QuMUSE: Generative Pretraining from Local Measurements for Characterizing Quantum Many-Body Ground States

Anonymous authors
Paper under double-blind review

Abstract

Characterizing quantum many-body ground states is fundamental to understanding quantum matter and advancing quantum technologies. However, most existing learning methods remain specialized to a particular Hamiltonian family, system size, or tasks, limiting knowledge reuse across physical systems. We introduce QuMUSE, a local measurement-native generative foundation model that learns unified, reusable representations of many-body ground states from local observables, targeting Hamiltonian families of variable scales and lattice geometries. QuMUSE consists of a variational autoencoder that compresses variable-dimensional measurement data into a fixed-size, physical system-aware latent state representation that decodes local observables, and a conditional latent flow model that generates state representations from physical specifications alone. Therefore, one unified representation supports both characterization from measurements and prediction without target-state measurements. Jointly pretrained on five 1D and 2D Hamiltonian families with up to 100 qubits, QuMUSE achieves the best overall observable-reconstruction and conditional-generation performance among baselines. Despite training only on local observables, the representations support predicting nonlinear and global properties, data-efficient adaptation to unseen Hamiltonian families, and unsupervised phase boundary witnessing. These results demonstrate the potential of measurement-native generative pretraining to reuse physical information across quantum systems and learning tasks.

1 Introduction

Understanding the physics of a quantum many-body system does not require reconstructing its entire wavefunction. The most faithful representation of an N-qubit closed quantum system, the state vector description, contains 2^N complex amplitudes, yet many physically relevant quantities depend only on local observables. For local Hamiltonians, useful properties of their ground states can be evaluated from few-body Pauli expectation values, without recovering the full state (Huang et al., 2020). This motivates a different objective for quantum representation learning, i.e., to retain the observable information needed for physical prediction, rather than every wavefunction amplitude. Local measurements provide a natural interface for this objective, connecting experimentally accessible information to the physical behavior of a system. For families of many-body ground states, the opportunity extends beyond characterizing individual states: learning jointly from related systems could turn their measurement data into a unified, reusable representation that supports prediction and adaptation across physical settings.

Realizing this opportunity requires more than a compact encoding of each state. Recently, machine learning based methods have explored representing ground state wavefunctions (Carleo & Troyer, 2017; Sharir et al., 2020; Pfau et al., 2020; Zhang & Di Ventra, 2023; Chen & Heyl, 2024), generating state preparation circuits (Wauters et al., 2020; Du et al., 2022; Wu et al., 2023; Liao et al., 2025), predicting state properties (Zhu et al., 2022; Wu et al., 2024; Tang et al., 2024a; Lewis et al., 2024; Wang et al., 2025), and synthesizing measurement data from Hamiltonian parameters (Wang et al., 2024; Tang et al., 2025). However, these methods are limited to learning a single state or a structured state family. What remains underexplored is a shared representation that unifies measurement-based inference and conditional generation across Hamiltonian families, system sizes, and lattice geome-

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.

tries. This is challenging for current methods because measurement collections vary in system size and spatial organization, while a reusable representation must support different observable queries and be accessible from either measurements or physical specifications. Such a representation would allow the same framework to characterize measured states and predict the observables for which measurements are unavailable. This raises a central question: *can one pretrained model establish a unified, fixed-size representation of many-body ground states that can be inferred from local measurements, generated from physical conditions, and efficiently adapted to new quantum systems?*

We introduce **QuMUSE**, a measurement-native generative pretraining framework that addresses this question through a unified, reusable representation of quantum many-body ground states. MUSE refers to **M**easurement-Native **U**nified **S**tate **E**mbedding. Our framework does not attempt an informationally complete reconstruction of an exponentially large wavefunction from low-order observables. Instead, it learns a compact unified representation of the local-observable content available to the model, which makes the framework measurement-native. A variational autoencoder (Kingma & Welling, 2013) maps a variable collection of local measurement data into a fixed-size latent, while an observable-queryable decoder predicts requested local observables given query measurement operators. Hence, the same latent can answer different observable queries without training a separate predictor for each quantity. In parallel, a physical encoder maps the Hamiltonian family, coefficients, lattice geometry, and system size into a semantic space aligned with the measurement-derived state representation. A conditional latent flow model (Ho et al., 2020; Rombach et al., 2022) then learns to produce state latent representations from these physical descriptors. We further introduce a decoder-aware objective, which penalizes latent errors according to their consequences for physical observables rather than only by isotropic distance in latent space. The resulting unified representation supports two complementary routes through the same framework, namely measurements $\rightarrow$ latent $\rightarrow$ observables, and physical specification $\rightarrow$ latent $\rightarrow$ observables.

We evaluate one jointly pretrained framework on five representative Hamiltonian families, including the 1D and 2D transverse-field Ising models, the 1D anisotropic Heisenberg and cluster-Ising models, and the 2D J_1–J_2 model with up to 100 qubits. Our framework attains the best overall reconstruction and conditional generation performance over the baseline methods. We further transfer the pretrained model to triangular-lattice Ising, long-range Ising, and Kitaev honeycomb systems using few-shot transfer learning. Compared to training the same architecture from scratch with the same computational budget, pretraining exhibits overall better performance. Beyond these tasks, we demonstrate that the pretrained representation supports physical applications without retraining the base model. Linear probing directly on the state representations can support predicting nonlinear and global properties beyond training. The learned representations also provide unsupervised witnesses of phase boundaries without requiring phase labels. Together, these results show that measurement-native generative pretraining produces representations that are reusable across observable prediction, quantum-property estimation, phase analysis, and adaptation to unseen systems.

Our contributions are summarized as follows:

- We formulate a measurement-native pretraining paradigm for quantum many-body ground states, in which one unified representation connects local observables with physical descriptions without wavefunction amplitudes or per-instance variational optimization.
- We introduce QuMUSE, which integrates a unified observable-queryable representation, alignment between measurement and physical-condition semantics, conditional latent generation, and a decoder-aware objective that prioritizes observable estimation accuracy.
- We demonstrate accurate observable reconstruction and conditional generation with one jointly pretrained framework spanning five Hamiltonian families and systems of up to 100 qubits. The learned representations further support data-efficient adaptation to three unseen quantum systems, linear probing of linear, nonlinear, and nonlocal quantum properties, and unsupervised phase-boundary witnessing.

2 Related Work

Neural representations of quantum many-body states. Neural quantum states (NQS) provide compact parameterizations of many-body wavefunctions and have been successfully applied to ground-state optimization (Carleo & Troyer, 2017; Carrasquilla et al., 2019; Sharir et al., 2020;

Pfau et al., 2020; Zhang & Di Ventra, 2023; Chen & Heyl, 2024), solving excited states (Pfau et al., 2024; Ma et al., 2025) and quantum dynamics (Schmitt & Heyl, 2020; Gutiérrez & Mendl, 2022; Qi et al., 2026). These approaches typically model wavefunction amplitudes or density matrices explicitly, thereby enabling variational energy optimization and sampling but often requiring a new optimization procedure for each Hamiltonian or target state. Recent foundation NQS alleviate this limitation by conditioning a shared variational wavefunction on Hamiltonian couplings and other physical descriptors, allowing one set of parameters to represent multiple related Hamiltonians (Rende et al., 2025; Zaklama et al., 2026). In contrast, our framework does not attempt to represent the exponentially large wavefunction. It instead uses a measurement-native formulation, which provides a common interface across Hamiltonian families and system sizes and is designed primarily for observable generation and property prediction rather than wavefunction optimization.

Learning quantum states and properties from measurements. Machine learning has also been used to extract useful information directly from quantum measurements. Neural quantum-state tomography learns generative models of measurement distributions from data collected for quantum states tomography (Torlai et al., 2018; Carrasquilla et al., 2019; Cha et al., 2021; Ahmed et al., 2021; Schmale et al., 2022; Wei et al., 2024), while classical shadows provide compact randomized-measurement estimators for predicting many observables without reconstructing the full density matrix (Huang et al., 2020; Akhtar et al., 2023). For parameterized local Hamiltonians, theoretical results further show that measurement data from related systems can enable efficient prediction of ground-state properties throughout a quantum phase (Huang et al., 2022; Lewis et al., 2024). Beyond state-specific reconstruction, generative query networks learn shared representations from measurements of structurally related states that can predict previously unseen measurement queries (Zhu et al., 2022). Multi-task models based on short-range correlations similarly learn representations that support the prediction of global and nonlinear properties (Wu et al., 2024). SSL4Q (Tang et al., 2024b) uses semi-supervised learning for state classification. Huang et al. (2025) select adaptive measurement bases for better property estimation. LLM4QPE (Tang et al., 2024a) treats measurement outcomes as sequences and uses autoregressive pretraining followed by property-specific fine-tuning. AiDE-Q (Wang et al., 2025) generates synthetic labels for property estimation.

Generative modeling for quantum states. Generative models offer an alternative to direct property regression by learning a distribution from which measurement outcomes or state representations can be sampled. Wang et al. (2024) condition an autoregressive model on Hamiltonian parameters to generate measurement outcomes of ground states. Zhu et al. (2024) directly generate density matrices using a diffusion model. QuaDiM (Tang et al., 2025) applies conditional diffusion directly to quantum measurement distributions and estimates state properties from the generated samples.

Our approach instead follows the general paradigm of generative representation learning. A variational autoencoder learns compact probabilistic representations through reconstruction and latent regularization (Kingma & Welling, 2013), while self-supervised consistency objectives can encourage representations to preserve semantic information and avoid collapse (Bardes et al., 2022). Latent diffusion (Rombach et al., 2022) and flow-matching (Liu et al., 2022; Albergo & Vanden-Eijnden, 2023; Lipman et al., 2023) methods model complex data distributions in the compressed space of a pretrained autoencoder, reducing the dimensionality of the generative problem while retaining a structured decoder. We adapt this general framework to quantum data. Unlike direct diffusion over measurement records, the generative model operates on a shared latent space whose dimensionality does not increase with the system size.

