

Multimodal discrete diffusion for quantum measurement generation in the one-dimensional transverse-field Ising model

YanHong Li  and Ho-Kin Tang ^{*}

*School of Science, Harbin Institute of Technology, Shenzhen 518055, China
and Shenzhen Key Laboratory of Advanced Functional Carbon Materials Research
and Comprehensive Application, Shenzhen 518055, China*



(Received 16 February 2026; accepted 17 August 2026; published 31 August 2026)

Diffusion-based large language models have recently attracted significant interest as a promising class of nonautoregressive generative models. However, their applications to quantum many-body systems remain largely unexplored. We introduce M-Diffushadow, a multimodal diffusion language model for quantum many-body systems. In M-Diffushadow, we adopt a full attention mechanism alongside iterative decoding, which better aligns with the physical characteristics of quantum systems and thereby enables a more faithful modeling of such systems. As a multimodal approach, M-Diffushadow enables bidirectional cross-modal inference, reconstruction from partially observed multimodal records, and fully joint generation without prior measurement outcomes, aligning naturally with realistic experimental constraints where measurement access is limited or incomplete. We demonstrate the effectiveness of the framework on the one-dimensional transverse-field Ising model (TFIM), where generated data accurately reproduce ground-state energies and correlation functions across different phases. We further provide an additional numerical example on the XXZ spin chain to support the applicability of M-Diffushadow beyond the original TFIM benchmark. M-Diffushadow shows the potential to capture and analyze the complex quantum phenomena in quantum many-body systems.

DOI: [10.1103/4tft-d7gf](https://doi.org/10.1103/4tft-d7gf)

I. INTRODUCTION

Quantum many-body systems play a central role in modern physics, providing the foundation for a wide range of emergent phenomena, from high-temperature superconductivity to quantum magnetism and quantum information processing [1–4]. Understanding how collective behavior arises from microscopic quantum interactions remains a fundamental challenge, particularly in regimes where strong correlations and entanglement dominate.

Recent advances in quantum engineering have enabled unprecedented experimental access to large-scale quantum many-body systems through programmable quantum simulators and near-term quantum devices [5,6]. However, the exponential growth of the Hilbert space with system size renders a direct description of these systems intractable for classical computation. As a consequence, experimental studies of quantum many-body systems necessarily rely on indirect access through measurement processes rather than explicit wavefunction representations. Among existing

^{*}Contact author: denghaojian@hit.edu.cn

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

approaches, classical-shadow (CS) tomography [7] has emerged as a powerful framework for efficiently estimating physical observables from randomized measurements. By significantly reducing the measurement cost for a wide class of observables, classical shadows have enabled scalable characterization of quantum systems that would otherwise be inaccessible. Nevertheless, acquiring sufficiently rich experimental data remains resource-intensive, and classical-shadow methods are typically designed for estimating specific observables rather than modeling the full distribution of measurement outcomes [8]. More broadly, experimental data in quantum many-body systems take the form of discrete and stochastic measurement records obtained under specific measurement protocols. Such data are inherently heterogeneous: Different measurement modalities probe complementary aspects of the same underlying quantum state while exhibiting distinct statistical structures. These characteristics pose a fundamental challenge for unified data-driven modeling, reuse, and generation of quantum measurement data across different experimental settings.

From a broader perspective, these challenges suggest that the central difficulty lies not only in estimating individual observables, but in modeling the underlying distribution of quantum measurement data itself. Recent years have witnessed increasing interest in data-driven approaches for studying complex physical systems, driven by the growing availability of experimental and numerical data [9–11]. Unlike traditional analytical or simulation-based methods, machine learning provides a flexible framework for extracting statistical structure directly from data, without requiring an explicit low-dimensional parametrization of the underlying physical state [12]. It has proven particularly powerful in quantum many-body physics, where learning the full probability distribution over measurement outcomes enables not only state reconstruction but also the discovery of emergent patterns and phase transitions [13].

In particular, modern generative models aim to learn the full probability distribution of high-dimensional data [14,15], enabling not only prediction but also sampling of new data instances consistent with the learned distribution. This perspective is especially relevant for quantum many-body experiments, where accessible information is naturally expressed in the form of stochastic measurement records rather than explicit wavefunctions. Such measurement data are discrete, high dimensional, and intrinsically probabilistic, closely resembling structured sequences encountered in natural language and other symbolic domains [16]. Language-model-inspired architectures have therefore emerged as a natural candidate for modeling quantum measurement data, treating measurement outcomes as sequences of discrete tokens conditioned on experimental settings [17]. From this viewpoint, different measurement modalities correspond to distinct but correlated data channels probing the same underlying quantum system. A generative model capable of learning joint and conditional distributions over these heterogeneous measurement records would provide a powerful tool for data-efficient characterization and synthesis of quantum experimental outcomes.

Within the class of generative models, the diffusion language model has arisen as a promising paradigm recently [18,19] and offers a particularly natural abstraction for modeling complex, high-dimensional scientific data distributions. Rather than directly generating data in a single step, diffusion models learn to transform noise into structured samples through a sequence of stochastic refinement steps, providing a flexible and stable framework for probabilistic generation [20,21]. This paradigm is especially well matched to quantum measurement data. Measurement records are inherently discrete, stochastic, and conditioned on experimental settings, and their statistical structure can be naturally captured by a gradual noising-and-denoising process. By operating directly in discrete state spaces, discrete diffusion models enable principled modeling of symbolic measurement outcomes while preserving their probabilistic interpretation. These properties make diffusion-based generative modeling a suitable foundation for learning and synthesizing quantum experimental data across different measurement modalities.

Motivated by these considerations, we introduce *M-Diffushadow*, a multimodal discrete diffusion framework for modeling and generating quantum measurement data. Our approach learns joint and conditional probability distributions over heterogeneous measurement modalities, including classical shadow measurements and nearest-neighbor spin-product outcomes, while conditioning on experimental parameters. By operating directly on discrete measurement records, *M-Diffushadow*

provides a unified generative description of quantum experimental data without relying on explicit wavefunction representations. Since the measurement results serve only as intermediate quantities from which diverse physical properties can be derived or further processed, M-Diffushadow focuses on generating these results and delegates subsequent and task-specific operations to downstream physical analysis. This strategy eliminates the need to train an individual model for each task; instead, one model supports multiple applications, such as estimating ground-state energy and correlation functions. Crucially, once trained, the model acts as a practical data surrogate and rapidly generates large volumes of measurement samples on demand, eliminating the need for time-consuming and resource-intensive measurements on actual quantum hardware. Moreover, M-Diffushadow, once trained on a sparse set of g values, can generate measurement samples for arbitrary g values without additional experiments. This makes M-Diffushadow a practical data surrogate that significantly accelerates the data acquisition process. At the model level, M-Diffushadow is built upon an iterative denoising paradigm, in which quantum measurement records are progressively refined from noisy inputs toward physically consistent samples. An architecture with full attention [22] is also employed to capture nonlocal correlations across qubits by allowing for bidirectional information flow and to integrate information from different measurement modalities within a shared representation space. This design is intended for flexible conditioning and masking, allowing the model to seamlessly switch among joint generation, cross-modal inference, and partial reconstruction modes.

To our knowledge, this work represents one of the first applications of discrete diffusion-based generative modeling to quantum many-body measurement data, and the first to explicitly address multimodal quantum measurement generation within a unified framework. This design is specifically tailored to address practical challenges in experimental settings, such as being restricted to a single data modality by equipment limitations or cost, and data incompleteness arising from equipment failure or partial measurements. To tackle these issues, M-Diffushadow supports cross-modal conditional generation, partial measurement reconstruction under incomplete observations, and joint multimodal generation without prior measurement outcomes, providing a versatile approach to heterogeneous quantum data modeling. We demonstrate the effectiveness of M-Diffushadow on the one-dimensional transverse-field Ising model (TFIM), where generated data accurately reproduce ground-state properties and correlation functions across different phases. These results highlight the potential of multimodal generative diffusion modeling as a data-efficient tool for quantum many-body characterization and synthesis.

The remainder of this article is organized as follows. Section II A introduces the physical background and quantum measurement data considered in this work. In Sec. II B, we give a brief introduction of the generative models related to our work, with particular emphasis on the discrete diffusion model. In Sec. III, we first provide a detailed description of the distinctive architectural design of M-Diffushadow. These structural characteristics are precisely what endow the model with a strong suitability for quantum measurement modeling and lay the foundation for its subsequent training and generation processes. The training and generation details are then provided in Secs. III B and III C, respectively, where we describe how the model learns from noisy measurement data and produces physically consistent samples. In Sec. IV, we report numerical experiments and results of different generating modes of M-Diffushadow, followed by discussions. Experiments on the XXZ model are also included to further evaluate the applicability of M-Diffushadow beyond the original TFIM benchmark.

II. PRELIMINARIES

A. Physical settings and data

We will first briefly summarize the physical setting and notation used throughout this work.

