental results show that this method accurately imputes missing QoS values and provides reliable uncertainty estimates for end-to-end paths. It reduces error rates by at least 80% on the EEL dataset and 17% on the FCTE dataset while lowering sampling costs by 20%.

This paper is organized as follows. Section II reviews related work on sparse QoS measurement, reliability modeling, and graph-based imputation. Section III motivates our approach by characterizing spatiotemporal dependencies and uncertainty in edge-cloud QoS. Section IV formalizes the sparse measurement problem and introduces the learning objective. Section V presents CSQoS in detail, including uniform sampling, the variational BayGNN encoder with edge-enhanced GraphSAGE and residual paths, the MLP decoder, Monte Carlo uncertainty estimation, continual training, and a complexity analysis. Section VI reports experiments on the EEL and FCTE datasets, covering metrics, baselines, ablations, and two case studies on uncertainty patterns and temporal robustness. Section VII concludes and outlines future directions.

II. Related Work

This section reviews related work from two major perspectives: (1) **Sparse QoS Measurement**, which focuses on efficient imputation of incomplete network data using sparse modeling and learning-based approaches; and (2) **QoS Reliability Evaluation**, which addresses the uncertainty and robustness of network service quality under dynamic and unpredictable environments.

A. Sparse QoS Measurement

Sparse network QoS measurement originates from sparsified matrix completion techniques. Common approaches in this domain include compressive sensing [17], [18], matrix completion [19], [20], [21], [22], and tensor completion [23], [24], [25], [26]. Compressive Sensing (CS) is a technique that accurately reconstructs sparse vectors from a subset of samples, where the vector contains only a few nonzero elements. However, many matrices indicating that there is still significant room for improvement in enhancing the performance of low-cost data collection in network monitoring systems.

Matrix completion offers effective data recovery under conditions of low data sparsity. However, its capacity to capture the inherent spatio-temporal dynamics in network traffic data remains limited, with a marked decline in performance as the missing data ratio increases [25], [26]. Furthermore, traditional matrix factorization techniques are transductive by nature,

which constrains their generalization to previously unseen nodes and necessitates full retraining when new data points are introduced, thereby reducing their flexibility in dynamic environments. In all, existing methods still face critical trade-offs between dimensionality handling, computational efficiency, and adaptability to fluctuating network conditions.

To overcome these constraints, Inductive Matrix Completion (IMC) has been proposed to improve generalization. However, IMC's performance suffers when content quality is low or unavailable. More recent research has introduced Graph Neural Networks (GNNs) into matrix completion, resulting in the development of the Inductive Graph-based Matrix Completion (IGMC) model[27]. Wu et. al. [28] presents an effective graph-modeling approach using GNNs and contrastive learning, but it does not explicitly consider the temporal dynamics of QoS. Wang et al. [29] propose TPP-GNCF, a comprehensive framework that integrates GNNs, collaborative filtering, trust filtering and differential privacy for IoT-service QoS prediction. However, their work does not explicitly model the temporal dynamics of QoS metrics (e.g., time-varying latency/throughput patterns) and thus may be less effective in contexts where QoS evolves rapidly over time. Furthermore, tensor-based completion algorithms can better handle the missing data in high dimensions. Applying tensor completion to traffic recovery can better capture and exploit spatial-temporal features in data [30], [31], but tensor completion faces higher computational overhead, and although some work has been proposed to accelerate tensor completion using GPU [32], it is still difficult to be used in large-scale edge network monitoring.

In contrast, our work jointly addresses these gaps by (i) performing *uncertainty-aware* QoS imputation through a variational Bayesian encoder–decoder, which outputs both predicted QoS and confidence estimates; (ii) incorporating edge-level QoS values directly into the message-passing process via an edge-enhanced GraphSAGE encoder with dual residual paths, stabilizing learning under sparse sampling; and (iii) enabling continual adaptation across timesteps to capture non-stationary QoS patterns without full retraining. These innovations distinguish our framework from prior tensor completion and IGMC-based models, which typically assume static low-rank structures and lack probabilistic reliability estimation or temporal adaptability.

B. Reliability Evaluation of Network QoS

The concept of reliability in QoS-aware service composition has become a focal point in multiple research domains. To address the uncertainty in network QoS, various techniques and models have been proposed and widely applied in different scenarios.

Fuzzy logic-based methods[33], [34], [35] have also been widely explored to represent uncertainty in QoS attributes. These approaches use fuzzy numbers to handle the imprecision in service parameters, making them particularly useful in environments with fluctuating service levels. For instance, the use of trapezoidal and triangular fuzzy numbers allows for greater flexibility in managing QoS uncertainty. However, scalability remains a challenge, as these methods may struggle with large-scale service environments.

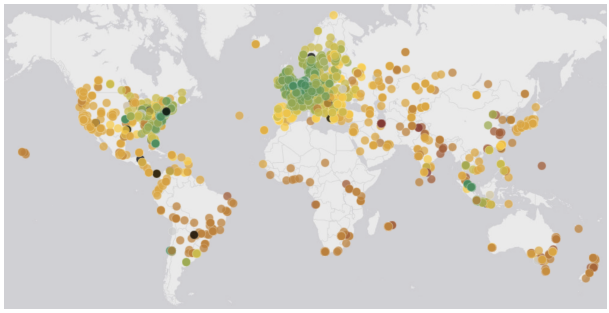


Fig. 2. Illustration of the necessity of QoS completion: RTT distribution based on RIPE Atlas probe data, demonstrating spatial-temporal correlation in network service quality.

Optimization-based models have been proposed to mitigate uncertainty in QoS composition through mathematical formulations. Robust optimization models such as Bertsimas and Sim-based techniques optimize QoS under uncertainty [36], offering near-to-optimal solutions with manageable time complexity. These models are particularly effective in IoT and cloud environments, where dynamic conditions prevail. However, optimization-based approaches can not be well-suited for environments where real-time adaptability is needed, as the models typically rely on pre-defined formulations that may not adjust quickly to rapid changes in system conditions.

Additionally, some bio-inspired algorithms [37], [38] like genetic algorithms and bee colony optimization have demonstrated success in finding optimal or near-optimal compositions under uncertain QoS settings. These algorithms provide efficient exploration of the service search space, albeit at the cost of higher time complexity due to their iterative nature.

Unlike these fuzzy or robust-optimization paradigms, our framework learns a probabilistic latent representation of the network state via variational inference, enabling data-driven, path-level reliability estimation through Monte Carlo sampling of the learned posterior. This provides calibrated QoS uncertainty that can guide adaptive resampling and resource scheduling in edge-cloud systems.

III. Motivation

While the previous section reviewed advances in sparse QoS measurement and reliability evaluation, existing studies still leave several open challenges unaddressed. In particular, most current approaches either overlook the inherent spatial-temporal dependencies in QoS data or lack adaptability to dynamic edge environments. These limitations motivate our work, which aims to develop a continual, adaptive, and graph-based framework for accurate QoS measurement under sparse sampling conditions.

A. Characterizing Spatial-Temporal Dependencies in QoS.

The motivation for this study is from the strong spatial-temporal correlation observed in end-to-end QoS metrics [39], [6], [40]. Empirical measurements reveal that network QoS exhibits structured dependencies across both spatial and temporal

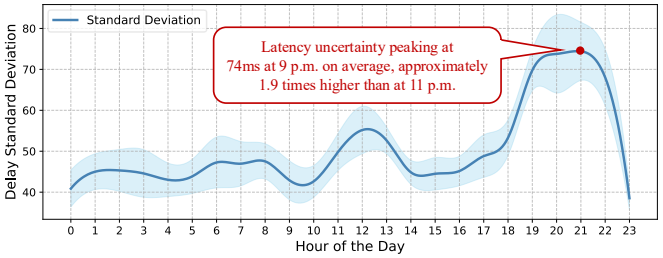
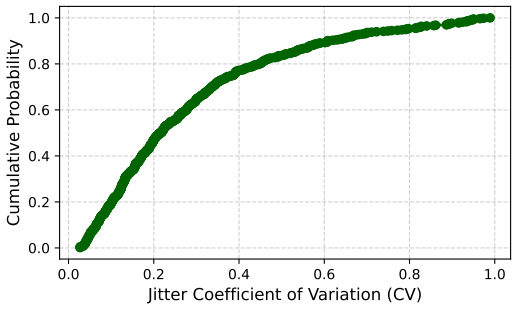
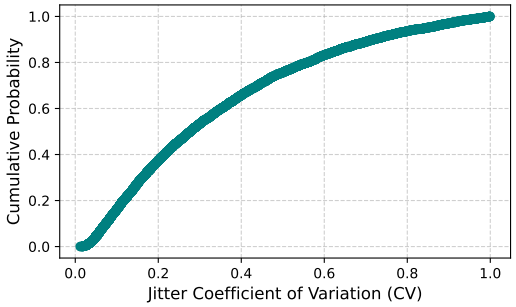


Fig. 3. Diurnal Latency Uncertainty (Std. Dev.) Trends with Evening Peaks in Edge Cloud Systems (EEL Dataset).