3 Generative Foundation Model for Quantum State Learning

We introduce QuMUSE, a measurement-native generative framework for learning unified representations of quantum many-body ground states. Given ground states of different Hamiltonian families and system sizes, we want to: (1) Learn unified representations for them; (2) Generate the state representations given descriptions of the physical system. The overall framework is shown in Figure 1. We first introduce the dataset construction procedure and the unified state representation in Section 3.1. The representation is learned by a query-conditioned VAE consisting of an encoder E_ϕ and a decoder D_ψ, as detailed in Section 3.2. After that, in Section 3.3, we introduce a conditional latent flow model v_θ, which learns to generate the latent representations of the states directly from

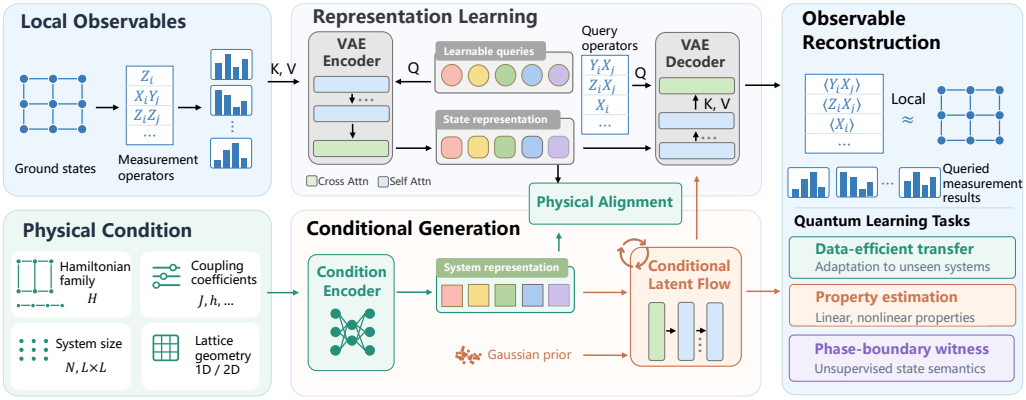

Figure 1: Overview of QuMUSE. A VAE with learnable queries compresses variable-length collections of local measurement data into a fixed-size latent representation and reconstructs the requested observables. A physical condition branch aligns the state representation with the physical system, enabling physical system awareness for the state representation. A conditional latent flow model subsequently generates state representations given Hamiltonian and lattice conditions, and the generated representations can be decoded back to physical observables with the decoder.

the physical condition $\mathbf{c}$. The resulting representation supports both observable prediction from measurement outcomes and conditional generation from physical specifications.

3.1 Measurement-Native Unified Representation of Ground States

Unlike neural quantum states (Carleo & Troyer, 2017; Torlai et al., 2018; Carrasquilla et al., 2019; Sharir et al., 2020; Zhang & Di Ventra, 2023; Chen & Heyl, 2024) that use a neural network to represent a quantum state, we propose the measurement-native unified representation, which directly acts on measurement outcomes and does not assume any knowledge of the Hamiltonian. This representation is agnostic to the system size, the type of the Hamiltonian, or the lattice geometry.

To learn this representation, we first construct a dataset $\mathcal{M}$ composed of the local measurement data $\mathbf{M_c}$ from ground states of Hamiltonians across multiple families and system sizes, $\mathcal{M} = \{\mathbf{M_c}\}_{\mathbf{c}}$. Let H_f denote a Hamiltonian within a Hamiltonian family $\mathcal{H}$ with identifier f, $\boldsymbol{\theta}$ its coefficients, and $\boldsymbol{\ell}$ the lattice specification. These quantities are combined into a physical condition $\mathbf{c} = [f, \boldsymbol{\theta}, \boldsymbol{\ell}]$. For the corresponding ground state $|\psi(\mathbf{c})\rangle$, we collect the local measurement data $\mathbf{M_c} = \{(\mathbf{m}_j^{(\mathbf{c})}, \mathbf{P}_j)\}_{j=1}^{M}$, where $\mathbf{P}_j = \{P_{j,a}\}_{a=1}^{B_j} \subset \mathcal{P}$ are B_j local measurement operators acting on a local support S_j of neighboring sites of qubits indexed by j and

$$m_{j,a}^{(\mathbf{c})} = \langle\psi(\mathbf{c})|P_{j,a}|\psi(\mathbf{c})\rangle\,, \qquad \mathbf{m}_j^{(\mathbf{c})} = \left(m_{j,1}^{(\mathbf{c})}, \ldots, m_{j,B_j}^{(\mathbf{c})}\right), \tag{1}$$

are the associated measurement outcomes. Note that we do not restrict B_j to be the same for all S_j because randomized and incomplete measurement is allowed in our framework.

To ensure the representation in fixed size and lattice-geometry agnostic, we take inspiration from QFormer (Li et al., 2023) and design a query-conditioned encoder network E_ϕ to map the variable-shape measurement dataset $\mathcal{M}$ into a unified fixed-size compact latent representation $\mathbf{z}$ for the quantum state. First, we sample a subset of measurement data $M_{\mathbf{c}} \subset \mathcal{M}$ corresponding to one state $|\psi(\mathbf{c})\rangle$ from the families, flatten and map it to a feature $\mathbf{V_c} \in \mathbb{R}^{B\times d_z}$, where B is the number of sampled measurement results. Then, we initialize K learnable queries $\mathbf{Q} \in \mathbb{R}^{K\times d_z}$, where K and d_z specify the dimension of the latent representation for a quantum state. The learnable queries interact with the measurement data via cross attention (Vaswani et al., 2017), which injects the information of the measurement data into them while keeping the dimension unchanged

$$\mathbf{o} = \mathrm{CrossAttn}(\mathbf{Q}, \mathbf{V_c}) = \mathrm{softmax}\left(\frac{\mathbf{Q}{\mathbf{V_c}}^T}{\sqrt{d_z}}\right)\mathbf{V_c} \in \mathbb{R}^{K\times d_z}. \tag{2}$$

Finally, after applying multiple self-attention transformer blocks to $\mathbf{o}$, we obtain the latent representation $\mathbf{z} \in \mathbb{R}^{K\times d_z}$ for the quantum state, independent of the system size. Note that this latent

does not explicitly encode any information about the physical system. To enable state representation generation from physical system descriptions using a latent flow model, we instead follow the best practice of latent diffusion Rombach et al. (2022) to model the latent as a Gaussian posterior,

$$q_\phi\left(\mathbf{z} \mid M_{\mathbf{c}}\right) = \mathcal{N}\left(\boldsymbol{\mu}_\phi(M_{\mathbf{c}}), \operatorname{diag} \boldsymbol{\sigma}_\phi^2(M_{\mathbf{c}})\right), \tag{3}$$

where $\boldsymbol{\mu}_\phi(M_{\mathbf{c}})$ and $\boldsymbol{\sigma}_\phi(M_{\mathbf{c}})$ are the mean and standard deviation predicted by the encoder. The latent is sampled through $\mathbf{z} = \boldsymbol{\mu}_\phi + \boldsymbol{\sigma}_\phi \odot \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and $\odot$ denotes element-wise multiplication.

3.2 Query-Conditioned VAE with Physical Condition Alignment

We now elaborate on how to learn the unified state representations for ground states from multiple Hamiltonian families. We train a variational autoencoder (VAE), in which the encoder compresses the variable-length measurement data into the unified state representation, and the decoder uses the measurement operators of interest to query the corresponding measurement values. Meanwhile, to incorporate the physical system information, we add a physical condition encoder branch and align the physical latent representation with the state represenation.

Observable-queryable decoder. For one state representation $\mathbf{z}_i$ introduced in Section 3.1, the decoder first maps it with stacks of Transformer blocks, obtaining feature $\mathbf{R}_i \in \mathbb{R}^{K \times d_o}$, where d_o is the feature dimension after the sequence of Transformer blocks. A query-based output head then predicts the measurement expectation values given queries of local operator $P_{j,a}$ via cross attention:

$$\widehat{m}_{i,j,a} = D_\psi\left(\mathbf{z}_i, P_{j,a}\right) = \text{Linear}(\text{CrossAttn}(\text{Rope}(\text{Linear}(P_{j,a}), g_j), \mathbf{R}_i)), \tag{4}$$

g_j specifies the local support the measurement operator acts on, and $\text{Rope}(\cdot, \cdot)$ is the rotary positional embedding (Su et al., 2023) to encode the position information. This design allows one shared decoder backbone to produce different output lengths and flexible shapes depending on the sizes and the lattice layouts of the target physical systems. Thereby, it enables joint learning for Hamiltonians across different families and system sizes.

To supervise the measurement value reconstruction, given a minibatch of samples $\mathcal{B}$, we compute

$$\mathcal{L}_{\text{rec}} = \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \frac{1}{|\Omega_i|} \sum_{(j,a) \in \Omega_i} \left(\widehat{m}_{i,j,a} - m_{i,j,a}\right)^2, \tag{5}$$

where Ω_i denote the valid local observables for state i. Per-state normalization prevents systems with larger size from dominating the objective because they contain more measurement values.

Alignment with physical conditions. A physical condition encoder maps the physical condition $\mathbf{c}$ to a diagonal Gaussian density, allowing the state representation be aware of the physical system, which benefits the generation stage. See Appendix C.2 for ablation study on the effectiveness of this design. The condition representation is also modeled as a Gaussian

$$p_\omega\left(\mathbf{s} \mid \mathbf{c}\right) = \mathcal{N}\left(\boldsymbol{\mu}_c, \operatorname{diag} \boldsymbol{\sigma}_c^2\right), \qquad \left(\boldsymbol{\mu}_c, \log \boldsymbol{\sigma}_c^2\right) = g_\omega(\mathbf{c}). \tag{6}$$

The condition encoder is trained to align with the vectorized state representation $\mathbf{r} = \text{vec}(\mathbf{z}) \in \mathbb{R}^{d_r}$:

$$\mathcal{L}_{\text{cond}} = \frac{1}{2d_r} \sum_{k=1}^{d_r} \left[\left(r_k - \mu_{c,k}\right)^2 \exp\left(-\log \sigma_{c,k}^2\right) + \log \sigma_{c,k}^2\right]. \tag{7}$$

The complete VAE objective is

$$\mathcal{L}_{\text{VAE}} = \mathcal{L}_{\text{rec}} + \lambda_{\text{cond}} \mathcal{L}_{\text{cond}} + \lambda_{\text{KL}} \mathcal{L}_{\text{KL}}, \tag{8}$$

where $\mathcal{L}_{\text{KL}}$ regularizes the latent space for the state representations towards isotropic Gaussian. After this stage, the VAE components are fixed. The learned representations are provided to the second stage conditional latent flow model for generative modeling.

3.3 Conditional Latent Flow with Decoder-Aware Reconstruction Loss

After training the VAE, the next stage learns to generate the state latent representation directly from the physical conditions. We instantiate this stage as a conditional Scalable Interpolant Transformer (SiT) (Ma et al., 2024) trained with a continuous flow-matching objective (Liu et al., 2022; Albergo & Vanden-Eijnden, 2023; Lipman et al., 2023). Operating in the unified latent space avoids applying diffusion directly to measurement values whose dimensionality changes with the system size.

Generative modeling with flow matching. Given the state representation $\mathbf{z}$ produced by the frozen VAE encoder, we normalize all training representations using a single scalar factor a_z estimated from the training set to construct the target $\mathbf{x}_1 = a_z\mathbf{z}$. The source noise is sampled from an isotropic Gaussian, $\mathbf{x}_0 = \sigma_0\boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. For timestep $t \sim \mathcal{U}[0, 1]$, we define a linear stochastic interpolant between the source and the target

$$\mathbf{x}_t = (1-t)\mathbf{x}_0 + t\mathbf{x}_1, \qquad \mathbf{u}_t = \mathbf{x}_1 - \mathbf{x}_0. \tag{9}$$

A conditional velocity model $v_\theta(\mathbf{x}_t, t, \mathbf{c})$ is trained to predict the target velocity $\mathbf{u}_t$

$$\mathcal{L}_{\text{flow}} = \mathbb{E}_{\mathcal{M},\mathbf{x}_0,t}\left[\left\|v_\theta\left(\mathbf{x}_t, t, \mathbf{c}\right) - \mathbf{u}_t\right\|^2\right]. \tag{10}$$

Because the latent shape is fixed, the same velocity model is used across all supported Hamiltonian families with different lattice geometries and system sizes.

