

Efficient measurement error mitigation with subsystem-balanced Pauli twirling

Xiao-Yue Xu, Chen Ding, and Wan-Su Bao*

Henan Key Laboratory of Quantum Information and Cryptography, Zhengzhou, Henan 450000, China



(Received 21 September 2025; accepted 7 January 2026; published 17 February 2026)

Measurement error mitigation (MEM) is essential for realizing reliable quantum computation. Model-free measurement error mitigation (MF-MEM) is an important class of MEM methods that employs Pauli twirling—typically with a random twirling set—to convert measurement noise into a state-independent scaling factor, thereby enabling error mitigation through simple calibration. However, such methods face prohibitive sampling overhead, limiting their scalability. To address this, we introduce subsystem-balanced Pauli twirling (SB-PT), a twirling method designed for MF-MEM that enforces Pauli operator balance on measuring subsystems to selectively suppress dominant measurement noise. Theoretically, for a weight- τ Pauli observable, SB-PT removes all independent error components using only $\mathcal{O}(4^\tau)$ random circuits, substantially reducing the sampling overhead over conventional Pauli twirling. This efficiency gain is most significant for sparse observables. To extend such resource-efficient mitigation to arbitrary observables, we generalize and integrate a hardware-efficient measurement transformation framework that converts high-weight Pauli operators into low-weight effective ones via linear-depth circuits. The circuit noise introduced during this transformation is jointly mitigated with native measurement noise using a unified twirling protocol, ensuring robust performance. Extensive numerical simulations demonstrate a greater than 16-fold improvement in sampling efficiency over conventional random twirling, with consistent performance gains across varying system sizes and error regimes. This work provides a resource-frugal and experimentally viable path toward high-fidelity measurement in near-term quantum devices.

DOI: [10.1103/z7lh-nlg3](https://doi.org/10.1103/z7lh-nlg3)

I. INTRODUCTION

Quantum computers hold significant potential for solving classically intractable problems across a wide range of fields, including quantum chemistry, optimization, and machine learning [1–3], yet are fundamentally limited by hardware imperfections. Among these, measurement errors pose a particularly critical challenge: They systematically corrupt the final readout after quantum state collapse, irreversibly destroying computational results. Crucially, the dimensionality of measurement noise scales exponentially with qubit number, requiring $\mathcal{O}(4^n)$ parameters for full characterization, rapidly becoming infeasible for practical devices.

This exponential barrier motivates simplified error models. The most direct approach leverages measurement's projective nature, treating readout errors as classical state transitions after state collapse [4–8]. We model them through a stochastic map with transfer matrix Λ relating ideal (p_{ideal}) and noisy (p_{noisy}) distributions via

$$p_{\text{ideal}} = \Lambda^{-1} p_{\text{noisy}}. \quad (1)$$

Quantum mechanically, this corresponds to a quantum channel $\mathcal{N} = \mathcal{E}_\Lambda \circ \mathcal{D}$, where the complete dephasing channel

$\mathcal{D}(\rho) = \sum_k \langle k | \rho | k \rangle |k\rangle \langle k|$ eliminates coherence, followed by the population noise channel $\mathcal{E}_\Lambda(\sigma) = \sum_i (\sum_j [\Lambda]_{ij} \sigma_j) |i\rangle \langle i|$ for the resulting diagonal state $\sigma = \sum_j \sigma_j |j\rangle \langle j|$ with $\sigma_j = \langle j | \sigma | j \rangle$, which modifies the diagonal elements according to the transfer matrix Λ . However, even this classical reduction remains exponentially costly as it requires $\mathcal{O}(2^n)$ parameters, motivating further simplification. The tensor Product Noise (TPN) model [5,9–11] assumes qubit-independent measurement noise, which implies a transfer matrix that factorizes as $\Lambda_{\text{TPN}} = \bigotimes_{i=1}^n \Lambda_i$. This reduces the complexity to $\mathcal{O}(n)$, making mitigation tractable for large systems, though at the cost of neglecting correlated and quantum coherent errors.

An alternative paradigm is model-free measurement error mitigation (MF-MEM) [12–16], which circumvents model-dependent limitations. This method employs Pauli twirling [17–20] to tailor the full measurement noise channel into a Pauli form, enabling its mitigation via state-independent rescaling without requiring explicit characterization. While effective for moderate-scale systems [21], the conventional approach using undirected random twirling becomes prohibitively resource intensive at scale.

Motivated by tailored twirling techniques that construct minimal twirling sets for specific error channels [22], we introduce Subsystem-Balanced Pauli twirling (SB-PT) for MF-MEM. SB-PT systematically exploits the structure of measurement noise by designing twirling sets in which each Pauli operator is balanced on the measured subsystems, thereby selectively suppressing the dominant independent error components. We rigorously prove that for any Pauli observable of weight $\tau < n$, SB-PT requires only $\mathcal{O}(4^\tau)$ random circuits to exactly cancel all independent measurement

*Contact author: bws@qiclab.cn

Published by the American Physical Society under the terms of the Creative Commons Attribution 4.0 International license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

errors, while also delivering strictly tighter sampling overhead bounds than conventional random twirling. To extend these advantages beyond sparse observables, we integrate a measurement transformation protocol that generalizes compression readout [23,24]; this protocol converts any m -weight Pauli observable into k -weight form ($k \leq m$) via $\mathcal{O}(n)$ -depth circuits while maintaining compatibility with twirling. This framework incorporates both measurement noise and circuit-induced noise into a unified effective noise model, enabling unified error mitigation. Numerical simulations confirm these advantages across multiple noise regimes and observable types, demonstrating significantly improved resource efficiency over MF-MEM with random twirling while maintaining comparable fidelity.

The remainder of this paper is organized as follows. Section II reviews the established theoretical foundation, introducing the Pauli-transfer-matrix representation, Pauli twirling, and the MF-MEM framework. This section reformulates these known concepts into a consistent notation that prepares the essential formalism for our protocols developed in later sections. In Sec. III, we develop the SB-PT protocol: We first characterize the algebraic structure of classical measurement noise within the Pauli-transfer-matrix representation, then construct the SB-PT method by exploiting the inherent symmetries. Section IV addresses mitigation scalability through measurement transformation techniques, enabling efficient SB-PT implementation for high-weight observables. Comprehensive numerical simulations validating our protocol under various noise regimes are presented in Sec. V. Finally, we conclude with discussions on broader implications and potential extensions.

II. TWIRLING-BASED MEASUREMENT ERROR MITIGATION

A. The Pauli-transfer-matrix formalism

Measurement noise systematically corrupts the expectation values of Pauli observables by introducing non-physical shifts. To establish this formally, we adopt the Pauli-transfer-matrix (PTM) representation [25,26]. Within this formalism, the quantum state ρ is encoded as the column vector $|\rho\rangle$, while the quantum channel $\mathcal{E}$ —completely positive trace-preserving (CPTP) [1] map—corresponds to a real matrix $\mathcal{M}_{\mathcal{E}}$ acting on the vector. The observable O maps to row vector $\langle\langle O|$ in the same basis. The expectation value $\langle O \rangle$ of an observable O , given that the state ρ has evolved under the channel $\mathcal{E}$, is given by the Hilbert-Schmidt inner product:

$$\langle O \rangle = \text{Tr}(O\mathcal{E}(\rho)) = \langle\langle O|\mathcal{M}_{\mathcal{E}}|\rho\rangle\rangle. \quad (2)$$

We order the n -qubit Pauli group $\{I, X, Y, Z\}^{\otimes n}$ by assigning to each index $p \in \{0, 1, \dots, 4^n - 1\}$ the unique Pauli operator $P_p = \bigotimes_{i=1}^n \sigma_{\mathbf{p}_i}$, where $\sigma_0 = I, \sigma_1 = X, \sigma_2 = Y, \sigma_3 = Z$, and $\mathbf{p} = (\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_n) \in \mathbb{Z}_4^n$ is the base-4 representation of the number p , where each $\mathbf{p}_i$ is a digit of p . Likewise, the subgroup $\{I, Z\}^{\otimes n}$ is ordered by an analogous binary map that assigns to each index $r \in \{0, 1, \dots, 2^n - 1\}$ the operator $Z_r = \bigotimes_{i=1}^n \sigma_{\mathbf{r}_i}$, where $\mathbf{r} \in \mathbb{Z}_2^n$ and $\sigma_0 = I, \sigma_1 = Z$. To embed the subgroup $\{I, Z\}^{\otimes n}$ into the full n -qubit Pauli group, we

introduce the index-preserving maps:

$$\begin{aligned} \phi : \{0, 1, \dots, 2^n - 1\} &\rightarrow \{0, 1, \dots, 4^n - 1\}, \\ \tilde{\phi} : \mathbb{Z}_2^n &\rightarrow \mathbb{Z}_4^n, \end{aligned} \quad (3)$$

where we flip every 1 in $\mathbb{Z}_2^n$ to 3 in $\mathbb{Z}_4^n$. Formally,

$$\begin{aligned} \tilde{\phi}(\mathbf{r}) &= (3\mathbf{r}_1, \dots, 3\mathbf{r}_n) \in \mathbb{Z}_4^n, \\ \phi(r) &= \sum_{i=1}^n 3\mathbf{r}_i 4^{i-1}. \end{aligned} \quad (4)$$

This construction ensures that $Z_r = P_{\phi(r)}$ for all r .

The matrix element of $\mathcal{M}_{\mathcal{E}}$ is defined for indices $p, q \in \{0, 1, \dots, 4^n - 1\}$ as

$$[\mathcal{M}_{\mathcal{E}}]_{pq} = \frac{1}{2^n} \text{Tr}(P_p \mathcal{E}(P_q)). \quad (5)$$

In particular, any Pauli observable P_q can be represented by the standard-basis row vector $\langle\langle P_q| = \mathbf{e}_q^T = (0, \dots, 1, \dots, 0)^T$ in the PTM representation, where $\mathbf{e}_q$ has a single nonzero entry (1 at index q corresponding to the position of P_q in the Pauli basis).

B. Model-free measurement error mitigation via Pauli twirling

We now formalize the impact of measurement noise on expectation-value estimation for Pauli observables and introduce the foundational principles of MF-MEM. Without loss of generality, we describe our protocol using Pauli-Z observables, as single-qubit Cliffords can map any Pauli observable to this form. For general measurement noise $\mathcal{R}$ within the PTM representation, the noisy expectation value for Pauli-Z observable Z_r is

$$\langle\tilde{Z}_r\rangle_{\rho} = \langle\langle Z_r|\mathcal{R}|\rho\rangle\rangle = 2^n \sum_{j=0}^{4^n-1} [\mathcal{R}]_{\phi(r)j} \varrho_j(\rho), \quad (6)$$

where $\varrho_j(\rho)$ is the j th element of $|\rho\rangle$, which equals the scaled noiseless expectation value $\langle\langle P_j|\rho\rangle\rangle = \frac{1}{2^n} \text{Tr}(\rho P_j)$. The measurement noise $\mathcal{R}$ results in additional systematic bias of the ideal value $\varrho_{\phi(r)}(\rho)$.

MF-MEM methods exploits Pauli twirling to convert the measurement noise channel $\mathcal{R}$ into a stochastic Pauli channel, enabling bias correction. For a randomly chosen twirling set $\mathcal{S}$ (a subset of the n -qubit Pauli group with $|\mathcal{S}| \leq 4^n$), the twirled noise channel [17–20] is

$$\mathcal{R}^{\text{twirled}} = \frac{1}{|\mathcal{S}|} \sum_{P_q \in \mathcal{S}} \mathcal{P}_q \mathcal{R} \mathcal{P}_q, \quad (7)$$

where $\mathcal{P}_q$ is the PTM of Pauli operator P_q , a diagonal matrix with element $[\mathcal{P}_q]_{ii} = \eta(P_q, P_i)$ for $i \in \{0, 1, \dots, 4^n - 1\}$. Here, $\eta(P_q, P_i) = \pm 1$ denotes the commutator sign satisfying $P_q P_i P_q^\dagger = \eta(P_q, P_i) P_i$. The latter $\mathcal{P}_q$ is virtually applied through postprocessing of the measurement outcomes. Specifically, when the Pauli term on a single qubit is X or Y , this virtual application is achieved by bit-flipping the measurement outcomes (0 and 1).

The PTM elements of the twirled channel simplify to

$$[\mathcal{R}^{\text{twirled}}]_{ij} = \frac{1}{|\mathcal{S}|} \sum_{P_q \in \mathcal{S}} \eta(P_q, P_i) \eta(P_q, P_j) [\mathcal{R}]_{ij}. \quad (8)$$

Defining the scaling factor,

$$\alpha_{ij}(\mathcal{S}) = \frac{1}{|\mathcal{S}|} \sum_{P_q \in \mathcal{S}} \eta(P_q, P_i P_j), \quad (9)$$

and noting that $\eta(P_q, P_i) \eta(P_q, P_j) = \eta(P_q, P_i P_j)$ holds under the standard Pauli multiplication convention, we obtain

$$[\mathcal{R}^{\text{twirled}}]_{ij} = \alpha_{ij}(\mathcal{S}) [\mathcal{R}]_{ij}. \quad (10)$$

Crucially, the diagonal elements remain unchanged since

$$\alpha_{ii}(\mathcal{S}) = \frac{1}{|\mathcal{S}|} \sum_{P_q \in \mathcal{S}} \eta(P_q, P_i P_i) = \frac{1}{|\mathcal{S}|} \sum_{P_q \in \mathcal{S}} \eta(P_q, I) = 1, \quad (11)$$

preserving the intrinsic noise characteristics, while off-diagonal elements are scaled by $\alpha_{ij}(\mathcal{S})$, with the scaling magnitude determined by $\mathcal{S}$.

