LEARNING STRUCTURED CORRECTIONS THROUGH STATE PERTURBATIONS FOR QUANTUM ERROR MITIGATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Noise remains a major obstacle to the reliable execution of algorithms on quantum devices, motivating the development of quantum error mitigation (QEM) methods. Existing learning-based QEM approaches often learn a direct mapping from noisy circuit observations to mitigated measurement expectation values and do not explicitly exploit the underlying structure of the mitigation problem. In this work, we introduce a different QEM framework based on pseudo perturbations of the input quantum state. By expanding the perturbation in a Pauli basis, this QEM is casted as a structured prediction problem, for which we develop a neural predictor that infers perturbation coefficients rather than directly predicting the mitigated output. To address the exponential number of Pauli basis, we derive a Pauli propagation based selection procedure. This approach identifies informative perturbation directions and constructs a reduced perturbation subspace. Experiments across diverse quantum tasks and noise settings show that our proposed framework outperform against state-of-the-art baselines.

1 INTRODUCTION

Quantum computing is an emerging computational paradigm with the potential to address a wide range of problems Shor (1999); Grover (1996); Bravo-Prieto et al. (2023); Peruzzo et al. (2014). Many quantum algorithms require the use of numerous qubits and quantum gates, thereby demanding highly reliable execution of these operations. However, existing quantum hardware, commonly referred to as Noisy Intermediate-Scale Quantum (NISQ) devices Preskill (2018), is inherently susceptible to noise, leading to imperfect and unreliable outcomes. These noise degrades the fidelity of quantum state and negatively affects the performance of quantum algorithm. To alleviate this, Quantum Error Mitigation (QEM) emerges as a prominent research direction.

Established quantum error mitigation (QEM) methods include Zero-Noise Extrapolation (ZNE) Giurgica-Tiron et al. (2020), which extrapolates expectation values measured at multiple noise levels to the zero-noise limit; Clifford Data Regression (CDR) Czarnik et al. (2021), which learns a correction from classically simulable near-Clifford circuits; and Probabilistic Error Cancellation (PEC) Temme et al. (2017), which uses quasiprobability sampling of noisy circuits to cancel errors under a characterized noise model. Despite their effectiveness, these methods face practical limitations. ZNE can amplify finite-shot uncertainty through extrapolation, CDR depends on the representativeness of its near-Clifford training circuits, and PEC requires accurate noise characterization. Recent studies have explored machine learning (ML) for QEM Kim et al. (2020); Liao et al. (2024); Bao et al. (2025); Liao et al. (2025). Liao et al. Liao et al. (2024) investigate several learning architectures, including graph neural networks (GNNs), which encode circuit topology through message passing. However, capturing long-range dependencies through local message passing can require multiple propagation layers. To address this limitation, Bao et al. Bao et al. (2025) introduce a non-message-passing graph transformer that combines structural encoding with global attention to model dependencies across quantum circuits.

Another challenge in learning-based QEM is obtaining ideal training targets for circuits that are difficult to simulate classically. Bao et al. Bao et al. (2025) propose circuit inverse composition, inspired by mirror-circuit benchmarking Proctor et al. (2022), to construct training circuits with

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

known ideal outputs. Similarly, Liao et al. Liao et al. (2025) introduce a data augmentation strategy based on fiducial processes, replacing single-qubit gates with cancelling gate pairs while preserving the circuit's entangling-gate structure.

Despite these advances, existing learning-based QEM methods predominantly learn direct mappings from noisy observations to mitigated expectation values, without explicitly exploiting the structure of observable-dependent corrections. Furthermore, ML based method (Liao et al., 2024; Bao et al., 2025) relies on circuit encoding which limit their applicability. To address this limitation, we introduce **QEM-STEP (Quantum Error Mitigation using STructured statE Perturbation)**, a framework that starts from the noisy measurement as its baseline and learns structured corrections through perturbations expanded in the Pauli basis. To make this formulation computationally feasible, we develop a principled criterion for selecting a truncated Pauli basis that balances approximation error, finite-shot uncertainty, and discrepancies between ideal and noisy responses.

The training-data generation strategies in Bao et al. (2025); Liao et al. (2025) rely on circuit inversion or cancelling gate pairs, which impose structural constraints on the generated training circuits. To alleviate this, we propose two data-generation strategies based on Pauli propagation and near-Clifford state construction. The former constructs classically feasible training circuits, while the latter generates pairs of input states for the swap test. These strategies aim to preserve the structural diversity of training circuits while enabling efficient classical computation of ideal training labels.

Overall, our main contributions are as follows:

- We introduce a quantum error mitigation framework based on pseudo perturbations of the input quantum state. By decomposing the perturbation into a Pauli basis, we cast mitigation as a regularized projection problem over Pauli response coefficients and learn them with a neural predictor. The resulting formulation is flexible and supports both circuit-based implementations and more general quantum evolutions.
- We develop a classically pragmatic method for selecting informative perturbation directions using Pauli propagation. The resulting support set captures the dominant Pauli components while controlling the approximation error, enabling efficient basis selection and reducing the projection problem in our perturbation framework from the exponential Pauli space to a practical subspace.
- To support the proposed perturbation formulation and Pauli support selection when ideal training targets are unavailable, we develop simulatable training-data generation strategies. By constructing quantum evolutions whose ideal expectation values remain tractable to simulate, the generated data provides paired noisy and ideal targets needed to train the proposed mitigation model. Numerical experiments further corroborate our findings.

2 PRELIMINARIES

2.1 QUANTUM COMPUTING

Let $\mathcal{H} = \mathbb{C}^d$ be a finite-dimensional complex Hilbert space of n-qubit system with $\dim(\mathcal{H}) = d = 2^n$. We denote by $\mathcal{B}(\mathcal{H})$ the vector space of all linear operators acting on $\mathcal{H}$, thus $\mathcal{B}(\mathcal{H}) \cong \mathbb{C}^{d\times d}$. Given $B \in \mathbb{C}^{d\times d}$, we denote its conjugate transpose as $B^\dagger$. In particular, $\|B\|_\infty$ is the spectral norm, $\|B\|_2 = \|B\|_{\text{HS}} = \sqrt{\text{Tr}(B^\dagger B)}$ is the Hilbert-Schidt norm and $\|B\|_1$ is the trace norm. In addition, we denote the set of all Hermitian (self-adjoint) matrices by $\text{Herm}(d) = \{B \in \mathbb{C}^{d\times d} \mid B = B^\dagger\}$ and set of all valid density operators is denoted by $\mathcal{D}(d) = \{\rho \in \mathbb{C}^{d\times d} \mid \rho \succeq 0, \text{Tr}(\rho) = 1\} \subset \text{Herm}(d)$. For any linear superoperator $\Phi : \mathcal{B}(\mathcal{H}) \to \mathcal{B}(\mathcal{H})$, there is a unique adjoint $\Phi^\dagger$ so that $\text{Tr}(A\Phi(B)) = \text{Tr}(B\Phi^\dagger(A))$ for any $A, B \in \mathcal{B}(\mathcal{H})$. Note that $(\Phi_1 \circ \Phi_2)^\dagger = \Phi_2^\dagger \circ \Phi_1^\dagger$.

Quantum evolution Let $\mathcal{U} : \mathcal{B}(\mathcal{H}) \to \mathcal{B}(\mathcal{H})$ be an unitary quantum evolution acting on quantum states ρ, defined as $\mathcal{U}[\rho] = U\rho U^\dagger$ where $U \in \mathbb{C}^{d\times d}$ and $UU^\dagger = I$. A quantum circuit is a particular realization of such an evolution obtained by composing L layers as

$$\boldsymbol{\mathcal{U}} = \mathcal{U}_L \circ \mathcal{U}_{L-1} \circ \cdots \circ \mathcal{U}_1, \tag{1}$$

which maps any input state ρ to a output state $\boldsymbol{\mathcal{U}}[\rho] = \boldsymbol{U}\rho\boldsymbol{U}^\dagger \in \mathcal{D}(d)$ where $\boldsymbol{U} = U_L \cdot U_{L-1} \cdots U_1$. We denote the Heisenberg picture of the unitary in equation 1 as $\boldsymbol{\mathcal{U}}^\dagger[O]$ where $\mathcal{U}_l^\dagger[O] = U_l^\dagger O U_l$.

Measurement To compute outcome of a quantum evolution, we compute the expectation value of an observable $O \in \mathrm{Herm}(d)$ with respect to the input quantum state ρ. The ideal expectation value of O is defined as: $\ell(O, \mathcal{U}, \rho) = \mathrm{Tr}\left(O\boldsymbol{\mathcal{U}}[\rho]\right)$, which represents the expected measurement outcome of the observable after applying the unitary evolution.

Pauli string Let $\mathcal{P}_n = \{I, X, Y, Z\}^{\otimes n}$ denote the set of n-qubit Pauli strings. For Pauli string $P = Q_1 \otimes \cdots \otimes Q_n$ where $Q_j \in \{I, X, Y, Z\}$, its Pauli weight is defined as $\mathrm{wt}(P) = |\{j \in [n] : Q_j \neq I\}|$, i.e., the number of non-identity tensor factors. Since $\mathcal{P}^{\otimes n}$ forms an orthogonal basis for $\mathcal{B}(\mathcal{H})$, any Hermitian operator B belong to this space can be decompose as

$$B = \sum_{P_i \in \mathcal{P}_n} \alpha(B, P) P_i, \tag{2}$$

where $\alpha(B, P_i) = \frac{1}{2^n} \langle B, P_i \rangle_{\mathrm{HS}} = \frac{1}{2^n} \mathrm{Tr}(B^\dagger P_i)$ is the Hilbert–Schmidt inner product Georges et al. (2025). We further denote by $\boldsymbol{\alpha}(B) := (\alpha(B, P_i))_{i=1}^{4^n}$ the vector of Pauli coefficients of B.

2.2 Quantum Noise Channels and Error Mitigation

Noise in unitary evolution is a linear superoperator $\mathcal{E} \colon \mathcal{B}(\mathcal{H}) \to \mathcal{B}(\mathcal{H})$. Additionally, noise in quantum circuit can be modeled as the composition García-Martín et al. (2024) $\boldsymbol{\mathcal{E}} = \mathcal{E}_L \circ \mathcal{U}_L \circ \cdots \circ \mathcal{E}_1 \circ \mathcal{U}_1$, where each $\mathcal{E}_i$ is (a) $\mathcal{E}_i$ is convex-linear (i.e, $\mathcal{E}_i\left[\sum_j p_j \rho_j\right] = \sum_j p_j \mathcal{E}_i[\rho_j]$) and (b) $\mathcal{E}_i$ is completely positive (i.e., $\mathcal{E}_i[\rho] \succeq 0$ and $(\mathcal{E}_i \otimes \mathcal{I}_k)[\rho] \succeq 0$ for any $\rho \succeq 0$ and $k \geq 1$) and trace-preserving (CPTP) (i.e., $\mathrm{Tr}(\mathcal{E}_i[\rho]) = \mathrm{Tr}(\rho) = 1$ for any $\rho \in \mathcal{D}(d)$) Nielsen and Chuang (2010); Watrous (2018). Such a channel always admits a Kraus representation $\mathcal{E}_i[\rho] = \sum_k K_{i,k} \rho K_{i,k}^\dagger$, for some $K_k \in \mathcal{B}(\mathcal{H})$ satisfying the completeness relation $\sum_k K_k^\dagger K_k = I_d$. The adjoint is given by $\boldsymbol{\mathcal{E}}^\dagger = \mathcal{U}_1^\dagger \circ \mathcal{E}_1^\dagger \circ \cdots \circ \mathcal{U}_L^\dagger \circ \mathcal{E}_L^\dagger$. See Appendix B for detailed description of noise model we use in this paper. An input state ρ is now mapped to a noisy output $\boldsymbol{\mathcal{E}}[\rho]$ and the corresponding measurement in the presence of noise is $\ell(O, \mathcal{E}, \rho) = \mathrm{Tr}(O\boldsymbol{\mathcal{E}}[\rho])$.

