High-Fidelity, Multiqubit Generalized Measurements with Dynamic Circuits

Petr Ivashkov[1,*], Gideon Uchehara[2], Liang Jiang[3], Derek S. Wang[4], and Alireza Seif[4]

[1]*Department of Information Technology and Electrical Engineering, ETH Zürich, Zürich, Switzerland*
[2]*Department of Electrical and Computer Engineering, University of British Columbia, Vancouver, Canada*
[3]*Pritzker School of Molecular Engineering, University of Chicago, Chicago, Illinois, USA*
[4]*IBM Quantum, IBM T.J. Watson Research Center, Yorktown Heights, New York, USA*

Generalized measurements, also called positive operator-valued measures (POVMs), can offer advantages over projective measurements in various quantum information tasks. Here, we realize a generalized measurement of one and two superconducting qubits with high fidelity and in a single experimental setting. To do so, we propose a hybrid method, the "Naimark-terminated binary tree," based on a hybridization of Naimark's dilation and binary tree techniques that leverages emerging hardware capabilities for midcircuit measurements and feed-forward control. Furthermore, we showcase a highly effective use of approximate compiling to enhance POVM fidelity in noisy conditions. We argue that our hybrid method scales better toward larger system sizes than its constituent methods and demonstrate its advantage by performing detector tomography of symmetric, informationally complete POVM (SIC POVM). Detector fidelity is further improved through a composite error-mitigation strategy that incorporates twirling and a newly devised conditional readout error mitigation. Looking forward, we expect improvements in approximate compilation and hardware noise for dynamic circuits to enable generalized measurements of larger multiqubit POVMs on superconducting qubits.

I. INTRODUCTION

Measuring information accurately and efficiently from inherently probabilistic systems is a central challenge of quantum physics. Projective, or von Neumann, measurements are often used in experiments because of their comparatively simple realization on many quantum computing platforms [1]. At the same time, they can be suboptimal in such tasks as quantum state discrimination [2,3], where no projective measurement can unambiguously tell two nonorthogonal states apart with a single shot. Generalized measurements or positive operator-valued measures (POVMs) define the most general framework for quantum measurements, including projective measurement as a special case. Among various broad areas [4–6], POVMs allow for unambiguous state discrimination [2,7], optimal state tomography [8,9,11], entanglement detection [12], Bell's inequalities [13], quantum machine-learning algorithms [14], and improved observable estimates in variational

quantum algorithms [15]. Therefore, it is crucial to have a deterministic protocol, which can be realized in a single experimental setting to implement a general POVM [16].

Despite the high utility of POVMs, realizing them on superconducting quantum systems is challenging. In principle, any POVM can be realized by a projective measurement in an extended Hilbert space through Naimark's dilation [17,18]. Recent efforts involved embedding the system in the qudit space of superconducting qubits [19] and trapped ions [20]. However, this requires efficient discrimination of qudit states, which adds a level of experimental complexity and suffers from readout errors. Alternatively, a POVM may be realized by coupling the system to a number of auxiliary qubits that scales with the size of the POVM [21,22]. Such implementations, however, raise concerns about circuit complexity and, hence, scalability to multiqubit systems. For instance, the large number of auxiliary qubits required for Naimark's dilation may not be readily available or directly connected to the qubits one intends to measure. Overall, Naimark's dilation encounters practical implementation issues due to the complex unitary operations required in the extended Hilbert space.

A promising alternative is a binary search, which employs only a single auxiliary qubit to realize general multiqubit POVMs [23,24]. It involves a sequence of conditional two-outcome POVMs, requiring cutting-edge

---

*Contact author: ivashkov.peter@gmail.com

hardware capabilities such as midcircuit measurements and feed-forward control [25,26] comprising "dynamic" or "adaptive" circuits. This scheme has been recently demonstrated in a specialized experiment with a single microwave cavity coupled to a transmon qubit [27]. However, extending the implementation of this scheme to multiqubit programmable quantum processors requires circuits with potentially large depths and many feed-forward operations that limit its fidelity.

To address the limitation imposed on POVM fidelity by circuit noise, we propose a novel method for single-setting POVMs on multiqubit systems using dynamic circuits. Furthermore, we introduce innovative use of approximate compiling to implement measurements, enhancing POVM fidelity under noisy conditions [28]. In Sec. III, we present a hybridization of Naimark's dilation with the binary search—an approach that we call "Naimark-terminated binary tree." Our hybrid method results in shorter-depth circuits and scales better toward larger systems. In Sec. IV, we implement all three methods on an IBM quantum device and demonstrate the advantage of our hybrid approach by performing detector tomography of symmetric, informationally complete POVM (SIC POVM) [8,9]. Using our hybrid approach with a composite error-mitigation strategy, including twirling and newly devised conditional readout error mitigation (CREM), we improve the fidelity of two qubit SIC POVM to $70.4 \pm 0.1\%$ from 52% and 40% of bare Naimark and binary-tree approaches, respectively. Our code and data are available in the Supplemental Material [10].

II. POVM

Formally, a POVM is a set $F = \{F_i\}$ of M positive semidefinite Hermitian operators, called POVM elements. Each element F_i corresponds to a measurement outcome i with probability $P(i) = \mathrm{Tr}(F_i \rho)$, where ρ is the state of the system. POVM elements must satisfy the completeness relation $\sum_{i=1}^{M} F_i = I$ to have a normalized probability distribution. Unlike projective measurements, POVM elements are not necessarily orthogonal. We restrict our attention to POVMs whose elements are linearly independent rank-one operators $F_i = |\psi_i\rangle\langle\psi_i|$ where $|\psi_i\rangle$ is not necessarily normalized. Any higher-rank POVMs can be obtained by relabeling and mixing the outcomes of rank-one POVMs with a maximum of d^2 elements, where d is the dimension of the system [29]. Note that a POVM defines only the measurement statistics $P(i)$ but not the postmeasurement state of the system, which depends on how the POVM is realized. In fact, we disregard the postmeasurement state in applications like observable estimation or quantum state tomography (see Appendix E), where the system is measured only once at the end of the experiment. Therefore, this paper focuses only on measurement statistics. This allows us to measure both the

system and the auxiliary qubits directly using Naimark's dilation and achieve a higher fidelity than binary search at the cost of destroying the postmeasurement state.

III. NAIMARK-TERMINATED BINARY TREE

In this section, we propose a novel combination of two previously known methods for general POVMs, binary search and Naimark's dilation. This new hybrid scheme, the Naimark-terminated binary tree, is more efficient than its constituent methods. In a nutshell, we perform the binary search by repeatedly dividing the set of POVM elements in half to narrow down the search range. When the number of remaining POVM elements corresponds to the dimension of the compound system and auxiliary Hilbert space, we interrupt the binary search and apply Naimark's dilation.

We begin with the binary search. As originally detailed by Andersson and Oi [23], binary search can realize general POVM with only a single auxiliary qubit. It, therefore, reduces the complexity of manipulating an extended system and saves the quantum memory space when M is large. We adopt the notation from Shen *et al.* [24] to briefly outline the key steps below; see Appendix H for the full treatment.

To construct the binary search tree, we begin by padding our set of POVM elements with zero operators until M is the nearest power of two. In the first step, we split the original POVM into two sets of $M/2$ elements, for example, as

$$B_0 := \sum_{i=1}^{M/2} F_i \quad \text{and} \quad B_1 := \sum_{i=M/2+1}^{M} F_i. \tag{1}$$

The ordering of nonzero F_i may be arbitrary and corresponds to relabeling the measurement outcomes. The set $\{B_0, B_1\}$ constitutes a valid POVM, which can be realized via an indirect measurement of the system using a single auxiliary qubit. Specifically, we measure the auxiliary qubit after it has interacted with the system via a suitable coupling unitary, effectively implementing a completely positive map [1]. The corresponding Kraus operators must satisfy the isometry condition $A_0^\dagger A_0 + A_1^\dagger A_1 = I$. Since the POVM $\{B_0, B_1\}$ is complete, we can always find suitable Kraus operators by taking the square root of the corresponding POVM elements: $A_0 = \sqrt{B_0}$ and $A_1 = \sqrt{B_1}$. Finally, we construct the coupling unitary U^B by stacking together the two binary Kraus operators A_0 and A_1 and completing the remaining matrix elements:

$$U^B = \begin{pmatrix} A_0 & * \\ A_1 & * \end{pmatrix} \in \mathbb{U}(2d). \tag{2}$$

The unitary operation is followed by the measurement of the auxiliary qubit in the computational basis. Finally, the auxiliary qubit is reset in the $|0\rangle$ state.

After this first partial filtering, we again split the remaining $M/2$ POVM elements in each branch into $M/4$ elements. This time, however, to implement the next step of partial filtering by measuring a binary POVM, we have to account for the postmeasurement state by modifying the Kraus operators for the subsequent steps $l \geq 2$ as follows:

$$A_{b^{(l)}} = K_{b^{(l)}} K_{b^{(l-1)}}^{-1}, \qquad (3)$$

where we use a binary string $b^{(l)}$ of length l to denote the sequence of measurement outcomes leading to the current branch in the binary search tree, e.g., $b^{(1)} \in \{0, 1\}$. Here, $K_{b^{(l)}} = \sqrt{\sum_{i=a}^{b} F_i}$ is obtained by aggregating the POVM elements located in the last level of the branch that starts from $b^{(l)}$ with indices ranging from a to b. Importantly, at every filtering step, we partition POVM elements such that each branch receives at least d nonzero POVM elements. As a result, the matrix $K_{b^{(l)}}$ is full rank for $l \leq m$, making it invertible. It is also worth noting that a unitary transformation $K_{b^{(l)}} \rightarrow W_{b^{(l)}} K_{b^{(l)}}$ with arbitrary unitary $W_{b^{(l)}}$ leaves the measurement statistics invariant and, thus, may have an optimization potential for constructing the coupling unitary in Eq. (2). As an example, Fig. 1(b) illustrates how a 16-element POVM on a two-qubit system is realized using the hybrid method. The blue box in Fig. 1(b) aggregates the corresponding POVM elements.

