QuSSET: Shot Set Encoder for Quantum Property Estimation of Variable-Length States

Anonymous authors
Paper under double-blind review

Abstract

Estimating many observables from a small measurement record is difficult when each new quantum state can be sampled only a few times. We ask whether measurements from one variational-circuit family can support property estimation across both observables and system lengths. QuSSET is a noncausal, variable-length shot-set encoder that preserves spatial order within each shot while aggregating repeated shots as an exchangeable set. Masked-outcome pretraining learns an ansatz-specific prior from unlabeled random-Pauli records without property labels or circuit parameters. Separate supervised heads then map the same property-agnostic records to either a symmetric isotropic spin-correlation matrix or an antisymmetric Dzyaloshinskii–Moriya (DM) correlation matrix. Pretraining and supervision use only lengths $L \leq 16$; the resulting estimators are evaluated without further adaptation on new states at $L = 18, 20$, using 64 shots per state. On a depth-4 long-range circuit family, QuSSET outperforms learned and calibrated statistical estimators using the same records, as well as deterministic estimators granted separately acquired, target-specific measurements, across both observables. Crucially, this advantage persists at the unseen lengths $L = 18, 20$: without length-specific adaptation, each estimator processes longer shot records and predicts correlation matrices larger than any encountered during training. Matched controls show that masked pretraining provides its largest benefit with few labels and a persistent gain for long-range correlations. Together, these results show that one ansatz-level measurement corpus and pretrained initialization can support distinct property estimators and extrapolate to chain lengths absent from training.

1 Introduction

Quantum processors are pursued in part because representing and evolving quantum systems can become intractable on classical hardware (Feynman, 1982). An L-qubit pure state is described by 2^L complex amplitudes, so storing and evolving a generic state vector requires resources exponential in L. Classical structure remains powerful in important regimes. Tensor networks efficiently treat many low-entanglement one-dimensional systems (Vidal, 2003; Schollwöck, 2011), and neural quantum states provide flexible trainable approximations (Carleo & Troyer, 2017). These representations are not uniformly compact or tractable for arbitrary quantum circuit outputs. The states that motivate quantum hardware may therefore include states for which no faithful classical description is readily available.

Even when a quantum device can prepare such a state, it does not expose the corresponding amplitudes. One measurement repetition, or *shot*, returns a stochastic classical outcome in a chosen basis. Yet applications usually seek a property of the underlying state rather than the samples themselves. Energy expectations provide objectives for variational quantum algorithms (Peruzzo et al., 2014), while correlations reveal spatial and collective structure. These quantities are expectation values that must be inferred from a finite measurement record. We use *quantum property estimation* (QPE) for this inference problem. Access to a useful quantum state is thus also a statistical problem.

Full quantum tomography reconstructs a classical state description, but often solves a broader problem than is needed for selected observables (D'Ariano et al., 2003). Neural tomography replaces an explicit density matrix with a learned outcome distribution, but still models an individual state generatively (Torlai et al., 2018). Classical shadows instead use randomized measurements to pre-

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

dict many properties that may be selected after acquisition (Huang et al., 2020). This broad scope is valuable, although the standard inverse-measurement estimator does not exploit which states are likely to occur (Elben et al., 2023). In many experiments, states are produced by varying the parameters of one fixed circuit template, called an *ansatz*. They form a structured distribution rather than the unrestricted state space. Learning that distribution offers a route to a more informative prior when measurements of each new state are scarce.

Machine learning can infer such an implicit prior from data. Generative models reconstruct quantum measurement statistics from samples (Carrasquilla et al., 2019), and learned estimators can amortize property prediction across related systems (Huang et al., 2022). LLM4QPE brings task-agnostic measurement pretraining to QPE and adapts the resulting model using limited property labels (Tang et al., 2024). QuaDiM develops the same broad direction with nonautoregressive diffusion modeling (Tang et al., 2025). These works establish that an unlabeled measurement corpus can encode useful physical regularities. We build on this insight and ask a controlled question. Can the raw corpus and pretrained initialization from one ansatz be held fixed while the requested observable changes, and can the resulting estimator operate at system lengths absent from all training?

Our goal is to amortize training-data acquisition at the ansatz level. Masked-outcome pretraining first fits a shared encoder to unlabeled local random-Pauli records. Each downstream property then copies that encoder and jointly finetunes it with a task-specific head on a small labeled subset. The states and measurement outcomes are reused across properties, while the labels and resulting finetuned weights differ. Exact state vectors at tractable sizes can provide several such labels without recollecting the measurement inputs. We restrict property supervision to a small labeled subset to evaluate the contribution of unlabeled pretraining. In experimental settings, obtaining reference labels can require additional high-precision measurements or classical simulation.

The second part concerns reuse across system size. The preceding label assumption becomes restrictive as the system grows. State-vector simulation becomes exponentially costly, while obtaining reference labels of comparable precision on hardware would itself consume additional measurements. A useful estimator should therefore learn where labels remain accessible and operate at longer lengths without length-specific finetuning. We study this extrapolation within one ansatz distribution, targeting the regime in which classical label generation eventually becomes the bottleneck.

We introduce QuSSET (Quantum Shot-Set Encoder for Transfer), which implements this two-stage procedure and is illustrated in Figure 1. For each target state, it maps a set of $K = 64$ local random-Pauli shots, including every basis and outcome, to the requested property. This record has two different structural axes. Shot order is physically irrelevant, motivating permutation-equivariant set aggregation (Zaheer et al., 2017), whereas site order encodes chain geometry, motivating bidirectional spatial attention over each fully observed shot (Vaswani et al., 2017). Relative positional encoding and shared weights let the same spatial blocks operate across chain lengths (Su et al., 2024). A Perceiver bottleneck then summarizes the variable-size record (Jaegle et al., 2021), and shared pairwise queries produce a matrix whose size follows the input length (Jaegle et al., 2022). We pretrain this backbone by masked-outcome prediction, which requires neither exact properties nor circuit parameters and follows masked representation learning (Devlin et al., 2019).

We evaluate two physically distinct pairwise targets on a depth-4 circuit whose power-law couplings and noncommuting rotations distribute parameter dependence across the chain. Isotropic spin correlations combine same-axis Pauli products, whereas Dzyaloshinskii–Moriya (DM) correlations are antisymmetric and use cross-axis products. Both tasks use a common measurement corpus and pretrained initialization. Pretraining and supervision use $L \leq 16$, with $L = 18, 20$ held out for length-OOD evaluation. QuSSET outperforms learned and statistical estimators using the same records and deterministic estimators granted target-specific measurements.

Our contributions are the following.

- **QuSSET for finite-shot property estimation.** Our noncausal, variable-length encoder respects site order and shot exchangeability, producing dense pairwise matrices for new states and chain lengths.
- **Length extrapolation beyond the training range.** Models pretrained and supervised only at $L \leq 16$ outperform the tested baselines on new states at $L = 18, 20$ without adaptation, even though these inputs contain more sites and require larger output matrices than

any training example. Shared spatial blocks and pairwise queries enable this extrapolation without length-specific retraining, offering a path toward sizes at which exact classical labels are no longer readily available.

- **Reusable ansatz-level measurement corpus.** Both tasks share the unlabeled pretraining corpus and pretrained initialization. Their supervised datasets also use identical states and 64-shot measurement records, with task-specific labels. With limited labels for each target, QuSSET improves on estimators using those records and on label-free deterministic estimators granted separate target-specific measurements.
- **Controlled attribution.** Matched controls quantify the effects of masked pretraining and access to circuit variables.

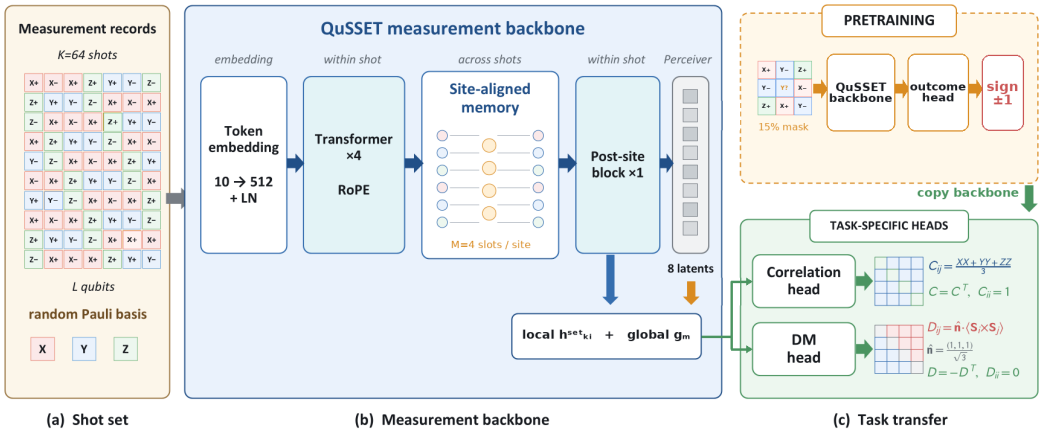


Figure 1: Overview of QuSSET. The backbone combines spatial attention within each shot, site-aligned aggregation across shots, and a Perceiver summary. It learns by reconstructing masked outcome signs, then initializes separate heads for (C) and (D). States, records, splits, and initialization are shared; labels and final weights are task-specific.

2 Background and Problem Definition

2.1 From quantum states to classical measurement records

A pure state of L qubits is a normalized vector $|\psi\rangle = \sum_{z\in\{0,1\}^L} a_z|z\rangle$, with 2^L complex amplitudes; a possibly mixed state is represented by a density matrix ρ. This exponential description is not returned by a quantum device. Instead, one prepares the state, chooses a measurement basis, and obtains one classical outcome. The measured copy is consumed, so another outcome requires another preparation. We call one such preparation-and-measurement event a *shot*.