(a)



(b)

Fig. 4. QoS uncertainty, expressed as the coefficient of variation (C.V.), on the EEL dataset: (a) spatial distribution across source-destination pairs; (b) temporal variation across different dates.

dimensions, which can be leveraged to optimize measurement strategies. To illustrate this correlation, we analyze the Round Trip Time (RTT) distribution using global data from RIPE Atlas probes, as shown in Fig. 2. The visualization highlights a clear geographical pattern, where RTT values exhibit smooth transitions. This spatial coherence suggests that network service quality in adjacent regions tends to be similar, reinforcing the notion that network QoS follows a structured distribution rather than a purely random process.

Such spatial-temporal dependencies provide a strong theoretical foundation for adopting sparse sampling and imputation techniques in QoS measurement. Instead of performing exhaustive measurements across all edge nodes, an intelligent sampling strategy can exploit these correlations to infer missing values while maintaining high accuracy. This insight directly informs our proposed approach, which integrates spatial-temporal patterns into an efficient and adaptive QoS measurement framework.

B. Variability and Uncertainty in Edge Cloud QoS

Network QoS is quantified through metrics such as latency, jitter, throughput, and reliability [36]. In edge cloud systems, however, these metrics exhibit pronounced spatiotemporal variability due to two primary factors: (1) Heterogeneity in computational resources (e.g., hardware capabilities, network topologies, and software configurations) introduces inherent performance discrepancies across edge nodes, even under identical operational conditions; (2) Dynamic network fluctuations—including congestion, interference, and bandwidth instability—further amplify QoS volatility, particularly in wireless and mobile edge environments.

To characterize the volatility of network service quality, we quantify uncertainty using the standard deviation of latency. This approach provides a measurable representation of service fluctuation. As depicted in Fig. 3, the EEL dataset demonstrates a distinct diurnal dependency, with latency uncertainty peaking at 74ms during evening hours (e.g., 9 p.m.). Such temporal instability critically impacts latency-sensitive applications, where the seamless orchestration of dependent subservices is crucial. We have also measured the C.V. on EEL dataset, from the aspects of spatial and temporal as shown in Fig. 4, which shows a significant uncertainty.

This analysis validates the challenges outlined in our framework: the spatial-temporal volatility of QoS metrics and the imperative for reliability estimation. Our `CSQoS` model addresses these by jointly optimizing variational Bayesian inference and topological feature extraction, enabling precise uncertainty quantification while minimizing sampling costs.

IV. Problem Definition

This section presents the overall methodology for addressing the QoS sparse measurement problem. Our goal is to develop a continual learning framework that can accurately impute missing QoS values from sparsely sampled observations while adapting to temporal variations in edge-cloud networks. To this end, we first construct a graph-based representation of the edge-cloud system to effectively model spatial dependencies among nodes. Next, we describe the network sampling process that reflects real-world sparse measurement scenarios. Finally, we introduce the continual learning–based imputation framework, which incrementally updates the model to maintain accuracy and efficiency over time. Tab. I summarizes the notations in this paper.

*Definition 1 (**Full QoS Graph**):* Given a directed acyclic graph $\mathcal{G}_t = (\mathcal{V}, \mathcal{E}_t)$ at time t, representing the quality of service within an edge cloud system of n nodes, where: $v_i \in \mathcal{V}$ denotes the i-th node, $e_{ij}^t \in \mathcal{E}_t$ represents the service quality when accessing node v_j from node v_i at time t, including self-loops. The adjacency matrix $\mathcal{A}_t \in \mathbb{R}^{n\times n}$ is defined by $a_{ij} = e_{ij}^t$. The total number of edges is $|\mathcal{E}_t| = n \times n$.

*Definition 2 (**Network QoS Sampling**):* At each time t, we generate a binary random sampling mask $\mathcal{M}_t^\alpha \in \{0,1\}^{n\times n}$ based on the sampling rate α, where each element m_{ij} is independently set to 1 with probability α. The sampled

TABLE I
SUMMARY OF NOTATIONS

Symbol	Description
$\mathcal{G}_t = (\mathcal{V}, \mathcal{E}_t)$	Full QoS graph at time t, where $\mathcal{V}$ is the set of nodes (edge/cloud nodes) and $\mathcal{E}_t$ is the set of directed edges representing QoS between node pairs.
$e_{ij}^t \in \mathcal{E}_t$	Directed edge from node v_i to v_j at time t, whose value encodes measured QoS (e.g., latency, throughput).
$\mathbf{A}_t \in \mathbb{R}^{n \times n}$	Adjacency or QoS matrix at time t, where $(\mathbf{A}_t)_{ij} = a_{ij} = e_{ij}^t$.
$\lvert\mathcal{E}_t\rvert = n \times n$	Number of edges (including self-connections) in the full QoS graph at time t.
α	Sampling rate or probing budget, i.e., the fraction of QoS pairs actually measured at each t.
$\mathbf{M}_t^\alpha \in \{0,1\}^{n \times n}$	Binary sampling mask at time t under sampling rate α. Each entry $(\mathbf{M}_t^\alpha)_{ij} = m_{ij}$ is 1 if the QoS value on edge (i,j) is observed or measured, and 0 otherwise.
m_{ij}	Bernoulli sample indicator for edge (i,j): $m_{ij} = 1$ with probability α and $m_{ij} = 0$ otherwise.
$\mathcal{G}_t' = (\mathcal{V}, \mathcal{E}_t')$	Sampled subgraph at time t after sparse measurement, where $\mathcal{E}_t' = \{e_{ij}^t \in \mathcal{E}_t \mid m_{ij} = 1\}$.
$\mathbf{E}_t$	Ground-truth (dense) QoS matrix at time t containing all pairwise QoS values.
$\mathbf{E}_t'$	Observed (sampled) QoS values at time t, defined only on edges where $m_{ij} = 1$.
$\hat{\mathbf{E}}_t$	Reconstructed or imputed QoS matrix at time t, predicted by the model.
f_t	QoS imputation function at time t, which predicts missing QoS values from sampled data: $\hat{\mathbf{E}}_t = f_t(\mathcal{V}, \mathcal{E}_t')$.
$g(\cdot)$	Continual update function that transfers knowledge from previous timesteps to initialize or adapt f_t using $\{f_1, \ldots, f_{t-1}\}$. Enables continual learning over time.
$\mathbf{Z}_t = \{\mathbf{z}_i^t\}_{i=1}^n$	Latent variable set (latent graph representation) for all nodes at time t.
$p(\mathbf{Z}_t)$	Prior distribution over latent variables $\mathbf{Z}_t$, modeled as a factorized multivariate Gaussian.
$q(\mathbf{Z}_t \mid \mathbf{X}, \mathcal{E}_t')$	Variational posterior distribution approximating $p(\mathbf{Z}_t \mid \mathbf{X}, \mathcal{E}_t')$, parameterized by mean $\boldsymbol{\mu}_t$ and standard deviation $\boldsymbol{\sigma}_t$.
EGS	Edge-enhanced GraphSAGE encoder that incorporates edge QoS values e_{ij} into neighborhood aggregation for better structural representation.
$\text{Res}_\alpha, \text{Res}_\beta$	Residual paths in the encoder. Res_α: per-layer skip connection to mitigate over-smoothing. Res_β: global skip from encoder input to output to stabilize gradients.
$p(\hat{\mathbf{E}}_t \mid \hat{\mathbf{Z}}_t)$	Decoder likelihood over the reconstructed QoS matrix, factorized across node pairs.

subgraph is $\mathcal{G}_t' = (\mathcal{V}, \mathcal{E}_t')$, where: $\mathcal{E}_t' = \{e_{ij}^t \in \mathcal{E}_t \mid m_{ij} = 1\}$, The number of sampled edges is $|\mathcal{E}_t'| = \alpha|\mathcal{E}_t|$.

*Definition 3 (**Sparse Measurement Problem**):* Our goal is to develop an imputation function f_t that predicts the missing QoS values using the sampled data $\mathcal{E}_t'$. To enhance prediction accuracy and reduce training costs, we employ a continual learning approach, updating f_t based on previous models $f_1, \ldots, f_{t-1}$ via a function g.

The optimization problem is formulated as:

$$\begin{aligned} \arg\min_{f_t} \quad & \mathbb{E}_{\mathcal{M}} \left[\left\| \mathcal{E}_t - \hat{\mathcal{E}}_t \right\|_F^2 \right] \\ \text{s.t.} \quad & \mathcal{E}_t' \subset \mathcal{E}_t \\ & \hat{\mathcal{E}}_t = f_t(\mathcal{V}, \mathcal{E}_t') \\ & |\mathcal{E}_t'| = \alpha|\mathcal{E}_t| \\ & f_t = g(f_1, \ldots, f_{t-1}) \end{aligned} \tag{1}$$

where: $\hat{\mathcal{E}}_t$ is the set of imputed QoS values for missing edges. The expectation $\mathbb{E}_{\mathcal{M}}$ is over the randomness of the sampling mask $\mathcal{M}_t^\alpha$. The Frobenius norm $\|\cdot\|_F$ measures the imputation error.

