SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

November 2026, Vol. 69, Iss. 11, 212301:1–212301:15
https://doi.org/10.1007/s11432-025-5152-9

Quantization-aware Bayesian inference framework for unsourced random access in backhaul-limited cell-free massive MIMO systems

Yanxiang JIANG[1], Yige HUANG[1], Zhanghao PAN[1], Tao LI[1], Fu-Chun ZHENG[1,2] & Pengcheng ZHU[1*]

[1]*National Mobile Communications Research Laboratory, Southeast University, Nanjing 211189, China*
[2]*School of Electronic and Information Engineering, Harbin Institute of Technology, Shenzhen, Shenzhen 518055, China*

Received 6 December 2025/Revised 10 June 2026/Accepted 2 September 2026/Published online 26 October 2026

Abstract In this paper, we consider joint active user detection, channel estimation and data decoding for unsourced random access (URA) in cell-free massive multiple-input multiple-output (CF-mMIMO) systems with backhaul constraints, where the access points (APs) may receive concurrent transmissions from hundreds of low-power terminal devices. In such scenarios, the coarse quantization of baseband signals at APs introduces nonlinear distortion that severely degrades joint detection and decoding. To address this issue, we develop a Bayesian inference framework that models the quantized observations as a generalized linear model (GLM). By building upon this model, we extend the vector approximate message passing net (VAMP-Net) and MIMO low-density parity-check net (MIMO-LDPC-Net) into generalized VAMP-Net and generalized MIMO-LDPC-Net, enabling quantization-aware joint activity detection, channel estimation, and data decoding. Simulation results demonstrate that the proposed approach achieves at least an order of magnitude lower detection and decoding error rates than the existing linear-model-based algorithms, particularly under low-bit backhaul quantization, while maintaining moderate additional computational overhead compared with SLM-based unfolded receivers, making it suitable for large-scale CF-mMIMO systems.

Keywords cell-free massive MIMO, unsourced random access, generalized linear model, Bayesian inference, backhaul quantization

Citation Jiang Y X, Huang Y G, Pan Z H, et al. Quantization-aware Bayesian inference framework for unsourced random access in backhaul-limited cell-free massive MIMO systems. Sci China Inf Sci, 2026, 69(11): 212301, https://doi.org/10.1007/s11432-025-5152-9

1 Introduction

Cell-free massive multiple-input multiple-output (CF-mMIMO) has emerged as a promising architecture for future wireless networks [1–3]. The core advantage of CF-mMIMO lies in its ability to eliminate cell boundaries and provide uniformly high spectral efficiency (SE) across a wide coverage area [1]. In this architecture, a large number of distributed access points (APs) jointly serve all users through coherent transmission, offering significant gains in throughput, reliability, and energy efficiency compared with conventional cellular systems [2,3]. Recent work has also investigated cell-free architectures for emerging 6G services, such as multi-static integrated sensing and communication (ISAC) based on network-assisted full-duplex cell-free networks [4]. However, the fully digital implementation of CF-mMIMO is often hindered by high hardware cost and substantial energy consumption. Such energy and implementation constraints are also central to sustainable 6G designs that aim to integrate AI while reducing network energy consumption [5]. To alleviate these practical challenges, the hybrid analog-digital (HAD) architecture is widely adopted. This architecture integrates low-dimensional digital baseband processing with energy-efficient analog phase shifter (PS) networks, providing a practical trade-off between performance and implementation complexity in CF-mMIMO systems [6, 7]. The mmWave band is particularly relevant to CF-mMIMO deployments because it enables large-scale antenna arrays and wide bandwidth, but also imposes stringent hardware and power-consumption constraints. In mmWave APs, hybrid analog-digital architectures and low-resolution ADCs/DACs are commonly adopted to reduce the number of high-resolution RF chains and the amount of data forwarded from APs to the CPU. Moreover, mmWave channels are typically sparse in the angular domain, which naturally matches the sparse Bayesian inference structure used for activity detection and channel estimation in massive access. Therefore,

* Corresponding author (email: p.zhu@seu.edu.cn)

© Science China Press 2026

the mmWave CF-mMIMO scenario provides a representative setting where backhaul-limited quantized observations, sparse channel structures, and centralized cell-free processing coexist.

Beyond maximizing SE, a paramount goal for next-generation wireless systems is to support the exponential increase in the number of connected devices, primarily driven by the Internet of Things (IoT). This requires efficient support for sporadic data transmission across massive numbers of devices, leading to the critical research area of massive access [8, 9], which provides efficient multiple access techniques for accommodating an extremely large number of intermittently active devices. The field of massive access has been accelerated by the introduction of model-driven neural architectures [10–12], which seek to unroll iterative inference algorithms into deep neural networks to leverage the power of data-driven learning and guarantee fast convergence [13]. This direction aligns with recent overviews of AI and communication for 6G networks, which highlight embedding AI capabilities across network layers to optimize efficiency and enhance system robustness [14].

1.1 Related work

Early research on massive access mainly focused on sourced random access (SRA), which assigns each device a unique non-orthogonal pilot to support numerous devices' communication. However, it is impractical to assign each device a unique pilot in the massive access scenario with a huge number of potential devices, and it is a waste of pilot resources, especially with sporadic traffic. To overcome the limitations of conventional scheduling, unsourced random access (URA) has appeared as a low-overhead multiple-access paradigm [15]. In URA, active devices transmit short packets using a common codebook without pre-assigned identities. This novel approach shifts the complexity from resource allocation to the receiver side, where the APs must perform joint activity detection, channel estimation, and data decoding without knowing which device transmitted which packet. This challenging receiver task is typically addressed through Bayesian inference [16], which underlies a family of high-performance algorithms based on approximate message passing (AMP) [9,17,18]. The authors in [9] proposed the structured sparsity-based generalized approximate message passing (SS-GAMP) algorithm by utilizing the expectation maximization (EM) algorithm to learn unknown hyperparameters and applying it to the orthogonal frequency division multiplexing (OFDM) system. The authors in [17] proposed the multiple measurement vector AMP (MMV-AMP) algorithm. Recent advances have further accelerated this field with the introduction of model-driven neural architectures for robust URA, including the vector approximate message passing network (VAMP-Net) for joint activity detection and channel estimation (JADCE) [18] and the MIMO low-density parity-check net (MIMO-LDPC-Net) for data decoding [18].

Nevertheless, most existing model-driven deep-learning receivers for URA are unfolded from inference algorithms developed under the standard linear model (SLM). Although these SLM-based networks are effective when the receiver has access to unquantized or high-resolution observations, they do not explicitly incorporate the nonlinear likelihood induced by coarse quantization. As a result, the quantization distortion is usually treated as additional additive noise, which leads to mismatched extrinsic information and degraded detection, channel estimation, and decoding performance under low-resolution backhaul links. This limitation becomes particularly pronounced in backhaul-limited CF-mMIMO systems, where a large number of distributed APs must forward quantized baseband samples to the CPU. These observations motivate us to reformulate the receiver design under a generalized linear model (GLM), where the quantization operation is incorporated into the Bayesian inference procedure through the likelihood function.

The concept of URA has been actively explored within the context of centralized Massive MIMO, with many previous studies proposing efficient joint active user detection, channel estimation and data decoding schemes under the assumption of an SLM. Specifically, early studies mainly assumed ideal, high-capacity backhaul links for perfectly forwarding received signals to the central processing unit (CPU). However, the architecture of CF-mMIMO, while highly beneficial for coverage and SE, introduces a unique set of challenges. Due to the large number of geographically distributed APs and the requirement for centralized joint processing [19], CF-mMIMO imposes a substantial aggregation burden on the backhaul infrastructure. Consequently, providing high-capacity backhaul to every AP is often economically and energetically prohibitive in dense deployments [20]. This practical limitation necessitates the use of cost-efficient strategies, such as the adoption of low-resolution analog-to-digital converters (ADCs) at the APs [21], which significantly reduce the backhaul link data load and hardware power consumption.

The use of low-resolution ADCs introduces coarse quantization on the received baseband signals, which results in significant nonlinear distortion and quantization noise that severely degrade joint detection and decoding performance [22–24]. The canonical AMP algorithm and its SLM-based variants are fundamentally limited as they do not explicitly account for this quantization noise, making them sub-optimal when applied to the quantized compressed sensing (QCS) problem. To address this algorithmic limitation at the inference level, Rangan [25]

formulated quantized measurements within a GLM and derived the generalized AMP (GAMP) algorithm. This approach explicitly models the observation non-linearity. Subsequent work [26] provided a unified Bayesian inference interpretation that connects AMP, VAMP, and their extensions under a common theoretical framework, offering algorithmic tools for robust inference under various nonlinear measurement models. Recent studies on multiple access and resource allocation for MIMO systems have also highlighted the importance of efficient receiver-side processing and resource-efficient network design in supporting massive connectivity [27, 28].