In QEM, the goal is to estimate the ideal expectation value of a quantum circuit from noisy observations. Let $\mathcal{T} = \{(C_t, \rho_t, O_t, y_t^{\mathrm{noisy}}, y_t^{\mathrm{ideal}})\}_{t=1}^N$ be a dataset of N examples, where $C_t = (\boldsymbol{\mathcal{U}^{(t)}}, \boldsymbol{\mathcal{E}^{(t)}})$ denotes the ideal and noisy evolution, ρ_t the input state, O_t the measured observable, y_t^{noisy} the noisy expectation value ($\ell(O, \mathcal{E}, \rho)$), and y_t^{ideal} the corresponding ideal expectation value ($\ell(O, \mathcal{U}, \rho)$), when available. A learned QEM model predicts $\hat{y}_t$. based on the dataset $\mathcal{T}$. The supervised QEM objective is then

$$\min_{\phi} \frac{1}{N} \sum_{t=1}^{N} \left(y_t^{\mathrm{ideal}} - \hat{y}_t\right)^2. \tag{3}$$

3 Quantum Error Mitigation using State Perturbation

This section introduces the proposed pseudo state perturbation framework for quantum error mitigation under noisy quantum evolution. We first formalize the state perturbation formulation and its associated mitigation objective. We then develop a procedure for constructing a reduced Pauli subspace while accounting for approximation and finite-shot errors. Finally, we cast state perturbation as a regularized projection problem and introduce a neural predictor for learning its solution.

Pseudo State Perturbation Framework Firstly, the pseudo perturbation is defined as $\rho + \delta$ where δ is the perturbation quantity. We refer to $\rho + \delta$ as a pseudo perturbation because it is not required to satisfy all conditions of a valid density matrix. Under this perturbation, the resulting expectation value for input state ρ is formally given by $\mathrm{Tr}(O\boldsymbol{\mathcal{E}}[\rho + \delta])$. Our goals is to find a δ such that

$$\mathrm{Tr}(O\boldsymbol{\mathcal{E}}[\rho + \delta]) \approx \mathrm{Tr}(O\boldsymbol{\mathcal{U}}[\rho]). \tag{4}$$

The perturbation quantity δ can be expanded into operation basis $\sum_i q_i P_i$ where $P_i \in \mathcal{P}_n$ is the set of Pauli basis and setting the perturb representation $c_i := \mathrm{Tr}(O\boldsymbol{\mathcal{E}}[P_i])$. By linearity of $\boldsymbol{\mathcal{E}}$ and the trace

$$\mathrm{Tr}(O\boldsymbol{\mathcal{E}}[\rho + \delta]) = \mathrm{Tr}(O\boldsymbol{\mathcal{E}}[\rho]) + \sum_i q_i \mathrm{Tr}(O\boldsymbol{\mathcal{E}}[P_i]) = \mathrm{Tr}(O\boldsymbol{\mathcal{E}}[\rho]) + \sum_i q_i c_i. \tag{5}$$

Equation 5 reduces the effect of the pseudo perturbation to a linear combination of Pauli response. QEM is therefore reduced to two problems: determining an appropriate perturbation coefficient vector $\boldsymbol{q}$ (Section 3.2), and constructing the perturb representation vector $\boldsymbol{c}$ (Section 3.1).

3.1 Construction of the perturb representation

The formulation in equation 5 assumes access to the perturb representation vector $\boldsymbol{c}$. However, an n-qubit system contains 4^n Pauli directions, and each perturb direction c_i must be estimated through noisy quantum measurements. To make this estimation practical, we restrict the perturbation to a support set $\mathcal{S}_K \subseteq \mathcal{P}_n$ containing K Pauli directions, where $K \ll 4^n$ denotes the support size. We next show how individual perturb directions can be physically estimated and how to identify an informative support $\mathcal{S}_K$ through Pauli propagation.

3.1.1 Physical implementation of Pauli response

We consider only nonidentity Pauli basis elements $P_i \in \mathcal{P}_n \setminus \{I^{\otimes n}\}$. For Pauli basis operators, direct evaluation of $\boldsymbol{c}$ is generally not physically realizable because P_i is not a valid density matrix. To overcome this issue, we decompose P_i into projectors onto its two eigenspaces. Let Π_+ and Π_- denote the projectors onto the $+1$ and -1 eigenspaces of P_i, respectively. Then

$$\mathrm{Tr}(O\boldsymbol{\mathcal{E}}[P_i]) = \mathrm{Tr}(O\boldsymbol{\mathcal{E}}[\Pi_+]) - \mathrm{Tr}(O\boldsymbol{\mathcal{E}}[\Pi_-]). \tag{6}$$

Here, $\Pi_\pm = \sum_k |\psi_{k,\pm}\rangle \langle\psi_{k,\pm}|$, where each $|\psi_{k,\pm}\rangle$ is an eigenvector of P_i satisfying $P_i |\psi_{k,\pm}\rangle = \pm |\psi_{k,\pm}\rangle$. Equivalently, the projectors can be written as $\Pi_\pm = \frac{\mathcal{I} \pm P_i}{2}$.

Although $\Pi_\pm$ are valid positive semidefinite projectors, they are not normalized density matrices for $n > 1$. To obtain valid quantum states, we normalize them as $\rho_{i,\pm} = \frac{\Pi_\pm}{2^{n-1}}$. Substituting these normalized states into equation 6 yields

$$\mathrm{Tr}(O\boldsymbol{\mathcal{E}}[P_i]) = 2^{n-1} \left(\mathrm{Tr}(O\boldsymbol{\mathcal{E}}[\rho_{i,+}]) - \mathrm{Tr}(O\boldsymbol{\mathcal{E}}[\rho_{i,-}])\right). \tag{7}$$

Lemma 3.1 (Finite-shot variance of the perturb representation) *Let $\hat{c}_i$ be the estimator obtained from equation 7, using M independent shots for each of $\rho_{i,+}$ and $\rho_{i,-}$, with the two shot sets sampled independently. If $\|O\|_\infty \leq 1$, then*

$$\sigma_{c_i} := \sqrt{\mathrm{Var}(\hat{c}_i)} \leq \frac{2^{n-\frac{1}{2}}}{\sqrt{M}}. \tag{8}$$

3.1.2 Pauli propagation based support selection

Although each perturb representation can be estimated physically, evaluating all 4^n Pauli directions is impractical. We therefore seek a reduced support containing the most informative directions. Firstly, we relate perturb representation vector to the Pauli coefficients of the observable in Heisenberg picture.

Remark 3.2 (Perturb representation and Pauli coefficient equivalence) *Given noisy quantum evolution $\boldsymbol{\mathcal{E}}$ and Pauli observable O, the perturbation representation vector satisfies $\boldsymbol{c} = 2^n \boldsymbol{\alpha}\left(\boldsymbol{\mathcal{E}}^\dagger[O]\right)$. Let $\hat{\alpha}_i = \frac{\hat{c}_i}{2^n}$ denote the finite-shot estimator of $\alpha\left(\boldsymbol{\mathcal{E}}^\dagger[O], P_i\right)$. By Lemma 3.1,*

$$\sigma_{\alpha,i} := \sqrt{\mathrm{Var}(\hat{\alpha}_i)} \leq \frac{1}{\sqrt{2M}}. \tag{9}$$

Remark 3.2 establishes that the perturb representation and the Pauli coefficient of the noisy Heisenberg picture observable differ only by a global factor 2^n. Therefore, selecting a support that controls the discarded Pauli mass of $\boldsymbol{\mathcal{E}}^\dagger[O]$ is equivalent to controlling the discarded perturb direction. For Hermitian operator B, given the definition of Pauli coefficient α in equation 2, let $\hat{\alpha}(.) = \alpha(.) + \zeta$ denote an estimate obtained from M measurement shots where ζ is its finite-shot error. We next derive an upper bound on the approximation error by truncation and finite-shot error using the discarded tail term (Shao et al., 2025) $\Delta(K) := \sum_{j=K+1}^{4^n} |\alpha(B, P_j)|^2$.

Theorem 3.3 (Approximation and finite-shot error) *Let $\boldsymbol{\mathcal{U}} = \mathcal{U}_L \circ \cdots \circ \mathcal{U}_1$ be an n-qubit unitary channel, $\boldsymbol{\mathcal{E}}$ be its noisy implementation, and O be a Pauli observable. Initialize $\widetilde{O}_L = O$ and, for $l = L, \ldots, 1$, define $A_{l-1} = \mathcal{U}_l^\dagger[\widetilde{O}_l]$, and $\widetilde{O}_{l-1} = \text{Top-}K(A_{l-1})$, where Top-$K$ retains the K Pauli terms with the largest coefficient magnitudes (see equation 2). We define $\Delta_{l-1}(K) := \sum_{j>K} |\alpha(A_{l-1}, P_j)|^2$, and $E(K) := \sum_{l=1}^{L} \sqrt{\Delta_{l-1}(K)}$, where the Pauli strings are ordered so that $|\alpha(A_{l-1}, P_1)| \geq |\alpha(A_{l-1}, P_2)| \geq \cdots \geq |\alpha(A_{l-1}, P_{4^n})|$. Then*

$$\mathbb{E}_\zeta \left[\left\| \boldsymbol{\alpha}\left(\boldsymbol{\mathcal{E}}^\dagger[O]\right) - \widehat{\boldsymbol{\alpha}}_K\left(\boldsymbol{\mathcal{E}}^\dagger[O]\right) \right\|_2^2 \right] \leq 2 \left\| \boldsymbol{\alpha}\left(\boldsymbol{\mathcal{E}}^\dagger[O]\right) - \boldsymbol{\alpha}\left(\boldsymbol{\mathcal{U}}^\dagger[O]\right) \right\|_2^2 + 2E(K)^2 + \frac{K}{2M}. \quad (10)$$

The proof of this Theorem is provided in Appendix D.3. While Shao et al. (2025) bound the error of a single Top-K truncation, our theorem accounts for errors accumulated across sequential propagation steps and finite-shot estimation. It thus captures the trade-off between retaining more Pauli terms and estimating more coefficients with a fixed measurement budget. Since $\boldsymbol{c}$ and $\boldsymbol{\alpha}(\boldsymbol{\mathcal{E}}^\dagger(O))$ differ only by a constant scaling factor, and $\left\| \boldsymbol{\alpha}\left(\boldsymbol{\mathcal{E}}^\dagger[O]\right) - \boldsymbol{\alpha}\left(\boldsymbol{\mathcal{U}}^\dagger[O]\right) \right\|$ is independent of K, Theorem 3.3 provides a criterion for selecting the support size. We use Pauli propagation to select the support by setting the truncation level to

$$K^* = \arg\min_K \{2E(K)^2 + \frac{K}{2M}\}$$

The detail of Pauli propagation algorithm is presented in Appendix E.2.