V. METHODOLOGY

A. QoS sampling

To enhance the model's ability to capture the distributional characteristics of the entire QoS matrix and improve the accuracy of imputation, it is crucial to perform uniform sampling across the entire QoS matrix. Assuming that the factors influencing network service quality are uniformly distributed throughout the edge cloud, we model the QoS sampling matrix as independently and identically distributed (i.i.d.). To facilitate feasible analysis and ensure sampling uniformity, we adopt a Bernoulli sampling approach [41] with a given sampling rate α. This method approximates a uniform distribution and ensures an expected number of samples equal to $n \times n \times \alpha$, thereby satisfying the required sampling size. Accordingly, the sampling mask can be represented as:

$$m_{ij} = \begin{cases} 1, u \geq \alpha \ and \ u \sim \mathcal{U}(0,1) \\ 0, u < \alpha \ and \ u \sim \mathcal{U}(0,1) \end{cases} \tag{2}$$

where $u_{ij} \sim \mathcal{U}(0,1)$ denotes a random variable uniformly distributed over the interval $(0,1)$. This formulation ensures that each element of the sampling mask is independently drawn from a Bernoulli distribution with parameter α, effectively serving as an approximation to uniform sampling over the QoS matrix.

B. Variational Modeling and Optimization

1) Prior and Variational Posterior Distributions: In an edge cloud system, the QoS imputation problem can be considered as a regression task. Inspired by the auto-encoder architecture, we here set up two processes to encode the features of node and edge into a latent variable $\mathbf{z}$ and decode the representation into QoS matrix by a probabilistic decoder $p_\theta(\mathbf{x}|\mathbf{z})$ with parameters θ. Next, we will detail the processes. We set the prior distribution of the node latent representations to be a multivariate Gaussian random vector empirically:

$$p(\mathbf{z}) = \prod_{i=1}^{n} p(\mathbf{z}_t^{v_i}) \tag{3}$$

Computing the true posterior $p(\mathbf{z}|\mathcal{X}, \mathcal{E}'_t)$ is intractable due to the complexity of integrating over all possible latent variables. Therefore, we employ variational inference to approximate the posterior with a variational distribution q:

$$q(\mathcal{Z}_t|\mathcal{X}, \mathcal{E}'_t) = \prod_{i=1}^{n} q(\mathbf{z}_t^{v_i}|\mathcal{X}, \mathcal{E}'_t), \tag{4}$$

Since q is the variational distribution of p, they are considered to be of the same distribution. Therefore, we represent the variational distribution q using the mean μ and standard deviation σ. The subsequent Edge-Enhanced Graph Convolutional Network with Residual Paths is then employed to learn and generate these parameters μ and σ, thereby constructing the latent distribution.

2) Optimization: By introducing the variational replacement q, the log-likelihood can be decomposed as Eq. (5). The parameters can be learned and optimized by maximizing the evidence lower bound (ELBO), where $\mathbb{E}_{q(\mathcal{Z}_t|X,\mathcal{E}'_t)}[\log p(\hat{\mathcal{E}}_t|\mathcal{Z}_t)]$ is expected log-likelihood (ELL) with prior distribution and $\mathbb{D}_{KL}[q(\cdot)||p(\cdot)]$ is the Kullback-Leibler divergence between $q(\cdot)$ and $p(\cdot)$.

$$\begin{aligned} L_{ELBO} =& \mathbb{E}_{q(\mathcal{Z}_t|X,\mathcal{E}'_t)}[\log p(\hat{\mathcal{E}}_t|\mathcal{Z}_t)] \\ & - \mathbb{D}_{KL}[q(\mathcal{Z}_t|\mathcal{X}, \mathcal{E}'_t)\|p(\mathcal{Z}_t)] \end{aligned} \tag{5}$$

The first part of the equation represents the ELL loss, which indicates the similarity between the generated latency matrix and the original matrix. The second term captures the discrepancy between the original distribution of $\mathcal{Z}$ and its variational distribution. Maximizing the ELBO loss aims to achieve a more accurate generation of the latency matrix while reducing the distance between the original distribution of $\mathcal{Z}$ and its variational distribution. The overall algorithm can be summarized as Algorithm 1.

3) Reparameterization Trick: Since neural networks require gradient backpropagation during training, directly computing the expectation of the Gaussian distribution formed by μ and σ is not feasible, as it involves the parameters of the deep GNN network. To address this issue and enable backpropagation through the stochastic sampling process, we employ the reparameterization trick. This trick expresses samples from the variational posterior as a deterministic function of the network parameters and a noise variable:

$$\mathbf{z}_i^t = \boldsymbol{\mu}_i^t + \boldsymbol{\sigma}_i^t \odot \boldsymbol{\epsilon}_i, \tag{6}$$

where $\boldsymbol{\epsilon}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\odot$ denotes Hadamard (element-wise) product.

In practice, only part of the QoS data between edge nodes can be measured due to cost. The latent variable $\mathbf{z}$ represents hidden factors describing network conditions. The prior $p(\mathbf{z})$ models our general expectation of these factors, while the variational posterior $q(\mathbf{z}|\mathcal{X}, \mathcal{E}'_t)$ updates this belief after observing sampled data. Maximizing the ELBO helps the model reconstruct missing QoS values while keeping predictions reliable. The reparameterization trick allows the model to learn this process efficiently through backpropagation.

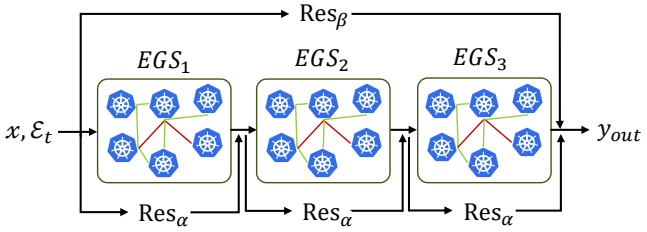


Fig. 5. Illustration of residual paths Res_α and Res_β.

C. Bayesian Edge-enhanced Graph Convolutional Network with Residual Paths

Building on the Bayesian Neural Network's ability to quantify predictive uncertainty, we incorporate graph structures into the encoder by leveraging an *edge-enhanced* GCN architecture. This design captures how QoS relationships vary across nodes and edges, allowing the framework to account for both uncertainty and network topology. By jointly modeling latent distributions over node embeddings and integrating edge-based information, the proposed EGCN yields robust QoS imputation even when data is sparse or partially missing.

1) Edge-enhanced GraphSAGE Encoder: To generate the parameters $\boldsymbol{\mu}_t = \{\boldsymbol{\mu}_i\}$ and $\boldsymbol{\sigma}_t = \{\boldsymbol{\sigma}_i\}$, we extend the standard GraphSAGE model [42] to incorporate edge features, resulting in the *Edge-enhanced GraphSAGE* (EGS). Unlike GraphSAGE, which ignores edge attributes, EGS integrates the QoS values present on the edges. The aggregation function in EGS is redefined as:

$$\text{AGG}_{v_i}^l = \frac{1}{|\mathcal{N}_{v_i}|} \sum_{v_j \in \mathcal{N}_{v_i}} \sigma\left(\mathbf{W}_1^l \cdot \text{concat}\left(\mathbf{x}_j^{(l-1)}, e_{ij}\right)\right), \tag{7}$$

where: l is the current layer index; $\mathcal{N}_{v_i}$ denotes the set of neighboring nodes of v_i; σ is the sigmoid activation function; $\mathbf{W}_1^l$ is a learnable weight matrix; concat$(\cdot)$ represents vector concatenation; $\mathbf{x}_j^{(l-1)}$ is the feature representation of node v_j from the previous layer; e_{ij} is the QoS value on the edge from v_i to v_j.

The hidden state is then updated as:

$$\mathbf{h}_{v_i}^l = \sigma\left(\mathbf{W}_2^l \cdot \text{concat}\left(\mathbf{h}_{v_i}^{(l-1)}, \text{AGG}_{v_i}^l\right)\right), \tag{8}$$

where $\mathbf{W}_2^l$ is another learnable weight matrix.

The latent representation vectors for each node are obtained by stacking the outputs of EGS:

$$\begin{aligned} \boldsymbol{\mu}_t &= \kappa\left(\text{EGS}_\mu(\mathcal{X}, \mathcal{E}'_t)\right), \\ \log\left(\boldsymbol{\sigma}_t\right) &= \kappa\left(\text{EGS}_\sigma(\mathcal{X}, \mathcal{E}'_t)\right), \end{aligned} \tag{9}$$

where $\kappa(\cdot)$ denotes the stacking operator.