When employing the full twirling set $\mathcal{S} = \{I, X, Y, Z\}^{\otimes n}$, the off-diagonal elements vanish ($\alpha_{ij}(\mathcal{S}) = 0$ for $i \neq j$), resulting in a diagonal matrix $\mathcal{R}^{\text{twirled}}$. Consequently, the twirled noisy expectation value for Z_r ,

$$v(\rho) = 2^n \sum_{j=0}^{4^n-1} \alpha_{\phi(r)j}(\mathcal{S}) [\mathcal{R}]_{\phi(r)j} \mathcal{Q}_j(\rho), \quad (12)$$

then simplifies to

$$v(\rho) = 2^n [\mathcal{R}]_{\phi(r)\phi(r)} \cdot \mathcal{Q}_j(\rho) = [\mathcal{R}]_{\phi(r)\phi(r)} \text{Tr}(\rho Z_r). \quad (13)$$

This shows that the noisy expectation value $v(\rho)$ is equal to a scalar multiple of the ideal expectation value $\langle Z_r \rangle_\rho = \text{Tr}(\rho Z_r)$ [15]. The multiplicative factor $\lambda = [\mathcal{R}]_{\phi(r)\phi(r)}$ depends solely on the measurement noise channel $\mathcal{R}$ and is independent of the state ρ . Since $\text{Tr}(\rho_0 Z_r) = 1$ with $\rho_0 = |0\rangle\langle 0|^{\otimes n}$ for any r (as $|0\rangle\langle 0|$ is a $+1$ eigenstate of Z), we obtain $\lambda = v(\rho_0)$. The mitigated expectation value is then given by

$$a(\rho) = \frac{v(\rho)}{v(\rho_0)} = \frac{\lambda \langle Z_r \rangle_\rho}{\lambda} = \langle Z_r \rangle_\rho, \quad (14)$$

which exactly recovers the ideal expectation value.

However, when a reduced twirling set $\mathcal{S}$ (with $|\mathcal{S}| < 4^n$) is employed, the twirling becomes imperfect. This results in nonvanishing off-diagonal elements, meaning $\alpha_{\phi(r)j}(\mathcal{S}) \neq 0$ for some $j \neq \phi(r)$. Consequently, certain noise terms are not fully canceled via Eq. (14), introducing error terms of the form $\alpha_{\phi(r)j}(\mathcal{S}) [\mathcal{R}]_{\phi(r)j} \mathcal{Q}_j(\rho)$ into $v(\rho)$. These errors cause the mitigated estimate $a(\rho)$ to deviate from the ideal expectation value.

As part of the theoretical foundation, we formalize the estimation error bound in Theorem 1 (see Appendix A 1 for proof), preparing the necessary tools for our subsequent analysis.

Theorem 1. For any n -qubit quantum state ρ and Pauli- Z observable Z_r , with probability at least $1 - \delta$ over the choice of random $\mathcal{S}$, the estimation error $\epsilon(\rho) = |a(\rho) - \text{Tr}(\rho Z_r)|$ is bounded by

$$\epsilon(\rho) \leq \frac{2\kappa(\mathcal{S}) \sum_{j \neq \phi(r)} |[\mathcal{R}]_{\phi(r)j}|}{|[\mathcal{R}]_{\phi(r)\phi(r)} - \kappa(\mathcal{S}) \sum_{j \neq \phi(r)} |[\mathcal{R}]_{\phi(r)j}|}, \quad (15)$$

where

$$\kappa(\mathcal{S}) = \sqrt{\frac{2}{|\mathcal{S}|} (2n \ln 2 + \ln(2/\delta))}, \quad (16)$$

under the diagonal dominance condition

$$|[\mathcal{R}]_{\phi(r)\phi(r)}| > \kappa(\mathcal{S}) \sum_{j \neq \phi(r)} |[\mathcal{R}]_{\phi(r)j}|. \quad (17)$$

This diagonal dominance condition holds under moderate noise levels, as demonstrated in our numerical simulations using real device calibration data. The above theorem provides a performance guarantee for the MF-MEM methods with a reduced twirling set.

III. SUBSYSTEM-BALANCED PAULI TWIRLING

A. Structure of classical measurement noise model

The efficacy of any tailored twirling strategy hinges on a precise understanding of the noise it aims to suppress. Unlike general quantum processes, classical measurement noise exhibits a highly constrained algebraic structure. This structure arises because the measurement process projects the quantum state into a classical outcome, confining noise effects to a specific subset of elements in the PTM. Furthermore, as evidenced by experimental characterizations of superconducting quantum processors [27–29], this noise is not unstructured; it predominantly follows a hierarchical model where local error processes dominate. This characteristic structure motivates the design of twirling sets that target and cancel these dominant errors efficiently, rather than treating the noise as a black box.

To set up our framework, we define the Pauli- Z subset: the set of index pairs where each pair corresponds to an operator restricted to $\{I, Z\}^{\otimes n}$. Formally, the Pauli- Z subset of index pairs is defined as

$$\mathbb{I}_Z := \{(\phi(r), \phi(s)) | \mathbf{r}, \mathbf{s} \in \mathbb{Z}_2^n\}. \quad (18)$$

For the PTM element $[\mathcal{R}]_{ij}$ with $i, j \in \{0, 1, \dots, 4^n - 1\}$, it holds that $[\mathcal{R}]_{ij} = 0$ when $(i, j) \notin \mathbb{I}_Z$. Then we can define the reduced matrix $\mathcal{R}_Z$ as the restriction of the full PTM $\mathcal{R}$ to this Pauli- Z subset $\mathbb{I}_Z$, as illustrated in Fig. 1(a), which forms a $2^n \times 2^n$ matrix that completely characterizes the classical measurement noise. For $\mathbf{r}, \mathbf{s} \in \mathbb{Z}_2^n$, the PTM encodes noise statistics through

$$[\mathcal{R}]_{\phi(r), \phi(s)} = [\mathcal{R}_Z]_{\mathbf{r}, \mathbf{s}} = \frac{1}{2^n} \sum_{j=0}^{2^n-1} (-1)^{\mathbf{s} \cdot \mathbf{j}} \sum_{k=0}^{2^n-1} [\Lambda]_{kj} (-1)^{\mathbf{r} \cdot \mathbf{k}}, \quad (19)$$

where the dot product $\mathbf{r} \cdot \mathbf{k} = \bigoplus_i \mathbf{r}_i \mathbf{k}_i$ is computed modulo 2. This expression decomposes physically into measurement-induced dephasing channel $\mathcal{D}$ and classical bit-flip noise

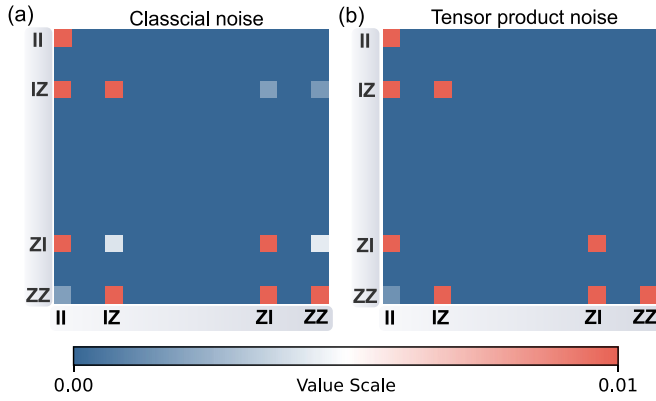


FIG. 1. Pauli Transfer Matrix structure for classical measurement noise models. To reveal the nonzero patterns more clearly, the heatmaps display the absolute values of the PTM elements. This removes visual clutter caused by the signs of the values. Explicit operator annotations on axes show the correspondence between n -qubit Pauli operators and index positions. (a) The PTM of classical measurement noise for two-qubit system. Nonzero elements are confined to the Pauli-Z subset $\mathbb{I}_Z$, with axes annotated by their corresponding Pauli-Z operators. (b) The PTM of TPN for two-qubit system. Nonzero elements satisfy the trigger condition $\text{supp}(Z_r) \subseteq \text{supp}(Z_s)$, forming the TPN subset $\mathbb{T}$. Here, we define the support of Z_r as $\text{supp}(Z_r) = \{k | \mathbf{r}_k = 1\}$. The reduced structure results from spatial measurement noise independence. Both PTMs rely on calibration data from *Zuchongzhi* 2.1 taken directly from two-qubit measurements and assembled from tensor products of single-qubit data, respectively.

channel $\mathcal{E}_\Lambda$, as introduced in Eq. (1) and the accompanying text. The term $\sum_j (-1)^{s_j}$ captures $\mathcal{D}$, representing wavefunction collapse to computational basis states during measurement. Meanwhile, $\sum_k [\Lambda]_{kj} (-1)^{\mathbf{r} \cdot \mathbf{k}}$ corresponds to $\mathcal{E}_\Lambda$, computing the expectation value of Z_r after classical readout errors governed by the transfer matrix Λ .

Especially, under the TPN model, spatial independence simplifies the noise structure significantly. The overall PTM factorizes into single-qubit components:

$$\mathcal{R}_Z = \bigotimes_{k=1}^n \mathcal{R}_Z^{(k)}, \quad \text{where} \quad \mathcal{R}_Z^{(k)} = \begin{pmatrix} 1 & 0 \\ \omega_k & \zeta_k \end{pmatrix}, \quad (20)$$

with local parameters $\omega_k = a_k + b_k - 1$ and $\zeta_k = a_k - b_k$, where $a_k = \mathbb{P}(\text{noisy } 0 | \text{ideal } 0)$ is the probability of obtaining the noisy outcome 0 given the ideal state $|0\rangle$ and $b_k = \mathbb{P}(\text{noisy } 1 | \text{ideal } 1)$ is the probability of obtaining 1 given the ideal $|1\rangle$. The constraint $|\omega_k| \leq 1 - |\zeta_k|$ ensures physical realizability.

This tensor decomposition imposes a strict *trigger condition* that determines which matrix elements can be nonzero. To characterize this condition, we define the support of a Pauli operator Z_r as

$$\text{supp}(Z_r) = \{k | \mathbf{r}_k = 1\}. \quad (21)$$

Then, an element $[\mathcal{R}_Z]_{rs}$ of $\mathcal{R}_Z$ is nonzero if and only if the support of Z_s is contained within that of Z_r

$$[\mathcal{R}_Z]_{rs} \neq 0 \iff \text{supp}(Z_s) \subseteq \text{supp}(Z_r). \quad (22)$$

Physically, this condition implies that the measurement outcome of Z_s can be affected by Z_r only if Z_r is measured on all qubits included in the measurement of Z_s .

The collection of all operator pairs satisfying this support condition defines the TPN subset:

$$\mathbb{T} = \{(r, s) | \text{supp}(Z_s) \subseteq \text{supp}(Z_r)\}, \quad (23)$$

as illustrated in Fig. 1(b).

For any $(r, s) \in \mathbb{T}$ and $r \neq 0$, the matrix elements decompose into physically interpretable factors:

$$[\mathcal{R}_Z]_{rs} = \left(\prod_{k: \mathbf{s}_k=1} \zeta_k \right) \times \left(\prod_{k: \mathbf{r}_k=1, \mathbf{s}_k=0} \omega_k \right). \quad (24)$$

Here, the first product quantifies bit-flip suppression on qubits measured by both operators, while the second captures the noise contribution from qubits measured by Z_r but not by Z_s (see Appendix B for a detailed derivation).

B. Subsystem-balanced Pauli twirling tailored for TPN subspace

Building upon the TPN subset $\mathbb{T}$ defined above, we now present a decomposition of the noise PTM $\mathcal{R}_Z$ that separates its elements based on membership in $\mathbb{T}$:

$$\mathcal{R}_Z = \mathcal{R}_Z^{\mathbb{T}} + \mathcal{R}_Z^{\overline{\mathbb{T}}}, \quad (25)$$

where the two components are defined as follows:

(1) $\mathcal{R}_Z^{\mathbb{T}}$ consists of all elements whose indices (r, s) belong to $\mathbb{T}$, i.e., $[\mathcal{R}_Z^{\mathbb{T}}]_{rs} = [\mathcal{R}_Z]_{rs}$ if $(r, s) \in \mathbb{T}$, and 0 otherwise.

(2) $\mathcal{R}_Z^{\overline{\mathbb{T}}}$ contains all elements whose indices are in the complement set $\overline{\mathbb{T}}$, i.e., $[\mathcal{R}_Z^{\overline{\mathbb{T}}}]_{rs} = [\mathcal{R}_Z]_{rs}$ if $(r, s) \notin \mathbb{T}$, and 0 otherwise.