3.2 Regularized projection problem

For a dataset $\mathcal{T} = \{(C_t, \rho_t, O_t, y_t^{\text{noisy}}, y_t^{\text{ideal}})\}_{t=1}^N$, each instance is assigned its own perturbation operator δ_t. The corrected prediction for the t-th example becomes $y_t^{\text{noisy}} + \boldsymbol{q}_t^\top \boldsymbol{c}_t$. In this formulation, the component of perturb representation vector $\boldsymbol{c}_t$ are restricted to that of $\mathcal{S}_{K^*}$ as selected using the procedure in Section 3.1. To solve this problem, the objective in equation 3 is converted to finding

$$\{\boldsymbol{q}_t^*\}_{t=1}^N = \arg\min_{\{\boldsymbol{q}_t\}_{t=1}^N} \frac{1}{N} \sum_{t=1}^N \left(y_t^{\text{ideal}} - y_t^{\text{noisy}} - \boldsymbol{q}_t^\top \boldsymbol{c}_t \right)^2 + \frac{\beta}{N} \sum_{t=1}^N \|\boldsymbol{q}_t\|_2^2. \quad (11)$$

For each instance t, the corresponding coefficient vector is obtained by solving

$$\boldsymbol{q}_t^* = \arg\min_{\boldsymbol{q}_t} \left(y_t^{\text{ideal}} - y_t^{\text{noisy}} - \boldsymbol{q}_t^\top \boldsymbol{c}_t \right)^2 + \beta \|\boldsymbol{q}_t\|_2^2.$$

This gives the closed-form solution

$$\boldsymbol{q}_t^* = \frac{\left(y_t^{\text{ideal}} - y_t^{\text{noisy}} \right)}{\|\boldsymbol{c}_t\|_2^2 + \beta} \boldsymbol{c}_t. \quad (12)$$

Although 12 provides the optimal perturbation for a training instance with known y_t^{ideal}, it cannot be use for an unseen noisy circuit. We therefore learn a neural predictor g_ϕ parameterized by ϕ that maps perturb representation to the corresponding perturbation coefficients. We define $\hat{\boldsymbol{q}}_t = g_\phi(y_t^{\text{noisy}}, \boldsymbol{c}_t, \boldsymbol{p})$, where $\boldsymbol{p}_j$ denote the one-hot encoding of P_j. The corrected prediction become $y_t^{\text{noisy}} + \hat{\boldsymbol{q}}_t^\top \boldsymbol{c}_t$. The neural perturbation model is trained by minimizing

$$\phi^* = \arg\min_\phi \frac{1}{N} \sum_{t=1}^N \left(y_t^{\text{ideal}} - y_t^{\text{noisy}} - g_\phi(y_t^{\text{noisy}}, \boldsymbol{c}_t, \boldsymbol{p})^\top \boldsymbol{c}_t \right)^2 + \beta \frac{1}{N} \sum_{t=1}^N \|\hat{\boldsymbol{q}}_t\|_2^2.$$

Now we turn to constructing the neural architecture for solving equation 11. For each instance, the model takes as input the noisy expectation y_t^{noisy}, the one-hot encoding of each selected Pauli basis $P_j \in \mathcal{S}_K$, the corresponding perturbation response coefficient $c_{t,j}$, and the response-vector norm $\|\boldsymbol{c}_t\|_2$. The model predicts one coefficient $\hat{q}_{t,j}$ for each $P_j \in \mathcal{S}_K$, which are collected into the coefficient vector $\hat{\boldsymbol{q}}_t = (\hat{q}_{t,1}, \ldots, \hat{q}_{t,K})^\top$.

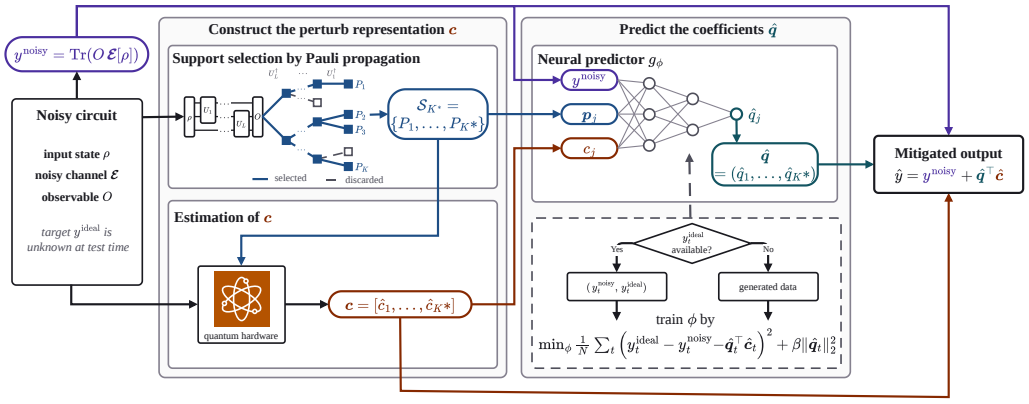


Figure 1: Overview of the proposed QEM-STEP framework. The perturbation representation is constructed by selecting compact Pauli directions via propagation and estimating their effects on quantum hardware. A neural predictor, trained using either available ideal targets or generated data, predicts the perturbation coefficients and produces the mitigated expectation value.

Perturbation representation context (PR context) Since each coefficient is associated with a particular Pauli basis, we first construct a global context that summarizes the selected perturbation responses. Let $\boldsymbol{p}_j$ denote the one-hot encoding of P_j. For each $P_j \in \mathcal{S}_K$, the perturbation representation context is obtained by

$$\boldsymbol{u}_{t,j} = \sigma(W_{\text{context}}[\boldsymbol{p}_j \,\|\, c_{t,j}] + \boldsymbol{b}_{\text{context}})\,, \qquad \boldsymbol{c}_t^{\text{context}} = \frac{1}{K}\sum_{P_j \in \mathcal{S}_K} \boldsymbol{u}_{t,j},$$

where σ denotes the ReLU activation and $|$ denotes concatenation. For each selected Pauli basis P_j, we then concatenate the local features with this shared context, $\boldsymbol{x}_{t,j} = \left[y_t^{\text{noisy}} \,|\, \boldsymbol{p}_j \,|\, c_{t,j} \,|\, \|\boldsymbol{c}_t\|_2 \,|\, \boldsymbol{c}_t^{\text{context}}\right]$. A shared linear transformation with ReLU activation is then applied to obtain the Pauli-specific hidden representation $\boldsymbol{h}_{t,j} = \sigma(W_h \boldsymbol{x}_{t,j} + \boldsymbol{b}_h)\,,$

Shared coefficient context (SC context) To allow the prediction of each coefficient to depend on the learned representations of the other selected Pauli bases, we further aggregate the hidden representations through mean pooling, $\bar{\boldsymbol{h}}_t = \frac{1}{K}\sum_{P_j \in \mathcal{S}_K} \boldsymbol{h}_{t,j}$. The resulting shared context is concatenated with each local hidden representation and passed through a final linear layer, $\hat{q}_{t,j} = W_{\text{out}}\left[\boldsymbol{h}_{t,j} \,\|\, \bar{\boldsymbol{h}}_t\right] + b_{\text{out}}$. Thus, each predicted coefficient incorporates both Pauli-specific information and a summary of the other coefficients in the selected perturbation support. We initialize the weights (W_{out}) and bias (b_{out}) of the final output layer to zero, so the model initially starts from the noisy baseline. The predicted coefficients are assembled into $\hat{\boldsymbol{q}}_t$ and used in the mitigation rule $\hat{y}_t = y_t^{\text{noisy}} + \hat{\boldsymbol{q}}_t^\top \boldsymbol{c}_t$.

4 Numerical Experiments

We first describe the experimental setup across the considered QEM tasks and introduce the comparison baselines. We then evaluate QEM-STEP against these baselines on the standard benchmarks. After that, we introduce two training-data generation strategies for settings in which ideal expectation values are unavailable during training.

4.1 Experiment Setup

We evaluate QEM-STEP across a diverse set of QEM tasks. We first consider standard circuit-based benchmarks, including Trotterized dynamics of the transverse-field Ising model (TIFM) and randomly generated quantum circuits. We further evaluate QEM-STEP under several challenging settings, including circuit-scale extrapolation to out-of-distribution circuits, noise-scale amplification which mimics ZNE, the swap-test benchmark of Liao et al. (2025), and coherent noise. Finally,

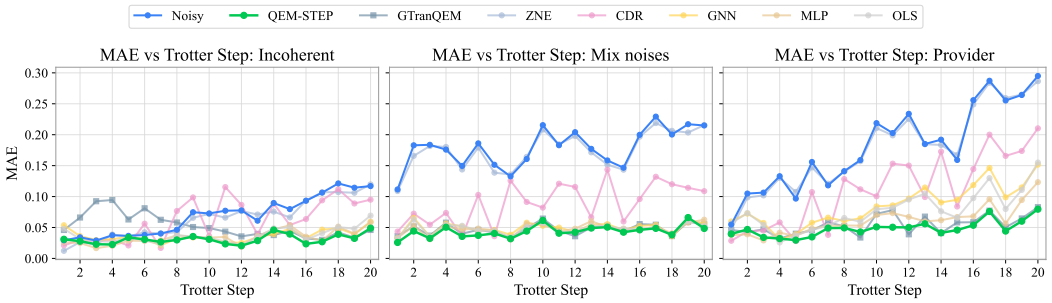


Figure 2: MAE of QEM-STEP, baseline methods, and unmitigated noisy measurements versus the number of Trotter steps. Results are shown for All-6-Trotter under three noise models in the standard setting in Table 1.

Table 1: Comparison of QEM methods across different circuit and noise settings. The best value in each column are bold and highlighted in gray while the second best is underlined.