Decoder-aware reconstruction objective. The isotropic latent-space flow error does not necessarily correspond to a comparable error in physical observables because some latent directions may be amplified more strongly than others by the VAE decoder. Hence, we augment flow matching with a decoder-aware reconstruction loss. This leads to better observable prediction accuracy. Details of the ablation can be found in Appendix C.3.

For a sampled interpolation time t, the predicted velocity defines the one-step estimate

$$\widehat{\mathbf{x}}_1 = \mathbf{x}_t + (1-t)v_\theta\left(\mathbf{x}_t, t, \mathbf{c}\right). \tag{11}$$

We undo the latent normalization and pass the resulting latent through the frozen observable decoder

$$\widehat{\mathcal{M}} = D_\psi\left(a_z^{-1}\widehat{\mathbf{x}}_1, \mathcal{P}\right), \tag{12}$$

where $\mathcal{P}$ are the query measurement operators. We supervise $\widehat{\mathcal{M}}$ with the same reconstruction loss as Equation 5. The final generative training objective is

$$\mathcal{L}_{\text{gen}} = \mathcal{L}_{\text{flow}} + \lambda_{\text{rec}}\mathcal{L}_{\text{rec}}. \tag{13}$$

Gradients from $\mathcal{L}_{\text{rec}}$ are propagated through the frozen decoder with respect to its latent input, but the decoder parameters themselves are not updated. The auxiliary term therefore teaches the latent flow model which latent errors are physically consequential while preserving the observable representation learned during VAE pretraining.

Conditional generation. At inference time, we sample noise $\mathbf{x}(0) \sim \mathcal{N}(\mathbf{0}, \sigma_0^2\mathbf{I})$ and solve the conditional probability-flow ordinary differential equation

$$\frac{\mathrm{d}\mathbf{x}(t)}{\mathrm{d}t} = v_\theta\left(\mathbf{x}(t), t, \mathbf{c}\right), \qquad t \in [0, 1]. \tag{14}$$

The terminal latent is transformed back to the VAE scale and decoded at the given queries by $\widetilde{\mathcal{M}} = D_\psi\left(a_z^{-1}\mathbf{x}(1), \mathcal{P}\right)$. Thus, given only the physical condition including Hamiltonian identifier, coefficients, and lattice specification, the model generates a fixed-size latent representation, which can be queried into a complete set of local measurement expectation value predictions.

4 Experiments

4.1 Observable Reconstruction and Conditional Generation

We first train a single unified model on five Hamiltonian families across three system sizes. The Hamiltonian families include the 1D and 2D transverse-field Ising models (TFIM), the 1D cluster-Ising model, the 1D anisotropic Heisenberg XXZ model, and the 2D J_1–J_2 model. Experimental configurations and implementation details are described in Appendix A and B. During training, we use the full set of local observables with measurement expectation values. For VAE inference, only 100 randomly selected **incomplete** observables are used as the input. We evaluate our framework for both reconstruction and generation, which estimates a complete set of local Pauli observables. In the observable reconstruction setting, the VAE receives the local measurement data and reconstructs the

Table 1: Observable reconstruction and conditional generation RMSE with standard deviation across all test states, reported in units of 10^{-3}. Best is highlighted in **bold**.

		Observable reconstruction				Conditional generation	
Family	Qubits	Classical shadow	GQNQ	LLM4QPE	Ours	QuaDiM	Ours
1D TFIM	9	17.517 ± 0.448	74.643 ± 25.138	19.949 ± 6.680	$\mathbf{5.058 \pm 2.855}$	60.905 ± 26.751	$\mathbf{5.674 \pm 3.048}$
	36	17.454 ± 0.215	76.409 ± 24.270	16.951 ± 4.212	$\mathbf{4.012 \pm 1.929}$	56.977 ± 26.629	$\mathbf{6.009 \pm 3.409}$
	100	17.325 ± 0.137	75.612 ± 22.672	16.694 ± 3.098	$\mathbf{5.525 \pm 3.503}$	52.747 ± 23.141	$\mathbf{7.022 \pm 4.536}$
2D TFIM	9 (3×3)	17.254 ± 0.248	82.948 ± 36.792	20.784 ± 6.730	$\mathbf{4.592 \pm 0.858}$	55.658 ± 14.640	$\mathbf{6.044 \pm 1.298}$
	36 (6×6)	17.459 ± 0.126	79.644 ± 29.500	22.268 ± 7.989	$\mathbf{7.002 \pm 1.232}$	64.952 ± 28.616	$\mathbf{8.803 \pm 2.221}$
	100 (10×10)	17.488 ± 0.089	94.440 ± 29.605	23.594 ± 7.257	$\mathbf{8.155 \pm 2.185}$	61.747 ± 19.817	$\mathbf{11.335 \pm 3.908}$
1D XXZ	9	17.571 ± 0.354	91.940 ± 16.940	58.852 ± 19.940	$\mathbf{6.044 \pm 2.712}$	113.931 ± 57.153	$\mathbf{34.511 \pm 27.077}$
	36	17.331 ± 0.200	64.923 ± 7.983	31.215 ± 8.442	$\mathbf{10.200 \pm 7.267}$	63.294 ± 14.706	$\mathbf{13.667 \pm 10.275}$
	100	$\mathbf{17.454 \pm 0.133}$	55.027 ± 16.019	29.116 ± 8.471	20.564 ± 13.285	64.924 ± 29.037	$\mathbf{27.300 \pm 24.621}$
Cluster-Ising	9	27.750 ± 0.487	136.961 ± 67.279	27.361 ± 8.777	$\mathbf{11.667 \pm 9.089}$	63.527 ± 30.165	$\mathbf{15.597 \pm 8.798}$
	36	27.818 ± 0.157	139.882 ± 64.608	30.962 ± 16.279	$\mathbf{17.083 \pm 18.583}$	68.334 ± 27.830	$\mathbf{25.626 \pm 19.947}$
	100	27.777 ± 0.081	137.716 ± 63.313	29.924 ± 11.302	$\mathbf{14.542 \pm 13.110}$	75.828 ± 33.745	$\mathbf{28.463 \pm 24.476}$
2D J_1–J_2	9 (3×3)	21.049 ± 0.218	178.969 ± 30.305	29.772 ± 4.590	$\mathbf{2.173 \pm 0.973}$	200.418 ± 35.271	$\mathbf{8.584 \pm 10.915}$
	36 (6×6)	21.215 ± 0.173	191.683 ± 14.346	31.678 ± 3.740	$\mathbf{8.535 \pm 4.451}$	205.078 ± 13.371	$\mathbf{14.886 \pm 18.403}$
	100 (10×10)	$\mathbf{21.107 \pm 0.091}$	196.865 ± 23.675	51.256 ± 28.812	36.853 ± 16.380	204.887 ± 22.529	$\mathbf{50.830 \pm 40.865}$
Overall		20.238 ± 4.042	111.844 ± 58.870	29.358 ± 16.400	$\mathbf{10.800 \pm 12.259}$	94.214 ± 63.271	$\mathbf{17.623 \pm 21.793}$

queried complete set of local observables through the fixed-size latent bottleneck. In the conditional-generation setting, the model receives only the Hamiltonian and lattice condition and generates a latent representation that is mapped to observables by the frozen VAE decoder.

We compute the root mean squared error (RMSE) with standard deviation between the predicted and ground-truth measurement expectation values of observables across 300 held-out states from different Hamiltonian families and system sizes. The baseline methods we compare against for reconstruction are GQNQ (Zhu et al., 2022), LLM4QPE (Tang et al., 2024a), and classical shadows (Huang et al., 2020) using their best evaluated experiment settings. For generation, we compare against QuaDiM (Tang et al., 2025). All baselines are trained separately for each experiment setting.

Table 1 shows the reconstruction and generation results. Note that classical shadow receives the same set of complete local measurement operators acting on the same support as the query operators. For observable reconstruction, our framework achieves the best overall RMSE. Across the systems, the unified fixed-size bottleneck accurately preserves local observables despite substantial variation in lattice geometry, observable layout, and input dimensionality. For conditional generation, our framework outperforms QuaDiM in every setting, with an 81.3% reduction in overall RMSE. This suggests that generating a latent state representation and subsequently decoding it into observables can be substantially more effective than directly modeling system-dependent measurement records, supporting the advantage of latent-space generative modeling for quantum state learning.

4.2 Few-Shot Transfer Learning to Unseen Hamiltonian Families

We next investigate whether the representation learned by the unified QuMUSE model can be efficiently adapted to Hamiltonian families outside the pretraining distribution. This setting evaluates the reuse of learned representations across changes in interaction structure and lattice geometry, extending beyond prediction at unseen coupling values within familiar Hamiltonian families. We consider three target systems with long-range interactions or unseen lattice geometry: triangular-lattice antiferromagnetic transverse-field Ising model, long-range Ising model, and Kitaev honeycomb model. Details of these Hamiltonians are introduced in Appendix A.4. For each target system, we use 6 randomly sampled states for transfer learning. Evaluation is performed on 20 held-out states of disjoint parameter points. We compare three initialization strategies:

- **Zero-shot:** directly evaluate the pretrained framework without target-system adaptation.
- **Transfer:** fine-tune the pretrained VAE and SiT with few-shot adaptation in 2500 steps.
- **Scratch:** train the same architecture on the same target states and for the same number of optimization steps, but with random initialization.

We evaluate the observable reconstruction and the conditional generation performance on the target Hamiltonian families. As presented in Table 2, our framework shows some zero-shot transferabil-

Table 2: Few-shot transfer to unseen Hamiltonian families. We compute RMSE $\pm$ standard deviation across all test states, in units of 10^{-3}. Zero-shot: Evaluate the pretrained framework without fine-tuning. Transfer: Data-efficient transfer to target families using 6 states. Scratch: Train from scratch using the same computational budget as the transfer learning. Best is highlighted in **bold**.

	Observable reconstruction			Conditional generation		
Target system	Zero-shot	Transfer	Scratch	Zero-shot	Transfer	Scratch
Triangular Ising	44.699 ± 45.625	$\mathbf{4.023 \pm 2.478}$	6.308 ± 3.997	129.337 ± 90.470	$\mathbf{7.717 \pm 9.688}$	10.372 ± 11.852
Long-range Ising	33.769 ± 3.155	$\mathbf{3.043 \pm 0.473}$	9.364 ± 3.850	73.707 ± 43.259	$\mathbf{2.998 \pm 0.527}$	9.266 ± 3.784
Kitaev honeycomb	184.425 ± 120.890	$\mathbf{5.254 \pm 2.384}$	8.017 ± 3.669	461.627 ± 45.545	9.362 ± 10.295	$\mathbf{8.343 \pm 3.803}$

Table 3: Quantum property estimation RMSE with standard deviation, evaluated on five properties. We additionally compute mean Normalized RMSE (NRMSE) and mean R^2 averaging the corresponding metrics over the prediction of all five Hamiltonian families. Best is highlighted in **bold**.