The Pauli operators X, Y, Z are the standard single-qubit observables. Measuring Pauli b_i on qubit i returns a sign $s_i \in \{-1, +1\}$. Measuring all L qubits produces a basis string $\mathbf{b}$ and an outcome string $\mathbf{s}$, analogous to one row of a classical dataset but with strong correlations between columns. Repeating this procedure gives only samples from the state, not its amplitudes. Consequently, even a deterministic quantum state induces statistical uncertainty when the number of shots is finite.

For an L-qubit state ρ, a local Pauli shot consists of bases $\mathbf{b} = (b_1, \ldots, b_L)$, $b_i \in \{X, Y, Z\}$, and signs $\mathbf{s} \in \{-1, +1\}^L$, sampled according to

$$p(\mathbf{s} \mid \mathbf{b}, \rho) = \mathrm{Tr}\left[\rho \bigotimes_{i=1}^{L} \frac{I + s_i b_i}{2}\right]. \tag{1}$$

This is the Born rule: it specifies the probability of each classical string but does not make that probability directly observable. An expectation value such as $\langle Z_i Z_j \rangle_\rho = \mathrm{Tr}(\rho Z_i Z_j)$ must be inferred from repeated outcomes.

2.2 QUANTUM PROPERTY ESTIMATION

Throughout this paper, QPE means *quantum property estimation*, not the quantum phase-estimation algorithm. Given a finite record $\mathcal{D}_\rho = \{(\mathbf{b}_k, \mathbf{s}_k)\}_{k=1}^K$, QPE predicts a classical property of the underlying state without reconstructing all 2^L amplitudes. Here $K = 64$, and every basis is chosen uniformly and independently across sites and shots. Our first property is the symmetric normalized spin-correlation matrix

$$C_{ii} = 1, \qquad C_{ij} = \frac{4}{3}\langle \mathbf{S}_i \cdot \mathbf{S}_j \rangle_\rho = \frac{\langle X_i X_j \rangle + \langle Y_i Y_j \rangle + \langle Z_i Z_j \rangle}{3}, \quad i < j. \tag{2}$$

where $\mathbf{S}_i = (X_i, Y_i, Z_i)/2$ is the spin-$1/2$ operator. Thus C_{ij} averages the tendency of two sites to align along the three spatial axes; positive and negative values correspond to aligned and anti-aligned correlations, respectively. We learn the $L(L-1)/2$ nontrivial entries; the diagonal is fixed by definition. The output dimension grows quadratically with L, which makes a fixed-size regression head unsuitable when the chain length changes.

Our second property is the component of vector spin chirality along $\hat{\mathbf{n}} = (1,1,1)/\sqrt{3}$,

$$\begin{aligned} D_{ij} &= \hat{\mathbf{n}} \cdot \langle \mathbf{S}_i \times \mathbf{S}_j \rangle_\rho \\ &= \frac{1}{4\sqrt{3}} \langle Y_i Z_j - Z_i Y_j + Z_i X_j - X_i Z_j + X_i Y_j - Y_i X_j \rangle_\rho, \qquad i < j, \\ D_{ji} &= -D_{ij}, \qquad D_{ii} = 0. \end{aligned} \tag{3}$$

An interaction $\mathbf{D}_{ij} \cdot (\mathbf{S}_i \times \mathbf{S}_j)$ is the antisymmetric exchange term introduced phenomenologically by Dzyaloshinsky and derived microscopically by Moriya from spin–orbit-coupled superexchange (Dzyaloshinsky, 1958; Moriya, 1960). Here we do not add this interaction to the state-generating circuit. Instead, Equation 3 is a distinct observable of the same states: unlike C_{ij}, it changes sign under $i \leftrightarrow j$ and depends only on cross-axis Pauli products. It therefore tests whether one measurement corpus can support a qualitatively different pairwise query.

For either target $T \in \{C, D\}$, the supervised loss over the independent upper triangle is

$$\mathcal{L}_{\text{term}}^{(T)} = \frac{2}{L(L-1)} \sum_{i<j} \text{SmoothL1}_{\beta=0.05}(\widehat{T}_{ij} - T_{ij}). \tag{4}$$

Thus $\mathcal{L}_{\text{term}}$ refers to observable-entry regression, not Hamiltonian energy. One supervised example is an entire physical state: its 64-shot record is the input and its exact target matrix is the label. Accordingly, N_t counts *labeled states*, not shots or individual matrix entries. Unlabeled pretraining states provide measurement records but neither target. Length OOD asks the same estimator to process more qubits and produce a larger matrix than it encountered during either pretraining or supervised training.

3 RELATED WORK

Randomized measurements. Shadow tomography and classical shadows formalize predicting many observables without full state reconstruction (Aaronson, 2018; Huang et al., 2020; Elben et al., 2023). When targets are known, measurement grouping can reduce variance and VQE measurement overhead (Verteletskyi et al., 2020). Our grouped-XYZ baseline measures every site along a common axis for C, while the balanced cross-basis schedule covers the six ordered axis pairs needed for D. Both are unbiased, target-specific alternatives to the reusable local random-Pauli corpus. Robust shadow variants extend randomized estimation to device noise (Chen et al., 2021; Koh & Grewal, 2022).

Learning quantum-state priors. Neural quantum states and neural tomography learn descriptions of individual states or families (Carleo & Troyer, 2017; Torlai et al., 2018; Carrasquilla et al., 2019). Learned property estimators instead amortize mappings from measurements or physical variables to observables. LLM4QPE pretrains a condition-aware autoregressive model (Tang et al., 2024); its correlation benchmark fixes the Hamiltonian scales, varies one power-law exponent, and trains separate models for $L \in \{8, 10, 12\}$. Tang et al. report improved correlation estimates when

condition embeddings are included. We study a more broadly parameterized regime: each circuit state varies $8L + 5$ continuous parameters, long-range correlations depend on variables distributed across the chain, and one estimator must also extrapolate to unseen lengths. For isotropic correlations, exact-variable-only controls fail to improve on zero prediction under the tested architectures and training budgets (Appendix F). Adding exact variables to QuSSET likewise yields no consistent improvement across targets and label budgets, highlighting that the utility of conditioning depends on the estimator and evaluation setting. Our LLM4QPE-like baseline retains causal measurement modeling and circuit-variable conditioning in this setting, whereas canonical QuSSET treats shots as an unordered set and uses measurements alone. QuaDiM instead uses conditional diffusion (Tang et al., 2025). Bayesian shadow tomography uses a Set Transformer to correct classical-shadow estimates of scalar fidelity and Rényi purity (Cha et al., 2025); QuSSET retains the spatial site axis and produces a length-dependent pairwise output. AiDE-Q selects synthetic labels from hybrid-quality data (Wang et al., 2025), a data-engine strategy complementary to our unlabeled representation learning.

Masked set representations. Masked prediction provides transferable representations without task labels (Devlin et al., 2019). Attention can amortize inference over sampled tasks (Müller et al., 2022), and latent bottlenecks support structured inputs and outputs (Jaegle et al., 2021; 2022). Quantum records add a two-axis symmetry: shots are exchangeable, whereas qubit position is meaningful. Our architecture preserves this distinction rather than flattening the record into an arbitrary sequence.

4 METHOD

4.1 ANSATZ DISTRIBUTION AND DATA

An *ansatz* is a parameterized circuit template: fixing its gates and varying its continuous angles produces a family of states. This is the natural repeated-query unit in variational quantum algorithms. We keep the layer structure and parameter distribution fixed while changing both the angles (a new state) and the number of qubits (a new chain length).

Each state begins from the product state with Bloch vector $(1, 1, 1)/\sqrt{3}$. Four layers apply long-range ZZ evolution followed by site-dependent X and Z rotations,

$$U_r = \exp\left(-i \sum_i \zeta_{r,i} Z_i\right) \prod_i \exp(-i\beta_{r,i} X_i) \exp\left(-i\gamma_r \sum_{i<j} |i-j|^{-\alpha_c} Z_i Z_j\right). \tag{5}$$

We use $e^{-i\theta P}$, and sample $\alpha_c \sim U(0.5, 2)$, $\gamma_r \sim U(-0.5, 0.5)$, and $\beta_{r,i}, \zeta_{r,i} \sim U(-\pi/2, \pi/2)$. At depth four, this gives $8L + 5$ independently sampled continuous variables, including $8L$ site-dependent rotation angles. The interaction in every layer acts on all qubit pairs, so shallow depth does not impose a nearest-neighbor light cone; the intervening X rotations also do not commute with the ZZ evolution. Their combined action maps an $O(L)$-dimensional circuit specification to $O(L^2)$ state- and pair-dependent targets, with long-distance correlations controlled by variables distributed across the chain. We choose this broad family to avoid a low-dimensional or local parameter-to-property shortcut. It is neither a ground-state ensemble nor one optimization trajectory. Measurements and exact labels are generated from the state vector, and independent splits are checked for duplicate parameter fingerprints.

State-vector simulation lets us sample the same measurement strings that an ideal device would return and compute exact labels for evaluation. Crucially, the network never receives the state vector or Born probabilities. Its downstream input is always the finite 64-shot record. The C and D tasks use identical states, random-Pauli bases, outcomes, splits, and labeled-state subsets; we derive only a new exact target matrix from the simulator. Exact targets are exposed only for the selected supervised states, while the larger pretraining split contains measurement records alone. Table 1 separates these resources explicitly.

