

Calibration-conditioned featurewise linear modulation decoders for low-latency decoding of quantum error correction evaluated on IBM repetition-code experiments

Samuel Stein ^{1,*} Shuwen Kan ^{1,2} Chenxu Liu ¹ Adrian Harkness ^{1,3} Sean Garner ^{1,4}
Zefan Du ² Yufei Ding ⁵ Ying Mao ² and Ang Li ^{1,4}

¹*Physical and Computational Sciences, Pacific Northwest National Laboratory, Richland, Washington, USA*

²*Department of Computer and Information Sciences, Fordham University, New York City, USA*

³*Department of Industrial and Systems Engineering, Lehigh University, Bethlehem, USA*

⁴*Department of Electrical & Computer Engineering, University of Washington, Seattle, USA*

⁵*Department of Computer Science & Engineering, University of California San Diego, San Diego, USA*



(Received 19 January 2026; accepted 9 July 2026; published 17 August 2026)

Real-time decoding of quantum error correction is essential for enabling fault-tolerant quantum computation. A practical decoder must operate with high accuracy at low latency, while remaining robust to spatial and temporal variations in hardware noise. We introduce a hardware-conditioned neural decoder framework designed to exploit the natural separation of timescales in superconducting processors, where calibration drifts occur over hours while error correction requires microsecond-scale responses. By processing calibration data through a graph-based encoder and conditioning a lightweight convolutional backbone via featurewise linear modulation (FiLM), we decouple the heavy processing of device statistics from the low-latency syndrome decoding. We evaluate this approach using the repetition code as a test bed on IBM *Fez*, *Kingston*, and *Pittsburgh* processors, collecting over 2.7 million experimental shots spanning distances up to $d = 11$. Training one decoder per code distance, basis, and round count (d, r), we demonstrate that each trained model generalizes to unseen qubit chains and new calibration data acquired days later without retraining. The FiLM-conditioned decoder achieves up to a $11.1\times$ reduction in logical error rate relative to modified minimum-weight perfect matching. Cross-conditioning ablations further show that the calibration input actively improves decoding, with performance degrading when the decoder is conditioned on incorrect backend data or the conditioning is disabled. We observe that by employing a network architecture that exploits the highly asynchronous nature of system calibration and decoding, hardware-conditioned neural decoding demonstrated promising, adaptive performance with negligible latency overhead relative to unconditioned baselines.

DOI: [10.1103/ftdh-t3yg](https://doi.org/10.1103/ftdh-t3yg)

I. INTRODUCTION

Quantum computers are inherently noisy, and this noise limits their ability to perform meaningful large-scale computation. Quantum error correction (QEC) is the leading path to suppressing physical errors to the levels required to execute computationally meaningful quantum algorithms [1,2]. By encoding logical qubits into a larger Hilbert space and repeatedly performing syndrome extraction, fault-tolerant protocols can exponentially reduce logical error rates (LERs) with increasing code distance [3,4]. Realizing this promise hinges not only on hardware but also on the accompanying classical stack. A decoder must correctly predict the required corrections at the cadence of repeated syndrome extraction such that we can drive computation [5]. If the decoder lags or produces incorrect corrections, the code-decoder setting forfeits its utility.

Beyond decoding accuracy, a challenge shared by practical QEC architectures, independent of the underlying hardware platform, is that decoders must adapt to noise that is spatially inhomogeneous and drifts over time, while maintaining accuracy and decoding performance. These demands are naturally decoupled by their data generation velocities: Syndrome data are produced at the cadence of syndrome extraction, whereas device characterization data are generated and refreshed on substantially slower timescales. A decoding strategy that adapts to the hardware state on the slow timescale, while keeping the fast decoding path lightweight, can serve both demands simultaneously. This separation of timescales is what we target and exploit in this work.

Within the hardware landscape of quantum computing, superconducting processors are a leading architecture for realizing large-scale fault-tolerant quantum computation. However, they face operational challenges pertinent to decoding: Qubit performance is spatially heterogeneous, noise channels are highly asymmetric, and the intrinsic speed of superconducting processors necessitates decoding latencies on the microsecond timescale. To maintain fault tolerance under these conditions, decoders must be high throughput, accurate, and capable of adapting to heterogeneous noise profiles without the operational cost of constant retraining.

*Contact author: samuel.stein@pnnl.gov

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

Accordingly, it is important to assess decoders in experimental settings, since commonly used simulated noise models often omit precisely the latent hardware dynamics that challenge decoder performance.

Publicly available IBM superconducting devices currently feature a heavy-hex lattice topology. Devices that support midcircuit reset connectivity are restricted to degree-2 and degree-3 qubit coupling, which precludes the native embedding of higher-weight stabilizer codes, such as the surface code, without significant compilation overhead. Consequently, the repetition code serves as a strong candidate for studying decoding dynamics on native hardware [6–10]. Although the repetition code protects only a single Pauli basis, it shares the stabilizer formalism foundational to all stabilizer codes, including Calderbank-Shor-Steane (CSS) codes [11]. Our use of the repetition code is motivated by the view that progress toward deployed QEC requires investigating and resolving decoding challenges in simpler systems first, as these challenges propagate upward in code complexity: Performant decoding of surface codes and other topologically local codes builds on the problems the repetition code already poses, just as high-weight nonlocal quantum low-density parity-check codes (qLDPC) build on competence with local codes. This, alongside the repetition code being natively accessible at scale on today’s hardware, motivates it as our experimental platform; extension to more complete QEC codes is discussed in Sec. V. Developing decoders that satisfy hardware operational requirements while performing well on these codes is a prerequisite for scalable QEC.

A practical decoder faces three simultaneous demands: (1) decoding throughput to match computational demand of the underlying code, (2) accuracy in matching error syndromes to the underlying error, and (3) adaptability to spatiotemporally varying noise and hardware without bespoke retuning and redesign. Classical analytic decoders such as minimum-weight perfect matching (MWPM) set strong baselines for topological codes [12], but they rely on an approximated error model, explicitly derived from hardware statistics; deviations from idealized assumptions such as bias, inhomogeneity, and correlated errors all can degrade performance.

Motivated by these trade-offs, multiple recent experiments have turned to machine-learning decoders that infer noise structure directly from experimental data [13,14]. Convolutional models exploit locality and translational invariance of detector events, while more expressive transformer [15] and graph models [16] can capture longer-range space-time correlations. For instance, a recurrent transformer-based decoder trained and then fine-tuned on real syndromes outperformed state-of-the-art algorithmic decoders on Google’s Sycamore data at distances $d = 3$ and $d = 5$, attaining a 2.748% LER against the MWPM-correlated 3.597% at $d = 5$ [13]. Graph-neural decoders trained on detector graphs have surpassed matching under circuit-level noise in simulation and reached parity with MWPM on experimental repetition-code data [16]. These results are compelling, and the open questions surrounding how these models generalize are well motivated. These models are typically trained on a single fixed device and noise distribution, requiring retraining or fine-tuning when conditions change. Furthermore, low-latency constraints can limit model complexity

and architectural choices. For a decoder to scale, it must generalize beyond the specific experiments on which it was trained.

We pursue a route to decoder adaptability that directly exploits this separation of timescales between hardware drift and decoding. Device characteristics such as coherence times and gate fidelities drift over hours or days [17], whereas syndrome decoding in superconducting systems operates at the microsecond scale. To exploit this, we employ *featurewise linear modulation* (FiLM) [18]. FiLM applies channelwise transformations to intermediate features based on external side information. By using the varying calibration data to generate these modulation parameters asynchronously, we can condition a lightweight convolutional backbone on the current device state without treating the hardware state as a real-time input. This strategy allows the decoder to adapt to spatial and temporal noise variations, while preserving the low-latency lightweight neural model required for the QEC cycle.

In this work, we propose a calibration-conditioned neural decoder that couples a graph-neural system encoder with a FiLM-modulated convolutional backbone operating on detection-event tiles. The graph encoder ingests per-qubit and per-edge calibration features (e.g., T_1/T_2 , readout assignment errors, gate error rates) and produces a latent conditioning vector that modulates the convolutional feature maps through FiLM layers. We train one model per (d, r) configuration (distance d and number of rounds r), with experiment data generated from three IBM devices and multiple contiguous physical qubit subsets, and evaluate the trained model a week later on contiguous sets of physical qubits with calibration data. The model has not seen these data during training. All results reported, including the validation set, are excluded from the training process and solely used for evaluation. In this work, we target generalization across hardware noise configurations, namely, unseen qubit chains and calibration snapshots, with each (d, r) configuration served by its own dedicated lightweight decoder; generalization across code distances and syndrome depths is a complementary direction for learned decoders that we leave to future work.

Our work draws on a large experimental corpus, comprising 2 760 704 repetition-code shots. We utilize 400 calibration snapshots from three IBM processors, spanning code distances up to $d = 11$. We evaluate the model on contiguous sets of physical qubits with calibration data acquired one week after training. The model maintains high performance without retraining. Specifically, we observe a performance crossover at $(d, r) \approx (7, 5)$ for the Z basis, beyond which the FiLM-conditioned decoder consistently outperforms comparative baselines. At the largest code size ($d = 11, r = 11$), the decoder achieves a $11.1\times$ improvement in logical error rate relative to hardware-informed MWPM.

In this work, we put forward three key contributions. *First*, we propose a decoding framework that effectively decouples the processing of slow-varying noise characteristics from fast-path syndrome decoding. By integrating a graph-neural hardware encoder with a FiLM-modulated convolutional backbone, we enable hardware-aware inference that adapts to changing device dynamics at latencies comparable with unmodified convolutional neural networks (CNNs). *Second*, we demonstrate that this calibration-conditioned

approach sustains logical-error suppression across both spatial and temporal hardware variations at fixed (d, r) . In experiment, the proposed decoder generalizes to unseen qubit chains and calibration snapshots acquired a week later, achieving up to a $7.41 \times$ reduction in logical error rate relative to minimum-weight perfect matching decoders. We further probe the conditioning mechanism directly through cross-conditioning ablations, deliberately mismatching the calibration input with randomized, training-period, or wrong-device calibration data, and observe that decoding performance degrades progressively as we scramble the conditioned information in the large-circuit regime. *Third*, we provide our experiment dataset and raw IBM experiment archives, comprising $\approx 2.7 \times 10^6$ repetition-code shots across 400 hardware snapshots for future works to utilize.

The remainder of this paper is organized as follows. Section II reviews the repetition code, superconducting-hardware noise biases, and prior decoding strategies. Section III details the FiLM-conditioned architecture and its integration with graph-based system features. Section IV describes the experimental pipeline and dataset, and presents performance comparisons and latency implications, including sensitivity analyses of FiLM parameters and cross-calibration transfer. Section V discusses implications for scalable QEC and extensions to higher-dimensional codes, and Sec. VI discusses natural extensions to this work and concludes.

II. BACKGROUND AND PROBLEM SETTING

In this section, we introduce the required preliminaries and notation of our problem setting. We refer the reader to Nielsen and Chuang [19] and Gottesman [4] for a detailed introduction to quantum error correction.

A. Quantum error correction

A stabilizer code encodes k logical qubits into n physical qubits by defining an Abelian subgroup $\mathcal{S} \subset \mathcal{P}_n$ of the n -qubit Pauli group $\mathcal{P}_n$ that excludes $-I$. The simultaneous $+1$ eigenspace of all elements in $\mathcal{S}$ forms the protected codespace. Logical operators act nontrivially on this codespace and are drawn from the normalizer $\mathcal{N}(\mathcal{S}) = \{P \in \mathcal{P}_n \mid PSP^\dagger = S\}$ but not from $\mathcal{S}$ itself, so that $\mathcal{N}(\mathcal{S})/\mathcal{S}$ defines the logical operators. The *distance* d of a stabilizer code is the minimum weight of any nontrivial logical operator, that is, the smallest number of physical qubits on which a logical operator acts nontrivially. These properties are often summarized as $[[n, k, d]]$. A code of distance d can detect up to $d - 1$ and correct up to $\lfloor (d - 1)/2 \rfloor$ arbitrary single-qubit errors on distinct qubits. Syndrome measurements project the system onto stabilizer eigenspaces, revealing parity information from which a decoder infers the most likely recovery operation.

Repetition code. The distance- d repetition code encodes one logical qubit ($k = 1$) into n physical data qubits. Let X_i and Z_i denote Pauli operators on data qubit i . For the bit-flip-protecting (Z -type) realization, the stabilizer group is generated by

$$S_Z = \langle Z_i Z_{i+1} : 1 \leq i \leq d - 1 \rangle, \quad (1)$$

with logical operator $\bar{X} = X^{\otimes d}$; the dual ($X_i \leftrightarrow Z_i$), phase-flip-protecting (X -type) realization interchanges X and Z with

$S_X = \langle X_i X_{i+1} : 1 \leq i \leq d - 1 \rangle$ and $\bar{Z} = Z^{\otimes d}$. In either case, the code corrects up to $\lfloor (d - 1)/2 \rfloor$ errors of the protected Pauli type [4]. Syndrome extraction is performed by measuring each stabilizer generator repeatedly using ancilla qubits. In this work, we implement this on hardware as a single dynamic circuit, where ancilla qubits are measured and reset in real time for r consecutive rounds. These measurements project the system into stabilizer eigenspaces and reveal where local parity constraints have been violated, forming the basis for the detection events that are used to decode and recover the corrected Pauli frame.