Method	All-6-Trotter						Max-6-Random					
	Incoherent		Mixed Noise		Fake Provider		Incoherent		Mixed Noise		Fake Provider	
	MAE ↓	IR ↑	MAE ↓	IR ↑	MAE ↓	IR ↑	MAE ↓	IR ↑	MAE ↓	IR ↑	MAE ↓	IR ↑
Raw Noisy	0.0687	–	0.1790	–	0.1809	–	0.0368	–	0.0915	–	0.0430	–
ZNE	0.0458	0.5933	0.1108	0.8000	0.1273	0.7633	0.0407	0.3367	0.0635	0.6567	0.0375	0.5000
CDR	0.0625	0.6300	0.0897	0.7467	0.1121	0.7500	0.0324	0.5300	0.0463	**0.6933**	0.0315	0.6167
OLS	0.0374	0.6500	0.0479	0.7867	0.0773	0.7500	0.0357	0.4667	0.0418	0.6367	0.0295	0.5533
MLP	0.0342	0.7000	0.0483	0.7767	0.0614	0.7367	0.0350	0.4700	0.0418	0.6333	0.0299	0.5467
GNN	0.0368	0.6433	0.0489	0.7800	0.0852	0.7200	0.0336	0.5133	0.0419	0.6367	0.0294	0.5433
GTranQEM	0.0524	0.5533	0.0464	**0.8233**	0.0543	0.7500	0.0311	**0.5333**	0.0395	0.6633	0.0300	0.5367
QEM-STEP	**0.0310**	**0.7233**	**0.0438**	0.8067	**0.0483**	**0.7933**	**0.0283**	0.5133	**0.0373**	0.6833	**0.0265**	**0.6233**

to assess the applicability of QEM-STEP beyond the circuit-based model, we consider the many-body dynamics experiment of Liao et al. (2025). More details of the benchmark are provided in Appendix C. Across these experiments, we compare QEM-STEP against a range of QEM baselines, including ZNE Giurgica-Tiron et al. (2020), CDR Czarnik et al. (2021), GNN, MLP, and OLS Liao et al. (2024), GTranQEM Bao et al. (2025), and DAEM Liao et al. (2025).

4.2 Standard Setting

This setting focuses on two types of circuits: Trotterized circuits and random circuits. The former benchmark generates structured circuits following the Trotterized dynamics of the one-dimensional TFIM Chakrabarti et al. (1996). Meanwhile, the latter generates circuits with random structures to evaluate the generality of QEM-STEP. Specifically, we randomly generate circuits with at most 6 qubits and 150 gates. We consider three noise settings for both benchmarks. The incoherent-noise setting uses depolarizing errors and Pauli-X, while the mixed-noise setting combines Pauli-X errors, depolarizing errors, and readout errors. We additionally consider IBM's `FakeHanoiV2` backend which mimic the noise profiles in real hardware. The generators for both benchmarks are based on the implementation of Bao et al. (2025) and additional description is in Appendix C.1.

Figure 2 shows that QEM-STEP maintains low MAE as the Trotter steps increases across all three noise models, demonstrating robustness to increasing circuit depth. Table 1 summarizes the results across all six combinations of circuit and noise settings. QEM-STEP achieves the lowest MAE in every setting, demonstrating consistent improvements in mitigation accuracy over the baselines. It also remains competitive in terms of IR, ranking first in three settings and second in two.

4.3 Not-available Ideal Expectation Value Settings

This setting assumes that ideal expectation values are unavailable for the training data. The first benchmark follows Bao et al. (2025) where the test circuits are generated using the same Trotterized construction as in the standard setting, but no training circuits with ideal targets are provided. We

additionally evaluate on the swap-test benchmark of Liao et al. (2025). The swap test estimates the overlap between two quantum states by applying controlled-SWAP operations to the input registers and measuring an ancillary control qubit. We consider pairs of five-qubit states, with noise introduced before each controlled-SWAP operation. Because expectation values on quantum hardware are estimated from a finite number of measurement shots Han et al. (2025); Endo et al. (2018), we evaluate finite-shot estimates alongside the exact expectation value simulations of Liao et al. (2025). To enable supervised training in the absence of ideal training targets, we introduce two strategies for generating classically tractable training data. We compare our strategies with that originally proposed in Bao et al. (2025); Liao et al. (2025).

Iterative Gate Addition We iteratively construct each circuit while exactly propagating the target observable backward in the Heisenberg picture using Pauli propagation Rudolph et al. (2025). Consistent with the preceding notation, we initialize $O_L = O$ and update $O_{l-1} = \mathcal{U}_l^\dagger[O_l]$, At each step, the evolved observable admits the Pauli expansion $O_l = \sum_{P_i \in \mathcal{P}_n} \alpha(O_l, P_i) P_i$, where the number of nonzero Pauli coefficients is K_l. Although K_l can grow exponentially with n, the cost of each Pauli propagation step is polynomial in the input size and linear in the number of retained Pauli terms Shao et al. (2025). We therefore retain a candidate gate only if the resulting circuit satisfies $K_l \leq \mathrm{poly}(n, |\boldsymbol{\mathcal{U}}|)$ where $|\boldsymbol{\mathcal{U}}|$ denotes the number of gates in the circuit; otherwise, construction is terminated. Hence, the resulting ideal expectation value can be evaluated exactly with polynomial classical resources.

Near-Clifford State Pair The original swap-test experiments in Liao et al. (2025) do not provide training ideal expectation values. We therefore introduce a fixed-circuit data-generation strategy that constructs pairs of near-Clifford states with a restricted number of non-Clifford gates. Such circuits admit efficient classical simulation when the non-Clifford content is sufficiently limited Aaronson and Gottesman (2004); Bravyi and Gosset (2016); Bravyi et al. (2019). In particular, this allows both the generated states and their pairwise fidelity to be computed efficiently. This construction enables us to generate diverse pairs of input states while retaining classically polynomial time ideal targets for supervised training.

Table 2: Performance comparison without ideal training expectation values using Inverse-Circuit and Iterative Gate Addition training-data generation strategies. Test circuits follow the 6-qubit Trotterized mixed-noise setting. Bold gray cells indicate the overall best result for each metric, while underlining is the second best. Raw Noisy, ZNE, and CDR results are shared across strategies.

Metric	Strategy	Raw Noisy	ZNE	CDR	OLS	MLP	GNN	GTranQEM	QEM-STEP
MAE ↓	Inverse-Circuit	0.1485	0.0909	0.0974	0.0733	0.1052	0.1068	0.0719	–
	Iterative Gate Addition				0.0414	0.0396	0.0404	<u>0.0382</u>	**0.0363**
IR ↑	Inverse-Circuit	–	0.7567	0.6000	0.7600	0.6933	**0.8067**	0.6833	–
	Iterative Gate Addition				0.7233	0.7267	0.7233	<u>0.8033</u>	0.7733

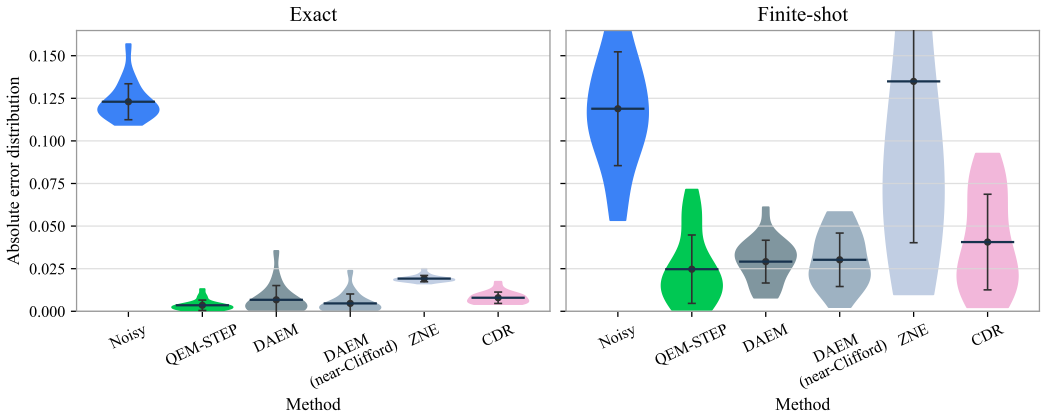


Figure 3: Mitigation performance of QEM methods on swap-test setting.The x-axis displays different QEM methods,while the y-axis denotes absolute error distribution

Table 2 evaluates the methods on the six-qubit Trotter benchmark under mixed noise. Compared with inverse-circuit training data, our iterative gate-addition strategy substantially reduces the MAE of every learning-based baseline. Using the generated training data, QEM-STEP achieves the lowest overall MAE of 0.0363. QEM-STEP also obtains the second-highest IR, trailing only GTranQEM. Figure 3 reports the results on the swap-test benchmark. Under exact expectation values, replacing the original DAEM training scheme with our near-Clifford state pairs reduces its MAE from 0.0061 to 0.0046. QEM-STEP further achieves the lowest MAE of 0.0035. Under finite-shot measurements, QEM-STEP again obtains the lowest MAE. These results demonstrate that our data-generation strategy provides more effective supervision when ideal training targets are unavailable.

4.4 Additional result

Finally, we evaluate QEM-STEP beyond the circuit model using the many-body dynamics benchmark of Liao et al. (2025). This experiment considers the real-time dynamics of a 50-qubit transverse-field Ising system, with the evolution generated directly by the Ising Hamiltonian rather than by an explicit gate-based circuit. Noise is applied after the unitary evolution, and the mitigation task targets expectation values of two-qubit Pauli observables on nearest-neighbor qubits under phase-damping and amplitude-damping noise. Similar to the swap-test we also evaluate finite-shot estimates alongside the exact expectation value simulations of Liao et al. (2025). The implementation for these benchmark is based on Liao et al. (2025).

Table 3: MAE and IR for different methods under the Many-body setting. The best value in each column are bold and highlighted in gray while the second best is underlined

Method	Many-body				Many-body (finite-shot)			
	Amplitude Damping		Dephasing Damping		Amplitude Damping		Dephasing Damping	
	MAE ↓	IR ↑	MAE ↓	IR ↑	MAE ↓	IR ↑	MAE ↓	IR ↑
Raw Noisy	0.0138	–	0.0138	–	<u>0.0282</u>	–	0.0304	–
ZNE	**0.0000**	**0.9998**	**0.0013**	**0.8889**	0.0341	0.4222	0.0311	0.4173
DAEM	<u>0.0005</u>	0.8952	<u>0.0015</u>	0.6707	0.0503	<u>0.4780</u>	<u>0.0273</u>	<u>0.6669</u>
QEM-STEP	0.0047	<u>0.8956</u>	0.0059	<u>0.6949</u>	**0.0140**	**0.8381**	**0.0148**	**0.8551**

Table 3 compares QEM-STEP with the baselines on the many-body dynamics benchmark. Although ZNE performs better when exact expectation values are available, QEM-STEP demonstrates a clear advantage in the more realistic finite-shot setting. Under both amplitude damping and dephasing, QEM-STEP achieves the best MAE and IR, outperforming both ZNE and DAEM. These results highlight the effectiveness of QEM-STEP when expectation values are estimated from finite measurements, as required on quantum hardware.

5 Conclusions

We introduced a pseudo state perturbation framework for quantum error mitigation, casting the task as a regularized projection problem in a Pauli basis. Building on this formulation, QEM-STEP combines a practical classical basis-selection procedure based on backward Pauli propagation with a neural predictor that learns instance-dependent perturbation coefficients. Experiments on quantum circuits and many-body dynamics demonstrate competitive mitigation performance across diverse tasks and noise models. These results highlight the potential of structured state perturbations as a flexible foundation for learning-based quantum error mitigation.

AI USE STATEMENT

We used generative AI tools to edit language, reorganize the manuscript, cross-check algebra, and review the code for the experiments reported in this paper. We did not use these tools to generate research content. We reviewed all AI-assisted work and take responsibility for the final manuscript, including any text, claims, and artifacts produced with AI assistance.