4.2 SHOT-SET ENCODER

Figure 1 follows the complete data flow. For a batch of B states, the token array has shape $B \times K \times L$; below we suppress the leading batch axis. A shared lookup table and LayerNorm embed each basis–

Table 1: Data shared by both tasks. Each supervised fit uses a nested subset $N_t \in \{20, 50, 90, 180, 405\}$ of the 600-state pool. Pretraining has no property labels, and 64 shots denotes the cost per downstream state rather than per task.

Split	Available states	Shots/state	Lengths	Labels
Unlabeled pretraining	1,200	1,024	6,8,10,12,14,16	no
Supervised pool (N_t used/fit)	600	64	6,8,10,12,14,16	yes
ID evaluation	600	64	6,8,10,12,14,16	yes
Length-OOD evaluation	400	64	18,20	yes

sign token into $h^{(0)} \in \mathbb{R}^{K\times L\times H}$, with $H = 512$. QuSSET preserves qubit order but treats the K repeated shots as an exchangeable set. Its three stages can be summarized as

$$\begin{aligned} h^{\mathrm{sp}} &= \mathrm{Spatial}_{1:4}(h^{(0)}) \in \mathbb{R}^{K\times L\times H}, \\ (M_i, \overline{H}_i) &= \mathrm{SiteAggregate}(h^{\mathrm{sp}}_{:,i}) \in \mathbb{R}^{M\times H} \times \mathbb{R}^{K\times H}, \\ h^{\mathrm{set}} &= \mathrm{Spatial}_{5}(\mathrm{stack}_i\, \overline{H}_i) \in \mathbb{R}^{K\times L\times H}, \\ V &= [\mathrm{flat}_{k,i}(h^{\mathrm{set}}_{ki}); \mathrm{flat}_{i,a}(M_{i,a})] \in \mathbb{R}^{(KL+LM)\times H}, \\ g &= \mathrm{Perceiver}(g^{(0)}; V) \in \mathbb{R}^{P\times H}, \qquad M = 4, \quad P = 8. \end{aligned} \tag{6}$$

The spatial blocks are noncausal Transformers (Vaswani et al., 2017) with rotary positions only along the qubit axis (Su et al., 2024). Site-aligned learned queries compress and return evidence across shots without shot positional encodings. The Perceiver latents (Jaegle et al., 2021) summarize both the final local tokens and site memories. Downstream heads retain both h^{set} and g. Full residual updates, attention directions, and tensor reshaping are given in Appendix A.1.

Masked-measurement pretraining. For a uniformly sampled mask set $\mathcal{M}$ containing 15% of positions, the input token preserves the measured basis but hides its ± 1 sign. Concretely, X-, Y-, and Z-basis measurements use distinct mask tokens. Thus $\widetilde{x}$ has shape $K \times L$ and retains every basis choice. The backbone returns the tensors in Equation 6; the pretext head maps $F = \mathrm{flat}_{k,i}(h^{\mathrm{set}}_{ki}) \in \mathbb{R}^{KL\times H}$ and $g \in \mathbb{R}^{P\times H}$ to sign logits $\ell = H_{\theta_M}(F, g) \in \mathbb{R}^{KL\times 2}$. Only masked positions contribute to

$$\mathcal{L}_{\mathrm{mask}} = -\frac{1}{|\mathcal{M}|} \sum_{(k,i)\in\mathcal{M}} \log p_{\theta_B,\theta_M}(s_{ki} \mid \widetilde{x})\,. \tag{7}$$

It uses hard outcomes, not Born probabilities, property labels, or circuit parameters. The pretext head parameters θ_M are discarded after training; only the backbone parameters θ_B initialize supervised estimation.

Pairwise observable queries. Let $z_i = K^{-1}\sum_k h^{\mathrm{set}}_{ki} \in \mathbb{R}^H$ and $c_{ij} = K^{-1}\sum_k h^{\mathrm{set}}_{ki} \odot h^{\mathrm{set}}_{kj} \in \mathbb{R}^H$. For each of the $N = L(L-1)/2$ pairs, the symmetric and antisymmetric heads respectively form

$$\begin{aligned} q^C_{ij} &= W_C[z_i + z_j; c_{ij}] + u_C(\phi_C(L,i,j)) \in \mathbb{R}^H, \\ q^D_{ij} &= W_D[z_i; z_j; c_{ij}] + u_D(\phi_D(L,i,j)) \in \mathbb{R}^H, \\ Q_T &= (q^T_{ij})_{i<j} \in \mathbb{R}^{N\times H}, \\ Q'_T &= \mathrm{PairBlock}(Q_T + \mathrm{MHA}(Q_T; g)) \in \mathbb{R}^{N\times H}, \end{aligned} \tag{8}$$

$$\widehat{t}_T = a_T \tanh(f_T(Q'_T)) \in \mathbb{R}^N, \qquad a_C = 1, \quad a_D = \tfrac{1}{2}. \tag{9}$$

Here $\phi_T \in \mathbb{R}^{12}$ contains target-appropriate length and relative-geometry features. For C, the head symmetrizes the predicted upper triangle and fixes $C_{ii} = 1$. For D, an ordered pair projection completes the matrix by $D_{ji} = -D_{ij}$ and fixes $D_{ii} = 0$. The two heads have the same query–attention structure but separate parameters. Parameter sharing across all pairs makes the output size follow the input length. Canonical QuSSET receives measurements only. The optional conditioned ablation applies a typed gated residual from an encoder of the exact circuit variables at the query stage. It is statistically comparable but not consistently superior, so measurement-only QuSSET is

Table 2: QuSSET-family RMSE (lower is better). ID is $L \leq 16$, OOD is unseen $L = 18, 20$, and bold marks the best three-seed mean per target and column.

Target	Model	$N_t = 20$		$N_t = 50$		$N_t = 90$		$N_t = 180$		$N_t = 405$	
		ID	OOD	ID	OOD	ID	OOD	ID	OOD	ID	OOD
C	QuSSET	**0.0786**	**0.0708**	0.0769	0.0691	0.0762	0.0676	0.0756	0.0670	**0.0738**	**0.0655**
	+ circuit variables	0.0791	0.0713	**0.0767**	**0.0689**	**0.0757**	**0.0673**	**0.0754**	**0.0668**	0.0740	0.0656
	Scratch	0.1008	0.0928	0.0773	0.0699	0.0764	0.0686	0.0759	0.0678	0.0746	0.0664
D	QuSSET	0.0525	0.0476	**0.0480**	**0.0429**	0.0478	**0.0425**	0.0474	**0.0419**	**0.0462**	**0.0410**
	+ circuit variables	**0.0523**	**0.0474**	0.0485	0.0435	0.0478	**0.0425**	0.0477	0.0421	0.0463	**0.0410**
	Scratch	0.0632	0.0584	0.0518	0.0468	**0.0474**	0.0426	**0.0470**	0.0421	0.0463	0.0415

our primary model. Both it and scratch use the identical Equations 8–9; only backbone initialization differs.

5 Experiments

5.1 Protocol and comparisons

We use nested labeled subsets $N_t \in \{20, 50, 90, 180, 405\}$ and three seeds. All measurement-based learned methods receive the same 64 local random-Pauli shots. The architecture-matched *scratch* control omits pretraining, *circuit variables only* removes measurements, and *QuSSET + circuit variables* adds the optional variable encoder. *LLM4QPE-like* is a causal, circuit-conditioned Transformer adapted to our data rather than a reproduction of its original benchmark.

Simple estimators comprise unbiased classical shadows, distance-bin shrinkage, and cross-validated measurement-feature ridge. Ridge uses the raw matched-basis estimate, empirical Pauli-term means and counts, and pair geometry; its penalty is selected by state-grouped five-fold cross-validation. The two calibrated estimators are defined in Appendix A.3. Target-aware estimators instead receive separate 64-shot records: cyclic 22/21/21 global $X/Y/Z$ shots for C, and balanced site-wise assignments covering the six cross-axis terms for D. Because a learned estimator using target-aware bases would require a separate property-specific measurement corpus and thus fall outside our corpus-reuse setting, we include target-aware acquisition only as a label-free deterministic reference.

The primary metric pools unique off-diagonal entries before computing RMSE. Long distance means $|i - j| \geq \lceil L/2 \rceil$. Error bars show one sample standard deviation; analytic estimators are evaluated once. Training and batching details are in Appendix A.

At $N_t = 20$, we use a predeclared training schedule because a separate validation split would leave too few training states. The resulting scratch gap is a fixed-budget comparison (Appendix E); distance diagnostics use $N_t = 90$.

5.2 One measurement corpus, two observable families

Figure 2 gives the central reuse result. Its top and bottom rows differ in exact labels and task head while holding the corpus, split, labeled-state indices, and pretrained initialization fixed. Each row reports raw RMSE in its target's natural scale, so the comparison concerns rankings within a target rather than the absolute difficulty of C versus D. One estimator represents each external comparison class; Table 2 resolves the QuSSET family at every label budget.

Across both targets and both evaluation regimes, QuSSET outperforms the LLM4QPE-like model, calibrated shrinkage and ridge estimators, classical shadows, and deterministic estimators using separately acquired target-aware measurements. The latter comparison is especially relevant to corpus reuse: grouped-XYZ and balanced cross-basis measurements are tailored to one observable, whereas QuSSET uses the same general random-Pauli acquisition for C and D. ID and OOD magnitudes are not directly comparable because their target distributions differ, but the within-panel ranking is stable.