1.2 Contributions

We note that the GLM formulation itself is a well-established tool for nonlinear measurement models. The technical focus of this work is not to propose a new GLM inference principle, but to integrate quantization-aware GLM inference into the complete CF-mMIMO URA receiver chain. In the considered system, the CPU receives quantized observations from multiple distributed APs and jointly performs activity detection, channel estimation, and LDPC-based data decoding. This requires embedding the quantization likelihood into the extrinsic-information exchange of unfolded VAMP-Net and MIMO-LDPC-Net architectures, so that the resulting receiver can operate directly on quantized AP-to-CPU observations. Motivated by the aforementioned discussions, we develop a unified Bayesian inference framework for URA in CF-mMIMO systems under backhaul capacity constraints. It is worth emphasizing that the novelty of this work does not lie in the GLM formulation itself, which is a well-established tool for nonlinear measurement models. Rather, the technical challenge lies in integrating GLM-based quantization-aware inference into the complete CF-mMIMO URA receiver chain. In the considered system, the CPU receives stacked quantized observations from multiple distributed APs through finite-capacity AP-to-CPU links, and must jointly perform activity detection, channel estimation, and LDPC-based data decoding. This requires the nonlinear quantization likelihood to be embedded into the extrinsic-information exchange between the denoising module and the SLM solver, while preserving compatibility with VAMP-Net and MIMO-LDPC-Net. Moreover, because the likelihood-aware denoiser involves quantization-bin operations and is not directly trained end-to-end, parameter sharing and modular unfolding are adopted to obtain compact and trainable generalized receiver architectures. By introducing GLM, our proposed design incorporates quantization effects into the end-to-end detection and decoding process. The main contributions of this work are summarized as follows.

• GLM-based modeling for backhaul-limited CF-mMIMO. We introduce GLM to accurately capture the nonlinear distortion caused by coarse backhaul quantization. This formulation unifies joint activity detection, channel estimation, and data decoding under a consistent Bayesian inference perspective.

• Alternating Bayesian inference framework. Targeting the proposed GLM, we develop an alternating-update Bayesian inference framework to address the joint activity detection, channel estimation, and decoding problems under backhaul constraints. Through exchanging extrinsic information between the denoiser and SLM solver modules, this framework effectively handles the scenario where SLM is transformed into the GLM due to the non-linear effects of low-resolution bit quantization.

• Unified quantization-aware unfolded receiver architectures. By building upon the modular Bayesian inference framework, we reformulate the existing VAMP-Net and MIMO-LDPC-Net structures into generalized counterparts that exchange extrinsic information with the quantization-aware GLM denoiser. This enables both activity/channel recovery and LDPC-based data decoding to operate directly on quantized AP-to-CPU observations.

The rest of this paper is organized as follows. In Section 2, the backhaul-limited CF-mMIMO systems are described. In Section 3, the unified Bayesian framework and generalized algorithms are detailed. In Section 4, simulation results of the proposed methods are shown. The final conclusion is drawn in Section 5.

1.3 Notations

Notations used in this paper are explained as follows. Lower-case and upper-case boldface letters represent vectors and matrices, respectively. $(\cdot)^{\mathrm{T}}$ denotes the transpose; a_{pq} denotes the (p,q)-th entry of the matrix $\mathbf{A}$. $\mathcal{CN}(\cdot,\cdot)$ and $\mathcal{N}(\cdot,\cdot)$ denote a circularly symmetric complex Gaussian distribution and a real Gaussian distribution. $\mathrm{diag}\{\cdot\}$, $\mathrm{tr}\{\cdot\}$, $\mathbb{E}\{\cdot\}$, and $\mathbb{E}\{\cdot \mid \cdot\}$ are the functions of diagonalization, trace, expectation, and conditional expectation, respectively. $\mathrm{Re}\{\cdot\}$ and $\mathrm{Im}\{\cdot\}$ represent the real part and imaginary part of a complex number, respectively. $p(\cdot)$ denotes the probability density function (PDF) of a random variable; $p(\cdot \mid \cdot)$ denotes the conditional PDF.

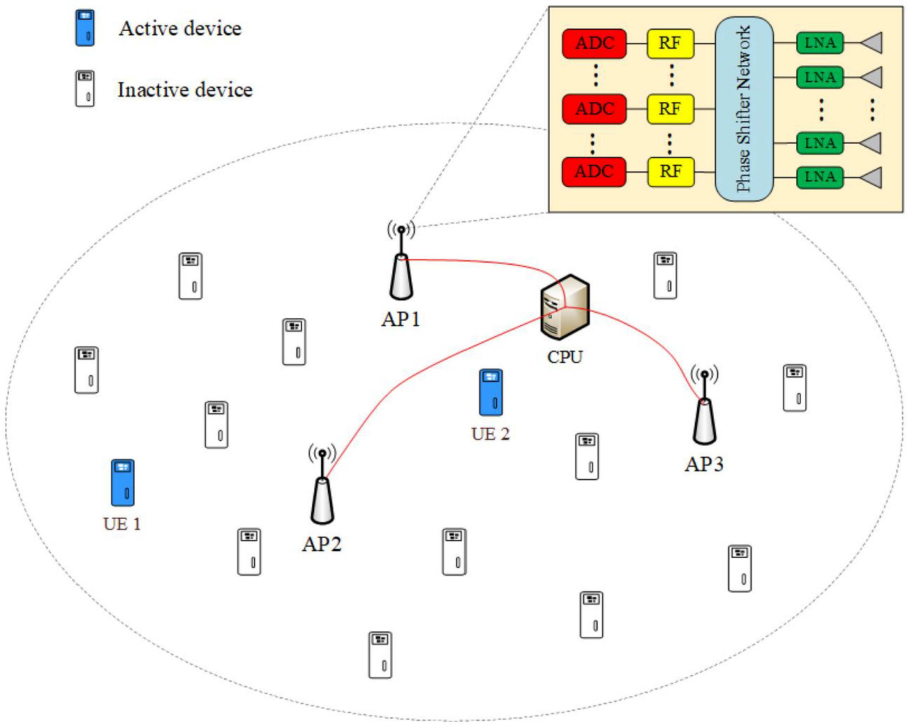

Figure 1 (Color online) Architecture of CF-mMIMO system with unsourced random access. Only a sparse subset of devices is active (e.g., UE 1 and UE 2) and transmits data using a common codebook without pre-assigned identities.

2 System model

This section establishes the mathematical model for the uplink of the backhaul-limited CF-mMIMO systems operating in the URA paradigm. Due to the finite capacity of backhaul links, each AP must quantize its received signals before forwarding them to the CPU. This model incorporates the low-precision backhaul and shows how the resulting nonlinearity converts the classic SLM into GLM.

2.1 System architecture and uplink transmission model

As illustrated in Figure 1, we consider a mmWave CF-mMIMO system with K single-antenna devices and M_{ap} APs. Each AP is equipped with M_{an} antennas connected to N_{rf} radio-frequency (RF) chains. All APs are connected to a CPU through finite-capacity backhaul links. All AP antennas are connected to low-noise amplifiers (LNAs). Each RF chain is connected to an ADC, and all RF chains are connected to the antennas through PS networks [29]. The total number of antennas is defined as $M = M_{\text{an}} \times M_{\text{ap}}$, and the total number of RF chains is defined as $N_r = M_{\text{ap}} \times N_{\text{rf}}$.

It is important to distinguish the adopted CF-mMIMO model from a conventional cellular massive-MIMO model. In a conventional cellular system, the received samples are typically collected and processed at a single co-located BS, and the receiver mainly observes the users associated with that cell. In contrast, in the considered CF-mMIMO architecture, multiple geographically distributed APs simultaneously receive the uplink signals from the same set of potential devices. Each AP first obtains its local low-dimensional RF-chain observations and then forwards the quantized baseband samples to the CPU through finite-capacity AP-to-CPU links. The CPU stacks the quantized observations from all APs along the receive dimension and performs centralized joint activity detection, channel estimation, and data decoding. Therefore, the receive dimension in our matrix model is not the number of RF chains at a single BS, but the aggregate number of RF chains over all distributed APs, i.e., $N_r = M_{\text{ap}} N_{\text{rf}}$.

Under this cell-free formulation, the sparse matrix $\mathbf{X} \in \mathbb{C}^{K \times N_r}$ has a common row support across the distributed AP/RF-chain observations: each row corresponds to one potential device, and an active device contributes channel/effective-signal components to multiple AP/RF-chain observations. Thus, the proposed receiver exploits the distributed multi-AP observations for joint sparse inference at the CPU. When $M_{\text{ap}} = 1$, the model reduces to a conventional single-cell or co-located massive-MIMO receiver model. Hence, conventional cellular-based approaches can be viewed as special cases of the adopted formulation, whereas the focus of this paper is on quantization-aware

Bayesian inference for the distributed AP-to-CPU observation model in backhaul-limited CF-mMIMO systems.

In each coherence time, only K_a ($K_a \ll K$) devices are active. Assumptions include that devices are strictly synchronized with APs and each device has a unique nonorthogonal pilot. Moreover, the CPU is assumed to know the pilots of all users, where L_p, L_c, and L_d denote the pilot lengths for activity detection, channel estimation, and data decoding, respectively[1)]. The active devices transmit pilots for joint active detection and channel estimation in the uplink transmission and the CPU identifies active devices and estimates channels based on the received signals. For reliable data recovery in the URA scenario, the information bits transmitted by each active device are protected by LDPC code [30, 31], which is subsequently decoded by the CPU to recover the original messages.