REPRODUCIBILITY STATEMENT

We describe the implementations of the noise models and baselines and cite the sources from which these implementations were taken/adapted. The main text and appendix provides the experimental configurations, a full description of our method, and its assumptions. For theoretical result, the definition, assumptions and description are provided in the main text while proofs are provided in the appendix.

REFERENCES

Peter W Shor. Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer. *SIAM review*, 41(2):303–332, 1999.

Lov K Grover. A fast quantum mechanical algorithm for database search. In *Proceedings of the twenty-eighth annual ACM symposium on Theory of computing*, pages 212–219, 1996.

Carlos Bravo-Prieto, Ryan LaRose, Marco Cerezo, Yigit Subasi, Lukasz Cincio, and Patrick J Coles. Variational quantum linear solver. *Quantum*, 7:1188, 2023.

Alberto Peruzzo, Jarrod McClean, Peter Shadbolt, Man-Hong Yung, Xiao-Qi Zhou, Peter J Love, Alán Aspuru-Guzik, and Jeremy L O'brien. A variational eigenvalue solver on a photonic quantum processor. *Nature communications*, 5(1):4213, 2014.

John Preskill. Quantum computing in the nisq era and beyond. *Quantum*, 2:79, 2018.

Tudor Giurgica-Tiron, Yousef Hindy, Ryan LaRose, Andrea Mari, and William J Zeng. Digital zero noise extrapolation for quantum error mitigation. In *2020 IEEE international conference on quantum computing and engineering (QCE)*, pages 306–316. IEEE, 2020.

Piotr Czarnik, Andrew Arrasmith, Patrick J Coles, and Lukasz Cincio. Error mitigation with clifford quantum-circuit data. *Quantum*, 5:592, 2021.

Kristan Temme, Sergey Bravyi, and Jay M Gambetta. Error mitigation for short-depth quantum circuits. *Physical review letters*, 119(18):180509, 2017.

Changjun Kim, Kyungdeock Daniel Park, and June-Koo Rhee. Quantum error mitigation with artificial neural network. *IEEE Access*, 8:188853–188860, 2020.

Haoran Liao, Derek S Wang, Iskandar Sitdikov, Ciro Salcedo, Alireza Seif, and Zlatko K Minev. Machine learning for practical quantum error mitigation. *Nature Machine Intelligence*, 6(12): 1478–1486, 2024.

Tianyi Bao, Xinyu Ye, Hang Ruan, Chang Liu, Wenjie Wu, and Junchi Yan. Beyond circuit connections: A non-message passing graph transformer approach for quantum error mitigation. In *The Thirteenth International Conference on Learning Representations*, 2025.

Manwen Liao, Yan Zhu, Giulio Chiribella, and Yuxiang Yang. Noise-agnostic quantum error mitigation with data augmented neural models. *npj Quantum Information*, 11(1):8, 2025.

Timothy Proctor, Kenneth Rudinger, Kevin Young, Erik Nielsen, and Robin Blume-Kohout. Measuring the capabilities of quantum computers. *Nature Physics*, 18(1):75–79, 2022.

Timothy N Georges, Bjorn K Berntson, Christoph Sünderhauf, and Aleksei V Ivanov. Pauli decomposition via the fast walsh-hadamard transform. *New Journal of Physics*, 27(3):033004, 2025.

Diego García-Martín, Martín Larocca, and Marco Cerezo. Effects of noise on the overparametrization of quantum neural networks. *Physical Review Research*, 6(1):013295, 2024.

Michael A Nielsen and Isaac L Chuang. *Quantum computation and quantum information*. Cambridge university press, 2010.

John Watrous. *The theory of quantum information*. Cambridge university press, 2018.

Yuguo Shao, Song Cheng, and Zhengwei Liu. Pauli propagation: Simulating quantum spin dynamics via operator complexity. *arXiv preprint arXiv:2510.22311*, 2025.

Bikas K Chakrabarti, Amit Dutta, and Parongama Sen. *Quantum Ising phases and transitions in transverse Ising models*. Springer, 1996.

Mancheon Han, Hyowon Park, and Sangkook Choi. Quantum zeno monte carlo for computing observables. *npj Quantum Information*, 11(1):46, 2025.

Suguru Endo, Simon C Benjamin, and Ying Li. Practical quantum error mitigation for near-future applications. *Physical Review X*, 8(3):031027, 2018.

Manuel S Rudolph, Tyson Jones, Yanting Teng, Armando Angrisani, and Zoë Holmes. Pauli propagation: A computational framework for simulating quantum systems. *arXiv preprint arXiv:2505.21606*, 2025.

Scott Aaronson and Daniel Gottesman. Improved simulation of stabilizer circuits. *Physical Review A—Atomic, Molecular, and Optical Physics*, 70(5):052328, 2004.

Sergey Bravyi and David Gosset. Improved classical simulation of quantum circuits dominated by clifford gates. *Physical review letters*, 116(25):250501, 2016.

Sergey Bravyi, Dan Browne, Padraic Calpin, Earl Campbell, David Gosset, and Mark Howard. Simulation of quantum circuits by low-rank stabilizer decompositions. *Quantum*, 3:181, 2019.

Jarrod R McClean, Jonathan Romero, Ryan Babbush, and Alán Aspuru-Guzik. The theory of variational hybrid quantum-classical algorithms. *New Journal of Physics*, 18(2):023023, 2016.

Kunal Sharma, Sumeet Khatri, Marco Cerezo, and Patrick J Coles. Noise resilience of variational quantum compiling. *New Journal of Physics*, 22(4):043006, 2020.

Daniel Stilck França and Raul Garcia-Patron. Limitations of optimization algorithms on noisy quantum devices. *Nature Physics*, 17(11):1221–1227, 2021.

Samson Wang, Enrico Fontana, Marco Cerezo, Kunal Sharma, Akira Sone, Lukasz Cincio, and Patrick J Coles. Noise-induced barren plateaus in variational quantum algorithms. *Nature communications*, 12(1):6961, 2021.

Marco Schumann, Frank K Wilhelm, and Alessandro Ciani. Emergence of noise-induced barren plateaus in arbitrary layered noise models. *Quantum Science and Technology*, 9(4):045019, 2024.

Phattharaporn Singkanipa and Daniel A Lidar. Beyond unital noise in variational quantum algorithms: noise-induced barren plateaus and limit sets. *Quantum*, 9:1617, 2025.

Hanrui Wang, Jiaqi Gu, Yongshan Ding, Zirui Li, Frederic T Chong, David Z Pan, and Song Han. Quantumnat: quantum noise-aware training with noise injection, quantization and normalization. In *Proceedings of the 59th ACM/IEEE design automation conference*, pages 1–6, 2022.

Frank Arute, Kunal Arya, Ryan Babbush, Dave Bacon, Joseph C Bardin, Rami Barends, Rupak Biswas, Sergio Boixo, Fernando GSL Brandao, David A Buell, et al. Quantum supremacy using a programmable superconducting processor. *nature*, 574(7779):505–510, 2019.

Steven R White. Density matrix formulation for quantum renormalization groups. *Physical review letters*, 69(19):2863, 1992.

Jacob C Bridgeman and Christopher T Chubb. Hand-waving and interpretive dance: an introductory course on tensor networks. *Journal of physics A: Mathematical and theoretical*, 50(22):223001, 2017.

Cornelius Lanczos. An iteration method for the solution of the eigenvalue problem of linear differential and integral operators. *Journal of research of the National Bureau of Standards*, 45(4): 255–282, 1950.

Steven R White and Adrian E Feiguin. Real-time evolution using the density matrix renormalization group. *Physical review letters*, 93(7):076401, 2004.

Andrew John Daley, Corinna Kollath, Ulrich Schollwöck, and Guifré Vidal. Time-dependent density-matrix renormalization-group using adaptive effective hilbert spaces. *Journal of Statistical Mechanics: Theory and Experiment*, 2004(04):P04005, 2004.

Masuo Suzuki. Generalized trotter's formula and systematic approximants of exponential operators and inner derivations with applications to many-body problems. *Communications in Mathematical Physics*, 51(2):183–190, 1976.

Kaifeng Bu, Dax Enshan Koh, Roy J Garcia, and Arthur Jaffe. Classical shadows with pauli-invariant unitary ensembles. *npj Quantum Information*, 10(1):6, 2024.

A Related works

Effect of noise Noise affects quantum algorithms in several ways. Imperfect gates and state preparation alter the state produced by a circuit, while readout errors distort measurement outcomes. Consequently, measured expectation values can differ from their ideal values, affecting tasks that depend on accurate observables Temme et al. (2017). In variational algorithms, noise can also change the cost landscape and hinder optimization. Some errors are naturally suppressed by variational optimization McClean et al. (2016), and certain noise models preserve the optimal parameters in variational quantum compiling Sharma et al. (2020). This resilience, however, depends on the setting. Noise can limit optimization performance Stilck França and Garcia-Patron (2021), induce barren plateaus with vanishing gradients Wang et al. (2021); Schumann et al. (2024); Singkanipa and Lidar (2025), and make quantum neural network states less sensitive to parameter changes García-Martín et al. (2024). It can also reduce classification accuracy Wang et al. (2022).

Pauli propagation Pauli propagation simulates a quantum process in the Heisenberg picture by evolving a measured observable backward as a weighted sum of Pauli strings Rudolph et al. (2025). Clifford gates map each Pauli string to another Pauli string, whereas non-Clifford gates can generate additional terms. As the number of terms may grow rapidly, approximate methods truncate the expansion, for example by retaining the largest Pauli coefficients at each step Shao et al. (2025). Prior work uses this representation to estimate observables through classical simulation. We use it for two purposes in QEM: to select Pauli directions that influence the measured observable and to construct training circuits whose ideal expectation values remain classically tractable.

B Quantum noise model

In this section, We describe the noise models used throughout the experiments. The incoherent gate, readout, and coherent error settings follow Bao et al. (2025), whereas the damping channels follow Liao et al. (2025).

B.1 Incoherent Errors

This paper consider two types of incoherent errors: bit-flip and depolarizing error

Bit-flip error A bit-flip channel with error probability p_X is defined as

$$\mathcal{E}_{r_X}[\rho] = (1 - r_X)\rho + r_X X\rho X.$$

Thus, the input state remains unchanged with probability $1-r_X$ and undergoes a Pauli-X error with probability r_X.

Depolarizing error For a single qubit, the depolarizing channel with error probability p_{dep} is

$$\mathcal{E}_{r_{\mathrm{dep}}}[\rho] = (1 - r_{\mathrm{dep}})\rho + r_{\mathrm{dep}}\frac{I}{2}.$$

Equivalently,

$$\mathcal{E}_{r_{\mathrm{dep}}}[\rho] = \left(1 - \frac{3r_{\mathrm{dep}}}{4}\right)\rho + \frac{r_{\mathrm{dep}}}{4}\left(X\rho X + Y\rho Y + Z\rho Z\right).$$

The corresponding Kraus operators are $K_0 = \sqrt{1 - \frac{3r_{\mathrm{dep}}}{4}}\, I, K_1 = \frac{\sqrt{r_{\mathrm{dep}}}}{2}X, K_2 = \frac{\sqrt{r_{\mathrm{dep}}}}{2}Y, K_3 = \frac{\sqrt{r_{\mathrm{dep}}}}{2}Z$.