We mainly consider a transverse-field Ising model, where the transverse-field strength g serves as a global control parameter. The one-dimensional TFIM is a paradigmatic quantum many-body

system that has been extensively studied in condensed matter physics [1,23]. Owing to its rich physical behavior and conceptual simplicity, it provides a natural setting for investigating quantum phase transitions and correlated spin dynamics.

We consider an N -qubit chain with periodic boundary conditions, described by the Hamiltonian

$$H = - \sum_{i=1}^N (1 - g) Z_i Z_{i+1} - g \sum_{i=1}^N X_i, \quad (1)$$

where Z_i and X_i denote Pauli operators acting on the i th qubit and $g \in [0, 1]$ is the dimensionless transverse-field strength.

The model captures the competition between two fundamental physical mechanisms: The nearest-neighbor ZZ interaction favors ferromagnetic ordering, while the transverse field induces quantum fluctuations that suppress spin alignment. This competition gives rise to a quantum phase transition at the critical point $g_c = 0.5$, separating a ferromagnetic phase for $g < g_c$ from a paramagnetic phase for $g > g_c$ [24].

We primarily consider classical shadows and nearest-neighbor spin products (SPs), which constitute the two modalities studied by our model. The first modality is based on the classical-shadow protocol [7,25], which has been shown to provide an efficient and scalable representation of quantum states through randomized measurements. For each experimental shot, the classical-shadow data are represented as a triplet $(g, \mathbf{P}, \mathbf{b})$, where g denotes the transverse-field strength of the Hamiltonian, $\mathbf{P} = (p_1, \dots, p_N)$ is a sequence of randomly sampled local Pauli measurement bases with $p_j \in \{X, Y, Z\}$, and $\mathbf{b} = (b_1, \dots, b_N)$ is the corresponding measurement outcome sequence with $b_j \in \{0, 1\}$. This modality captures local measurement information conditioned on both the Hamiltonian parameter and the chosen measurement bases.

In addition to classical shadows, we introduce a second modality consisting of nearest-neighbor spin-product measurements. For each nearest-neighbor pair indexed by j with $(j = 1, 2, \dots, N)$, this modality records the measurement bases $(p_j^{(1)}, p_j^{(2)})$ applied to adjacent spins together with the corresponding spin-product outcome d_j . For notational convenience, we denote the paired measurement bases sequence $(p_j^{(1)}, p_j^{(2)})$ by $\mathbf{P}_c$, in analogy to the single-site basis sequence $\mathbf{P}$ in the classical-shadow modality, and denote the associated spin-product outcomes by $\mathbf{d}$. Such observables directly probe short-range correlations in the system and remain local and experimentally accessible. Beyond providing sensitivity to interaction-driven correlations [26,27], this “native” SP modality offers a structurally distinct perspective that complements the randomized snapshots of classical shadows. M-Diffushadow is designed to learn the cross-modal relationships between such disparate measurement records. By treating each modality as an abstract token sequence, the architecture remains modality agnostic, allowing it to perform cross-modal inference and reconstruct full physical pictures from incomplete experimental access. This design serves as a representative proof of concept that is naturally extensible to other measurement combinations or Hamiltonians.

B. Generative modeling

The measurement outcomes obtained from quantum many-body systems are inherently stochastic, even when the underlying Hamiltonian parameters are fixed. As a result, learning to model such systems naturally calls for a probabilistic generative framework that captures the full distribution of measurement data, rather than deterministic predictions of specific observables [9,12].

In this work, we adopt a generative modeling perspective and aim to learn the joint distribution of experimentally accessible measurement outcomes conditioned on the Hamiltonian parameter and measurement settings. This formulation allows the model to represent the intrinsic randomness of quantum measurements and to generate consistent samples that reflect the statistical structure of the underlying quantum state. On the other hand, the measurement bases and corresponding outcomes considered in our setting take values in discrete and finite sets. This naturally motivates

a formulation in a discrete state space, which allows for a physically faithful representation of quantum measurement data.

Discrete diffusion models provide a natural framework for generative modeling in such discrete state spaces, where data generation is cast as a progressive denoising process [21,28]. Within this paradigm, measurement outcomes are first partially corrupted through a stochastic noising procedure, and a parametrized model is trained to iteratively reconstruct the original data by conditioning on the remaining information.

In the present work, we adopt this discrete diffusion perspective to model multimodal quantum measurement data, where both classical shadow measurements and nearest-neighbor spin-product observables are treated as discrete variables and jointly generated through conditional denoising. Detailed model architectures and training procedures are described in the following section.

III. METHOD

We introduce a multimodal diffusion model named M-Diffushadow based on classical shadow tomography, focusing on classical shadow and nearest-neighbor spin product. In this section, we provide a detailed introduction to M-Diffushadow, encompassing its architectural features (Sec. III A), training paradigm (Sec. III B), and generation paradigm (Sec. III C). It is precisely these unique designs that enable it to effectively model quantum systems and make predictions regarding their physical properties.

A. The architecture of M-Diffushadow

M-Diffushadow is built on the foundation of the transformer architecture [22]. The attention mechanism is the main point of difference from the architecture of mainstream autoregressive models. Along with the well-designed structured multimodal embedding, this attention mechanism makes M-Diffushadow better suited for quantum many-body problem measurements.

A defining architectural feature of M-Diffushadow is its use of full attention [22], in contrast to the causal attention used by mainstream autoregressive models like GPT [16]. Full attention means that every token in the sequence can attend to all tokens before and after it. In contrast, causal attention only allows the current token to attend to previous tokens, without visibility to subsequent ones. The choice of full attention over causal attention is motivated by the physical structure of quantum measurement data. Unlike natural language, where tokens are generated sequentially and a causal ordering exists, quantum measurement outcomes are jointly distributed with no intrinsic causal sequence. Full attention allows each token to attend to all others without directional constraints, providing a more natural framework for modeling the nonlocal correlations inherent in such data. In contrast, causal attention imposes a unidirectional information flow that may introduce an artificial bias without a clear physical justification in the context of simultaneous quantum measurements.

The embedding module is also carefully designed to effectively capture the underlying physical laws and relationships between different modalities. We design a structured multimodal embedding layer that encodes these two heterogeneous data types into discrete symbolic sequences, which can be processed by the diffusion model. This unified representation provides a consistent embedding space for the subsequent diffusion denoising process.

For the classical-shadow modality, we design different embedding modules for each type of data. For the parameter g , we use an independent linear mapping component, since g is a real-valued parameter between 0 and 1. This linear mapping avoids unnecessary dimensional expansion, preventing confusion for M-Diffushadow, while also reducing the overall complexity. For P and b , we use two separate embedding layers to map them independently. This is because they are different types of data, each carrying distinct information. After embedding g , P , and b , the resulting embeddings g' , P' , and b' are concatenated and input into a specially designed transformer module

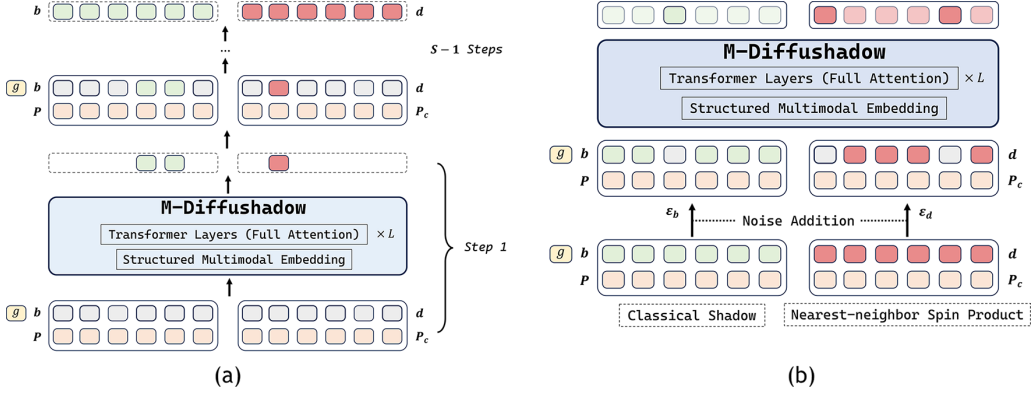


FIG. 1. The training and generating processes of M-Diffushadow. (a) The model is trained to predict the correct outcomes for the masked positions. (b) The generation process proceeds from a fully masked state to the final output through S iterative unmasking steps.

with full attention mechanisms. In addition, position encoding and type encoding are incorporated to further distinguish different types of information and enrich the input representation.