This decomposition isolates $\mathcal{R}_Z^{\mathbb{T}}$, which captures all local independent errors conforming to the TPN model, as well as a subset of correlated errors consistent with the TPN support structure. The remaining term, $\mathcal{R}_Z^{\overline{\mathbb{T}}}$, contains only correlated errors not captured by the TPN model and is typically weak in magnitude.

To specifically target and eliminate the dominant TPN-related errors captured in $\mathcal{R}_Z^{\mathbb{T}}$, we propose SB-PT. This method constructs tailored twirling sets that are balanced across measured subsystems, enabling efficient mitigation of TPN-structured errors without requiring prior noise characterization. By exploiting the spatial locality of target observables, SB-PT significantly reduces the resource overhead compared to conventional random twirling. The overall structure of MF-MEM combined with SB-PT is illustrated in Fig. 2(a).

To realize this tailored twirling strategy, we first characterize the spatial extent of the target observable. We define the weight of a Pauli-Z operator Z_r as the number of qubits on which it acts nontrivially

$$\tau(r) \equiv |Z_r| = |\text{supp}(Z_r)| = \text{Hamm}(\mathbf{r}, \mathbf{0}), \quad (26)$$

where $\text{Hamm}(\mathbf{r}, \mathbf{0}) = \sum_{i=1}^n \mathbf{r}_i$ denotes the Hamming weight of the binary vector $\mathbf{r}$. This weight $\tau(r)$ corresponds to the number of qubits measured.

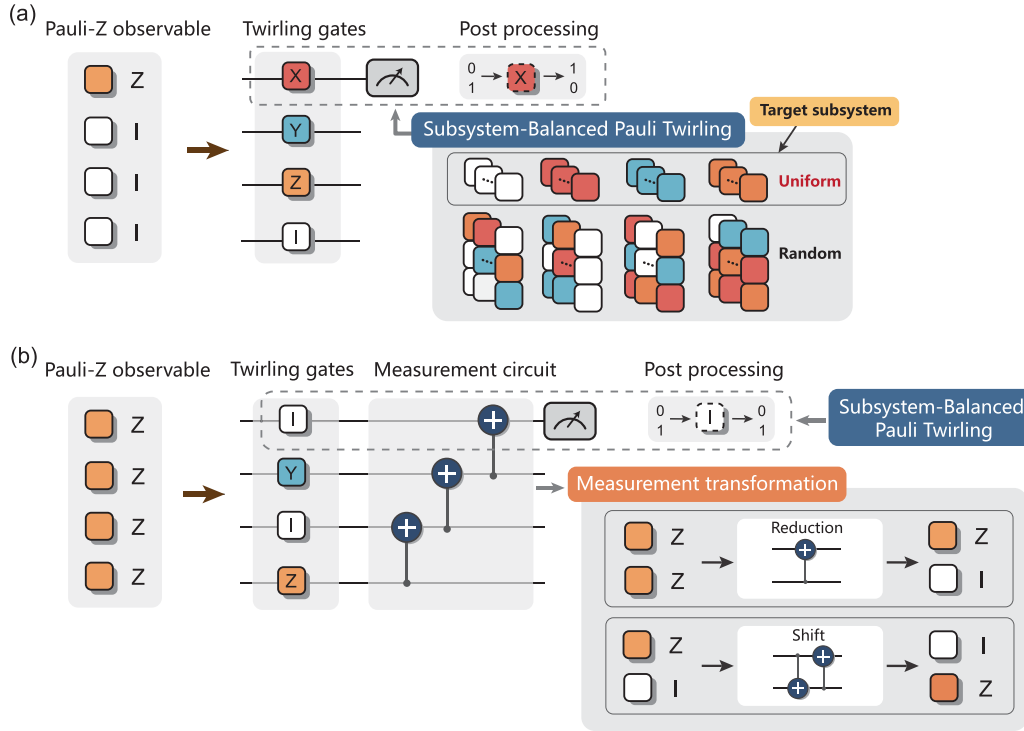


FIG. 2. The framework for MF-MEM with SB-PT. (a) For sparse Pauli observables, we directly apply SB-PT by uniformly sampling Pauli operators within the observable's support and random Pauli gates on remaining qubits. The first sequence of twirling gates is physically applied; the second is virtually implemented via post process (measurement bit-flips). (b) For high-weight observables, we employ measurement transformation to transform the observable into a sparse Pauli form. Then, we apply SB-PT to simultaneously twirl the composite noise originating from both the measurement circuit and the native measurement process. The measurement transformation technology comprises two modalities: weight reduction (converting contiguous Pauli-Z operators to sparse ones) and location shift (relocating Pauli-Z operators within the observable).

The SB-PT set $\mathcal{S}^{\text{sub}}$ is constructed via a concrete procedure that guarantees uniformly random Pauli gates on the measured qubits. Given the set size $|\mathcal{S}^{\text{sub}}| = c4^{\tau(r)}$ with $c \geq 1$, the construction proceeds as follows. For the measured subsystem of $\tau(r)$ qubits, we generate the Pauli operators by replicating the full Pauli group c times and applying a random permutation to the resulting sequence of Pauli operators. This ensures both complete coverage and genuine randomness across the measured qubits. We then assign Pauli operators to the remaining $n - \tau(r)$ qubits by independently sampling uniformly from $\{I, X, Y, Z\}$ for each qubit, thereby forming the complete n -qubit twirling operators.

Based on this construction, each resulting twirling operator $P_q \in \mathcal{S}^{\text{sub}}$ naturally exhibits the structure:

$$P_q = \left(\bigotimes_{i \in \text{supp}(Z_r)} \sigma_{q_i} \right) \otimes \left(\bigotimes_{i' \notin \text{supp}(Z_r)} \sigma_{q_{i'}} \right), \quad (27)$$

where the first tensor factor acts on the $\tau(r)$ -qubit measured subsystem, and the second on the remaining qubits. Moreover, our specific construction method ensures perfect balance: Since we include each possible Pauli string on the measured subsystem exactly c times, for any individual qubit $i \in \text{supp}(Z_r)$ and any Pauli operator $\Gamma \in \{I, X, Y, Z\}$, the marginal distribution is uniform. Specifically, among the $c4^{\tau(r)}$ operators, the count where $\sigma_{q_i} = \Gamma$ is precisely $c4^{\tau(r)-1}$, as the other $\tau(r) - 1$ qubits can form $4^{\tau(r)-1}$ independent combinations.

This directly yields the marginal uniformity condition:

$$\frac{|\{q : \sigma_{q_i} = \Gamma\}|}{|\mathcal{S}^{\text{sub}}|} = \frac{c4^{\tau(r)-1}}{c4^{\tau(r)}} = \frac{1}{4}. \quad (28)$$

For qubits outside $\text{supp}(Z_r)$, the independent random assignment ensures their individual marginal distributions are also uniform.

This full-system twirling design serves two key purposes. First, by ensuring balanced twirling on the measured subsystem, it enables the exact cancellation of dominant local noise components. Second, and crucially, by applying twirling to the complement of the observable's support, it actively suppresses correlated errors that span across the measured and unmeasured subsystems. These cross-system correlations are transformed by the random Pauli operators into locally correctable noise on the measured qubits. Specifically, the protocol ensures that all error components in $\mathcal{R}_Z^{\mathbb{T}}$ acting on $\text{supp}(Z_r)$ are exactly canceled, as formalized in Theorem 2 (see Appendix A2 for proof).

Theorem 2. For any Pauli-Z observable Z_r , we define the trigger set $\mathcal{J}_r$ as a subset of $\{0, \dots, 2^n - 1\} \setminus \{r\}$, where

$$s \in \mathcal{J}_r \Leftrightarrow (r, s) \in \mathbb{T}. \quad (29)$$

Then for any $s \in \mathcal{J}_r$, the scaling factor of SB-PT twirling set $\mathcal{S}^{\text{sub}}$ satisfies

$$\alpha_{\phi(r)\phi(s)}(\mathcal{S}^{\text{sub}}) = 0. \quad (30)$$

This induces complete cancellation of all noise components in $\mathcal{R}_Z^{\mathbb{T}}$ acting on $\text{supp}(Z_r)$.

This targeted annihilation within the TPN subset concentrates resources on suppressing the dominant noise component $\mathcal{R}_Z^{\mathbb{T}}$, substantially improving the sampling efficiency for finite-sized twirling sets. As a necessary consequence, noise components outside this subspace—specifically, elements in $\mathcal{R}_Z^{\mathbb{T}}$ and quantum coherent errors—experience weaker suppression compared to random twirling. Nevertheless, by increasing the integer scaling factor c , our method progressively suppresses these residual components, driving all error terms toward zero while maintaining balanced Pauli sampling over the subsystem.

When integrated into the MF-MEM framework, the SB-PT set $\mathcal{S}^{\text{sub}}$ yields a tighter bound on the estimation error than the general result in Theorem 1, as we establish in Theorem 3 (see Appendix A 3 for the full proof).

Theorem 3. We define the extension set as $\Phi(\mathcal{J}_r) = \{\phi(s) | s \in \mathcal{J}_r\} \subseteq \{0, \dots, 4^n - 1\}$. For any n -qubit quantum state ρ and Pauli-Z observable Z_r , with probability $\geq 1 - \delta$ over SB-PT set $\mathcal{S}^{\text{sub}}$, the estimation error $\epsilon(\rho) = |a(\rho) - \text{Tr}(\rho Z_r)|$ is bounded by

$$\epsilon_{\text{SB-PT}}(\rho) \leq \frac{2\kappa_{\text{SB}}(\mathcal{S}^{\text{sub}}) \sum_{j \notin \Phi(\mathcal{J}_r)} |[\mathcal{R}]_{\phi(r)j}|}{|[\mathcal{R}]_{\phi(r)\phi(r)}| - \kappa_{\text{SB}}(\mathcal{S}^{\text{sub}}) \sum_{j \notin \Phi(\mathcal{J}_r)} |[\mathcal{R}]_{\phi(r)j}|}, \quad (31)$$

where

$$\kappa_{\text{SB}}(\mathcal{S}^{\text{sub}}) = \sqrt{\frac{2}{|\mathcal{S}^{\text{sub}}|} (2m \ln 2 + \ln(2/\delta))}, \quad (32)$$

$$m = n - \tau(r) = n - |\text{supp}(Z_r)|,$$

under the diagonal dominance condition:

$$|[\mathcal{R}]_{\phi(r)\phi(r)}| > \kappa_{\text{SB}}(\mathcal{S}^{\text{sub}}) \sum_{j \notin \Phi(\mathcal{J}_r)} |[\mathcal{R}]_{\phi(r)j}|. \quad (33)$$

This establishes an error bound that is strictly less than the one presented in Eq. (15).

The scaling of our bound with $\tau(r)$ complicates direct comparison across observables of different weights, since the size of the twirling set $|\mathcal{S}^{\text{sub}}| = c4^{\tau(r)}$ itself grows with the observable support. Nevertheless, for any fixed $\tau(r)$ and fixed twirling set size, our bound is strictly tighter than that of random twirling (given in Theorem 1), demonstrating the practical benefit of SB-PT under finite sampling.

For high-weight observables, however, the performance gap between SB-PT and random twirling narrows, particularly when the size of the twirling set is smaller than the full Pauli group on the measured subsystem, i.e., when $|\mathcal{S}^{\text{sub}}| < 4^{\tau(r)}$. To mitigate this limitation, we next introduce measurement transformation techniques that reduce the effective weight of such observables. Although these techniques introduce additional circuit noise, they offer a favorable trade-off that preserves overall error mitigation performance, thereby extending the scope of our method to broader classes of observables.

IV. MEASUREMENT TRANSFORMATION

To extend the advantages of SB-PT from sparse observables to global ones, we employ a hardware-efficient measurement transformation (MT) framework. This framework converts arbitrary high-weight Pauli-Z observables into measurement-equivalent low-weight counterparts, enabling the subsequent application of our efficient SB-PT protocol. The integrated MT protocol leverages two complementary techniques:

(1) Weight reduction (WR), as proposed in prior work [23,24], compresses contiguous Pauli-Z operators into single ones.

(2) Location shift (LS), which we introduce here, spatially redistributes the support of Pauli-Z operators to facilitate WR.

As illustrated in Fig. 2(b), the resulting measurement circuit consists entirely of nearest-neighbor CX gates. The following sections formalize this framework and analyze its critical synergy with SB-PT, demonstrating how we mitigate the resulting composite noise and preserve low sampling complexity.

A. Weight reduction

We begin by structurally decomposing the Pauli-Z observable Z_r into its fundamental components. This involves grouping consecutive nonidentity (Z) and identity (I) operators into contiguous segments.

Formally, we define the following two sets of segments:

(1) $\Gamma_k^r = Z^{(i_k^1)} \otimes \dots \otimes Z^{(i_k^{\tau_k(r)})}$ is the set of nonidentity segments, where each Γ_k^r is a tensor product of $\tau_k(r)$ Pauli-Z terms.

(2) $\Upsilon_k^r = I^{(i_k^1)} \otimes \dots \otimes I^{(i_k^{\beta_k(r)})}$ is the set of identity segments, where each Υ_k^r is a tensor product of $\beta_k(r)$ identity terms.