Method	Magnetization	Correlation	Rényi-2 mutual information	Cluster string order	Half-system Rényi-2 entropy	Mean NRMSE $\downarrow$	Mean $R^2 \uparrow$
GQNQ	0.20387 ± 0.01477	0.08720 ± 0.00547	0.08150 ± 0.01084	0.07622 ± 0.01152	0.19139 ± 0.01323	0.3282	0.8846
Classical shadow	0.08806 ± 0.01738	0.06256 ± 0.00336	0.10364 ± 0.01123	0.14458 ± 0.01783	0.21468 ± 0.01381	0.3408	0.8630
LLM4QPE	0.05614 ± 0.00952	0.02033 ± 0.00213	0.06658 ± 0.00822	0.09510 ± 0.01922	0.13057 ± 0.01001	0.2035	0.9478
AiDE-Q	0.05597 ± 0.00913	0.02033 ± 0.00211	0.06692 ± 0.00808	0.08560 ± 0.01649	0.13072 ± 0.01025	0.1983	0.9505
NTK	0.25181 ± 0.01471	0.02593 ± 0.00343	$\mathbf{0.05477 \pm 0.01466}$	0.24377 ± 0.04537	0.18869 ± 0.03110	0.3804	0.8081
QuaDiM	0.22793 ± 0.01448	0.06969 ± 0.00263	0.13221 ± 0.01069	0.40330 ± 0.07438	0.28921 ± 0.02735	0.6022	0.5304
Ours	$\mathbf{0.05408 \pm 0.00933}$	$\mathbf{0.00565 \pm 0.00111}$	0.05605 ± 0.00809	$\mathbf{0.02175 \pm 0.00400}$	$\mathbf{0.10346 \pm 0.01193}$	**0.1292**	**0.9749**

ity to unseen systems, but still requires fine-tuning for accurate estimation. Initializing from the unified model for fine-tuning substantially improves data-efficient adaptation. Relative to zero-shot evaluation, few-shot adaptation outperforms in both reconstruction and generation for every target, demonstrating that only a small number of target states is required to specialize the pretrained model. Moreover, compared with training from scratch under the same training budget, transfer initialization consistently improves reconstruction performance. This indicates that the pretrained framework provides a useful initial representation even when the target lattice or coupling structure was not present during pretraining.

5 Applications

5.1 Quantum Property Estimation via Linear Probing

We evaluate whether the pretrained representation captures physical information beyond the local observables used during representation learning. Specifically, we freeze each state encoder and train an independent mapping for properties of interest. The properties we consider are detailed in Appendix A.5, including local, global, linear and nonlinear properties. The mapping operates directly on the state latent representations. Given a frozen vectorized representation $\mathbf{r}$ for the quantum state and an ansatz function f, the prediction for property p can be formulated as a regression task:

$$\widehat{y}^{(p)} = f(\mathbf{r}). \tag{15}$$

Here we consider the simplest ansatz for prediction to demonstrate the faithfulness of the pretrained representation. We use the linear probing $\widehat{y}^{(p)} = \mathbf{w}_p^\top \mathbf{r} + b_p$, where only $\mathbf{w}_p$ and b_p are parameters to be optimized. Through this formulation, the fixed latent space is augmented so it can predict global and nonlinear properties even though the representations are trained from only local observables. We train each property on 480 states and evaluate RMSE on 2000 unseen states in total.

We compare our results against baseline methods including GQNQ (Zhu et al., 2022), Classical shadow (Huang et al., 2020), LLM4QPE (Tang et al., 2024a), AiDE-Q (Wang et al., 2025), NTK (Huang et al., 2022) and QuaDiM (Tang et al., 2025). Table 3 shows the property-estimation results. The learning model used in AiDE-Q is the same as LLM4QPE. Individual properties use RMSE as metric, and we additionally compute the overall Normalized RMSE (NRMSE) and R^2 score. Our framework achieves the best overall performance among all methods. The improvement is observed not only for the local linear quantities magnetization and correlation function, but also for the nonlinear half-system entropy and for the many-body cluster string order. Thus, although our framework is pretrained from local observables, its latent representation organizes these measurements in a form from which both nonlinear and nonlocal properties can be recovered using only a

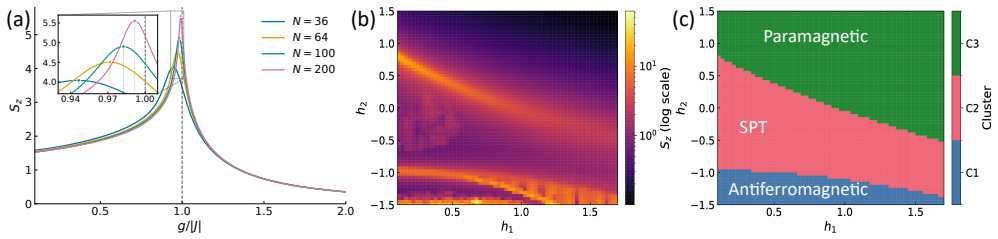

Figure 2: Unsupervised witness of the phase boundary using the learned state representations. The phase indicator S_z suggests a phase boundary. (a) Changes in state representations w.r.t. 1D Ising coefficients g. (b) Changes in state representations w.r.t. cluster-Ising coefficients h_1 and h_2. (c) Clustering applied to state representations recovers the phase diagram for the cluster-Ising model.

linear prediction head. This avoids additional computation overhead of retraining a new base model for predicting unseen new quantum properties.

5.2 Unsupervised Witnesses of Quantum Phase Boundaries

We finally examine whether the pretrained representation can reveal quantum phase boundaries without using phase labels. Using the trained conditional generative model, we generate state representations directly from Hamiltonian and lattice specifications, avoiding additional state preparation and measurement at the queried coefficient points. We quantify changes in these representations using a phase indicator S_z and apply unsupervised clustering to identify distinct regions of the coefficient space. As defined in Appendix A.6, S_z measures how rapidly the latent representation changes with Hamiltonian coefficients, suggesting possible phase transitions. Figure 2 demonstrates this application for the 1D transverse-field Ising and cluster-Ising models. Without using phase labels for supervision (Cong et al., 2019; Liu et al., 2023) or fine-tuning the representation with phase-aware properties (Wu et al., 2024), our analysis reuses the pretrained representations. These results suggest that pretraining on local observables captures physical structure relevant to phase organization, while conditional generation enables this structure to be explored without acquiring additional measurements during the coefficient scan.

1D Ising model. Figure 2(a) shows S_z as a function of $g/|J|$ of Ising model with 4 system sizes N. The indicator develops a pronounced peak near the transition between the ferromagnetic and paramagnetic phases. As the system size increases, the peak becomes sharper and moves toward the thermodynamic critical point $g/|J| = 1$, indicating that the representation captures increasingly rapid changes in the ground state near criticality. Analogous shifts of effective critical points defined by entanglement-entropy maxima have been established for finite Ising chains (Iglói & Lin, 2008). Therefore, we interpret the peak as a finite-size witness of the phase transition.

Cluster-Ising model. Figures 2(b) and (c) extend this analysis to the two-dimensional parameter space (h_1, h_2). The response map in panel (b), exhibits enhanced responses along bands separating distinct regimes. Unsupervised k-means clustering (MacQueen, 1967) produces the three regions shown in panel (c), which correspond to the antiferromagnetic, symmetry-protected topological (SPT), and paramagnetic phases. This coincides with the established cluster-Ising phase diagram (Cong et al., 2019). The correspondence between the representation change and cluster boundaries shows that the pretrained features retain information relevant to phase organization, supporting their use for unsupervised phase-boundary witnessing.

6 Conclusion

We introduce QuMUSE, a foundational generative pretraining framework for quantum many-body ground-state representations from local measurements. Our framework enables observable reconstruction and conditional generation through a unified, fixed-size representation. Experiments demonstrate accurate observable prediction with one model spanning different Hamiltonian families and system sizes. The pretrained representations also support data-efficient transfer to unseen Hamiltonian families, quantum property estimation, and unsupervised phase-boundary witnessing.

AI USE STATEMENT

In this work, we used generative AI tools to assist in identifying relevant literature, improving the clarity and readability of the manuscript, drafting and editing portions of the paper, formatting tables, assisting coding the method and experiments, supporting qualitative and quantitative data analysis. We have not used AI to generate synthetic data sets, develop theoretical models or conceptual frameworks, formulate mathematical claims, propose or refine hypotheses, design or provide feedback on research methodology or experiments, clean or reformat dataset. We have reviewed all AI-assisted writing for accuracy, verified the identified references, and cross-checked the numerical values in the formatted tables against our experimental results. Also, we have checked the correctness of the AI-generated codes and data analysis. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We have included the experimental configurations in Appendix A, the implementation and training details of the model in Appendix B. We will release the source code for ground states simulation, model training and inference upon acceptance to facilitate reproducibility and further research in quantum state learning.

REFERENCES

Shahnawaz Ahmed, Carlos Sánchez Muñoz, Franco Nori, and Anton Frisk Kockum. Quantum state tomography with conditional generative adversarial networks. *Physical Review Letters*, 127(14), 2021. ISSN 1079-7114. doi: 10.1103/physrevlett.127.140502. URL http://dx.doi.org/10.1103/PhysRevLett.127.140502.

Ahmed A. Akhtar, Hong-Ye Hu, and Yi-Zhuang You. Scalable and flexible classical shadow tomography with tensor networks. *Quantum*, 7:1026, 2023. ISSN 2521-327X. doi: 10.22331/q-2023-06-01-1026. URL http://dx.doi.org/10.22331/q-2023-06-01-1026.

Michael S. Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants, 2023. URL https://arxiv.org/abs/2209.15571.

Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning, 2022. URL https://arxiv.org/abs/2105.04906.

Giuseppe Carleo and Matthias Troyer. Solving the quantum many-body problem with artificial neural networks. *Science*, 355(6325):602–606, 2017. doi: 10.1126/science.aag2302. URL https://www.science.org/doi/abs/10.1126/science.aag2302.

Juan Carrasquilla, Giacomo Torlai, Roger G. Melko, and Leandro Aolita. Reconstructing quantum states with generative models. *Nature Machine Intelligence*, 1(3):155–161, March 2019. ISSN 2522-5839. doi: 10.1038/s42256-019-0028-1. URL http://dx.doi.org/10.1038/s42256-019-0028-1.

Peter Cha, Paul Ginsparg, Felix Wu, Juan Carrasquilla, Peter L McMahon, and Eun-Ah Kim. Attention-based quantum tomography. *Machine Learning: Science and Technology*, 3(1): 01LT01, November 2021. ISSN 2632-2153. doi: 10.1088/2632-2153/ac362b. URL http://dx.doi.org/10.1088/2632-2153/ac362b.

Ao Chen and Markus Heyl. Empowering deep neural quantum states through efficient optimization. *Nature Physics*, 20(9):1476–1481, 2024. ISSN 1745-2481. doi: 10.1038/s41567-024-02566-1. URL http://dx.doi.org/10.1038/s41567-024-02566-1.