For the nearest-neighbor spin-product modality, it is crucial that its information is efficiently understood by M-Diffushadow and connected with the classical-shadow modality. In this context, we introduce a combination encoder to handle the measurement basis pairs $(p_j^{(1)}, p_j^{(2)})$ in the spin-product data. Initially, for $p_j^{(1)}$ and $p_j^{(2)}$, we use the same embedding components as for the classical-shadow modality, obtaining the measurement basis embeddings $e_j^{(1)}$ and $e_j^{(2)}$. Next, we design a trainable combination encoder to learn the distinct physical meaning of different measurement basis combinations, as shown in Eq. (2):

$$ep_j = W_2 \times \text{GELU}(W_1 \times e_j^{(1)} + e_j^{(2)}) + b_1 + b_2, \quad (2)$$

where W_1, W_2, b_1, b_2 are trainable parameters. This combination encoder allows us to retain the feature representations of the measurement basis combinations in the nearest-neighbor spin-product modality while also establishing their relationship with the classical-shadow modality. For the measurement results d_j , we apply the same embedding module as used for b of the classical-shadow modality. Finally, the various numerical embeddings from both modalities are concatenated to form a sequence that is fed into the subsequent transformer module. In addition, we incorporate position encoding and type encoding to further distinguish between different types of information and enrich the input representation. Notably, the same type encoding is used for both the classical-shadow modality's measurement basis and the adjacent spin-product modality's measurement basis combinations, emphasizing their shared data type.

B. The training of M-Diffushadow

As illustrated in Fig. 1(a), the input sequence of training data consists of three parts, including the parameter of Hamiltonian g , classical-shadow modality $(p_j, b_j)_{j=1}^N$, and nearest-neighbor spin-product modality $(p_j^{(1)}, p_j^{(2)}, d_j)_{j=1}^N$.

To train M-Diffushadow, we construct a training dataset comprising measurement records of both modalities generated from the ground states of the TFI model at eight values of the transverse-field strength $g \in \{0.0, 0.25, 0.4, 0.5, 0.6, 0.75, 0.9, 1.0\}$ from the interval $[0, 1]$. All training samples, spanning different g values, are mixed and jointly used to train M-Diffushadow. This enables the model to learn different physical regimes across different physical phases within a single unified framework.

To ensure that the model efficiently learns knowledge from both modalities and their interrelationships, while simultaneously preventing overfitting, we have designed a dynamic reshuffling data loading strategy. The dynamic reshuffling of paired multimodal data forces the model to learn the intrinsic relationships between modalities rather than memorizing specific combinations. Specifically, for each instance sharing the same parameter g , the classical shadow modality sequence and the nearest-neighbor spin-product modality sequence are randomly paired to form an input sequence: $[g, (p_j, b_j)_{j=1}^N, (p_j^{(1)}, p_j^{(2)}, d_j)_{j=1}^N]$. In the subsequent epoch, the original pairings are disrupted and randomly recombined to generate a new input sequence. By reshuffling the combinations across consecutive epochs, each specific modality pairing is guaranteed to appear only once, thereby maximizing data utilization efficiency while preventing the model from memorizing fixed pairings. In practice, the number of reshuffling epochs i is treated as a hyperparameter. This dynamic recombination strategy forces the model to establish semantic correspondences between the two modalities in the latent space. Since each pairing is randomized, the model cannot rely on any specific sample combination and must instead learn the mapping between the internal features of each modality and the shared physical state g .

After the original training data are constructed, they need to be masked to facilitate model training. While preserving the conditioning information g , $\mathbf{p}_j$, and $\mathbf{p}_j^{(i)}$, $i = 1, 2$, we employ a stochastic masking procedure that corrupts the measurement outcome sequence of both modalities $\mathbf{b}_j$ and $\mathbf{d}_j$. Specifically, for each training sample of classical-shadow modality in a batch, we sample an independent masking probability $p_{\text{mask}} \sim U[\epsilon_b, 1]$, where ϵ_b is a small hyperparameter ensuring minimal masking. We then generate a binary mask sequence $\mathbf{M} \in \{0, 1\}^N$ by independently drawing each $\mathbf{M}[i]$ from a Bernoulli distribution: $m_i \sim \text{Bernoulli}(p_{\text{mask}})$, where N is the number of qubits. And the measurement outcome sequence of classical-shadow modality with masks is constructed as

$$\mathbf{b}_j[i] = \begin{cases} [\mathbf{M}], & \text{if } \mathbf{M}[i] = 1, \\ [\mathbf{b}_0[i]], & \text{otherwise,} \end{cases} \quad (3)$$

where b_0 is the original measurement outcome of classical-shadow modality. The measurement outcome sequence of nearest-neighbor spin-product modality d_j follows the same noise addition process as that of b_j . The only difference between the two modalities lies in the independently configured hyperparameters ϵ_b and ϵ_d , which control the degree of noise applied to each modality.

The objective of training is to predict the masked measurement outcomes given g , $\mathbf{P}$, $\mathbf{P}_c$, and the rest of the measurement outcomes of both modalities. The loss function of training is presented as [Eq. (4)]

$$\mathcal{J}_t = -\mathbb{E}_{g, P, O_0, O_t} \left[\sum_{i=1}^N \mathbf{1}[O_t^i = \mathbf{M}] \ln p_\theta(O_0^i | g, P, O_t) \right]. \quad (4)$$

In Eq. (4), $\mathbf{1}[\cdot]$ denotes the indicator function, O_0 represents the measurement outcome sequence without mask, O_t represents the measurement outcome sequence with random masks, and θ represents the parameters of Diffushadow being trained. For simplicity, let P denote the measurement basis and O denote the measurement outcome, corresponding, respectively, to the measurement basis sequences and the measurement outcome sequences for both of the two modalities. Under this loss formulation, the gradient updates are exclusively computed based on the masked positions.

C. The generating of M-Diffushadow

As illustrated in Fig. 2, M-Diffushadow supports three distinct generation modes. First, M-Diffushadow supports cross-modal conditional generation, where one modality is generated conditioned on the observation of the other. Specifically, given the classical-shadow measurement outcomes b_j or the nearest-neighbor spin-product measurement outcomes d_j , M-Diffushadow infers and generates the missing modality. This capability closely aligns with realistic experimental scenarios in which only partial modality data may be accessible due to hardware limitations or cost

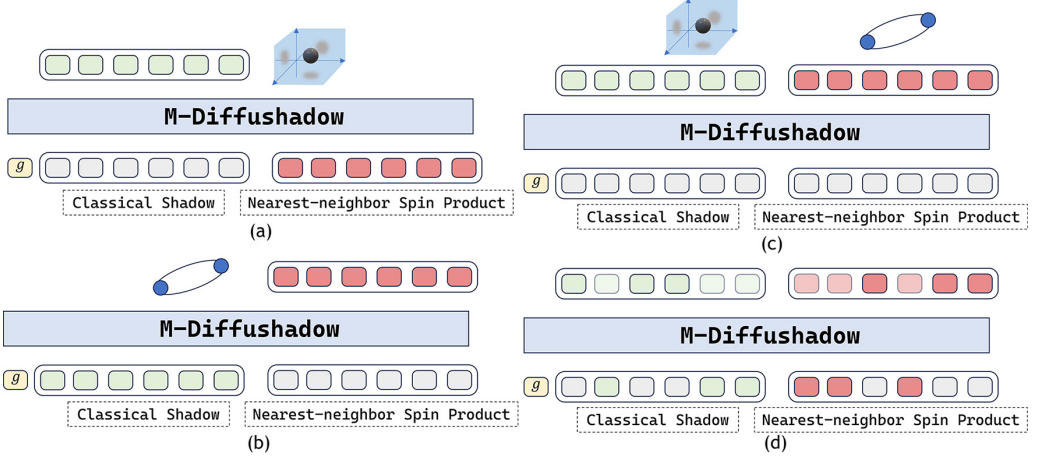


FIG. 2. Different generation modes of M-Diffushadow: (a) cross-modal inference (from nearest-neighbor spin-product to classical shadow); (b) cross-modal inference (from classical shadow to nearest-neighbor spin product); (c) joint generation; and (d) partial measurement reconstruction.

constraints. By enabling cross-modal inference, the model can reconstruct missing observational information from available data, providing a flexible and efficient tool for experimental design and data analysis. Second, M-Diffushadow supports partial measurement reconstruction within both modalities. In this setting, measurement outcomes from both modalities are randomly masked, and the model generates the missing entries conditioned on the remaining observed data across modalities. This corresponds to experimentally realistic scenarios in which measurement records are incomplete due to finite sampling, noise, or data corruption affecting multiple measurement channels. By leveraging correlations both within and across modalities, the model reconstructs incomplete records while preserving the underlying statistical structure. Third, M-Diffushadow enables the simultaneous generation of observations from both modalities. Given the system parameter g and the measurement basis sequence of both modalities, the model jointly generates the classical-shadow measurement outcomes b_j and the nearest-neighbor spin-product measurement outcomes d_j . This multimodal generation paradigm enables efficient and coherent synthesis of heterogeneous experimental data from the same physical system, thereby significantly improving experimental efficiency.

As shown in Fig. 1(b), the generating process of M-Diffushadow starts with a completely masked measurement outcome sequence of both modalities. Given the parameter of Hamiltonian g , randomly sampled Pauli measurement bases for classical shadow and nearest-neighbor spin product, respectively, and mask sequence, M-Diffushadow is expected to predict the corresponding measurement outcomes. M-Diffushadow generates predictions through iterative denoising, with the positions unmasked in each step being dynamically determined. In contrast, autoregressive models generate predictions one by one from left to right, which potentially introduce a false causal assumption. M-Diffushadow does not impose a predefined order. It acknowledges that all measurement outcomes emerge simultaneously from the same quantum state $\rho(g)$, which aligns with the fundamental nature of quantum systems.

