SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

August 2026, Vol. 69, Iss. 8, 180504:1–180504:16
https://doi.org/10.1007/s11432-025-5016-2

Special Topic: Quantum Information

Q-Tag: watermarking quantum circuit generative models

Yang YANG[1,2], Yuzhu LONG[1], Han FANG[3*], Zhaoyun CHEN[2], Zhonghui LI[1], Weiming ZHANG[3*] & Guoping GUO[4,5]

[1]*School of Electronic Information Engineering, Anhui University, Hefei 230039, China*
[2]*Institute of Artificial Intelligence, Hefei Comprehensive National Science Center, Hefei 230088, China*
[3]*School of Cyber Science and Technology, University of Science and Technology of China, Hefei 230026, China*
[4]*Laboratory of Quantum Information, University of Science and Technology of China, Hefei 230026, China*
[5]*Origin Quantum Computing Technology Company, Hefei 230026, China*

Received 31 December 2025/Revised 30 March 2026/Accepted 26 June 2026/Published online 28 July 2026

Abstract Quantum cloud platforms have become the most widely adopted and mainstream approach for accessing quantum computing resources, due to the scarcity and operational complexity of quantum hardware. In this service-oriented paradigm, quantum circuits, which constitute high-value intellectual property, are exposed to risks of unauthorized access, reuse, and misuse. Digital watermarking has been explored as a promising mechanism for protecting quantum circuits by embedding ownership information for tracing and verification. However, driven by recent advances in generative artificial intelligence, the paradigm of quantum circuit design is shifting from individually and manually constructed circuits to automated synthesis based on quantum circuit generative models (QCGMs). In such generative settings, protecting only individual output circuits is insufficient, and existing post hoc, circuit-centric watermarking methods are not designed to integrate with the generative process, often failing to simultaneously ensure stealthiness, functional correctness, and robustness at scale. These limitations highlight the need for a new watermarking paradigm that is natively integrated with quantum circuit generative models. To the best of our limited knowledge, we present the first watermarking framework for QCGMs, which embeds ownership signals into the generation process while preserving circuit fidelity. We introduce a symmetric sampling strategy that aligns watermark encoding with the model's Gaussian prior, and a synchronization mechanism that counteracts adversarial watermark attack through latent drift correction. Empirical results confirm that our method achieves high-fidelity circuit generation and robust watermark detection across a range of perturbations, paving the way for scalable, secure copyright protection in AI-powered quantum design.

Keywords quantum circuit, IP protection, latent diffusion model, quantum circuit generative models, watermarking

Citation Yang Y, Long Y Z, Fang H, et al. Q-Tag: watermarking quantum circuit generative models. Sci China Inf Sci, 2026, 69(8): 180504, https://doi.org/10.1007/s11432-025-5016-2

1 Introduction

Quantum machine learning (QML) has emerged as a promising frontier in quantum computing [1, 2], offering the potential to accelerate tasks in classification, regression, generative modeling, and beyond by exploiting quantum parallelism [3–7]. These capabilities are poised to revolutionize fields ranging from chemistry and cryptography to machine learning [8–10]. As quantum algorithms grow increasingly complex and application-driven, quantum circuits, the executable representations of quantum algorithms, play a central role in bridging theoretical models and real hardware implementation [11]. In essence, quantum circuits form the concrete runtime layer of QML, translating high-level abstractions into physical gate operations. Given their function-specific structure and optimization, these circuits are not generic but encode significant intellectual property (IP), often the result of proprietary design or computationally expensive synthesis [12].

Quantum cloud platforms have become the most widely adopted and mainstream approach for accessing quantum computing resources, largely due to the scarcity and operational complexity of quantum hardware. Leading platforms such as IBM Quantum [13], Amazon Braket [14], and Origin Quantum [15] provide standardized cloud-based interfaces that allow users to remotely submit and execute quantum circuits without maintaining physical quantum devices, thereby significantly lowering the barrier to quantum computing and enabling scalable access to quantum resources [16,17]. However, outsourcing circuit execution to quantum cloud platforms also introduces non-negligible intellectual property (IP) risks. Once uploaded to the cloud, quantum circuits may be stored, inspected, replicated, or reused by untrusted third parties without the designer's consent or awareness. All of these threats highlight the

* Corresponding author (email: fanghan@ustc.edu.cn, zhangwm@ustc.edu.cn)