Iris Cong, Soonwon Choi, and Mikhail D. Lukin. Quantum convolutional neural networks. *Nature Physics*, 15(12):1273–1278, August 2019. ISSN 1745-2481. doi: 10.1038/s41567-019-0648-8. URL http://dx.doi.org/10.1038/s41567-019-0648-8.

Yuxuan Du, Tao Huang, Shan You, Min-Hsiu Hsieh, and Dacheng Tao. Quantum circuit architecture search for variational quantum algorithms. *npj Quantum Information*, 8(1), May 2022. ISSN

2056-6387. doi: 10.1038/s41534-022-00570-y. URL http://dx.doi.org/10.1038/s41534-022-00570-y.

William Falcon and The PyTorch Lightning team. PyTorch Lightning, March 2019. URL https://github.com/Lightning-AI/lightning.

Irene López Gutiérrez and Christian B. Mendl. Real time evolution with neural-network quantum states. *Quantum*, 6:627, January 2022. ISSN 2521-327X. doi: 10.22331/q-2022-01-20-627. URL http://dx.doi.org/10.22331/q-2022-01-20-627.

P. Hauke and L. Tagliacozzo. Spread of correlations in long-range interacting quantum systems. *Physical Review Letters*, 111(20), November 2013. ISSN 1079-7114. doi: 10.1103/physrevlett.111.207202. URL http://dx.doi.org/10.1103/PhysRevLett.111.207202.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020. URL https://arxiv.org/abs/2006.11239.

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. ISSN 1745-2481. doi: 10.1038/s41567-020-0932-7. URL http://dx.doi.org/10.1038/s41567-020-0932-7.

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), 2022. ISSN 1095-9203. doi: 10.1126/science.abk3333. URL http://dx.doi.org/10.1126/science.abk3333.

Jiaxin Huang, Yan Zhu, Giulio Chiribella, and Ya-Dong Wu. Sequence-model-guided measurement selection for quantum state learning, 2025. URL https://arxiv.org/abs/2507.09891.

Ferenc Iglói and Yu-Cheng Lin. Finite-size scaling of the entanglement entropy of the quantum ising chain with homogeneous, periodically modulated and random couplings. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(06):P06004, 2008. ISSN 1742-5468. doi: 10.1088/1742-5468/2008/06/p06004. URL http://dx.doi.org/10.1088/1742-5468/2008/06/P06004.

Abhinav Kandala, Antonio Mezzacapo, Kristan Temme, Maika Takita, Markus Brink, Jerry M. Chow, and Jay M. Gambetta. Hardware-efficient variational quantum eigensolver for small molecules and quantum magnets. *Nature*, 549(7671):242–246, 2017. ISSN 1476-4687. doi: 10.1038/nature23879. URL http://dx.doi.org/10.1038/nature23879.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Alexei Kitaev. Anyons in an exactly solved model and beyond. *Annals of Physics*, 321(1):2–111, January 2006. ISSN 0003-4916. doi: 10.1016/j.aop.2005.10.005. URL http://dx.doi.org/10.1016/j.aop.2005.10.005.

Laura Lewis, Hsin-Yuan Huang, Viet T. Tran, Sebastian Lehner, Richard Kueng, and John Preskill. Improved machine learning algorithm for predicting ground state properties. *Nature Communications*, 15(1), January 2024. ISSN 2041-1723. doi: 10.1038/s41467-024-45014-7. URL http://dx.doi.org/10.1038/s41467-024-45014-7.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023. URL https://arxiv.org/abs/2301.12597.

Manwen Liao, Yan Zhu, Weitian Zhang, and Yuxiang Yang. Reinforced learning explicit circuit representations for quantum state characterization from local measurements. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 37623–37641. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/liao25j.html.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling, 2023. URL https://arxiv.org/abs/2210.02747.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow, 2022. URL https://arxiv.org/abs/2209.03003.

Yu-Jie Liu, Adam Smith, Michael Knap, and Frank Pollmann. Model-independent learning of quantum phases of matter with quantum convolutional neural networks. *Phys. Rev. Lett.*, 130:220603, Jun 2023. doi: 10.1103/PhysRevLett.130.220603. URL https://link.aps.org/doi/10.1103/PhysRevLett.130.220603.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. URL https://arxiv.org/abs/1711.05101.

Nanye Ma, Mark Goldstein, Michael S. Albergo, Nicholas M. Boffi, Eric Vanden-Eijnden, and Saining Xie. Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers, 2024. URL https://arxiv.org/abs/2401.08740.

Yixuan Ma, Chang Liu, Weikang Li, Shun-Yao Zhang, L. M. Duan, Yukai Wu, and Dong-Ling Deng. Solving excited states for long-range interacting trapped ions with neural networks, 2025. URL https://arxiv.org/abs/2506.08594.

J. B. MacQueen. Some methods for classification and analysis of multivariate observations. In L. M. Le Cam and J. Neyman (eds.), *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pp. 281–297. University of California Press, 1967.

Michael A. Nielsen and Isaac L. Chuang. *Quantum Computation and Quantum Information*. Cambridge University Press, Cambridge, UK, 2010.

D. Perez-Garcia, F. Verstraete, M. M. Wolf, and J. I. Cirac. Matrix product state representations. *Quantum Info. Comput.*, 7(5):401–430, July 2007. ISSN 1533-7146.

David Pfau, James S. Spencer, Alexander G. D. G. Matthews, and W. M. C. Foulkes. Ab initio solution of the many-electron schrödinger equation with deep neural networks. *Phys. Rev. Res.*, 2:033429, Sep 2020. doi: 10.1103/PhysRevResearch.2.033429. URL https://link.aps.org/doi/10.1103/PhysRevResearch.2.033429.

David Pfau, Simon Axelrod, Halvard Sutterud, Ingrid von Glehn, and James S. Spencer. Accurate computation of quantum excited states with neural networks. *Science*, 385(6711), August 2024. ISSN 1095-9203. doi: 10.1126/science.adn0137. URL http://dx.doi.org/10.1126/science.adn0137.

Zihao Qi, Christopher Earls, and Yang Peng. Neural operator quantum state: A foundation model for quantum dynamics, 2026. URL https://arxiv.org/abs/2603.25066.

Xiangjian Qian and Mingpu Qin. Absence of spin liquid phase in the J_1–J_2 heisenberg model on the square lattice. *Physical Review B*, 109(16), April 2024. ISSN 2469-9969. doi: 10.1103/physrevb.109.l161103. URL http://dx.doi.org/10.1103/PhysRevB.109.L161103.

Riccardo Rende, Luciano Loris Viteritti, Federico Becca, Antonello Scardicchio, Alessandro Laio, and Giuseppe Carleo. Foundation neural-networks quantum states as a unified ansatz for multiple hamiltonians. *Nature Communications*, 16(1), August 2025. ISSN 2041-1723. doi: 10.1038/s41467-025-62098-x. URL http://dx.doi.org/10.1038/s41467-025-62098-x.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. URL https://arxiv.org/abs/2112.10752.

Tobias Schmale, Moritz Reh, and Martin Gärttner. Efficient quantum state tomography with convolutional neural networks. *npj Quantum Information*, 8(1), 2022. ISSN 2056-6387. doi: 10.1038/s41534-022-00621-4. URL http://dx.doi.org/10.1038/s41534-022-00621-4.

Markus Schmitt and Markus Heyl. Quantum many-body dynamics in two dimensions with artificial neural networks. *Physical Review Letters*, 125(10), 2020. ISSN 1079-7114. doi: 10.1103/physrevlett.125.100503. URL http://dx.doi.org/10.1103/PhysRevLett.125.100503.

Or Sharir, Yoav Levine, Noam Wies, Giuseppe Carleo, and Amnon Shashua. Deep autoregressive models for the efficient variational simulation of many-body quantum systems. *Physical Review Letters*, 124(2), January 2020. ISSN 1079-7114. doi: 10.1103/physrevlett.124.020503. URL http://dx.doi.org/10.1103/PhysRevLett.124.020503.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding, 2023. URL https://arxiv.org/abs/2104.09864.

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

Yehui Tang, Nianzu Yang, Mabiao Long, and Junchi Yan. Ssl4q: semi-supervised learning of quantum data with application to quantum state classification. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024b.

Yehui Tang, Mabiao Long, and Junchi Yan. Quadim: A conditional diffusion model for quantum state property estimation. In *The Thirteenth International Conference on Learning Representations*, 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(5):447–450, February 2018. ISSN 1745-2481. doi: 10.1038/s41567-018-0048-5. URL http://dx.doi.org/10.1038/s41567-018-0048-5.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Haoxiang Wang, Maurice Weber, Josh Izaac, and Cedric Yen-Yu Lin. Predicting properties of quantum systems with conditional generative models, 2024. URL https://arxiv.org/abs/2211.16943.

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, 2025. URL https://arxiv.org/abs/2509.26109.

Matteo M. Wauters, Emanuele Panizon, Glen B. Mbeng, and Giuseppe E. Santoro. Reinforcement-learning-assisted quantum optimization. *Physical Review Research*, 2(3), 2020. ISSN 2643-1564. doi: 10.1103/physrevresearch.2.033446. URL http://dx.doi.org/10.1103/PhysRevResearch.2.033446.

Victor Wei, W. A. Coish, Pooya Ronagh, and Christine A. Muschik. Neural-shadow quantum state tomography. *Physical Review Research*, 6(2), 2024. ISSN 2643-1564. doi: 10.1103/physrevresearch.6.023250. URL http://dx.doi.org/10.1103/PhysRevResearch.6.023250.

Steven R. White. Density matrix formulation for quantum renormalization groups. *Phys. Rev. Lett.*, 69:2863–2866, Nov 1992. doi: 10.1103/PhysRevLett.69.2863. URL https://link.aps.org/doi/10.1103/PhysRevLett.69.2863.

Wenjie Wu, Ge Yan, Xudong Lu, Kaisen Pan, and Junchi Yan. Quantumdarts: differentiable quantum architecture search for variational quantum algorithms. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Ya-Dong Wu, Yan Zhu, Yuexuan Wang, and Giulio Chiribella. Learning quantum properties from short-range correlations using multi-task networks. *Nature Communications*, 15(1):8796, 2024.

Timothy Zaklama, Daniele Guerci, and Liang Fu. Attention-based foundation model for quantum states, 2026. URL https://arxiv.org/abs/2512.11962.

Shang-Shun Zhang, Gábor B. Halász, and Cristian D. Batista. Theory of the kitaev model in a [111] magnetic field. *Nature Communications*, 13(1), January 2022. ISSN 2041-1723. doi: 10.1038/s41467-022-28014-3. URL http://dx.doi.org/10.1038/s41467-022-28014-3.

Yuan-Hang Zhang and Massimiliano Di Ventra. Transformer quantum state: A multipurpose model for quantum many-body problems. *Physical Review B*, 107(7), February 2023. ISSN 2469-9969. doi: 10.1103/physrevb.107.075147. URL http://dx.doi.org/10.1103/PhysRevB.107.075147.

Yan Zhu, Ya-Dong Wu, Ge Bai, Dong-Sheng Wang, Yuexuan Wang, and Giulio Chiribella. Flexible learning of quantum states with generative query neural networks. *Nature Communications*, 13 (1), October 2022. ISSN 2041-1723. doi: 10.1038/s41467-022-33928-z. URL http://dx.doi.org/10.1038/s41467-022-33928-z.