The entire observable is then given by the tensor product of all these segments:

$$Z_r = \left(\bigotimes_{k=1}^{n_Z} \Gamma_k^r \right) \otimes \left(\bigotimes_{k=1}^{n_I} \Upsilon_k^r \right), \quad (34)$$

The total weight $\tau(r)$ is the sum of the lengths of all nonidentity segments: $\tau(r) = \sum_{k=1}^{n_Z} \tau_k(r)$.

To reduce the measured qubits for the observable Z_r , we construct a Clifford circuit U_R that maps Z_r to an effective observable $Z_r^{\text{eff}} = U_R Z_r U_R^\dagger$ with lower weight. This is achieved by compressing each high-weight segment Γ_k^r ($\tau_k(r) > 1$) onto a single qubit t_k chosen from its support $\{i_k^1, \dots, i_k^{\tau_k(r)}\}$.

The core component of this compression is a chain of CX gates [24], denoted as $U_C(i, j)$, which is defined as follows:

$$U_C(i, j) = \begin{cases} \text{CX}_{i,i+1} \text{CX}_{i+1,i+2} \dots \text{CX}_{j-1,j}, & \text{if } i < j, \\ \text{CX}_{(i,i-1)} \text{CX}_{(i-1,i-2)} \dots \text{CX}_{(j+1,j)}, & \text{if } i > j, \\ I^{\otimes n}, & \text{if } i = j. \end{cases} \quad (35)$$

For each segment Γ_k^r , we define the Clifford transformation as $U_R(\Gamma_k^r, t_k) = U_C(i_k^{\tau_k(r)}, t_k) U_C(i_k^1, t_k)$, which ensures

$$U_R(\Gamma_k^r, t_k) \Gamma_k^r U_R^\dagger(\Gamma_k^r, t_k) = Z^{(t_k)}, \quad (36)$$

as proven in Appendix C 1.

The full circuit U_R is then formed by concatenating the segmentwise transformations, yielding the effective observable:

$$\begin{aligned} Z_r^{\text{eff}} &= U_R Z_r U_R^\dagger \\ &= \left(\bigotimes_{\tau_k(r)=1} Z^{(i_k)} \right) \otimes \left(\bigotimes_{\tau_k(r)>1} Z^{(i_k)} \right) \otimes \left(\bigotimes_{k=1}^{n_f} \Upsilon_k^r \right), \end{aligned} \quad (37)$$

which has a reduced effective weight $\tau^{\text{eff}} = n_Z$. Measuring Z_r on state ρ is equivalent to measuring Z_r^{eff} on the state $U_R \rho U_R^\dagger$.

B. Location shift

However, when the Pauli-Z terms in Z_r are noncontiguously distributed, the direct application of WR is hindered; we therefore introduce LS to address this limitation. It reorganizes the support of Z_r into contiguous segments without altering its total weight, thereby enabling subsequent WR.

The fundamental operation of LS is to shift a Z operator to an adjacent qubit. Consider a pair of adjacent qubits $(i, i+1)$. If the original observable contains $Z^{(i)}$ on qubit i and $I^{(i+1)}$ on qubit $i+1$, we can define the shift operation

$$U_S(i, i+1) = CX_{i,i+1} CX_{i+1,i}. \quad (38)$$

The transformation of $Z^{(i)} \otimes I^{(i+1)}$ under the action of $U_S(i, i+1)$ yields

$$U_S(Z^{(i)} \otimes I^{(i+1)}) U_S^\dagger = I^{(i)} \otimes Z^{(i+1)}, \quad (39)$$

effectively shifting the Z support from qubit i to qubit $i+1$ (see Appendix C 2 for proof). The reverse shift, from $I^{(i-1)} \otimes Z^{(i)}$ to $Z^{(i-1)} \otimes I^{(i)}$, is achieved by the conjugation operation $U_S(i, i-1)$.

By sequentially applying these nearest-neighbor shift operations, spatially separated Pauli-Z operators can be consolidated into contiguous blocks. Once contiguity is established, the WR protocol can compress each block for full measurement compression.

C. Measurement circuit construction for arbitrary effective Pauli-Z observable

By integrating the two techniques above, we can construct a Clifford circuit U_{MT} for any Pauli-Z observable Z_r , transforming it into an effective low-weight observable:

$$Z_r^{\text{eff}} = U_{\text{MT}} Z_r U_{\text{MT}}^\dagger. \quad (40)$$

This allows strategic control over the number and locations of Pauli-Z operators in the measurement basis. Moreover, by redistributing the observable's support, the circuit reduces spatial proximity among measured qubits, thereby inherently suppressing correlated measurement errors that depend on physical layout.

The measurement circuit U_{MT} can be constructed to map Z_r to a single-qubit observable on a target qubit $j \in \{1, \dots, n\}$:

$$U_{\text{MT}} Z_r U_{\text{MT}}^\dagger = I^{\otimes(j-1)} \otimes Z^{(j)} \otimes I^{\otimes(n-j)}. \quad (41)$$

This construction is not unique; various sequences of LS and WR operations can achieve the same goal. A typical approach is to first apply LS to group all Pauli-Z terms into one contiguous

segment, and then apply WR to compress this segment into a single qubit.

More generally, the framework extends naturally to arbitrary effective weight $\tau^{\text{eff}} > 1$ by partitioning the qubits into k disjoint subsets $\{H_1, H_2, \dots, H_k\}$, each containing one qubit designated as the measurement site. A similar measurement circuit construction can be applied independently to each subset to achieve the desired low-weight effective observable.

To estimate the expectation value $\langle Z_r \rangle_\rho$, we prepare the state $U_{\text{MT}} \rho U_{\text{MT}}^\dagger$, measure in the computational basis, and post-process the outcomes to compute

$$\langle Z_r \rangle_\rho = \text{Tr}(\rho Z_r) = \text{Tr}((U_{\text{MT}} \rho U_{\text{MT}}^\dagger) Z_r^{\text{eff}}). \quad (42)$$

D. Noise-resilient measurement transformation with subsystem-balanced Pauli twirling

Although MT provides inherent resilience to measurement noise when $\tau^{\text{eff}} < \tau$, the physical implementation of the measurement circuit inevitably introduces additional gate errors. To characterize the effect of noise, we model the measurement circuit under the standard assumption of gate-independent, CPTP noise. Each ideal gate $\mathbb{U}_i$ is accompanied by a noise channel $\mathcal{E}_i$, yielding the noisy implementation:

$$\tilde{\mathbb{U}} = (\mathcal{E}_l \mathbb{U}_l) \circ \dots \circ (\mathcal{E}_2 \mathbb{U}_2) \circ (\mathcal{E}_1 \mathbb{U}_1). \quad (43)$$

To analyze the cumulative effect of these errors, we propagate each noise channel to the end of the measurement circuit through conjugation by the subsequent ideal gates. This transformation is achieved as follows: for any noise channel $\mathcal{E}_i$ occurring at the i th step, we conjugate it by all subsequent gates $\mathbb{U}_{i+1:l} = \mathbb{U}_l \circ \dots \circ \mathbb{U}_{i+1}$, effectively shifting it to the end of the circuit:

$$\tilde{\mathbb{U}} = (\mathcal{E}_l \circ (\mathbb{U}_l \mathcal{E}_{l-1} \mathbb{U}_l^\dagger)) \circ \dots \circ (\mathbb{U}_{2:l} \mathcal{E}_1 \mathbb{U}_{2:l}^\dagger) \circ \mathbb{U}. \quad (44)$$

This expression defines the effective noise channel:

$$\mathcal{E}^{\text{eff}} = \mathcal{E}_l \circ (\mathbb{U}_l \mathcal{E}_{l-1} \mathbb{U}_l^\dagger) \circ \dots \circ (\mathbb{U}_{2:l} \mathcal{E}_1 \mathbb{U}_{2:l}^\dagger), \quad (45)$$

which captures the combined effect of all gate errors during the measurement circuit.

In the PTM representation, the noisy measurement circuit is modeled as $\tilde{\mathcal{U}} = \mathcal{C}\mathcal{U}$, where $\mathcal{C}$ is the PTM of the effective error channel $\mathcal{E}^{\text{eff}}$. The resulting noisy expectation value is given by

$$\langle \tilde{Z}_r \rangle_\rho = \langle \langle Z_r^{\text{eff}} | \mathcal{R} \mathcal{C} \mathcal{U} | \rho \rangle \rangle, \quad (46)$$

with $\mathcal{R}$ denoting the measurement noise channel. This expression captures the full effect of noise throughout the measurement process and leads to a systematic deviation from the ideal expectation value $\langle Z_r \rangle_\rho$. The diagonal elements of $\mathcal{C}$, notably $[\mathcal{C}]_{\phi(r)\phi(r)}$, multiply the ideal expectation value by a constant factor. This factor can be readily mitigated through calibration, as shown in Eq. (14). However, the off-diagonal terms—including coherent errors—introduce a nontrivial bias that cannot be corrected by simple scaling.

To address both measurement noise and the gate errors introduced by the measurement circuit, we combine them into a single effective noise channel $\tilde{\mathcal{R}} = \mathcal{R}\mathcal{C}$. We then apply SB-PT tailored to the effective observable Z_r^{eff} . For each selected

Pauli operator $P_q \in \mathcal{S}^{\text{sub}}$, we first apply the physically transformed operator $P_{q'} = \mathbb{U}(P_q) = U_{\text{MT}} P_q U_{\text{MT}}^\dagger$ prior to the noisy measurement circuit $\tilde{\mathbb{U}}$, and then virtually apply the original P_q during classical postprocessing as described in Sec. II B. This unified mitigation framework suppresses errors in $\overline{\mathcal{R}}$, enabling accurate estimation of $\langle Z_r \rangle_\rho$ even in the presence of coherent errors in both measurement and circuit operations.

V. NUMERICAL SIMULATIONS

To assess the performance of our method, we perform extensive numerical simulations using two classes of noise-free quantum states: (1) random computational-basis product states, and (2) Haar-random pure states.

A. Noise settings

Measurement noise is systematically incorporated via two distinct types of measurement transfer matrices (see detailed illustration in Appendix D 1). The first type is reconstructed from empirical calibration data of qubits $\{Q02, Q05, Q06, Q09, Q10, Q11\}$ on the *Zuchongzhi* 2.1 superconducting quantum processor [28,30], which exhibits a measurement error rate (MER) of 0.0452 ± 0.0143 .

However, since this experimental characterization is limited to six qubits, we introduce a second, programmable synthetic noise model to investigate larger systems. This synthetic noise model incorporates two tunable components, independent and correlated measurement errors, both designed to capture the nonuniform noise profiles observed in real devices. The independent part is constructed as a tensor product of single-qubit transfer matrices, where each error strength is independently sampled from a normal distribution with mean 0.015 and standard deviation 0.01. Similarly, the correlated component is applied to each of the $n - 1$ adjacent qubit pairs via a two-qubit noise operator based on the continuous-time Markov process model [5]:

$$\Lambda = \exp \left(\sum_{i=1}^{4(n-1)} \lambda_i G_i \right), \quad (47)$$

where each λ_i is also drawn independently from a normal distribution with mean 0.008 and standard deviation 0.005. These $\{\lambda_i\}_i$ correspond to the strengths of the generators $\{G_i\}_i$, which belong to four fundamental types of two-qubit readout error processes:

- (1) $|10\rangle\langle 01| - |01\rangle\langle 01|$ refers to readout error $01 \rightarrow 10$,
- (2) $|01\rangle\langle 10| - |10\rangle\langle 10|$ refers to readout error $10 \rightarrow 01$,
- (3) $|11\rangle\langle 00| - |00\rangle\langle 00|$ refers to readout error $00 \rightarrow 11$,
- (4) $|00\rangle\langle 11| - |11\rangle\langle 11|$ refers to readout error $11 \rightarrow 00$.

The overall synthetic transfer matrix is constructed as a combination of these two noise components. By sampling both independent and correlated error strengths probabilistically, the model introduces qubit-dependent variations that closely emulate the behavior of real superconducting quantum processors.

To model gate noise in the measurement circuit, we employ a hybrid approach combining both incoherent and coherent errors. Incoherent noise is introduced via a global depolarizing channel, with error rates set to 5×10^{-4} per single-qubit gate

and 5×10^{-3} per two-qubit gate. Coherent errors are incorporated using a unitary perturbation of the form $\exp(-i\frac{\beta}{2}XX)$, where the noise strength is set to $\beta = 0.01$.

Unless otherwise stated, the first type of transfer matrix is used by default.

B. Results

We begin by numerically validating the performance of SB-PT with MF-MEM mentioned in Sec. II B, using six-qubit Haar-random pure states. We evaluate three Pauli-Z observables with increasing weights τ : $ZIIIII$ ($\tau = 1$), $ZZIIII$ ($\tau = 2$), and $ZZZIII$ ($\tau = 3$), comparing the absolute error between ideal expectation values and estimates from four methods: direct noisy measurement (Noisy), MEM with TPN model (TPN), MF-MEM with random twirling (MF), and SB-PT [MF(sub)]. Notably, when the size of the twirling set is less than 4^τ , SB-PT strategically constructs the set to ensure that Pauli terms acting on the measured subsystem are nonrepeating.