2) Incorporating Residual Connections: While the enhanced aggregation in EGS effectively captures node representations, deep graph neural networks can suffer from over-smoothing and vanishing gradients [43]. To address these issues, we introduce residual connections, inspired by ResNet [44], to facilitate better information flow. We introduce two residual paths, Res_α and Res_β, as illustrated in Figure 5:

Algorithm 1 GNN-Variational Bayesian Encoder-Decoder

Input: Graph $G = (V, E)$, node features X, initial embeddings H_0
Output: Updated embeddings H
Initialize parameters θ
for each iteration **do**
  for each node $v \in V$ **do**
    AGG: $h^{\mu}_{N(v)} = \text{AGG}(h^{\mu}_{N(v)}, \{h_u | u \in N(v), e_{uv}\})$
    AGG: $h^{\sigma}_{N(v)} = \text{AGG}(h^{\sigma}_{N(v)}, \{h_u | u \in N(v), e_{uv}\})$
    Update: $h^{\mu}_v = RELU(W \cdot \mathcal{C}(h^{\mu}_v, h^{\mu}_{N(v)}))$
    Update: $h^{\sigma}_v = RELU(W \cdot \mathcal{C}(h^{\sigma}_v, h^{\mu}_{N(v)}))$
    Reparam. : $z_v = h^{\mu}_v + \epsilon \odot h^{\sigma}_v$
  end for
  for each node $v \in V$ and $u \in V$ **do**
    Reconstruct: $\hat{x}_{uv} = \text{InnerProduct}(z_u, z_v)$
  end for
  Reconstruction loss: $\mathcal{L}_{\text{recon}} = \frac{1}{|V'|} \sum_{v \in V'} ||x_v - \hat{x}_v||^2$
  KL divergence loss: $\mathcal{L}_{\text{KL}} = D_{KL}(q_\phi(z|X)||p(z))$
  Loss: $\mathcal{L} = \mathcal{L}_{\text{recon}} + \mathcal{L}_{\text{KL}}$
  Update parameters: $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}$
end for
Return H

- Res_α connects the input and output of each EGS block, mitigating over-smoothing by preserving original features.
- Res_β connects the encoder's input directly to its output, alleviating gradient vanishing and allowing dynamic information to pass through to the decoder.

D. Decoding the QoS Matrix from Latent Representations

After obtaining the latent representation distribution $\mathcal{Z}_t = \{\mathbf{z}_i^t\}_{i=1}^n$ for each edge cloud node, we proceed to reconstruct the original QoS matrix from these representations. With the concrete latent representations $\hat{\mathcal{Z}}_t$, we aim to reconstruct the QoS matrix $\hat{\mathcal{E}}_t$. We design a decoder that models the conditional probability of each QoS value given the latent representations:

$$p(\hat{\mathcal{E}}_t | \hat{\mathcal{Z}}_t) = \prod_{i=1}^{n} \prod_{j=1}^{n} p(e_{ij} | \hat{\mathbf{z}}_i^t, \hat{\mathbf{z}}_j^t). \tag{10}$$

We define the decoder function using an inner product between the latent representations of node pairs, followed by a ReLU activation function to ensure non-negativity of the QoS values:

$$\hat{e}_{ij} = \text{ReLU}\left((\hat{\mathbf{z}}_i^t)^\top \hat{\mathbf{z}}_j^t\right). \tag{11}$$

This formulation leverages the relational information captured in the latent space to reconstruct the QoS values between nodes. Fig. 6 illustrates the architecture of the variational Bayesian graph autoencoder used in our model. By decoding the latent representations through the designed decoder model, we reconstruct the QoS matrix and address the sparse measurement problem.

E. QoS Uncertainty Estimation

We employ a Monte Carlo sampling methodology to systematically sample from the posterior distribution generated by the model, providing a rigorous means to analyze the uncertainty inherent in the original distribution. By drawing multiple independent samples $q_i \sim q$ from the posterior, we are able to explore the distributional characteristics of the model's predictions, facilitating a more nuanced understanding of the stochastic nature of the underlying QoS predictions.

Monte Carlo sampling offers a scalable and flexible mechanism for uncertainty quantification, especially in scenarios where the posterior distribution exhibits non-standard or complex forms that are difficult to characterize analytically. The variance estimates derived from the sampling process are then used as proxies for the uncertainty in the QoS predictions, aiding in the identification of model limitations and informing subsequent refinements. This method ultimately enables more informed decision-making and more resilient prediction frameworks, particularly in dynamic, real-world network environments where QoS metrics fluctuate over time.

F. Continual Model Training and Inference

We want to use the existing knowledge to accelerate the convergence and enhance the generalizability of the model. Previous mathematical methods cannot infer new matrices with existing knowledge. We hope to be able to utilize what we have learned currently for the next round of training and inference. Here we introduce another function g, to fully utilize the weights from t to $t-1$, to initialize the weights at t. At the beginning of training at time t, we first initialize the model parameters with Eq. (12), and in this way, the temporal information and historical model can be memorized and retained for current imputation.

$$f = g(f_1, ...f_{n-1}). \tag{12}$$

G. Computational Complexity Analysis

In the analysis of computational complexity, we assume that the number of hidden units is constant. The size of the time window T is a fixed number during the training and evaluating process. These parameters have no relationship with the input scale, so we don't consider them. Since we used two independent Bayesian QoS graph encoders for feature distribution extraction and MLP for feature reduction, the overall computational complexity can be represented as $\mathcal{O}(2 \times |V| \cdot |E| \cdot |X|) = \mathcal{O}(|V| \cdot |E| \cdot |X|)$, where:

- $|V|$ represents the number of nodes in every graph $\mathcal{G}_j^i$. Since we want to aggregate the nodes in a hotspot graph, we should traverse every node in $\mathcal{G}_j^i$.
- $|E|$ represents the number of edges in every graph $\mathcal{G}_j^i$. For every node in V, we use node sampling to get the node sets. In this study, we sample the $K = 2$ depth of neighbor nodes. However, the $K = 1$ neighbor nodes are uncertain but there exists an upper bound $|V|$ of the neighbor nodes, in which $|V|$ represents the number of edges in V.

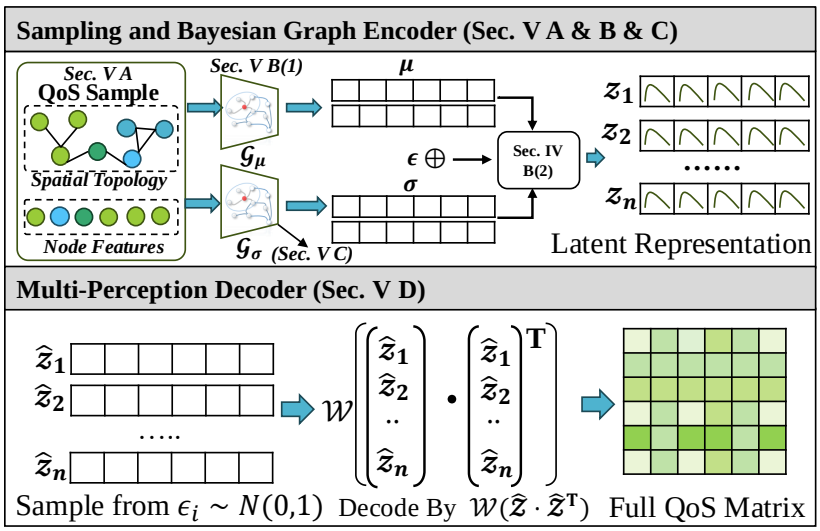


Fig. 6. Illustration of Variational Bayesian Graph Auto-Encoder.

- $|X|$ represents the cost of aggregation. $|X|$ represents the number of node features. Since aggregating is applied to node features, and only simple operations in the aggregation, the complexity can be denoted $\mathcal{O}(|X|)$.