Following Bao et al. (2025), the error rates are set to 0.16% for single-qubit gates and 0.62% for two-qubit gates, based on the reported characteristics of the Sycamore device Arute et al. (2019).

B.2 Readout errors

Readout errors occur when a measurement device reports an outcome different from the ideal measurement outcome. We assume that the measurement outcome of each qubit is independently flipped

with probability r_{ro}. Thus, an n-qubit measurement may contain errors on any subset of its measured bits. FollowingBao et al. (2025), we set $r_{\mathrm{ro}} = 0.038$. Readout errors are combined with the incoherent gate errors in the mixed-noise setting.

B.3 Coherent errors

Coherent errors arise from systematic imperfections in unitary gate operations. For a gate generated by a Pauli operator P, an over-rotation by angle ϵ is represented by

$$U_{\mathrm{err}}(\epsilon) = \exp\left(-\frac{\mathrm{i}\epsilon}{2}P\right).$$

Unlike incoherent noise, this error preserves the purity of the state but introduces a systematic deviation that can accumulate coherently with circuit depth. Following the coherent-noise setting of Bao et al. (2025), systematic over-rotations are applied to the two-qubit CX, CY, CZ, and SWAP gates, with an average over-rotation angle of 0.02π. These coherent errors are combined with the noise model extracted from the IBM `FakeLima` provider.

B.4 Phase damping

Phase damping, also referred to as dephasing, suppresses the off-diagonal elements of a density matrix while leaving its populations unchanged. Following the swap-test experiment of Liao et al. (2025), the channel is defined as

$$\mathcal{E}_\lambda^{\mathrm{PD}}[\rho] = K_0\rho K_0^\dagger + K_1\rho K_1^\dagger,$$

where

$$K_0 = \begin{pmatrix} 1 & 0 \\ 0 & e^{-2\lambda} \end{pmatrix}, \qquad K_1 = \begin{pmatrix} 0 & 0 \\ 0 & \sqrt{1-e^{-4\lambda}} \end{pmatrix}.$$

For the swap-test benchmark, this channel is applied before each controlled-SWAP operation using $\lambda \in \{0.05, 0.08, 0.12, 0.15\}$.

Meanwhile, in the many-body benchmark it is applied after the complete Ising evolution and immediately before measurement. Following Liao et al. (2025), the many-body experiments consider noise strengths $\lambda \in [0.05, 0.29]$ with step 0.02

B.5 Amplitude damping

Amplitude damping models energy relaxation from the excited state $|1\rangle$ to the ground state $|0\rangle$. Following Liao et al. (2025), the channel is

$$\mathcal{E}_\lambda^{\mathrm{AD}}[\rho] = K_0\rho K_0^\dagger + K_1\rho K_1^\dagger,$$

with

$$K_0 = \begin{pmatrix} 1 & 0 \\ 0 & \sqrt{1-\lambda} \end{pmatrix}, \qquad K_1 = \begin{pmatrix} 0 & \sqrt{\lambda} \\ 0 & 0 \end{pmatrix},$$

where $\lambda \in [0, 1]$ is the decay probability. Following Liao et al. (2025), the many-body experiment is evaluated over 13 noise strengths $\lambda \in [0.05, 0.29]$ with step 0.02

C Detailed experiment design and additional result

C.1 Standard settings

Trotterized circuits. Following Bao et al. (2025) and Chakrabarti et al. (1996), we construct Trotterized circuits that simulate the dynamics of the one-dimensional transverse-field Ising model,

$$H = -J\sum_j Z_j Z_{j+1} + h\sum_j X_j,$$

where J is the nearest-neighbor coupling strength and h is the transverse-field strength. The circuits approximate the time-evolution operator $e^{-\mathrm{i}Ht}$ by repeatedly applying local interaction and

transverse-field evolution gates. We generate different circuit instances by randomly selecting J, h, and the evolution time t, with coupling parameters sampled uniformly within the paramagnetic regime $J < h$. For the All-6-Trotter benchmark, we use six-qubit circuits with 1 to 20 Trotter steps, comprising 750 training circuits, 150 validation circuits, and 300 test circuits.

Random circuits. To evaluate mitigation performance across diverse circuit structures, we construct the Max-6-Random benchmark following Bao et al. (2025). Each circuit contains at most six qubits and 150 gates, selected from

$$\{X, Y, Z, H, R_X, R_Y, R_Z, \mathrm{CNOT}, \mathrm{CZ}, \mathrm{SWAP}, R_{ZZ}\}.$$

The gate types, their ordering, and the qubits on which they act are randomized within each circuit. For parameterized rotation gates, the rotation angles are randomly sampled from $[0, 2\pi]$. The dataset contains 750 training circuits, 150 validation circuits, and 300 test circuits.

C.2 Non-standard settings

We evaluate QEM-STEP beyond the standard benchmarks through three additional settings: circuit-scale extrapolation, coherent noise, and noise amplification. These experiments follow Bao et al. (2025) implementation, with details provided below.

Circuit-scale extrapolation. We assess generalization to greater circuit depths using the same six-qubit circuit generator as in the standard Trotterized benchmark. The dataset contains 260 training circuits with 1 - 15 Trotter steps, 40 validation circuits, and 100 test circuits with 16 - 20 Trotter steps. This separation introduces a distribution shift in circuit depth, allowing us to evaluate mitigation on deeper circuits than those encountered during training.

Coherent noise. We evaluate four-qubit Trotterized circuits subject to systematic over-rotation errors combined with the noise model of the IBM `FakeLima` provider. The over-rotation errors are applied to two-qubit gates, including CX, CY, CZ, and SWAP, with an average angle of 0.02π. This setting assesses mitigation under systematic gate imperfections alongside simulated device noise.

Noise amplification. We augment the training data for four-qubit Trotterized circuits by executing each circuit at noise-scaling factors $\lambda \in \{1, 2, 3\}$. Each execution produces a distinct noisy input paired with the same ideal expectation value. Applying these factors to 200 training circuits yields 600 training examples, with an additional 200 circuits used for testing. This setting evaluates robustness to varying noise strengths while keeping the underlying circuit structures fixed across noise levels.

Table 4: MAE and IR for different methods under different experimental settings. The best value in each column are bold and highlighted in gray while the second best is underlined

	Circuit-Scale Extrapolation		**Noise-Scale Amplification**		**Coherent Noise**	
Method	**MAE ↓**	**IR ↑**	**MAE ↓**	**IR ↑**	**MAE ↓**	**IR ↑**
Raw Noisy	0.1679	–	0.1921	–	0.2616	–
ZNE	0.1192	<u>0.8100</u>	0.1320	**0.8000**	0.2600	0.5466
CDR	0.1320	0.6300	0.1092	0.7150	0.1697	0.7033
OLS	0.0646	0.7500	0.1091	0.6100	0.1240	0.7000
MLP	<u>0.0600</u>	0.7700	<u>0.0725</u>	0.7300	<u>0.1143</u>	0.7233
GNN	0.0820	0.6700	0.1068	0.6200	0.1289	0.7133
GTranQEM	0.0642	0.7900	0.1222	0.5800	0.1170	<u>0.7633</u>
QEM-STEP	**0.0589**	**0.8300**	**0.0589**	<u>0.7500</u>	**0.0630**	**0.8300**

Table 4 further evaluates the robustness of QEM-STEP under circuit-scale extrapolation, noise-scale amplification, and coherent noise. QEM-STEP achieves the lowest MAE in all three settings, while attaining the highest IR under circuit-scale extrapolation and coherent noise. The advantage is particularly pronounced under coherent noise, where QEM-STEP achieves an MAE of 0.0630, compared with 0.1143 for the strongest baseline, while also obtaining the highest IR of 0.8300. These results indicate that QEM-STEP remains effective under shifts in both circuit scale and noise characteristics.

C.3 SWAP-TEST

The swap-test benchmark of Liao et al. (2025) estimates the overlap between two five-qubit pure states $|\psi\rangle$ and $|\phi\rangle$. The circuit uses eleven qubits, having two five-qubit input registers and an ancillary control qubit initialized in $|0\rangle$. A Hadamard gate is applied to the ancilla, followed by five controlled-SWAP gates acting on corresponding qubits of the two registers, and a final Hadamard gate on the ancilla. Measuring the observable $O = Z \otimes I^{\otimes 10}$ yields the ideal expectation value $\langle O \rangle^{\text{ideal}} = |\langle \psi | \phi \rangle|^2$, which equals the fidelity between the input states.

Noise is introduced before each controlled-SWAP operation. Following Liao et al. (2025), we consider the phase-damping channel (see Appendix B.4) with noise strengths $\lambda \in \{0.05, 0.08, 0.12, 0.15\}$. The mitigation task is to recover the ideal state overlap from the noisy ancilla expectation values.

C.4 MANY-BODY DYNAMICS IMPLEMENTATION DETAILS

Following Liao et al. (2025), the dynamics are governed by the one-dimensional transverse-field Ising Hamiltonian

$$H_{\text{Ising}} = -g \sum_{j=1}^{n} X_j - J \sum_{j=1}^{n-1} Z_j Z_{j+1}.$$

The target evolution is generated by $U(t) = \exp(-\mathrm{i} H_{\text{Ising}} t)$, with $n = 50$, $J = 1$, $g = 2$, and $t = 5$. The input states ρ are ground states of the same Ising Hamiltonian family, generated by varying J over $[-2, 2]$ while fixing $g = 1$. We use 100 input states for training and 20 input states for testing. The ground states are computed using the density matrix renormalization group (DMRG) White (1992), in which the Hamiltonian and quantum state are represented as a matrix product operator (MPO) and a matrix product state (MPS), respectively Bridgeman and Chubb (2017). Following Liao et al. (2025), we perform two DMRG sweeps and two Lanczos iterations Lanczos (1950) for each local optimization, with a maximum Krylov-subspace dimension of 4. The dynamics are subsequently simulated using time-evolving block decimation (TEBD) White and Feiguin (2004); Daley et al. (2004). Based on the Trotter-Suzuki decomposition Suzuki (1976), the Hamiltonian is decomposed into local terms, whose corresponding evolution operators are sequentially applied to the initial MPS. We use a time step of $dt = 2$, a maximum bond dimension of 25, and a singular-value cutoff of 10^{-10}.

The noise channel is applied after the unitary evolution and immediately before measurement. Consequently, for an input state ρ and observable O, the noisy expectation value can be evaluated in the Heisenberg picture as

$$\begin{aligned} \mathrm{Tr}(O\,\mathcal{E}\,[\mathcal{U}(t)[\rho]]) &= \mathrm{Tr}\big(\mathcal{E}^\dagger[O]\,\mathcal{U}(t)[\rho]\big)\,, \\ \mathcal{E}^\dagger[O] &= \sum_k K_k^\dagger O K_k, \end{aligned} \tag{13}$$

where $\{K_k\}_k$ are the Kraus operators of $\mathcal{E}$. Following the implementation of Liao et al. (2025), we model the post-evolution noise as a product of independent local single-qubit channels. Let $\mathcal{V} \subseteq \{1, \ldots, n\}$ denote the support of a given Pauli observable, i.e., the subsystem on which it acts nontrivially, and let $\bar{\mathcal{V}}$ denote its complement. Since the observable acts as the identity on $\bar{\mathcal{V}}$ and the adjoint of a trace-preserving channel leaves the identity invariant, the adjoint noise transformation needs to be evaluated only on the sites in $\mathcal{V}$. In this experiment, $\mathcal{V}$ consists of a neighboring qubit pair, and we evaluate the corresponding two-qubit Pauli observables under local phase-damping and amplitude-damping channels. In the exact-expectation setting, the noisy expectation value is computed directly from the evolved matrix-product state. In the finite-shot setting, it is estimated by sampling outcomes from the corresponding Pauli measurement distribution.