2.2 Low-bit quantization in the activity-detection phase

Let $\mathbf{Y} \in \mathbb{C}^{L_p \times N_r}$ denote the received pilot matrix used for activity detection. After b-bit uniform scalar quantization, the CPU observes

$$\overline{\mathbf{Y}} = \mathcal{Q}_c(\mathbf{Y}) \;=\; \mathcal{Q}_c(\mathbf{A}\mathbf{X} + \mathbf{N}), \tag{1}$$

where $\mathcal{Q}_c(\,\cdot\,)$ denotes the complex b-bit quantizer that operates element-wise, $\mathbf{A} \in \mathbb{C}^{L_p \times K}$ is the known pilot matrix, $\mathbf{X} \in \mathbb{C}^{K \times N_r}$ is the row-sparse device-state matrix that stacks the effective channel/signal components from all AP RF chains, and $\mathbf{N} \in \mathbb{C}^{L_p \times N_r}$ is the additive white Gaussian noise (AWGN) matrix. Accordingly, $\mathbf{Y}, \overline{\mathbf{Y}} \in \mathbb{C}^{L_p \times N_r}$. The k-th row of $\mathbf{X}$ collects the effective components of the k-th potential device across all AP RF chains. Therefore, $\mathbf{X}$ is row-sparse with a common support over the distributed AP observations, because only a small subset of devices is active in each coherence block. For any scalar $x \in \mathbb{C}$, we have

$$\mathcal{Q}_c(x) = \mathcal{Q}\big(\mathrm{Re}\{x\}\big) + \mathrm{j}\,\mathcal{Q}\big(\mathrm{Im}\{x\}\big), \tag{2}$$

where $\mathcal{Q}(\cdot)$ is a symmetric uniform real-valued quantizer with dynamic range $[y_{\min}, y_{\max}]$. The step size is

$$\Delta = \frac{y_{\max} - y_{\min}}{2^{\,b}}, \tag{3}$$

and the quantization rule is

$$\mathcal{Q}(y) = \begin{cases} \Delta \left\lfloor \dfrac{y}{\Delta} \right\rfloor, & y \geqslant 0, \\ \Delta \left\lceil \dfrac{y}{\Delta} \right\rceil, & y < 0. \end{cases} \tag{4}$$

When the quantizer is not in the saturation region, the real- and imaginary parts of quantization noise $\overline{\mathbf{N}}$ can be modelled as mutually independent uniform random variables in $\left(-\frac{\Delta}{2}, \frac{\Delta}{2}\right)$ with variance $\Delta^2/12$ per component. It should be emphasized that this uniform-noise description is used only to characterize the quantization error under the non-saturation condition. In the proposed GLM-based receiver, the quantizer is not replaced by an additive Gaussian-noise model. Instead, the nonlinear quantization operation is incorporated through the likelihood function $p(\overline{y}_{lr}|y_{lr})$, which is determined by the corresponding quantization interval. The Gaussian distributions used in the message-passing updates represent the extrinsic beliefs of the pre-quantized linear variables produced by the linear module, rather than the distribution of the quantization noise itself. Therefore, the proposed algorithm remains quantization-aware even though Gaussian approximations are used for extrinsic-message representation. Because the non-linear mapping $\mathcal{Q}_c$ is embedded in (1), activity detection under limited backhaul must be formulated as a sparse inference problem under GLM [16], rather than SLM assumed in most previous work.

AP-to-CPU signaling load. In this paper, the terms backhaul and fronthaul both refer to the AP-to-CPU links used to forward the received baseband observations in the considered CF-mMIMO architecture. Since the real and imaginary parts are quantized separately, a complex sample quantized with b bits per real dimension requires $2b$ bits for AP-to-CPU forwarding. Therefore, for a phase with length L_s, the signaling load of forwarding quantized baseband samples from all APs to the CPU is approximately.

$$R_{\mathrm{FH}}(L_s, b) = 2b\, M_{\mathrm{ap}} N_{\mathrm{rf}} L_s \quad \text{bits}, \tag{5}$$

1) The above pilot and synchronization assumptions are adopted to focus on the impact of coarse AP-to-CPU quantization on joint activity detection, channel estimation, and data decoding. In dense CF deployments, pilot contamination or highly correlated non-orthogonal pilots increase the coherence of the sensing matrix $\mathbf{A}$, which may degrade sparse recovery performance. Similarly, timing offsets may destroy the aligned linear mixing model in (1) and introduce additional inter-symbol interference or delayed pilot components. The proposed Bayesian framework can in principle be extended to such cases by using an augmented sensing dictionary that includes delayed pilot replicas or by incorporating timing-offset parameters into the likelihood model. A detailed treatment of asynchronous access and pilot-contamination-aware inference is left for future work.

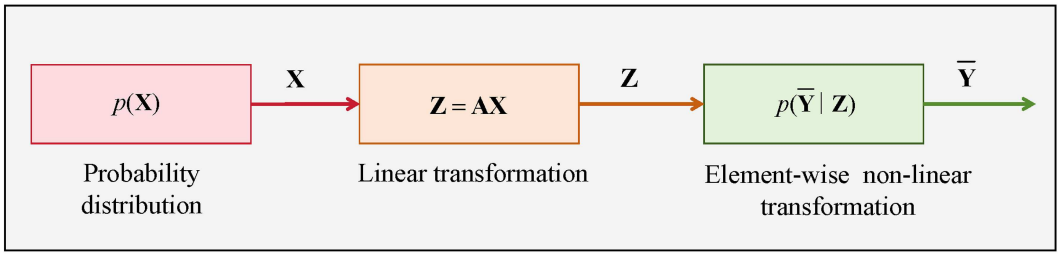

Figure 2 (Color online) Illustration of the GLM. GLM extends SLM by incorporating a nonlinear transformation in the final observation step.

where L_s can be L_p, L_c, or L_d for activity detection, channel estimation, or data decoding, respectively. Compared with forwarding b_0-bit high-resolution samples, the quantized architecture reduces the AP-to-CPU sample-forwarding load by a factor of b/b_0, excluding protocol overhead. The proposed GLM-based receiver is designed to operate directly on these quantized observations, thereby improving inference accuracy without increasing the AP-to-CPU signaling dimensionality.

2.3 Channel estimation and data decoding phase

Following the activity-detection phase, similar quantization effects arise in the subsequent channel-estimation and data-detection phase. During the channel estimation phase, the quantized received pilot matrix can be expressed as follows:

$$\overline{\mathbf{R}} = \mathcal{Q}_c(\mathbf{R}) = \mathcal{Q}_c\big(\mathbf{G}\,\mathbf{H}_{\mathrm{af}} + \mathbf{N}\big), \tag{6}$$

where $\mathbf{H}_{\mathrm{af}}$ denotes the channel matrix in the angular domain, and $\mathbf{G} \triangleq \mathbf{A}\mathbf{F}^{\mathrm{H}}$ the effective sensing matrix, with $\mathbf{F}$ denoting the block discrete Fourier transform matrix.

During the data decoding phase, the quantized observations can be expressed as follows:

$$\overline{\mathbf{Y}}_d = \mathcal{Q}_c(\mathbf{Y}_c) = \mathcal{Q}_c\Big(\sum_{k\in\kappa_a} f(\mathbf{v}_k)\,\mathbf{h}_k^{\mathrm{T}} + \mathbf{N}\Big), \tag{7}$$

where κ_a denotes the active-user set, $\mathbf{v}_k$ the binary message sequence generated by each active-user k, $f(\cdot)$ the encoding function which maps the binary sequence into a complex codeword used for transmission, and $\mathbf{h}_k$ the equivalent channel for active-user k.

The classical SLM is thus recovered only in the ideal case where $b \to \infty$. With finite b, the non-linear distortion in (4)–(7) can severely degrade any algorithm that implicitly assumes linear observations. Therefore, this paper focuses on robust inference under coarse quantization, aiming to recover the active-user channel matrix $\hat{\mathbf{H}}_{\mathrm{af}}$ with high accuracy and reliably decode the transmitted bit sequence $\mathcal{L}(\overline{\mathbf{Y}}_d)$, despite the presence of quantization noise.

3 Joint active detection, channel estimation, and data decoding algorithms

In this section, we first introduce the structure of the unified Bayesian inference framework based on GLM and the update rules of the framework. Then, we detail the design of generalized VAMP-Net and generalized MIMO-LDPC-Net to enable joint detection, channel estimation, and data decoding.

3.1 Bayesian inference framework

Consider the GLM depicted in Figure 2. The device-state matrix $\mathbf{X} \in \mathbb{C}^{K\times N_r}$ is drawn from the prior density $p(\mathbf{X})$. A linear transform $\mathbf{A} \in \mathbb{C}^{L_p\times K}$ produces the intermediate matrix $\mathbf{Z} = \mathbf{A}\mathbf{X} \in \mathbb{C}^{L_p\times N_r}$, after which each element of $\mathbf{Z}$ passes through an element-wise random nonlinear mapping to yield the quantized observation $\overline{\mathbf{Y}} \in \mathbb{C}^{L_p\times N_r}$. This mapping relationship can be described by a conditional probability density function, which can be written as follows:

$$p(\overline{\mathbf{Y}} \mid \mathbf{Z}) = \prod_{r=1}^{N_r}\prod_{l=1}^{L_p} p\big(\overline{y}_{lr} \mid z_{lr}\big) = \prod_{r=1}^{N_r}\prod_{l=1}^{L_p} p\Big(\overline{y}_{lr} \,\Big|\, \sum_{k=1}^{K} A_{lk}x_{kr}\Big), \tag{8}$$

where $\overline{y}_{lr}$, A_{lk}, and x_{kr} denote the corresponding elements of $\overline{\mathbf{Y}}$, $\mathbf{A}$, and $\mathbf{X}$, respectively.