For POVMs with more than four elements, we continue bisecting the remaining POVM elements within each branch until each branch has at most $2d$ elements in its leaves. Precisely we arrive at this point after $m = \log_2(M/2d)$ iterations. For example, a 16-element POVM on two-qubits requires a single iteration of binary search, as shown in Fig. 1(b). At level m the cumulative Kraus operator is given as

$$A_{b^{(m)}}^c = A_{b^{(m)}} \cdots A_{b^{(2)}} A_{b^{(1)}} = K_{b^{(m)}}, \qquad (4)$$
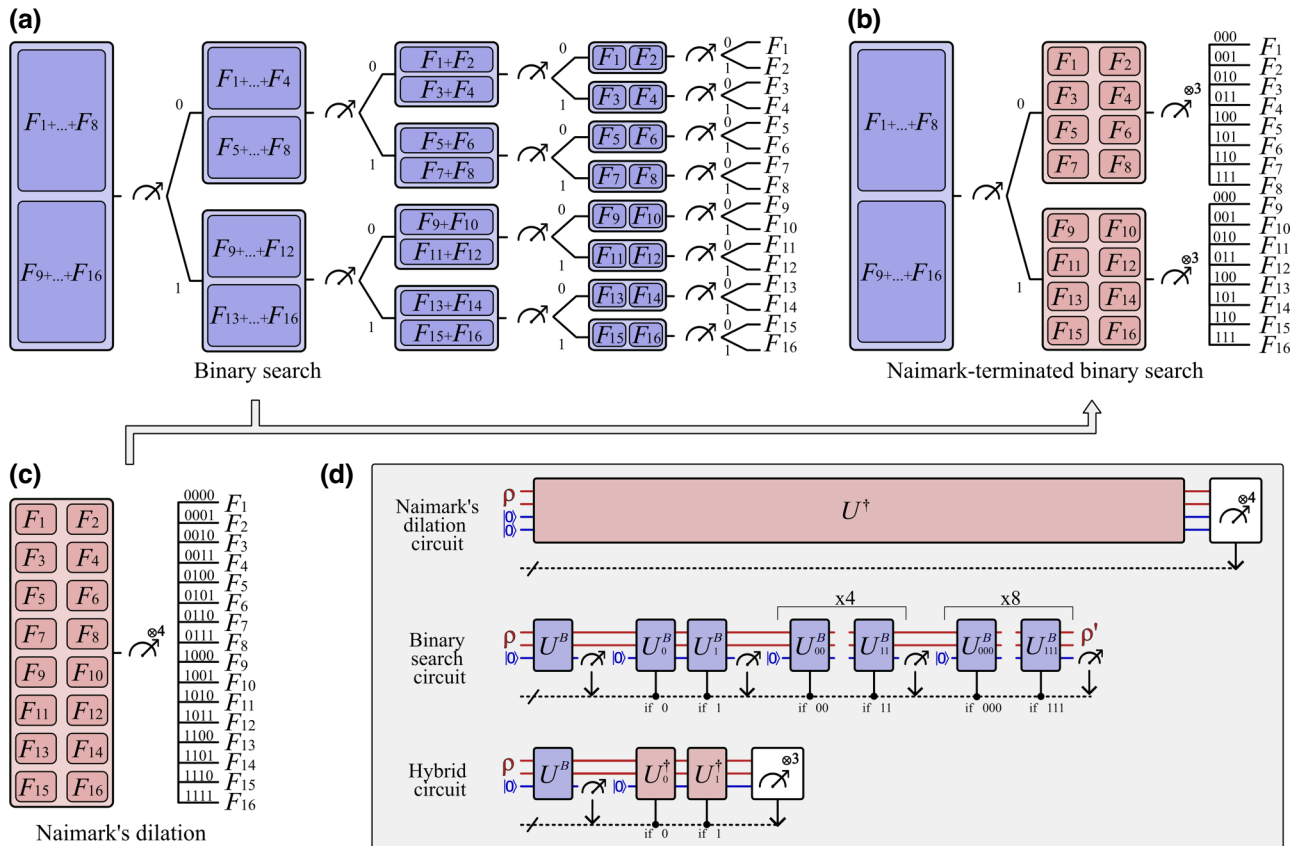

FIG. 1. Schematic of a two-qubit POVM with $M = 16$ POVM elements realized through (a) binary tree (blue), (b) hybrid scheme (blue and red), and (c) Naimark's dilation (red). Branching occurs after the measurement conditional on the outcome, with the trajectory defined by a bitstring of previous outcomes. Each scheme terminates with a bitstring of length four corresponding to one of 16 POVM elements. The number of qubits n and the number of POVM elements M can be arbitrary. Gray panel (d) shows quantum circuit implementations for methods (a)–(c), ordered from bottom to top. System qubits are in red, and auxiliary qubits are in blue. Binary search tree and hybrid circuit include midcircuit measurements and unitaries conditional on the classical register (dotted line). Longer unitary blocks correspond to the higher circuit depth.

and the conditional postmeasurement state of the system is

$$\rho_{s^{(m)}} = \frac{K_{b^{(m)}} \rho K_{b^{(m)}}^{\dagger}}{\mathrm{Tr}(K_{b^{(m)}} \rho K_{b^{(m)}}^{\dagger})}. \qquad (5)$$

In principle, we could continue dividing the set of POVM elements in half, narrowing down the search range until the target element is found, as shown in Fig. 1(a). This approach would require $\log_2 M - m$ additional iterations and a modified construction of Kraus operators for subsequent iterations, as detailed in Appendix H.

Instead, we observe that at level m, each branch has at most $2d$ elements in its leaves. For example, as illustrated in Fig. 1(b), each red box aggregates at most eight POVM elements. Therefore, at this point, the single auxiliary qubit suffices to apply Naimark's dilation. To account for the postmeasurement state, we modify the vectors $|\psi_i\rangle$ to $|\tilde{\psi}_i\rangle = K_{b^{(m)}}^{-1\dagger}|\psi_i\rangle$. The new POVM defined by $\tilde{F}_i = |\tilde{\psi}_i\rangle\langle\tilde{\psi}_i|$ satisfies the completeness relation because

$$\sum_{i=a}^{b} |\tilde{\psi}_i\rangle\langle\tilde{\psi}_i| = K_{b^{(m)}}^{-1\dagger} \left(\sum_{i=a}^{b} F_i \right) K_{b^{(m)}}^{-1} = I, \qquad (6)$$

where a and b stand for the first and last indices of aggregated POVM elements at the level m with $b - a + 1 = 2d$. Therefore, the modified vectors $|\tilde{\psi}_i\rangle$ can be arranged as column vectors to form a $d \times 2d$ array whose rows are orthonormal $2d$-dimensional vectors. Such an array can always be extended to a $2d \times 2d$ unitary matrix U, as detailed in Appendix G. By projectively measuring the compound system in the computational basis after the unitary transformation $U^{\dagger}$, as shown in Fig. 1(b), we obtain the outcomes $|\tilde{\psi}_i^{\,\mathrm{ext}}\rangle\langle\tilde{\psi}_i^{\,\mathrm{ext}}|$, where $|\tilde{\psi}_i^{\,\mathrm{ext}}\rangle$ are the columns of U. The conditional probability of observing the outcome i given that a string of previous measurement outcomes $b^{(m)}$ has been obtained is given by

$$P(i|b^{(m)}) = \mathrm{Tr}(|\tilde{\psi}_i\rangle\langle\tilde{\psi}_i|\rho_{s^{(m)}}) = \frac{\mathrm{Tr}(F_i\rho)}{\mathrm{Tr}(K_{b^{(m)}}\rho K_{b^{(m)}}^{\dagger})}, \qquad (7)$$

which results in the correct measurement statistics $P(i) = \mathrm{Tr}(F_i\rho)$.

Overall, the hybrid scheme involves $m = \log_2(M/2d)$ binary search steps and a final level that applies Naimark's dilation. This results in a total of $m + 1$ steps where $(n + 1)$-qubit unitaries are applied. In contrast, a standard binary search requires $\log_2 M$ layers of $(n + 1)$-qubit unitaries. Therefore, a hybrid circuit has $\log_2 d$ fewer layers. For instance, the 16-element POVM in Fig. 1 is implemented through the binary search in $\log_2 M = 4$ steps, each involving a unitary acting on $n + 1 = 3$ qubits. In contrast, hybrid involves only a single ($m = \log_2(M/2d) = 1$) binary search step and an additional level of conditional Naimark

unitaries, thus requiring two steps of three-qubit unitaries. Bare Naimark's dilation realizes the same POVM with a single application of a four-qubit unitary. Finally, it is worth noting that when $M \le 2d$, the hybrid scheme simplifies to a single application of Naimark's dilation without the need for any binary search steps. For example, a four-element POVM on a single qubit is most efficiently realized with a single application of Naimark's dilation, as detailed in Appendix A.

IV. EXPERIMENT

In this section, we compare the three methods by implementing SIC POVMs on one and two qubits. The choice of SIC POVMs as our benchmark is motivated by their extensive study in the literature, particularly for their unique tomographic properties [9]. Formally, a SIC POVM consist of $M = d^2$ elements $F_i = 1/d|\phi_i\rangle\langle\phi_i|$, which have equal pairwise overlap with each other $|\langle\phi_i|\phi_j\rangle|^2 = 1/(d + 1)$ for $i \ne j$. This symmetric property makes this class of POVMs particularly important in the context of optimal state tomography [20,30] and quantum key distribution [31,32]. However, practical implementation of SIC POVMs becomes challenging for more than one qubit.

A. Detector tomography of SIC POVMs

In practice, the realized POVM will differ from the ideal SIC POVM due to noise in the circuit [33]. To quantify the quality of POVM implementation, we perform detector tomography by preparing an overcomplete set of initial states and reconstructing the realized POVM from measurement statistics, as described in Appendix B. To calculate the fidelity, we represent the POVM as a measurement channel [34,35]:

$$\mathcal{E}_F(\rho) = \sum_{i=1}^{M} \mathrm{Tr}(F_i\rho) |i\rangle\langle i|, \qquad (8)$$

where the output state $\mathcal{E}_F(\rho)$ represents the POVM outcome probabilities. The POVM fidelity between the target POVM F and the realized POVM $\tilde{F}$ can be defined as

$$\mathcal{F}_D(\mathcal{E}_F, \mathcal{E}_{\tilde{F}}) = \mathcal{F}_S(\Lambda_{\mathcal{E}_F}, \Lambda_{\mathcal{E}_{\tilde{F}}}), \qquad (9)$$

where $\mathcal{F}_S(\rho, \sigma) = \mathrm{Tr}\left(\sqrt{\sqrt{\rho}\sigma\sqrt{\rho}}\right)^2$ is the conventional state fidelity. Here, $\Lambda_{\mathcal{E}_F}$ stands for the normalized Choi matrix of the quantum channel:

$$\Lambda_{\mathcal{E}_F} = \frac{1}{d} \sum_{i,j=1}^{d} |i\rangle\langle j| \otimes \mathcal{E}_F(|i\rangle\langle j|). \qquad (10)$$

The POVM fidelity can be reduced by different noise sources in the circuit. To mitigate the impact of coherent noise, we employ Pauli twirling on CNOTs, taking five

twirled instances for each circuit [36–40]. Furthermore, the effect of imprecise readout is mitigated using readout error mitigation [41,42]. In circuits involving midcircuit measurement and feed forward, measurement errors can propagate through the conditional operations and further reduce the fidelity. To address these errors, we extended the standard readout error mitigation to account for error propagation in dynamic circuits. As detailed in Appendix C, our CREM technique involves taking measurement statistics from additional circuits that provide the necessary information for the deconvolution of noisy outcome probabilities. Finally, we note that the fidelity of a POVM with completely random outcomes is $1/2^n$. Therefore, a non-trivial realization of a POVM should yield a fidelity higher than $1/2^n$. For example, for one qubit, the baseline fidelity is 50%, and for two qubits, it is 25%.