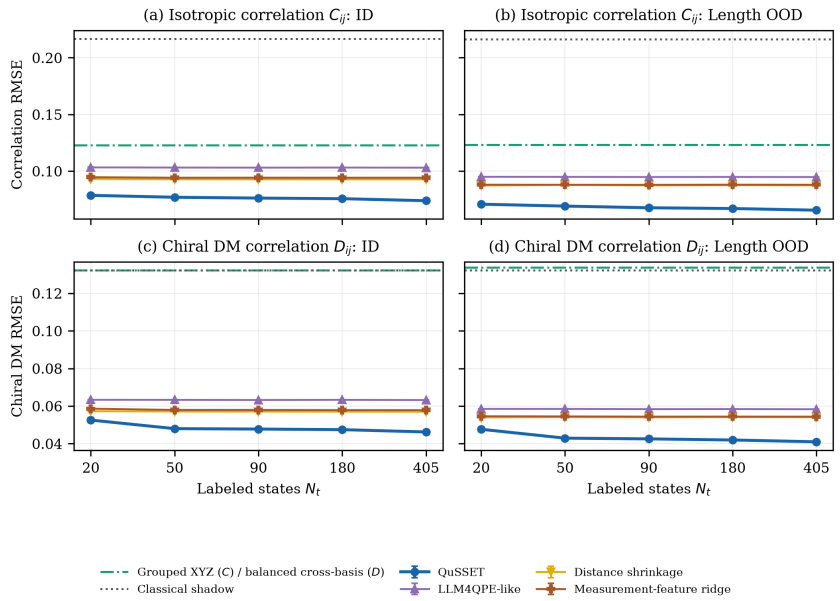


Figure 2: One random-Pauli corpus, two targets. Rows show C and D; columns show ID and length OOD. RMSE scales are shared only within each row. Learned curves show three-seed mean $\pm$ standard deviation; target-aware estimators use separately acquired grouped-XYZ (C) or balanced cross-basis (D) records.

The matched scratch control separates this overall advantage from the effect of initialization. Pretraining matters most when labels are scarce: at $N_t = 20$, it reduces ID/OOD RMSE by 22.0%/23.7% for C and 16.9%/18.4% for D. As the labeled set grows, scratch closes most of the gap, while the pretrained model retains a smaller advantage at every OOD budget for both targets. At larger label budgets, most of the separation from external baselines is retained by the scratch estimator; pretraining provides a larger benefit in the low-label, fixed-budget setting and a smaller residual gain thereafter. Adding exact circuit variables has no consistent effect across either property, split, or label budget, which motivates the measurement-only model as canonical. Figure 3 resolves the remaining pretraining effect by correlation range.

5.3 Pretraining preferentially improves long-range correlations

At $N_t = 20$, the top row of Figure 3 shows isotropic gains rising from about 18% at nearest neighbors to above 30% at maximum separation. After this low-label transient, positive gains remain concentrated at intermediate and long distances: pooled long-distance RMSE improves by 1.0–1.6% ID and 2.5–3.6% OOD at $N_t = 90, 405$.

At $N_t = 90$, both isotropic estimators beat the zero predictor and explain variation beyond the distance-bin mean (Appendix Figure 7); complementary ID results are in Appendix Figure 6.

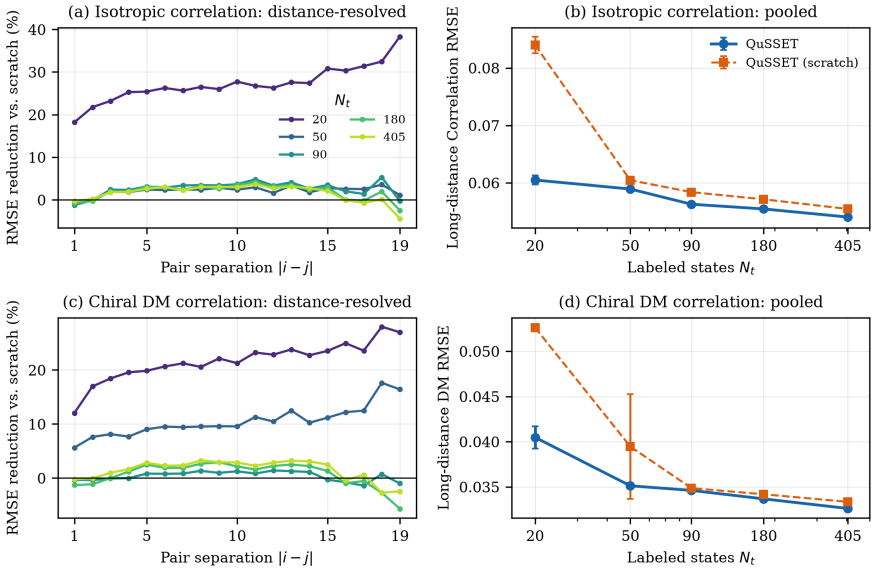


Figure 3: Masked-pretraining gains under length OOD for C and D. Left: distance-resolved RMSE reduction over scratch. Right: RMSE pooled over $|i-j| \geq \lceil L/2 \rceil$.

5.4 Chiral correlations retain a long-range OOD gain

DM combines six cross-axis rather than three same-axis correlators, yet the bottom row of Figure 3 shows the same aggregate OOD tendency. Pretraining lowers long-distance RMSE by 23.1%/11.0% at $N_t = 20/50$, and by 0.7%/1.5%/2.2% at $N_t = 90/180/405$. For both targets, the pooled trend is the primary statistic because individual maximum-distance bins have little support and fluctuate across label budgets.

6 Discussion

A demanding nonlocal estimation problem. The benchmark maps $8L+5$ independently sampled circuit variables to pair observables through noncommuting long-range dynamics. Adding exact variables to QuSSET yields no consistent improvement across the two targets. For isotropic correlations, the tested variable-only regressors remain at or below zero-predictor performance despite more labels, capacity, and training, while measurement-only QuSSET achieves positive held-out R^2 (Appendix F). This supports measurement-based estimation in this ensemble without implying an intrinsic limit of parameter-based prediction.

Extrapolation beyond the training range. One parameterization expands its output and retains its advantage at $L = 18, 20$, supporting extrapolation beyond the labeled training range.

One corpus, distinct properties. Separate heads adapted from one corpus succeed on symmetric C and antisymmetric D, and outperform baselines with favorable, separately acquired measurements.

What self-supervision adds. Scratch retains most of the external-baseline advantage. Pretraining helps most with few labels and retains a smaller long-range benefit under length OOD; Figure 3 resolves this remaining difference by pair separation.

Conclusion. Across two pairwise observables in a noiseless circuit family, one corpus supports 64-shot estimators trained at $L \leq 16$ and evaluated at $L = 18, 20$. Pretraining uses $1{,}200 \times 1{,}024$ unlabeled shots and is clearest at long range; each adaptation uses N_t exact labels.

Reproducibility Statement

The anonymous supplement contains generation, training, evaluation, and plotting code; fixed splits; configs; commands; and machine-readable results for every table and figure. Code and generated data will be released after double-blind review.

AI Use Statement

We used generative AI tools to assist with drafting and editing portions of the manuscript, preparing figures, and writing or revising research code. We also used them to discuss experimental design and the interpretation of results. The authors independently checked the resulting text, figures, code, analyses, and citations against the underlying methods and results. The authors take responsibility for all content and claims in this paper.

References

Scott Aaronson. Shadow tomography of quantum states. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pp. 325–338, 2018. doi: 10.1145/3188745.3188802.

Giuseppe Carleo and Matthias Troyer. Solving the quantum many-body problem with artificial neural networks. *Science*, 355(6325):602–606, 2017.

Juan Carrasquilla, Giacomo Torlai, Roger G Melko, and Leandro Aolita. Reconstructing quantum states with generative models. *Nature Machine Intelligence*, 1(3):155–161, 2019.

Hyunho Cha, Wonjung Kim, and Jungwoo Lee. Scalable bayesian shadow tomography for quantum property estimation with set transformers. *arXiv preprint arXiv:2509.18674*, 2025. doi: 10.48550/arXiv.2509.18674. URL https://arxiv.org/abs/2509.18674.

Senrui Chen, Wenjun Yu, Pei Zeng, and Steven T. Flammia. Robust shadow estimation. *PRX Quantum*, 2(3):030348, 2021. doi: 10.1103/PRXQuantum.2.030348.

G Mauro D'Ariano, Matteo GA Paris, and Massimiliano F Sacchi. Quantum tomography. In Peter W. Hawkes (ed.), *Advances in Imaging and Electron Physics*, volume 128, pp. 205–308. Elsevier, 2003. doi: 10.1016/S1076-5670(03)80065-4.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pp. 4171–4186, 2019.

I. Dzyaloshinsky. A thermodynamic theory of "weak" ferromagnetism of antiferromagnetics. *Journal of Physics and Chemistry of Solids*, 4(4):241–255, 1958. doi: 10.1016/0022-3697(58)90076-3.

Andreas Elben, Steven T. Flammia, Hsin-Yuan Huang, Richard Kueng, John Preskill, Benoît Vermersch, and Peter Zoller. The randomized measurement toolbox. *Nature Reviews Physics*, 5:9–24, 2023.

Richard P. Feynman. Simulating physics with computers. *International Journal of Theoretical Physics*, 21(6–7):467–488, 1982. doi: 10.1007/BF02650179.

Hsin-Yuan Huang, Richard Kueng, and John Preskill. Predicting many properties of a quantum system from very few measurements. *Nature Physics*, 16(10):1050–1057, 2020.

Hsin-Yuan Huang, Richard Kueng, Giacomo Torlai, Victor V. Albert, and John Preskill. Provably efficient machine learning for quantum many-body problems. *Science*, 377(6613):eabk3333, 2022. doi: 10.1126/science.abk3333.

Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and João Carreira. Perceiver: General perception with iterative attention. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4651–4664. PMLR, 2021. URL https://proceedings.mlr.press/v139/jaegle21a.html.

Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, Olivier Hénaff, Matthew M. Botvinick, Andrew Zisserman, Oriol Vinyals, and João Carreira. Perceiver IO: A general architecture for structured inputs & outputs. In *International Conference on Learning Representations (ICLR)*, 2022. URL https://openreview.net/forum?id=fILj7WpI-g.

Dax Enshan Koh and Sabee Grewal. Classical shadows with noise. *Quantum*, 6:776, 2022. doi: 10.22331/q-2022-08-16-776.

Tôru Moriya. Anisotropic superexchange interaction and weak ferromagnetism. *Physical Review*, 120(1):91–98, 1960. doi: 10.1103/PhysRev.120.91.

Samuel Müller, Noah Hollmann, Sebastian Pineda Arango, Josif Grabocka, and Frank Hutter. Transformers can do bayesian inference. In *International Conference on Learning Representations (ICLR)*, 2022. URL https://openreview.net/forum?id=KSugKcbNf9.

Alberto Peruzzo, Jarrod McClean, Peter Shadbolt, Man-Hong Yung, Xiao-Qi Zhou, Peter J. Love, Alán Aspuru-Guzik, and Jeremy L. O'Brien. A variational eigenvalue solver on a photonic quantum processor. *Nature Communications*, 5:4213, 2014. doi: 10.1038/ncomms5213.

Ulrich Schollwöck. The density-matrix renormalization group in the age of matrix product states. *Annals of Physics*, 326(1):96–192, 2011. doi: 10.1016/j.aop.2010.09.012.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Yehui Tang, Hao Xiong, Nianzu Yang, Tailong Xiao, and Junchi Yan. Towards LLM4QPE: Unsupervised pretraining of quantum property estimation and a benchmark. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=vrBVFXwAmi.

Yehui Tang, Mabiao Long, and Junchi Yan. QuaDiM: A conditional diffusion model for quantum state property estimation. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=P7f55HQtV8.

Giacomo Torlai, Guglielmo Mazzola, Juan Carrasquilla, Matthias Troyer, Roger Melko, and Giuseppe Carleo. Neural-network quantum state tomography. *Nature Physics*, 14:447–450, 2018.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Vladyslav Verteletskyi, Tzu-Ching Yen, and Artur F. Izmaylov. Measurement optimization in the variational quantum eigensolver using a minimum clique cover. *The Journal of Chemical Physics*, 152(12):124114, 2020. doi: 10.1063/1.5141458.

Guifré Vidal. Efficient classical simulation of slightly entangled quantum computations. *Physical Review Letters*, 91(14):147902, 2003. doi: 10.1103/PhysRevLett.91.147902.

Xinbiao Wang, Yuxuan Du, Zihan Lou, Yang Qian, Kaining Zhang, Yong Luo, Bo Du, and Dacheng Tao. AiDE-Q: Synthetic labeled datasets can enhance learning models for quantum property estimation. *arXiv preprint arXiv:2509.26109*, 2025. doi: 10.48550/arXiv.2509.26109. URL https://arxiv.org/abs/2509.26109.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabás Póczos, Ruslan Salakhutdinov, and Alexander J. Smola. Deep sets. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, pp. 3391–3401, 2017. URL https://papers.neurips.cc/paper/2017/hash/f22e4747da1aa27e363d86d40ff442fe-Abstract.html.

A Implementation and Evaluation Details

A.1 Detailed shot-set encoder

This section expands the compact computation in Equation 6. Restoring the batch axis, tokens and embedded features have shapes $B \times K \times L$ and $B \times K \times L \times H$, respectively. All QuSSET-family models use $H = 512$, eight attention heads of dimension 64, and dropout 0.1.

Embedding and within-shot blocks. A shared lookup table and LayerNorm give

$$h_{ki}^{(0)} = \mathrm{LN}(\mathrm{Embed}(x_{ki})) \in \mathbb{R}^H. \tag{10}$$

For each shot independently, four pre-LayerNorm, noncausal Transformer blocks mix the L sites:

$$\begin{aligned} \widetilde{h}_{k,:}^{(r)} &= h_{k,:}^{(r-1)} + \mathrm{MHA}_{\mathrm{RoPE}}\Big(\mathrm{LN}(h_{k,:}^{(r-1)})\Big), \\ h_{k,:}^{(r)} &= \widetilde{h}_{k,:}^{(r)} + \mathrm{FFN}_{4H}\Big(\mathrm{LN}(\widetilde{h}_{k,:}^{(r)})\Big), \qquad r = 1, \dots, 4. \end{aligned} \tag{11}$$

Rotary positions affect queries and keys only along the qubit axis. The output $h^{\mathrm{sp}} = h^{(4)}$ has shape $K \times L \times H$ per state.

Site-aligned aggregation. The learned bank $\mu^{(0)} \in \mathbb{R}^{M \times H}$, with $M = 4$, is shared across sites and copied to each site before processing a state. For $H_i^{(4)} = h_{:,i}^{(4)} \in \mathbb{R}^{K \times H}$, memory queries first compress the K observations and then return their summary to the shot tokens:

$$\begin{aligned} M_i' &= \mu^{(0)} + \mathrm{MHA}\Big(\mathrm{LN}(\mu^{(0)}); \mathrm{LN}(H_i^{(4)})\Big), \qquad & M_i &= M_i' + \mathrm{FFN}_{2H}(\mathrm{LN}(M_i')), \\ \widetilde{H}_i &= H_i^{(4)} + \mathrm{MHA}\Big(\mathrm{LN}(H_i^{(4)}); \mathrm{LN}(M_i)\Big), \qquad & \overline{H}_i &= \widetilde{H}_i + \mathrm{FFN}_{2H}(\mathrm{LN}(\widetilde{H}_i)). \end{aligned} \tag{12}$$

Here $M_i \in \mathbb{R}^{M \times H}$, $\overline{H}_i \in \mathbb{R}^{K \times H}$, and $\mathrm{MHA}(Q; KV)$ denotes cross-attention. No shot positional encoding is used. The shared queries therefore give shot-permutation equivariance at cost $O(LKM)$, rather than $O(LK^2)$. Stacking the sites gives $\overline{h} \in \mathbb{R}^{K \times L \times H}$, after which a fifth spatial block produces

$$h_{k,:}^{\mathrm{set}} = \mathrm{SpatialBlock}_5(\overline{h}_{k,:}), \qquad h^{\mathrm{set}} \in \mathbb{R}^{K \times L \times H}. \tag{13}$$

Global summary. The Perceiver key–value sequence retains both local and memory features,

$$V = [\mathrm{flat}_{k,i}(h_{ki}^{\mathrm{set}}); \mathrm{flat}_{i,a}(M_{i,a})] \in \mathbb{R}^{(KL+LM) \times H}. \tag{14}$$

With $P = 8$ learned queries $g^{(0)} \in \mathbb{R}^{P \times H}$, the Perceiver block is

$$\begin{aligned} \widetilde{g} &= g^{(0)} + \mathrm{MHA}\Big(\mathrm{LN}(g^{(0)}); \mathrm{LN}(V)\Big), \\ \widehat{g} &= \widetilde{g} + \mathrm{MHA}(\mathrm{LN}(\widetilde{g}); \mathrm{LN}(\widetilde{g})), \\ g &= \widehat{g} + \mathrm{FFN}_{4H}(\mathrm{LN}(\widehat{g})) \in \mathbb{R}^{P \times H}. \end{aligned} \tag{15}$$

The second attention operation is latent self-attention. The task heads receive both h^{set} and g, rather than forcing local evidence through the global bottleneck.

Masked-outcome head. For $F = \mathrm{flat}_{k,i}(h_{ki}^{\mathrm{set}}) \in \mathbb{R}^{KL \times H}$, each local feature queries the global latents before the two-way sign classifier:

$$\begin{aligned} U &= F + \mathrm{MHA}(\mathrm{LN}(F); g) \in \mathbb{R}^{KL \times H}, \\ R &= U + \mathrm{FFN}_{2H}(\mathrm{LN}(U)) \in \mathbb{R}^{KL \times H}, \\ \ell_{ki} &= W_{\mathrm{out}} R_{ki} + b_{\mathrm{out}} \in \mathbb{R}^2, \qquad & p_{\theta_B, \theta_M}(s_{ki} \mid \widetilde{x}) &= \mathrm{softmax}(\ell_{ki})_{s_{ki}}. \end{aligned} \tag{16}$$

Logits are evaluated at all positions, but Equation 7 trains only on the masked set.

Table 3: Neural training configurations. Batch size counts 64-shot chunks during pretraining and physical states during supervision. All schedules are cosine after the listed warmup. LR and WD are the AdamW peak learning rate and weight decay.

Stage	Batch	Updates	LR	WD	Warmup	Final LR
QuSSET MLM	64	5,400	10^{-4}	10^{-2}	200	10^{-6}
QuSSET/scratch supervised	16	$S(N_t)$	10^{-4}	10^{-3}	$\min(200, \lfloor S/10 \rfloor)$	10^{-6}
LLM4QPE-like AR	128	2,760	10^{-4}	10^{-2}	200	10^{-6}
LLM4QPE-like supervised	16	$S(N_t)$	10^{-4}	10^{-3}	0	0

Pairwise heads. The projected queries in Equation 8 have shape $N \times H$ per state. They cross-attend to $g \in \mathbb{R}^{P\times H}$, then pass through one self-attention and $2H$-width feed-forward refinement over the N pairs. For optional circuit conditioning, a separate variable encoder produces site and global context in $\mathbb{R}^H$; a typed gated residual is added after the measurement-only cross-attention and pair refinement. Canonical QuSSET omits this path. The isotropic and DM heads use separate projection and prediction parameters, and scratch changes only the initialization of the otherwise identical measurement backbone and head.