Syndrome extraction and detection events. Stabilizers are measured repeatedly for r rounds using parity checks over ancilla qubits. Let $s_{t,i} \in \{0, 1\}$ denote the parity outcome of stabilizer i at round t ($1 \leq t \leq r, 1 \leq i \leq d - 1$). Consecutive rounds are xored to form detection events,

$$\chi_{t,i} = s_{t,i} \oplus s_{t-1,i}, \quad s_{0,i} = 0. \quad (2)$$

This localizes changes caused by data or measurement faults. Single faults create short, local space-time chains of detection events; longer chains indicate multiple faults or measurement errors across rounds.

Decoding objective. Decoding is tasked with inferring what errors occurred that generated the syndrome map observed [20–22]. Given the full detection-event history χ and a hardware prior over Pauli errors $P(E)$, the information-theoretic optimum is the recovery map R^* that maximizes the probability of returning to the codespace without flipping the logical bit:

$$R^*(\chi) = \arg \max_{L \in \{\mathbb{I}, \bar{X} \text{ (or } \bar{Z})\}} \sum_{E \in \mathcal{P}_n: EL \in \mathcal{S}} P(E|\chi). \quad (3)$$

In this work, we target a per-qubit marginal estimate. The decoder outputs $p \in [0, 1]^d$ with $p_i \approx \Pr(E_i = 1|\chi)$ for the protected error channel. With a threshold $\tau = 0.5$, $\hat{c} = \mathbb{1}[p > \tau]$ updates a tracked Pauli frame, and the reported logical is extracted from the frame-adjusted final measurements.

B. IBM hardware

All experiments and results in this work are conducted on IBM superconducting quantum processors based on transmon qubits arranged in heavy-hex lattice topologies, specifically the IBM FEZ, KINGSTON, and PITTSBURGH. This architecture provides nearest-neighbor connectivity of weight 2 and 3 over a heavy-hex grid. Notably, this sparse connectivity prevents the native tiling of surface-code plaquettes or higher-weight connectivity, motivating the use of linear repetition codes as an effective low-overhead topological choice without incurring compilation or swapping overhead. IBM supports routine access to calibrated qubit parameters through IBM's open calibration interface. Each calibration snapshot includes coherence times (T_1, T_2), single- and two-qubit gate error rates, and readout assignment errors.

C. Comparable decoders

Neural network-based decoding has garnered attention for its potential to address the complex hardware noise profiles beyond standard algorithmic approaches. Krastanov and

Jiang [23] utilized deep neural networks to estimate conditional error probabilities for stabilizer codes under simulated depolarizing noise. Maskara *et al.* [24] applied feedforward networks to topological color codes, while Varsamopoulos *et al.* [25] investigated long short-term memory networks to compare high- and low-level decoding strategies in simulated environments. To address temporal correlations, Baireuther *et al.* [26] employed recurrent neural networks trained on density-matrix simulations of surface-code circuits. More recently, Lange *et al.* [16] utilized graph neural networks (GNNs) to capture the topological dependencies of quantum codes. While these studies demonstrate the viability of machine learning for decoding, they primarily rely on simulated data. It is increasingly recognized that a key advantage of machine learning is its ability to infer noise and error dynamics directly from experimental data [27–29], capturing device-specific and hardware-dependent effects that idealized models often omit by training on real noisy outputs rather than on preconceived noise models.

To evaluate our calibration-conditioned approach, we compare against two baselines, a modified MWPM decoder, and a CNN baseline. Both the modified MWPM decoder and the FiLM decoder consume the same calibration snapshot (T_1 , T_2 , gate error rates, and readout assignment errors) and comprise two models with differing inductive biases. The FiLM decoder is tasked with learning relationships between calibration data and observed syndrome statistics via statistical correlations in experiment data, allowing latent statistics and correlations that fall outside an analytic noise model to be captured.

1. Modified MWPM (calibration informed)

We derive a circuit-level noise model from the transpiled schedule executed on the chosen contiguous chain and the calibration snapshot. Specifically:

(1) *Gate errors*: Each one-qubit (two-qubit) gate is followed by a Pauli-twirled depolarizing channel with probability p_{1q} (p_{2q}) taken from the device calibration for that gate.

(2) *Measurement*: Readout assignment error is inserted using the calibrated assignment probability for the measured qubit(s).

(3) *Idling*: At the end of each round, every data qubit experiences a Pauli-twirled channel whose parameters are computed from (T_1 , T_2) and the effective idle/reset durations in that round.

From this model, we construct a detector graph whose edges connect pairs of detectors (or a detector to a boundary). Each edge is assigned a weight given by the negative log-likelihood derived from the corresponding fault probability. Running minimum-weight perfect matching on this graph returns a most likely set of errors [30]. We implement MWPM using PyMatching with edge fault probabilities derived from the circuit-level noise detector graph above. A separate detector graph and edge-weight assignment is constructed for each (device, basis, d , r , calibration snapshot) so that MWPM has access to the same calibration information as the FiLM decoder, namely, T_1 , T_2 , gate error rates, and readout assignment errors.

2. CNN baseline (unconditioned)

The CNN baseline is architecturally identical to the FiLM decoder's convolutional backbone in Fig. 1 and output head but excludes the hardware encoder and FiLM generator. The detection tensor $\chi \in \{0, 1\}^{r \times (d-1)}$ is fed directly through the convolutional blocks to produce per-qubit flip probabilities. We train one model per (d , r) and basis using the same training/validation split, loss, optimizer, and inference threshold as the FiLM model. This baseline therefore controls for network architecture and training protocol, isolating the contribution of calibration conditioning.

III. CALIBRATION-CONDITIONED FiLM DECODER ARCHITECTURE

ALGORITHM 1. Calibration-conditioned FiLM convolutional neural decoder architecture.

```

Input : Syndrome tensor  $\chi \in \{0, 1\}^{r \times (d-1)}$ .
         Hardware graph  $G$  with node calibration
         features.
Output: Per-qubit correction logits  $\hat{y} \in \mathbb{R}^d$ 

/* 1. System Encoder (GNN) - Embeds hardware
   properties */
1  $H^{(0)} \leftarrow G$ ; // Initial features from the graph
2 for  $\ell = 1$  to 3 do
3    $H^{(\ell)} \leftarrow \text{ReLU}(\text{GCNConv}_\ell(H^{(\ell-1)}))$ 
4  $\mathbf{z} \leftarrow \text{GlobalMeanPool}(H^{(3)})$ ; // Latent hardware
   vector  $\mathbf{z} \in \mathbb{R}^{256}$ 

/* 2. FiLM Generator (MLP) - Generates
   conditioning parameters */
// This branch is pre-computable
5  $\theta \leftarrow W_2 \cdot \text{ReLU}(W_1 \mathbf{z} + b_1) + b_2$ 
6  $(\gamma_1, \beta_1), (\gamma_2, \beta_2), (\gamma_3, \beta_3) \leftarrow \text{Partition}(\theta)$ ; // For CNN
   blocks 1, 2, 3

/* 3. Decoder Arm (FiLMed CNN) - Processes
   detection events */
7  $x \leftarrow \chi$ 
8 for  $\ell = 1$  to 3 do
   // Modulate features with hardware-specific
   // params
9    $x \leftarrow \text{ReLU}(\gamma_\ell \odot \text{Conv}_\ell(x) + \beta_\ell)$ 

/* 4. Output Head - Generates final logits */
10  $x_{\text{flat}} \leftarrow \text{Flatten}(x)$ 
11  $\text{logits}_{\text{flat}} \leftarrow W_{\text{head}} x_{\text{flat}} + b_{\text{head}}$ 
12  $\hat{y} \leftarrow \text{logits}_{\text{flat}}$ 
13 return  $\hat{y}$ 

```

We employ a calibration-conditioned FiLM decoder architecture motivated by three key requirements:

(1) *Generalization*: The decoder must be robust to spatial and temporal variations in noise without requiring retraining for every specific qubit chain or calibration cycle.

(2) *Latency*: The conditioning mechanism must add minimal computational overhead at inference time.

(3) *Correlations*: The model should leverage learned noise structures from data, capturing latent dynamics that uncorrelated hardware calibration statistics may omit.

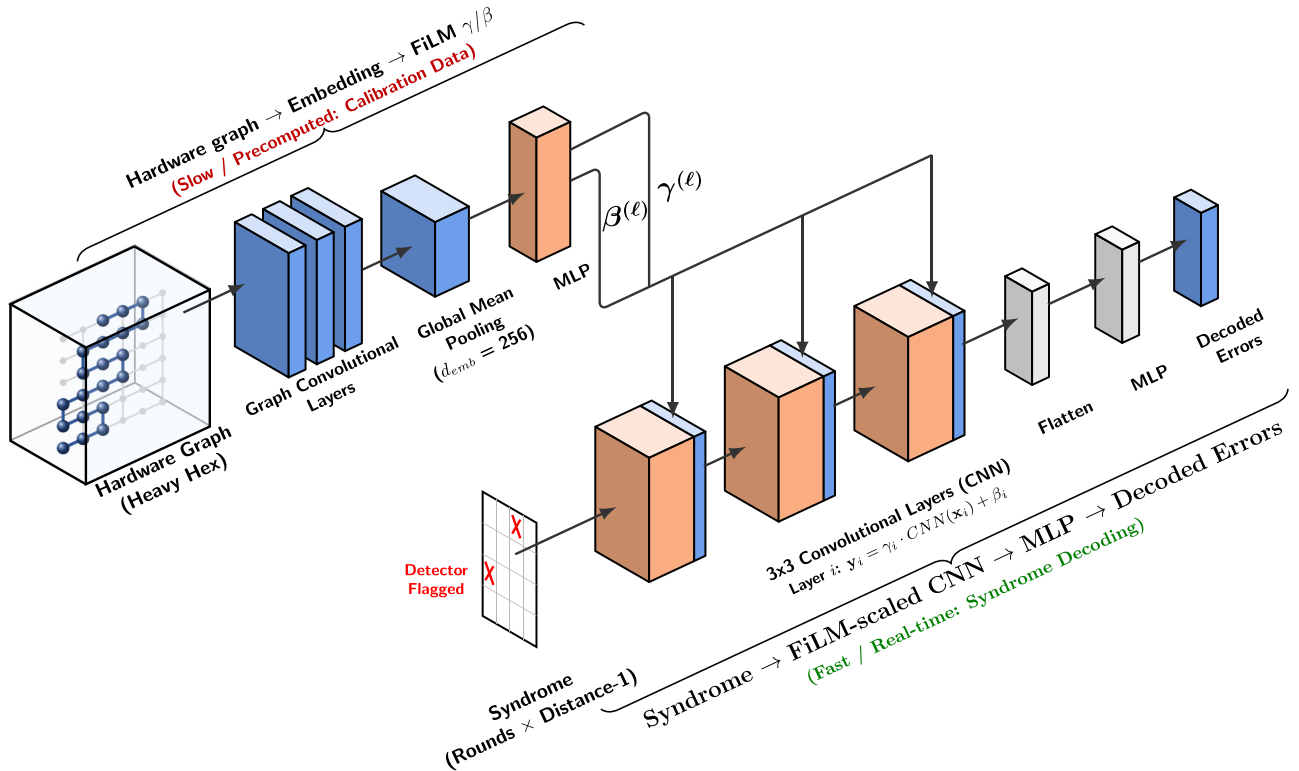


FIG. 1. Architecture of the calibration-conditioned FiLM neural decoder. The framework consists of three core components. (Upper left) Hardware encoder: An experimental shot defines a calibration subgraph $G = (V, E)$, extracted from the target IBM heavy-hex topology device, where node features include normalized T_1, T_2 , and gate errors. This graph is processed by a three-layer GCN and pooled to form a latent calibration embedding $\mathbf{z} \in \mathbb{R}^{256}$. (Upper right) FiLM generator: A multilayer perceptron (MLP) projects the latent embedding $\mathbf{z}$ into layerwise modulation parameters $\gamma^{(\ell)}$ (scale) and $\beta^{(\ell)}$ (shift). (Right) Decoder arm: The error syndrome tensor $\chi \in \{0, 1\}^{r \times (d-1)}$ is processed by a CNN. The feature maps of the CNN layers are modulated via the FiLM operation: $Z^{(\ell)} = \sigma(\gamma^{(\ell)} \odot \text{Conv}(Z^{(\ell-1)}) + \beta^{(\ell)})$. The output CNN features are flattened and mapped to per-qubit correction probabilities via a dense output head.