As shown in Fig. 3, SB-PT significantly outperforms random twirling, achieving superior sampling efficiency and estimation accuracy. A key observation is the pronounced performance leap when the number of random circuits exceeds 4^τ . This threshold behavior aligns with the theoretical expectation that SB-PT requires significantly fewer samples than random twirling once the subsystem symmetry is adequately exploited. These results conclusively demonstrate the enhanced efficiency of SB-PT, particularly for low-weight Pauli-Z observables, establishing its advantage over conventional mitigation techniques.

We next extend our analysis to global Pauli-Z observables, where the observable's weight equals the system size n . In this regime, SB-PT alone fails to provide advantages due to the high weight of the observable. To address this challenge, we incorporate MT to reduce the global observable to weight-1 using a circuit of depth $n - 1$ (see Appendix D 2 for detailed circuit implementation). This approach leads to two hybrid protocols: MF-MEM with MT and random twirling [MT(rnd)], and MF-MEM with MT and SB-PT [MT(sub)]. The results presented in Fig. 4(c) are obtained using the second type of synthetic measurement noise model.

As presented in Figs. 4(a) and 4(b), MT(sub) demonstrates a substantial performance advantage, achieving near-optimal error mitigation with very few random circuit instances. Notably, MT(sub) using only $r_i = 4$ attains accuracy comparable to that of the second-best method—MT(rnd) in (a) and MF in (b)—evaluated at $r_i = 64$, corresponding to a 16-fold improvement in sampling efficiency.

The scaling properties of the methods are further investigated using Haar-random states, as shown in Figs. 4(c)–4(e). Figure 4(c) reveals a critical divergence in scaling behavior with increasing qubit number n , using the relative error (defined as the logarithm of the ratio between mitigated and unmitigated absolute errors). The relative error for MT(sub) rises gradually, confirming its favorable scaling for larger systems. In contrast, the relative error for the MF method increases abruptly for $n > 10$, indicating limited scalability. The shot-number dependence in Fig. 4(d) demonstrates that

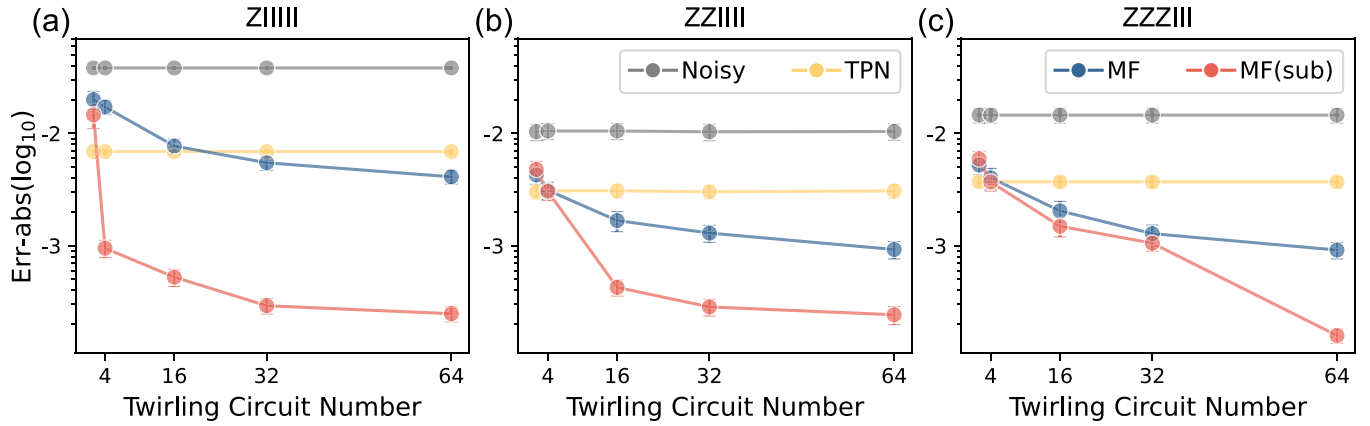


FIG. 3. The performance comparison between random Pauli twirling and SB-PT with increasing random twirling circuit number. Absolute error between ideal expectation values and estimated results for three six-qubit Pauli observables: (a) ZIIIII, (b) ZZIIII, and (c) ZZZIII, using Haar-random pure states. Four methods are compared: direct noisy measurement (Noisy), measurement error mitigation with tensor product noise model (TPN), model-free mitigation with random Pauli twirling (MF), and model-free mitigation with subsystem-balanced Pauli twirling [MF(sub)]. Error bars represent the SEM computed from 100 independent replicates.

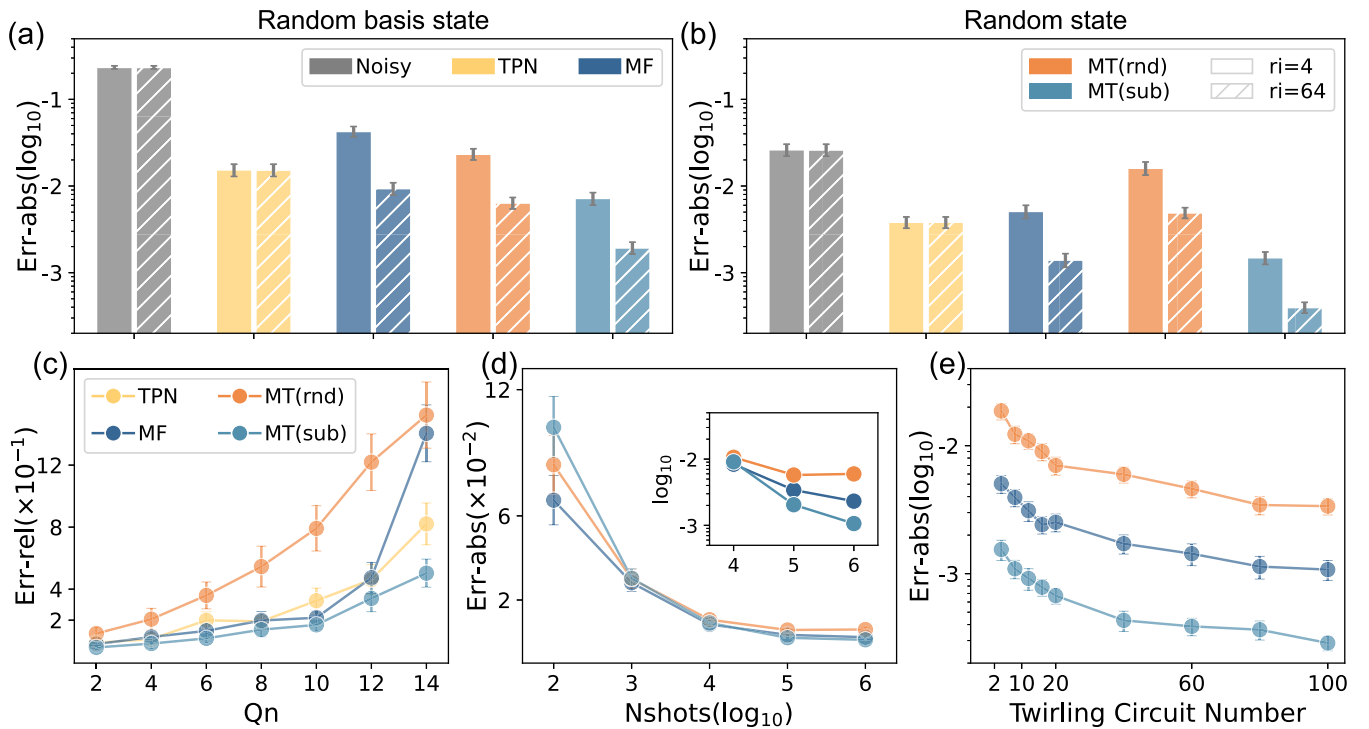


FIG. 4. Comprehensive performance evaluation of measurement error mitigation methods for global Pauli-Z observables. Absolute error between ideal expectation values and mitigated estimates is compared across five methods: direct noisy measurement (Noisy), measurement error mitigation with tensor product noise model (TPN), model-free mitigation (MF), and model-free mitigation with measurement transformation (MT) using either random Pauli twirling [MT(rnd)] or subsystem-balanced Pauli twirling [MT(sub)]. (a) Results for six-qubit random computational-basis states with different random circuit numbers ($ri = 4$ and 64). (b) Results for a six-qubit Haar-random pure states with different random circuit numbers ($ri = 4$ and 64). (c) Relative error (logarithm of the ratio between mitigated and unmitigated absolute errors) for random states with increasing qubit number (Qn). (d) Performance with increasing measurement shot number for three model-free methods. (e) Performance with increasing random twirling circuit number for three twirling-based model-free mitigation methods: MF, MT(rnd), and MT(sub). For all panels except (d), the measurement shot number is infinite. The measurement noise model is constructed as a tensor product of independent and two-local matrices. Results presented in panel (c) are obtained using a synthetic measurement noise model; others are obtained using reconstructed noise from calibrated data on the *Zuchongzhi* 2.1 superconducting quantum processor. And in simulations for panels (c)–(e), we use Haar-random pure states. Error bars represent the standard error of the mean computed from 100 independent replicates.

the absolute error of MT(sub) continues to improve up to 10^6 shots, at which point it exceeds the precision of MT(rnd) by an order of magnitude. This confirms its enhanced capability to utilize increased sampling for uncertainty reduction. As shown in Fig. 4(e), MT(sub) exhibits significantly higher sampling efficiency than all other methods: it achieves at $r_i = 4$ an accuracy comparable to that of MF at $r_i = 100$. Collectively, these results affirm that the MT(sub) protocol—enabled by SB-PT—offers superior scaling in both system size and sampling efficiency, advancing scalable MEM. Further exploration of the relation between MT(sub) performance and noise parameters is provided in Appendix D 3.

To examine its flexibility under realistic hardware constraints, where two-qubit gate performance degrades rapidly with circuit depth, we evaluate the trade-off between mitigation accuracy and the depth of measurement circuit using Haar-random states.

We consider two structured Pauli-Z observables: $o_1 = ZZIZZI$ and $o_2 = ZIZIZI$. For each observable, we compare mitigation using MT with the effective weight reduced to $\tau^{\text{eff}} = 1$ [MT($\tau^{\text{eff}} = 1$)] and to $\tau^{\text{eff}} = 2$ [MT($\tau^{\text{eff}} = 2$)], both implemented via tailored sequences of WR and LS operations and both employing SB-PT (see Appendix D 2 for more details). For observable o_1 , the support of the effective observable is reduced to qubit {1} under $\tau^{\text{eff}} = 1$, and to qubits {1, 2} under $\tau^{\text{eff}} = 2$. Similarly, for o_2 , the support is restricted to qubit {1} at $\tau^{\text{eff}} = 1$, and to qubits {1, 3} at $\tau^{\text{eff}} = 2$.

As shown in Fig. 5, MT($\tau^{\text{eff}} = 2$) nearly matches the error rates of MT($\tau^{\text{eff}} = 1$), while reducing circuit depth by over 50%: from 5 to 1 for o_1 and from 6 to 3 for o_2 . Notably, despite this slight compromise in precision, MT($\tau^{\text{eff}} = 2$) still significantly outperforms conventional methods such as TPN and MF methods. These results establish the efficacy of our method in mitigating errors under stringent depth constraints while largely preserving accuracy.

VI. CONCLUSION

This work establishes an efficient and scalable framework for MEM on near-term quantum devices through the introduction of SB-PT. SB-PT embodies a resource-aware design that strategically concentrates twirling on the measured subsystem, leveraging the empirical prior of hierarchical noise structure to maximize mitigation efficacy under finite sampling constraints, while remaining effective in the general model-free MEM context. Its conceptual advance is supported by a rigorously derived sampling bound that is strictly tighter than that of random Pauli twirling under finite sampling.

We further integrate MT to extend these advantages to global observables, resulting in a unified architecture that maintains both scalability and hardware efficiency. The resulting framework preserves linear circuit depth in the number of qubits and relies on hardware-friendly nearest-neighbor CX gates, making it especially suitable for superconducting quantum platforms. Crucially, our model offers a unified treatment of native measurement noise and MT-induced gate noise through noise propagation equivalence,

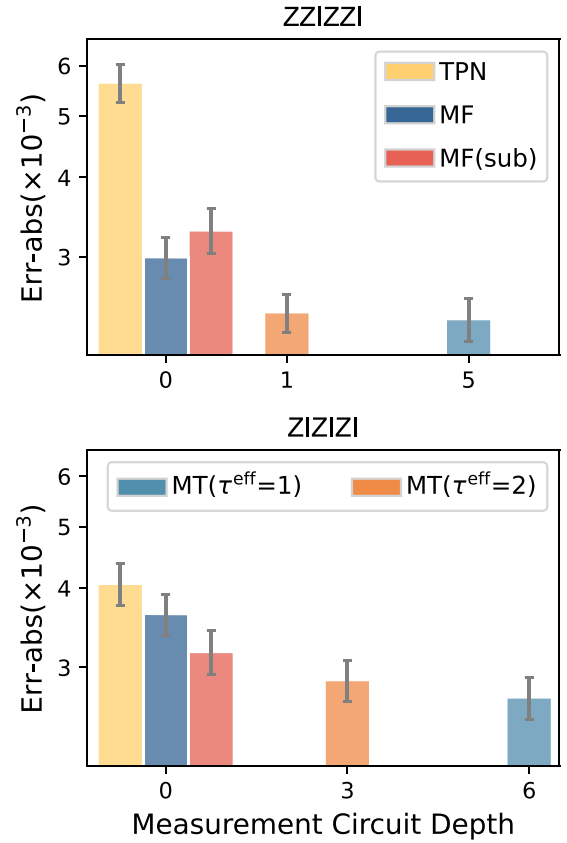


FIG. 5. Performance comparison of error mitigation methods for structured Pauli-Z observables. Numerical simulations use ideal six-qubit Haar-random pure states measured under a noise model characterized by calibration data from *Zuchongzhi* 2.1. Two observables are evaluated: $ZZIZZI$ and $ZIZIZI$. MT($\tau^{\text{eff}} = 1$) denotes the protocol converting the observable to effective weight 1 via measurement transformation, while MT($\tau^{\text{eff}} = 2$) converts to weight 2. Both employ SB-PT. Error bars indicate the SEM over 100 independent replicates.

enabling holistic error mitigation within a coherent MF-MEM pipeline.