Geometry features. Let $p_i = (i-1)/(L-1)$ and $e_i = \min(p_i, 1-p_i)$ for sites indexed from 1 to L. Both heads receive six length features,

$$\lambda_L = \left[L/32, (L/32)^2, \log L/\log 32, L^{-1}, \cos(\pi L/2), \sin(\pi L/2)\right]. \tag{17}$$

For the symmetric target, the remaining six entries are $[|p_i - p_j|, |p_i - p_j|^2, \mathbf{1}[i = j], (p_i + p_j)/2, p_i p_j, \min(e_i, e_j)]$. For the ordered DM target they are $[p_j - p_i, |p_j - p_i|, p_i, p_j, e_i, e_j]$. Concatenating either list with λ_L gives the 12-dimensional $\phi_C(L, i, j)$ or $\phi_D(L, i, j)$ in Equation 8.

Optimization and batching. Supervised budgets use

$$S(N_t) = 25\,\lceil N_t/16 \rceil \in \{50, 100, 150, 300, 650\} \tag{18}$$

steps for $N_t = 20, 50, 90, 180, 405$. Batches are length-homogeneous, nominally size 16, and retain each length group's incomplete batch. One traversal contains $\{6, 6, 6, 12, 30\}$ batches, so schedules are approximately $\{8.33, 16.67, 25, 25, 21.67\}$ traversals. The last traversal is partial when steps are not divisible by the batch count. This is expected and identical across matched methods. Label subsets are nested, shuffles are seeded, and evaluation does not select checkpoints. QuSSET-family results use the final raw weights of their predeclared schedules.

All neural models use AdamW and clip the gradient norm to 1. Table 3 records the remaining optimization settings. In masked pretraining, every 1,024-shot state record is partitioned into 16 nonoverlapping 64-shot chunks; all chunks are consumed rather than subsampling 64 shots once. A length-stratified 90/10 split leaves 1,080 training and 120 validation states. Twenty passes over the expanded chunks give 5,400 optimizer updates, with 15% of the shot–site outcomes masked.

The three seeds 42, 43, and 44 change model initialization, training order, the pretraining train/validation partition, and stochastic masking. Downstream physical datasets, evaluation splits, and nested labeled-state indices are fixed; each finetuning run is paired with the pretrained checkpoint of the same seed. During supervised training the seed therefore changes the newly initialized task head and data order, but not which labeled states are available.

Metrics. RMSE and R^2 pool upper-triangular entries unless resolved by length or distance. Zero-predictor RMSE in Figure 7 is $\sqrt{\mathbb{E}[C_{ij}^2]}$ within each bin. The within-distance R^2 uses the empirical bin mean as reference, so positive values indicate improvement over the distance-bin mean. Pair counts decrease at extreme distance; we report support and avoid interpreting one endpoint as the aggregate trend.

A.2 LLM4QPE-LIKE IMPLEMENTATION

The baseline uses a 256-dimensional, four-layer, eight-head causal Transformer with dropout 0.1 and a GRU-based encoder of the exact circuit variables. Autoregressive pretraining predicts each

six-valued local-Pauli outcome from the preceding sites of its shot and the circuit condition. At adaptation, a one-layer LSTM aggregates the 64 outcomes independently at each site, the causal Transformer mixes the site sequence, and a shared pair head predicts the requested matrix. Correlation and DM supervision minimize pooled off-diagonal RMSE. Pretraining uses all sixteen 64-shot chunks per state for 20 epochs; supervised fits use the same labeled states, records, and $S(N_t)$ update budgets as QuSSET. Table 3 gives its optimizer settings.

A.3 Calibrated Simple Baselines

Distance shrinkage and measurement-feature ridge use the same $K = 64$ local random-Pauli records and the same N_t labeled states as the learned estimators. They receive neither circuit variables nor unlabeled pretraining data. Both first reduce the shots to transparent pairwise statistics. For state s, pair $i < j$, and ordered Pauli axes $a, b \in \{X, Y, Z\}$, define

$$n_{sij}^{ab} = \sum_{k=1}^{K} \mathbf{1}[b_{ski} = a,\, b_{skj} = b],$$
$$m_{sij}^{ab} = \frac{1}{\max(n_{sij}^{ab}, 1)} \sum_{k:\, b_{ski}=a,\, b_{skj}=b} o_{ski} o_{skj}, \tag{19}$$

so an unobserved term has $m_{sij}^{ab} = 0$. The raw matched-basis estimates are

$$r_{sij}^{C} = \tfrac{1}{3}(m_{sij}^{XX} + m_{sij}^{YY} + m_{sij}^{ZZ}),$$
$$r_{sij}^{D} = \frac{m_{sij}^{YZ} - m_{sij}^{ZY} + m_{sij}^{ZX} - m_{sij}^{XZ} + m_{sij}^{XY} - m_{sij}^{YX}}{4\sqrt{3}}. \tag{20}$$

The implementation then enforces symmetry for C and antisymmetry for D. The effective count n_{sij} sums the three same-axis counts for C or the six cross-axis counts appearing in Equation 20 for D.

Balanced cross-basis acquisition. For each length, one 64-shot schedule is fixed and reused for all evaluation states, using base seed 61027. A candidate shot randomly permutes the approximately balanced site template $X, Y, Z, X, Y, Z, \ldots$. From 256 candidate schedules we first maximize the minimum count of every ordered cross-axis basis YZ, ZY, ZX, XZ, XY, YX over all pairs, then minimize the summed squared deviation of those counts from $64/9$. The selected schedule has nonzero coverage for all six terms of every pair. Its estimator is exactly r_{sij}^{D} in Equation 20, using the empirical mean from the shots assigned to each ordered basis. Each such mean is unbiased for its Pauli-product expectation, so their fixed linear combination is unbiased for D_{ij}; measurements on the other sites only specify the joint basis and do not change the pairwise marginal.

Distance shrinkage. Let $\delta_{ij} = (j - i)/(L - 1)$, and assign each pair to one of eight uniformly spaced bins $q = \min(7, \lfloor 8\delta_{ij} \rfloor)$. From the labeled training pairs in bin q, with exact target y_{sij}, we estimate

$$\mu_q = \mathrm{mean}(y_{sij}), \qquad \tau_q^2 = \mathrm{var}(y_{sij}), \qquad \nu_q = \mathrm{mean}\left[\max(n_{sij}, 1)(r_{sij} - y_{sij})^2\right]. \tag{21}$$

For a test pair in the same bin, the estimator is

$$\widehat{y}_{sij} = \mu_q + w_{sij}(r_{sij} - \mu_q), \qquad w_{sij} = \frac{\tau_q^2}{\tau_q^2 + \nu_q / \max(n_{sij}, 1)}, \tag{22}$$

with $w_{sij} = 0$ if no relevant term was observed. Thus poorly sampled pairs revert toward the learned distance prior, while well-sampled pairs retain more of their measurement estimate. Empty or singleton bins use the global fitted moments. Predictions are clipped to $[-1, 1]$ for C and $[-1/2, 1/2]$ for D, then completed with the appropriate matrix symmetry.

Measurement-feature ridge. One shared linear model predicts every pair and length. For C, its 15-dimensional feature vector is

$$x_{sij}^{C} = \left[r_{sij}^{C},\ (m_{sij}^{aa})_a,\ (n_{sij}^{aa}/K)_a,\ (\log(1 + n_{sij}^{aa}))_a,\ \delta_{ij}, \delta_{ij}^2, (j - i)^{-1}, L^{-1}, (i + j + 1)/(2L)\right]. \tag{23}$$

For D, the analogous 33-dimensional vector replaces each three-entry same-axis block with all nine ordered (a, b) term means and counts, while retaining r^D_{sij} and the same five geometry features. Features are standardized using training-set statistics. With an unpenalized intercept, ridge fits

$$(\widehat{\beta}_0, \widehat{\beta}) = \arg\min_{\beta_0,\beta} \sum_{s,i<j} \left(y_{sij} - \beta_0 - \beta^\top \widetilde{x}_{sij}\right)^2 + \lambda\|\beta\|_2^2. \tag{24}$$

We select $\lambda \in \{0, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 100, 1000\}$ by five-fold cross-validation grouped by state, so no pair from a held-out state appears in its fitting folds. The selected model is refit on all N_t labeled states and clipped and symmetrized as above. This baseline tests whether a small supervised calibration of explicit measurement statistics and geometry suffices, without a learned shot-set representation.

B Complete Estimator Comparison

Main Figure 2 retains one representative from each estimator class so that the two-property comparison remains legible. Figure 4 instead audits the isotropic result against every implemented baseline.

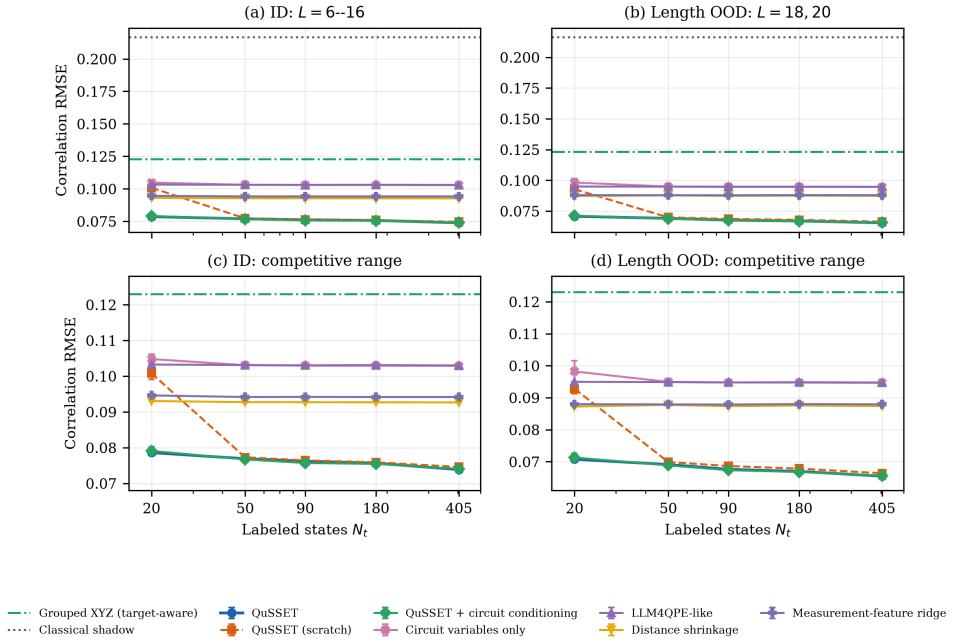


Figure 4: Complete isotropic-correlation comparison. Top: all methods; bottom: the competitive range. Trained curves show mean ± one sample standard deviation over three seeds, and horizontal lines are label-free estimators. Grouped XYZ uses its separately acquired target-aware dataset; all other measurement-based estimators use the shared local random-Pauli records.