Our neural architecture is outlined in Fig. 1. Each experimental shot provides a detection-event tensor and a hardware graph of the underlying qubits on which the experiment is run. The neural decoder factors into three core components: (1) a hardware encoder that embeds the calibration subgraph into a learned latent vector representation, (2) a FiLM generator that maps this latent vector context to per-convolutional layer feature map scale/bias, and (3) a central convolutional neural pipeline that consumes detection events generated by the underlying experiment modulated by FiLM.

For a repetition-code experiment of distance d , comprising $2d - 1$ contiguous physical qubits and r rounds, we form a detection tensor $\chi \in \{0, 1\}^{r \times (d-1)}$ (time, stabilizer index) and a calibration subgraph $G = (V, E)$. This subgraph is extracted from the backend's heavy-hex lattice, visualized as the highlighted chain in Fig. 1 (HARDWARE GRAPH), in representing the specific contiguous qubit chain on which the code is executed. Node features include normalized (T_1, T_2) , gate errors, and readout-assignment error. The network outputs per-qubit flip probabilities $p \in [0, 1]^d$ for the protected error channel. Thresholding at $\tau = 0.5$ yields a binary correction frame $\hat{c} = \mathbf{1}[p > \tau]$ that is xored into the tracked Pauli frame before readout. Final logical state is decoded by xoring the data-qubit measurement at the end of the experiment with the tracked Pauli frame and majority voting.

A. FiLM generator and system encoder

We embed the calibration subgraph $G = (V, E)$, whose vertices correspond to the qubits in the contiguous chain and whose node features include normalized (T_1, T_2) , readout assignment error, and single-qubit gate error rates, with edge features given by two-qubit gate error rates. A three-layer graph convolutional network (GCN) with ReLU activations and global mean pooling produces a hardware-dependent latent embedding

$$h = \frac{1}{|V|} \sum_{v \in V} \text{GCN}(X_V, E)_v \in \mathbb{R}^{d_{\text{emb}}}, \quad (4)$$

using an embedding dimension of $d_{\text{emb}} = 256$.

While one could concatenate calibration features to the syndrome input, we employ FiLM as a more effective strategy to modulate internal feature maps of the underlying CNN. This allows the network to reweight the decoder's filters based on the hardware state, adjusting the processing logic of the detection events rather than appending additional inputs, increasing per-shot compute.

The latent vector is passed through a two-layer MLP that produces FiLM conditioning parameters for each convolutional block $\ell \in \{1, 2, 3\}$:

$$\{(\gamma^{(\ell)}, \beta^{(\ell)})\}_{\ell=1}^3 = f_{\theta}(h), \quad \gamma^{(\ell)}, \beta^{(\ell)} \in \mathbb{R}^{C_{\ell}}, \quad (5)$$

where C_ℓ is the channel width of block ℓ and (γ, β) broadcast over spatial dimensions. As these parameters depend only on the hardware graph G and not the detection tensor χ , they are precomputable.

Given FiLM parameters, the detection-event tensor $\chi \in \{0, 1\}^{r \times (d-1)}$ is processed by a sequence of FiLM-modulated convolutional blocks. Each block applies the same generic transformation

$$Z^{(\ell)} = \sigma(\gamma^{(\ell)} \odot \text{Conv}^{(\ell)}(Z^{(\ell-1)}) + \beta^{(\ell)}), \quad Z^{(0)} = \chi. \quad (6)$$

The learned parameters γ and β apply a channelwise transformation to the feature maps, amplifying or suppressing specific features based on the hardware calibration state.

After flattening the final feature map $Z^{(3)}$, a linear output head produces per-qubit flip probabilities

$$\hat{p}_i = \text{Pr}(\text{flip on data qubit } i | \chi, G), \quad 1 \leq i \leq d. \quad (7)$$

Thresholding at $\tau = 0.5$ yields the correction frame $\hat{c} = \mathbf{1}[\hat{p} > \tau]$, which is xored into the tracked Pauli frame before extracting the logical state.

B. Training objective and protocol

We optimize binary cross entropy over per-qubit flip targets, asking only:

“Given the detection history and calibration context, what is the probability that data qubit i ’s target Pauli frame should be XORED?”

Formally, with predictions $\hat{p}_{n,i} \in [0, 1]$ for shot n and data qubit i , we compute the binary cross entropy as

$$\mathcal{L} = -\frac{1}{N} \sum_{n=1}^N \sum_{i=1}^{d_n} [y_{n,i} \ln \hat{p}_{n,i} + (1 - y_{n,i}) \ln(1 - \hat{p}_{n,i})]. \quad (8)$$

At inference, thresholding $\hat{c} = \mathbf{1}[\hat{p} > \frac{1}{2}]$ updates the tracked Pauli frame before computing the logical state.

The network never receives the target logical state, final data-qubit readouts, or the logical outcome. Each training example consists only of the detection tensor $\chi \in \{0, 1\}^{r \times (d-1)}$ and the calibration subgraph G . Targets $y \in \{0, 1\}^d$ are obtained from the observed measurements and the corresponding corrections that recover the initial logical state.

For each experimental shot n , we construct a per-data-qubit flip target $y_n \in \{0, 1\}^d$. Let

$$q_n = (q_{n,1}, \dots, q_{n,d}) \in \{0, 1\}^d$$

denote the prepared logical bit string, with $q_{n,1} = q_{n,2} = \dots = q_{n,d}$ for our repetition-code experiments, and let

$$m_n = (m_{n,1}, \dots, m_{n,d}) \in \{0, 1\}^d$$

denote the corresponding measured bit string on the data qubits at the final round. We define

$$y_n = q_n \oplus m_n,$$

where $\oplus$ denotes bitwise xor. Equivalently, $y_{n,i} = 1$ if the measurement on data qubit i disagrees with the prepared logical value and $y_{n,i} = 0$ otherwise. Thus, y_n specifies the Pauli-frame update that, when applied to m_n , recovers the prepared logical state q_n , and this is the Pauli-frame update

that the decoder is trained to predict from the detection history and calibration features.

With respect to training, for each basis (X and Z separately) and each (d, r) , we merge shots across IBM *Fez*, *Kingston*, and *Pittsburgh* and split 70%/30% for train/validation. We use Adam (initial learning rate 5×10^{-3}) with cosine annealing for 100 epochs. Model parameters are described in Table II.

We compare against (1) CNN only (FiLM disabled; same network backbone, loss, and training) and (2) a noise-aware MWPM baseline reported in Sec. IV.

C. Complexity of FiLM decoding architecture

A significant advantage of the FiLM architecture is the complete decoupling of the hardware-conditioning overhead from the real-time decoding loop. As the FiLM parameters $\{\gamma^{(\ell)}, \beta^{(\ell)}\}$ are derived solely from the calibration graph G , they do not need to be recomputed for every syndrome tensor χ .

We employ a weight folding strategy, whereby when the convolutional feature map must be updated, the newly generated modulation parameters $(\{\gamma^{(\ell)}, \beta^{(\ell)}\})$ are mapped directly into the respective convolutional weights and biases. This transformation converts the conditional operations into standard convolutions with updated convolutional kernel values. Consequently, the per-shot inference path maintains computational complexity identical to the unconditioned CNN baseline, allowing the decoder to benefit from hardware-aware feature maps with no additional latency penalty during active decoding.

IV. RESULTS

A. Experimental setup and evaluation

We evaluate repetition-code decoders for code distances $d \in \{3, 5, 7, 9, 11\}$ and syndrome depths $r \in \{1, 3, \dots, d\}$ in both X - and Z -basis encodings, using IBM *Fez*, *Kingston*, and *Pittsburgh* processors (heavy-hex connectivity). A total of 2 760 704 shots were collected over 400 contiguous-qubit-chain calibration snapshots.

The data are used as follows. For each (d, r) and basis, all training-period shots matching that configuration are pooled across devices and contiguous qubit chains, and randomly partitioned into a 70:30 train/validation split at the shot level. The 70% partition is used for training and the 30% used for validation, serving as a validation split during training and the held-out validation data split results reported in Sec. IV B. Entirely separate from both partitions, a second collection of experiments was executed one week later on different contiguous qubit chains, selected independently of the training data, after the devices had been recalibrated; these unseen recalibrated datasets are used purely for evaluation in Sec. IV C and are never used for training, model selection, or hyperparameter tuning. Per (d, r) and basis configuration, the unseen recalibrated datasets comprise between 16 384 and 24 576 shots. As such, all decoding performance reported is on data the models were not trained on. All experiments follow the high-level structure depicted in Fig. 2, adjusting the PREP, describing the choice of initial logical state, SYNDR, EXTR, based on the repetition code form, and R, defined as the

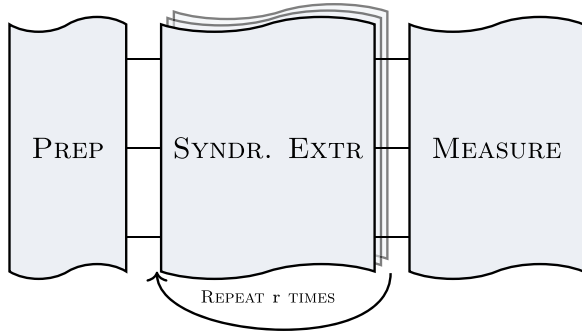


FIG. 2. Overall repetition-code process deployed, where PREP initializes the logical code word, either $|0\rangle_L$ or $|1\rangle_L$ for a Z-repetition code, or $|+\rangle_L$ or $|-\rangle_L$ for an X-repetition code. Syndrome extraction is repeated r times, after which the data qubits are measured. See Fig. 2 for a compiled example over IBM FEZ basis gates.

number of syndrome extraction rounds. Prior to measurement, we map the target observable of each qubit onto the Z basis. An example logical $|+\rangle$ experiment is outlined in Fig. 3, whereby we prepare the $|+\rangle$ state, measure $d - 1$ syndromes over r rounds, and finally measure the data qubits. The information attained from the syndromes is decoded and xored with the data qubits to attain a final logical state measurement.

We compare FiLM+CNN to (1) MWPM and (2) CNN (identical convolutional backbone with FiLM disabled, trained without FiLM). We report LER and plot results with binomial 95% confidence intervals. To quantify relative changes, we use

$$\rho_{A/B} \equiv \frac{\text{LER}_A}{\text{LER}_B}, \quad \mathcal{G}_{A \leftarrow B} \equiv \frac{1}{\rho_{A/B}} = \frac{\text{LER}_B}{\text{LER}_A},$$

so that $\rho_{A/B} < 1$ ($\mathcal{G}_{A \leftarrow B} > 1$) indicates that decoder A improves on B.

B. Validation data results

We train one decoder model per (d, r) pair on 70% of the training data, and first evaluate logical error rates on the 30% validation data. Figures 4 and 5 plot LER versus r for all

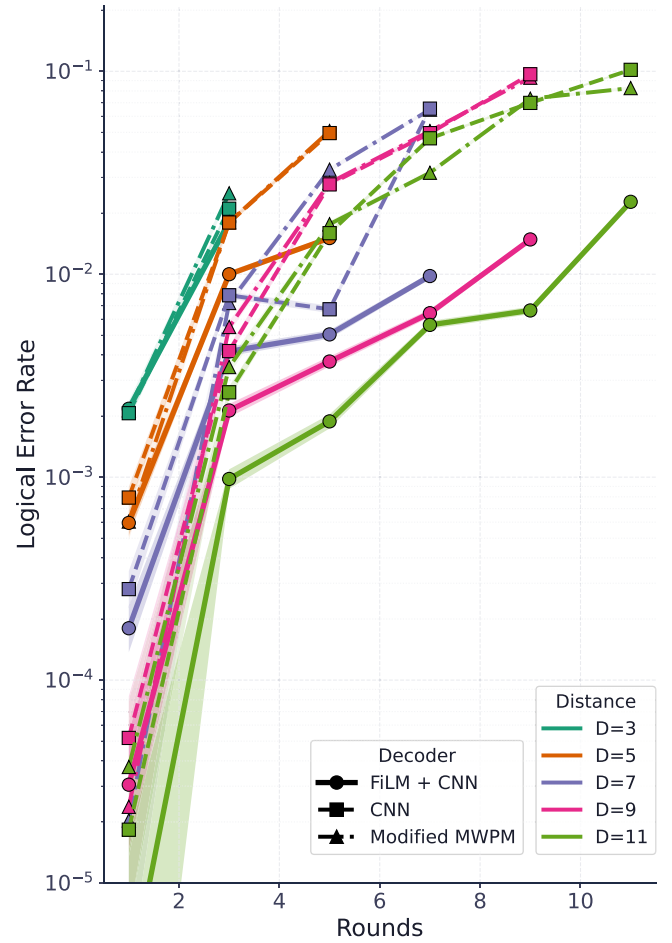


FIG. 4. Validation results: Logical error rates in the X basis as a function of measurement rounds. Logical states $|+\rangle_L$ and $|-\rangle_L$ are tested. Results are shown for code distances $d = 3, 5, 7, 9, 11$ for experiment rounds $r = 1, 3, \dots, d$ states. We compare the performance of the proposed FiLM+CNN decoder (solid lines) against a standard CNN (dashed lines) and modified MWPM (dash-dotted lines).