We now describe the implementation of QEM-STEP in this experiment. Since the equation 13 apply only one noise channel after evolution, we set $\rho' := \mathcal{U}[\rho]$ then the after perturbation will be $\mathrm{Tr}\big(\mathcal{E}^\dagger[O]\,(\rho' + \delta)\big) = \mathrm{Tr}\big(\mathcal{E}^\dagger[O]\,(\rho')\big) + \mathrm{Tr}\big(\mathcal{E}^\dagger[O](\delta)\big)$. For the many-body experiment considered here, $|\mathcal{V}| = 2$. The global observable can therefore be written as $O = O_{\mathcal{V}} \otimes \mathcal{I}_{\bar{\mathcal{V}}}$, where $O_{\mathcal{V}}$ is the local Pauli observable acting on the measured subsystem. Since the noise is introduced after the many-body evolution, we consider a post-evolution noise channel that does not couple $\mathcal{V}$ to its

complement,

$$\mathcal{E} = \mathcal{E}_{\mathcal{V}} \otimes \mathcal{E}_{\bar{\mathcal{V}}},$$

where $\mathcal{E}_{\bar{\mathcal{V}}}$ is trace preserving. For a global Pauli perturbation direction $P_i = P_{i,\mathcal{V}} \otimes P_{i,\bar{\mathcal{V}}}$, the corresponding response coefficient is

$$\begin{aligned} c_i &= \mathrm{Tr}\left[O\,\mathcal{E}(P_i)\right] \\ &= \mathrm{Tr}\left[(O_{\mathcal{V}} \otimes \mathcal{I}_{\bar{\mathcal{V}}})\left(\mathcal{E}_{\mathcal{V}}(P_{i,\mathcal{V}}) \otimes \mathcal{E}_{\bar{\mathcal{V}}}(P_{i,\bar{\mathcal{V}}})\right)\right] \\ &= \mathrm{Tr}_{\mathcal{V}}\left[O_{\mathcal{V}}\,\mathcal{E}_{\mathcal{V}}(P_{i,\mathcal{V}})\right]\mathrm{Tr}_{\bar{\mathcal{V}}}\left[\mathcal{E}_{\bar{\mathcal{V}}}(P_{i,\bar{\mathcal{V}}})\right] \\ &= \mathrm{Tr}_{\mathcal{V}}\left[O_{\mathcal{V}}\,\mathcal{E}_{\mathcal{V}}(P_{i,A\mathcal{V}})\right]\mathrm{Tr}_{\bar{\mathcal{V}}}\left[P_{i,\bar{\mathcal{V}}}\right], \end{aligned}$$

where the last equality follows from the trace-preserving property of $\mathcal{E}_{\bar{\mathcal{V}}}$. Every nonidentity Pauli string is traceless. Therefore, if $P_{i,\bar{\mathcal{V}}} \neq \mathcal{I}_{\bar{\mathcal{V}}}$ then $\mathrm{Tr}_{\bar{\mathcal{V}}}(P_{i,\bar{\mathcal{V}}}) = 0$, and consequently $c_i = 0$. Hence, only Pauli perturbation directions whose support lies entirely inside the measured subsystem $\mathcal{V}$ can contribute to the mitigation term. The full correction is therefore exactly equivalent to a sum over the local Pauli basis on $\mathcal{V}$. For the two-qubit observables used in this experiment, $|\mathcal{V}| = 2$, and the number of nonidentity Pauli operators is $4^{|\mathcal{V}|} - 1 = 4^2 - 1 = 15$.

C.5 Ablation studies

Table 5: Ablation study of the context components and Top-K selection in QEM-STEP. The best value for each metric is shown in bold and highlighted in gray.

Index	PR Context	SC Context	Top$K$	MAE	CIR
0				0.0375	0.6900
1			✓	0.0410	0.6400
2		✓		0.0376	0.6500
3		✓	✓	0.0398	0.6133
4	✓			0.0350	0.6667
5	✓		✓	0.0360	0.6733
6	✓	✓		0.0333	0.6800
7	✓	✓	✓	**0.0310**	**0.7233**

Table 5 examines the contribution of the PR context, SC context, and Top-K selection. We run this ablation study on the All-6-Trotter incoherent noise. Among the individual components, the PR context provides the clearest reduction in MAE, decreasing it from 0.0375 to 0.0350. Combining the PR and SC contexts further reduces the MAE to 0.0333, suggesting that the two context representations provide complementary information. Compared with the configuration without any of the three components, the complete model reduces MAE by 17.3% and increases CIR from 0.6900 to 0.7233.

D Additional proof

D.1 Proof of Lemma 3.1

Proof. Let $\mu_{i,\pm} = \mathrm{Tr}(O\,\boldsymbol{\mathcal{E}}[\rho_{i,\pm}])$, and let $\hat{\mu}_{i,\pm}$ denote their empirical estimates obtained from M independent shots. By equation 7, $\hat{c}_i = 2^{n-1}\left(\hat{\mu}_{i,+} - \hat{\mu}_{i,-}\right)$. Assuming the measurements used to estimate the two terms are independent,

$$\mathrm{Var}(\hat{c}_i) = 2^{2n-2}\left[\mathrm{Var}(\hat{\mu}_{i,+}) + \mathrm{Var}(\hat{\mu}_{i,-})\right].$$

Since $\|O\|_\infty \leq 1$, each single-shot outcome $X_{i,\pm}$ satisfies $|X_{i,\pm}| \leq 1$, and hence $\mathrm{Var}(X_{i,\pm}) \leq \mathbb{E}[X_{i,\pm}^2] \leq 1$. Therefore,

$$\mathrm{Var}(\hat{\mu}_{i,\pm}) = \frac{\mathrm{Var}(X_{i,\pm})}{M} \leq \frac{1}{M}.$$

Substituting into the previous expression yields

$$\mathrm{Var}(\hat{c}_i) \leq 2^{2n-2}\left(\frac{1}{M} + \frac{1}{M}\right) = \frac{2^{2n-1}}{M}.$$

Taking the square root gives equation 8. □

D.2 Proof of Remark 3.2

Proof. By the definition of the adjoint channel,

$$c_i = \mathrm{Tr}(O\,\boldsymbol{\mathcal{E}}[P_i]) = \mathrm{Tr}\Big(\boldsymbol{\mathcal{E}}^\dagger[O]P_i\Big) = 2^n\alpha\Big(\boldsymbol{\mathcal{E}}^\dagger[O], P_i\Big)\,.$$

The vector identity follows componentwise. We have $\hat{\alpha}_i = \frac{\hat{c}_i}{2^n}$. Therefore, $\mathrm{Var}(\hat{\alpha}_i) = \frac{1}{2^{2n}}\,\mathrm{Var}(\hat{c}_i)$. Applying Lemma 3.1, $\mathrm{Var}(\hat{\alpha}_i) \le \frac{1}{2^{2n}}\frac{2^{2n-1}}{M} = \frac{1}{2M}$. Taking the square root gives equation 9. □

Lemma D.1 (Accumulated Top-K truncation) *Given an n-qubit unitary $\boldsymbol{\mathcal{U}}$ and observable O. Let the sequentially truncated propagation be initialized by $\widetilde{O}_L = O$. At propagation step l, let A_{l-1} denote the observable obtained before truncation, and let $\widetilde{O}_{l-1}$ denote its Top-K approximation:*

$$A_{l-1} = \mathcal{U}_l^\dagger(\widetilde{O}_l), \qquad \widetilde{O}_{l-1} = \textit{Top-}K(A_{l-1})$$

Using Pauli decomposition $A_{l-1} = \sum_j \alpha(A_{l-1}, P_j)P_j$ (equation 2), assuming $|\alpha(A_{l-1}, P_1)| \ge |\alpha(A_{l-1}, P_2)| \ge \cdots \ge |\alpha(A_{l-1}, P_{4^n})|$, and using the discarded squared Pauli mass $\Delta_{l-1}(K) := \sum_{j>K} |\alpha(A_{l-1}, P_j)|^2$ (Shao et al., 2025), the error of the sequentially truncated observable satisfies

$$\Big\|\boldsymbol{\alpha}(\boldsymbol{\mathcal{U}}^\dagger[O]) - \boldsymbol{\alpha}(\widetilde{O}_0)\Big\|_2 \le \sum_{l=1}^{L}\sqrt{\Delta_{l-1}(K)}. \tag{14}$$

Proof. By Pauli decomposition of $\boldsymbol{\alpha}(A_{l-1})$ and $\boldsymbol{\alpha}(\widetilde{O}_{l-1})$, we have

$$\Big\|\boldsymbol{\alpha}(A_{l-1}) - \boldsymbol{\alpha}(\widetilde{O}_{l-1})\Big\|_2^2 = \sum_{j>K}|\alpha(A_{l-1}, P_j)|^2 = \Delta_{l-1}(K).$$

Moreover, unitary conjugation preserves the Euclidean norm of the Pauli coefficient vector, since for any Hermitian operator B, $\|\boldsymbol{\alpha}(B)\|_2^2 = 2^{-n}\,\mathrm{Tr}(B^\dagger B)$. (Bu et al., 2024) Using $A_{l-1} = \mathcal{U}_l^\dagger[\widetilde{O}_l]$, the triangle inequality therefore gives

$$\begin{aligned}
&\Big\|\boldsymbol{\alpha}\Big((\mathcal{U}_l^\dagger \circ \cdots \circ \mathcal{U}_L^\dagger)[O]\Big) - \boldsymbol{\alpha}(\widetilde{O}_{l-1})\Big\|_2 \\
&\quad\le \Big\|\boldsymbol{\alpha}\Big((\mathcal{U}_l^\dagger \circ \cdots \circ \mathcal{U}_L^\dagger)[O]\Big) - \boldsymbol{\alpha}(A_{l-1})\Big\|_2 + \sqrt{\Delta_{l-1}(K)} \\
&\quad= \Big\|\boldsymbol{\alpha}\Big((\mathcal{U}_{l+1}^\dagger \circ \cdots \circ \mathcal{U}_L^\dagger)[O]\Big) - \boldsymbol{\alpha}(\widetilde{O}_l)\Big\|_2 + \sqrt{\Delta_{l-1}(K)},
\end{aligned}$$

where the empty composition for $l = L$ is the identity map. Iterating this inequality and using $\widetilde{O}_L = O$ yields

$$\Big\|\boldsymbol{\alpha}\Big(\boldsymbol{\mathcal{U}}^\dagger[O]\Big) - \boldsymbol{\alpha}(\widetilde{O}_0)\Big\|_2 \le \sum_{l=1}^{L}\sqrt{\Delta_{l-1}(K)}.$$

□

D.3 Proof of Theorem 3.3

Proof. Let $\mathcal{S}_K$ denote the Pauli support obtained from the sequential Top-K propagation in Lemma D.1. Let $\boldsymbol{\alpha}_K$ and $\widehat{\boldsymbol{\alpha}}_K$ denote the exact and finite-shot Pauli coefficient vectors of $\boldsymbol{\mathcal{E}}^\dagger[O]$ restricted to $\mathcal{S}_K$, with all coefficients outside $\mathcal{S}_K$ set to zero. Hence, $\boldsymbol{\alpha}_K(\boldsymbol{\mathcal{E}}^\dagger[O])$ is the projection of $\boldsymbol{\alpha}(\boldsymbol{\mathcal{E}}^\dagger[O])$ onto the coordinate subspace supported on $\mathcal{S}_K$. Since $\boldsymbol{\alpha}(\widetilde{O}_0)$ is also supported on $\mathcal{S}_K$, this projection is at least as close to $\boldsymbol{\alpha}(\boldsymbol{\mathcal{E}}^\dagger[O])$ as $\boldsymbol{\alpha}(\widetilde{O}_0)$.