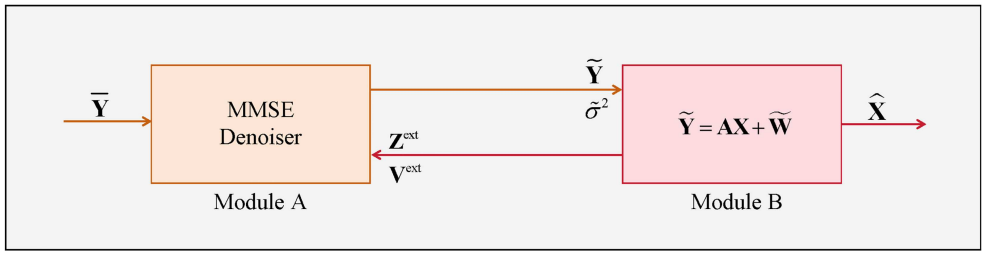

Figure 3 (Color online) Schematic diagram of the Bayesian inference framework. MMSE denoiser module and the SLM solver module iteratively exchange extrinsic information for Bayesian inference.

The goal of the GLM model is to compute the posterior mean estimation of $\mathbf{X}$ given the quantized observations $\overline{\mathbf{Y}}$, denoted as $\mathbb{E}\{\mathbf{X} \mid \overline{\mathbf{Y}}\}$. We assume that both the prior $p(\mathbf{X})$ and the likelihood $p(\overline{\mathbf{Y}} \mid \mathbf{Z})$ are perfectly known. It is noteworthy that SLM is a special case of GLM when the likelihood reduces to AWGN, i.e., no quantization.

To address the inference problem under GLM, Bayesian inference [26] serves as a principled theoretical foundation. To implement this inference efficiently, a modular Bayesian inference framework is proposed herein.

Modular Bayesian inference. The schematic diagram of the modular Bayesian inference framework for sparse recovery under GLM is presented in Figure 3. The inference is achieved by iteratively exchanging extrinsic means and variances between two main modules: Module A (MMSE Denoiser) and Module B (SLM Solver). Module A processes the quantized observation $\overline{\mathbf{Y}}$, extrinsic mean $\mathbf{Z}^{\text{ext}}$ and extrinsic variance $\mathbf{V}^{\text{ext}}$ of $\mathbf{Z}$, and outputs the "denoised" measurements $\tilde{\mathbf{Y}}$ with residual noise variance $\tilde{\sigma}^2$. Module B then treats this as a standard linear model $\tilde{\mathbf{Y}} = \mathbf{AX} + \tilde{\mathbf{W}}$, performs sparse linear inversion, and returns the updated extrinsic statistics.

Update rules for a single element. The entry z_{lr} of $\mathbf{Z}$ can be expressed as follows:

$$z_{lr} = \sum_{n=1}^{N} A_{ln} x_{nr}. \tag{9}$$

According to the central limit theorem, z_{lr} is approximately Gaussian. Let $z_{lr}^{\text{ext}}(t-1)$ and $v_{lr}^{\text{ext}}(t-1)$ denote the extrinsic mean and variance entering Module A at iteration t, respectively. When the quantized observation $\overline{y}_{lr}$ is known, the posterior mean and variance coming out of Module B can be expressed as follows:

$$z_{lr}^{\text{post}}(t) = \mathbb{E}\big\{z_{lr} \mid z_{lr}^{\text{ext}}(t-1), v_{lr}^{\text{ext}}(t-1)\big\}, \tag{10}$$

$$v_{lr}^{\text{post}}(t) = \text{Var}\big\{z_{lr} \mid z_{lr}^{\text{ext}}(t-1), v_{lr}^{\text{ext}}(t-1)\big\} \tag{11}$$

with posterior probability density function $p(z_{lr})$ proportional to

$$p(z_{lr}) \propto p(\overline{y}_{lr} \mid z_{lr}) \mathcal{CN}\big(z_{lr}; z_{lr}^{\text{ext}}(t-1), v_{lr}^{\text{ext}}(t-1)\big). \tag{12}$$

Extrinsic statistics sent to Module B can be obtained by removing the a priori contribution as follows:

$$\tilde{\sigma}_{lr}^2(t) = \Big(\frac{1}{v_{lr}^{\text{post}}(t)} - \frac{1}{v_{lr}^{\text{ext}}(t-1)}\Big)^{-1}, \tag{13}$$

$$\tilde{y}_{lr}(t) = \tilde{\sigma}_{lr}^2(t)\Big(\frac{z_{lr}^{\text{post}}(t)}{v_{lr}^{\text{post}}(t)} - \frac{z_{lr}^{\text{ext}}(t-1)}{v_{lr}^{\text{ext}}(t-1)}\Big). \tag{14}$$

Module B, in turn, solves an SLM problem under the observation model. $\tilde{y}_{lr}(t) = z_{lr} + \tilde{z}$ with $\tilde{z} \sim \mathcal{CN}\big(0, \tilde{\sigma}_{lr}^2(t)\big)$, computes $z_{lr}^{\text{post}}(t)$ and $v_{lr}^{\text{post}}(t)$, and sends back new extrinsic statistics as follows:

$$v_{lr}^{\text{ext}}(t) = \Big(\frac{1}{v_{lr}^{\text{post}}(t)} - \frac{1}{\tilde{\sigma}_{lr}^2(t)}\Big)^{-1}, \tag{15}$$

$$z_{lr}^{\text{ext}}(t) = v_{lr}^{\text{ext}}(t)\Big(\frac{z_{lr}^{\text{post}}(t)}{v_{lr}^{\text{post}}(t)} - \frac{\tilde{y}_{lr}(t)}{\tilde{\sigma}_{lr}^2(t)}\Big). \tag{16}$$

Correspondingly, the GLM inference task is reduced to an iterative cascade of SLM sub-problems. The complete iteration procedures for the Bayesian inference framework within the GLM framework are depicted in Algorithm 1.

Algorithm 1 Bayesian inference framework under GLM.

Initialize $\mathbf{Z}^{\text{ext}}(0)$, $\mathbf{V}^{\text{ext}}(0)$;
for $t = 1, 2, \ldots, T$ **do**
  Update $\mathbf{Z}^{\text{post}}$, $\mathbf{V}^{\text{post}}$ via (10) and (11);
  Compute $\tilde{\sigma}^2(t)$ and $\widetilde{\mathbf{Y}}(t)$ via (13) and (14);
  SLM solver: obtain $\mathbb{E}\{\mathbf{X} \mid \widetilde{\mathbf{Y}}(t)\}$ and $\text{Var}\{\mathbf{X} \mid \widetilde{\mathbf{Y}}(t)\}$;
  Update $\mathbf{V}^{\text{ext}}$, $\mathbf{Z}^{\text{ext}}$ via (15) and (16);
end for
Output $\mathbb{E}\{\mathbf{X} \mid \widetilde{\mathbf{Y}}\}$.

3.2 Joint detection, channel estimation, and data decoding algorithms based on GLM

The previous subsection introduced a Bayesian inference framework for the GLM. Since Module A (the element-wise non-linear denoiser) and Module B (the SLM-solver) depend on the specific non-linearity and the selected sparse-recovery algorithm, their update rules differ slightly across different scenarios. In the following, we specialize the framework to quantized observations, focusing on the sparse signal reconstruction problem with quantization noise.

Posterior mean and variance in Module A. For quantized observations $\overline{\mathbf{Y}} = \mathcal{Q}_c(\mathbf{Y})$, the posterior density can be expressed as follows:

$$p(\mathbf{Y} \mid \overline{\mathbf{Y}}) \;\propto\; p(\overline{\mathbf{Y}} \mid \mathbf{Y})\, \mathcal{CN}(\mathbf{Y}; \mathbf{Y}^{\text{ext}}, V^{\text{ext}}), \tag{17}$$

where $\mathbf{Y}$ is approximately complex Gaussian according to the central-limit theorem, and $p(\overline{\mathbf{Y}} \mid \mathbf{Y})$ factorises over the matrix entries. Because practical receivers quantize the real and imaginary parts separately, we build two independent real-valued Gaussian denoisers inside Module A. Moreover, since Module A performs independent denoising processing on each element, the expectation $\mathbb{E}\{\mathbf{Y} \mid \overline{\mathbf{Y}}\} = \int \mathbf{Y} p(\mathbf{Y} \mid \overline{\mathbf{Y}})\, \mathrm{d}\mathbf{Y}$ can be decomposed element-wise. Focusing on the real part (the imaginary part is handled identically), the MMSE estimate of a single entry can be expressed as follows:

$$\mathbb{E}\{y_{lr} \mid \overline{y}_{lr}\} \;=\; \int y_{lr}\, p(y_{lr} \mid \overline{y}_{lr})\, \mathrm{d}y_{lr}. \tag{18}$$

Let $\{t_k\}_{k=0}^{2^b}$ be the b-bit quantizer thresholds and $c_k = (t_{k-1}+t_k)/2$ be the reproduction level in interval $(t_{k-1}, t_k]$, respectively. For $y_{lr} \sim \mathcal{N}(y_{lr}^{\text{ext}},\, V^{\text{ext}}/2)$, the conditional probability that y_{lr} is quantized to the level $\overline{y}_{lr} = c_k$ can be expressed as follows:

$$P\big(\overline{y}_{lr} = c_k \mid y_{lr}\big) = \int_{t_{k-1}}^{t_k} \mathcal{N}\big(v; y_{lr}^{\text{ext}}, V^{\text{ext}}/2\big)\, \mathrm{d}v. \tag{19}$$