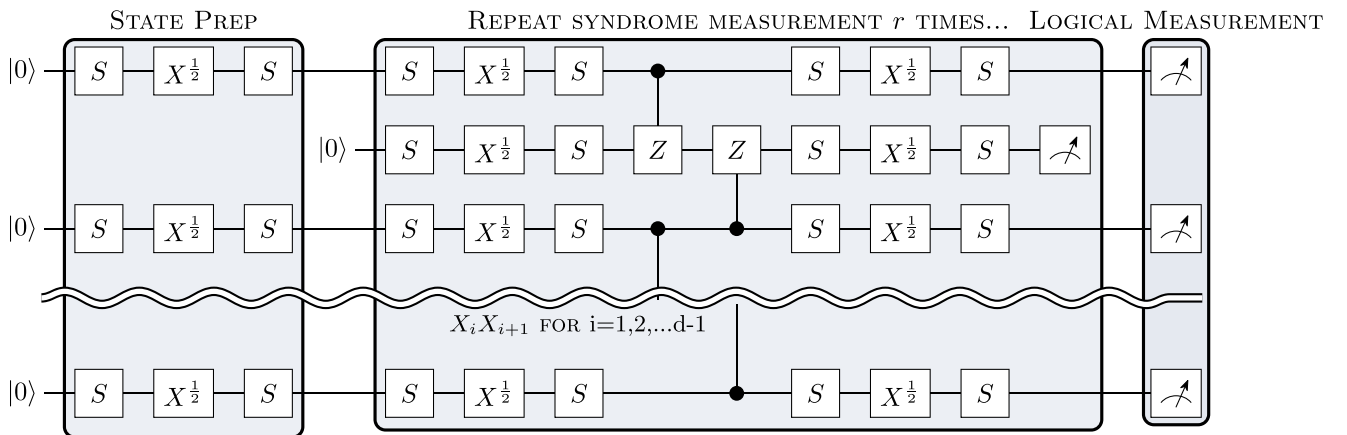


FIG. 3. Compiled syndrome extraction circuit onto device hardware for X-repetition code for a distance d repetition code with r syndrome extraction rounds. State prep in this example prepares the logical $|+\rangle$ state.

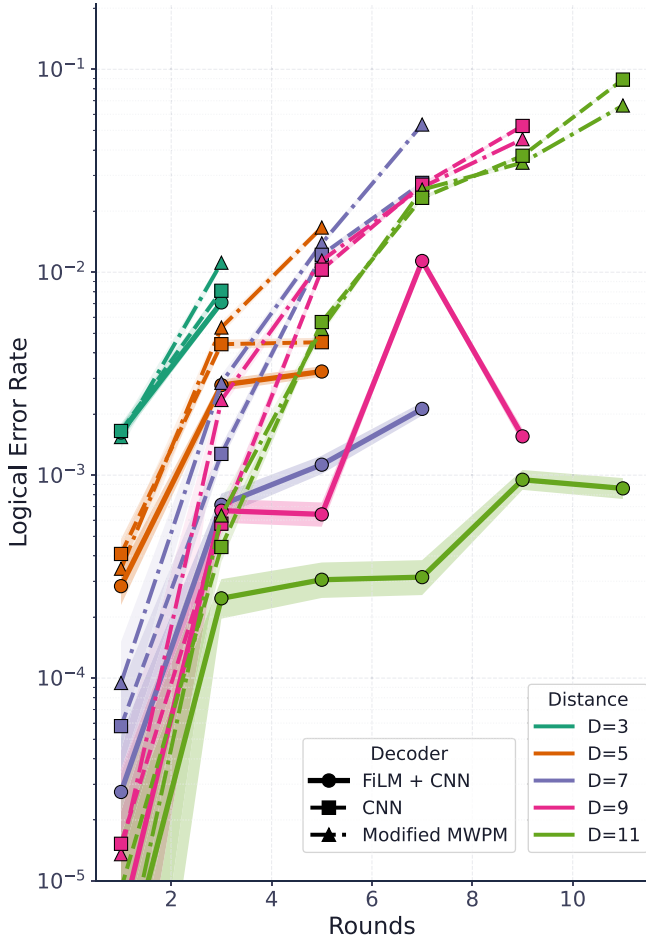


FIG. 5. Validation results: Logical error rates in the Z basis as a function of measurement rounds. Logical states $|0\rangle_L$ and $|1\rangle_L$ are tested. Results are shown for code distances $d = 3, 5, 7, 9, 11$ for experiment rounds $r = 1, 3, \dots, d$ states. We compare the performance of the proposed FiLM+CNN decoder (solid lines) against a standard CNN (dashed lines) and modified MWPM (dash-dotted lines).

(d, r) , with per-configuration values tabulated in Tables III and IV. Data points for distance d experiments over rounds $r \in \{1, 3, \dots, d\}$ are plotted. We observe two trends:

(1) *Shallow circuits favor MWPM*: For small d and r , the analytic model in MWPM is competitive or best, especially in the X basis.

(2) *Deeper circuits favor FiLM+CNN*: As d and r increase, FiLM+CNN overtakes both MWPM and CNN. At $(d, r) = (11, 11)$, we observe a $\mathcal{G}_{\text{FiLM} \leftarrow \text{MWPM}} = 11.11\times$ reduction in LER over MWPM in the Z basis and a $\mathcal{G}_{\text{FiLM} \leftarrow \text{MWPM}} = 5.92\times$ reduction in the dual X basis (Tables III and IV).

At shallow code depths and few syndrome-extraction rounds, the MWPM decoder retains an advantage. At these depths and small distances, the error graph is sparse. MWPM performs well in these regimes, where underlying noise and error profile match the generating features for MWPM. Neural networks, conversely, potentially have (1) less data to train on in these regimes due to lower χ dimensionality, though more importantly, (2) while there also being fewer “complex” error

patterns to learn as non-Markovian errors and other parasitic correlated noise channels are yet to accumulate. As the circuit volume grows, these assumptions that underpin MWPM start to break down. Temporal and spatial correlations compound, producing error structures that pairwise matching struggles to capture. In the larger-system regime ($d \geq 7, r \geq 5$), the learned decoders, and especially the hardware-conditioned FiLM+CNN model, consistently outperform both MWPM and the unconditioned CNN. With respect to the unconditioned CNN model, CNN improves over MWPM in several mid- and large-volume regimes, particularly in the Z basis. However, FiLM+CNN consistently achieves lower LER than CNN across most larger-volume experiments with $r \geq 5$ in our experiments on the validation dataset, reflecting the benefit of calibration-aware modulation over a purely data-driven but unconditioned convolutional backbone.

We observe the Z-basis configuration achieves a median gain of $4.55\times$ (range $1.99\times$ to $11.11\times$) over MWPM in the large-system regime ($d \geq 7, r \geq 5$), whereas the X basis achieves a median gain of $3.19\times$ (range $1.69\times$ to $8.77\times$).

The Z-basis experiments outperforming the X-basis experiments aligns with the noise characteristics of the underlying hardware, where $T_1 > T_2$ across the majority of experiments. While the Z-repetition code detects bit flips (X errors) and is theoretically robust to dephasing errors, our FiLM feature analysis (Tables X and XI) indicates that T_2 remains a significant driver of feature modulation in the early layers of the Z-basis decoder. This sensitivity likely arises from the hardware compilation of the syndrome extraction circuit, whereby the native gate set requires basis-changing operations, such as CZ gates, transiently exposing the Z code to dephasing errors during syndrome extraction. Consequently, the decoder appears to learn that T_2 is a relevant predictor of correlated errors even for the Z code. In the X basis, which is continuously sensitive to dephasing errors during idling, the model exhibits a similar pattern, with T_1 being a relevant feature. This suggests that the hardware-conditioned architecture attempts to learn the latent dynamics of the circuit implementation, evident by the impact of transient noise during syndrome extraction, rather than relying on an error model based solely on the code’s target logical protection. This transient noise exposure can be seen in Fig. 3 within the syndrome extraction block.

C. Unseen recalibrated experiment results

To verify that the performance gains observed on the validation split are not an artifact of train-validation partitioning, and that the decoder adapts to unseen hardware calibrations, we evaluate the trained decoders on experiments executed on different contiguous qubit chains taken from the IBM KINGSTON one week later. These chains were selected independently of the training data and have been recalibrated multiple times, resulting in different hardware noise fingerprints. The decoders are never retrained or fine-tuned between these datasets. Results are plotted and tabulated in Figs. 6 and 7 and Tables V and VI.

We observe that the models trained continue to outperform MWPM and CNN in experiments, observing a consistent performance crossover at $\approx(7, 5)$ for Z-basis experiments and $\approx(7, 7)$ for the dual X experiments, beyond which the

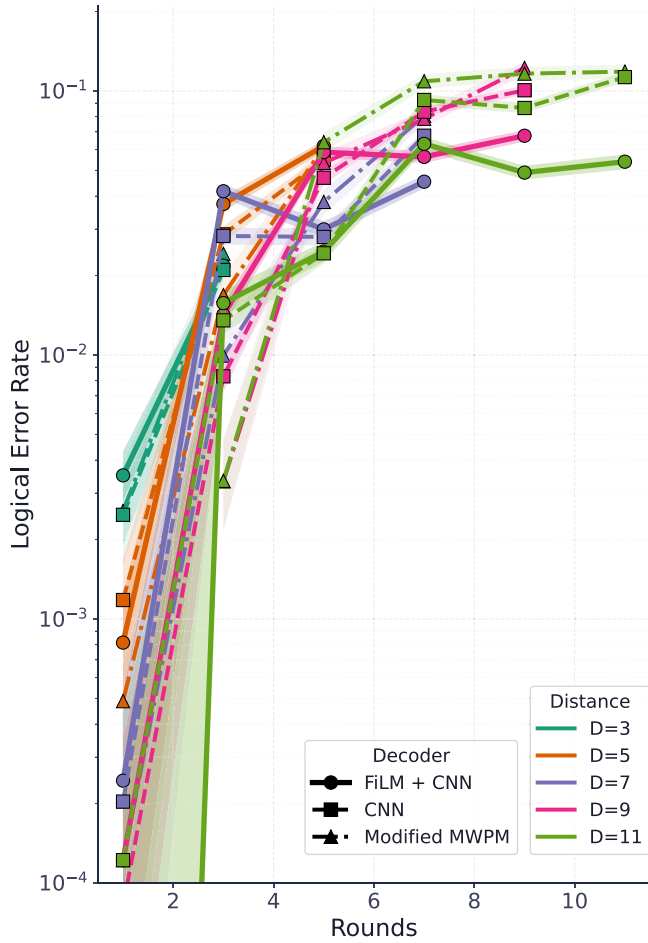


FIG. 6. Unseen recalibrated results: Logical error rates in the X basis as a function of measurement rounds for $D \in \{3, 5, 7, 9, 11\}$. FiLM+CNN (solid), CNN only (dashed), and Modified MWPM (dash dot). Shaded bands: 95% binomial confidence interval (CIs) across calibration snapshots.

learned decoder consistently outperforms the baselines. At $(d, r) = (11, 11)$, we reproduce the strong gains observed on the validation split: FiLM + CNN reduces LER by $7.41\times$ in the Z basis and $2.19\times$ in the X basis relative to MWPM (Tables V and VI).

Performance improvements past (d, r) pairs of $(d \geq 7, r \geq 5)$ is observed to be as follows:

(1) Z basis: FiLM+CNN improves on MWPM past this regime with a gain of $\mathcal{G}_{\text{FiLM} \leftarrow \text{MWPM}} = 1.89\times$ at $(7, 5)$, up to $\mathcal{G}_{\text{FiLM} \leftarrow \text{MWPM}} = 7.41\times$ at $(11, 11)$.

(2) X basis: FiLM+CNN improves on MWPM past this regime with a gain of $\mathcal{G}_{\text{FiLM} \leftarrow \text{MWPM}} = 1.27\times$ at $(7, 5)$, up to $\mathcal{G}_{\text{FiLM} \leftarrow \text{MWPM}} = 2.19\times$ at $(11, 11)$.

Comparing to the unconditioned CNN model, we observe that the FiLM+CNN decoder yields up to a $8.33\times$ reduction in LER for the Z basis and up to a $2.09\times$ reduction for the X basis at $(d, r) = (11, 11)$ (Tables V and VI). This highlights the specific value of calibration conditioning in high-noise regimes, where the unconditioned model struggles to generalize.