On the other hand, we design a specific decoding strategy for this dynamic iterative decoding process. M-Diffushadow predicts probabilities for all masked positions at each denoising step. Instead of unmasking all positions simultaneously, we employ an adaptive strategy that selectively uncovers the most confident predictions in each iteration. In particular, at step i , M-Diffushadow replaces the masks at positions S_i with the latest predicted results based on their predicted confidence, while the remaining positions continue to be masked. The unmasked positions S_i , together with all previously unmasked positions, are then used as informational references for generation in step $i + 1$. This

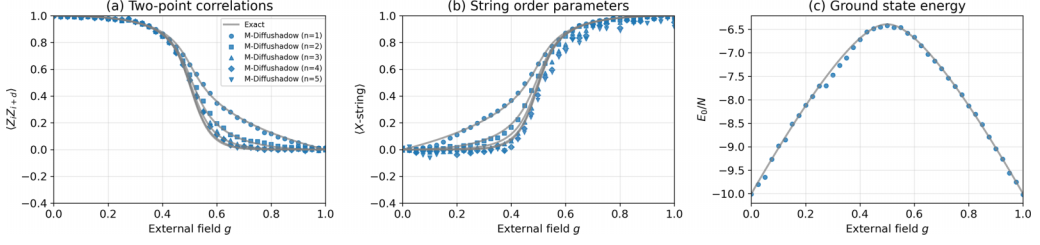


FIG. 3. The estimation results of TFI model with $N = 10$ qubits calculated by classical-shadow data generated by M-Diffushadow. (a) Correlation function $\langle Z_i Z_{i+n} \rangle$, $n = 1, \dots, 5$ of order parameters; (b) correlation function $\langle \prod_{i=1}^n X_i \rangle$, $n = 1, \dots, 5$ of disorder parameters; and (c) ground-state energy E_0 .

process is repeated at step $i + 1$ until all positions in the measurement outcome sequence have been unmasked. The decoding strategy allows the model to resolve the most certain predictions first, using these high-confidence outcomes as contextual anchors to facilitate more challenging predictions, resulting in a more robust generation trajectory.

IV. RESULTS AND DISCUSSIONS

In this section, we systematically evaluate M-Diffushadow under generation settings with progressively reduced observational information. To our knowledge, there is currently no existing generative model that addresses the multimodal quantum measurement generation task studied in this work. We have therefore selected ShadowGPT [29] as the most relevant baseline, which generates classical-shadow measurement sequences using a causal autoregressive architecture. In contrast to M-Diffushadow, ShadowGPT is a single-modal model and does not incorporate nearest-neighbor spin-product measurements. Specifically, we consider three scenarios introduced in Sec. III C: (1) cross-modal conditional generation, where one measurement modality is provided to generate the other; (2) partial measurement reconstruction, where a subset of measurement tokens is masked and reconstructed; and (3) joint generation, where both modalities are generated without conditioning. We estimate the physical properties of the quantum system using the generated data and compare them with the exact solutions. Three properties estimated by the measurement data generated by the models are included in the experiment: ground-state energy E_0 , correlation function $\langle Z_i Z_{i+n} \rangle$, $n = 1, \dots, 5$ of order parameters, and correlation function $\langle \prod_{i=1}^n X_i \rangle$, $n = 1, \dots, 5$ of disorder parameters [29]. This illustrates that a single generative model can serve as a universal data source for diverse downstream analyses.

A. Cross-modal conditional generation

We first investigate the performance of M-Diffushadow in the cross-modal conditional setting, where one measurement modality is provided and the other is generated. Specifically, we consider two directional tasks: shadow-to-spin generation and spin-to-shadow generation. This setup corresponds to a partially observed experimental scenario in which only a single type of measurement data is accessible. For the classical-shadow modality, three physical properties are estimated using the classical-shadow protocol, including the ground-state energy and the correlation functions defined above. For the nearest-neighbor spin-product modality, the nearest-neighbor correlation function is evaluated.

Figure 3 presents the physical properties computed from classical-shadow data generated by M-Diffushadow when conditioned on nearest-neighbor spin-product measurements (spin-to-shadow generation). In the figure, the solid blue dots represent the predictions from M-Diffushadow, while the gray solid line denotes the exact solution. Across all the values sampled for g , the predictions from M-Diffushadow align closely with the exact values, indicating that the model achieves high accuracy. Particularly, the region around $g = 0.5$ corresponds to the phase transition of the

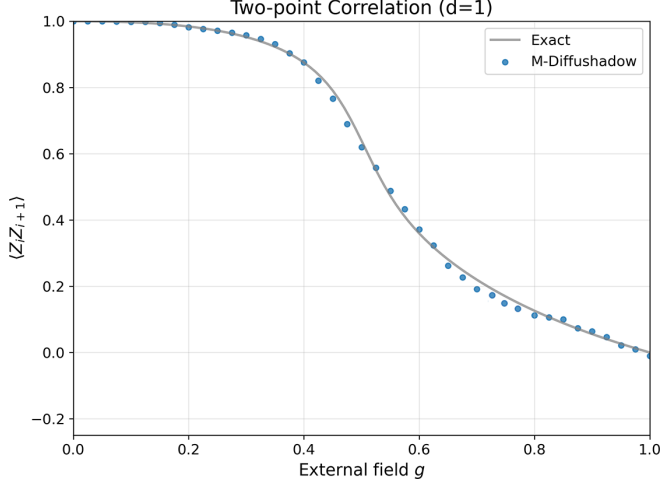


FIG. 4. The nearest-neighbor correlation function of TFI model with $N = 10$ qubits calculated by classical-shadow data generated by M-Diffushadow.

transverse-field Ising model, where internal interactions are complex and prediction is difficult. Even under such conditions, M-Diffushadow still achieves high predictive accuracy for all three physical properties, demonstrating its effectiveness. This shows that M-Diffushadow can accurately represent the classical-shadow measurements of quantum systems and reliably generate high-quality classical-shadow data for estimating physical properties. Figure 4 shows the nearest-neighbor correlation function estimated from spin-product data generated by M-Diffushadow when conditioned on classical-shadow measurements (shadow-to-spin generation). Same as the classical-shadow modality, the predictions from M-Diffushadow remain highly consistent with the exact values.

Notably, as shown in Figs. 3 and 4, the model is tested on 41 uniformly sampled values of g across $[0, 1]$, which is denser than the 8 values used during training. Considering that our model is trained only on a sparse set of g , the results show that M-Diffushadow achieves high accuracy even on g values that were not seen during the training phase, which exhibits promising generalization capabilities. This is significant for the efficient generation of a large amount of shadow data with diverse g values.

To further assess the cross-modal fidelity of M-Diffushadow, we examine observables that are accessible to both modalities. Specifically, we compute $\langle Z_i Z_{i+1} \rangle$ from the classical-shadow data generated in the CS generation mode, and $\langle \prod_{i=1}^2 X_i \rangle$ and $\langle \prod_{i=1}^4 X_i \rangle$ from the spin-product data generated in the SP generation mode. As reported in Table I, the mean squared errors (MSEs) for these quantities are comparable to those obtained under the opposite generation mode, confirming that M-Diffushadow faithfully captures the underlying quantum state and generates high-fidelity data across both cross-modal directions.

Taken together, these results demonstrate that M-Diffushadow effectively learns the cross-modal correlations between classical-shadow measurements and nearest-neighbor spin-product data. The model is therefore capable of reconstructing physically meaningful observables from a single available modality, enabling reliable inference in partially observed quantum measurement scenarios.