B. One-qubit SIC POVM

The one-qubit SIC POVM is implemented using Naimark's dilation and binary search, both requiring a single auxiliary qubit (see Appendix A). In either case, the circuit can be exactly compiled with a small number of CNOTs. Naimark's dilation requires three CNOTs and binary search results in an average CNOT depth of 4.5. Figure 2(a) summarizes the highest fidelities achieved with each method and the impact of readout error mitigation on ibmq_kolkata [43]. While both Naimark's dilation and binary tree require one auxiliary qubit for a single-qubit SIC POVM, there are additional sources of noise that limit the fidelity of the latter. Specifically, the binary tree applies two unitaries, which leads to a higher circuit depth. Moreover, it suffers from additional idle time due to midcircuit readout and the delay in processing measurement outcomes to determine the next unitary operation [26]. This reduces the fidelity of the binary-tree approach. After readout error mitigation, Naimark achieves a fidelity of $98.4 \pm 0.2\%$ while binary tree reaches $90.6 \pm 0.2\%$ using CREM. Naimark's fidelity increases by 2% due to readout error mitigation, while binary's increases by 1.3%. The smaller enhancement from readout error mitigation in binary tree is due to two main factors. Firstly, in our experiment, the auxiliary qubit has a lower readout error than the system qubit, and therefore, the binary-tree approach is less affected by those errors compared to Naimark. Secondly, qubit readout deviates from being a perfect quantum non-demolition (QND) measurement [44]. Thus, the postmeasurement state may differ from the recorded measurement outcome. Such discrepancies are not corrected by CREM as detailed in Appendix C.

C. Two-qubit SIC POVM

For the two-qubit SIC POVM, exact compilation of required unitary operations into native gates results in
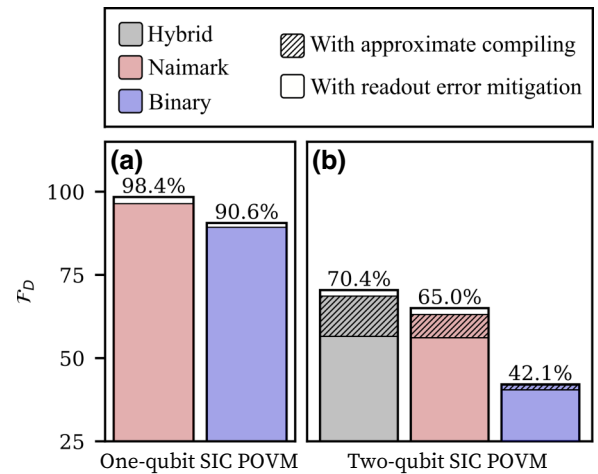

FIG. 2. Comparison of the highest achieved POVM fidelities for different implementations of one- (left) and two-qubit (right) SIC POVMs. The colored bars represent three different methods: Naimark (red), binary (blue), and hybrid (gray). White bars represent the improvement from readout error mitigation and hatched bars represent the improvement from approximate compiling. Error bars are obtained from bootstrapping and are too narrow to be visible.

circuits with a large CNOT count. This limits their applicability on near-term quantum hardware. To overcome this challenge, we use approximate compiling, which aims to find shorter circuits for a given unitary at the cost of only approximating the target unitary. In this approach, we constrain both the number and connectivity of CNOT gates within the circuit, which are interleaved with single-qubit rotation gates. An optimizer then tries to find the rotation angles for these single-qubit gates such that the resulting circuit implements the target unitary as accurately as possible [28]. We use approximate compiling for all unitaries involved in the construction of each of the three schemes. For Naimark's dilation, the optimization involves compiling a single four-qubit unitary, while for the binary search and hybrid methods, multiple three-qubit unitaries at different levels are compiled separately and then combined together to realize the measurement sequence.

When using approximate compiling, we expect a trade-off between circuit depth and the accuracy of the approximated POVM. For example, a low number of CNOT results in a large approximation error, as illustrated in the ideal simulation in Fig. 3. However, deeper circuits will encounter more noise, and *a priori* it is unclear which CNOT depth will result in the optimal performance. To approach this, we compile and run each algorithm at ten different CNOT depths, ranging from about 9 to 35, and for each algorithm, we pick the circuit with the highest POVM fidelity in the experiment, as shown in Fig. 3. Additional details of data acquisition can be found in Appendix J. Finally, Fig. 2(b) summarizes the highest
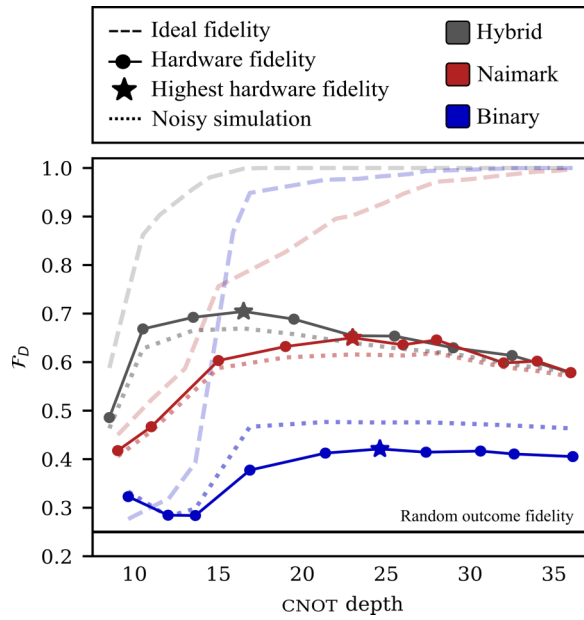
FIG. 3. POVM implementation fidelity across varying CNOT depths for binary tree (blue), Naimark's dilation (red), and hybrid (gray) schemes. The line styles represent ideal fidelity (dashed), depolarizing noise model simulation (dotted), and hardware fidelity (solid), with the peak hardware fidelity denoted by a star. The fidelity of a random outcome POVM is denoted by a horizontal line at $\mathcal{F}_D = 0.25$. Error bars are obtained from bootstrapping and are smaller than markers.

fidelities achieved on hardware with each method, along with the effect of readout error mitigation. In addition, the improvement from approximate compiling is represented by the difference between the optimal CNOT depth and the highest CNOT depth. For hybrid and Naimark, the improvement from approximate compiling is substantial. However, its effect is limited for binary tree due to long idle times in feed-forward operations, which are the main limiting factors for fidelity. We observe that the hybrid method results in the highest fidelity of $70.4 \pm 0.1\%$. Binary tree achieves a maximum of $42.1 \pm 0.1\%$ fidelity, significantly lower than the best fidelity of $65.0 \pm 0.1\%$ using Naimark's dilation. However, it is still higher than the baseline fidelity of 25%. Even though the binary tree is more efficient in terms of required CNOT depth, its fidelity is strongly affected by the high number of feed-forward operations in the three midcircuit measurements, causing longer idle times and coherence loss. We confirm this interpretation with numerical simulations of these experiments with a noise model inspired by the hardware in Sec. IV D. Finally, as in the one-qubit experiment, the lower improvement of only 0.2% from CREM in the binary tree as compared to the 1.9% for Naimark and 1.8% for the hybrid circuit is attributed to the comparatively low readout errors on the auxiliary qubit as well as the non-QND errors in the auxiliary qubit readout.

D. Noise analysis

CNOT gates are significantly noisier than single-qubit gates. Therefore, their count is a key noise metric. However, in the context of dynamic circuits, fidelity may further degrade due to idle times from midcircuit measurements and feed-forward operations. Each conditional unitary in binary and hybrid circuits involves verifying the unitary's condition, with a time cost comparable to that of a CNOT gate's duration.

To provide a phenomenological explanation for the two-qubit SIC POVM results, we perform a noisy simulation using a depolarizing noise model. Firstly, to model the fidelity decay due to CNOT depth, we attach a depolarizing error $\epsilon_{CNOT} = 1.5\%$ to every CNOT gate. Secondly, we associate a depolarizing error $\epsilon_{idle} = 5\%$ with each measurement and feed-forward operation. This is motivated by the comparatively long measurement and feed-forward times, usually multiple times longer than the CNOT gate time [26]. We choose the depolarizing errors heuristically, and the error values are consistent with recent experiments on analogous devices [26,44,45]. The binary tree is impacted the most by this second kind of error because it has three midcircuit measurements and 14 feed-forward cases, as shown in Fig. 1(d). In contrast, the hybrid only has one midcircuit measurement and two feed-forward cases. On the other hand, Naimark's circuit requires the highest CNOT depth for approximate compilation and is, therefore, additionally affected by the approximation error, as shown in Fig. 3. Overall, the hybrid scheme outperforms its constituents across all CNOT depths, consistent with the experimental results. Despite the phenomenological nature of our noise model, we achieved notable agreement with the experimental data, suggesting that the error model captures the two major noise sources. An extended analysis of the performance of the three schemes in different noise regimes can be found in Appendix I.

E. Scaling to larger systems

We now discuss the resource costs of the three algorithms when scaling to larger systems. As an illustrative example, we focus on POVMs with $M = d^2$ elements, such as informationally complete POVMs [8], and extend this discussion in Appendix D. Naimark's dilation realizes a POVM with $M = d^2$ elements through a single unitary acting on $\log_2 M = 2n$ qubits followed by a layer of end-circuit measurements. For an n-qubit system, Naimark thus requires n auxiliary qubits. Notably, end-circuit measurements can usually be executed in parallel, effectively counting as a single measurement step. Binary search utilizes $\log_2 M = 2n$ layers of $(n+1)$-qubit unitaries, interleaved by $2n - 1$ midcircuit measurements and a final end-circuit measurement. Of these $2n$ unitaries, $2n - 1$ are conditional. In contrast, the hybrid uses $\log_2(M/d) = n$ layers of $(n+1)$-qubit unitaries, interleaved by $n - 1$

midcircuit measurements and terminating with $n + 1$ end-circuit measurements. Thus, the hybrid approach requires half as many layers compared to binary. Additionally, its final $n + 1$ end-circuit measurements can also be parallelized, providing yet another advantage over binary. For both binary and hybrid, each midcircuit measurement is followed by resetting the auxiliary qubit. This reset can be achieved at minimal cost by applying an X gate conditional on the measurement outcome (active reset).