Yuchen Zhu, Tianrong Chen, Evangelos A. Theodorou, Xie Chen, and Molei Tao. Quantum state generation with structure-preserving diffusion model, 2024. URL https://arxiv.org/abs/2404.06336.

A Hamiltonians and Experimental Configuration

A.1 Hamiltonian Families of Interest

We consider quantum systems with open boundary conditions. Throughout this section, X_i, Y_i, and Z_i denote Pauli operators acting on qubit (site) i. The joint QuMUSE model is trained on five Hamiltonian families spanning one- and two-dimensional lattices and three system sizes. The Hamiltonians we consider during training are as follows.

One-dimensional transverse-field Ising model. The one-dimensional transverse-field Ising model is

$$H^{\mathrm{1D}}_{\mathrm{TFIM}}(g) = J \sum_{i=1}^{N-1} Z_i Z_{i+1} + g \sum_{i=1}^{N} X_i, \tag{16}$$

where $J = -1$ and $g \in [-3, 3]$. We consider chains with $N \in \{9, 36, 100\}$ qubits.

Two-dimensional transverse-field Ising model. On an $L \times L$ square lattice, the Hamiltonian is

$$H^{\mathrm{2D}}_{\mathrm{TFIM}}(g) = J \sum_{\langle i,j \rangle} Z_i Z_j + g \sum_{i=1}^{L^2} X_i, \tag{17}$$

where $\langle i, j \rangle$ denotes nearest-neighbor lattice bonds, $J = 1$, and $g \in [-4, 4]$. We use $L \in \{3, 6, 10\}$, corresponding to 9, 36, and 100 qubits.

One-dimensional Anisotropic Heisenberg XXZ model. The anisotropic Heisenberg model is

$$H_{\mathrm{XXZ}}(J_z) = J \sum_{i=1}^{N-1} (X_i X_{i+1} + Y_i Y_{i+1}) + J_z \sum_{i=1}^{N-1} Z_i Z_{i+1}, \tag{18}$$

where $J = -1$, $J_z \in [-2, 2]$, and $N \in \{9, 36, 100\}$.

One-dimensional cluster-Ising model. The cluster-Ising Hamiltonian is

$$\begin{aligned} H_{\mathrm{CI}}(h_1, h_2) = &- \sum_{i=1}^{N-2} Z_i X_{i+1} Z_{i+2} - h_1 \sum_{i=1}^{N} X_i \\ &- h_2 \sum_{i=1}^{N-1} X_i X_{i+1}, \end{aligned} \tag{19}$$

where $(h_1, h_2) \in [-2, 2]^2$ and $N \in \{9, 36, 100\}$.

Two-dimensional J_1–J_2 model. The square-lattice J_1–J_2 Hamiltonian is

$$H_{J_1\text{-}J_2} = J_1 \sum_{\langle i,j \rangle} \boldsymbol{\sigma}_i \cdot \boldsymbol{\sigma}_j + J_2 \sum_{\langle\langle i,j \rangle\rangle} \boldsymbol{\sigma}_i \cdot \boldsymbol{\sigma}_j, \tag{20}$$

where

$$\boldsymbol{\sigma}_i \cdot \boldsymbol{\sigma}_j = X_i X_j + Y_i Y_j + Z_i Z_j. \tag{21}$$

The first sum runs over nearest-neighbor horizontal and vertical bonds, while the second sum represents the exchange interaction to the next nearest-neighbor. We set $J_1 = 1$, vary $J_2 \in [0, 1]$, and use $L \in \{3, 6, 10\}$, corresponding to 9, 36, and 100 qubits.

A.2 Ground State Preparation

To construct the ground states, we use matrix-product states (MPS) (Perez-Garcia et al., 2007) to represent quantum states and optimize them using density matrix renormalization group (DMRG) (White, 1992). The data-generation pipeline uses a maximum MPS bond dimension of

Table 4: Local-observable layouts used by QuMUSE. Local support represents the local sites the measurement operators act on. Observables per support indicates the types of local operators for each Hamiltonian family and the total number of them.

Hamiltonian	Local support	Observables per support
1D TFIM	Nearest-neighbor pair	$\{I, X, Y, Z\}^{\otimes 2}$; 16
2D TFIM	Nearest-neighbor bond	$\{I, X, Y, Z\}^{\otimes 2}$; 16
1D XXZ	Nearest-neighbor pair	$\{I, X, Y, Z\}^{\otimes 2}$; 16
Cluster-Ising	Contiguous three-site window	$\{I, X, Y, Z\}^{\otimes 3}$; 64
2D J_1–J_2	Nearest or diagonal bond	$\{X, Y, Z\}^{\otimes 2}$; 9

64, the maximum dimension of Krylov space of 10, and an energy-change stopping tolerance of 10^{-6}. We have verified that the bond dimension is large enough for accurate representing the studied ground states, and increasing bond dimension further does not yield better ground state energy. For the 2D Hamiltonians datasets, DMRG continuation is performed in both increasing and decreasing directions of the second coefficient, i.e., J_2 for the J_1–J_2 model and g for the 2D Ising model.

A.3 Local-Observable Construction

For a local support $\mathcal{S}$ that defines a volume of qubits to be measured, and a measurement operator $P_A(\mathcal{S})$ for measurement, we compute the measurement outcome

$$m_{A,\mathcal{S}} = \langle\psi|P_A(\mathcal{S})|\psi\rangle\,. \tag{22}$$

These values are computed by efficient tensor-network contractions. The measurement outcomes (observables) are stored together with their support position specifications S_j without a pre-defined order to form the measurement data. Throughout the experiments, we consider local Pauli operators as the measurement operators, whose expectation values can be estimated using single-qubit basis rotations followed by computational-basis measurements (Kandala et al., 2017; Huang et al., 2020). This makes them experimentally accessible on quantum devices. Other measurement operators can be decomposed into a weighted sum of Pauli operators (Nielsen & Chuang, 2010). Table 4 shows the details of the local observable layout used for each Hamiltonian family in the experiments.

A.4 Few-Shot Transfer Hamiltonians

In section 4.2 of the main text, we additionally evaluate adaptation to three quantum systems whose lattice geometry or interaction structure is absent from joint training. The Hamiltonians considered are given below.

Triangular-lattice transverse-field Ising model. The first target is the open triangular lattice with antiferromagnetic Ising interactions:

$$H_{\triangle}(g) = J_{\triangle} \sum_{\langle i,j\rangle\in\mathcal{E}_{\triangle}} Z_i Z_j + g \sum_i X_i, \qquad J_{\triangle} > 0, \tag{23}$$

where $\mathcal{E}_{\triangle}$ contains the open-boundary triangular-lattice edges, $J_{\triangle} = 1$ and $g \in [0.2, 3.2]$. Compared to the 2D square lattice, this Hamiltonian has additional connectivity at the diagonal of each grid. We set the system size to 3×3. This target is used to demonstrate the transferability of our framework to Hamiltonian of changed interaction graph from the training data.

Long-range Ising model. The second target is the open chain Ising model with long range interactions (Hauke & Tagliacozzo, 2013):

$$H_{\mathrm{LR}}(\alpha) = \frac{J_{\mathrm{LR}}}{\mathcal{N}_\alpha} \sum_{1\le i<j\le N} \frac{Z_i Z_j}{|i-j|^\alpha} + g_{\mathrm{LR}} \sum_{i=1}^{N} X_i, \tag{24}$$

where $\mathcal{N}_\alpha$ is the Kac normalization used by the data-generation implementation and $J_{\mathrm{LR}} = g_{\mathrm{LR}} = 1$, $\alpha \in [0.5, 3.5]$. This target is used to demonstrate the transferability of our framework to Hamiltonian of changed coupling range from the training data.

Kitaev honeycomb model. The third target is the open honeycomb model with bond-dependent interactions (Kitaev, 2006):

$$\begin{aligned} H_{\mathrm{K}}(J_z) =& J_x \sum_{\langle i,j\rangle_x} X_i X_j + J_y \sum_{\langle i,j\rangle_y} Y_i Y_j + J_z \sum_{\langle i,j\rangle_z} Z_i Z_j \\ &+ \frac{h_{111}}{\sqrt{3}} \sum_i \left(X_i + Y_i + Z_i\right), \end{aligned} \tag{25}$$

where the three sums run over the corresponding x-, y-, and z-type honeycomb bonds. We use $J_x = J_y = 1$, vary $J_z \in [0.4, 2.0]$, and set $h_{111} = 0.05$ and the system size to 12. The weak field along the $[111]$ direction breaks time-reversal symmetry, opens a topological gap in the Majorana fermion spectrum, and drives transitions into exotic spin-liquid phases (Zhang et al., 2022).

A.5 Quantum Properties of Interest in the Linear Probing Application

For the quantum property estimation application in Section 5.1, we consider the following properties, including local, global, linear, and nonlinear quantities.

The magnetization and the correlation function are

$$m_x = \frac{1}{N} \sum_{i=1}^{N} \langle X_i \rangle, \tag{26}$$

and

$$C_{\mathrm{nn}} = \frac{1}{|\mathcal{E}|} \sum_{(i,j)\in\mathcal{E}} \frac{\langle X_i X_j \rangle + \langle Y_i Y_j \rangle + \langle Z_i Z_j \rangle}{3}, \tag{27}$$

where $\mathcal{E}$ denotes the set of physical nearest-neighbor edges. We additionally predict the nearest-neighbor Rényi-2 mutual information

$$I_2(i:j) = S_2(i) + S_2(j) - S_2(ij), \tag{28}$$

where $S_2(i)$ is the second-order Rényi entropy on the reduced density matrix of qubit site i. We average over all $(i, j) \in \mathcal{E}$, where $\mathcal{E}$ denotes the connectivity. For two-dimensional systems, $\mathcal{E}$ contains Manhattan-neighbor bonds, and $S_2(A)$ is evaluated at the central cut of the saved row-major snake ordering.

We also consider the cluster string order O_{ZXZ}

$$O_{ZXZ} = \left\langle Z_0 \left(\prod_k X_{2k+1} \right) Z_{N-1} \right\rangle. \tag{29}$$

The cluster string order is a many-body product-of-stabilizers string with Z endpoints and alternating interior X operators. Finally, we consider the half-system Rényi-2 entropy

$$S_2(A) = -\log \sum_k \lambda_k^4, \tag{30}$$

where $\{\lambda_k\}$ are the Schmidt coefficients at the central matrix-product-state cut.

A.6 Specification of the Phase Indicator

In Section 5.2, we apply a phase indicator on the learned state representations for witness of quantum phase boundaries. The indicator S_z quantifies how rapidly the VAE state representation changes as the Hamiltonian coefficients vary. Let $z(\boldsymbol{\theta})$ denote the state representation associated with the ground state at coefficients $\boldsymbol{\theta}$. We compute the indicator from neighboring representations in a parameter scan, without using phase labels. Note that the parameter scan is done by the generative model without obtaining and measuring the actual ground states.