B. Partial measurement reconstruction

We next examine the performance of M-Diffushadow in the partial measurement reconstruction setting, where a subset of measurement entries is masked and generated conditioned on the remaining observed data. This setup corresponds to a partially observed experimental scenario in which only incomplete measurement records are available. In addition to the observable-level

TABLE I. Mean squared error of physical observables estimated from data generated by M-Diffushadow under different generation modes. Joint generation refers to simultaneous generation of CS and nearest-neighbor SP data. CS generation and SP generation correspond to cross-modal inference, where one modality is generated conditioned on the other. Dashes indicate observables that are not accessible from the corresponding modality.

g	Model	Generation mode	E_0	$\langle \Pi_{i=1}^1 X_i \rangle$	$\langle \Pi_{i=1}^2 X_i \rangle$	$\langle \Pi_{i=1}^3 X_i \rangle$	$\langle \Pi_{i=1}^4 X_i \rangle$	$\langle \Pi_{i=1}^5 X_i \rangle$	$\langle Z_{i+1} \rangle$
0.1	ShdwGPT	–	3.63×10^{-4}	7.55×10^{-3}	9.20×10^{-5}	6.80×10^{-4}	8.06×10^{-3}	2.77×10^{-3}	2.36×10^{-5}
	M-Diffu	CS generation	4.77×10^{-5}	7.06×10^{-5}	7.61×10^{-4}	1.81×10^{-4}	2.55×10^{-4}	8.23×10^{-4}	3.66×10^{-6}
		SP generation	–	–	1.11×10^{-3}	–	5.74×10^{-5}	–	1.22×10^{-6}
0.4	ShdwGPT	Joint generation	5.62×10^{-5}	7.14×10^{-5}	1.42×10^{-4}	1.28×10^{-5}	1.16×10^{-4}	1.64×10^{-3}	3.61×10^{-5}
		–	2.36×10^{-2}	8.14×10^{-5}	4.03×10^{-3}	7.33×10^{-4}	2.02×10^{-2}	3.81×10^{-4}	1.07×10^{-3}
	M-Diffu	CS generation	8.71×10^{-4}	1.37×10^{-5}	1.05×10^{-3}	1.31×10^{-3}	4.50×10^{-3}	5.20×10^{-3}	6.49×10^{-5}
0.5	ShdwGPT	SP generation	–	–	1.10×10^{-3}	–	4.06×10^{-4}	–	8.30×10^{-5}
		Joint generation	1.53×10^{-2}	7.01×10^{-5}	2.92×10^{-3}	6.53×10^{-4}	1.20×10^{-3}	2.41×10^{-4}	1.09×10^{-5}
	M-Diffu	–	1.45×10^{-2}	6.31×10^{-5}	1.61×10^{-2}	2.15×10^{-4}	3.31×10^{-3}	2.77×10^{-4}	6.30×10^{-3}
0.6	ShdwGPT	CS generation	7.59×10^{-3}	3.37×10^{-5}	4.86×10^{-5}	1.43×10^{-4}	2.88×10^{-2}	1.12×10^{-2}	6.22×10^{-6}
		SP generation	–	–	6.75×10^{-5}	–	4.03×10^{-2}	–	1.62×10^{-5}
	M-Diffu	Joint generation	1.65×10^{-2}	6.09×10^{-5}	5.23×10^{-3}	2.47×10^{-3}	4.54×10^{-2}	3.19×10^{-2}	2.42×10^{-6}
0.9	ShdwGPT	–	1.64×10^{-2}	4.81×10^{-6}	2.48×10^{-4}	3.07×10^{-4}	1.25×10^{-3}	2.05×10^{-3}	2.66×10^{-3}
		CS generation	2.39×10^{-4}	2.49×10^{-5}	3.49×10^{-5}	6.03×10^{-4}	2.77×10^{-3}	2.03×10^{-2}	4.88×10^{-6}
	M-Diffu	SP generation	–	–	1.27×10^{-5}	–	1.12×10^{-2}	–	9.05×10^{-5}
0.9	ShdwGPT	Joint generation	3.37×10^{-3}	6.27×10^{-7}	3.10×10^{-5}	1.53×10^{-3}	4.28×10^{-3}	8.44×10^{-3}	1.37×10^{-5}
		–	6.79×10^{-4}	1.28×10^{-5}	2.48×10^{-5}	4.76×10^{-5}	3.30×10^{-3}	7.57×10^{-5}	6.27×10^{-3}
	M-Diffu	CS generation	4.05×10^{-4}	2.58×10^{-6}	2.49×10^{-5}	4.50×10^{-4}	1.27×10^{-3}	4.15×10^{-3}	1.15×10^{-4}
0.9	M-Diffu	SP generation	–	–	3.45×10^{-7}	–	8.24×10^{-7}	–	3.98×10^{-5}
		Joint generation	5.15×10^{-4}	1.30×10^{-5}	1.17×10^{-4}	2.67×10^{-4}	1.58×10^{-3}	4.35×10^{-4}	1.51×10^{-5}

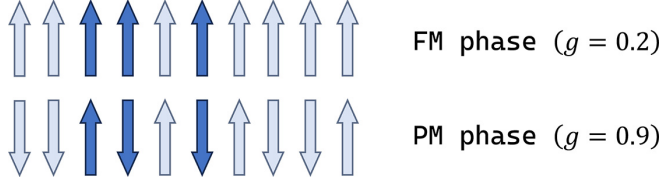


FIG. 5. One instance of partial measurement reconstruction. The dark-colored arrows represent the values that were originally missing and subsequently predicted by the model.

evaluations presented above, we further investigate the direct reconstruction accuracy of M-Diffushadow at the level of measurement outcomes. This experiment complements the other two studies: While cross-modal and joint generation are evaluated indirectly through derived physical observables, here we assess the correctness of the predicted measurement records themselves.

We focus on the ferromagnetic regime ($g < 0.5$), where the ground state exhibits strong alignment along the Z direction. All measurement bases are fixed to Z, and a subset of measurement entries is randomly masked. The model is then required to reconstruct the missing outcomes conditioned on the remaining observed entries. Since the ferromagnetic ground state is strongly polarized, the masked entries of the classical shadows are expected to take the value as the unmasked ones, and the masked entries of the nearest-neighbor spin products are expected to be 1 with high probability. To quantify reconstruction performance, we introduce *tokenwise accuracy*, defined as the fraction of masked entries that are correctly predicted.

In Fig. 5, we show one instance of generated result in two phases; dark-colored arrows indicate the originally missing entries reconstructed by the model, whereas light-colored arrows correspond to the initially known input. For the FM phase, the reconstructed spins align consistently with the known ones, all pointing in the same direction; for the PM phase, however, the reconstructed spins exhibit no clear correlation with the observed ones, appearing randomly oriented. This behavior aligns precisely with the physical nature of the two phases. We also studied the tokenwise accuracy of M-Diffushadow for different mask ratios across the ferromagnetic regime. For the classical shadow, the accuracy remains nearly 100% across the entire ferromagnetic regime and for all masking levels, indicating that the model can almost perfectly reconstruct missing entries in this modality. For the nearest-neighbor spin-product modality, the reconstruction accuracy also stays significantly above the random baseline (50%) throughout the parameter range. Specifically, the accuracy declines as the mask ratio increases, reaching its lowest point at a mask ratio of 0.8, where the accuracy is 90% at $g = 0.15$ and 0.2. The average accuracy at a mask ratio of 0.8 is 98.78%, compared to 99.6% at other mask ratios. These results confirm that M-Diffushadow effectively captures the underlying conditional structure of measurement data in partially observed settings, achieving high reconstruction fidelity at the level of individual measurement outcomes.

C. Joint generation

Finally, we evaluate the ability of M-Diffushadow to simultaneously generate measurement data for both modalities without conditioning on any prior measurement outcomes. In contrast to the cross-modal and partial reconstruction settings discussed above, joint generation requires the model to sample directly from the learned joint distribution. For the classical-shadow modality, we estimate three different physical properties using the classical-shadow protocol. As for the nearest-neighbor spin-product modality, we estimate the nearest-neighbor correlation function for evaluation.

The mean squared errors for different generation strategies are summarized in Table I. Across all considered values of g , joint generation yields consistently low estimation errors for all observables. The errors are slightly higher than those in the cross-modal conditional setting, but they are comparable to or even better than those of ShadowGPT. Only a few metrics fall slightly short of ShadowGPT. These results demonstrate that M-Diffushadow can simultaneously generate

high-quality data for both classical shadow and nearest-neighbor spin-product modalities, enabling accurate estimation of multiple physical observables.

We further verify the internal consistency of the jointly generated multimodal samples. Although the two modalities are generated simultaneously, they are expected to describe the same underlying quantum state and therefore should yield compatible estimates for observables accessible to both. To check this, we compute the nearest-neighbor correlation $\langle Z_i Z_{i+1} \rangle$ independently from the classical-shadow and spin-product modalities within each joint sample. Across all values of g , the two estimates are in close agreement, with a mean absolute discrepancy below 2.92×10^{-3} . This confirms that M-Diffushadow produces mutually consistent measurement records across modalities, and that the joint generation mode does not introduce spurious incompatibilities between the two data channels.

D. Additional validation on the XXZ model

Furthermore, M-Diffushadow is not restricted to the TFI model but can be extended to other Hamiltonians and combinations of modalities. To verify this, we perform an additional numerical experiment on the one-dimensional XXZ spin chain. The Hamiltonian is given by

$$H_{XXZ} = J_{xy} \sum_i (X_i X_{i+1} + Y_i Y_{i+1} + \Delta Z_i Z_{i+1}), \quad (5)$$

where Δ is the anisotropy parameter. Compared with the TFIM considered in the main experiments, the XXZ model provides a different spin-chain Hamiltonian involving multiple local-correlation channels.