We consider all unitaries in the binary search and Naimark's dilation as generic. Therefore, an upper bound of $\mathcal{O}(4^n)$ CNOTs required for decomposing a generic n-qubit unitary [46] serves as a common cost unit for the circuit depth of all three schemes. Consequently, the CNOT depths for Naimark, binary, and hybrid schemes scale as $\mathcal{O}(16^n)$, $\mathcal{O}(2n \times 4^{n+1})$, and $\mathcal{O}(n \times 4^{n+1})$, respectively. We conclude that the hybrid scheme requires asymptotically the shortest circuit. Furthermore, since the number of midcircuit measurements and conditional operations increases only linearly with system size, the CNOT depth emerges as the critical cost factor for larger systems. Finally, in Appendix I, we provide further analysis on the impact of noise from midcircuit measurements and conditional operations in the two-qubit experiment.

V. CONCLUSION

In conclusion, we introduce a new approach for implementing single-setting POVMs on multiqubit superconducting systems using dynamic circuits. Our method results in shorter-depth circuits and outperforms both Naimark's dilation and binary tree in implementing a two-qubit SIC POVM. We further demonstrate that approximate compiling is an effective approach to realizing generalized measurements under noisy conditions. In addition, we devise a new CREM technique to combat error propagation in dynamic circuits and enhance the fidelity. We limit our implementation to two-qubit POVMs, as our attempts to extend this approach to three-qubit POVMs resulted in prohibitively large circuit depths in the order of hundreds of CNOTs, surpassing any previously successful implementation on similar hardware. We, therefore, expect that implementing higher-dimensional POVMs using our approach will require improvements in hardware, such as faster feed forward and more efficient approximate compiling techniques. For example, an optimized compiling strategy could exploit the unitary freedom in the definition of Kraus operators in Eq. (3) or try permutations of POVM elements. Nevertheless, our results open new possibilities in the near future. For example, parallel execution of products of two-qubit POVMs would allow multiple qubits to be covered on the chip, a strategy with potential implications for multiqubit state tomography or classical shadows [20]. We also suggest incorporating our hybrid approach in the development of hardware-efficient,

parametric POVMs, which has proven to be an effective technique for targeted applications [15].

Note added.—Recently, we became aware of a related but independently developed error-mitigation technique for mid-circuit measurements [47].

APPENDIX A: ONE-QUBIT POVM IMPLEMENTATION

A single-qubit rank-one POVM can have at most $M = 4$ linearly independent elements. Using Naimark's dilation, we can realize such a POVM with a single auxiliary qubit by applying a suitable coupling unitary and measuring the system and auxiliary qubit in the computational basis, as detailed in Appendix G. Therefore, a single-qubit POVM is a special case where the dimension of unitaries for the binary tree and Naimark's dilation coincide. In this case, we do not expect binary to offer any advantage over Naimark because it requires two layers of two-qubit unitaries, as shown in Fig. 4. Finally, our hybrid approach simplifies to a single application of Naimark's dilation because there is no need for partial filtering.

The advantage of Naimark over binary is confirmed by the higher fidelities obtained in the one-qubit realization of a SIC POVM in Sec. IV B.

APPENDIX B: DETECTOR TOMOGRAPHY

Detector tomography is used to reconstruct an unknown detector, or a POVM in the present work. In many aspects, it is similar to state tomography, where an unknown state is
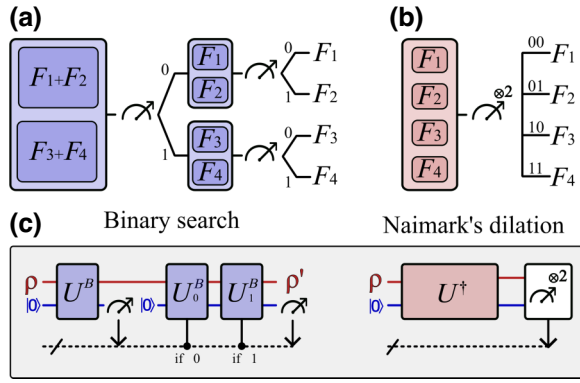
FIG. 4. Schematic representation of a single qubit POVM with $M=4$ POVM elements realized through (a) binary tree (blue), and (b) Naimark's dilation (red). Branching occurs after the measurement conditional on the outcome, with the trajectory defined by a bitstring of previous outcomes. Each scheme terminates with a bitstring of length two corresponding to one of four POVM elements. Gray panel (c) shows quantum circuit implementations for methods (a) and (b). System qubits are in red, and auxiliary qubits are in blue. Binary tree includes a single midcircuit measurement and two unitaries conditional on the classical register (dotted line).

reconstructed from the measurement statistics of a known detector. In contrast, in detector tomography, we aim to reconstruct an unknown detector by preparing and measuring a set of known initial states. To elaborate, for a specific known state ρ_i, the probability of getting a measurement outcome F_m is given by $p_{mi} = \text{tr}(F_m \rho_i)$. Every F_m can be expressed using multiqubit Pauli matrices: $F_m = \sum_{k=1}^{d^2} f_{mk}\sigma_k$. This leads to the probability equation:

$$p_{mi} = \sum_{k=1}^{d^2} f_{mk}\text{tr}(\sigma_k\rho_i), \tag{B1}$$

which is linear and analogous to state tomography. We define the matrix S as the matrix containing trace values that are specific to the set of initial states:

$$S = \begin{pmatrix} \text{tr}(\sigma_1\rho_1) & \cdots & \text{tr}(\sigma_{d^2}\rho_1) \\ \text{tr}(\sigma_1\rho_2) & \cdots & \text{tr}(\sigma_{d^2}\rho_2) \\ \vdots & \vdots & \vdots \\ \text{tr}(\sigma_1\rho_n) & \cdots & \text{tr}(\sigma_{d^2}\rho_n) \end{pmatrix}. \tag{B2}$$

Then, by writing $\vec{f}_m$ as the vector of coefficients in front of Pauli matrices in the expansion of F_m and $\vec{p}_m$ as the vector of corresponding probabilities determined from the experiment, we find the solution to $\vec{p}_m = S\vec{f}_m$ by computing the least-squares estimate:

$$\vec{f}_m = (S^\dagger S)^{-1}S^\dagger \vec{p}_m. \tag{B3}$$

The procedure above has to be repeated for every unknown POVM element F_m. In general, the least-squares estimate is not guaranteed to produce a non-negative F_m, especially with few measurement samples. Therefore, the Choi matrix in Eq. (10) of the reconstructed POVM may have small negative eigenvalues. To impose the non-negativity, one can rescale the eigenvalues of the Choi matrix from least squares, effectively projecting it onto the set of quantum states [48,49]. This procedure corresponds to computing the maximum-likelihood estimate under the assumption of Gaussian noise [48]. In our experiments, however, we take sufficient measurement samples such that rescaling is only necessary for a few POVM elements. In particular, rescaling has no effect on fidelities within the statistical uncertainty.

For a two-qubit experiment, 36 Pauli basis states were used, making the set of initial states overcomplete. Consequently, the matrix S had dimensions 36×16. We also used an overcomplete set of six Pauli basis states for a one-qubit experiment, resulting in a 6×4 matrix. In principle, only d^2 initial states are fundamentally required to form a complete basis. However, using extra states improves the estimation accuracy and compensates for the state-dependent performance of a particular POVM implementation.

APPENDIX C: CONDITIONAL READOUT ERROR MITIGATION

Imprecise multiqubit measurements can be described using a classical probabilistic model. Measurement errors arise from random misclassification of correct outcomes, represented by a confusion matrix M, where $M_{ij} = \text{p}(i|j)$ is the probability of misclassifying j as i. Here, i and j are computational basis states. Observed probabilities (P) are obtained by applying the confusion matrix to ideal probabilities (Q): $P = MQ$. For example, for the case of measuring one qubit ($i \in 0, 1$), the confusion matrix is given as

$$M = \begin{pmatrix} 1 - \epsilon_0 & \epsilon_1 \\ \epsilon_0 & 1 - \epsilon_1 \end{pmatrix}. \tag{C1}$$

For instance, the probability of obtaining 0 is the sum of the probability $(1 - \epsilon_0)$ of correctly identifying 0 as 0 and the probability ϵ_1 of misclassifying 1 as 0. Error-free probabilities Q can be obtained by inverting the confusion matrix: $Q = M^{-1}P$. In general, measuring n qubits in the computational basis can result in any of 2^n computational states with some unique probability. Therefore, to characterize the full confusion matrix M, one needs to obtain 4^n matrix elements through calibration. Often, however, it is justified to assume that readout errors are local, i.e., the probability of misclassifying the state of some qubit is independent of the states of other qubits. This allows us to construct the

global confusion matrix M as the tensor product of n 2×2 confusion matrices M_i, one for each qubit:

$$M = M_n \otimes M_{n-1} \otimes \ldots \otimes M_1. \qquad \text{(C2)}$$

The situation is different with dynamic circuits, which enable midcircuit measurements, allowing subsequent gates to be conditional on these measurement outcomes. For instance, consider the two-qubit circuit in Fig. 5. A unitary operation U prior to the first midcircuit measurement encompasses the circuit's unitary segment. After measuring qubit q_1, the result is stored in the first bit of the classical register. Subsequently, the circuit applies U_0 if the measurement reports 0 and U_1 if 1. We see that if a bit-flip error occurs during the midcircuit measurement, then not only is the wrong measurement outcome recorded, but also the wrong condition is triggered, and the wrong unitary is applied. Moreover, the unitaries now act on the flipped postmeasurement state of qubit q_1. In this way, the measurement error from midcircuit measurement propagates through the circuit, affecting the evolution of the quantum state. To account for this error propagation, consider the second circuit in Fig. 5. The second circuit, which we call a conditional calibration circuit, differs from the original circuits in that the conditions of the two unitaries are inverted, and an X gate flips the postmeasurement state. Notice that if a bit flip occurs during the midcircuit measurement in the second circuit, the subsequent state evolution will correspond to the case when no bit-flip occurs in the first circuit. Denote ϵ_i^k as the probability of misclassifying the state i

on qubit q_k; P and Q as the noisy and ideal probabilities for the original circuit; and $\tilde{P}$ and $\tilde{Q}$ as the corresponding probabilities for the conditional calibration circuit. Also, P_{ij} denotes the probability of measuring q_1 in state j and q_2 in state i. We can now express the noisy probability P_{00} of obtaining the outcome 00 as

$$P_{00} = (1 - \epsilon_0^2)(1 - \epsilon_0^1)Q_{00} + (1 - \epsilon_0^2)\epsilon_1^1 \tilde{Q}_{01}$$
$$+ \epsilon_1^2(1 - \epsilon_0^1)Q_{10} + \epsilon_1^2\epsilon_1^1 \tilde{Q}_{11}.$$

We see that the noisy probability P_{00} includes four different scenarios:

(1) The ideal case scenario when no readout error occurs during both midcircuit and end-circuit measurement with probability $(1 - \epsilon_0^2)(1 - \epsilon_0^1)$.
(2) The outcome 01 is misclassified as 00 with probability $(1 - \epsilon_0^2)\epsilon_1^1$ due to the error on the midcircuit measurement. However, the subsequent state evolution corresponds to the error-free scenario.
(3) The outcome 10 is misclassified as 00 with probability $\epsilon_1^2(1 - \epsilon_0^1)$ due to the error on the end-circuit measurement.
(4) The outcome 11 is misclassified as 00 with probability $\epsilon_1^2\epsilon_1^1$ due to errors on both measurements, whereby the effective state evolution is unaffected by measurement errors.