Representation normalization. For a scan containing M coefficient points, let $z_i = z(\boldsymbol{\theta}_i)$. We compute the mean representation and the root-mean-square distance from this mean over the full scan:

$$\bar{z} = \frac{1}{M}\sum_{i=1}^{M} z_i, \qquad r = \left(\frac{1}{M}\sum_{i=1}^{M}\|z_i - \bar{z}\|_2^2\right)^{1/2}. \tag{31}$$

The normalized representations are

$$\tilde{z}_i = \frac{z_i - \bar{z}}{r}. \tag{32}$$

This normalization uses a single scalar r shared by all latent coordinates, so that the mean squared norm of the centered, normalized representations is one. For the Ising comparison across system sizes, both $\bar{z}$ and r are computed separately for each size using its full field scan. For cluster-Ising, they are computed over the full two-dimensional coefficient grid.

One-dimensional Ising model. We evaluate successive latent vectors on a uniform field grid with spacing $\Delta g = 0.001$. The indicator is assigned to the midpoint of each adjacent pair:

$$S_z\left(\frac{g_i + g_{i+1}}{2}\right) = \frac{\|\tilde{z}(g_{i+1}) - \tilde{z}(g_i)\|_2}{g_{i+1} - g_i}, \qquad i = 1, \ldots, M-1. \tag{33}$$

This finite-difference score approximates $\|d\tilde{z}/dg\|_2$. With $|J| = 1$, the field spacing also equals the spacing in the plotted coordinate $g/|J|$.

Cluster-Ising model. For $\mathbf{h} = (h_1, h_2)$, we combine changes along both coefficient directions. Let $\mathcal{N}_\alpha(\mathbf{h})$ contain the immediately adjacent grid points that differ from $\mathbf{h}$ only in coordinate h_α. Thus, $\mathcal{N}_1$ contains the available left and right neighbors, while $\mathcal{N}_2$ contains the available lower and upper neighbors. We define

$$S_z(\mathbf{h}) = \left[\sum_{\alpha=1}^{2} \frac{1}{|\mathcal{N}_\alpha(\mathbf{h})|} \sum_{\mathbf{h}' \in \mathcal{N}_\alpha(\mathbf{h})} \frac{\|\tilde{z}(\mathbf{h}') - \tilde{z}(\mathbf{h})\|_2^2}{|h'_\alpha - h_\alpha|^2}\right]^{1/2}. \tag{34}$$

For each direction, we average the squared finite-difference norms over the available neighbors, then sum the two directional contributions and take the square root. At a grid boundary, only the available neighbor in the corresponding direction is used. For a smooth representation and sufficiently small grid spacings, this score approximates the Frobenius norm of the latent Jacobian:

$$S_z(\mathbf{h}) \approx \left(\left\|\frac{\partial \tilde{z}}{\partial h_1}\right\|_2^2 + \left\|\frac{\partial \tilde{z}}{\partial h_2}\right\|_2^2\right)^{1/2}. \tag{35}$$

Large values of S_z indicate rapid changes in the normalized representation, with peaks or ridges serving as candidate phase-boundary witnesses.

B Implementation Details

B.1 Physical Condition Representation

Here we specify the physical condition representation, serving as input to the additional branch of the VAE encoder as well as the conditional input to the generative model. The physical system can be described by a fixed-dimensional vector

$$\mathbf{c} = [f, \boldsymbol{\theta}, \boldsymbol{\ell}], \tag{36}$$

where f is a Hamiltonian-family identifier, $\boldsymbol{\theta}$ contains Hamiltonian coefficients, and

$$\boldsymbol{\ell} = [d, n_{\mathrm{row}}, n_{\mathrm{col}}, N] \tag{37}$$

describes the lattice geometry, where d specifies the dimension, n_{row} and n_{col} represent the number of rows and columns for the Hamiltonian sites, and N represents the system size.

The coefficient encodings are

$$\begin{aligned} \boldsymbol{\theta}_{\text{1D TFIM}} &= [g, 0, 0], \\ \boldsymbol{\theta}_{\text{2D TFIM}} &= [J, g, 0], \\ \boldsymbol{\theta}_{\text{XXZ}} &= [-1, J_z, 0], \\ \boldsymbol{\theta}_{\text{CI}} &= [h_1, h_2, 0], \\ \boldsymbol{\theta}_{J_1\text{-}J_2} &= [J_1, J_2, 0]. \end{aligned} \tag{38}$$

For one-dimensional systems, the lattice vector is $\boldsymbol{\ell} = [1, 1, N, N]$. For square lattices, it is $\boldsymbol{\ell} = [2, L, L, L^2]$. The condition encoder is composed of a sequence of linear projection, multilayer perceptron, and LayerNorm, which map the normalized condition to a physical system representation of 128 dimension. During the VAE training stage, this representation is aligned with the state representation after mapping to the same dimension via an additional linear layer.

B.2 Model Configuration

Variational autoencoder. Throughout the experiments, we set the number of learnable queries to 32. The unified VAE uses a latent tensor of shape 32×16. The encoder is composed of 4 transformer blocks and the decoder is composed of 8 transformer blocks. We set the hidden size to 768 and the number of heads to 8. For the loss function, we set $\lambda_{\text{cond}} = 0.1$. To penalize the distribution of the latent, we set $\lambda_{\text{KL}} = 10^{-5}$.

Conditional Scalable Interpolant Transformer. The SiT operates on the 32×16 VAE latent tensor. Each latent token is projected to width 256, and the denoiser contains 6 transformer blocks with 8 attention heads. Fixed positional embeddings are applied to the 32 latent positions, while a sinusoidal embedding encodes the interpolation time.

The VAE latents are normalized using one scalar inverse standard deviation estimated over the training set. The standard deviation is lower bounded by 0.05. No mean subtraction or channel-wise scaling is applied.

B.3 Optimization and Data Splits

All models are implemented with PyTorch Lightning (Falcon & The PyTorch Lightning team, 2019). We use batch size of 16, float32 numerical precision, and gradient-norm clipping at 1.0. Both stages of VAE and SiT use linear learning-rate warmup followed by cosine decay.

VAE training. The VAE is optimized with AdamW (Loshchilov & Hutter, 2019) using

$$(\beta_1, \beta_2) = (0.9, 0.99), \qquad \epsilon = 10^{-6}, \qquad \text{weight decay} = 10^{-2}. \tag{39}$$

The training uses a learning rate of 2×10^{-4} for $16{,}000$ optimizer steps, with 100 warmup steps and a final cosine-decay factor of 10^{-3}. KL regularization, posterior sampling noise, and conditional density alignment are included between steps $6{,}000$ and $9{,}000$. Backpropagation from the physical condition representation into the VAE backbone is activated between steps $6{,}000$ and $12{,}000$.

SiT training. The SiT is optimized with AdamW using

$$(\beta_1, \beta_2) = (0.9, 0.99), \qquad \epsilon = 10^{-6}, \qquad \text{weight decay} = 10^{-3}. \tag{40}$$

We use a learning rate of 10^{-4}, 100 warmup steps, cosine decay to a factor of 10^{-3}, and $10{,}000$ optimizer steps. The VAE remains fixed, while the SiT condition encoder is trained jointly with the denoiser.

Table 5: Robustness to finite measurement shots. RMSE for reconstruction and generation is reported as mean ± standard deviation across test states. Raw denotes the RMSE of direct finite-shot measurement estimates.

	Observable RMSE		
Shots	Raw	Reconstruction	Generation
Exact	–	0.010800 ± 0.012259	0.017623 ± 0.021793
128	0.081878	0.055308 ± 0.019518	0.035330 ± 0.021330
256	0.058045	0.033192 ± 0.013014	0.028824 ± 0.033778
512	0.040885	0.020337 ± 0.010379	0.021495 ± 0.021871
1024	0.028993	0.014000 ± 0.010620	0.022020 ± 0.031689
2048	0.020487	0.010952 ± 0.011160	0.018825 ± 0.025275

Joint-training split. The joint experiment contains all 15 combinations of five Hamiltonian families and three system sizes. Each family–size cell contributes 80 training states, and 20 test states. The resulting split contains 1,200 training states and 300 test states. All splits are disjoint, and all family–size cells receive equal sampling weight during joint training.

B.4 Inference

For VAE inference, the input is 100 randomly selected local observables. This randomization is applied to both measurement operator type and the specific sites (local support) the operator acts on.

For SiT inference, at generation time, the learned velocity ordinary differential equation is integrated from the $\mathcal{N}(\mathbf{0}, 0.01\mathbf{I})$ prior using 50 Euler steps. For each physical condition, the primary generation metric averages the decoded observable vectors from eight independent latent samples.

C Ablation Study

C.1 Robustness to Finite Measurement Shots

We evaluate the robustness of our framework to finite-shot measurement noise. The experiment settings are mostly the same as the experiment section in the main text. One difference is that, for each randomly selected operator and local support, we measure the quantum state with finite number of shots. Finite-shot models are trained on the noisy measurements, while exact expectation values are reserved for final evaluation. At test time, reconstruction uses finite-shot measurements, whereas conditional generation receives only the Hamiltonian and lattice condition and averages eight decoded samples per state.

Table 5 shows that the VAE consistently reduces observable error relative to the raw measurement estimates, demonstrating denoising despite training on noisy targets. Reconstruction RMSE decreases monotonically as the shot budget increases. Conditional-generation RMSE also decreases overall, although the trend is not strictly monotonic. At 2048 number of measurement shots, both tasks approach the exact-expectation reference. These results indicate that the learned representation can suppress measurement noise and support accurate conditional generation even when training data are obtained from finite-shot measurements.

C.2 Physical-Condition Alignment During VAE Training

We investigate the role of physical-condition alignment by comparing VAE training with and without $\mathcal{L}_{\mathrm{cond}}$. We evaluate both observable reconstruction from measurements and conditional generation from physical system specifications.

Table 6 reveals a trade-off between the two tasks: removing alignment improves reconstruction accuracy but increases conditional-generation error. Conversely, including alignment improves generation at the cost of higher reconstruction error. This is consistent with alignment encouraging a latent representation that is easier to predict from physical conditions, even when it does not

Table 6: Effect of physical-condition alignment during VAE training.

VAE training	Reconstruction RMSE	Generation RMSE
w/o alignment	**0.008680**	0.019533
w/ alignment	0.010800	**0.017623**

Table 7: Conditional-generation comparison with and without auxiliary reconstruction loss.

Latent flow training objective	Generation RMSE
$\mathcal{L}_{\text{flow}} + \lambda_{\text{rec}}\mathcal{L}_{\text{rec}}$	**0.017623**
$\mathcal{L}_{\text{flow}}$	0.021575

minimize measurement reconstruction error. Thus the role of alignment is to prepare the learned representation for conditional generation, rather than improving reconstruction alone.

C.3 Auxiliary Reconstruction Loss During Latent Flow Training

We compare conditional generation using the combined objective $\mathcal{L}_{\text{flow}} + \lambda_{\text{rec}}\mathcal{L}_{\text{rec}}$ with flow matching alone ($\lambda_{\text{rec}} = 0$). We report the result of joint training with the same configuration as the experiment presented in the main text. RMSE is computed over the same test states and average eight decoded samples per state. As shown in Table 7, the model trained with the combined objective achieves lower RMSE than the flow-only variant. This suggests that the auxiliary reconstruction loss during latent flow training helps achieve better performance for generation, even if the parameters of the VAE decoder are frozen.