The expanded view preserves the ranking emphasized in the main text. It also exposes the distinction between improving the estimator on the common random-Pauli corpus and changing acquisition itself: Grouped XYZ receives favorable target-specific measurements, whereas QuSSET does not.

Table 4: Measurement and supervision accounting. N_t means exact labels for that many states.

Method	Test shots	Acquisition	Labels	Pretraining
QuSSET	64	local random Pauli	N_t	unlabeled shots
QuSSET + circuit variables	64	local random Pauli	N_t	unlabeled shots
QuSSET (scratch)	64	local random Pauli	N_t	none
LLM4QPE-like	64	local random Pauli	N_t	model pretraining
Distance shrinkage	64	local random Pauli	N_t	none
Measurement-feature ridge	64	local random Pauli	N_t	none
Classical shadow	64	local random Pauli	0	none
Grouped XYZ	64	cyclic 22/21/21 X/Y/Z	0	none
Balanced cross-basis	64	target-balanced site bases	0	none
Circuit variables only	0	measurements ignored	N_t	none

C Length-Wise Extrapolation

The main results pool the held-out lengths $L = 18, 20$. Figure 5 separates every length, showing behavior across the training boundary and at both held-out endpoints.

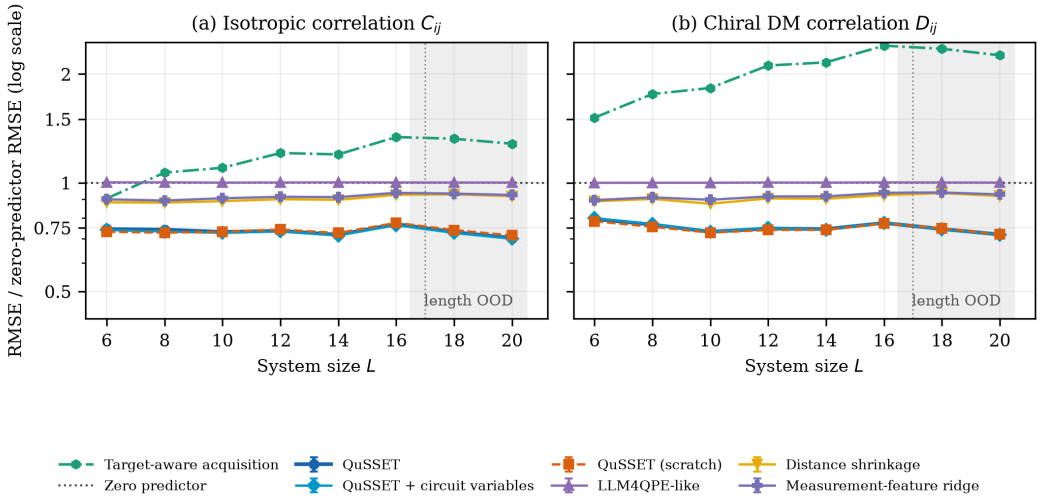


Figure 5: Variable-length adaptation at $N_t = 90$ for isotropic and chiral DM correlations. RMSE is normalized by the zero-predictor RMSE at each length and shown on a logarithmic scale. The same model parameters are evaluated at every length; the shaded $L = 18, 20$ region is absent from unlabeled and labeled training. Target-aware curves use separately acquired grouped-XYZ or balanced cross-basis measurements.

Both observable families change smoothly through the training range and do not abruptly deteriorate at 18 or 20. The per-length curves therefore confirm the pooled extrapolation result across both held-out sizes despite their length-dependent target distributions.

D Diagnosing the Long-Range Pretraining Gain

D.1 Complementary In-Distribution Result

Main Figure 3 focuses on length OOD. Figure 6 reports only the complementary ID calculation and asks whether the same distance dependence is already visible at the supervised lengths.

The ID result follows the same qualitative pattern as the OOD result: pretraining is most visible when labels are scarce, while its remaining higher-budget contribution is concentrated in longer-range entries rather than uniform over the matrix.

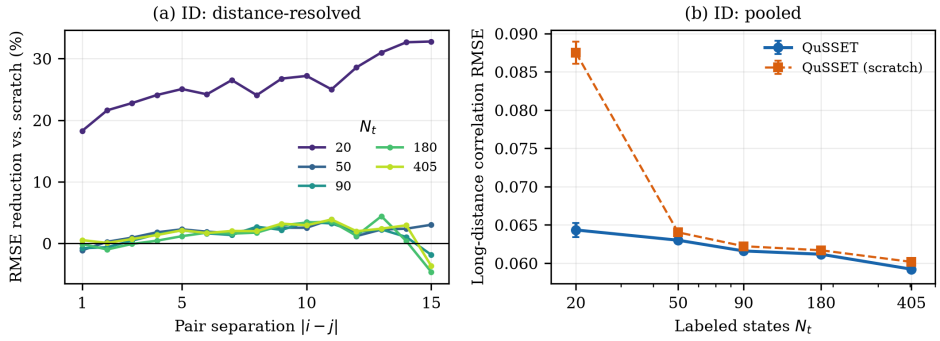


Figure 6: In-distribution isotropic-correlation pretraining diagnostic. Left: distance-resolved RMSE reduction relative to the architecture-matched scratch estimator for every label budget. Right: raw RMSE pooled over long-distance entries. The large low-label gain becomes smaller but remains positive in aggregate after the overall estimators have nearly converged.

D.2 Variation beyond distance-bin means

A distance-dependent gain could be misleading if long-range targets were merely small and the model learned to predict zero. We therefore choose $N_t = 90$, after the low-label optimization transient, and compare error with the zero-predictor scale as well as variation around each distance-bin mean.

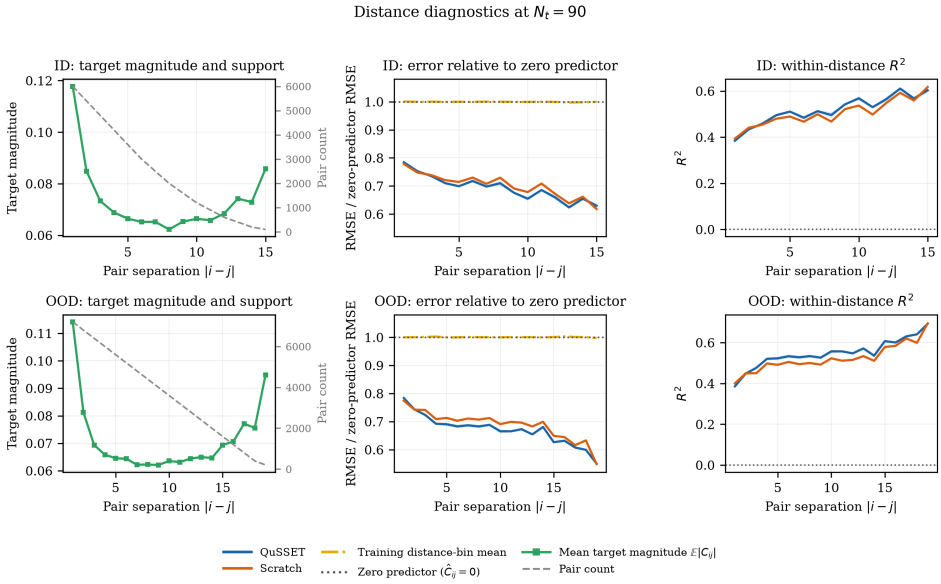


Figure 7: Isotropic distance diagnostics at $N_t = 90$. Left: mean $|C_{ij}|$ and pair support. Middle: distance-bin RMSE divided by zero-predictor RMSE; values below one improve on $\widehat{C}_{ij} = 0$. Right: R^2 relative to the mean within each distance bin. Bands show one standard deviation over seeds.

Positive within-distance R^2 indicates that the estimators explain variation beyond the distance-bin mean. The decreasing pair support at the largest separation is why the main claim concerns an aggregate long-range region instead of every endpoint bin.

E Optimization-Path Audit at $N_t = 20$

The main experiment trains $N_t = 20$ for a complete 50-step cosine schedule with five warmup steps. The continuous 600-step audit uses a distinct cosine schedule with 60 warmup steps. Because checkpoints from these schedules do not share the same learning-rate history, Figure 8 reports the extended trajectories and the main 50-step endpoints separately.