These results suggest that FiLM's calibration-driven modulation has learned a form of hardware-adaptive feature

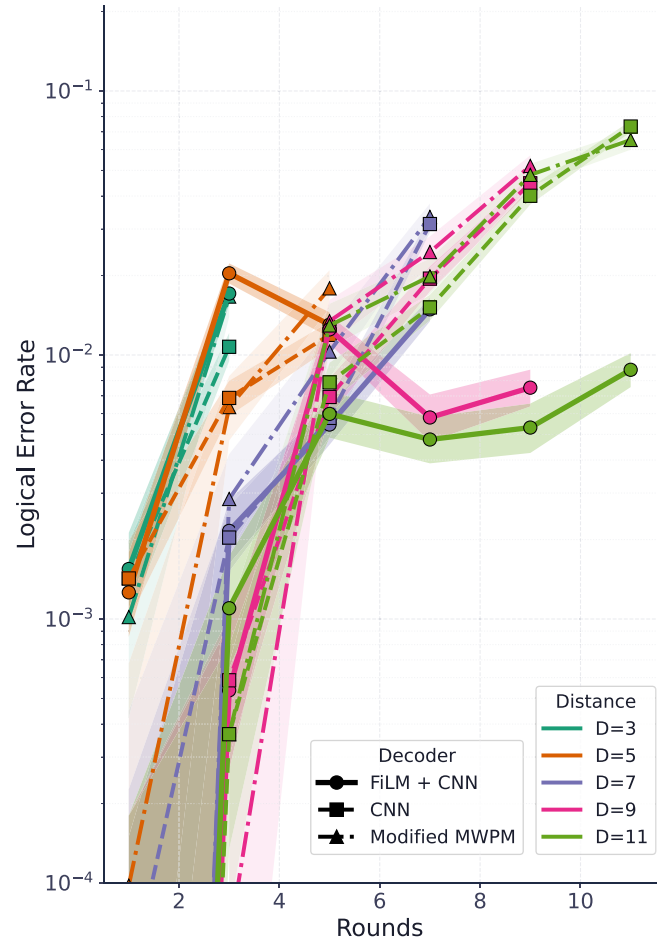


FIG. 7. Unseen recalibrated results: Logical error rates in the Z basis matching Fig. 6. Separation between FiLM+CNN and baselines increases at larger r and d , consistent with phase-noise bias.

selection. The convolutional backbone learns a rich set of reusable spatial-temporal features, while the FiLM parameters selectively amplify or suppress channels depending on the current device calibration. This mechanism helps explain the capability to generalize across days and qubit-chain selections without retraining.

D. Cross-conditioning ablation: Quantifying the role of calibration information

In this subsection, we investigate how the calibration information supplied through FiLM affects decoding performance. We perform a cross-conditioning ablation on the unseen recalibrated Kingston datasets, reevaluating each trained FiLM decoder while the conditioning input is perturbed or mismatched and the syndrome data are held fixed. We consider four conditioning inputs. The first adds zero-mean Gaussian noise with standard deviation $\sigma = 0.25$ or $\sigma = 1.0$ to the standardized calibration graph features, where $\sigma = 1.0$ noise is of the same magnitude as the features themselves. The second conditions with the stale training-period Kingston calibration snapshot, acquired approximately one week before the evaluation data. The third conditions with calibration graphs drawn from matched (d, r) experiments on FEZ or PITTSBURGH. The

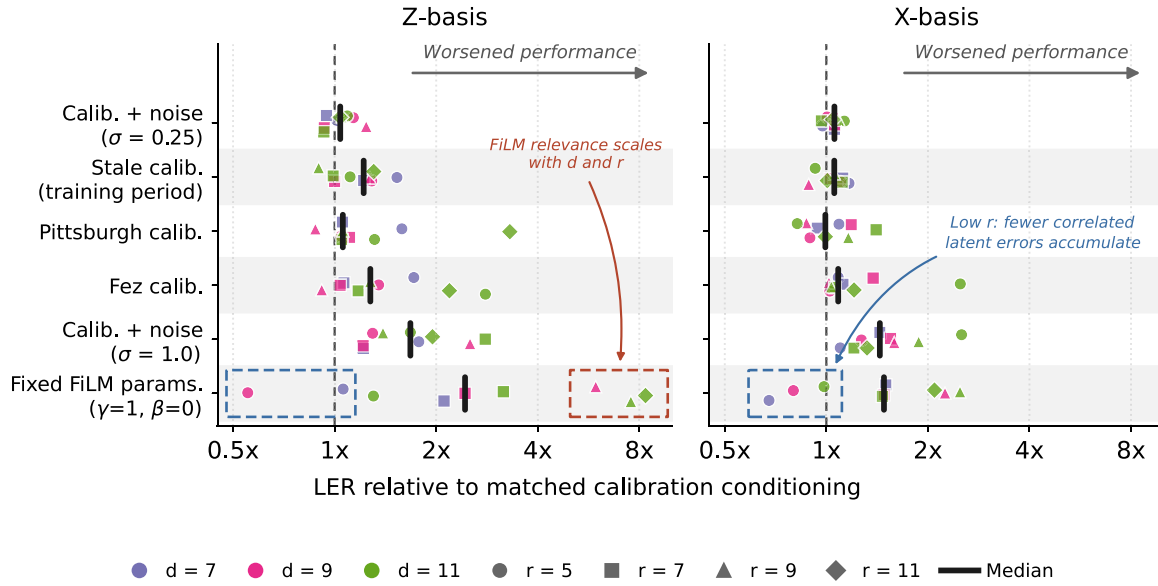


FIG. 8. Cross-conditioning ablation on unseen recalibrated KINGSTON data, for the large-circuit regime ($d \geq 7$, $r \geq 5$). Each marker is one (d, r) configuration, placed at the LER of the corrupted-conditioning decoder relative to the matched-conditioning decoder (logarithmic axis; the dashed line at $1\times$ denotes parity with matched conditioning). Marker color encodes code distance d and marker shape encodes the number of syndrome rounds r . Conditioning corruptions are ordered top to bottom by increasing severity; black bars mark per-variant medians. Decoding degrades progressively as the conditioned calibration information is scrambled, and within each corruption the penalty grows with circuit volume: The deepest circuits (triangles and diamonds) and $d = 11$ configurations carry the largest ratios. Dashed boxes highlight the two extremes of this trend for the decoder with fixed FiLM parameters. At low r (blue), few correlated latent errors have accumulated and conditioning matters least, whereas at the largest d (green) and r (diamond), the learned latent noise structure is most consequential.

fourth fixes the FiLM parameters to $\gamma = 1$, $\beta = 0$, bypassing the conditioning pathway.

Results are summarized in Fig. 8. All results are reported as the ratio of the perturbed-conditioning LER to the matched-conditioning LER for the same configuration, such that values above unity indicate degradation attributable to the conditioning input alone. Table VII condenses these results, and per-configuration ratios are tabulated in Tables VIII and IX. In the large-circuit regime ($d \geq 7$, $r \geq 5$), Gaussian noise at $\sigma = 0.25$ shifts the LER by a median ratio of $1.05\times$. Stale and wrong-device calibration yield median ratios from approximately $0.99\times$ to $1.28\times$, with the largest penalties concentrated at $d = 11$, reaching $3.3\times$. Gaussian noise at $\sigma = 1.0$ yields a median ratio of $1.50\times$ and degrades all but three configurations. Fixing the FiLM parameters is the most damaging, with a median ratio of $1.80\times$ and a maximum of $8.3\times$, collapsing the decoder onto the unconditioned CNN baseline, whose learned corrections in this regime rarely alter the majority-vote outcome of the final data-qubit readout. Within each conditioning input, the largest ratios are carried by the deepest circuits, consistent with conditioning errors compounding over rounds.

We observe that the conditioning pathway actively informs decoding. Perturbing the calibration input at the scale of the calibration features themselves degrades logical error rates in the regime where conditioning provides its benefit, and the penalty grows with both perturbation strength and circuit volume. Smaller perturbations, week-old snapshots, and same-generation wrong-device calibrations shift performance more mildly. This suggests the decoder interpolates smoothly across calibration space rather than depending on an exactly

matched calibration input, consistent with the generalization to recalibrated devices observed in Sec. IV C, and with FiLM learning a calibration-indexed family of decoding policies rather than a single static rule. In the shallow regime ($d < 7$ or $r < 5$), conditioning confers no measurable benefit, consistent with the performance crossover discussed in Sec. IV B.

E. Inference latency

To quantify the computational overhead of calibration conditioning, we benchmark the per-shot forward-pass latency. We compare three configurations to isolate the cost of the FiLM mechanism:

- (1) *No-FiLM (CNN)*: The baseline convolutional decoder without hardware conditioning.
- (2) *FiLM (Folded)*: Calibration-dependent parameters (γ, β) are precomputed and mapped into the convolutional weights.
- (3) *FiLM (Dynamic)*: Hardware encoder (GNN) and FiLM generator are executed for every sample.

Benchmarks were conducted using NVIDIA RTX 5000 graphics processing unit (GPU). We report the mean and standard deviation over 2000 decoding iterations following a 500 iteration warm-up.

As detailed in Table I, the FiLM (dynamic) approach increases latency by over an order of magnitude, resulting in a per-shot latency of ≈ 1.4 ms, with execution latency dominated by the graph neural network execution. However, the FiLM (folded) strategy successfully eliminates this overhead. The folded latencies ($\approx 85\text{--}95\text{ }\mu\text{s}$) track the no-FiLM baseline within the reported measurement error. Notably, latency

TABLE I. Per-shot inference latency (batch size = 1). Comparison of execution time across decoder strategies. *Dynamic* computes FiLM parameters per shot; *folded* precomputes them, baking the modulation into the CNN weights. The folded approach incurs negligible overhead compared to the unconditioned baseline.

(d, r)	Decoder mode	Avg (μ s)	Std (μ s)
(3, 3)	FiLM (dynamic)	1419.2	280.9
	FiLM (folded)	96.7	27.0
	No-FiLM (CNN)	87.4	18.1
(5, 5)	FiLM (dynamic)	1419.8	151.4
	FiLM (folded)	84.9	11.7
	No-FiLM (CNN)	83.3	14.5
(7, 7)	FiLM (dynamic)	1430.1	227.0
	FiLM (folded)	84.4	18.5
	No-FiLM (CNN)	84.9	16.2
(9, 9)	FiLM (dynamic)	1426.7	203.7
	FiLM (folded)	93.0	17.3
	No-FiLM (CNN)	98.0	13.3
(11, 11)	FiLM (dynamic)	1380.1	185.2
	FiLM (folded)	84.6	15.8
	No-FiLM (CNN)	81.1	12.4

remains relatively constant as the code distance scales from $d = 3$ to $d = 11$; in this regime, run-time is dominated by fixed GPU kernel launch overheads rather than computational complexity. This implies that the computational workload is minimal, and on dedicated control hardware such as field programmable gate array (FPGAs) or application specific integrated chipset (ASICs), which eliminate kernel scheduling overheads, we anticipate the latency would drop significantly, governed by the logic and depth of the network [31,32]. These results highlight that the proposed architecture can embed hardware-conditioning performance gains with next to zero additional latency cost at inference time.

F. Interpreting FiLM features for decoding

To understand how the graph encoder modulates the convolutional backbone, we analyze the sensitivity of the generated FiLM parameters $\phi = \{\gamma, \beta\}$ to the input calibration features $\mathbf{f}$. We compute the full experimental dataset averaged Jacobian $\mathbf{J} = \mathbb{E}[\partial\phi/\partial\mathbf{f}]$ and perform a singular value decomposition (SVD) to extract the dominant modes of the Jacobian.

Detailed statistics are reported in Table XI (Z basis) and Table X (X basis). These tables present the singular values (σ) indicating the mode's strength and the normalized projection of the singular vector onto specific hardware features ($T_1, T_2, \epsilon_g, \epsilon_r$). Focusing on the largest (d, r) size ($d = 11$), we observe three distinct features:

(1) *Physical feature extraction (Layer 1)*: Across both experiments, the first convolutional layer acts as a physical feature extractor, modulating filters based on mixtures of coherence parameters or gate and measurement errors. In the X basis (Table X), the dominant mode ($\gamma = 1.81, \beta = 3.18$) exhibits a coupled sensitivity, driven primarily by dephasing ($T_2, \gamma = +0.81, \beta = +0.83$) but significantly counterweighted by relaxation ($T_1, \gamma = -0.58, \beta = -0.54$). Similarly, in the

Z basis (Table XI), the encoder identifies both coherence channels as prevalent features. While the first mode isolates readout error, subsequent modes explicitly target T_1 (mode 2) and T_2 (mode 3). This provides evidence that layer 1 utilizes the full calibration profile (T_1 and T_2) to inform decision-making, irrespective of code type, rather than relying on a single dominant error channel.