C.4 Number of Latent Tokens

We investigate the impact of the number of latent tokens K on reconstruction and generation performance. Note that the number of latent tokens also equals the number of learnable queries in the VAE encoder. This parameter controls the capacity of the state representation and the number of tokens modeled by the conditional latent flow network.

As shown in Table 8, larger token budgets generally improve observable reconstruction, with the largest tested representation achieving the lowest reconstruction error. Conditional-generation error, however, varies nonmonotonically with token count, and the default configuration achieves the lowest reported generation RMSE. Thus, increasing representation capacity and improving reconstruction do not necessarily improve conditional generation. From the results, $K = 32$ is the best for generation, while larger representations offer improved reconstruction accuracy.

C.5 Effectiveness of the Latent Flow Generative Stage

Our framework uses a conditional latent flow model to generate state representations from Hamiltonian conditions. A simpler alternative is direct interpolation in the VAE latent space. To assess the contribution of the latent flow model, we compare it with an interpolation baseline that constructs a latent representation for a target Hamiltonian condition from reference latent vectors and predicts local observables through the VAE decoder. Both approaches are evaluated under the same experimental settings as in the main text.

Table 9 shows that conditional generation on average achieves better performance compared to the interpolation baseline. Especially for cluster-Ising, the RMSE reduces 76.6–88.0% across the three system sizes. Interpolation remains competitive in several settings, including the Ising models, and achieves lower mean error in some cases. However, its substantially larger errors show that the tested interpolation scheme does not adequately capture the variation of ground-state observables in this family. These results support the benefit of learning a conditional generator beyond directly interpolating the encoded states.

Table 8: Effect of the number of latent tokens. The default configuration is used in the experiments of the main text.

Latent tokens K	Reconstruction RMSE	Generation RMSE
8	0.011979	0.020985
16	0.011032	0.022091
32 (default)	0.010800	**0.017623**
64	0.007286	0.019001
128	**0.005920**	0.019115

Table 9: Conditional generation versus direct interpolation in the VAE latent space. Observable RMSE is reported in units of 10^{-3} as mean $\pm$ standard deviation across test states.

		RMSE	
Hamiltonian	Qubits	Interpolation	Generation
1D TFIM	9	9.949 ± 1.676	$\mathbf{5.674 \pm 3.048}$
	36	7.284 ± 1.915	$\mathbf{6.009 \pm 3.409}$
	100	7.892 ± 3.835	$\mathbf{7.022 \pm 4.536}$
2D TFIM	9	8.206 ± 0.846	$\mathbf{6.044 \pm 1.298}$
	36	$\mathbf{8.715 \pm 1.085}$	8.803 ± 2.221
	100	$\mathbf{9.713 \pm 2.163}$	11.335 ± 3.908
XXZ	9	50.970 ± 32.535	$\mathbf{34.511 \pm 27.077}$
	36	14.891 ± 8.493	$\mathbf{13.667 \pm 10.275}$
	100	$\mathbf{25.802 \pm 16.725}$	27.300 ± 24.621
Cluster-Ising	9	130.190 ± 27.517	$\mathbf{15.597 \pm 8.798}$
	36	123.753 ± 32.393	$\mathbf{25.626 \pm 19.947}$
	100	121.863 ± 37.306	$\mathbf{28.463 \pm 24.476}$
J_1–J_2	9	10.276 ± 0.552	$\mathbf{8.584 \pm 10.915}$
	36	24.407 ± 37.366	$\mathbf{14.886 \pm 18.403}$
	100	$\mathbf{41.527 \pm 19.797}$	50.830 ± 40.865
Overall	–	39.696 ± 49.171	$\mathbf{17.623 \pm 21.793}$

Moreover, in situations where measurement is not available to the ground states, interpolation cannot be achieved because the state representations from the nearby Hamiltonian coefficients cannot be constructed. Nevertheless, for generative framework, once trained, the conditional generator produces state representations directly from Hamiltonian conditions without accessing reference-state representations at inference time.

D Zero-shot Transfer to Unseen System Sizes

We examine the model's performance on zero-shot transfer to unseen system sizes. The unified model is pretrained on systems containing 9, 36, and 100 qubits. We evaluate it directly on systems with 25, 49, and 81 qubits across the same five Hamiltonian families. For two-dimensional systems, these correspond to 5×5, 7×7, and 9×9 lattices. The model processes the resulting observable collections using its shared encoder and decoder while retaining the same latent representation size.

Table 10 shows that zero-shot accuracy varies substantially across Hamiltonian families. Both reconstruction and conditional generation achieve their lowest errors on the Ising families, indicating effective transfer to unseen chain lengths and lattice dimensions in these settings. Errors are higher for XXZ, Cluster-Ising, and J_1–J_2 systems. The results demonstrate that the shared representation supports zero-shot evaluation at unseen intermediate sizes, while also revealing that accurate transfer across system sizes remains dependent on the Hamiltonian family.

Table 10: Zero-shot performance on unseen system sizes. Results are averaged across 25, 49 and 81 qubits.

Hamiltonian	Reconstruction RMSE	Generation RMSE
1D TFIM	0.004412 ± 0.002228	0.006658 ± 0.004071
2D TFIM	0.011697 ± 0.002291	0.013579 ± 0.006610
1D XXZ	0.028655 ± 0.018894	0.032447 ± 0.024660
Cluster-Ising	0.048751 ± 0.061647	0.054393 ± 0.059764
2D J_1–J_2	0.080412 ± 0.032953	0.072760 ± 0.041803

E Visualization of the Learned State Representations

We examine the physical structure captured by the learned state representations through t-SNE (Van der Maaten & Hinton, 2008) visualizations of 36-qubit systems in Figure 3. (a)–(e) show individual Hamiltonian families, colored by the values of the Hamiltonian coefficients, while (f) presents a joint view colored by Hamiltonian family. The visualization exhibits systematic variation with physical parameters, together with compact groups and distinct branches in several families. We now analyze and provide some physical interpretations for each of the visualization results.

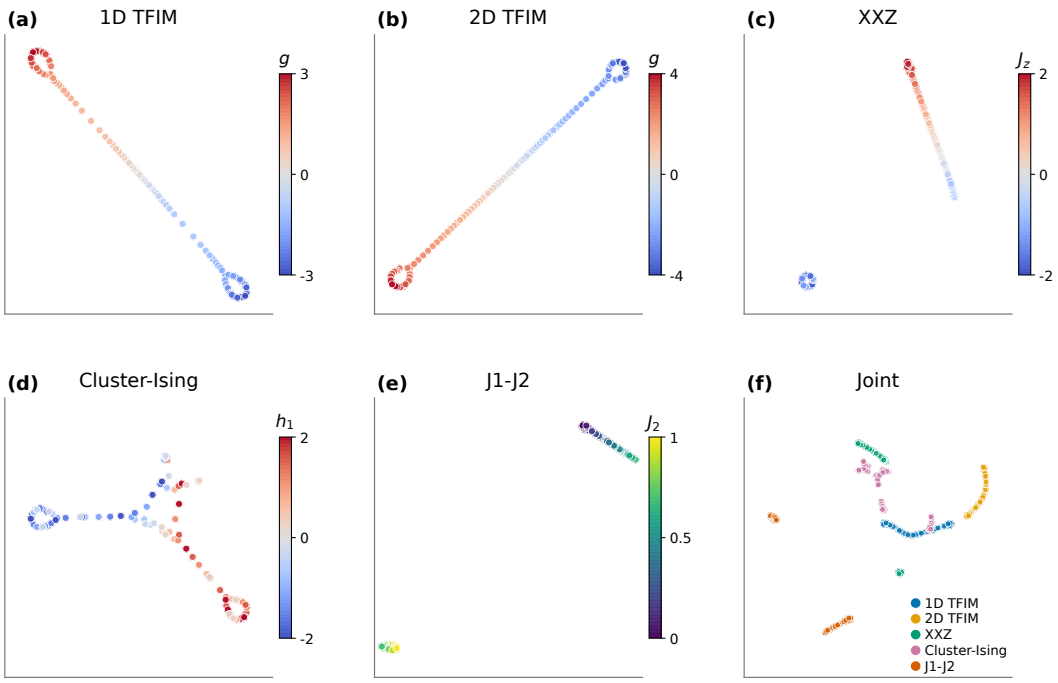

Figure 3: t-SNE visualization of the learned state representations.

Transverse-field Ising models. The 1D and 2D TFIM representations form extended curves with a clear color progression as the transverse field changes. This organization is consistent with the evolution of local magnetization and spin correlations as the balance between Ising interactions and the transverse field varies. The compact end regions are compatible with polarization saturation: for the convention $H = J\sum_{\langle i,j\rangle} Z_iZ_j + g\sum_i X_i$, sufficiently large $|g|$ drives the ground state toward X polarization with $\langle X_i\rangle \simeq -\operatorname{sgn}(g)$, making local observables less sensitive to further changes in field strength.

Anisotropic Heisenberg XXZ model. The XXZ visualization contains a compact group at the most negative J_z values, separated from a branch covering the remaining parameter range. Under our convention $J = -1$, the regime $J_z < -1$ admits fully z-polarized ferromagnetic ground states.

Their local observables are largely insensitive to J_z within this regime once a common symmetry representative is chosen, providing a physical explanation for similar representations in the compact group. The extended branch is consistent with changing longitudinal and transverse correlations outside this regime.

Cluster-Ising model. The Cluster-Ising representations exhibit several branches and a more heterogeneous arrangement. This family depends on both h_1 and h_2, where we we color the points by h_1. Consequently, states with similar colors can have different local correlations because their second coupling differs. With the change of coefficients, the competition among the three-spin cluster interaction, transverse field, and nearest-neighbor XX interaction supports distinct correlation regimes, including a symmetry-protected topological cluster regime. The branching behavior of the visualization result is compatible with this richer dependence on physical parameters.

J_1–J_2 model. The visualization result illustrates a separation of a branch at smaller and intermediate J_2 from a compact group at larger J_2. Physically, increasing the antiferromagnetic coupling introduces frustration and changes the competition between Néel and stripe correlations (Qian & Qin, 2024). The observed separation is therefore compatible with a reorganization of local bond correlations across these regimes.

Joint Hamiltonian families. The t-SNE result on the joint Hamiltonian families retains substantial organization by Hamiltonian family, while cluster-Ising representations occupy several regions, including locations near the Ising representations. This suggests that the shared latent space retains both family-dependent structure and variation within each family. Family separation could also reflect differences in observable layouts and the physical-condition alignment used during training. Therefore, we can interpret the visualization as qualitative evidence of physically structured representations.

F Limitations and Future Work

Our current evaluation remains limited to selected ground-state families, and generalization varies across physical regimes. Future work will expand the diversity and scale of pretraining data, strengthen robustness to experimental noise and distribution shifts, and investigate extensions to finite-temperature states and quantum dynamics.