In a similar manner we can obtain expressions for P_{01}, P_{10}, and P_{11}:

$$P_{01} = \epsilon_0^1(1 - \epsilon_0^2)\tilde{Q}_{00} + (1 - \epsilon_0^2)(1 - \epsilon_1^1)Q_{01} + \epsilon_0^1\epsilon_1^2\tilde{Q}_{10} + (1 - \epsilon_1^1)\epsilon_1^2 Q_{11},$$
$$P_{10} = (1 - \epsilon_0^1)\epsilon_0^2 Q_{00} + \epsilon_0^2\epsilon_1^1\tilde{Q}_{01} + (1 - \epsilon_0^1)(1 - \epsilon_1^2)Q_{10} + \epsilon_1^1(1 - \epsilon_1^2)\tilde{Q}_{11},$$
$$P_{11} = \epsilon_0^1\epsilon_0^2\tilde{Q}_{00} + \epsilon_0^2(1 - \epsilon_1^1)Q_{01} + \epsilon_0^1(1 - \epsilon_1^2)\tilde{Q}_{10} + (1 - \epsilon_1^1)(1 - \epsilon_1^2)Q_{11}.$$

Following the same logic, we obtain the expression for $\tilde{P}_{00}$, $\tilde{P}_{01}$, $\tilde{P}_{10}$, and $\tilde{P}_{11}$. It is convenient to write

$$\begin{pmatrix} P_{00} \\ \tilde{P}_{01} \\ P_{10} \\ \tilde{P}_{11} \end{pmatrix} = \left[\begin{pmatrix} 1 - \epsilon_0^1 & \epsilon_1^1 \\ \epsilon_0^1 & 1 - \epsilon_1^1 \end{pmatrix} \otimes \begin{pmatrix} 1 - \epsilon_0^2 & \epsilon_1^2 \\ \epsilon_0^2 & 1 - \epsilon_1^2 \end{pmatrix} \right] \begin{pmatrix} Q_{00} \\ \tilde{Q}_{01} \\ Q_{10} \\ \tilde{Q}_{11} \end{pmatrix} = [M_2 \otimes M_1] \begin{pmatrix} Q_{00} \\ \tilde{Q}_{01} \\ Q_{10} \\ \tilde{Q}_{11} \end{pmatrix}$$

and

$$\begin{pmatrix} \tilde{P}_{00} \\ P_{01} \\ \tilde{P}_{10} \\ P_{11} \end{pmatrix} = \left[\begin{pmatrix} 1 - \epsilon_0^1 & \epsilon_1^1 \\ \epsilon_0^1 & 1 - \epsilon_1^1 \end{pmatrix} \otimes \begin{pmatrix} 1 - \epsilon_0^2 & \epsilon_1^2 \\ \epsilon_0^2 & 1 - \epsilon_1^2 \end{pmatrix} \right] \begin{pmatrix} \tilde{Q}_{00} \\ Q_{01} \\ \tilde{Q}_{10} \\ Q_{11} \end{pmatrix} = [M_2 \otimes M_1] \begin{pmatrix} \tilde{Q}_{00} \\ Q_{01} \\ \tilde{Q}_{10} \\ Q_{11} \end{pmatrix}.$$
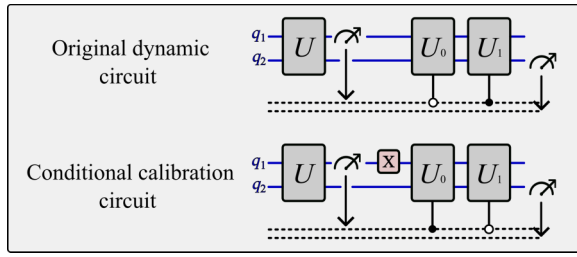
FIG. 5. Conditional readout error mitigation. (a) Original two-qubit circuit with a unitary applied to q_1 and q_2, followed by a midcircuit measurement, with subsequent operations conditioned on the measurement's outcome. (b) Conditional calibration circuit replicating the original with the feed-forward conditions exchanged, and the postmeasurement state flipped.

Here, M_i denotes the confusion matrix on qubit q_i. In a similar manner to the standard readout error mitigation, we can invert the Kronecker product $(M_2 \bigotimes M_1)^{-1} = M_2^{-1} \bigotimes M_1^{-1}$ to obtain error-free probabilities Q. As a byproduct, we also obtain error-free probabilities $\tilde{Q}$ for the conditional calibration circuit. It is worth noting that the described postprocessing requires obtaining probabilities for two circuits (original and calibration circuit), which doubles the total number of samples.

The described procedure extends to multiple midcircuit measurements. The number of required conditional calibration circuits (including the original circuit) is given by the number of possible combinations of measurement errors, which is $2^{n_{\mathrm{mid}}}$ where n_{mid} is the number of midcircuit measurements. Therefore, the sampling overhead scales exponentially with n_{mid}, which makes this technique feasible only for a few numbers of midcircuit measurements. Note that the underlying assumption we made about the nature of readout errors is that readout is perfectly QND. In this setting, CREM can correct readout errors. In reality, the readout of qubits is not perfectly QND, which can result in inconsistencies between the postmeasurement state and the measurement outcome. For example, a qubit can decay during the measurement pulse. Moreover, applying measurements can cause leakage from the computation subspace to other states outside of the computational space. These errors contribute to non-QND measurement errors,
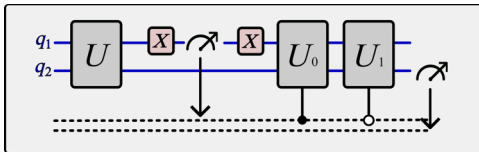

FIG. 6. Two-qubit circuit modeling non-QND readout errors as postmeasurement bit flip with probability ϵ_{QND}. The "correctable" error is modeled as a premeasurement bit flip with probability ϵ.
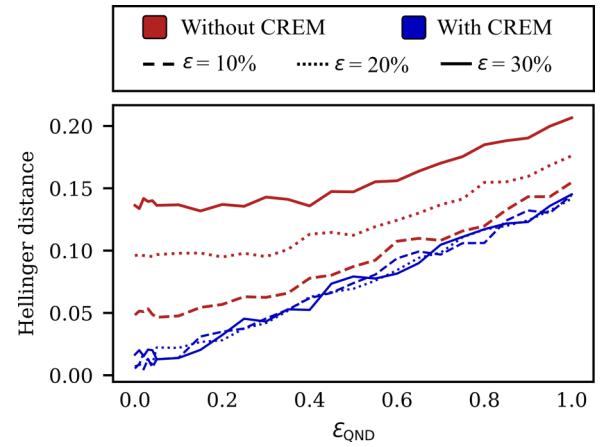

FIG. 7. Illustration of the increase in Hellinger distance between the noisy and noiseless probability distributions with the rise of non-QND error rate ϵ_{QND}, comparing outcomes with (blue) and without (red) CREM across different "correctable" error rates ϵ. In the presence of both errors, CREM removes the contribution of "correctable" errors.

which alter the state after the measurement [44]. In midcircuit measurements, these errors affect the subsequent computations, whereas in final measurements they manifest as initialization errors in the next round. CREM does not account for non-QND errors. To illustrate how CREM performs in the presence of non-QND errors, we use a simple noise model that includes random bit flips before and after the noiseless measurement pulse. The corresponding two-qubit circuit is shown in Fig. 6. The random bit flip before the measurement occurs with probability ϵ and represents the error that misclassifies the qubit state but leaves the postmeasurement state consistent with the measurement outcome. The random bit flip with probability ϵ_{QND} after the measurement represents the non-QNDness of the readout. The measurement on the second qubit is ideal, and the unitaries in the circuit are chosen randomly. By sweeping the non-QND error ϵ_{QND}, as shown in Fig. 7, we observe that the Hellinger distance between the noisy and ideal measurement outcome distributions increases for both mitigated and unmitigated results. However, the error of unmitigated results also grows with ϵ, while mitigated results are independent of ϵ. Therefore, in a realistic setting, where both kinds of errors are present, CREM takes out the contribution of the first type of error.

APPENDIX D: RESOURCE ESTIMATE

In Sec. IV E of the main text, we discuss the upper bound on the number of CNOTs for all three schemes for POVMs with $M = d^2$. Here, we extend the discussion to POVMs with $M \leq d^2$, still assuming that POVM elements are linearly independent rank-one operators. As follows from the main text, Naimark's dilation requires a single $\lceil \log_2 M \rceil$-qubit unitary. The binary-tree scheme requires $n + 1$-qubit

TABLE I.   Upper bounds on the number of CNOTs required for POVM implementation.

M	Naimark	Binary	Hybrid
$2^n < M \leq 2^{n+1}$	4^{n+1}	$(n+1) \times 4^{n+1}$	4^{n+1}
$2^{n+1} < M \leq 4^n$	M^2	$\log_2 M \times 4^{n+1}$	$(\log_2 M - n) \times 4^{n+1}$
Example for a minimal IC POVM			
$M = 4^n$	16^n	$2n \times 4^{n+1}$	$n \times 4^{n+1}$

unitaries to be implemented $\lceil \log_2 M \rceil$ times. Finally, the hybrid scheme requires $n + 1$-qubit unitaries to be implemented $\lceil \log_2 \frac{M}{2^n} \rceil$ times. Table I provides asymptotic upper bounds on the number of CNOTs. For clarity, we treat the case when $2^n < M \leq 2^{n+1}$ separately. We first pad the POVM with zero operators to the nearest power of two. That is, if a given POVM initially has $2^n < M < 2^{n+1}$ elements, we pad it until $M = 2^{n+1}$. The POVM with $M = 2^{n+1}$ elements can be most efficiently implemented with Naimark's dilation using a single auxiliary qubit. Binary search requires $(n + 1)$ partial filtering steps, while the hybrid scheme coincides with Naimark's dilation, as in the case of one-qubit SIC POVM in Appendix IV B. An important example is a minimal informationally complete POVM, i.e., a POVM with 4^n linearly independent POVM elements, such as SIC POVM. In this case, the hybrid scheme results in a circuit with half the length compared to the binary search, as shown in Table I.