H. Theoretical Imputation Error Bound

To quantify the accuracy of the proposed CSQoS framework in predicting unmeasured QoS values, we provide a theoretical analysis of the expected reconstruction error and its upper bound. We assume that the true QoS matrix admits a low-rank approximation $E_t^{(d)} = Z^*(Z^*)^\top$ of rank d, and that the decoder function is B-Lipschitz continuous with respect to its inputs. Under these assumptions, the expected imputation error of our model satisfies the following upper bound:

$$\mathbb{E}\big[(e_{ij} - \hat{e}_{ij})^2\big] \le 2 \cdot \frac{\|E_t - E_t^{(d)}\|_F^2}{n^2} + 2B^2 \cdot \mathbb{E}_q\big[\|z_i - \mu_i\|_2^2 + \|z_j - \mu_j\|_2^2\big], \tag{13}$$

where μ_i is the posterior mean of node i under $q(z_i \mid X, E_t')$, and $\|\cdot\|_F$ denotes the Frobenius norm.

a) Low-Rank Approximation Bias: The first term $\|E_t - E_t^{(d)}\|_F^2/n^2$ measures the approximation bias—how well the true QoS matrix can be represented by a d-dimensional latent structure. Empirical studies of large-scale Internet and edge-cloud networks have shown that end-to-end QoS matrices exhibit low effective rank due to shared routing paths and congestion bottlenecks. Consequently, this bias term is typically small in practice.

b) Bayesian Estimation Variance: The second term quantifies the estimation variance caused by the stochasticity of the variational posterior. A tighter posterior (with smaller variance around μ_i) directly reduces this term, thereby tightening the error bound. Our training objective,

$$\mathcal{L}_{\text{ELBO}} = \mathbb{E}_q\Big[\log p(\hat{E}_t \mid Z_t)\Big] - D_{\text{KL}}(q(Z_t \mid X, E_t') \,\|\, p(Z_t)), \tag{14}$$

implicitly minimizes both components: the likelihood term reduces the reconstruction bias, while the KL divergence regularizer suppresses posterior variance.

c) Temporal Continuity Regularization: In addition, our continual learning initialization,

$$f_t^{(0)} = (1-\beta) f_{t-1} + \beta f_{\text{init}}, \tag{15}$$

preserves temporal smoothness between consecutive timesteps. Assuming bounded drift of the optimal parameters, $\|f_t^\star - f_{t-1}^\star\|_2 \le \Delta$, this warm-start strategy effectively constrains cumulative bias propagation across time, leading to a tighter and temporally stable error bound.

In summary, Eq. (13) shows that the imputation error of CSQoS is jointly controlled by (i) the low-rank representability of the underlying QoS structure and (ii) the uncertainty of the variational posterior, while the continual learning mechanism maintains this bound stably over time by exploiting temporal smoothness in evolving edge-cloud environments. The posterior variance estimated by our model thus serves as a quantitative indicator of prediction reliability under sparse sampling conditions.

VI. Experiments

A. Datasets

EEL Dataset. [1] The EEL dataset was collected by Zhang [6] in collaboration with PPIO, a nationwide edge cloud provider in China. This dataset comprises 5,174 end-to-end latency measurements across a crowdsourced edge cloud. The data spans from November 27, 2021, to December 17, 2021, containing a total of 943,191,155 data rows.

FCTE Dataset. [2] The FCTE dataset was collected by Xu [45] and measures a leading public edge platform that is densely deployed nationwide. This dataset contains a total of 331,098,145 data rows, with an estimated geographic distribution density of approximately 500 per $10^6 mi^2$.

[1]https://github.com/henrycoding/IWQoS23EdgeMeasurements
[2]https://github.com/xumengwei/EdgeWorkloadsTraces

B. Preprocessing

To construct sampling matrices from both datasets, we performed a series of preprocessing steps. Due to the high sparsity of the FCTE dataset, we aggregated three consecutive 5-minute measurement points into a single matrix to ensure more robust and representative data. Consequently, we obtained:

- 721 fully sampled matrices with a 5-minute interval on the EEL dataset.
- 504 approximately fully sampled matrices with a 15-minute interval on the FCTE dataset.

These matrices serve as the foundation for subsequent QoS imputation and uncertainty estimation experiments. The preprocessing ensures data consistency and mitigates sparsity issues, enabling more reliable model training and evaluation.

C. Experiment Settings

We use PyTorch to implement the framework, and we train for 3000 epochs with a learning rate of 0.001. For every QoS Matrix imputation task, we use a stacked 3-layer EGS block with 64 hidden units and RELU activation. We set the size of hidden units $\|\{\mathcal{Z}\}\| = 64$.We define g as a combination of *warm-start initialization* and *proximity regularization*. Specifically, the continual model at step t is initialized as

$$f_t^{(0)} = (1-\beta) f_{t-1} + \beta f_{\text{init}},$$

where $\beta \in [0,1]$ controls the balance between the previous model and the base initialization. We set β to 0.5 in our experiments, which provides a balanced trade-off between model stability (preserving previously learned knowledge) and adaptability (allowing sufficient update from the new initialization).

D. Evaluation Metrics

Relative Error: We use relative root mean square (RMSE) to evaluate the accuracy of network QoS reconstruction. The relative RMSE is defined as Eq. (16).

$$relative - RMSE = \frac{\sqrt{\sum_{(i,j)\in\bar{\mathcal{G}}}(e_{ij}-\hat{e}_{ij})^2}}{\sqrt{\sum_{(i,j)\in\bar{\mathcal{G}}}(e_{ij})^2}}, \qquad (16)$$

E. Baseline Models

We compare our model with the following imputation methods:

- **Mean.** Fills missing values with the mean of measured samples.
- **KNN** [46], [47]. Uses observed values from k nearest neighbors, weighted by Euclidean distance.
- **SVD** [48], [49]. Matrix completion via low-rank SVD decomposition.
- **MICE** [50], [51]. Iteratively models each missing variable conditioned on observed ones using multiple regression.
- **Spectral** [52], [53]. Nuclear norm–regularized matrix completion with iterative soft-thresholded SVD.
- **GCN** [42], [54]. GraphSAGE-based node embeddings for inductive learning.
- **GAT** [55], [56]. Graph attention networks with masked self-attentional layers for graph-structured data.

F. Performance Evaluation

We evaluate the recovery performance of model QoS data based on the relative Root Mean Square Error ($relative - RMSE$) metric for each sampling rate. We calculate the $relative - RMSE$ metric for each sampling ratio on the two datasets. We collect the performance metrics of baseline models and our model into Tab. II.

Our model shows the best performance than the baselines. By horizontally comparing each baseline model in the table, we observe that our model achieves the best sampling performance at the same sampling rate. This indicates that our model effectively leverages the spatio-temporal correlations within the dataset for learning and inference, thereby achieving superior data recovery performance compared to other models.

We find that our model achieves a reduction in error rates of at least 80% on the EEL dataset and 17% on the FCTE dataset. There is a slight disparity in recovery performance between the two datasets, attributed to partial data loss in the FCTE dataset, resulting in an actual sampling rate lower than the estimated sampling rate.

As the sampling rate increases, the recovery error tends to converge to a fixed value. By comparing columns in the table, we observe varying degrees of convergence in sampling errors, regardless of the imputation methods employed, including mean imputation, mathematical-based imputation methods (KNN, SVD, MICE, Spectral), or machine learning-based approaches (GCN, GAT, Ours), among others. This suggests that for each sampling rate, there exists an optimal sampling upper limit. Beyond this limit, increasing the sampling rate does not significantly improve the final prediction results. This also underscores the rationale behind sparse sampling, indicating its efficacy in practical applications.

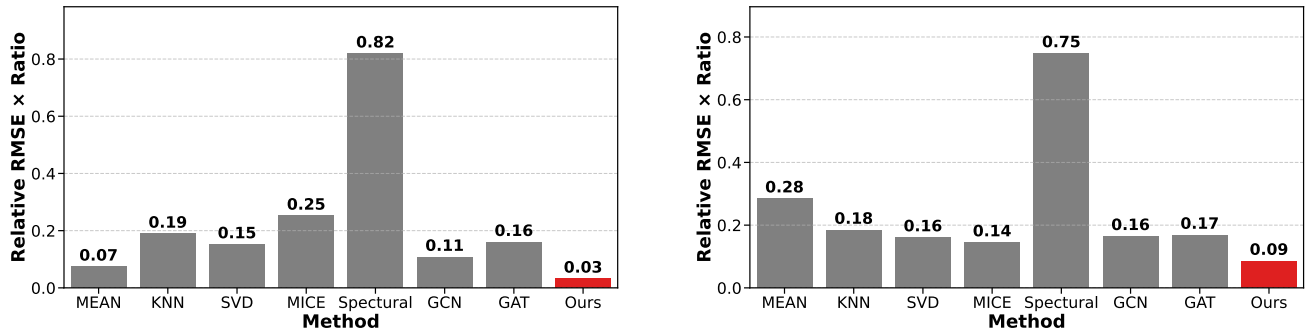
Even as sampling approaches complete coverage of the dataset, factors such as measurement noise, dynamically changing network conditions, and data inaccuracies can prevent the imputation error from converging to zero. Therefore, the residual gap between the predicted values and actual measurements reflects not only the impact of incomplete sampling, but also the underlying stochasticity and uncertainty inherent in real-world network monitoring processes.

Our model adapts well to lower sampling rates, resulting in significant cost reductions. We measured the relative-RMSE values after the recovery error converges and the corresponding sample ratio for each baseline model, as shown in Fig. 7. Compared to other models, ours achieves high recovery accuracy with only approximately 20% data sampling on the EEL dataset and optimal recovery accuracy with approximately 50% data sampling on the FCTE dataset. Our model's convergence sampling rate is at least 20% lower than that of other models, indicating a reduction of at least 20% in sampling costs.