(2) *Partitioning of measurement noise (Layer 1)*: The Z-basis analysis reveals a seemingly particular separation of error sources in the first layer. Mode 1 is driven primarily by readout error (ϵ_r , weight $\gamma = +0.95, \beta = -0.95$), while mode 2 captures relaxation (T_1 , weight $\gamma = +0.96, \beta = +0.98$). This is consistent as well in the X-basis analysis, whereby the second mode is dedicated almost exclusively to readout error ($\gamma = +0.95, \beta = +0.96$).

(3) *Transition from specific feature modulation to learned kernel modulation (L2/L3)*: In deeper layers, the association with specific error channels seemingly diminishes. By layer 3 of the X-basis model, the dominant mode ($\gamma = 0.53, \beta = 4.01$) exhibits mixed sensitivity to noise model features, coupling T_2 ($\gamma = +0.76, \beta = +0.74$), T_1 ($\gamma = -0.42, \beta = -0.40$), and readout errors ($\gamma = +0.47, \beta = +0.53$) simultaneously. A similar pattern emerges in the Z-basis experiments, whereby the top mode ($\gamma = 0.65, \beta = 3.45$) couples to a mixture of T_1 ($\gamma = -0.41, \beta = -0.04$), T_2 ($\gamma = +0.41, \beta = -0.25$), and readout errors ($\gamma = +0.81, \beta = -0.97$). This evolution motivates our hypothesis that the network hierarchy is learning the decoding task process. Layer 1 performs calibration-aware feature extraction (isolating physical error sources), while deeper layers perform the logical inference, reweighting these extracted feature maps to resolve particular detector patterns where the distinction between a measurement error and a data-qubit error is context dependent.

The Jacobian analyses suggest that FiLM acts as a soft attention mechanism, whereby the weights are continuous features generated via a learned hardware embedding. Rather than learning a static decoding rule, the network learns a library of filters and a dynamic policy for deploying them. The hardware encoder provides decoding context based on prior extracted hardware features from which the model learns underlying correlations. This mechanism helps explain the model's ability to generalize to unseen chains without re-training, as the convolutional weights provide the decoding capability, while the calibration embedding provides the decoding context.

V. DISCUSSION

The spatiotemporal variability of quantum hardware typically forces a trade-off between decoder accuracy and the operational cost in either pretraining or induction. In this work, we put forward a calibration-conditioned neural decoder as a strategy to mitigate this challenge. Our experimental results suggest that conditioning on a learned embedding of calibration data allows the decoder to maintain performance across hardware changes, with promising results in the domain of adaptive decoding. However, we acknowledge the results are limited to the repetition code, and hence as hardware evolves further tests on larger codes are needed.

A. Generalization via calibration conditioning

An observation of our work is the decoder's continued performant capabilities on unseen qubit chains and recalibrated devices. The ability to function on hardware configurations suggests that the model is not strictly memorizing the spatial locations of erroneous qubits or specific recurring error channels from the training set. Instead, the results are consistent with the network learning a generalized conditional probability distribution within the calibration embedding space. The model appears to learn how specific calibrations, such as a drop in readout fidelity or an increase in T_1 relaxation, correlate with error likelihoods, regardless of their specific location.

While the network is only explicitly provided standard calibration metrics that are used to derive error rates for MWPM (T_1 , T_2 , gate errors), the performance gains suggest that the GCN encoder may also be inferring latent noise properties of the system based on the learned correlation between calibration data, topology information, and experiment data. This capability is particularly important for fault-tolerant computation where logical qubits are not static entities. Protocols such as lattice surgery or code deformations inherently require decoders to dynamically shift their physical support during computation. As such, maintaining a bespoke trained decoder for every physical qubit permutation is unreasonable. A single, adaptable decoder that generalizes sufficiently across hardware provides the necessary flexibility for these dynamic operations, effectively decoupling the decoding policy from specific physical qubits. The cross-conditioning ablations of Sec. IV D support this view. The learned representation tolerates perturbations at the scale of natural calibration drift and week-old snapshots with only mild degradation, while remaining sensitive to the calibration content itself, with penalties that grow with circuit volume. This generalization operates at the level of hardware configurations for a given (d, r) ; extending a single conditioned model across code distances and syndrome depths is a natural next step for this framework.

B. Scope of the repetition-code test bed and the path to higher-dimensional codes

In scaling toward universal fault tolerance, controlled studies on smaller codes serve as valuable intermediate points. Successfully deploying a surface code will likely require first resolving the decoding challenges the repetition code already presents, in the same way that high-weight, long-range qLDPC codes build on the challenges of topologically local codes. The repetition code additionally admits native, routing compilation-free embedding at scale on heavy-hex devices, enabling the multidevice, multichain, multisnapshot corpus on which our conditioning results rest, although we note that it protects only a single Pauli basis and is inherently limited relative to higher-dimensional codes.

A natural next step is extending calibration-conditioned decoding to a suitable quantum code such as the surface code [12] or the iceberg code [33]. The calibration-graph encoder, the FiLM generator, the separation between

asynchronous calibration processing and real-time syndrome decoding, and the weight-folding strategy all carry over directly, as these codes are likewise embedded in hardware graphs with spatially varying calibration data. The detector structure is similarly mirrored, with stabilizer measurements compared across consecutive rounds to form detection events, and as such the architecture requires no structural changes, only a restructuring of the detector input shape and output targets to the geometry of the code.

C. Calibration-conditioned decoding in the broader hardware-aware stack

More broadly, this work is part of an emerging direction in which device characterization data are no longer used solely to construct simplified analytic error models, but are instead consumed directly—and adaptively—by downstream algorithms. Recent work has integrated gate-set tomography data into generative quantum circuit synthesis [34], sharing the philosophy that hardware characterization should be embedded into the computational workflow itself to confer robustness against complex, correlated, and drifting noise. Within decoding specifically, recent circuit-level frameworks emphasize decoders that are fast, accurate, and adaptable to the structure of the logical circuit being executed, where logical gates introduce nontrivial error correlations [35,36]. Calibration-conditioned modulation of the kind proposed here is complementary to such circuit-aware decoders: The FiLM mechanism conditions on *hardware* state, while circuit-aware architectures condition on *computation* structure, and the two conditioning signals could in principle be combined within a single modulated backbone. We view characterization-conditioned decoding as one instance of this larger pattern, alongside characterization-driven compilation, control, and circuit synthesis.

D. Latency and implementation

From an operational standpoint, the proposed architecture demonstrates that a system noise model could potentially be integrated without incurring a latency penalty. Our empirical benchmarks demonstrate that by folding the precomputed FiLM parameters (γ, β) directly into the convolutional weights, the real-time inference complexity becomes approximately identical to the unconditioned baseline. This effectively decouples the computational cost of the hardware encoder from the critical low-latency path of the decoder.

This neural architecture exploits the natural separation of timescales in superconducting quantum processors. Calibration data shifts occur over hours or days, whereas syndrome extraction occurs at the microsecond level. By performing the heavy lifting of the graph neural network asynchronously, the real-time decoding path remains lightweight.

VI. CONCLUSION AND FUTURE WORK

In this work, we introduced a hardware-conditioned neural decoder that leverages FiLM to adapt dynamically to the spatiotemporal variability of superconducting quantum

processors. By evaluating over 2.7 million experimental shots on IBM devices, we observed that this approach yields lower logical error rates compared to matching-based and unconditioned neural baselines, particularly in large-distance regimes ($d \geq 7$) where error correlations are most pronounced. Furthermore, we demonstrate that a single model per (d, r) configuration is capable of generalizing to unseen contiguous qubit chains and future calibration snapshots without retraining, suggesting the decoder can learn a transferable representation of hardware noise. Cross-conditioning ablations corroborate this mechanism directly: Corrupting or mismatching the calibration input degrades decoding progressively with corruption severity in the large-circuit regime.

A natural extension and future work is to extend this conditioning framework to higher-weight stabilizer codes, such as the surface code, where the calibration graph naturally generalizes to support dual X - and Z -basis decoding. We anticipate that the latency benefits of the architecture will persist, as the heavy lifting of the graph neural network is performed via precomputable FiLM parameters. A primary challenge lies in designing topological backbones capable of navigating more complex code stabilizer graphs while retaining the efficiency of the conditioned convolutional approach.

ACKNOWLEDGMENTS

This material is based upon work supported by the U.S. Department of Energy, Office of Science, National Quantum Information Science Research Centers, Quantum Science Center (QSC). This research was supported by PNNL’s Quantum Algorithms and Architecture for Domain Science (QuAADS) Laboratory Directed Research and Development (LDRD) Initiative. The Pacific Northwest National Laboratory is operated by Battelle for the U.S. Department of Energy under Contract No. DE-AC05-76RL01830. This research used resources of the Oak Ridge Leadership Computing Facility (OLCF), which is a DOE Office of Science User Facility supported under Contract No. DE-AC05-00OR22725. This research used resources of the National Energy Research Scientific Computing Center (NERSC), a U.S. Department of Energy Office of Science User Facility located at Lawrence Berkeley National Laboratory, operated under Contract No. DE-AC02-05CH11231.

DATA AVAILABILITY

The data supporting this work are available in Ref. [37].

APPENDIX: NEURAL NETWORK PARAMETERS

The FiLM-conditioned decoder couples a three-layer 3×3 CNN with FiLM scalings derived from a three-layer GCN over hardware graphs.

TABLE II. FiLM+CNN decoder architecture and training summary.

Parameter	Value
<i>Architecture</i>	
Main network arch.	2D convolutional (three blocks)
Conv kernel config	$k = 3$, padding = 1, ReLU
Channel depth	$128 \rightarrow 256 \rightarrow 512$
Hardware encoder	three-layer GCN
Latent dimension	256
FiLM generator	MLP ($256 \rightarrow 256 \rightarrow \text{out}$)
Output head	Sigmoid
<i>Training</i>	
Batch size	4096
Epochs	100
Data split	70% train/30% val
Optimizer	Adam ($\beta_1 = 0.9$, $\beta_2 = 0.999$)
Learning rate	5×10^{-3}
Scheduler	Cosine annealing ($\eta_{\min} = 0$)
Loss function	Binary cross entropy

1. Logical error rates

Tables III–VI present the detailed breakdown of LER across all experimental configurations. Tables III and IV correspond to the unseen validation set, while Tables V and VI correspond to the unseen data on the Kingston device, sampled one week later. Values of “—” and 0.00 indicate regimes where no logical errors were observed for the baseline decoder, rendering ratio comparisons undefined. These configurations correspond to shallow circuits where all decoders succeed post majority vote.

2. Cross-conditioning ablation tables

Table VII summarizes the cross-conditioning ablation of Sec. IV D over the large-circuit regime, and Tables VIII and IX tabulate the per-configuration LER ratios underlying the cross-conditioning ablation of Sec. IV D. All evaluations use the unseen recalibrated KINGSTON datasets.

3. Interpretation of learned FiLM modes

To investigate the physical mechanism of the decoder, we perform an SVD on the Jacobian of the FiLM generator. Table XI (Z basis) and Table X (X basis) detail the dominant singular modes.

We decompose the sensitivity of the FiLM parameters ($\theta \in \{\gamma, \beta\}$) to the input hardware calibration features. The table separates the normalized contributions of relaxation (T_1), dephasing (T_2), gate error (ϵ_g), and readout assignment error (ϵ_r) under features (**f**). Dominant drivers for each mode are highlighted in bold.

TABLE III. Z-basis decoder comparison on validation set. Comparison of LER for FiLM+CNN, modified MWPM, and unconditioned CNN. The ratio $\rho = \text{LER}_{\text{FiLM}}/\text{LER}_{\text{Base}}$ indicates performance gain (values < 1.0 indicate FiLM improvement). Values of “—” and 0.00 indicate no logical errors. Bold indicates the FiLM+CNN model improved upon the relative model it is being compared to.