In summary, the hybrid scheme yields a shorter circuit than the binary search tree for any system size and number of POVM elements. It outperforms Naimark's dilation for $M > 2^{n+1}$ while coinciding with Naimark's dilation when $M \leq 2^{n+1}$.

APPENDIX E: STATE TOMOGRAPHY

Quantum state tomography is one of the prominent examples where the choice of measurement plays a central role. A typical approach to perform state reconstruction of a multiqubit system is to perform tomographic measurements on each individual qubit using each of the Pauli bases. For an n-qubit system, this requires 3^n measurement settings, a number which grows exponentially with the system size, making it impractical for large numbers of qubits. Moreover, Pauli bases are known to be suboptimal for state tomography, requiring more samples than necessary due to the informational redundancy of its projectors. The last aspect is investigated in more detail in Appendix F.

SIC POVMs, in contrast, are known to be sample-optimal for state tomography [9]. In the previous sections, we performed the detector tomography by assuming perfect knowledge of the prepared states to obtain the tomographic information about the implemented POVM. Conversely, we could assume the perfect knowledge of the detector (SIC POVM) to reconstruct the prepared quantum states. Of course, since the implemented POVM is imperfect, the resulting reconstruction fidelity will deviate from the ideal.

Notably, we do not need to perform any additional measurements to the ones obtained for detector tomography. We simply recast the problem and use the ideal SIC POVM and the measurement statistics to perform the linear inversion. In practice, the reconstructed states may not be physical due to small negative eigenvalues. To address this problem without significant computational overhead, we rescale the eigenvalues of each reconstructed density matrix to make it positive semidefinite [48]. Figure 8 shows the average reconstruction fidelities of the six single-qubit Pauli basis states and the 36 two-qubit Pauli basis states. Here, for the two-qubit experiment, we used the optimal CNOT depths determined in Sec. IV. We see that, overall, the achieved peak fidelities by each algorithm agree closely with the corresponding POVM fidelities.

APPENDIX F: RESOURCE OVERHEAD OF STATE TOMOGRAPHY

In the context of state tomography with Pauli basis, we need 3^n different measurement settings, each yielding d outcomes. The total number of measurement outcomes
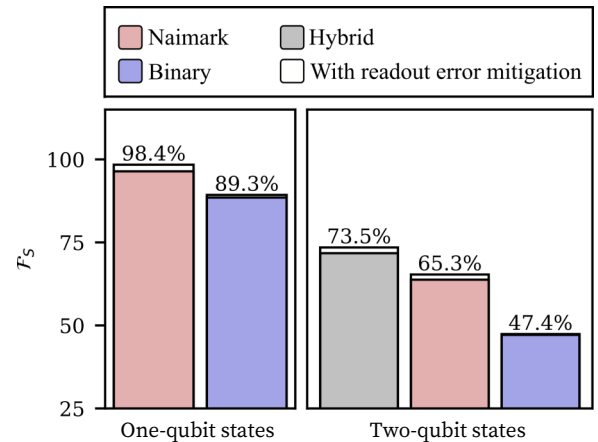

FIG. 8.   Comparison of the highest achieved (average) state fidelities for different implementations of one- (left) and two-qubit (right) SIC POVMs. The colored bars represent three different methods: Naimark (red), binary (blue), and hybrid (gray). White bars represent the improvement from readout error mitigation.
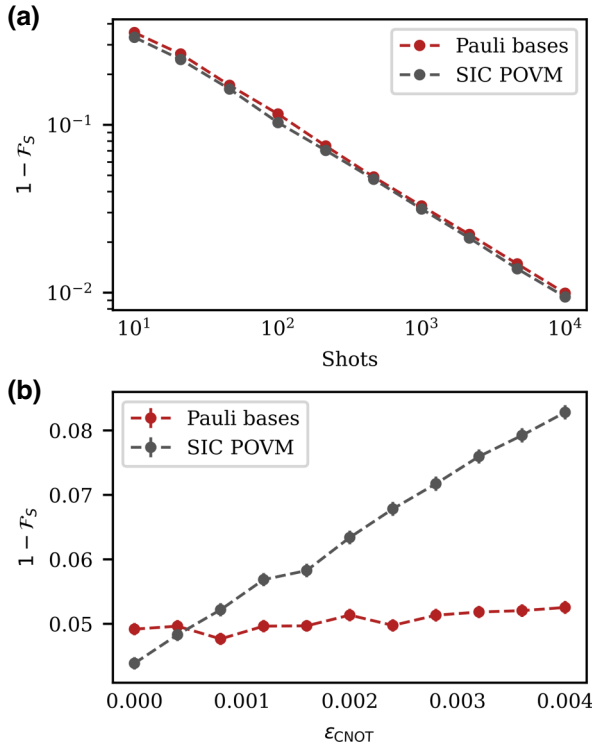
FIG. 9. Comparison of the infidelity of state reconstruction between Pauli bases (red) and SIC POVM (gray) as a function of the number of measurement shots (a) and the CNOT gate error rate (b). In (a), the simulation is noiseless, and in (b), the number of shots is fixed at 500.

exceeds the $d^2 - 1$ parameters of a d-dimensional density matrix. Therefore, Pauli bases provide redundant information, which results in a higher variance in the fidelity of the reconstructed state [31].

SIC POVMs, however, are proven optimal for state tomography [9]. This is observed in Fig. 9 by comparing the infidelity of reconstruction for a two-qubit SIC POVM with varying numbers of shots using a noiseless Qiskit simulator. At lower shots, SIC POVM exhibits a slight advantage, around 1% at 100 shots and 0.15% at 1000 shots. As the number of shots increases, both methods converge to zero infidelity at a rate of $N^{\frac{1}{2}}$, evident from the $-1/2$ slope of the curves. However, due to noise in the circuit that implements a SIC POVM, its performance is considerably lower than that of the Pauli bases. Nevertheless, if the noise level of the device is sufficiently low, using SIC POVMs for state tomography will be more efficient.

We investigate the performance of SIC POVM under a simplistic noise model that applies depolarizing noise to the CNOT gates. The purpose of this analysis is to showcase a crossover point where SIC POVM reconstruction yields a higher fidelity than Pauli bases for a fixed number of 500 shots. Figure 9 shows the corresponding trade-off. At low

depolarizing noise, SIC POVM results in lower infidelity, which grows as we increase the error. Pauli basis tomography is expectedly unaffected by this noise model because Pauli basis measurements do not include any CNOTs. A more realistic noise model would include single-qubit and measurement errors.

APPENDIX G: DETAILS OF NAIMARK'S DILATION

Any POVM can be realized by a projective measurement in a higher-dimensional Hilbert space by introducing a number of auxiliary qubits [17]. In general, the dimension of the required auxiliary system dimension will depend on the number of POVM elements, the dilation method, and the ability to perform direct measurements on the system [50]. One way to extend the Hilbert space is by, for example, coupling the qubit system to neighboring qubits on a superconducting chip. Moreover, in the scope of this paper, we are only interested in the measurement statistics $P(i)$ and disregard the postmeasurement state of the system. Therefore, we consider the situation where both the system and auxiliary qubit are measured directly. The extension of the Hilbert space requires a minimal number of n_A auxiliary qubits so that the dimension of the compound system $2^{n+n_A} \geq M$. Therefore, the auxiliary qubit resource is best utilized when M is a power of 2. Otherwise, we have to pad our set of POVM elements with zero operators until M is a power of 2. For a POVM of M elements we find M unnormalized vectors $|\psi_i\rangle$ such that $F_i = |\psi_i\rangle\langle\psi_i|$. These vectors can be arranged as column vectors to form an $d \times M$ array whose rows are orthonormal M-dimensional vectors. This fact follows from the completeness of POVM: $\sum_{i=1}^{M} F_i = I$. Such an array can always be extended to a $M \times M$ unitary matrix, as shown in Eq. (G1):

$$U = \begin{pmatrix} \psi_{11} & \cdots & \psi_{M1} \\ \vdots & & \vdots \\ \psi_{1d} & \cdots & \psi_{Md} \\ * & \vdots & * \end{pmatrix} \in \mathbb{U}(M). \qquad \text{(G1)}$$

The circuit implementing Naimark's dilation is constructed by preparing the auxiliary qubit in the $|0\rangle^{\otimes n_A}$ state and applying the coupling unitary $U^\dagger$. Finally, by projectively measuring the compound system in the computational basis after the unitary transformation $U^\dagger$, we obtain the outcomes $|\psi_i^{\text{ext}}\rangle\langle\psi_i^{\text{ext}}|$, where $|\psi_i^{\text{ext}}\rangle$ are the columns of $U^\dagger$. The probability of observing an outcome i is given by taking the trace over the auxiliary qubits and

the system:

$$P(i) = \text{Tr}(|\psi_i^{\text{ext}}\rangle\langle\psi_i^{\text{ext}}|(\rho_A \otimes \rho))$$
$$= \text{Tr}(|\psi_i\rangle\langle\psi_i|\rho)$$
$$= \text{Tr}(F_i\rho).$$

With this, we fully recover the original POVM measurement statistics $P(i) = \text{Tr}(F_i\rho)$.

Finally, we illustrate the procedure in Fig. 4(b) for a four-element POVM on a single-qubit system with the corresponding circuit depicted in Fig. 4(c). Moreover, an example of Naimark's dilation applied to a two-qubit system to realize a 16-element POVM is illustrated in Fig. 1(c) with the circuit shown in Fig. 1(d).

In conclusion, it is worth noting that the way in which the dilation is realized can be different. For example, suppose the system of interest harbors more dimensions than the dimension of its computational subspace, such as higher-energy states of superconducting qubits [19] or trapped ions [20]. In that case, the extended Hilbert space can be a direct sum $\mathcal{H}_{\text{ext}} = \mathcal{H} \oplus \mathcal{H}_A$ of the system and auxiliary spaces [50,51]. Often, such dilation is rather difficult because one needs to discriminate between qudit states efficiently.

APPENDIX H: DETAILS OF THE BINARY-TREE CONSTRUCTION

In this section, we outline the key steps of the binary tree protocol using the notation from Shen *et al.* [24]. A detailed treatment can be found in Andersson *et al.* [23] and Shen *et al.* [24]. To construct the binary search tree, we begin by padding our set of POVM elements with zero operators until M is the nearest power of 2. At the first level $l = 1$, we find a suitable Kraus operator A_0 for the coupling unitary from the diagonalization of $B_0 = V_0 D_0^2 V_0^\dagger$, resulting in $A_0 = V_0 D_0 V_0^\dagger$. Here, D_0 is a diagonal matrix with non-negative eigenvalues because B_0 is positive Hermitian. We analogously construct $A_1 = V_1 D_1 V_1^\dagger$ from $B_1 = V_1 D_1^2 V_1^\dagger$. As detailed in the main text, we implement this binary POVM via an indirect measurement of the system by constructing a suitable coupling unitary.