Define the finite-shot estimation noise by $\boldsymbol{\zeta} := \widehat{\boldsymbol{\alpha}}_K\left(\boldsymbol{\mathcal{E}}^\dagger[O]\right) - \boldsymbol{\alpha}_K\left(\boldsymbol{\mathcal{E}}^\dagger[O]\right)$. Thus, $\widehat{\boldsymbol{\alpha}}_K\left(\boldsymbol{\mathcal{E}}^\dagger[O]\right) = \boldsymbol{\alpha}_K\left(\boldsymbol{\mathcal{E}}^\dagger[O]\right) + \boldsymbol{\zeta}$, and assuming $\mathbb{E}[\boldsymbol{\zeta}] = 0$.

We first bound the deterministic error induced by restricting the noisy Heisenberg-picture observable to $\mathcal{S}_K$. By the projection property above,

$$\begin{aligned}\left\|\boldsymbol{\alpha}\left(\boldsymbol{\mathcal{E}}^{\dagger}[O]\right)-\boldsymbol{\alpha}_K\left(\boldsymbol{\mathcal{E}}^{\dagger}[O]\right)\right\|_2 &\le \left\|\boldsymbol{\alpha}\left(\boldsymbol{\mathcal{E}}^{\dagger}[O]\right)-\boldsymbol{\alpha}(\widetilde{O}_0)\right\|_2 \\ &\le \left\|\boldsymbol{\alpha}\left(\boldsymbol{\mathcal{E}}^{\dagger}[O]\right)-\boldsymbol{\alpha}\left(\boldsymbol{\mathcal{U}}^{\dagger}[O]\right)\right\|_2+\left\|\boldsymbol{\alpha}\left(\boldsymbol{\mathcal{U}}^{\dagger}[O]\right)-\boldsymbol{\alpha}(\widetilde{O}_0)\right\|_2 \\ &\le \left\|\boldsymbol{\alpha}\left(\boldsymbol{\mathcal{E}}^{\dagger}[O]\right)-\boldsymbol{\alpha}\left(\boldsymbol{\mathcal{U}}^{\dagger}[O]\right)\right\|_2+E(K), \end{aligned} \tag{15}$$

where the second inequality follows from the triangle inequality, and the last inequality follows from the Lemma D.1. Consequently,

$$\left\|\boldsymbol{\alpha}\left(\boldsymbol{\mathcal{E}}^{\dagger}[O]\right)-\boldsymbol{\alpha}_K\left(\boldsymbol{\mathcal{E}}^{\dagger}[O]\right)\right\|_2^2 \le \left(\left\|\boldsymbol{\alpha}\left(\boldsymbol{\mathcal{E}}^{\dagger}[O]\right)-\boldsymbol{\alpha}\left(\boldsymbol{\mathcal{U}}^{\dagger}[O]\right)\right\|_2+E(K)\right)^2 \tag{16}$$

$$\le 2\left\|\boldsymbol{\alpha}\left(\boldsymbol{\mathcal{E}}^{\dagger}[O]\right)-\boldsymbol{\alpha}\left(\boldsymbol{\mathcal{U}}^{\dagger}[O]\right)\right\|_2^2+2E(K)^2. \tag{17}$$

Since $\boldsymbol{\zeta}$ has at most K nonzero entries and each retained coefficient estimator has variance at most $1/(2M)$,

$$\mathbb{E}\left[\|\boldsymbol{\zeta}\|_2^2\right]=\sum_{i\in\mathcal{S}_K}\mathrm{Var}(\widehat{\alpha}_i)\le K/(2M). \tag{18}$$

Finally, from equation 17 and equation 18,

$$\begin{aligned}\mathbb{E}\left[\left\|\boldsymbol{\alpha}\left(\boldsymbol{\mathcal{E}}^{\dagger}[O]\right)-\widehat{\boldsymbol{\alpha}}_K\left(\boldsymbol{\mathcal{E}}^{\dagger}[O]\right)\right\|_2^2\right] &= \mathbb{E}\left[\left\|\boldsymbol{\alpha}\left(\boldsymbol{\mathcal{E}}^{\dagger}[O]\right)-\boldsymbol{\alpha}_K\left(\boldsymbol{\mathcal{E}}^{\dagger}[O]\right)-\boldsymbol{\zeta}\right\|_2^2\right] \\ &= \left\|\boldsymbol{\alpha}\left(\boldsymbol{\mathcal{E}}^{\dagger}[O]\right)-\boldsymbol{\alpha}_K\left(\boldsymbol{\mathcal{E}}^{\dagger}[O]\right)\right\|_2^2+\mathbb{E}\left[\|\boldsymbol{\zeta}\|_2^2\right], \end{aligned} \tag{19}$$

where the cross term vanishes because $\mathbb{E}[\boldsymbol{\zeta}]=0$. Applying the previous two bounds gives

$$\mathbb{E}\left[\left\|\boldsymbol{\alpha}\left(\boldsymbol{\mathcal{E}}^{\dagger}[O]\right)-\widehat{\boldsymbol{\alpha}}_K\left(\boldsymbol{\mathcal{E}}^{\dagger}[O]\right)\right\|_2^2\right]\le 2\left\|\boldsymbol{\alpha}\left(\boldsymbol{\mathcal{E}}^{\dagger}[O]\right)-\boldsymbol{\alpha}\left(\boldsymbol{\mathcal{U}}^{\dagger}[O]\right)\right\|_2^2+2E(K)^2+\frac{K}{2M}.$$

□

E Additional implementation details

In this section, we provide additional detail on metric description, the architecture of neural network in Section 3.2 and the Pauli propagation in Section 3.1.

E.1 Metrics setup

Mean Absolute Error (MAE) represents the average error magnitude between the mitigated expectation value and the actual ground truth. **Improvement Ratio (IR)** represents the number of successfully improved instance.

$$\text{MAE}=\frac{1}{N}\sum_{t=1}^{N}\|\hat{y}_t-y_t^{\text{ideal}}\|;\quad \text{IR}=\frac{\text{Number of improved instance}}{N}$$

E.2 Pauli propagation

In this work, we modify the Pauli propagation in Shao et al. (2025). The original work uses it to classically simulate ideal expectation value while we use it to select the support set for perturb direction.

E.3 Architecture setup

Table 6 reports the model dimensions and parameter counts used in the circuit-based, swap-test, and many-body experiments.

Algorithm 1 Adaptive Pauli support selection by Pauli propagation

Require: Ideal circuit $\boldsymbol{\mathcal{U}} = \mathcal{U}_L \circ \cdots \circ \mathcal{U}_1$, observable O, shot M
Ensure: Selected support $\mathcal{S}_{K^*}$, accumulated discarded error $E(K^*)$

$\mathrm{Score}_{\mathrm{best}} \leftarrow \infty$
for $K = 1, 2, \ldots, 4^n$ **do**
  $\widetilde{O}_L \leftarrow O$, $E(K) \leftarrow 0$
  for $l = L, L-1, \ldots, 1$ **do**
    $A_{l-1} \leftarrow 0$
    for each P with $\alpha(\widetilde{O}_l, P) \neq 0$ **do**
      $A_{l-1} \leftarrow A_{l-1} + \alpha(\widetilde{O}_l, P)\, U_l^\dagger P U_l$
    end for
    $\widetilde{O}_{l-1} \leftarrow \mathrm{TopK}_K(A_{l-1})$
    $\Delta_{l-1}(K) \leftarrow \sum_{P \in \mathcal{P}_n, \alpha(\widetilde{O}_{l-1}, P) = 0} |\alpha(A_{l-1}, P)|^2$
    $E(K) \leftarrow E(K) + \sqrt{\Delta_{l-1}(K)}$
  end for
  $\mathrm{Score}(K) \leftarrow 2E(K)^2 + K/2M$
  if $\mathrm{Score}(K) < \mathrm{Score}_{\mathrm{best}}$ **then**
    $\mathrm{Score}_{\mathrm{best}} \leftarrow \mathrm{Score}(K)$
    $K^* \leftarrow K$
    $E(K^*) \leftarrow E(K)$
    $\mathcal{S}_{K^*} \leftarrow \{P \in \mathcal{P}_n : \alpha(\widetilde{O}_0, P) \neq 0\}$
  end if
  if $E(K) \leq 0$ **then**
    break ▷ No truncation error remains
  end if
  if $(K+1)/2M \geq \mathrm{Score}_{\mathrm{best}}$ **then**
    break ▷ All later scores are at least this large
  end if
end for
return $\mathcal{S}_{K^*}, E(K^*)$

Table 6: Model size for each experimental setting.

(a) Circuit-based experiments.

	GTranQEM			QEM-STEP		
Experiment Name	**Layer**	**Hidden Dim**	**# of Params**	**PR dim**	**SC dim**	**# of Params**
All-6-Trotter, Mixed Noise	1	24	14665	64	128	14658
All-6-Trotter, Incoherent Noise	1	128	126465	64	128	14658
All-6-Trotter, Provider Noise	1	32	19425	64	128	14658
Max-6-Random, Mixed Noise	1	64	44865	32	32	3234
Max-6-Random, Incoherent Noise	2	64	82433	32	32	3234
Max-6-Random, Provider Noise	2	64	82433	32	32	3234
Circuit-Scale Extrapolation	1	64	44865	32	512	33954
Noise Amplification	1	64	44609	64	128	12738
6-Qubit, No Ideal Expectation Values	1	64	44865	64	128	14658

(b) Swap-test and many-body experiments.

Experiment Name	**DAEM**	**QEM-STEP**		
	# of Params	**PR dim**	**SC dim**	**# of Params**
Swap-test	1782785	128	512	102658
Swap-test (finite-shot)	1782785	128	512	102658
Many-body, Amplitude Damping	1783809	256	256	181506
Many-body, Amplitude Damping (finite-shot)	1783809	256	256	181506
Many-body, Phase Damping	1783809	256	256	181506
Many-body, Phase Damping (finite-shot)	1783809	256	256	181506