Comprehensive numerical simulations confirm these advances: SB-PT demonstrates a clear performance separation beyond the theoretical threshold of 4^r random circuits for weight- τ observables, underscoring its superior sampling efficiency. When integrated with MT, the framework achieves up to a 16-fold reduction in required random circuit instances while preserving accuracy across increasing system sizes. Furthermore, reducing the effective observable weight from 1 to 2 yields over 50% savings in circuit depth with marginal precision loss, providing practical flexibility for depth-constrained hardware.

Looking forward, this work opens multiple paths for extension—including adaptive MT compilation tuned to specific states and devices, experimental validation on superconducting quantum processors, and generalization to tasks such as one-shot verification. By unifying theoretical soundness with hardware-aware design, SB-PT provides a viable pathway toward scalable quantum advantage in the noisy intermediate-scale quantum era.

ACKNOWLEDGEMENT

This work was supported by the Henan Science and Technology Major Project of the Department of Science & Technology of Henan Province (No. 241100210400).

DATA AVAILABILITY

The data that support the findings of this article are not publicly available. The data are available from the authors upon reasonable request.

APPENDIX A: THE PROOFS OF THEOREMS

1. The proof of Theorem 1

For any quantum state ρ and Pauli-Z observable Z_r , the estimation error is

$$\begin{aligned} \epsilon(\rho) &= |a(\rho) - \text{Tr}(\rho Z_r)| = \left| \frac{v(\rho)}{v(\rho_0)} - \text{Tr}(\rho Z_r) \right| \\ &= \frac{|v(\rho) - \text{Tr}(\rho Z_r)v(\rho_0)|}{|v(\rho_0)|}. \end{aligned} \quad (\text{A1})$$

By expanding the numerator of the given formula, we derive

$$\begin{aligned} v(\rho) - \text{Tr}(\rho Z_r)v(\rho_0) &= 2^n \left(\alpha_{\phi(r)\phi(r)}(\mathcal{S})[\mathcal{R}]_{\phi(r)\phi(r)}\varrho_{\phi(r)}(\rho) + \sum_{j \neq \phi(r)} \alpha_{\phi(r)j}(\mathcal{S})[\mathcal{R}]_{\phi(r)j}\varrho_j(\rho) - \varrho_{\phi(r)}(\rho) \sum_{j=0}^{4^n-1} \alpha_{\phi(r)j}(\mathcal{S})[\mathcal{R}]_{\phi(r)j} \right) \\ &= 2^n \left(\sum_{j \neq \phi(r)} \alpha_{\phi(r)j}(\mathcal{S})[\mathcal{R}]_{\phi(r)j}\varrho_j(\rho) - \varrho_{\phi(r)}(\rho) \sum_{j \neq \phi(r)} \alpha_{\phi(r)j}(\mathcal{S})[\mathcal{R}]_{\phi(r)j} \right). \end{aligned} \quad (\text{A2})$$

Here, we use the fact that $\alpha_{\phi(r)\phi(r)} = 1$ and $\varrho_j(\rho_0) = \frac{1}{2^n}$ for $\rho_0 = |0\rangle\langle 0|^{\otimes n}$ and all $j \in \{0, 1, \dots, 4^n - 1\}$. As $|\varrho_j(\rho)| \leq \frac{1}{2^n}$, we obtain

$$\begin{aligned} |v(\rho) - \text{Tr}(\rho Z_r)v(\rho_0)| &\leq 2^n \left(\sum_{j \neq \phi(r)} |\alpha_{\phi(r)j}(\mathcal{S})| |[\mathcal{R}]_{\phi(r)j}| |\varrho_j(\rho)| + |\varrho_{\phi(r)}(\rho)| \sum_{j \neq \phi(r)} |\alpha_{\phi(r)j}(\mathcal{S})| |[\mathcal{R}]_{\phi(r)j}| \right) \\ &\leq 2 \sum_{j \neq \phi(r)} |\alpha_{\phi(r)j}(\mathcal{S})| |[\mathcal{R}]_{\phi(r)j}| \\ &\leq 2 \max_{j \neq \phi(r)} |\alpha_{\phi(r)j}(\mathcal{S})| \sum_{j \neq \phi(r)} |[\mathcal{R}]_{\phi(r)j}|. \end{aligned} \quad (\text{A3})$$

According to Hoeffding's inequality, for a fixed $j \neq \phi(r)$ and twirling set $\mathcal{S}$ chosen randomly from $\{I, X, Y, Z\}^{\otimes n}$, it gives

$$\mathbb{P}(|\alpha_{\phi(r)j}(\mathcal{S})| \geq \epsilon) \leq 2 \exp\left(-\frac{|\mathcal{S}|\epsilon^2}{2}\right). \quad (\text{A4})$$

Taking a union bound over $4^n - 1$ off-diagonal indices $j \neq \phi(r)$

$$\mathbb{P}(\max_{j \neq \phi(r)} |\alpha_{\phi(r)j}(\mathcal{S})| \geq \epsilon) \leq 2 \times 4^n \exp\left(-\frac{|\mathcal{S}|\epsilon^2}{2}\right). \quad (\text{A5})$$

Setting $\delta = 2 \cdot 4^n \exp(-\frac{|\mathcal{S}|\epsilon^2}{2})$ and solving for ϵ

$$\epsilon = \sqrt{\frac{2}{|\mathcal{S}|} (2n \ln 2 + \ln(2/\delta))}. \quad (\text{A6})$$

Thus, with probability $\geq 1 - \delta$,

$$\max_{j \neq \phi(r)} |\alpha_{\phi(r)j}(\mathcal{S})| \leq \kappa(\mathcal{S}) \triangleq \sqrt{\frac{2}{|\mathcal{S}|} (2n \ln 2 + \ln(2/\delta))}. \quad (\text{A7})$$

Then, we can bound the numerator as

$$|v(\rho) - \text{Tr}[\rho Z_r]v(\rho_0)| \leq 2\kappa(\mathcal{S}) \sum_{j \neq \phi(r)} |[\mathcal{R}]_{\phi(r)j}|. \quad (\text{A8})$$

Assuming that $[\mathcal{R}]_{\phi(r)\phi(r)}$ is the dominant term (i.e., the diagonal elements of the noise matrix are significantly larger),

and the cross terms are relatively small, we derive bounds that are independent of ρ . To this end, we assume that the measurement noise $\mathcal{R}$ satisfies $|[\mathcal{R}]_{\phi(r)\phi(r)}|$ being substantially larger than the other terms:

$$|[\mathcal{R}]_{\phi(r)\phi(r)}| > \kappa(\mathcal{S}) \sum_{j \neq \phi(r)} |[\mathcal{R}]_{\phi(r)j}|. \quad (\text{A9})$$

Then, for the denominator of Eq. (A1), we can directly obtain

$$\begin{aligned} |v(\rho_0)| &\geq |[\mathcal{R}]_{\phi(r)\phi(r)}| - \sum_{j \neq \phi(r)} |\alpha_{\phi(r)j}(\mathcal{S})| |[\mathcal{R}]_{\phi(r)j}| \\ &\geq |[\mathcal{R}]_{\phi(r)\phi(r)}| - \kappa(\mathcal{S}) \sum_{j \neq \phi(r)} |[\mathcal{R}]_{\phi(r)j}|. \end{aligned} \quad (\text{A10})$$

Combining the results from Eqs. (A3) and (A10), we can derive the upper bound as given in Theorem 1.

2. The proof of Theorem 2

For Pauli-Z observable Z_r , the scaling factor for any $s \in \mathcal{J}_r$ is

$$\alpha_{\phi(r)\phi(s)}(\mathcal{S}^{\text{sub}}) = \frac{1}{|\mathcal{S}^{\text{sub}}|} \sum_{P_q \in \mathcal{S}^{\text{sub}}} \eta(P_q, Z_r Z_s). \quad (\text{A11})$$

We denote the restriction of P_q to the subset $\text{supp}(Z_r)$ as $P_q|_r = \bigotimes_{i \in \text{supp}(Z_r)} \sigma_{\mathbf{q}_i}$. Similarly, the restriction to the complement

of Z_r is denoted as $P_q|_{\bar{r}} = \bigotimes_{i' \notin \text{supp}(Z_r)} \sigma_{q_{i'}}$. This restriction notation naturally extends to any Pauli operator, including $Z_r Z_s$. As for any $s \in \mathcal{J}_r$, $\text{supp}(Z_s) \subseteq \text{supp}(Z_r)$ which indicates $(Z_r Z_s)|_{\bar{r}} = I$, and $(Z_r Z_s)|_r \neq I$ ($Z_r \neq Z_s$), then we can arrive at

$$\begin{aligned} \alpha_{\phi(r)\phi(s)}(\mathcal{S}^{\text{sub}}) &= \frac{1}{|\mathcal{S}^{\text{sub}}|} \sum_{P_q \in \mathcal{S}^{\text{sub}}} \eta(P_q|_r, (Z_r Z_s)|_r) \\ &\quad \times \eta(P_q|_{\bar{r}}, (Z_r Z_s)|_{\bar{r}}) \\ &= \frac{1}{|\mathcal{S}^{\text{sub}}|} \sum_{P_q \in \mathcal{S}^{\text{sub}}} \eta(P_q|_r, (Z_r Z_s)|_r) \eta(P_q|_{\bar{r}}, I) \\ &= \frac{1}{|\mathcal{S}^{\text{sub}}|} \sum_{P_q \in \mathcal{S}^{\text{sub}}} \eta(P_q|_r, (Z_r Z_s)|_r) \\ &= 0. \end{aligned} \quad (\text{A12})$$

In the derivation above, the third equality follows from the identity $\eta(P_q, I) = 1$ for all $q \in \{0, \dots, 4^n - 1\}$. The fourth equality is obtained by noting that the restricted operator $P_q|_r = \bigotimes_{i \in \text{supp}(Z_r)} \sigma_{q_i}$ is uniformly distributed over the twirling set $\mathcal{S}^{\text{sub}}$ of size $c \cdot 4^{\tau(r)}$, thereby forming a complete and balanced sampling of the Pauli subgroup on the support of Z_r .

3. The proof of Theorem 3

The proof of Eq. (31) is similar to that in Theorem 1, and we only detail the derivation of $\kappa_{\text{SB}}(\mathcal{S}^{\text{sub}})$. As given in Theorem 2, for any $s \in \mathcal{J}_r$, $\alpha_{\phi(r)\phi(s)}(\mathcal{S}^{\text{sub}})$ can be regarded as a random variable with

- (1) $\mathbb{E}[\alpha_{\phi(r)\phi(s)}] = 0$ (for $j \neq \phi(r)$),
- (2) $|\alpha_{\phi(r)\phi(s)}| \leq 1$.

We can expand the indices in $\mathcal{J}_r$ which satisfies Eq. (A12) beyond the Pauli-Z subset, since $(Z_r P_q)|_{\bar{r}} = I$ holds for any $\text{supp}(P_q) \subseteq \text{supp}(Z_r)$. We define the extension set as $\Phi(\mathcal{J}_r) = \{\phi(s) | s \in \mathcal{J}_r\} \subseteq \{0, \dots, 4^n - 1\}$. Applying Hoeffding inequality combined with the union bound over $|\Phi(\mathcal{J}_r)| = 4^{n-\tau(r)}$,

$$\mathbb{P}\left(\max_{j \notin \Phi(\mathcal{J}_r)} |\alpha_{\phi(s)j}| \geq \epsilon\right) \leq 2 \times 4^{n-\tau(r)} \exp\left(-\frac{|\mathcal{S}^{\text{sub}}|\epsilon^2}{2}\right), \quad (\text{A13})$$

and we can solve for ϵ via setting

$$\delta = 2 \times 4^{n-\tau(r)} \exp\left(-\frac{|\mathcal{S}^{\text{sub}}|\epsilon^2}{2}\right), \quad (\text{A14})$$

we have

$$\kappa_{\text{SB}}(\mathcal{S}^{\text{sub}}) \triangleq \epsilon = \sqrt{\frac{2}{|\mathcal{S}^{\text{sub}}|} (2m \ln 2 + \ln(2/\delta))}, \quad (\text{A15})$$

where $m = n - \tau(r) = n - |\text{supp}(Z_r)|$.