The superior performance of CSQoS under varying sampling rates can be attributed to three key design factors.

TABLE II
PERFORMANCE OF RELATIVE-RMSE

	EEL								FCTE							
RATIO	MEAN	KNN	SVD	MICE	Spectural	GCN	GAT	**Ours**	MEAN	KNN	SVD	MICE	Spectural	GCN	GAT	**Ours**
0.1	0.38	0.46	0.76	0.40	1.05	0.36	0.30	**0.19**	0.71	0.72	0.84	0.61	1.08	0.42	0.42	**0.35**
0.2	0.37	0.35	0.52	0.35	1.04	0.36	0.31	**0.16**	0.7	0.6	0.57	0.46	1.06	0.42	0.41	**0.24**
0.3	0.37	0.31	0.41	0.33	1.03	0.35	0.31	**0.16**	0.7	0.4	0.39	0.29	1.05	0.42	0.41	**0.20**
0.4	0.37	0.29	0.38	0.31	1.02	0.35	0.32	**0.15**	0.7	0.32	0.30	0.25	1.03	0.42	0.41	**0.19**
0.5	0.37	0.28	0.38	0.31	1.01	0.35	0.32	**0.15**	0.7	0.28	0.25	0.22	1.01	0.42	0.41	**0.17**
0.6	0.37	0.28	0.38	0.31	0.99	0.35	0.32	**0.15**	0.7	0.26	0.23	0.21	0.99	0.42	0.42	**0.17**
0.7	0.37	0.27	0.39	0.31	0.98	0.35	0.33	**0.15**	0.7	0.24	0.21	0.19	0.96	0.42	0.42	**0.17**
0.8	0.37	0.27	0.43	0.30	0.95	0.35	0.33	**0.15**	0.7	0.23	0.2	0.18	0.91	0.42	0.42	**0.16**
0.9	0.37	0.27	0.39	0.28	0.91	0.35	0.33	**0.15**	0.7	0.23	0.2	0.18	0.83	0.41	0.41	**0.17**



(a) (b)

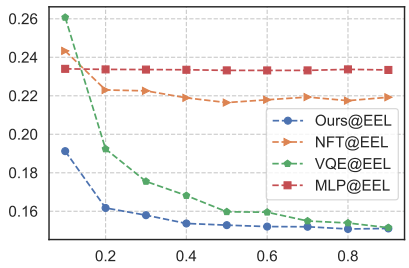
Fig. 7. Evaluation of Cost-Weighted Imputation Error (Relative RMSE × Ratio) for Various Methods on EEL (a) and FCTE (b). Smaller values indicate superior imputation quality. The proposed approach ("Ours") achieves the best trade-off between accuracy and missing-rate sensitivity.

The edge-enhanced GraphSAGE encoder captures high-order topological dependencies, enabling robust structural inference even when the number of observed QoS entries is small. The variational Bayesian mechanism regularizes the latent representation by modeling uncertainty, preventing overfitting to sparsely observed data and improving generalization at low sampling ratios. The continual learning mechanism allows model parameters to be incrementally updated with new observations, which stabilizes learning and accelerates convergence when the sampling rate increases. Together, these components enable CSQoS to maintain both accuracy and stability across a wide range of measurement densities.

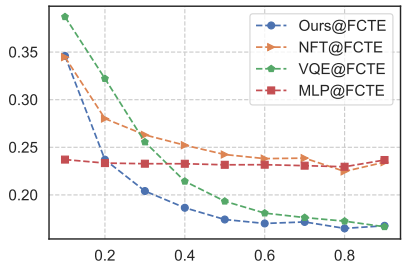
G. Ablation Study

To better understand the contributions of different components in our model, we conducted a comprehensive ablation study on both the EEL and FCTE datasets. The experimental setup includes the following variations:

- **No Fine-Tune. (NFT)** In No Fine-Tune, we ablate the continuous learning part of the model as a way to explore the role of continuous training and inference on accuracy.
- **MLP.** In the MLP, we only use the MLP decoder architecture. We wish to use this experiment to explore the impact of using topological relationships to impute the network quality matrix.
- **Variational QoS Encoder.(VQE)** In variational QoS encoder, we ablated the MLP decoder and used the variational QoS encoder block to output the complemented values directly. Since the output of the encoder is a normal distribution, we use the mean values as the imputed QoS.



(a)



(b)

Fig. 8. Ablation Study on (a) EEL and (b) FCTE datasets.

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.2025.3639892

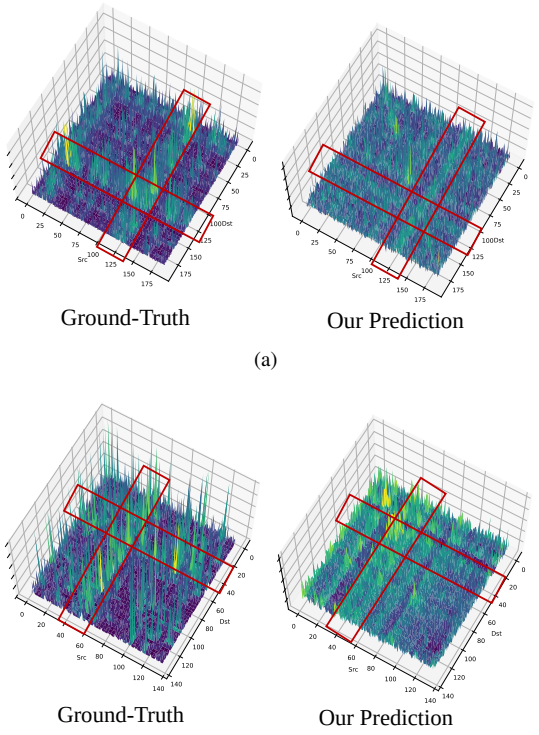


Fig. 9. QoS uncertainty estimation on (a) EEL and (b) FCTE datasets. Monte Carlo sampling (20 runs) is applied to assess prediction variance at 0.3 and 0.5 sampling ratios for EEL and FCTE. The plots depict model-estimated QoS variance against ground-truth variance derived from all samples.

The ablation study results, shown in Fig. 8, reveal several key insights regarding the effectiveness of different model components. Firstly, topological awareness plays a crucial role in QoS imputation, as evidenced by the consistently superior performance of our full model compared to the **MLP** variant. The **MLP** model, which lacks structural information, exhibits higher and more stable imputation errors across all sampling rates, particularly in high-density scenarios where node relationships are critical.

Secondly, continual learning significantly enhances model adaptability, as demonstrated by the **NFT** variant, which suffers from reduced accuracy, particularly in low-sampling regimes. The absence of continual adaptation prevents the model from effectively adjusting to evolving data distributions, leading to higher uncertainty in predictions.

Lastly, the **VQE** improves performance in low-sampling scenarios, leveraging probabilistic modeling to enhance imputation under sparse data conditions. However, as sampling rates increase, the full model outperforms **VQE**, suggesting that while variational encoding is beneficial for uncertainty estimation, the decoder component further refines prediction accuracy in high-density settings.

Overall, these results highlight the importance of integrating topological relationships, continual learning, and probabilistic modeling in QoS imputation tasks. Across different sampling ratios, the full CSQoS model consistently achieves the lowest relative RMSE because it effectively balances data sparsity and representation uncertainty. When sampling is extremely sparse,

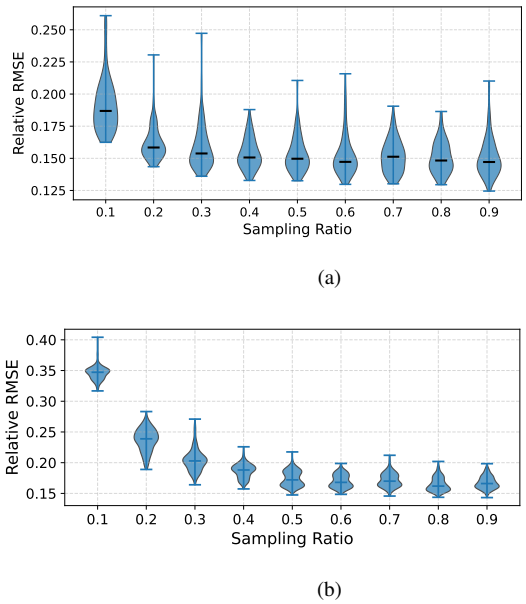


Fig. 10. QoS uncertainty on (a) EEL and (b) FCTE datasets.

Bayesian inference provides reliable uncertainty estimates that guide the model toward stable imputation. As the sampling ratio grows, the continual learning update reuses historical weights to refine latent representations without retraining from scratch, thereby improving computational efficiency and temporal adaptability.

The findings reinforce the need for a comprehensive approach that balances spatiotemporal learning and uncertainty estimation to achieve robust and reliable predictions.