The Gaussian term in (19) should be interpreted as the current extrinsic belief of the unquantized variable, while the likelihood term $P(\overline{y}_{lr}|y_{lr})$ explicitly enforces the quantization bin constraint. This distinction allows the denoiser to retain the nonlinear quantization information instead of treating quantization simply as additive Gaussian noise. Therefore, $p(y_{lr}|\bar{y}_{lr})$ can be expressed as follows:

$$p(y_{lr}|\bar{y}_{lr}) = \frac{P(\bar{y}_{lr}|y_{lr}) N(y_{lr}|y_{lr}^{\text{ext}}, V^{\text{ext}}/2)}{\int P(\bar{y}_{lr}|y_{lr}) N(y_{lr}|y_{lr}^{\text{ext}}, V^{\text{ext}}/2) \mathrm{d}y_{lr}}. \tag{20}$$

The denominator of (20) can be expressed as follows:

$$\int P(\bar{y}_{lr}|y_{lr}) N(y_{lr}|y_{lr}^{\text{ext}}, V^{\text{ext}}/2) \mathrm{d}y_{lr} = \Psi(\overline{y}_{lr}, y_{lr}^{\text{ext}}, V), \tag{21}$$

where

$$\Psi(\overline{y}_{lr}, y_{lr}^{\text{ext}}, V) = \tfrac{1}{2}\Big[\text{erf}\Big(\sqrt{\tfrac{1}{\sigma_w^2+V}}\,(t_k - y_{lr}^{\text{ext}})\Big) - \text{erf}\Big(\sqrt{\tfrac{1}{\sigma_w^2+V}}\,(t_{k-1} - y_{lr}^{\text{ext}})\Big)\Big]. \tag{22}$$

$\text{erf}(\cdot)$ in (22) denotes the complementary error function. For the quantization interval and quantization noise, we have

$$t_k = \text{sign}(\bar{y}_{lr}) \max\{|\bar{y}_{lr} - \Delta/2|\,, |\bar{y}_{lr} + \Delta/2|\}, \tag{23}$$

$$t_{k-1} = \text{sign}(\bar{y}_{lr}) \min\{|\bar{y}_{lr} - \Delta/2|\,, |\bar{y}_{lr} + \Delta/2|\}. \tag{24}$$

Substituting (20) into (18) yields

$$\mathbb{E}\{y_{lr} \mid \overline{y}_{lr}\} = y_{lr}^{\text{ext}} - \frac{V^{\text{ext}}\, \Phi(\overline{y}_{lr}, y_{lr}^{\text{ext}}, \sigma_w^2 + V^{\text{ext}}/2)}{2\, \Psi(\overline{y}_{lr}, y_{lr}^{\text{ext}}, V^{\text{ext}}/2)}, \tag{25}$$

where

$$\Phi(\overline{y}_{lr}, y_{lr}^{\text{ext}}, \sigma^2) = \mathcal{N}(t_k; y_{lr}^{\text{ext}}, \sigma^2) - \mathcal{N}(t_{k-1}; y_{lr}^{\text{ext}}, \sigma^2). \tag{26}$$

The posterior variance of each entry can be obtained from its second-order moment, which can be derived in the same manner as the posterior mean. Specifically, for $\overline{y}_{l,r}$, we have

$$\text{Var}(y_{lr} \mid \overline{y}_{lr}) = \mathbb{E}\{y_{lr}^2 \mid \overline{y}_{lr}\} - \big(\mathbb{E}\{y_{lr} \mid \overline{y}_{lr}\}\big)^2. \tag{27}$$

The posterior mean and variance expressions are subsequently used to update the extrinsic information exchanged between the denoising and linear modules.

VAMP inside Module B. To reconstruct sparse signals, we adopt the GLM-VAMP algorithm of [26] and embed it into the Bayesian framework. VAMP alternates between a linear MMSE (LMMSE) stage and an element-wise MMSE stage with a loop limit T to solve the sparse reconstruction problem in the form of $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{w}$, with $\mathbf{x} \sim \mathcal{CN}(\mathbf{r}_t, \tilde{\sigma}_t^2\mathbf{I})$ and $\mathbf{w} \sim \mathcal{CN}(\mathbf{0}, \sigma_w^2\mathbf{I})$. The expression for the LMMSE estimator is given by

$$\tilde{\boldsymbol{\eta}}(\tilde{\mathbf{r}}_t; \tilde{\sigma}_t) = \mathbf{V}\left(\text{diag}(\mathbf{s})^2 + \frac{\sigma_w^2}{\tilde{\sigma}_t^2}\mathbf{I}_r\right)^{-1}\left(\text{diag}(\mathbf{s})\,\mathbf{U}^{\text{H}}\mathbf{y} + \frac{\sigma_w^2}{\tilde{\sigma}_t^2}\mathbf{V}^{\text{H}}\tilde{\mathbf{r}}_t\right), \tag{28}$$

where $\{\mathbf{U}, \mathbf{s}, \mathbf{V}\}$ denotes the thin singular value decomposition (SVD) of $\mathbf{A}$:

$$\mathbf{A} = \mathbf{U}\,\text{diag}(\mathbf{s})\mathbf{V}^{\text{H}}. \tag{29}$$

The LMMSE output gives both the posterior mean of loop t, i.e., $\widetilde{\mathbf{x}}_t$, and the covariance $\widetilde{\mathbf{C}}$ of $\mathbf{x}$:

$$\widetilde{\mathbf{x}}_t = \tilde{\boldsymbol{\eta}}(\tilde{\mathbf{r}}_t; \tilde{\sigma}_t), \tag{30}$$

$$\widetilde{\mathbf{C}} = \left(\tfrac{\mathbf{A}^{\text{H}}\mathbf{A}}{\sigma_w^2} + \tfrac{1}{\tilde{\sigma}_t^2}\mathbf{I}_K\right)^{-1}. \tag{31}$$

Because $\mathbf{z} = \mathbf{A}\mathbf{x}$ in the GLM, after deriving the output $\widetilde{\mathbf{x}}_{T_{\text{SLM}}}$ of VAMP at inner-loop limit T_{SLM}, we obtain

$$\mathbf{z}^{\text{post}} = \mathbf{A}\,\widetilde{\mathbf{x}}_{T_{\text{SLM}}}, \tag{32}$$

$$\mathbf{v}^{\text{post}} = \text{diag}\big(\mathbf{A}\widetilde{\mathbf{C}}\mathbf{A}^{\text{H}}\big). \tag{33}$$

In practical updates, the posterior variance transmitted externally is the output obtained by averaging all elements as follows:

$$v^{\text{post}} = \tfrac{1}{L_p}\,\text{tr}\big(\mathbf{A}\widetilde{\mathbf{C}}\mathbf{A}^{\text{H}}\big). \tag{34}$$

Eqs. (32) and (34) are plugged into the extrinsic-information expressions given in (15) and (16) to obtain extrinsic statistics. The generalized VAMP algorithm for the GLM model is detailed in Algorithm 2.

Algorithm 2 Generalized VAMP for quantized observations.

Require: Quantized data $\overline{\mathbf{y}}$, sensing matrix $\mathbf{A}$, noise variance σ_w^2, outer-loop limit T, inner-loop (SLM) limit T_{SLM};
1: Initialize $\mathbf{z}^{\text{ext}}(0)$, $v^{\text{ext}}(0)$;
2: **for** $t = 1, 2, \ldots, T$, **do**
3: Compute $\mathbf{z}^{\text{post}}(t)$, $\mathbf{v}^{\text{post}}(t)$ via (10) and (11);
4: Update $\widetilde{\mathbf{Y}}(t)$, $\tilde{\sigma}^2(t)$ via (13) and (14);
5: Run VAMP for T_{SLM} iterations on $\widetilde{\mathbf{y}}(t) = \mathbf{A}\mathbf{x} + \widetilde{\mathbf{w}}(t)$, $\widetilde{\mathbf{w}}(t) \sim \mathcal{CN}\big(0, \tilde{\sigma}^2(t)\mathbf{I}_M\big)$ to obtain $\mathbf{z}^{\text{post}}(t)$ and $v^{\text{post}}(t)$;
6: Generate new extrinsic information $\mathbf{z}^{\text{ext}}(t)$, $v^{\text{ext}}(t)$ via (15) and (16);
7: **end for**
Ensure: The output $\widetilde{\mathbf{x}}_{T_{\text{SLM}}}$ of VAMP at outer-loop limit T.

Complexity discussion. Let $r = \text{rank}(\mathbf{A}) \leqslant \min(L_p, K)$. Because the pilot matrix $\mathbf{A}$ is fixed for all receive columns and all unfolded iterations, its thin SVD is computed only once, before online inference. For $L_p \leqslant K$, this preprocessing requires $O(L_p^2 K)$ operations and is amortized over all coherence blocks. The SVD form avoids explicitly constructing the $K \times K$ Gram matrix $\mathbf{A}^{\text{H}}\mathbf{A}$ and avoids a $K \times K$ inversion in every layer: the inverse in the LMMSE update reduces to r scalar diagonal divisions. Moreover, the scalar posterior variance in (34) is evaluated directly as

$$v^{\text{post}} = \frac{1}{L_p}\sum_{i=1}^{r}\frac{s_i^2}{s_i^2/\sigma_w^2 + 1/\tilde{\sigma}_t^2}, \tag{35}$$

Table 1 FLOP comparison for the activity-detection setting ($K = 2000$, $L_p = 110$, $N_r = 128$, $T = 20$, and $T_{\mathrm{SLM}} = 1$).