For this experiment, we use the same M-Diffushadow architecture and training strategy as in the TFIM experiments, while replacing the transverse-field parameter g with the anisotropy parameter Δ . The classical-shadow modality is kept as the randomized single-qubit measurement modality. For the other modality, we use a local-correlation modality composed of nearest-neighbor XX , YY , and ZZ Pauli-product channels. This setting allows us to test whether the same multimodal diffusion framework can be adapted to a different Hamiltonian and a different multichannel local-correlation measurement setting. We evaluate the quality of the generated XXZ measurement data by estimating the local correlations $\langle X_i X_{i+1} \rangle$, $\langle Y_i Y_{i+1} \rangle$, $\langle Z_i Z_{i+1} \rangle$, and the ground-state energy density E_0/N . These observables can be estimated from both the generated classical-shadow data and the generated local-correlation data. Therefore, in the joint generation mode, we estimate the same set of physical quantities separately from the generated classical-shadow branch and the generated local-correlation branch, which provides a consistency check between the two generated modalities.

The results are summarized in Table II for three representative anisotropy parameters ($\Delta = -1.5, 0, 1.5$), which correspond to qualitatively different correlation regimes of the XXZ chain. Across these regimes, M-Diffushadow consistently achieves small estimation errors for both local correlations and the ground-state energy density under cross-modal generation and joint generation settings. In the negative-anisotropy regime at $\Delta = -1.5$, the model accurately recovers the dominant local correlations, with all correlation MSEs remaining below 3.2×10^{-4} . Around the easy-plane region at $\Delta = 0$, where the XX , YY , and ZZ correlations contribute differently, the model still maintains low errors, with the largest correlation MSE on the order of 10^{-4} . In the strong positive-anisotropy regime at $\Delta = 1.5$, the estimation errors remain similarly small, and several correlation errors decrease to the 10^{-6} to 10^{-7} level. The energy-density errors are also controlled across all three regimes. These results indicate that the model not only performs well at a single parameter point but can also generate physically useful measurement data across different XXZ correlation regimes.

E. Attention mechanism of M-Diffushadow

The choice of the attention mechanism is a key design consideration for modeling quantum many-body systems. In this subsection, we designed experiments to compare the use of full attention

TABLE II. MSEs of physical observables estimated from generated XXZ measurement data under different generation modes and anisotropy parameters Δ . The reported observables include the ground-state energy density E_0/N and the nearest-neighbor local correlations $\langle X_i X_{i+1} \rangle$, $\langle Y_i Y_{i+1} \rangle$, and $\langle Z_i Z_{i+1} \rangle$. For joint generation, “CS” and “PP” indicate that the observables are estimated from the generated classical-shadow branch and the generated Pauli-product local-correlation branch, respectively.

Δ	Generation mode	E_0/N	$\langle X_i X_{i+1} \rangle$	$\langle Y_i Y_{i+1} \rangle$	$\langle Z_i Z_{i+1} \rangle$
-1.5	CS generation	7.71×10^{-5}	3.78×10^{-5}	1.43×10^{-4}	8.01×10^{-5}
	PP generation	4.03×10^{-4}	8.94×10^{-5}	1.00×10^{-4}	1.66×10^{-7}
	Joint generation (CS)	6.66×10^{-4}	5.85×10^{-5}	2.62×10^{-4}	2.56×10^{-6}
	Joint generation (PP)	5.67×10^{-4}	3.12×10^{-4}	2.00×10^{-4}	1.28×10^{-7}
0	CS generation	4.41×10^{-4}	1.66×10^{-5}	2.35×10^{-4}	4.06×10^{-6}
	PP generation	8.64×10^{-5}	4.28×10^{-5}	2.96×10^{-6}	1.14×10^{-4}
	Joint generation (CS)	1.06×10^{-3}	2.17×10^{-4}	5.28×10^{-4}	1.59×10^{-6}
	Joint generation (PP)	1.71×10^{-5}	6.18×10^{-6}	5.02×10^{-6}	1.91×10^{-5}
1.5	CS generation	1.80×10^{-4}	1.42×10^{-5}	7.02×10^{-5}	4.48×10^{-6}
	PP generation	2.13×10^{-4}	1.72×10^{-5}	9.73×10^{-7}	4.14×10^{-5}
	Joint generation (CS)	1.05×10^{-4}	1.51×10^{-4}	3.57×10^{-5}	1.00×10^{-7}
	Joint generation (PP)	1.30×10^{-4}	6.57×10^{-6}	9.91×10^{-6}	7.03×10^{-5}

and causal attention mechanisms in processing measurement data. In addition, we attempted to roughly analyze the internal information flow of M-Diffushadow through attention visualization.

To directly examine whether the choice of attention mechanism affects how the model processes quantum measurement correlations, we conduct a controlled experiment on causal and full attention on the classical-shadow modality. We construct a causal-attention variant by replacing the full attention with a causal attention mechanism, while keeping all other components unchanged, including the trained weights, embedding layers, and decoding strategy. We then fix the measurement outcome of a specific qubit to a definite value and observe how the predicted outcomes of the remaining qubits change in response, compared to the original predictions before fixation. For each unfixed qubit, we compute the Kullback-Leibler (KL) divergence between the predicted probability distributions before and after the fixation. A larger KL divergence indicates that the predicted distribution for that qubit changed substantially in response to the fixation.

The results of the experiment are shown in Fig. 6. Under causal attention, fixing the measurement outcome of a given qubit affects only the qubits appearing after it in the sequence, with the preceding qubits almost unchanged. This follows directly from the unidirectional information flow imposed by the causal mask. Under full attention, in contrast, fixing a single qubit influences the predicted outcomes on both sides, demonstrating bidirectional information flow. This contrast is relevant to quantum measurement data because measurement outcomes are jointly distributed according to $\rho(g)$, with no intrinsic causal ordering among qubits. Causal attention introduces an artificial order absent from the physical process being modeled, while full attention naturally respects this symmetry.

In a transformer architecture, each attention head computes a set of attention weights that quantify how strongly the representation at one token position (query) depends on representations at other token positions (keys). These attention weights are not static model parameters; rather, they are dynamically computed from the input tokens at each forward pass. Although attention weights do not directly correspond to physical observables, they provide a window into how the model organizes correlations internally for different input sequences. In this section, we analyze the attention weights in the trained M-Diffushadow to investigate how the internal information flow is organized across different physical regimes.

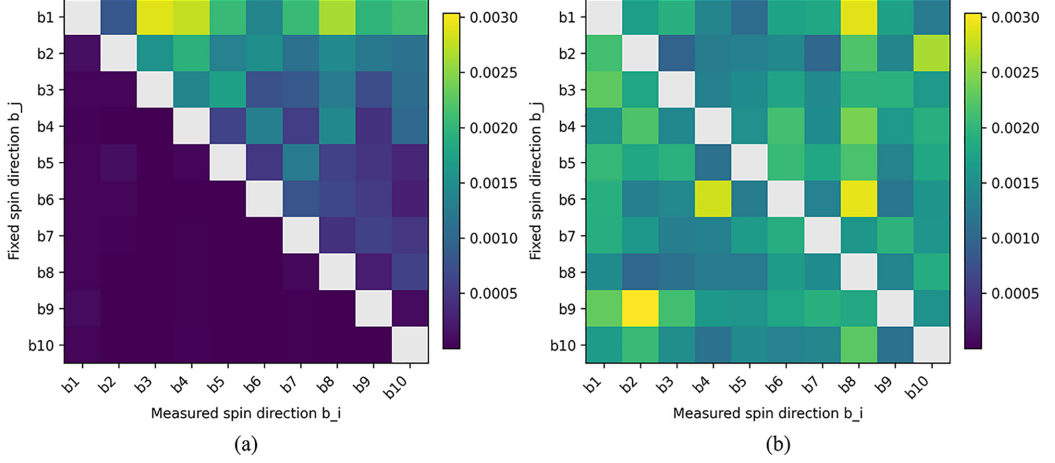


FIG. 6. The KL divergence of the remaining qubits when different specific qubits are fixed. (a) Results for M-Diffushadow with causal attention. (b) Results for M-Diffushadow with full attention.