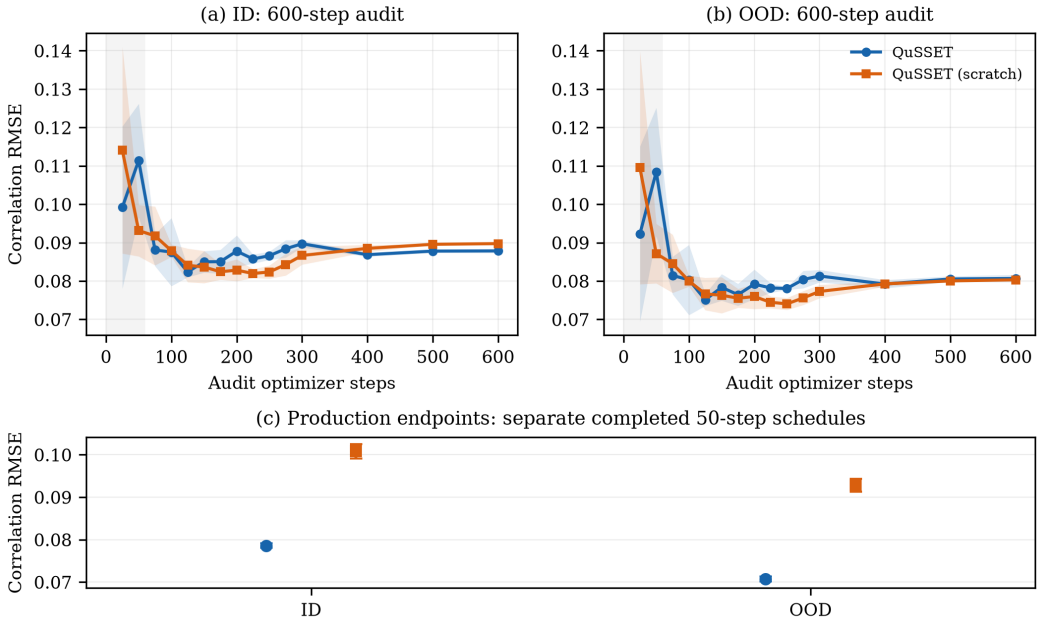


Figure 8: Optimization audit at $N_t = 20$. Top: checkpoints along one continuous 600-step schedule; the lightly shaded interval is its 60-step warmup. Bottom: final results from the separately trained production runs, each completing its own 50-step schedule with five warmup steps. All panels show raw weights and mean $\pm$ one standard deviation over three seeds.

In the extended runs, the pretrained estimator reaches its low-error region earlier than scratch. Scratch can reach comparable error with longer training, while both models eventually overfit the 20 labeled states. The low-label advantage therefore depends in part on the optimization schedule.

F Stress Test: Can Circuit Variables Replace Measurements?

The main discussion attributes success to state-specific measurement evidence rather than a simple deterministic map from circuit parameters. We test that interpretation under favorable conditions for the shortcut: the standard variable-only model uses $N_t = 405$ and 650 steps, while extended GRU-512, distance-512, and GRU-1024 variants use all 600 labels and 2,600 steps. The largest model has about 63 million parameters. None receives measurements at evaluation.

The standard variable model matches zero prediction: ID/OOD RMSE is $0.10289 \pm 0.00001/0.09465 \pm 0.00002$, with R^2 near zero. GRU-512 and GRU-1024 remain at this level after extended training, while the distance model is worse. In contrast, QuSSET at $N_t = 405$ reaches $0.07385 \pm 0.00011/0.06550 \pm 0.00012$ and $R^2 = 0.484/0.521$. This is an empirical stress test, not an impossibility or complexity result. It shows that the broadly sampled variables do not yield a useful held-out shortcut under the tested direct architectures.

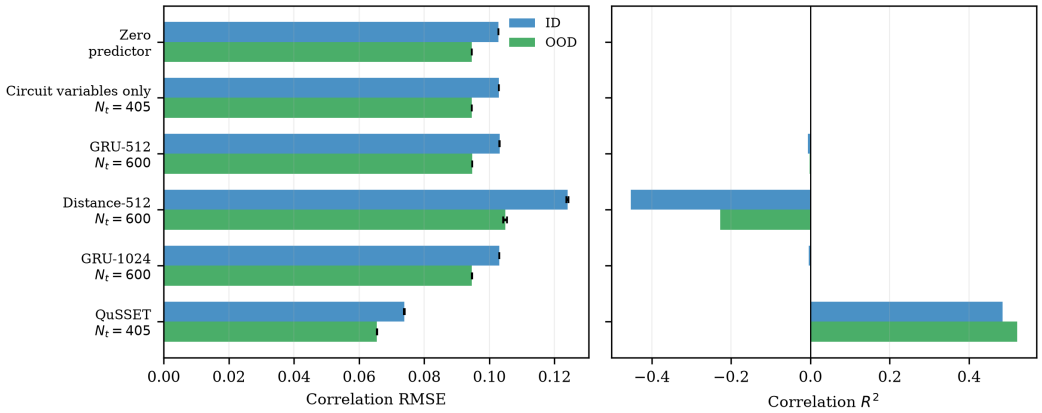


Figure 9: Circuit-variable-only stress test. Left: correlation RMSE; dashed markers denote zero predictors. Right: correlation R^2. Larger encoders, more labels, and longer optimization do not produce positive held-out R^2, whereas measurement-based QuSSET does. Error bars show one standard deviation over three seeds.

G Hamiltonian Contraction as an Auxiliary Analysis

This final analysis asks whether the predicted matrix can support a physically motivated scalar contraction. It is separate from the paper's two matrix-estimation tasks and is not used to support the main comparison. Define $Z_L = \sum_{i<j} |i-j|^{-\alpha_H}$. With $\alpha_H = 0.5$, we use the field-free, Kac-normalized Hamiltonian

$$H_L = -\frac{4L}{3Z_L} \sum_{i<j} |i-j|^{-\alpha_H} \mathbf{S}_i \cdot \mathbf{S}_j, \qquad e_L = \frac{\langle H_L \rangle}{L} = -\frac{1}{Z_L} \sum_{i<j} |i-j|^{-\alpha_H} C_{ij}. \tag{25}$$

For each N_t, a separate correlation estimator is trained from the masked-pretraining checkpoint or from scratch; it is not continued from the correlation-only fit. We use the same $S(N_t) = \{50, 100, 150, 300, 650\}$ update budgets and peak learning rate 10^{-4} as the main experiment, with loss $0.5\mathcal{L}_{\text{term}} + 0.5\,\text{MSE}(\widehat{e}_L, e_L)$. Target distributions differ between mixed ID and OOD lengths, so Figure 10 normalizes RMSE by target RMS separately and also reports R^2.

At $N_t = 20$, pretraining improves normalized ID energy RMSE from 1.037 ± 0.050 to 0.848 ± 0.059. At larger budgets scratch and pretrained models are close, and scratch is sometimes slightly better; we do not claim a uniform energy-specific pretraining advantage. This result is not an end-to-end VQE experiment: no optimization trajectory, ranking preservation, or final variational energy is evaluated.

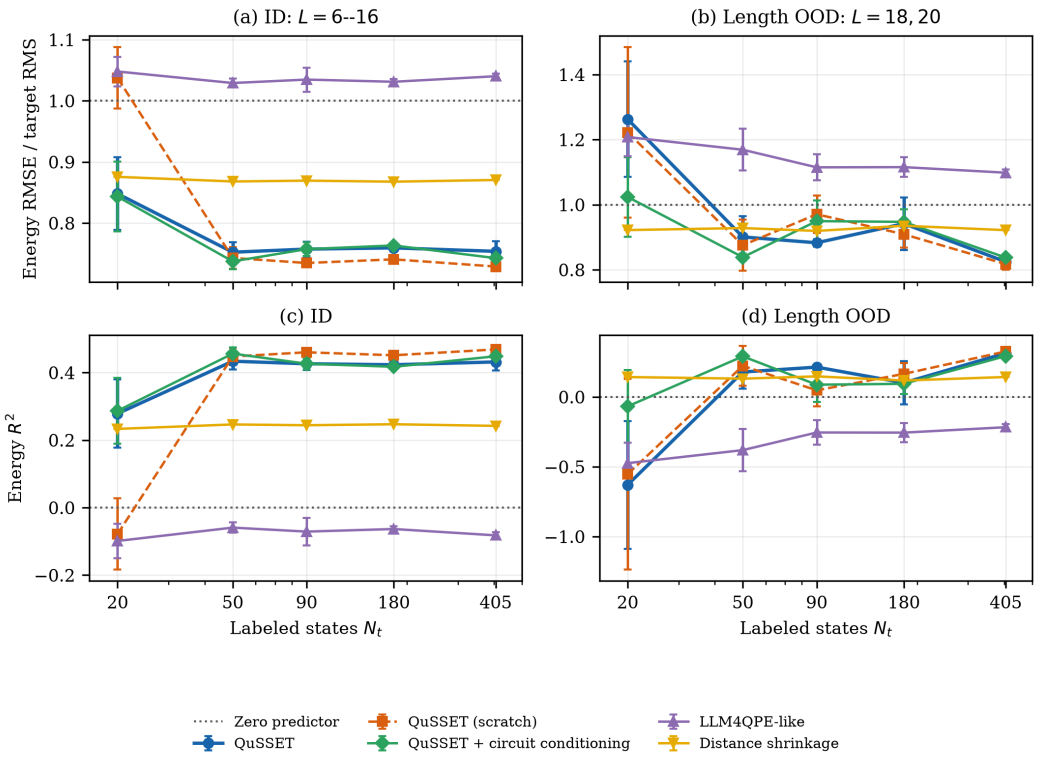


Figure 10: Auxiliary Heisenberg-energy estimation. Top: energy/site RMSE normalized by target RMS of each split. Bottom: R^2. Energy supervision uses separate hybrid-loss fits with the same labeled-state and update budgets as the main experiment and is not used in main results.