We next demonstrate that the error bound in Theorem 3 is tighter than that in Theorem 1. Defining the function $f(x) = \frac{2x}{a-x}$ for some constant a , it is straightforward to verify that f is monotonically increasing for $x < a$. Since $\kappa(\mathcal{S}) > \kappa_{\text{SB}}(\mathcal{S}^{\text{sub}})$ and $\sum_{j \neq \phi(r)} |\mathcal{R}|_{\phi(r)j}| \geq \sum_{j \notin \Phi(\mathcal{J}_r)} |\mathcal{R}|_{\phi(r)j}|$ with twirling set of the same size $|\mathcal{S}| = |\mathcal{S}^{\text{sub}}|$, we arrive at the conclusion that Theorem 3 provide a tighter error bound.

APPENDIX B: THE PAULI TRANSFER MATRIX OF THE TENSOR PRODUCT NOISE MODEL

In this Appendix, we will provide the detailed derivation of the PTM structure for the TPN model of classical measurement noise.

1. General form of the measurement noise in the PTM representation

We begin by deriving the general expression for a classical measurement noise $\mathcal{R}$ in the PTM representation. A classical noise model implies that the noisy measurement process can be described by a conditional probability distribution $[\Lambda]_{kj} = \mathbb{P}(\text{noisy } j | \text{ideal } k)$, for $k, j \in \{0, \dots, 2^n - 1\}$. The action of $\mathcal{R}$ on an ideal computational basis state $|k\rangle\langle k|$ is

$$\mathcal{R}|k\rangle\langle k| = \sum_{j=0}^{2^n-1} \Lambda_{kj} |j\rangle\langle j|, \quad (\text{B1})$$

where $|k\rangle\langle k|$ ($|j\rangle\langle j|$) is the PTM representation of state $|k\rangle\langle k|$ ($|j\rangle\langle j|$). For $p, q \in \{0, \dots, 4^n - 1\}$, the element $[\mathcal{R}]_{pq}$, which describes how the noise channel maps the Pauli operator P_p to the component of P_q in the output, is defined as

$$\mathcal{R}t_{pq} = \frac{1}{2^n} \langle\langle P_p | \mathcal{R} | P_q \rangle\rangle. \quad (\text{B2})$$

Since the output of $\mathcal{R}$ is always a diagonal matrix in the computational basis ($\sum_j \dots |j\rangle\langle j|$), the inner product $\langle\langle P_p | \cdot \rangle\rangle$ will be nonzero only if P_p itself is diagonal in the computational basis. This restricts P_p to the Pauli-Z operators, i.e., $P_p \in \{I, Z\}^{\otimes n}$. This justifies the definition of the Pauli-Z subset $\mathbb{I}_Z$ and the reduced matrix $\mathcal{R}_Z$. For the matrix elements $[\mathcal{R}]_{\phi(r), \phi(s)}$ where $P_{\phi(r)} = Z_r$ and $P_{\phi(s)} = Z_s$ are both Pauli-Z operators ($r, s \in \{0, \dots, 2^n - 1\}$), we derive

$$\begin{aligned} [\mathcal{R}]_{\phi(r), \phi(s)} &= [\mathcal{R}_Z]_{rs} \\ &= \frac{1}{2^n} \langle\langle Z_r | \mathcal{R} | Z_s \rangle\rangle \\ &= \frac{1}{2^n} \langle\langle Z_r | \mathcal{R} \left(\sum_{k=0}^{2^n-1} (-1)^{\mathbf{s} \cdot \mathbf{k}} |k\rangle\langle k| \right) \rangle\rangle \\ &= \frac{1}{2^n} \langle\langle Z_r | \left(\sum_{k=0}^{2^n-1} (-1)^{\mathbf{s} \cdot \mathbf{k}} \mathcal{R}|k\rangle\langle k| \right) \rangle\rangle \\ &= \frac{1}{2^n} \langle\langle Z_r | \left(\sum_{k=0}^{2^n-1} (-1)^{\mathbf{s} \cdot \mathbf{k}} \sum_{j=0}^{2^n-1} [\Lambda]_{kj} |j\rangle\langle j| \right) \rangle\rangle \\ &= \frac{1}{2^n} \sum_{k,j} [\Lambda]_{kj} (-1)^{\mathbf{s} \cdot \mathbf{k}} \langle\langle Z_r | j \rangle\rangle \\ &= \frac{1}{2^n} \sum_{k,j} [\Lambda]_{kj} (-1)^{\mathbf{s} \cdot \mathbf{k}} (-1)^{\mathbf{r} \cdot \mathbf{j}} \quad (\text{since } \langle\langle Z_r | j \rangle\rangle = (-1)^{\mathbf{r} \cdot \mathbf{j}}) \\ &= \frac{1}{2^n} \sum_{j=0}^{2^n-1} (-1)^{\mathbf{r} \cdot \mathbf{j}} \sum_{k=0}^{2^n-1} [\Lambda]_{kj} (-1)^{\mathbf{s} \cdot \mathbf{k}}, \end{aligned} \quad (\text{B3})$$

where Z_r denotes the PTM representation of Pauli-Z operator Z_r .

2. Factorization under the TPN model

The TPN model assumes that the noise is local and uncorrelated across qubits. The conditional probability matrix factorizes as $\Lambda = \bigotimes_{i=1}^n \Lambda^{(i)}$, where the single-qubit response matrix for the i th qubit is

$$\Lambda^{(i)} = \begin{pmatrix} a_i & 1 - b_i \\ 1 - a_i & b_i \end{pmatrix} = \begin{pmatrix} \mathbb{P}(0|0) & \mathbb{P}(0|1) \\ \mathbb{P}(1|0) & \mathbb{P}(1|1) \end{pmatrix}.$$

We introduce the parameters ω_i and ζ_i as defined in the main text:

$$\begin{aligned} \omega_i &= a_i + b_i - 1, \\ \zeta_i &= a_i - b_i. \end{aligned}$$

These parameters define the single-qubit PTM $\mathcal{R}_Z^{(i)}$ for the Pauli-Z operators $\{I, Z\}$ on the i th qubit:

$$\mathcal{R}_Z^{(i)} = \begin{pmatrix} 1 & 0 \\ \omega_i & \zeta_i \end{pmatrix}. \quad (\text{B4})$$

Due to the tensor product structure of the noise, the overall reduced PTM $\mathcal{R}_Z$ is the tensor product of the single-qubit PTMs:

$$\mathcal{R}_Z = \bigotimes_{i=1}^n \mathcal{R}_Z^{(i)} = \bigotimes_{i=1}^n \begin{pmatrix} 1 & 0 \\ \omega_i & \zeta_i \end{pmatrix}. \quad (\text{B5})$$

For the Pauli-Z operators $Z_r = \bigotimes_{i=1}^n Z^{r_i}$ and $Z_s = \bigotimes_{i=1}^n Z^{s_i}$ ($\mathbf{r}, \mathbf{s} \in \mathbb{Z}_2^n$), where we denote $Z^0 = I$ and $Z^1 = Z$, the matrix element $[\mathcal{R}_Z]_{rs}$ is a product over the qubits $i = 1 \dots n$:

$$[\mathcal{R}_Z]_{rs} = \prod_{i=1}^n [\mathcal{R}_Z^{(i)}]_{r_i s_i}. \quad (\text{B6})$$

For the overall product to be nonzero, we must avoid any factor where $\mathbf{r}_i = 0$ and $\mathbf{s}_i = 1$, as this gives $[\mathcal{R}_Z^{(i)}]_{0,1} = 0$. This is precisely the condition $\text{supp}(Z_s) \subseteq \text{supp}(Z_r)$, defining the TPN subset $\mathbb{T}$. For $(r, s) \in \mathbb{T}$, the product simplifies. We separate the qubits into three groups:

- (1) Qubits i , where $\mathbf{s}_i = 1$. Since $(r, s) \in \mathbb{T}$, we must have $\mathbf{r}_i = 1$ for these qubits. Their contribution to the product is $\prod_{i: \mathbf{s}_i=1} \zeta_i$.
- (2) Qubits i , where $\mathbf{r}_i = 1$ but $\mathbf{s}_i = 0$. Their contribution is $\prod_{i: \mathbf{r}_i=1, \mathbf{s}_i=0} \omega_i$.
- (3) Qubits i , where $\mathbf{r}_i = 0$ (and consequently $\mathbf{s}_i = 0$). Their contribution is $\prod_{i: \mathbf{r}_i=0} 1 = 1$.

Thus, the final expression for the matrix element is

$$[\mathcal{R}_Z]_{rs} = \left(\prod_{i: \mathbf{s}_i=1} \zeta_i \right) \times \left(\prod_{i: \mathbf{r}_i=1, \mathbf{s}_i=0} \omega_i \right), \quad \text{for } (r, s) \in \mathbb{T}. \quad (\text{B7})$$

This completes the derivation. The first product quantifies the faithful transmission of a Z error on the qubits in the support of Z_s , while the second product captures the noise introduced on qubits that are measured but did not have an underlying error.

APPENDIX C: THE MEASUREMENT CIRCUIT USED FOR MEASUREMENT TRANSFORMATION

1. The circuit for weight reduction

Let $\Gamma_k^r = Z^{(i_k^1)} \otimes \dots \otimes Z^{(i_k^{r_k(r)})}$ be a Pauli-Z operator acting on $\tau_k(r)$ contiguous qubits. The WR circuit is defined as

$$U_R(\Gamma_k^r, t_k) = U_C(i_k^{\tau_k(r)}, t_k) U_C(i_k^1, t_k), \quad t_k \in \{i_k^1, \dots, i_k^{\tau_k(r)}\}, \quad (\text{C1})$$

where $U_C(i, j)$ is a chain of CX gates:

$$U_C(i, j) = \begin{cases} \text{CX}_{i,i+1} \text{CX}_{i+1,i+2} \dots \text{CX}_{j-1,j}, & \text{if } i < j, \\ \text{CX}_{(i,i-1)} \text{CX}_{(i-1,i-2)} \dots \text{CX}_{(j+1,j)}, & \text{if } i > j, \\ I^{\otimes n}, & \text{if } i = j. \end{cases} \quad (\text{C2})$$

We now show that under this transformation, the effective Pauli observable becomes $Z^{(t_k)}$. The conjugate action of a CX gate $\text{CX}_{i,j}$, with qubit i as control and qubit j as target, on Pauli operators is

$$\text{CX}_{i,j}(Z^{(i)} \otimes Z^{(j)}) \text{CX}_{i,j}^\dagger = I^{(i)} \otimes Z^{(j)}. \quad (\text{C3})$$

Then, for the WR circuit, we have

$$\begin{aligned} U_R(\Gamma_k^r, t_k) \Gamma_k^r U_R^\dagger(\Gamma_k^r, t_k) &= U_C(i_k^{\tau_k(r)}, t_k) (U_C(i_k^1, t_k) \Gamma_k^r U_C^\dagger(i_k^1, t_k)) U_C^\dagger(i_k^{\tau_k(r)}, t_k) \\ &= U_C(i_k^{\tau_k(r)}, t_k) (Z^{(i_k)} \otimes \dots \otimes Z^{(i_k^{r_k(r)})}) U_C^\dagger(i_k^{\tau_k(r)}, t_k) \\ &= Z^{(t_k)}. \end{aligned} \quad (\text{C4})$$

2. The circuit for location shift

The LS circuit is defined as

$$U_S(i, i+1) = \text{CX}_{i,i+1} \text{CX}_{i+1,i}. \quad (\text{C5})$$

We now demonstrate that this unitary transformation effects the desired Pauli observable shift:

$$\begin{aligned} U_S(i, i+1) (Z^{(i)} \otimes I^{(i+1)}) U_S^\dagger(i, i+1) &= I^{(i)} \otimes Z^{(i+1)}. \end{aligned} \quad (\text{C6})$$

The transformation proceeds in two stages. First, conjugation by $\text{CX}_{i+1,i}$ yields

$$\text{CX}_{i+1,i}(Z^{(i)} \otimes I^{(i+1)}) \text{CX}_{i+1,i}^\dagger = Z^{(i)} \otimes Z^{(i+1)}. \quad (\text{C7})$$

Subsequent conjugation by $\text{CX}_{i,i+1}$, then produces the final result:

$$\text{CX}_{i,i+1}(Z^{(i)} \otimes Z^{(i+1)}) \text{CX}_{i,i+1}^\dagger = I^{(i)} \otimes Z^{(i+1)}. \quad (\text{C8})$$

The complementary case follows analogously:

$$U_S(i, i-1) (I^{(i-1)} \otimes Z^{(i)}) U_S^\dagger(i, i-1) = Z^{(i-1)} \otimes I^{(i)}. \quad (\text{C9})$$

APPENDIX D: SUPPLEMENTARY SIMULATION DETAILS

1. The measurement noise model employed in numerical simulations

In our numerical simulations, we employ two distinct classical measurement noise models, both characterized by transfer matrices as illustrated in Fig. 6.