Figures 7 and 8 show the attention maps of several attention heads for representative parameters in the ferromagnetic phase ($g = 0.2$) and the paramagnetic phase ($g = 0.9$), respectively. Specifically, the figures show head 2 in layer 1, head 1 in layer 2, and all the attention heads in the third layer. Across both values of the transverse field ($g = 0.2$ and $g = 0.9$), we observe robust

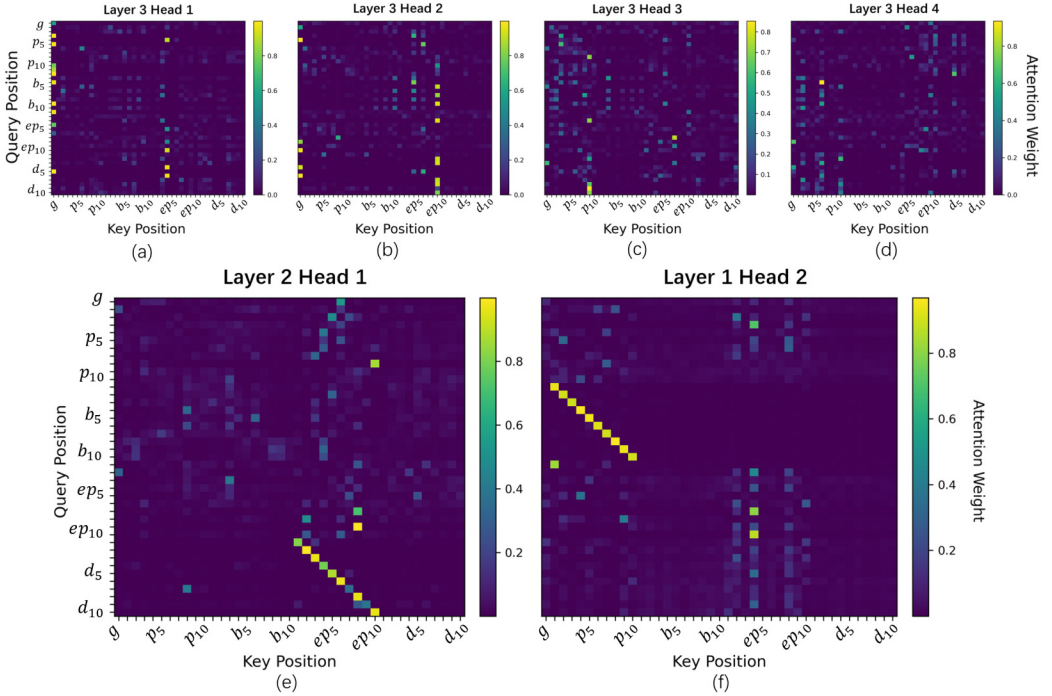


FIG. 7. Attention patterns of selected attention heads within M-Diffushadow at $g = 0.2$. (a)–(d) Heads 1–4 in layer 3, respectively; (e) head 1 in layer 2; and (f) head 2 in layer 1.

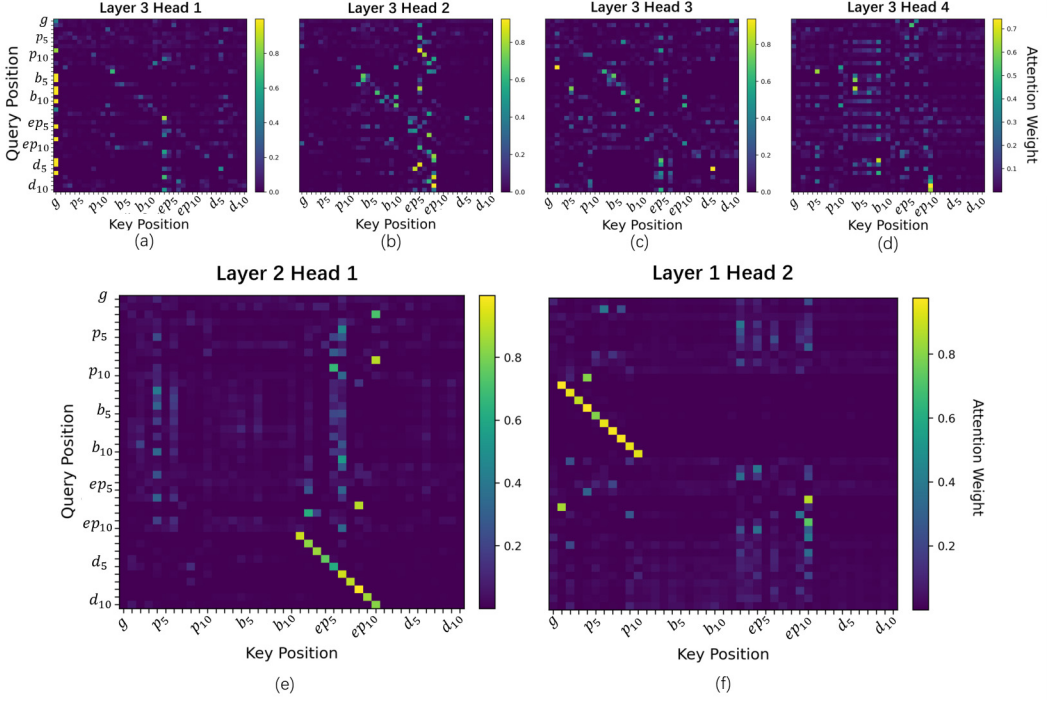


FIG. 8. Attention patterns of selected attention heads within M-Diffushadow at $g = 0.9$. (a)–(d) Heads 1–4 in layer 3, respectively; (e) head 1 in layer 2; and (f) head 2 in layer 1.

near-diagonal structures between operator tokens and their corresponding measurement outcomes. Specifically, in the attention map of layer 1, head 2, an approximately diagonal structure is observed between the measurement basis and outcomes of classical shadows. Similarly, in layer 2, head 1, the attention map reveals an approximately diagonal pattern between the measurement basis and outcomes of nearest-neighbor spin products. It demonstrates that distinct heads capture each modality. These diagonal alignments reflect the underlying generative structure of the data and indicate that the model has learned the local operator-outcome correspondence in a consistent manner across both regimes. In addition to the operator-outcome correspondence, we also observe interactions between different classes of measurement results. In particular, the attention map of head 4 in layer 3 displays noticeable weight linking single-site outcomes b and pair-product outcomes d . The presence of such cross-outcome attention suggests that the model does not treat these quantities independently, but instead builds internal representations that encode their mutual constraints. Furthermore, discernible qualitative differences are observed between the two phases. In the paramagnetic case ($g = 0.9$), several heads in layer 3 exhibit more pronounced diagonal trends and relatively concentrated attention patterns, whereas in the ferromagnetic case ($g = 0.2$) the attention appears comparatively more diffuse. This contrast may be related to differences in correlation length: Short-range correlations in the paramagnetic phase promote a localized operator-outcome correspondence, while long-range correlations in the ferromagnetic phase lead to more distributed attention. However, this observation remains preliminary and further systematic analysis is required.

These observed structured and phase-dependent patterns provide qualitative evidence that M-Diffushadow organizes and redistributes information in a manner consistent with the underlying correlation properties of the quantum many-body system.

V. DISCUSSION AND OUTLOOK

The results presented in the previous sections demonstrate that M-Diffushadow is not merely a generative tool for measurement interpolation, but a framework that captures the underlying physical correlations of the quantum system. In this section, we discuss the broader implications of our findings and outline directions for future work.

A particularly promising direction is the application of M-Diffushadow to readout error mitigation in NISQ devices [30]. The nonautoregressive nature of the diffusion process inherently allows for iterative self-correction [18,28]. Unlike autoregressive models, where errors at early steps propagate throughout the sequence, M-Diffushadow reevaluates the entire configuration at each denoising step. In practical scenarios, measurement records corrupted by hardware bit-flip noise can be treated as an initial noisy state that the model iteratively refines. By leveraging cross-modal consistency, the model can identify incompatible measurement outcomes where local correlations in the SP modality contradict the global information in the CS modality and use these physical constraints to correct the suspicious tokens. This iterative refinement process offers a natural path to project noisy hardware data back onto the physically consistent manifold.

Beyond the specific configurations explored in the main TFIM experiments, the M-Diffushadow framework is designed to be adaptable with respect to both the underlying Hamiltonian and the choice of measurement modalities. Although the nearest-neighbor spin-product modality is naturally suited to the TFIM benchmark, neither the TFIM Hamiltonian nor this specific auxiliary modality is an intrinsic requirement of the architecture. The additional XXZ-chain experiment provides preliminary numerical evidence for this flexibility, showing that the same framework can be applied to another spin-chain Hamiltonian together with a different multichannel local-correlation modality composed of nearest-neighbor XX , YY , and ZZ Pauli-product channels. For other Hamiltonians, the auxiliary modality can be tailored according to the dominant correlations and experimentally accessible measurement channels of the target system. For example, in Rydberg atom arrays, site-resolved occupation snapshots, density-density correlations, or blockade-constrained local configurations provide natural measurement channels for interacting spin models [31,32]. In Hubbard-type optical-lattice systems, spin-resolved or density-resolved snapshots, doublon-hole correlations, and string or parity observables could play an analogous role [33]. The framework can more generally accommodate a broader range of modality combinations, such as long-range correlations, multicomponent measurements, or higher-order Pauli strings, tailored to specific research requirements. Since these measurements can be represented as discrete token sequences with associated type and positional encodings, the same multimodal diffusion framework can be generalized across diverse physical platforms without being tied to the TFIM-specific nearest-neighbor spin-product choice. A promising direction for future exploration is therefore to leverage M-Diffushadow as a unified integration platform, capable of fusing fragmented records from distinct experimental setups, such as combining site-resolved snapshots from quantum gas microscopes with global spectroscopic data [34].