Setup		Logical error rate (LER)			Ratio (ρ)	
d	r	FiLM+CNN	MWPM	CNN	vs MWPM	vs CNN
3	1	1.30×10^{-3}	1.54×10^{-3}	1.07×10^{-3}	0.845	1.22
3	3	1.16×10^{-2}	1.11×10^{-2}	1.10×10^{-2}	1.05	1.06
5	1	7.52×10^{-4}	3.46×10^{-4}	7.76×10^{-4}	2.18	0.970
5	3	1.33×10^{-2}	5.34×10^{-3}	1.02×10^{-2}	2.48	1.29
5	5	1.72×10^{-2}	1.66×10^{-2}	1.58×10^{-2}	1.04	1.09
7	1	7.57×10^{-5}	9.46×10^{-5}	7.57×10^{-5}	0.800	1.00
7	3	2.59×10^{-3}	2.85×10^{-3}	2.09×10^{-3}	0.908	1.24
7	5	6.08×10^{-3}	1.39×10^{-2}	1.36×10^{-2}	0.438	0.447
7	7	1.19×10^{-2}	5.34×10^{-2}	3.18×10^{-2}	0.222	0.373
9	1	0.00	1.36×10^{-5}	0.00	0.000	—
9	3	9.03×10^{-4}	2.35×10^{-3}	9.77×10^{-4}	0.385	0.925
9	5	3.25×10^{-3}	1.14×10^{-2}	6.59×10^{-3}	0.284	0.493
9	7	3.42×10^{-3}	2.62×10^{-2}	3.48×10^{-2}	0.130	0.098
9	9	9.94×10^{-3}	4.53×10^{-2}	2.54×10^{-2}	0.220	0.392
11	1	0.00	3.21×10^{-6}	0.00	0.000	—
11	3	7.73×10^{-4}	6.33×10^{-4}	2.85×10^{-4}	1.22	2.71
11	5	2.65×10^{-3}	5.26×10^{-3}	5.09×10^{-3}	0.503	0.520
11	7	2.93×10^{-3}	2.55×10^{-2}	1.87×10^{-2}	0.115	0.157
11	9	5.17×10^{-3}	3.47×10^{-2}	4.98×10^{-2}	0.149	0.104
11	11	5.96×10^{-3}	6.63×10^{-2}	9.42×10^{-2}	0.090	0.063

TABLE IV. X-basis decoder comparison on validation set. Comparison of LER for FiLM+CNN, modified MWPM, and unconditioned CNN. The ratio $\rho = \text{LER}_{\text{FiLM}}/\text{LER}_{\text{Base}}$ indicates performance gain (values < 1.0 indicate FiLM improvement). Values of “—” and 0.00 indicate no logical errors. Bold indicates the FiLM+CNN model improved upon the relative model it is being compared to.

Setup		Logical error rate (LER)			Ratio (ρ)	
d	r	FiLM+CNN	MWPM	CNN	vs MWPM	vs CNN
3	1	2.76×10^{-3}	2.14×10^{-3}	2.33×10^{-3}	1.29	1.18
3	3	2.28×10^{-2}	2.50×10^{-2}	2.28×10^{-2}	0.912	1.00
5	1	8.79×10^{-4}	6.04×10^{-4}	6.59×10^{-4}	1.46	1.33
5	3	1.62×10^{-2}	1.80×10^{-2}	9.89×10^{-3}	0.897	1.64
5	5	1.96×10^{-2}	5.08×10^{-2}	2.38×10^{-2}	0.386	0.823
7	1	1.53×10^{-4}	2.04×10^{-5}	1.22×10^{-4}	7.50	1.25
7	3	1.07×10^{-2}	7.21×10^{-3}	3.35×10^{-3}	1.49	3.21
7	5	1.03×10^{-2}	3.28×10^{-2}	1.19×10^{-2}	0.313	0.862
7	7	3.84×10^{-2}	6.47×10^{-2}	2.87×10^{-2}	0.593	1.34
9	1	4.07×10^{-5}	2.37×10^{-5}	0.00	1.71	—
9	3	1.00×10^{-2}	5.49×10^{-3}	9.03×10^{-3}	1.82	1.11
9	5	1.04×10^{-2}	2.82×10^{-2}	1.40×10^{-2}	0.367	0.741
9	7	2.37×10^{-2}	5.07×10^{-2}	2.08×10^{-2}	0.467	1.14
9	9	1.83×10^{-2}	9.30×10^{-2}	4.96×10^{-2}	0.197	0.368
11	1	0.00	3.73×10^{-5}	0.00	0.000	—
11	3	3.17×10^{-3}	3.49×10^{-3}	1.22×10^{-3}	0.910	2.60
11	5	3.54×10^{-3}	1.76×10^{-2}	6.78×10^{-3}	0.201	0.523
11	7	3.60×10^{-3}	3.16×10^{-2}	1.67×10^{-2}	0.114	0.215
11	9	4.32×10^{-2}	7.33×10^{-2}	4.70×10^{-2}	0.590	0.919
11	11	1.40×10^{-2}	8.26×10^{-2}	3.63×10^{-2}	0.169	0.385

TABLE V. Z-basis decoder performance on Kingston unseen experimental data. Experiments are run one week later, on unseen contiguous qubit chains. Values of “—” and 0.00 indicate no logical errors. Bold indicates the FiLM+CNN model improved upon the relative model it is being compared to.

Setup		Logical error rate (LER)			Ratio (ρ)	
d	r	FiLM+CNN	MWPM	CNN	vs MWPM	vs CNN
3	1	1.55×10^{-3}	1.02×10^{-3}	1.42×10^{-3}	1.52	1.09
3	3	1.71×10^{-2}	1.66×10^{-2}	1.12×10^{-2}	1.03	1.53
5	1	1.26×10^{-3}	9.77×10^{-5}	1.30×10^{-3}	12.9	0.969
5	3	2.04×10^{-2}	6.35×10^{-3}	6.80×10^{-3}	3.21	3.00
5	5	1.30×10^{-2}	1.79×10^{-2}	9.97×10^{-3}	0.728	1.31
7	1	0.00	0.00	4.07×10^{-5}	—	0.000
7	3	2.16×10^{-3}	2.85×10^{-3}	2.04×10^{-3}	0.757	1.06
7	5	5.45×10^{-3}	1.03×10^{-2}	5.78×10^{-3}	0.529	0.944
7	7	1.49×10^{-2}	3.33×10^{-2}	3.13×10^{-2}	0.446	0.474
9	1	0.00	0.00	0.00	—	—
9	3	5.37×10^{-4}	6.10×10^{-5}	4.40×10^{-4}	8.80	1.22
9	5	1.25×10^{-2}	1.34×10^{-2}	6.93×10^{-3}	0.931	1.80
9	7	5.80×10^{-3}	2.45×10^{-2}	1.95×10^{-2}	0.236	0.298
9	9	7.52×10^{-3}	5.20×10^{-2}	4.47×10^{-2}	0.145	0.168
11	1	0.00	0.00	0.00	—	—
11	3	1.10×10^{-3}	3.66×10^{-4}	1.71×10^{-3}	3.00	0.643
11	5	5.98×10^{-3}	1.29×10^{-2}	7.87×10^{-3}	0.462	0.760
11	7	4.79×10^{-3}	1.98×10^{-2}	1.51×10^{-2}	0.241	0.316
11	9	5.31×10^{-3}	4.81×10^{-2}	4.01×10^{-2}	0.110	0.132
11	11	8.79×10^{-3}	6.52×10^{-2}	7.33×10^{-2}	0.135	0.120

TABLE VI. X-basis decoder performance on Kingston unseen experimental data. Experiments are run one week later, on unseen contiguous qubit chains. Values of “—” and 0.00 indicate no logical errors. Bold indicates the FiLM+CNN model improved upon the relative model it is being compared to.

Setup		Logical error rate (LER)			Ratio (ρ)	
d	r	FiLM+CNN	MWPM	CNN	vs MWPM	vs CNN
3	1	3.50×10^{-3}	2.56×10^{-3}	2.48×10^{-3}	1.37	1.41
3	3	2.18×10^{-2}	2.42×10^{-2}	1.91×10^{-2}	0.902	1.14
5	1	8.14×10^{-4}	4.88×10^{-4}	1.30×10^{-3}	1.67	0.625
5	3	3.73×10^{-2}	1.69×10^{-2}	3.33×10^{-2}	2.22	1.12
5	5	6.19×10^{-2}	6.14×10^{-2}	4.30×10^{-2}	1.01	1.44
7	1	2.44×10^{-4}	1.22×10^{-4}	2.44×10^{-4}	2.00	1.00
7	3	4.17×10^{-2}	9.95×10^{-3}	2.82×10^{-2}	4.19	1.48
7	5	2.99×10^{-2}	3.80×10^{-2}	2.02×10^{-2}	0.787	1.48
7	7	4.53×10^{-2}	8.04×10^{-2}	6.77×10^{-2}	0.563	0.669
9	1	1.22×10^{-4}	0.00	8.14×10^{-5}	—	1.50
9	3	1.42×10^{-2}	3.32×10^{-3}	6.63×10^{-3}	4.28	2.14
9	5	5.86×10^{-2}	5.34×10^{-2}	4.69×10^{-2}	1.10	1.25
9	7	5.63×10^{-2}	7.85×10^{-2}	8.35×10^{-2}	0.717	0.674
9	9	6.77×10^{-2}	1.22×10^{-1}	1.01×10^{-1}	0.554	0.672
11	1	0.00	0.00	1.22×10^{-4}	—	0.000
11	3	1.57×10^{-2}	3.34×10^{-3}	1.09×10^{-2}	4.71	1.44
11	5	2.47×10^{-2}	6.40×10^{-2}	2.43×10^{-2}	0.385	1.02
11	7	6.31×10^{-2}	1.09×10^{-1}	9.25×10^{-2}	0.579	0.682
11	9	4.91×10^{-2}	1.16×10^{-1}	8.62×10^{-2}	0.422	0.569
11	11	5.40×10^{-2}	1.18×10^{-1}	1.13×10^{-1}	0.456	0.478

TABLE VII. Cross-conditioning ablation summary over the large-circuit regime ($d \geq 7$, $r \geq 5$; nine configurations per basis). For each conditioning input, we report the median and range of the LER ratio vs matched conditioning.

Conditioning input	Z basis		X basis	
	Median	Range	Median	Range
Gaussian noise, $\sigma = 0.25$	$1.04\times$	[0.93, 1.24]	$1.06\times$	[0.97, 1.13]
Stale (training period)	$1.22\times$	[0.90, 1.53]	$1.06\times$	[0.89, 1.17]
PITTSBURGH calibration	$1.06\times$	[0.88, 3.31]	$0.99\times$	[0.82, 1.41]
FEZ calibration	$1.28\times$	[0.92, 2.80]	$1.08\times$	[1.02, 2.50]
Gaussian noise, $\sigma = 1.0$	$1.68\times$	[1.21, 2.80]	$1.44\times$	[1.10, 2.52]
Fixed FiLM parameters ($\gamma = 1$, $\beta = 0$)	$2.44\times$	[0.55, 8.34]	$1.48\times$	[0.68, 2.49]

TABLE VIII. Z-basis cross-conditioning ablation on unseen recalibrated KINGSTON data. Entries give the LER ratio relative to matched (current Kingston) calibration conditioning for each conditioning input of Fig. 8, where values above unity indicate worsened decoding and values below unity indicate improved decoding. Ratios are undefined (“–”) for configurations with no logical errors under matched conditioning; matched LERs are tabulated in Tables V and VI.

		LER ratio vs matched conditioning					
d	r	Stale	FEZ	PITTSB.	+Noise $\sigma = 0.25$	+Noise $\sigma = 1.0$	Fixed FiLM
3	1	1.24	0.84	0.95	1.00	1.32	0.63
3	3	0.84	0.85	0.84	1.05	1.03	0.52
5	1	1.06	1.10	1.06	1.06	1.00	0.87
5	3	0.57	0.42	0.38	0.97	0.87	0.24
5	5	0.94	1.01	1.08	1.01	1.27	0.98
7	1	–	–	–	–	–	–
7	3	1.64	1.42	1.38	0.94	1.89	1.47
7	5	1.53	1.72	1.58	1.01	1.78	1.06
7	7	1.22	1.07	1.05	0.95	1.22	2.11
9	1	–	–	–	–	–	–
9	3	1.00	1.00	1.00	1.00	1.00	0.82
9	5	1.29	1.35	1.04	1.14	1.30	0.55
9	7	1.00	1.04	1.11	0.93	1.21	2.44
9	9	1.27	0.92	0.88	1.24	2.52	5.94
11	1	–	–	–	–	–	–
11	3	1.00	1.00	1.00	1.00	1.00	0.33
11	5	1.11	2.80	1.31	1.09	1.68	1.30
11	7	0.99	1.17	1.05	0.93	2.80	3.16
11	9	0.90	1.28	1.06	1.05	1.39	7.55
11	11	1.31	2.19	3.31	1.04	1.95	8.34

TABLE IX. X -basis cross-conditioning ablation, in the format of Table VIII.

d	r	LER ratio vs matched conditioning					
		Stale	FEZ	PITTSB.	+Noise $\sigma = 0.25$	+Noise $\sigma = 1.0$	Fixed FiLM
3	1	0.78	0.94	0.88	0.90	0.86	0.71
3	3	0.96	1.06	0.95	1.03	1.06	1.08
5	1	1.10	1.05	1.50	1.20	1.30	0.40
5	3	1.14	1.25	1.12	0.97	1.15	0.77
5	5	1.13	0.99	0.85	0.97	1.06	0.89
7	1	1.00	1.00	1.50	1.00	1.50	0.50
7	3	1.12	1.34	1.08	1.04	1.19	0.68
7	5	1.17	1.08	1.09	0.97	1.10	0.68
7	7	1.12	1.12	0.94	1.06	1.44	1.50
9	1	0.67	2.67	1.67	0.67	1.67	0.33
9	3	0.93	1.41	0.64	0.94	1.65	0.44
9	5	1.10	1.02	0.89	1.00	1.27	0.80
9	7	1.03	1.38	1.18	1.06	1.55	1.48
9	9	0.89	1.02	0.87	1.06	1.59	2.25
11	1	—	—	—	—	—	—
11	3	1.24	1.17	1.13	1.06	1.15	0.70
11	5	0.93	2.50	0.82	1.13	2.52	0.98
11	7	1.11	1.06	1.41	0.97	1.21	1.47
11	9	1.06	1.03	1.16	1.09	1.88	2.49
11	11	1.01	1.21	0.99	1.05	1.32	2.09

TABLE X. X -basis dominant FiLM modes ($d = 11$). The feature columns show the sensitivity of the mode to specific hardware features. Values represent the normalized singular vector component (signed), with dominant drivers highlighted in bold.