After the initial two-outcome POVM, the postmeasurement state of the system, denoted as $\rho_{b^{(1)}}$, will depend on the measurement outcome. For instance, if the auxiliary qubit is measured in the $|0\rangle$ state, the postmeasurement state ρ_0 can be expressed as

$$\rho_0 = \frac{A_0 \rho A_0^\dagger}{\text{Tr}(A_0 \rho A_0^\dagger)}, \tag{H1}$$

where ρ represents the initial state of the system. Therefore, the measurement of the auxiliary qubit causes branching, effectively performing partial filtering. This procedure

may be seen as a quantum instrument that combines the classical measurement outcome with the conditional postmeasurement quantum state of the system [34]. The subsequent binary POVMs must take this postmeasurement state into account. Therefore, we modify the Kraus operators for all subsequent steps $l \geq 2$ as follows:

$$A_{b^{(l)}} = K_{b^{(l)}} K_{b^{(l-1)}}^+ + \frac{1}{\sqrt{2}} Q_{b^{(l-1)}}, \tag{H2}$$

where $K_{b^{(l)}} = \sqrt{\sum_{i=a}^{b} F_i}$ is obtained by aggregating the POVM elements located in the last level of the branch that starts from $b^{(l)}$ with indices ranging from a to b. The choice of $K_{b^{(l)}}$ is not unique because an arbitrary unitary transformation $K_{b^{(l)}} \to W_{b^{(l)}} K_{b^{(l)}}$ leaves the measurement statistics invariant. In contrast to the construction of Kraus operators for the hybrid scheme, $K_{b^{(l-1)}}$ will not be invertible if the number of remaining nonzero POVM elements is smaller than the system's dimension. Therefore, $K_{b^{(l-1)}}^+$ denotes the Moore-Penrose pseudoinverse of $K_{b^{(l-1)}}$, and $Q_{b^{(l-1)}}$ ensures that the pair of binary Kraus operators generates a complete POVM.

Overall, the construction of corresponding binary Kraus operators in Eq. (H2) follows a sequential process:

(1) First, we form $B_{b^{(l)}} = \sum_{i=a}^{b} F_i$ by aggregating the POVM elements located in the leaves of the branch that starts from $b^{(l)}$. This results in $B_{b^{(l)}} = V_{b^{(l)}} D_{b^{(l)}}^2 V_{b^{(l)}}^\dagger$.

(2) Next, we create $K_{b^{(l)}} = V_{b^{(l)}} D_{b^{(l)}} V_{b^{(l)}}^\dagger$ that satisfies the condition $K_{b^{(l)}}^\dagger K_{b^{(l)}} = B_{b^{(l)}}$.

(3) We compute $K_{b^{(l-1)}}^+$, which represents the Moore-Penrose pseudoinverse of $K_{b^{(l-1)}}$ which was obtained in the previous level.

(4) Then, we construct the support projection matrix of $D_{b^{(l-1)}}$, denoted as $P_{b^{(l-1)}}$, where $(P_{b^{(l-1)}})_{ij} = \text{sign}[(D_{b^{(l-1)}})_{ij}]$.

(5) Lastly, $Q_{b^{(l-1)}} = I - V_{b^{(l-1)}} P_{b^{(l-1)}} V_{b^{(l-1)}}^\dagger$ ensures that the pair of binary Kraus operators generates a valid POVM: $\sum_{\{0,1\}} A_{b^{(l)}}^\dagger A_{b^{(l)}} = I$.

This construction ensures that the matrix 2 is unitary and that the cumulative Kraus operator

$$A_{b^{(L)}}^c = A_{b^{(L)}} \cdots A_{b^{(2)}} A_{b^{(1)}} \tag{H3}$$

corresponds to the correct POVM element: $A_{b^{(L)}}^{c\dagger} A_{b^{(L)}}^c = F_i$ with $i = b^{(L)} + 1$, as shown in the Appendix to the paper by Shen *et al.* [24].

Each time, the coupling unitary applied will depend on the sequence of previous measurement outcomes that define the current branch. Therefore, the scheme requires midcircuit measurements and feed forward. It is worth
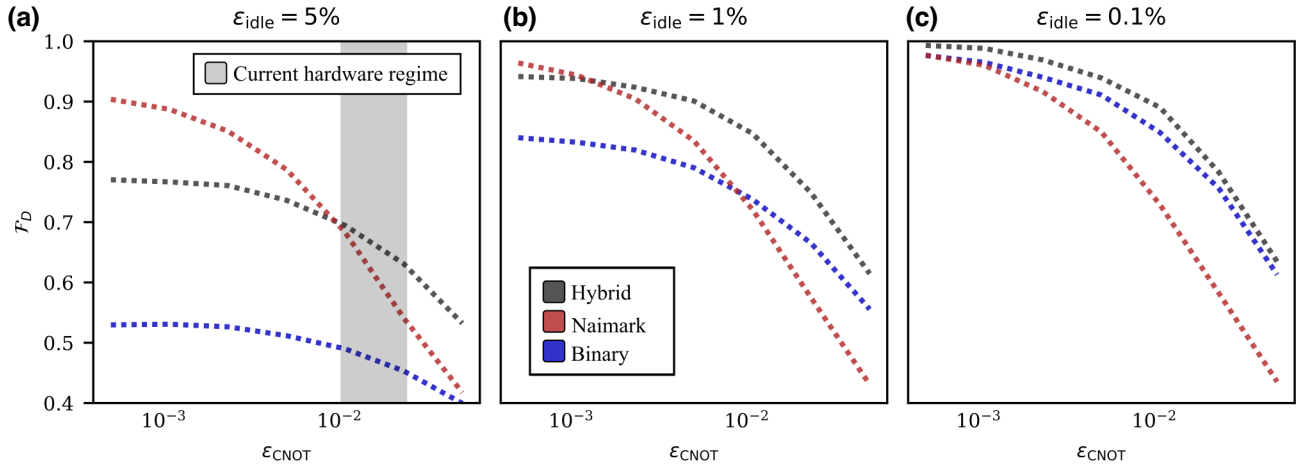
FIG. 10.   Noisy simulation of the two-qubit POVM for the binary tree (blue), Naimark's dilation (red), and hybrid (gray) schemes under various noise regimes. For each scheme, and for each combination of ϵ_{idle} and ϵ_{CNOT}, the fidelity is computed for 10 circuits at varying CNOT depths and the highest fidelity value is plotted. The gray-shaded area indicates the regime where the simulations best agree with the experimental results obtained on hardware.

noting that the system is never measured directly during the binary search. This fact makes this approach also applicable to quantum channels, as shown by Shen *et al.* [24].

In contrast to Naimark's dilation, which involves a single M-dimensional unitary applied to a set of $\log_2 M$ qubits, the binary scheme requires a total of $\log_2 M$ unitary operations, each acting on $n+1$ qubits.

APPENDIX I: EXTENDED NOISE ANALYSIS

In Sec. IV D of the main text, we employ a simple phenomenological noise model to explain the experimental results of the two-qubit experiment. The two hyperparameters of this model, ϵ_{CNOT} and ϵ_{idle}, represent the depolarizing errors associated with each CNOT gate and each measurement and feed-forward operation, respectively. While this model does not account for complex noise processes such as correlated errors, leakage, T_1 fluctuations, or measurement-induced control errors, which may occur on real superconducting qubit platforms [44], the close agreement with the experimental results in Fig. 3 suggests that it effectively captures the main error contributions from noisy CNOT gates and noise related to midcircuit measurements and idle time.

Here, we apply the same model to qualitatively understand the different noise regimes affecting the performance of our hybrid approach. Figure 10 displays noisy simulations of the two-qubit SIC POVM across various error regimes. For each combination of ϵ_{idle} and ϵ_{CNOT}, we perform simulations for all three schemes at different depths, selecting for each scheme the depth that yields the highest fidelity. These fidelity levels are depicted in Fig. 10.

Figure 10(a) represents the regime with comparatively high ϵ_{idle}, where the binary scheme is most affected due

to a large number of conditional operations and midcircuit measurements. In the range $\epsilon_{CNOT} \approx 1$–2%, we observe the best agreement with our experimental results. We thus conclude that the gray-shaded area in Fig. 10(a) corresponds to the current hardware regime. Figures 10(b) and 10(c) show lower ϵ_{idle} regimes, potentially achievable with faster midcircuit measurements and conditional operations or through effective error-suppression techniques such as dynamical decoupling [52]. In these conditions, we expect the hybrid scheme to offer a more significant advantage over Naimark. Importantly, in all scenarios, the fidelity of Naimark decays fastest with increasing ϵ_{CNOT} due to its higher CNOT depth.

Overall, while the fidelity improvements offered by the hybrid approach depend on noise from conditional operations and midcircuit measurements, the exponential growth of CNOT depth with the system size is more detrimental compared to the linear scaling of midcircuit measurements, as detailed in Sec. IV E. Thus, while it is possible that the hybrid scheme might not offer an advantage for smaller system sizes on some experimental platforms with, e.g., high-quality CNOTs and poor-quality conditional operations, we expect hybrid to outperform its constituents on larger systems. In particular, we demonstrate that the hybrid approach already provides an advantage in a two-qubit system for IBM quantum devices.

APPENDIX J: DATA ACQUISITION

We conducted both one- and two-qubit experiments on `ibmq_kolkata`, a 27-qubit quantum processor [43]. For the one-qubit experiment, exact compiling was utilized. For the two-qubit experiment, we applied approximate compiling across ten different CNOT depths for all three

methods, yielding different levels of approximation accuracy. Subsequently, for each circuit, we generated five twirled instances by inserting random Pauli gates before and after CNOT gates. For each twirled instance, we prepared 36 two-qubit Pauli basis states and took 4000 measurement samples for each state and instance, totaling 20 000 samples per circuit per basis state. The collected measurement statistics were then used to reconstruct the POVMs as outlined in Appendix B. From the reconstructed POVMs, we computed point estimates for the fidelities of each circuit. To estimate confidence intervals for the fidelity values, we used bootstrapping, resampling the measurement statistics to obtain a set of estimates from which we calculated the standard deviation. We executed 300 bootstrap instances for each circuit depth. The resulting confidence intervals, typically ranging from 0.1% to 0.2%, were too narrow to be visible on the plots in the main text. Overall, the experimental data was collected over the span of 2 h, well within the typical noise drift timescale. Nevertheless, to mitigate any potential biases from the slowly drifting noise environment, we randomized the order of all circuits before submission to the quantum backend.

---

[1] M. A. Nielsen and I. L. Chuang, *Quantum Computation and Quantum Information* (Cambridge University Press, Cambridge, 2012).

[2] J. A. Bergou, Discrimination of quantum states, J. Mod. Opt. **57**, 160 (2010).