Method	FLOP expression	Numerical FLOPs
Thin-SVD preprocessing (once)	$L_p^2 K$	2.420×10^7
SLM-based VAMP-Net	$TT_{\mathrm{SLM}} L_p K N_r$	5.632×10^8
Generalized VAMP-Net	$TT_{\mathrm{SLM}} L_p K N_r + TL_p N_r$	5.635×10^8

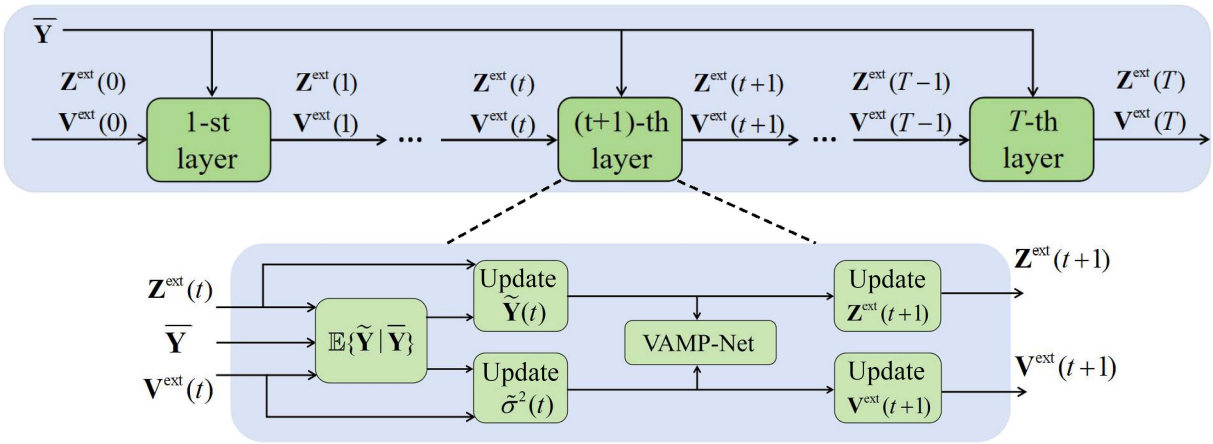

Figure 4 (Color online) Block diagram of the generalized VAMP-Net. Each unrolled block of the generalized VAMP-Net mirrors one generalized VAMP iteration.

which costs $O(r)$ operations and does not require forming either $\widetilde{\mathbf{C}}$ or $\mathbf{A}\widetilde{\mathbf{C}}\mathbf{A}^{\mathrm{H}}$.

For a fair comparison, the generalized VAMP-Net and SLM-based VAMP-Net use the same T unfolded blocks and T_{SLM} internal layers. In each outer iteration, the dominant complexity of the SLM solver in Module B is $O(T_{\mathrm{SLM}} L_p K N_r)$, while the element-wise GLM denoising in Module A requires $O(L_p N_r)$. Therefore, the online complexities of the SLM-based VAMP-Net and generalized VAMP-Net are $C_{\mathrm{SLM}} = O(TT_{\mathrm{SLM}} L_p K N_r)$ and $C_{\mathrm{GLM}} = O(TT_{\mathrm{SLM}} L_p K N_r + TL_p N_r)$, respectively. The one-time thin-SVD preprocessing has complexity of $C_{\mathrm{SVD}} = O(L_p^2 K)$ for $L_p \leqslant K$ and is excluded from the online complexities above. The FLOP counts are summarized and compared in Table 1.

From "generalized VAMP" to "generalized VAMP-Net". To reduce the computational complexity of the original iterative algorithm while preserving its inference structure, we unrolled the generalized VAMP into a neural network architecture, i.e., generalized VAMP-Net. Figure 4 illustrates the block diagram of our proposed generalized VAMP-Net. This architecture is derived by unrolling the T outer iterations of the generalized VAMP algorithm into T identical network blocks. Each block, corresponding to one outer iteration, consists of a non-linear Module A with output $\mathbb{E}\{\widetilde{\mathbf{Y}}|\overline{\mathbf{Y}}\}$ and a learnable T_{SLM}-layer VAMP-Net in [18] as Module B. The modules iteratively exchange updated extrinsic means $\mathbf{Z}^{\mathrm{ext}}(t)$ and variances $\mathbf{V}^{\mathrm{ext}}(t)$, effectively integrating the quantization-aware Bayesian inference into a deep learning framework for joint detection and channel estimation.

Training all blocks end-to-end would break backpropagation because Module A is non-differentiable and would also risk vanishing/exploding gradients. Instead, we share the parameters of the internal VAMP-Net across all blocks and train that sub-net only—greatly reducing model size and avoiding non-differentiability issues.

Training objective of generalized VAMP-Net. Let $\boldsymbol{\theta}_{\mathrm{V}}$ denote the trainable parameters shared by the internal VAMP-Net in Module B. For the i-th training sample, let $\mathbf{X}^{(i)}$ be the ground-truth sparse activity/channel matrix and let $\widehat{\mathbf{X}}^{(i)}_{\boldsymbol{\theta}_{\mathrm{V}}}$ denote the output of the generalized VAMP-Net after the outer GLM iterations. The generalized VAMP-Net is trained by minimizing the normalized mean-squared error (NMSE) between the estimated sparse matrix and the ground truth, i.e.,

$$\mathcal{L}_{\mathrm{GVAMP}}(\boldsymbol{\theta}_{\mathrm{V}}) = \frac{1}{N_{\mathrm{tr}}} \sum_{i=1}^{N_{\mathrm{tr}}} \frac{\left\| \widehat{\mathbf{X}}^{(i)}_{\boldsymbol{\theta}_{\mathrm{V}}} - \mathbf{X}^{(i)} \right\|_F^2}{\left\| \mathbf{X}^{(i)} \right\|_F^2}. \tag{36}$$

No additional quantization-specific penalty term is introduced in the loss function. Instead, the quantization effect is incorporated through the quantized input observations and the likelihood-aware MMSE denoising operation in Module A. Therefore, the training objective remains a standard reconstruction loss, while the network architecture and the inference procedure are quantization-aware.

Extension to LDPC decoding. The proposed GLM-based inference framework can also be applied to the decoding stage by extending the MIMO-LDPC-Net architecture introduced in [18]. In the original MIMO-LDPC-Net, the receiver jointly performs symbol detection and LDPC decoding through iterative message passing between the detection and decoding sub-networks. Here, we reformulate this architecture under the GLM model to explicitly account for coarse quantization, leading to the generalized MIMO-LDPC-Net.

In this extension, Module A remains unchanged, while Module B is replaced by an LDPC-based neural decoder that exchanges extrinsic information with Module A. Specifically, the decoder outputs bit-level log-likelihood ratios (LLRs), which are then converted into the posterior mean and variance of the corresponding BPSK symbols as follows:

$$\mathbb{E}\{s_{kl_1} \mid \mathbf{Y}_d\} = \tanh(L_{kl_1}), \tag{37}$$

$$\mathrm{Var}(s_{kl_1} \mid \mathbf{Y}_d) = 1 - \tanh^2(L_{kl_1}), \tag{38}$$

where L_{kl_1} denotes the estimated LLR of the l_1-th bit of user k, and $\mathbf{Y}_d$ the received signal in the LDPC decoding phase. These posterior statistics are subsequently used to update the extrinsic variables according to (15) and (16). To ensure numerical stability and efficient training, parameter sharing is applied across all internal LDPC sub-networks, yielding a compact and robust quantization-aware decoder.

Consequently, we arrive at a unified, quantization-aware receiver that jointly detects active devices, estimates their channels, and decodes their data in CF-mMIMO systems. At very low resolutions, such as 1–2 bits, or when the quantizer frequently operates in the saturation region, the quantization bins become too coarse, and the extrinsic Gaussian approximation may become less accurate. In such cases, the posterior information passed between the modules becomes less informative, which may slow convergence or degrade reconstruction accuracy. This behavior is consistent with the observed performance degradation under aggressive low-bit quantization.

4 Simulation results

In this section, we evaluate the performance of the generalized VAMP-Net and generalized MIMO-LDPC-Net under coarse quantization to validate the effectiveness of our proposed quantization-aware Bayesian inference framework. We first detail the key system parameters and training setup used for the numerical experiments. Subsequently, we evaluate the performance of the developed generalized VAMP-Net and generalized MIMO-LDPC-Net across various quantization resolutions, contrasting their robustness and inference accuracy with algorithms based on SLM. The simulations follow the considered mmWave CF-mMIMO setting, where the received signals from multiple distributed APs are processed in the angular-domain representation. This setting is consistent with the sparse propagation characteristics of mmWave channels and highlights the relevance of the proposed sparse Bayesian receiver under hybrid analog-digital architectures and quantized AP-to-CPU links.