VI. CONCLUSIONS

We introduce M-Diffushadow, a multimodal diffusion large language model for quantum system modeling and property estimation. To capture the complex internal correlations inherent in quantum systems, our model employs a full attention mechanism. Furthermore, we design a structured multimodal embedding to effectively process the diverse input modalities. To enable efficient learning of quantum system characteristics and representations, the training paradigm is carefully crafted, integrating three key components: a dynamic reshuffling data loading strategy, a dedicated mask addition mechanism, and an efficient loss function. M-Diffushadow’s generative framework exhibits remarkable flexibility, supporting cross-modal, partial-reconstruction, and joint generation modes. Its dynamic iterative unmasking strategy respects the simultaneous nature of quantum measurements, ensuring high-fidelity synthesis and physically consistent inference.

We evaluate M-Diffushadow across its different generation modes through a series of experiments on the TFI model. We begin by investigating the cross-modal inference performance of

M-Diffushadow. Experimental results confirm that, conditioned on a single observed modality, our model reliably generates high-quality estimates for the missing modality. Second, we evaluate our model on partial measurement reconstruction. The high accuracy in the ferromagnetic phase demonstrates the model’s ability to effectively reconstruct missing data. Finally, we task the model with jointly generating both modalities, which also yields highly accurate results. Additional experiments on the XXZ model provide numerical evidence supporting M-Diffushadow’s generalization beyond the TFI model.

As the first multimodal diffusion language model tailored for quantum systems, M-Diffushadow demonstrates promising capabilities in modeling quantum many-body systems and reconstructing incomplete measurement data. These results position it as a powerful tool with significant potential for advancing quantum experimental research.

ACKNOWLEDGMENTS

We acknowledge support from the Shenzhen Fundamental Research Program (No. JCYJ20250604145655074) and the Shenzhen Key Laboratory of Advanced Functional Carbon Materials Research and Comprehensive Application (No. ZDSYS20220527171407017).

DATA AVAILABILITY

The source code associated with this article is openly available at Ref. [35].

-
- [1] S. Sachdev, Quantum phase transitions, *Phys. World* **12**, 33 (1999).
 - [2] P. A. Lee, N. Nagaosa, and X.-G. Wen, Doping a Mott insulator: Physics of high-temperature superconductivity, *Rev. Mod. Phys.* **78**, 17 (2006).
 - [3] J. Preskill, Quantum computing in the NISQ era and beyond, *Quantum* **2**, 79 (2018).
 - [4] D. Basov, R. Averitt, and D. Hsieh, Towards properties on demand in quantum materials, *Nat. Mater.* **16**, 1077 (2017).
 - [5] D. Bourgund, T. Chalopin, P. Bojović, H. Schlömer, S. Wang, T. Franz, S. Hirthe, A. Bohrdt, F. Grusdt, I. Bloch, *et al.*, Formation of individual stripes in a mixed-dimensional cold-atom Fermi–Hubbard system, *Nature (London)* **637**, 57 (2025).
 - [6] A. D. King, A. Nocera, M. M. Rams, J. Dziarmaga, R. Wiersema, W. Bernoudy, J. Raymond, N. Kaushal, N. Heinsdorf, R. Harris, *et al.*, Beyond-classical computation in quantum simulation, *Science* **388**, 199 (2025).
 - [7] H.-Y. Huang, R. Kueng, and J. Preskill, Predicting many properties of a quantum system from very few measurements, *Nat. Phys.* **16**, 1050 (2020).
 - [8] G. Struchalin, Y. A. Zagorovskii, E. Kovlakov, S. Straupe, and S. Kulik, Experimental estimation of quantum state properties from classical shadows, *PRX Quantum* **2**, 010307 (2021).
 - [9] G. Carleo, I. Cirac, K. Cranmer, L. Daudet, M. Schuld, N. Tishby, L. Vogt-Maranto, and L. Zdeborová, Machine learning and the physical sciences, *Rev. Mod. Phys.* **91**, 045002 (2019).
 - [10] J. Carrasquilla and R. G. Melko, Machine learning phases of matter, *Nat. Phys.* **13**, 431 (2017).
 - [11] V. Dunjko and H. J. Briegel, Machine learning & artificial intelligence in the quantum domain: A review of recent progress, *Rep. Prog. Phys.* **81**, 074001 (2018).
 - [12] P. Mehta, M. Bukov, C.-H. Wang, A. G. Day, C. Richardson, C. K. Fisher, and D. J. Schwab, A high-bias, low-variance introduction to machine learning for physicists, *Phys. Rep.* **810**, 1 (2019).
 - [13] G. Torlai, G. Mazzola, J. Carrasquilla, M. Troyer, R. Melko, and G. Carleo, Neural-network quantum state tomography, *Nat. Phys.* **14**, 447 (2018).
 - [14] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, Generative adversarial nets, in *NIPS’14: Proceedings of the 28th International Conference on Neural Information Processing Systems*, NIPS’14, Vol. 2 (MIT Press, Cambridge, MA, 2014), pp. 2672–2680.

- [15] D. P. Kingma and M. Welling, Auto-encoding variational Bayes, in *Proceedings of the 2nd International Conference on Learning Representations (ICLR, Banff, 2014)*.
- [16] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, Language models are unsupervised multitask learners, [OpenAI Blog](#) **1**, 9 (2019).
- [17] Y. Du, Y. Zhu, Y.-H. Zhang, M.-H. Hsieh, P. Rebentrost, W. Gao, Y.-Z. You, J. Eisert, G. Chiribella, D. Tao, B. C. Sanders, and Y.-D. Wu, Artificial intelligence for representing and characterizing quantum systems, *Nat. Rev. Phys.* **8**, 579 (2026).
- [18] S. Nie, F. Zhu, Z. You, X. Zhang, J. Ou, J. Hu, J. Zhou, Y. Lin, J.-R. Wen, and C. Li, Large language diffusion models, *Adv. Neural Inf. Process. Syst.* **38**, 50608 (2025).
- [19] F. Zhu, R. Wang, S. Nie, X. Zhang, C. Wu, J. Zhou, Y. Lin, J.-R. Wen, and C. Li, LLaDA 1.5 variance-reduced preference optimization for large language diffusion models, in *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics* (Association for Computational Linguistics, San Diego, 2026), pp. 11425–11460.
- [20] J. Ho, A. Jain, and P. Abbeel, Denoising diffusion probabilistic models, in *Proceedings of the 34th International Conference on Neural Information Processing Systems* (Curran Associates Inc., Red Hook, NY, 2020), pp. 6840–6851.
- [21] J. Austin, D. D. Johnson, J. Ho, D. Tarlow, and R. van den Berg, Structured denoising diffusion models in discrete state-spaces, *Adv. Neural Inf. Process. Syst.* **34**, 17981 (2021).
- [22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, Attention is all you need, *Adv. Neural Inf. Process. Syst.* **30**, 5998 (2017).
- [23] P. Pfeuty, The one-dimensional Ising model with a transverse field, *Ann. Phys.* **57**, 79 (1970).
- [24] B. K. Chakrabarti, A. Dutta, and P. Sen, *Quantum Ising Phases and Transitions in Transverse Ising Models* (Springer, Berlin, Heidelberg, 1996).
- [25] H.-Y. Huang, R. Kueng, and J. Preskill, Efficient estimation of Pauli observables by derandomization, *Phys. Rev. Lett.* **127**, 030503 (2021).
- [26] L. Amico, R. Fazio, A. Osterloh, and V. Vedral, Entanglement in many-body systems, *Rev. Mod. Phys.* **80**, 517 (2008).
- [27] N. Laflorencie, Quantum entanglement in condensed matter systems, *Phys. Rep.* **646**, 1 (2016).
- [28] R. Yu, Q. Li, and X. Wang, Discrete diffusion in large language and multimodal models: A survey, [arXiv:2506.13759](#).
- [29] J. Yao and Y.-Z. You, ShadowGPT: Learning to solve quantum many-body problems from randomized measurements, [arXiv:2411.03285](#).
- [30] K. Temme, S. Bravyi, and J. M. Gambetta, Error mitigation for short-depth quantum circuits, *Phys. Rev. Lett.* **119**, 180509 (2017).
- [31] A. Browaeys and T. Lahaye, Many-body physics with individually controlled Rydberg atoms, *Nat. Phys.* **16**, 132 (2020).
- [32] S. Ebadi, T. T. Wang, H. Levine, A. Keesling, G. Semeghini, A. Omran, D. Bluvstein, R. Samajdar, H. Pichler, W. W. Ho, S. Choi, S. Sachdev, M. Greiner, V. Vuletić, and M. D. Lukin, Quantum phases of matter on a 256-atom programmable quantum simulator, *Nature (London)* **595**, 227 (2021).
- [33] C. Gross and I. Bloch, Quantum simulations with ultracold atoms in optical lattices, *Science* **357**, 995 (2017).
- [34] C. Gross and W. S. Bakr, Quantum gas microscopy for single atom and spin detection, *Nat. Phys.* **17**, 1316 (2021).
- [35] Y. Li and H.-K. Tang, M-Diffushadow Source code for Multimodal discrete diffusion for quantum measurement generation in the one-dimensional transverse-field Ising model, GitHub repository, 2026, <https://github.com/Miamm0v0/M-Diffushadow>.