H. Case Study A: QoS Uncertainty Evaluation and Spatiotemporal Fluctuations

We evaluate the model's uncertainty estimation capability in network QoS with Monte Carlo sampling. For the EEL dataset, we select the first QoS matrix with a 0.3 sampling ratio, and for the FCTE dataset, we use the first matrix with a 0.5 sampling ratio, generating 20 stochastic samples to assess prediction variance. The variance 3D plots for the sampled matrices, alongside the ground-truth matrices, are presented in Fig. 9. The ground truth is derived by computing the variance across all sampled matrices.

The results demonstrate that the model effectively captures the variance distribution of QoS values, with nodes highlighted by red boxes exhibiting accurately predicted uncertainty. Prominent striped patterns further reveal nodes with higher uncertainty, indicating significant service quality fluctuations in specific edge cloud nodes. These nodes are critical targets for resampling to improve measurement accuracy. Conversely, their poor network service quality stability suggests they should be deprioritized in scheduling.

The striped patterns highlight the spatiotemporal correlation of uncertainty in network QoS. Fluctuations in a node's service quality to a specific destination often indicate instability in its overall connectivity across the system. This underscores the need for adaptive network scheduling strategies to mitigate

uncertainty propagation. By leveraging uncertainty-aware predictions, the model enables dynamic resampling and network optimization, ensuring more reliable and efficient QoS measurement in edge computing environments.

I. Case Study B: Temporal Stability and Uncertainty Resilience in QoS Prediction

The preliminary analysis revealed that QoS uncertainty exhibits significant temporal variations, with fluctuations occurring at different times of the day. To further assess the robustness of our proposed model in handling these variations, we evaluate its hourly average prediction accuracy across a full day. This analysis helps determine whether the model can effectively mitigate uncertainty and provide reliable predictions under dynamically changing network conditions.

As illustrated in Fig. 10, despite the natural fluctuations in QoS uncertainty, the model consistently achieves stable and high prediction accuracy across all hours of the day for both datasets. This indicates that the model is resilient to hourly variations in network uncertainty, maintaining its predictive performance even during periods of heightened fluctuation. The consistency of the results suggests that our approach effectively captures the underlying spatiotemporal dependencies in network QoS while minimizing the propagation of uncertainty into the predictions. Furthermore, the ability to sustain low relative RMSE across different time periods demonstrates the model's capability to generalize well under real-world network conditions, ensuring reliable QoS estimation in edge computing environments.

VII. Conclusion

In this paper, we introduce a Bayesian variational encoder-decoder framework for QoS data imputation and reliability evaluation in distributed edge cloud environments. The proposed model leverages a Bayesian Edge-Enhanced GraphSAGE encoder to capture structural dependencies in the network, and an MLP-based decoder to reconstruct missing QoS values, enabling accurate prediction under sparse measurement. Through extensive empirical evaluation on real-world edge cloud datasets, our approach achieves state-of-the-art imputation accuracy with substantially reduced sampling cost. In addition, the framework provides calibrated QoS uncertainty estimates via Monte Carlo sampling, which supports more informed and risk-aware network management decisions. The ablation study further confirms the contribution of each component, showing in particular that continual learning improves stability and accuracy under low sampling ratios.

From a deployment perspective, CSQoS effectively reduces measurement overhead by requiring only a small fraction of QoS probes, though it introduces extra cost from continual adaptation and variational inference. This cost is manageable since updates can be scheduled periodically on centralized or resource-rich edge nodes. However, current limitations include unoptimized adaptation overhead and a focus mainly on latency metrics. We will explore adaptive update scheduling, lightweight variational approximations, and broader validation across multi-metric and cross-domain QoS scenarios.

References

[1] S. Zhang, M. Xu, W. Y. Bryan Lim, and D. Niyato, "Sustainable aigc workload scheduling of geo-distributed data centers: A multi-agent reinforcement learning approach," in *GLOBECOM 2023 - 2023 IEEE Global Communications Conference*, 2023, pp. 3500–3505.

[2] S. Duan, D. Wang, J. Ren, F. Lyu, Y. Zhang, H. Wu, and X. Shen, "Distributed artificial intelligence empowered by end-edge-cloud computing: A survey," *IEEE Communications Surveys & Tutorials*, vol. 25, no. 1, pp. 591–624, 2023.

[3] K. C. Serdaroglu, S. Baydere, B. Saovapakhiran, and C. Charnsripinyo, "Q-iot: Qos-aware multilayer service architecture for multiclass iot data traffic management," *IEEE Internet of Things Journal*, vol. 11, no. 17, pp. 28 330–28 340, 2024.

[4] T. Wang, Y. Liang, X. Shen, X. Zheng, A. Mahmood, and Q. Z. Sheng, "Edge computing and sensor-cloud: Overview, solutions, and directions," *ACM Comput. Surv.*, vol. 55, no. 13s, Jul. 2023.

[5] J. Chen, Y. Yang, C. Wang, H. Zhang, C. Qiu, and X. Wang, "Multitask offloading strategy optimization based on directed acyclic graphs for edge computing," *IEEE IoTJ*, 2022.

[6] H. Zhang and et. al., "How far have edge clouds gone? a spatial-temporal analysis of edge network latency in the wild," in *IEEE IWQoS 2023*, 2023.

[7] S. Shen, Y. Feng, M. Xu, C. Zhang, X. Wang, W. Wang, and V. C. Leung, "A holistic qos view of crowdsourced edge cloud platform," in *2023 IEEE/ACM 31st International Symposium on Quality of Service (IWQoS)*, 2023.

[8] J. Yin, Z. Tang, J. Lou, J. Guo, H. Cai, X. Wu, T. Wang, and W. Jia, "Qos-aware energy-efficient multi-uav offloading ratio and trajectory control algorithm in mobile-edge computing," *IEEE Internet of Things Journal*, vol. 11, no. 24, pp. 40 588–40 602, 2024.

[9] L. Zhao, Y. Liu, A. Hawbani, N. Lin, W. Zhao, and K. Yu, "Qos-aware multihop task offloading in satellite–terrestrial edge networks," *IEEE Internet of Things Journal*, vol. 11, no. 19, pp. 31 453–31 466, 2024.

[10] L. Yi and et. al., "QSFL: A two-level uplink communication optimization framework for federated learning," in *Proc. ICML*, vol. 162. PMLR, 2022, pp. 25 501–25 513.

[11] D. Hattori and M. Bandai, "Chunk grouping method for low-latency http-based live streaming," in *2022 IEEE International Conference on Consumer Electronics (ICCE)*, 2022, pp. 01–02.

[12] D. HATTORI and M. BANDAI, "Chunk grouping method to estimate available bandwidth for adaptive bitrate live streaming," *IEICE Transactions on Communications*, vol. E106.B, 07 2023.

[13] E. Figetakis and A. Refaey, "Autonomous mec selection in federated next-gen networks via deep reinforcement learning," in *GLOBECOM 2023*, 2023, pp. 2045–2050.

[14] S. Kumar, A. Goswami, R. Gupta, S. P. Singh, and A. Lay-Ekuakille, "A cost-effective and qos-aware user allocation approach for edge computing enabled iot," *IEEE Internet of Things Journal*, vol. 10, no. 2, pp. 1696–1710, 2023.

[15] H. Wu and et. al., "Network performance analysis of satellite-terrestrial vehicular network," *IEEE Internet of Things Journal*, 2024.

[16] V. Kyrylov, N. S. Bedi, and Q. Zang, "[Re] VAE Approximation Error: ELBO and Exponential Families," *ReScience C*, 2023.

[17] E. J. Candès and et. al., "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Transactions on information theory*, 2006.

[18] J. Haupt, W. U. Bajwa, M. Rabbat, and R. Nowak, "Compressed sensing for networked data," *IEEE Signal Processing Magazine*, 2008.

[19] J.-F. Cai, E. J. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM J. Optim.*, vol. 20, pp. 1956–1982, 2008.

[20] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Foundations of Computational mathematics*, 2009.

[21] R. H. Keshavan, A. Montanari, and S. Oh, "Matrix completion from a few entries," *IEEE Trans. Inf. Theory.*, 2010.

[22] Z. Wen, W. Yin, and Y. Zhang, "Solving a low-rank factorization model for matrix completion by a nonlinear successive over-relaxation algorithm," *Mathematical Programming Computation*, 2012.

[23] S. Gandy, B. Recht, and I. Yamada, "Tensor completion and low-n-rank tensor recovery via convex optimization," *Inverse problems*, 2011.

[24] J. Liu, P. Musialski, P. Wonka, and J. Ye, "Tensor completion for estimating missing values in visual data," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 208–220, 2013.

[25] K. Xie, L. Wang, X. Wang, G. Xie, J. Wen, G. Zhang, J. Cao, and D. Zhang, "Accurate recovery of internet traffic data: A sequential tensor completion approach," *IEEE Trans. Netw.*, 2018.