Param	Layer	Mode	σ	Hardware feature sensitivity			
				T_1	T_2	ϵ_g	ϵ_r
γ	L1	M1	1.81	−0.58	+0.81	+0.09	+0.08
		M2	1.38	+0.09	−0.00	−0.29	+0.95
		M3	0.56	+0.80	+0.55	+0.23	−0.01
		M4	0.36	−0.12	−0.22	+0.93	+0.29
γ	L2	M1	0.42	−0.49	+0.86	−0.04	+0.10
		M2	0.25	+0.30	+0.06	−0.16	+0.94
		M3	0.15	+0.80	+0.50	+0.25	−0.24
		M4	0.07	−0.18	−0.08	+0.96	+0.22
γ	L3	M1	0.53	−0.42	+0.76	−0.14	+0.47
		M2	0.37	+0.42	−0.32	−0.14	+0.84
		M3	0.13	+0.78	+0.53	−0.24	−0.23
		M4	0.10	+0.20	+0.20	+0.95	+0.14
β	L1	M1	3.18	−0.54	+0.83	−0.07	+0.13
		M2	2.42	+0.14	−0.08	−0.22	+0.96
		M3	1.19	+0.83	+0.56	+0.03	−0.07
		M4	0.69	−0.03	+0.03	+0.97	+0.23
β	L2	M1	2.95	−0.46	+0.77	−0.09	+0.44
		M2	2.27	+0.27	−0.36	−0.19	+0.87
		M3	1.10	+0.84	+0.53	+0.07	−0.02
		M4	0.60	−0.05	−0.04	+0.98	+0.21
β	L3	M1	4.01	−0.40	+0.74	−0.11	+0.53
		M2	3.03	+0.28	−0.47	−0.20	+0.82
		M3	1.50	+0.88	+0.48	+0.01	−0.02
		M4	0.85	+0.01	−0.01	+0.97	+0.23

TABLE XI. Z-basis dominant FiLM modes ($d = 11$). The feature columns show the sensitivity of the mode to specific hardware defects. Values represent the normalized singular vector component (signed), with dominant drivers highlighted in bold.

Param	Layer	Mode	σ	Hardware feature sensitivity			
				T_1	T_2	ϵ_g	ϵ_r
γ	L1	M1	0.43	-0.07	+0.31	-0.03	+0.95
		M2	0.19	+0.96	-0.22	-0.09	+0.14
		M3	0.12	-0.20	-0.86	+0.37	+0.28
		M4	0.05	+0.18	+0.34	+0.92	-0.07
γ	L2	M1	0.52	-0.21	-0.50	+0.03	-0.84
		M2	0.26	-0.95	+0.08	+0.25	+0.20
		M3	0.12	-0.21	-0.39	-0.86	+0.25
		M4	0.06	-0.14	+0.77	-0.44	-0.44
γ	L3	M1	0.65	-0.41	+0.41	-0.06	+0.81
		M2	0.22	+0.87	-0.10	-0.08	+0.48
		M3	0.06	+0.28	+0.88	+0.28	-0.28
		M4	0.03	-0.04	-0.24	+0.96	+0.17
β	L1	M1	2.59	-0.13	-0.25	+0.09	-0.95
		M2	1.50	+0.98	-0.20	-0.05	-0.09
		M3	0.57	-0.15	-0.92	+0.21	+0.28
		M4	0.29	+0.10	+0.22	+0.97	+0.02
β	L2	M1	2.87	-0.04	+0.28	-0.07	+0.96
		M2	1.68	+0.97	-0.19	-0.10	+0.09
		M3	0.70	+0.20	+0.93	-0.11	-0.28
		M4	0.33	+0.12	+0.11	+0.99	+0.05
β	L3	M1	3.45	-0.04	-0.25	+0.08	-0.97
		M2	2.09	+0.98	-0.18	-0.09	+0.00
		M3	0.84	-0.16	-0.93	+0.20	+0.26
		M4	0.39	+0.12	+0.20	+0.97	+0.03

- [1] D. A. Lidar and T. A. Brun, *Quantum Error Correction* (Cambridge University Press, Cambridge, England, 2013).
- [2] M. E. Beverland, P. Murali, M. Troyer, K. M. Svore, T. Hoeffler, V. Kliuchnikov, G. H. Low, M. Soeken, A. Sundaram, and A. Vashchillo, Assessing requirements to scale to practical quantum advantage, [arXiv:2211.07629](#).
- [3] B. M. Terhal, Quantum error correction for quantum memories, *Rev. Mod. Phys.* **87**, 307 (2015).
- [4] D. Gottesman, *Stabilizer Codes and Quantum Error Correction* (California Institute of Technology, Pasadena, California, 1997).
- [5] A. G. Fowler, A. C. Whiteside, and L. C. Hollenberg, Towards practical classical processing for the surface code: Timing analysis, *Phys. Rev. A* **86**, 042313 (2012).
- [6] J. Chiaverini, D. Leibfried, T. Schaetz, M. D. Barrett, R. B. Blakestad, J. Britton, W. M. Itano, J. D. Jost, E. Knill, C. Langer, R. Ozeri, and D. J. Wineland, Realization of quantum error correction, *Nature (London)* **432**, 602 (2004).
- [7] L. DiCarlo, M. D. Reed, L. Sun, B. R. Johnson, J. M. Chow, J. M. Gambetta, L. Frunzio, S. M. Girvin, M. H. Devoret, and R. J. Schoelkopf, Preparation and measurement of three-qubit entanglement in a superconducting circuit, *Nature (London)* **467**, 574 (2010).
- [8] P. Schindler, J. T. Barreiro, T. Monz, V. Nebendahl, D. Nigg, M. Chwalla, M. Hennrich, and R. Blatt, Experimental repetitive quantum error correction, *Science* **332**, 1059 (2011).
- [9] J. R. Wootton and D. Loss, Repetition code of 15 qubits, *Phys. Rev. A* **97**, 052313 (2018).
- [10] I. Besedin, M. Kerschbaum, J. Knoll, I. Hesner, L. Bödeker, L. Colmenarez, L. Hofele, N. Lacroix, C. Hellings, F. Swiadek, *et al.*, Realizing lattice surgery on two distance-three repetition codes with superconducting qubits, *Nat. Phys.* **22**, 189 (2026).
- [11] A. R. Calderbank and P. W. Shor, Good quantum error-correcting codes exist, *Phys. Rev. A* **54**, 1098 (1996).
- [12] A. G. Fowler, Proof of finite surface code threshold for matching, *Phys. Rev. Lett.* **109**, 180502 (2012).
- [13] J. Bausch, A. W. Senior, F. J. Heras, T. Edlich, A. Davies, M. Newman, C. Jones, K. Satzinger, M. Y. Niu, S. Blackwell, *et al.*, Learning high-accuracy error decoding for quantum processors, *Nature (London)* **635**, 834 (2024).
- [14] R. Acharya, *et al.* (Google Quantum AI and Collaborators), Quantum error correction below the surface code threshold, *Nature (London)* **638**, 920 (2025).
- [15] H. Wang, P. Liu, K. Shao, D. Li, J. Gu, D. Z. Pan, Y. Ding, and S. Han, Transformer-QEC: Quantum error correction code decoding with transferable transformers, [arXiv:2311.16082](#).
- [16] M. Lange, P. Havstrom, B. Srivastava, I. Bengtsson, V. Bergentall, K. Hammar, O. Heuts, E. van Nieuwenburg, and M. Granath, Data-driven decoding of quantum error correcting

- codes using graph neural networks, *Phys. Rev. Res.* **7**, 023181 (2025).
- [17] M. Carroll, S. Rosenblatt, P. Jurcevic, I. Lauer, and A. Kandala, Dynamics of superconducting qubit relaxation times, *npj Quantum Inf.* **8**, 132 (2022).
- [18] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville, Film: Visual reasoning with a general conditioning layer, in *Proceedings of the AAAI Conference on Artificial Intelligence* (AAAI, New Orleans, Louisiana, USA, 2018), Vol. 32.
- [19] M. A. Nielsen and I. L. Chuang, *Quantum Computation and Quantum Information* (Cambridge University Press, Cambridge, England, 2010).
- [20] G. S. Ravi, J. M. Baker, A. Fayyazi, S. F. Lin, A. Javadi-Abhari, M. Pedram, and F. T. Chong, Better than worst-case decoding for quantum error correction, in *Proceedings of the 28th ACM International Conference on Architectural Support for Programming Languages and Operating Systems* (ACM Press, New York, NY, 2023), Vol. 2, pp. 88–102.
- [21] S. J. Devitt, W. J. Munro, and K. Nemoto, Quantum error correction for beginners, *Rep. Prog. Phys.* **76**, 076001 (2013).
- [22] M.-H. Hsieh and F. Le Gall, NP-hardness of decoding quantum error-correction codes, *Phys. Rev. A* **83**, 052331 (2011).
- [23] S. Krastanov and L. Jiang, Deep neural network probabilistic decoder for stabilizer codes, *Sci. Rep.* **7**, 11003 (2017).
- [24] N. Maskara, A. Kubica, and T. Jochym-O'Connor, Advantages of versatile neural-network decoding for topological codes, *Phys. Rev. A* **99**, 052351 (2019).
- [25] S. Varsamopoulos, K. Bertels, and C. G. Almudever, Comparing neural network based decoders for the surface code, *IEEE Trans. Comput.* **69**, 300 (2019).
- [26] P. Baireuther, T. E. O'Brien, B. Tarasinski, and C. W. Beenakker, Machine-learning-assisted correction of correlated qubit errors in a topological code, *Quantum* **2**, 48 (2018).
- [27] A. Strikis, D. Qin, Y. Chen, S. C. Benjamin, and Y. Li, Learning-based quantum error mitigation, *PRX Quantum* **2**, 040330 (2021).
- [28] H. Liao, D. S. Wang, I. Sitdikov, C. Salcedo, A. Seif, and Z. K. Mineev, Machine learning for practical quantum error mitigation, *Nat. Mach. Intell.* **6**, 1478 (2024).
- [29] M. Liao, Y. Zhu, G. Chiribella, and Y. Yang, Noise-agnostic quantum error mitigation with data augmented neural models, *npj Quantum Inf.* **11**, 8 (2025).
- [30] O. Higgott, PyMatching: A Python package for decoding quantum codes with minimum-weight perfect matching, *ACM Trans. Quantum Comput.* **3**, 1 (2022).
- [31] Y. Hu, Y. Liu, and Z. Liu, A survey on convolutional neural network accelerators: GPU, FPGA and ASIC, in *Proceedings of the 2022 14th International Conference on Computer Research and Development (ICCRD)* (IEEE, Piscataway, NJ, 2022), pp. 100–107.
- [32] S. S. Lee, T. D. Nguyen, P. K. Meher, and S. Y. Park, Energy-efficient high-speed ASIC implementation of convolutional neural network using novel reduced critical-path design, *IEEE Access* **10**, 34032 (2022).
- [33] C. N. Self, M. Benedetti, and D. Amaro, Protecting expressive circuits with a quantum error detection code, *Nat. Phys.* **20**, 219 (2024).
- [34] K. Y. Yu, A. Sarkar, E. Hua, M. Rimbach-Russ, R. Ishihara, and S. Feld, From characterization to construction: Generative quantum circuit synthesis from gate set tomography data, *arXiv:2605.01367*.
- [35] J. P. Bonilla Ataides, A. Gu, S. F. Yelin, and M. D. Lukin, Neural decoders for universal quantum algorithms, *arXiv:2509.11370*.
- [36] Y. Zhou, C. Wan, Y. Xu, J. P. Zhou, K. Q. Weinberger, and E.-A. Kim, Learning to decode logical circuits, *Nat. Comput. Sci.* **5**, 1158 (2025).
- [37] S. Stein, Calibration-conditioned FiLM decoders for low-latency decoding of quantum error correction evaluated on IBM repetition-code experiments - Datasets (Version 0.1) [Dataset], Zenodo, 2026, <https://doi.org/10.5281/zenodo.20768087>.