[3] S. L. Braunstein and C. M. Caves, Statistical distance and the geometry of quantum states, Phys. Rev. Lett. **72**, 3439 (1994).

[4] N. Gisin and R. Thew, Quantum communication, Nat. Photonics **1**, 165 (2007).

[5] H. J. Briegel, D. E. Browne, W. Dür, R. Raussendorf, and M. V. den Nest, Measurement-based quantum computation, Nat. Phys. **5**, 19 (2009).

[6] G. Tóth and I. Apellaniz, Quantum metrology from a quantum information science perspective, J. Phys. A **47**, 424006 (2014).

[7] J. Bae and L.-C. Kwek, Quantum state discrimination and its applications, J. Phys. A: Math. Theor. **48**, 083001 (2015).

[8] A. J. Scott, Tight informationally complete quantum measurements, J. Phys. A **39**, 13507 (2006).

[9] J. M. Renes, R. Blume-Kohout, A. J. Scott, and C. M. Caves, Symmetric informationally complete quantum measurements, J. Math. Phys. **45**, 2171 (2004).

[10] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PRXQuantum.5.030315 for Python implementation of POVMs on multi-qubit systems and data files and documentation to reproduce the results.

[11] J. Haah, A. W. Harrow, Z. Ji, X. Wu, and N. Yu, in *Proceedings of the Forty-Eighth Annual ACM Symposium on Theory of Computing*, STOC '16 (Association for Computing Machinery, New York, NY, USA, 2016), p. 913.

[12] J. Shang, A. Asadian, H. Zhu, and O. Gühne, Enhanced entanglement criterion via symmetric informationally complete measurements, Phys. Rev. A **98**, 022309 (2018).

[13] T. Vértesi and E. Bene, Two-qubit Bell inequality for which positive operator-valued measurements are relevant, Phys. Rev. A **82**, 062115 (2010).

[14] W. J. Yun, H. Baek, and J. Kim, Projection valued measure-based quantum machine learning for multi-class classification, ArXiv:2210.16731.

[15] G. García-Pérez, M. A. Rossi, B. Sokolov, F. Tacchino, P. K. Barkoutsos, G. Mazzola, I. Tavernelli, and S. Maniscalco, Learning to measure: Adaptive informationally complete generalized measurements for quantum algorithms, PRX Quantum **2**, 040342 (2021).

[16] There also exist scalable probabilistic methods to implement general POVMs with a finite success probability [53,54].

[17] M. A. Naimark, A. I. Loginov, and V. S. Shul'man, Non-self-adjoint operator algebras in Hilbert space, J. Sov. Math. **5**, 250 (1976).

[18] A. Peres, Neumark's theorem and quantum inseparability, Found. Phys. **20**, 1441 (1990).

[19] L. E. Fischer, D. Miller, F. Tacchino, P. K. Barkoutsos, D. J. Egger, and I. Tavernelli, Ancilla-free implementation of generalized measurements for qubits embedded in a qudit space, Phys. Rev. Res. **4**, 033027 (2022).

[20] R. Stricker, M. Meth, L. Postler, C. Edmunds, C. Ferrie, R. Blatt, P. Schindler, T. Monz, R. Kueng, and M. Ringbauer, Experimental single-setting quantum state tomography, PRX Quantum **3**, 040310 (2022).

[21] Y. S. Yordanov and C. H. W. Barnes, Implementation of a general single-qubit positive operator-valued measure on a circuit-based quantum computer, Phys. Rev. A **100**, 062317 (2019).

[22] D. F. Pinto, M. S. Zanetti, M. L. W. Basso, and J. Maziero, Simulation of positive operator-valued measures and quantum instruments via quantum state-preparation algorithms, Phys. Rev. A **107**, 022411 (2023).

[23] E. Andersson and D. K. L. Oi, Binary search trees for generalized measurements, Phys. Rev. A **77**, 052104 (2008).

[24] C. Shen, K. Noh, V. V. Albert, S. Krastanov, M. H. Devoret, R. J. Schoelkopf, S. M. Girvin, and L. Jiang, Quantum channel construction with circuit quantum electrodynamics, Phys. Rev. B **95**, 134501 (2017).

[25] A. D. Córcoles, M. Takita, K. Inoue, S. Lekuch, Z. K. Minev, J. M. Chow, and J. M. Gambetta, Exploiting dynamic quantum circuits in a quantum algorithm with superconducting qubits, Phys. Rev. Lett. **127**, 100501 (2021).

[26] E. Bäumer, V. Tripathi, D. S. Wang, P. Rall, E. H. Chen, S. Majumder, A. Seif, and Z. K. Minev, Efficient long-range entanglement using dynamic circuits, ArXiv:2308.13065.

[27] W. Cai, J. Han, L. Hu, Y. Ma, X. Mu, W. Wang, Y. Xu, Z. Hua, H. Wang, Y. P. Song, J.-N. Zhang, C.-L. Zou, and L. Sun, High-efficiency arbitrary quantum operation on a high-dimensional quantum system, Phys. Rev. Lett. **127**, 090504 (2021).

[28] L. Madden and A. Simonetto, Best approximate quantum compiling problems, ACM Trans. Quantum Comput. **3**, 1 (2022).

[29] E. Haapasalo, T. Heinosaari, and J.-P. Pellonpää, Quantum measurements on finite dimensional systems: Relabeling and mixing, Quantum Inf. Process. **11**, 1751 (2011).

[30] H. Sosa-Martinez, N. K. Lysne, C. H. Baldwin, A. Kalev, I. H. Deutsch, and P. S. Jessen, Experimental study of optimal measurements for quantum state tomography, Phys. Rev. Lett. **119**, 150401 (2017).

[31] N. Bent, H. Qassim, A. A. Tahir, D. Sych, G. Leuchs, L. L. Sánchez-Soto, E. Karimi, and R. W. Boyd, Experimental realization of quantum tomography of photonic qudits via symmetric informationally complete positive operator-valued measures, Phys. Rev. X **5**, 041006 (2015).

[32] A. Tavakoli, I. Bengtsson, N. Gisin, and J. M. Renes, Compounds of symmetric informationally complete measurements and their application in quantum key distribution, Phys. Rev. Res. **2**, 043122 (2020).

[33] V. N. Ivanova-Rohling, N. Rohling, and G. Burkard, Optimal quantum state tomography with noisy gates, EPJ Quantum Technol. **10**, 1 (2023).

[34] M. M. Wilde, *Quantum Information Theory* (Cambridge University Press, Cambridge, 2017), 2nd ed.

[35] J. Z. Blumoff, K. Chou, C. Shen, M. Reagor, C. Axline, R. T. Brierley, M. P. Silveri, C. Wang, B. Vlastakis, S. E. Nigg, L. Frunzio, M. H. Devoret, L. Jiang, S. M. Girvin, and R. J. Schoelkopf, Implementing and characterizing precise multiqubit measurements, Phys. Rev. X **6**, 031041 (2016).

[36] E. Knill, Fault-tolerant postselected quantum computation: Schemes, ArXiv:quant-ph/0402171.

[37] O. Kern, G. Alber, and D. L. Shepelyansky, Quantum error correction of coherent errors by randomization, Eur. Phys. J. D-At., Mol., Opt. Plasma Phys. **32**, 153 (2005).

[38] C. H. Bennett, G. Brassard, S. Popescu, B. Schumacher, J. A. Smolin, and W. K. Wootters, Purification of noisy entanglement and faithful teleportation via noisy channels, Phys. Rev. Lett. **76**, 722 (1996).

[39] M. R. Geller and Z. Zhou, Efficient error models for fault-tolerant architectures and the pauli twirling approximation, Phys. Rev. A **88**, 012314 (2013).

[40] J. J. Wallman and J. Emerson, Noise tailoring for scalable quantum computation via randomized compiling, Phys. Rev. A **94**, 052325 (2016).

[41] P. D. Nation, H. Kang, N. Sundaresan, and J. M. Gambetta, Scalable mitigation of measurement errors on quantum computers, PRX Quantum **2**, 040326 (2021).

[42] E. van den Berg, Z. K. Minev, and K. Temme, Model-free readout-error mitigation for quantum expectation values, Phys. Rev. A **105**, 032620 (2022).

[43] IBM Quantum Platform: Compute resources, https://quantum-computing.ibm.com/ (accessed: 2023-09-30).

[44] L. C. G. Govia, P. Jurcevic, C. J. Wood, N. Kanazawa, S. T. Merkel, and D. C. McKay, A randomized benchmarking suite for mid-circuit measurements, New J. Phys. **25**, 123016 (2023).

[45] E. H. Chen, G.-Y. Zhu, R. Verresen, A. Seif, E. Bäumer, D. Layden, N. Tantivasadakarn, G. Zhu, S. Sheldon, A. Vishwanath, S. Trebst, and A. Kandala, Realizing the nishimori transition across the error threshold for constant-depth quantum circuits, ArXiv:2309.02863.

[46] R. Iten, R. Colbeck, I. Kukuljan, J. Home, and M. Christandl, Quantum circuits for isometries, Phys. Rev. A **93**, 032318 (2016).

[47] A. Hashim, A. Carignan-Dugas, L. Chen, C. Juenger, N. Fruitwala, Y. Xu, G. Huang, J. Wallman, and I. Siddiqi, Quasi-probabilistic readout correction of mid-circuit measurements for adaptive feedback via measurement randomized compiling, ArXiv:2312.14139.

[48] J. A. Smolin, J. M. Gambetta, and G. Smith, Efficient method for computing the maximum-likelihood quantum state from measurements with additive gaussian noise, Phys. Rev. Lett. **108**, 070502 (2012).

[49] M. Guţă, J. Kahn, R. Kueng, and J. A. Tropp, Fast state tomography with optimal error bounds, J. Phys. A: Math. Theor. **53**, 204001 (2020).

[50] P.-X. Chen, J. A. Bergou, S.-Y. Zhu, and G.-C. Guo, Ancilla dimensions needed to carry out positive-operator-valued measurement, Phys. Rev. A **76**, 060303 (2007).

[51] J. Preskill, Lecture notes for Ph219/CS219 quantum computation, http://theory.caltech.edu/ preskill/ph229/ (accessed: 2023-05-20).

[52] E. Bäumer, V. Tripathi, A. Seif, D. Lidar, and D. S. Wang, Quantum Fourier transform using dynamic circuits, ArXiv:2403.09514.

[53] T. Singal, F. B. Maciejewski, and M. Oszmaniec, Implementation of quantum measurements using classical resources and only a single ancillary qubit, npj Quantum Inf. **8**, 1 (2022).

[54] M. Oszmaniec, L. Guerini, P. Wittek, and A. Acín, Simulating positive-operator-valued measures with projective measurements, Phys. Rev. Lett. **119**, 190501 (2017).