[26] K. Xie, X. Wang, X. Wang, Y. Chen, G. Xie, Y. Ouyang, J. Wen, J. Cao, and D. Zhang, "Accurate recovery of missing network measurement data with localized tensor completion," *IEEE Trans. Netw.*, 2019.

[27] M. Zhang and Y. Chen, "Inductive matrix completion based on graph neural networks," *arXiv*, 2019.

[28] H. Wu, S. Tian, B. Jin, Y. Zhao, and L. Zhang, " Effective Graph Modeling and Contrastive Learning for Time-Aware QoS Prediction ," *IEEE Transactions on Services Computing*, vol. 17, no. 06, pp. 3513–3526, Nov. 2024.

[29] W. Wang, W. Ma, and K. Yan, "Trust-aware privacy-preserving qos prediction with graph neural collaborative filtering for internet of things services," *Complex & Intelligent Systems*, vol. 11, no. 4, pp. 1–18, 2025.

[30] K. Xie, C. Peng, X. Wang, G. Xie, and J. Wen, "Accurate recovery of internet traffic data under dynamic measurements," in *IEEE INFOCOM*. IEEE, 2017, pp. 1–9.

[31] K. Xie, L. Wang, X. Wang, G. Xie, J. Wen, and G. Zhang, "Accurate recovery of internet traffic data: A tensor completion approach," in *IEEE INFOCOM*. IEEE, 2016.

[32] K. Xie, Y. Chen, X. Wang, G. Xie, J. Cao, J. Wen, G. Yang, and J. Sun, "Accurate and fast recovery of network monitoring data with gpu-accelerated tensor completion," *IEEE Trans. Netw.*, 2020.

[33] X. Jian, Q. Zhu, and Y. Xia, "An interval-based fuzzy ranking approach for qos uncertainty-aware service composition," *Optik*, 2016.

[34] F. Seghir, "A genetic algorithm with an elitism replacement method for solving the nonfunctional web service composition under fuzzy qos parameters," in *2021 International Conference on Artificial Intelligence and Mechatronics Systems (AIMS)*, 2021, pp. 1–7.

[35] ——, "Fdmoabc: Fuzzy discrete multi-objective artificial bee colony approach for solving the non-deterministic qos-driven web service composition problem," *Expert Syst. Appl.*, vol. 167, no. C, Apr. 2021.

[36] M. Razian, M. Fathian, and R. Buyya, "Arc: Anomaly-aware robust cloud-integrated iot service composition based on uncertainty in advertised quality of service values," *Journal of Systems and Software*, 2020.

[37] S. Niu, G. Zou, Y. Gan, Y. Xiang, and B. Zhang, "Towards the optimality of qos-aware web service composition with uncertainty," *International Journal of Web and Grid Services*, vol. 15, no. 1, pp. 1–28, 2019.

[38] Y. Shu, J. Zhang, D. Zuo, and Q. Z. Sheng, "Interval-valued skyline web service selection on incomplete qos," in *2022 IEEE International Conference on Web Services (ICWS)*, 2022, pp. 361–366.

[39] S. Shen, Y. Feng, M. Xu, C. Zhang, X. Wang, W. Wang, and V. C. Leung, "A holistic qos view of crowdsourced edge cloud platform," in *2023 IEEE/ACM 31st International Symposium on Quality of Service (IWQoS)*. IEEE, 2023, pp. 01–10.

[40] M. Xu, Z. Fu, X. Ma, L. Zhang, Y. Li, F. Qian, S. Wang, K. Li, J. Yang, and X. Liu, "From cloud to edge: A first look at public edge platforms," in *IMC'21*, 2021.

[41] V. Kachitvichyanukul and B. W. Schmeiser, "Binomial random variate generation," vol. 31, no. 2, 1988.

[42] W. Hamilton, Z. Ying, and J. Leskovec, "Inductive representation learning on large graphs," *Advances in neural information processing systems*, vol. 30, 2017.

[43] Q. Li, Z. Han, and X.-M. Wu, "Deeper insights into graph convolutional networks for semi-supervised learning," in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, ser. AAAI'18/IAAI'18/EAAI'18. AAAI Press, 2018.

[44] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *arXiv preprint arXiv:1512.03385*, 2015.

[45] M. Xu, Z. Fu, X. Ma, L. Zhang, Y. Li, F. Qian, S. Wang, K. Li, J. Yang, and X. Liu, "From cloud to edge: a first look at public edge platforms," in *Proc.IMC*, 2021.

[46] D. M. P. Murti, U. Pujianto, A. P. Wibawa, and M. I. Akbar, "K-nearest neighbor (k-nn) based missing data imputation," in *2019 5th International Conference on Science in Information Technology (ICSITech)*, 2019.

[47] A. Havolli and et. al., "A comparative analysis of mlr, svr, and knn for improving quality of service in next generation network via machine learning regression," in *2024 13th Mediterranean Conference on Embedded Computing (MECO)*, 2024.

[48] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman, "Missing value estimation methods for dna microarrays," *Bioinformatics*, vol. 17, no. 6, pp. 520–525, 2001.

[49] P. Ashok and et. al., "Developed a machine learning and deep learning model for 5g mimo data based beam selection and intelligent network analytics," *SN Computer Science*, 2025.

[50] S. Van Buuren and K. Groothuis-Oudshoorn, "mice: Multivariate imputation by chained equations in r," *Journal of statistical software*, 2011.

[51] F. Loyola Lopes and et. al., "A label propagation approach for missing data imputation," *IEEE Access*, 2025.

[52] R. Mazumder and et. al., "Spectral regularization algorithms for learning large incomplete matrices," *The Journal of Machine Learning Research*, 2010.

[53] D. P. Isravel and et. al., "Enhanced multivariate singular spectrum analysis-based network traffic forecasting for real time industrial iot applications," *IET Networks*, 2024.

[54] R. Li, H. Shen, Q. Zhang *et al.*, "An edge-enhanced graphsage-based intrusion detection model for the internet of things," *Cluster Computing*, 2025.

[55] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," *arXiv preprint arXiv:1710.10903*, 2017.

[56] H. Latif-Martínez and et. al., "Graph attention networks for contextual anomaly detection in network monitoring," *Computers & Industrial Engineering*, 2025.

Heng Zhang Heng Zhang is currently pursuing a PhD degree from the School of Computer Science and Technology, College of Intelligence and Computing, Tianjin University, Tianjin, China. His current research interests include edge computing and edge clouds. He has published technical papers in AAAI, INFOCOM, TKDE, TMC, SIG KDD, TCC, CIKM, and IOT Journal.

Liping Yi (Member, IEEE) received her Ph.D. degree from the College of Computer Science, Nankai University, Tianjin, China. She is currently a tenure-track associate professor at the School of Artificial Intelligence, Tianjin University. Her research interests include federated learning and LLM-based multi-agent, she has authored technical papers in NeurIPS, ICML, ICDE, ICCV, MM, AAAI, IJCAI, KDD, WWW, ICASSP, ICWS, DASFAA, etc. conferences, and TMC, TSC, TACO, COMST, KBS journals. She served as the reviewer of ICML, NeurIPS, ICLR, KDD, WWW, AAAI, IJCAI, ICCV, CVPR, MM, ICASSP, ICME, FL-IJCAI'23 workshop, FL@FM-NeurIPS'23 workshop, FL@FM-TheWebConf'24 workshop, FL@FM-ICME'24 Workshop conferences, and TMC, AI, TNNLS, KBS, TGCN, Neurocomputing journals.

Xiaofei Wang (Senior Member, IEEE) received the B.S. degree from Huazhong University of Science and Technology, China, and the M.S. and Ph.D. degrees from Seoul National University, Seoul, South Korea. He was a Postdoctoral Fellow with The University of British Columbia, Vancouver, Canada, from 2014 to 2016. He is currently a Professor with the College of Intelligence and Computing, Tianjin University, Tianjin, China. Focusing on the research of edge computing, edge intelligence, and edge systems, he has published more than 200 technical papers in IEEE JSAC, TCC, ToN, TWC, IoTJ, COMST, TMM, INFOCOM, ICDCS and so on. He has received the Best Paper Awards of IEEE ICC, ICPADS, and in 2017, he was the recipient of the "IEEE ComSoc Fred W. Ellersick Prize", and in 2022, he received the "IEEE ComSoc Asia-Pacific Outstanding Paper Award".

Wenyu Wang Wenyu Wang is currently the co-founder of PPIO. In 2004, he led the design of PPLive, a P2P streaming platform with hundreds of millions of users, and served as the co-founder and chief architect of PPLive. In 2018, he co-founded PPIO to focus on advancing edge computing from concept to implementation, providing a distributed cloud infrastructure for next-generation applications. Focusing on edge computing, he has applied for more than 40 patents and published several technical papers in the IEEE Journal on Selected Areas in Communications, INFOCOM and so on.