4.1 Simulation setup

For fair comparison, we benchmark against their SLM-based counterparts (VAMP-Net and MIMO-LDPC-Net) [18]. We choose VAMP-Net and MIMO-LDPC-Net as the main numerical baselines because they are the closest model-driven unfolded receivers for the considered activity/channel-estimation and LDPC-decoding tasks. Conventional GLM/GAMP, quantized AMP, and one-bit compressed-sensing algorithms are highly relevant to quantized sparse recovery, but they are mainly designed for activity detection or generic sparse reconstruction, rather than the complete URA receiver chain with activity detection, channel estimation, and LDPC-based data decoding under quantized AP-to-CPU observations. Therefore, the comparison with SLM-based VAMP-Net and MIMO-LDPC-Net isolates the gain brought by replacing the SLM assumption with the proposed quantization-aware GLM formulation under the same unfolded receiver architecture. Both networks use a single internal (VAMP or LDPC) layer and 20 external iterations. This compact setting is adopted for both the SLM-based and GLM-based receivers to ensure a fair comparison under the same unfolded depth and to avoid attributing the performance gain to a larger network size. The quantization dynamic range is fixed for all compared methods under the same bit resolution, so that the observed performance difference mainly reflects the effect of quantization-aware inference rather than quantizer-range tuning. The key parameters for simulation are summarized in Tables 2 and 3.

Dataset generation. For each Monte-Carlo sample, $K_a = 100$ active devices are uniformly selected from the $K = 2000$ potential devices. The columns of the pilot matrix are generated as independent complex Gaussian sequences and then normalized to unit power. The sparse activity/channel matrix is generated according to the activity pattern, where the rows corresponding to inactive devices are set to zero, and the nonzero channel coefficients

Table 2 Parameters for simulation.

Parameter	Value
Total number of devices K	2000
Number of active devices K_a	100
Length of the pilot for active detection L_p	110
Length of the pilot for channel estimation L_c	80
Length of the pilot for data decoding L_d	256
Number of antennas per AP M_{an}	32
Number of RFs per AP N_{rf}	8
Number of APs M_{ap}	16

Table 3 Training hyperparameters for the neural receivers.

Parameter	Value
Optimizer	Adam
Initial learning rate for generalized VAMP-Net	1×10^{-3}
Initial learning rate for generalized MIMO-LDPC-Net	5×10^{-4}
Batch size	32
Maximum number of epochs	100
Early-stopping patience	10 epochs
Learning-rate decay	Factor 0.5 after 5 stagnant validation epochs
Weight decay	1×10^{-5}
Gradient clipping	ℓ_2-norm clipped to 5
Training samples	20000
Validation samples	2000
Testing samples	10000

of active devices are generated following the mmWave angular-domain sparse channel model. AWGN is added according to the target SNR, and the real and imaginary parts of the received baseband samples are separately quantized by the b-bit scalar quantizer in (2)–(4). The resulting quantized observations and their corresponding ground-truth sparse matrices are used to train the generalized VAMP-Net. For the generalized MIMO-LDPC-Net, the transmitted information bits are used as labels, and the network is trained using the bit-wise binary cross-entropy loss computed from the output LLRs. The SLM-based and GLM-based networks are trained and tested under the same activity, channel, SNR, and quantization settings for fair comparison.

The best model checkpoint is selected according to the validation loss. The generalized VAMP-Net is trained using the NMSE loss in the activity-detection/channel-estimation stage, while the generalized MIMO-LDPC-Net is trained using the bit-wise binary cross-entropy loss in the decoding stage. For a fair comparison, the proposed GLM-based networks and their SLM-based counterparts use the same training, validation, and testing datasets, as well as the same number of internal layers and external iterations.

4.2 Performance analysis for active detection and channel estimation phase

In this subsection, we evaluate the performance of the proposed generalized VAMP-Net specifically for the active detection and channel estimation phase, contrasting its robustness against the quantization effects with the conventional SLM-based VAMP-Net. We adopt error probability P_e and normalized mean square error (NMSE) as performance evaluation metrics. The NMSE is given by

$$\text{NMSE} = 10\log_{10}\left(\frac{\|\hat{\mathbf{X}} - \mathbf{X}\|_{\text{F}}^2}{\|\mathbf{X}\|_{\text{F}}^2}\right). \tag{39}$$

To further distinguish the effects of quantization-aware inference and learnable unfolding, we additionally include a non-learned generalized VAMP baseline using the analytical Module-B updates in Algorithm 2. This baseline and the generalized VAMP-Net share the same quantization-aware GLM denoiser in Module A, while Module B uses analytical VAMP updates in the former and learned unfolded updates in the latter. All three receivers are evaluated under the same system parameters, quantization settings, and iteration numbers.

Figure 5 plots the error probability P_e versus SNR for different quantization resolutions. The comparison with the non-learned generalized VAMP clarifies the contributions of the two components. Under coarse quantization, its

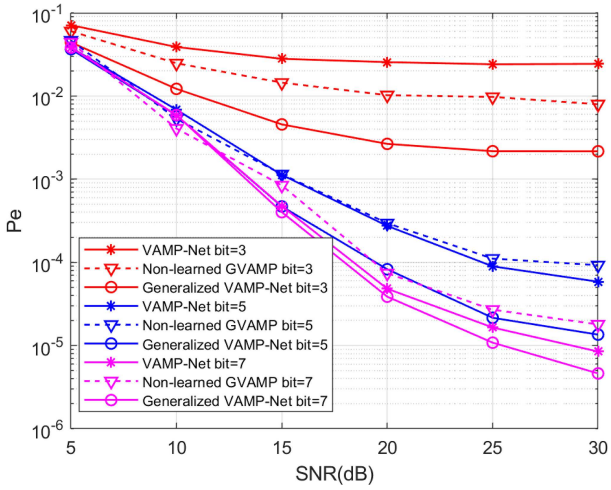

Figure 5 (Color online) Error probability P_e versus SNR for various quantization bit resolutions.

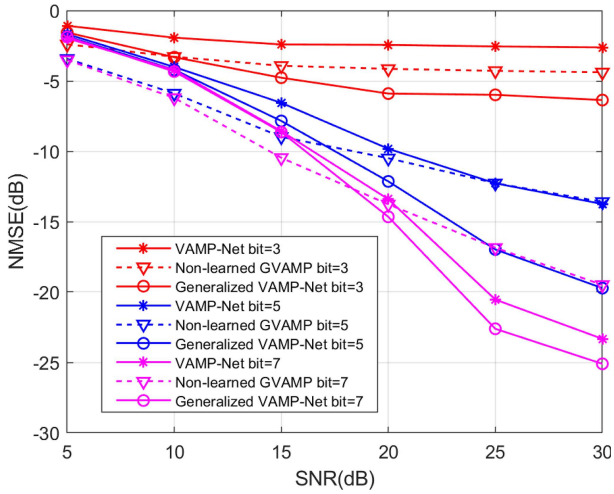

Figure 6 (Color online) NMSE versus SNR for various quantization bit-depths.

improvement over the SLM-based VAMP-Net demonstrates the benefit of explicitly incorporating the quantization likelihood through the GLM denoiser. The generalized VAMP-Net further reduces P_e, showing the additional gain provided by learnable unfolding. At higher quantization resolutions, the non-learned generalized VAMP does not uniformly outperform the SLM-based network, whereas the generalized VAMP-Net maintains the best high-SNR performance. This result indicates that the overall improvement arises from the combination of quantization-aware inference and learned network parameters.

Figure 6 shows the NMSE as a function of SNR. The NMSE results lead to the same conclusion as the error-probability comparison. Under coarse quantization, the gain of the non-learned generalized VAMP over the SLM-based VAMP-Net verifies the contribution of the GLM denoiser. The additional performance gap between the non-learned generalized VAMP and generalized VAMP-Net reflects the benefit of learning the unfolded Module-B parameters. Although the analytical baseline is not uniformly superior to the SLM-based network at higher quantization resolutions, the generalized VAMP-Net consistently achieves the lowest NMSE, confirming that quantization-aware likelihood modeling and learnable unfolding play complementary roles.

4.3 Performance analysis for LDPC decoding phase

Figure 7 compares the performance of MIMO-LDPC-Net and generalized MIMO-LDPC-Net, assuming perfect channel knowledge to isolate the decoding stage. To avoid ambiguity, we use $p_{\mathrm{md}} + p_{\mathrm{fa}}$, i.e., the sum of the miss-detection probability and the false-alarm probability, as the performance metric for multi-user message recovery in the LDPC decoding phase. Both methods improve monotonically with SNR, but the proposed generalized version maintains the lowest $p_{\mathrm{md}} + p_{\mathrm{fa}}$ values throughout the whole SNR range, demonstrating that explicit quantization modeling is equally beneficial in the decoder. At 3 bits per sample, MIMO-LDPC-Net nearly fails across the SNR range, whereas generalized MIMO-LDPC-Net still shows performance gains, which should be owing to its MMSE denoiser mitigating the quantization noise.

4.4 Performance for joint active user detection, channel estimation and LDPC decoding

Finally, we provide an end-to-end performance evaluation of the complete unified quantization-aware receiver, which jointly executes activity detection, channel estimation, and data decoding. Figure 8 shows the end-to-end error probability for joint activity detection, channel estimation, and data decoding.