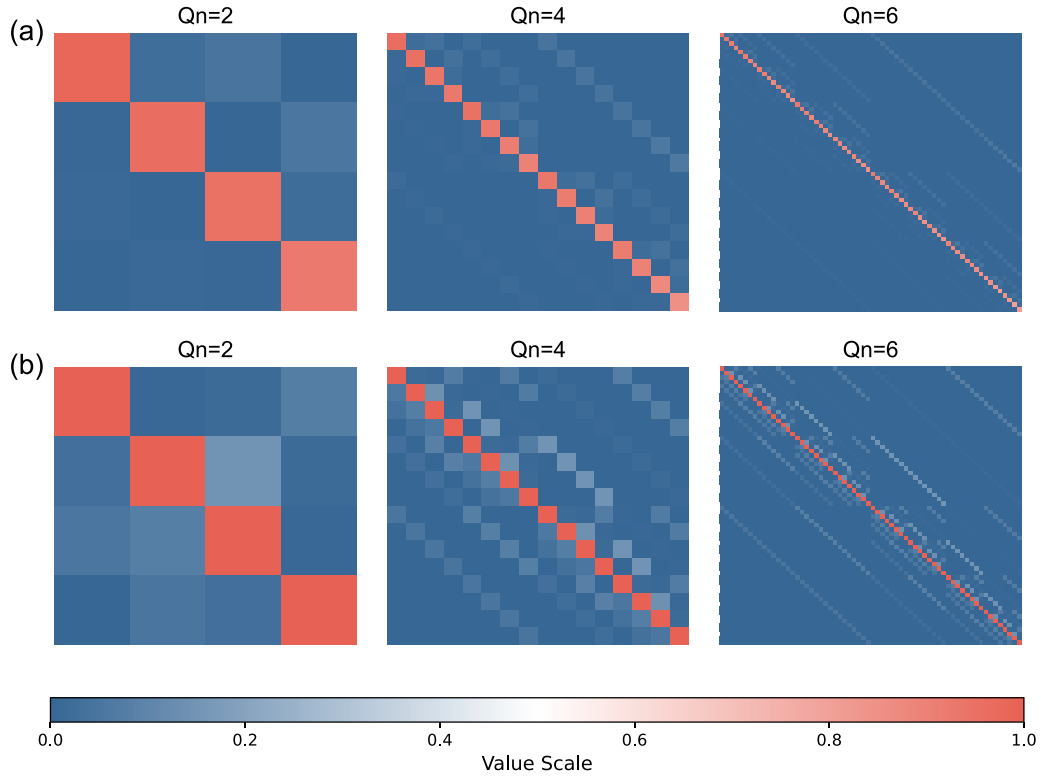


FIG. 6. The demonstration of measurement noise transfer matrices for varying qubit number. (a) The transfer matrix reconstructed from empirical calibration data of qubits $Q02, Q05, Q06, Q09, Q10, Q11$ on the *Zuchongzhi 2.1* superconducting processor [28,30], which exhibits an MER of 0.0452 ± 0.0143 . (b) The programmable synthetic transfer matrix which incorporates dual noise components—independent and correlated measurement errors.

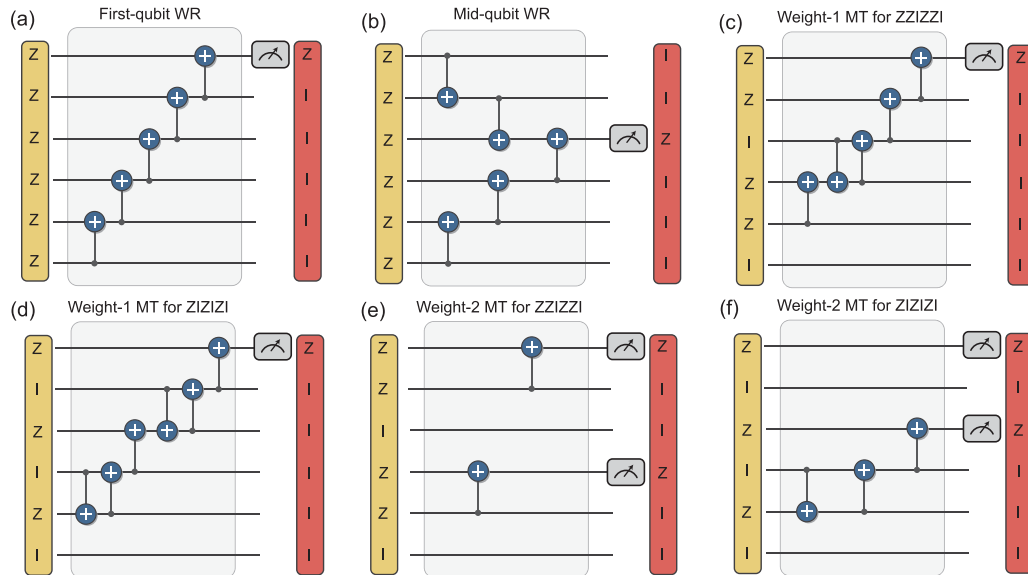


FIG. 7. The measurement circuits used for measurement transformation. Direct measurements of the targeted Pauli-Z observables (yellow) are equivalent to measuring the effective observables (red) via the circuits shown in panels (a)–(f). (a) The circuit for WR to the first qubit, which transfers global Pauli-Z to weight-1 effective Pauli-Z observable. (b) The circuit for WR to the third qubit. (c) The circuit for weight-1 MT for ZZIZZI, which transfers to the effective Pauli-Z with support {1}. (d) The circuit for weight-1 MT for ZIZIZI, which transfers to the effective Pauli-Z with support {1}. (e) The circuit for weight-2 MT for ZZIZZI, which transfers to the effective Pauli-Z with support {1, 4}. (f) The circuit for weight-2 MT for ZIZIZI, which transfers to the effective Pauli-Z with support {1, 3}.

2. Circuit implementation for simulations

The circuit implementation used in our simulations is illustrated in Fig. 7. It shows how direct measurements of targeted Pauli-Z observables can be equivalently performed using measurement circuits tailored to their effective observables.

3. Extended simulation results

In this Appendix, we investigate the robustness of the MT(sub) method against key noise parameters using a synthetic measurement noise model that incorporates both independent and correlated measurement errors, as well as coherent gate errors in the measurement circuit. We focus specifically on the correlated MER and the coherent gate error rate (GER), noting that independent measurement noise and incoherent gate noise are not considered here since they can be effectively suppressed by model-free error mitigation, as discussed in Sec. II B.

As shown in Fig. 8, MT(sub) exhibits strong resilience to both correlated measurement errors and coherent gate errors at moderate noise levels. Notably, when the correlated MER and GER remain below 10^{-2} , the performance degradation of MT(sub) is minimal, demonstrating its stability under realistic noise conditions. This robust behavior highlights the

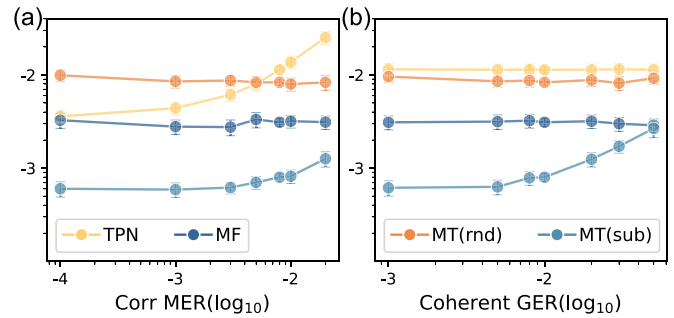


FIG. 8. Performance of mitigation methods under varying noise parameters for global Pauli-Z observable. Absolute error is compared across four methods: direct noisy measurement (Noisy), MF, and twirling with MT(rnd) or MT(sub). Simulations use six-qubit noiseless Haar-random states under the synthetic measurement noise model. (a) The absolute error with increasing corr MER. (b) The absolute error with increasing coherent GER. We set the standard ind MER as 0.015, the corr MER as 0.008, and the coherent GER as 0.01.

practicality of MT(sub) for near-term quantum devices where correlated and coherent errors are non-negligible.

-
- [1] M. A. Nielsen and I. Chuang, *Quantum Computation and Quantum Information*, 10th Anniversary Edition (Cambridge University Press, Cambridge, 2010).
 - [2] T. D. Ladd, F. Jelezko, R. Laflamme, Y. Nakamura, C. Monroe, and J. L. O'Brien, Quantum computers, *Nature (London)* **464**, 45 (2010).
 - [3] P. W. Shor, Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer, *SIAM Rev.* **41**, 303 (1999).
 - [4] Y. Chen, M. Farahzad, S. Yoo, and T.-C. Wei, Detector tomography on IBM quantum computers and mitigation of an imperfect measurement, *Phys. Rev. A* **100**, 052315 (2019).
 - [5] S. Bravyi, S. Sheldon, A. Kandala, D. C. McKay, and J. M. Gambetta, Mitigating measurement errors in multiqubit experiments, *Phys. Rev. A* **103**, 042605 (2021).
 - [6] F. B. Maciejewski, Z. Zimborás, and M. Oszmaniec, Mitigation of readout noise in near-term quantum devices by classical post-processing based on detector tomography, *Quantum* **4**, 257 (2020).
 - [7] H. Kwon and J. Bae, A hybrid quantum-classical approach to mitigating measurement errors in quantum algorithms, *IEEE Trans. Comput.* **70**, 1401 (2021).
 - [8] P. D. Nation, H. Kang, N. Sundaresan, and J. M. Gambetta, Scalable mitigation of measurement errors on quantum computers, *PRX Quantum* **2**, 040326 (2021).
 - [9] M. R. Geller, Rigorous measurement error correction, *Quantum Sci. Technol.* **5**, 03LT01 (2020).
 - [10] M. R. Geller and M. Sun, Efficient correction of multiqubit measurement errors, *Quantum Sci. Technol.* **6**, 025009 (2021).
 - [11] F. B. Maciejewski, F. Baccari, Z. Zimborás, and M. Oszmaniec, Modeling and mitigation of cross-talk effects in readout noise with applications to the quantum approximate optimization algorithm, *Quantum* **5**, 464 (2021).
 - [12] M. P. Harrigan, *et al.*, Quantum approximate optimization of non-planar graph problems on a planar superconducting processor, *Nat. Phys.* **17**, 332 (2021).
 - [13] R. Hicks, C. W. Bauer, and B. Nachman, Readout rebalancing for near-term quantum computers, *Phys. Rev. A* **103**, 022407 (2021).
 - [14] A. W. R. Smith, K. E. Khosla, C. N. Self, and M. S. Kim, Qubit readout error mitigation with bit-flip averaging, *Sci. Adv.* **7**, eabi8009 (2021).
 - [15] E. van den Berg, Z. K. Mineev, and K. Temme, Model-free readout-error mitigation for quantum expectation values, *Phys. Rev. A* **105**, 032620 (2022).
 - [16] S. Tang, C. Zheng, and K. Wang, Detecting and eliminating quantum noise of quantum measurements, *Phys. Scr.* **99**, 105129 (2024).
 - [17] M. R. Geller and Z. Zhou, Efficient error models for fault-tolerant architectures and the Pauli twirling approximation, *Phys. Rev. A* **88**, 012314 (2013).
 - [18] J. J. Wallman and J. Emerson, Noise tailoring for scalable quantum computation via randomized compiling, *Phys. Rev. A* **94**, 052325 (2016).
 - [19] A. Erhard, *et al.*, Characterizing large-scale quantum computers via cycle benchmarking, *Nat. Commun.* **10**, 5347 (2019).
 - [20] A. Hashim, *et al.*, Randomized compiling for scalable quantum computing on a noisy superconducting quantum processor, *Phys. Rev. X* **11**, 041039 (2021).
 - [21] Y. Kim, C. J. Wood, T. J. Yoder, S. T. Merkel, J. M. Gambetta, K. Temme, and A. Kandala, Scalable error mitigation for noisy

- quantum circuits produces competitive expectation values, *Nat. Phys.* **19**, 752 (2023).
- [22] Z. Cai and S. C. Benjamin, Constructing smaller Pauli twirling sets for arbitrary error channels, *Sci. Rep.* **9**, 11281 (2019).
- [23] H.-L. Huang and C. Ding, Quantum state compression shadow, *Eur. Phys. J.: Spec. Top.* **234**, 6203 (2025).
- [24] C. Ding, X.-Y. Xu, Y.-F. Niu, S. Zhang, W.-S. Bao, and H.-L. Huang, Noise-resistant quantum state compression readout, *Sci. China Phys. Mech. Astron.* **66**, 230311 (2023).
- [25] S. Endo, S. C. Benjamin, and Y. Li, Practical quantum error mitigation for near-future applications, *Phys. Rev. X* **8**, 031027 (2018).
- [26] D. Greenbaum, Introduction to quantum gate set tomography, [arXiv:1509.02921](https://arxiv.org/abs/1509.02921).
- [27] F. Arute, *et al.*, Quantum supremacy using a programmable superconducting processor, *Nature (London)* **574**, 505 (2019).
- [28] Y. Wu, *et al.*, Strong quantum computational advantage using a superconducting quantum processor, *Phys. Rev. Lett.* **127**, 180501 (2021).
- [29] A. Morvan, *et al.*, Phase transitions in random circuit sampling, *Nature (London)* **634**, 328 (2024).
- [30] Q. Zhu, *et al.*, Quantum computational advantage via 60-qubit 24-cycle random circuit sampling, *Sci. Bull.* **67**, 240 (2022).