Both SLM- and GLM-based receivers improve monotonically with SNR, and the generalized framework consistently yields the best performance. With only 3-bit quantization, neither receiver can decode reliably in the considered end-to-end URA setting, suggesting that the information loss caused by such aggressive quantization is too severe under the current system load and pilot/codeword lengths. This result does not contradict the benefit of quantization-aware inference; instead, it identifies the lower boundary of the feasible operating regime for the complete joint detection, channel-estimation, and decoding chain. At 5-bit resolution, the GLM-based framework exhibits a substantial performance improvement over the SLM-based receiver, while the AP-to-CPU signaling load

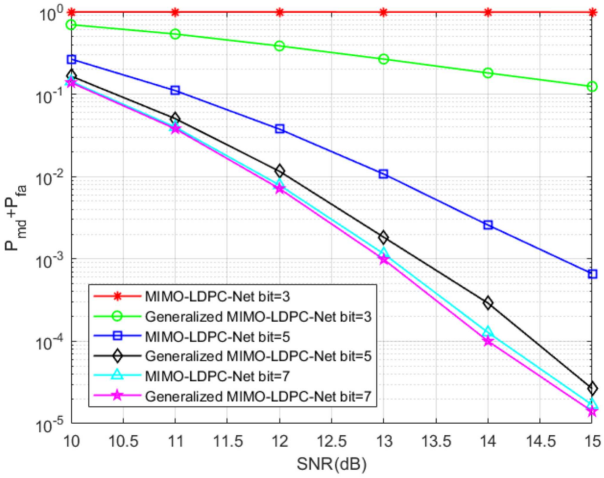

Figure 7 (Color online) Sum of miss-detection and false-alarm probabilities, $p_{\mathrm{md}} + p_{\mathrm{fa}}$, versus SNR in the LDPC decoding phase under various quantization bit resolutions.

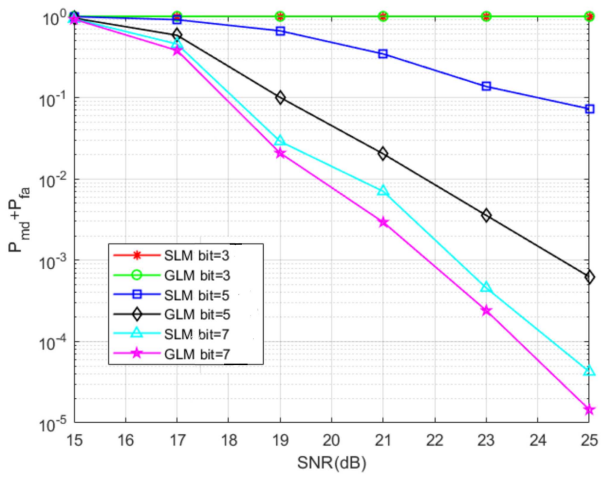

Figure 8 (Color online) End-to-end error probability (joint detection, channel estimation, and decoding) versus SNR.

remains much lower than that of high-resolution forwarding. When the resolution further increases, the performance gap gradually diminishes as the quantization effect becomes negligible. Therefore, under the considered setup, a practical operating regime is to use a resolution higher than 3 bits, with 5 bits per real dimension providing a favorable trade-off between end-to-end reliability and AP-to-CPU signaling load.

5 Conclusion

In this paper, we have proposed a unified quantization-aware receiver framework for joint activity detection, channel estimation, and data decoding in backhaul-limited CF-mMIMO systems. By formulating the quantized received signals at APs as GLM, our framework can unify the complex tasks under a consistent Bayesian inference perspective. Based on the iterative inference framework, we have successfully constructed the generalized VAMP-Net and the generalized MIMO-LDPC-Net, resulting in a robust and quantization-aware receiver with moderate additional complexity compared with SLM-based unfolded receivers. Numerical results have strongly validated the effectiveness of our proposed framework. Specifically, while SLM-based receivers suffer severe performance degradation under quantization, our generalized architecture maintains stable and continuous performance improvement due to the suppression of quantization noise by the Bayesian inference framework.

Acknowledgements This work was supported by Mobile Information Networks-National Science and Technology Major Project (Grant No. 2025ZD1305100) and National Natural Science Foundation of China (Grant No. 62671146).

References

1 You X, Wang C X, Huang J, et al. Towards 6G wireless communication networks: vision, enabling technologies, and new paradigm shifts. Sci China Inf Sci, 2021, 64: 110301
2 Ngo H Q, Ashikhmin A, Yang H, et al. Cell-free massive MIMO versus small cells. IEEE Trans Wirel Commun, 2017, 16: 1834–1850
3 Björnson E, Sanguinetti L. Making cell-free massive MIMO competitive with MMSE processing and centralized implementation. IEEE Trans Wirel Commun, 2020, 19: 77–90
4 Zeng F, Liu R, Sun X, et al. Multi-static ISAC based on network-assisted full-duplex cell-free networks: performance analysis and duplex mode optimization. Sci China Inf Sci, 2025, 68: 150303
5 You X H, Huang Y M, Zhang C, et al. When AI meets sustainable 6G. Sci China Inf Sci, 2025, 68: 110301
6 Feng C, Shen W, An J, et al. Weighted sum rate maximization of the mmWave cell-free MIMO downlink relying on hybrid precoding. IEEE Trans Wirel Commun, 2022, 21: 2547–2560
7 Li H, Jiang Y, Huang Y, et al. A multi-agent DRL method for distributed energy-efficient association and hybrid precoding in mmWave cell-free massive MIMO systems. IEEE Commun Lett, 2025, 29: 70–74
8 Chen X, Ng D W K, Yu W, et al. Massive access for 5G and beyond. IEEE J Sel Areas Commun, 2021, 39: 615–637
9 Ke M, Gao Z, Wu Y, et al. Massive access in cell-free massive MIMO-based Internet of Things: cloud computing and edge computing paradigms. IEEE J Sel Areas Commun, 2021, 39: 756–772
10 Johnston J, Wang X. Model-based deep learning for joint activity detection and channel estimation in massive and sporadic connectivity. IEEE Trans Wirel Commun, 2022, 21: 9806–9817
11 Cui Y, Li S, Zhang W. Jointly sparse signal recovery and support recovery via deep learning with applications in MIMO-Based grant-free random access. IEEE J Sel Areas Commun, 2021, 39: 788–803
12 Ma Z, Wu W, Gao F, et al. Model-driven deep learning for non-coherent massive machine-type communications. IEEE Trans Wirel Commun, 2024, 23: 2197–2211
13 Hu Q, Cai Y, Shi Q, et al. Iterative algorithm induced deep-unfolding neural networks: precoding design for multiuser MIMO systems. IEEE Trans Wirel Commun, 2021, 20: 1394–1410

14 Cui Q M, You X H, Wei N, et al. Overview of AI and communication for 6G network: fundamentals, challenges, and future research opportunities. Sci China Inf Sci, 2025, 68: 171301
15 Polyanskiy Y. A perspective on massive random-access. In: Proceedings of IEEE International Symposium on Information Theory (ISIT), Aachen, 2017. 2523–2527
16 Rangan S, Schniter P, Fletcher A K. Vector approximate message passing. IEEE Trans Inform Theor, 2019, 65: 6664–6684
17 Iimori H, Takahashi T, Ishibashi K, et al. Grant-free access via bilinear inference for cell-free MIMO with low-coherence pilots. IEEE Trans Wirel Commun, 2021, 20: 7694–7710
18 Li T, Jiang Y, Huang Y, et al. Model-based deep learning for massive access in mmWave cell-free massive MIMO system. In: Proceedings of IEEE International Conference on Communications Workshops (ICC Workshops), Denver, 2024. 828–833
19 Björnson E, Sanguinetti L. Scalable cell-free massive MIMO systems. IEEE Trans Commun, 2020, 68: 4247–4261
20 Ngo H Q, Tran L N, Duong T Q, et al. On the total energy efficiency of cell-free massive MIMO. IEEE Trans Green Commun Netw, 2018, 2: 25–39
21 Kim I, Bennis M, Choi J. Cell-free mmWave massive MIMO systems with low-capacity fronthaul links and low-resolution ADC/DACs. IEEE Trans Veh Technol, 2022, 71: 10512–10526
22 Tao Y, Jiang Y, Zheng F C, et al. Content popularity prediction based on quantized federated Bayesian learning in fog radio access networks. IEEE Trans Commun, 2023, 71: 893–907
23 Zhu W, Zhang S, Liu L. Joint transmission and compression design for 6G networked sensing with limited-capacity backhaul. In: Proceedings of IEEE Global Communications Conference (GLOBECOM), Cape Town, 2024. 187–192
24 Noh Y, Hong S. Compressed sensing based active user detection in MIMO systems with one-bit ADC. IEEE Trans Veh Technol, 2023, 72: 1313–1317
25 Rangan S. Generalized approximate message passing for estimation with random linear mixing. In: Proceedings of IEEE International Symposium on Information Theory (ISIT), St. Petersburg, 2011. 2168–2172
26 Meng X, Wu S, Zhu J. A unified Bayesian inference framework for generalized linear models. IEEE Signal Proc Let, 2018, 25: 398–402
27 Zhao Z, Yang Z, Chen M, et al. Energy-efficient probabilistic semantic communication over space-air-ground integrated networks. IEEE Trans Wirel Commun, 2025, 24: 8814–8829
28 Zhao Z, Yang Z, Hu Y, et al. Compression ratio allocation for probabilistic semantic communication with RSMA. IEEE Trans Commun, 2025, 73: 7304–7318
29 Guo J C, Yu Q Y, Sun W B, et al. Robust efficient hybrid pre-coding scheme for mmWave cell-free and user-centric massive MIMO communications. IEEE Trans Wirel Commun, 2021, 20: 8006–8022
30 Pradhan A K, Amalladinne V K, Vem A, et al. Sparse IDMA: a joint graph-based coding scheme for unsourced random access. IEEE Trans Commun, 2022, 70: 7124–7133
31 Li T, Wu Y, Zheng M, et al. Joint device detection, channel estimation, and data decoding with collision resolution for MIMO massive unsourced random access. IEEE J Sel Areas Commun, 2022, 40: 1535–1555