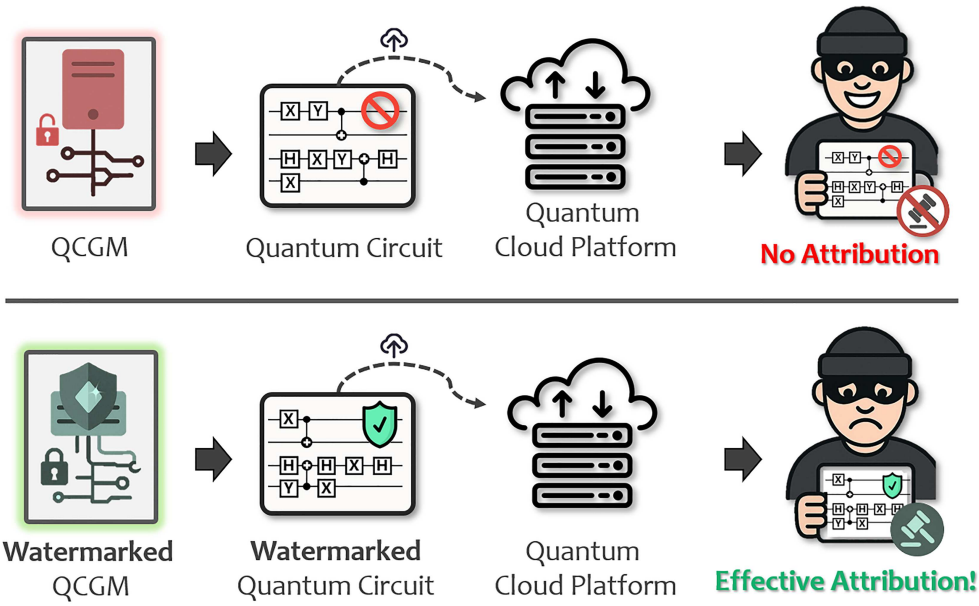

Figure 1 (Color online) Watermarking for quantum circuit generative model. Top: standard QCGMs generate circuits without embedded ownership, enabling misuse via cloud deployment with no attribution. Bottom: our watermarked QCGMs embed verifiable ownership into each generated circuit, enabling attribution and accountability even after unauthorized redistribution.

need for added protections and security measures that cloud providers must implement to ensure the confidentiality, integrity, and availability of quantum circuits [18]. To mitigate these risks, several studies have explored post hoc quantum circuit watermarking techniques that embed ownership information directly into quantum circuits via appending a rotation gate on ancilla qubits combined with other gates at the circuit output, or inserting a pair of random gates and their inverses in the circuit interior separated by barriers [19], enabling copyright verification and forensic tracing after deployment. These approaches represent early attempts to protect circuit-level intellectual property in quantum cloud environments.

Driven by recent advances in generative artificial intelligence, the paradigm of quantum circuit design is undergoing a fundamental shift from individually and manually constructed circuits to automated synthesis based on quantum circuit generative models (QCGMs) [6]. By leveraging data-driven and learning-based approaches, QCGMs enable efficient and scalable generation of quantum circuits conditioned on task objectives or optimization criteria, making them particularly suitable for deployment as cloud-based quantum design services. This shift in design paradigm fundamentally changes the requirements for copyright protection. In generative settings, quantum circuits are no longer isolated design artifacts but outputs sampled from a learned generative process, rendering the protection of individual circuit instances insufficient [20]. Instead, copyright protection must be tightly coupled with the generative mechanism itself, ensuring that all synthesized circuits inherently carry consistent and verifiable ownership information, as illustrated in Figure 1.

Under these requirements, the task of copyright protection shifts from safeguarding a static, finite set of designs to securing an infinite stream of model-generated outputs. This new scenario imposes stringent performance demands-specifically high generative fidelity, minimal hardware overhead, and robustness-that existing post-hoc, circuit-centric watermarking methods fail to meet for two fundamental reasons. First, these techniques typically necessitate explicit, intrusive modifications to the gate-level netlists or topological structures. Such structural perturbations can inadvertently shift the output away from the optimal generative distribution, potentially introducing subtle biases that compromise the fidelity and quality of the generated circuits. Furthermore, these modifications often incur non-trivial overheads in terms of power, performance, and area (PPA), which is undesirable in high-performance design. Second, their robustness is limited when circuits undergo common post-generation operations such as optimization, editing, or redistribution, making them unreliable for large-scale, automated generative settings. These challenges call for watermarking mechanisms that are natively integrated into quantum circuit generative models, enabling model-level copyright protection while preserving generation quality and robustness.

To address this emerging need, we present the first watermarking framework Q-Tag tailored to QCGMs, establishing a foundation for scalable and model-level copyright protection in AI-driven quantum design. Our method embeds ownership signals directly into the generative process by controlling the latent initialization through a symmetric sampling mechanism (SSM), ensuring statistical alignment with the model's Gaussian prior and avoiding

generation artifacts. Furthermore, to ensure robustness against watermark attacks, we introduce a synchronization restoration strategy that corrects latent drift via proactive zero-padding techniques. Empirical evaluations across multiple scenarios demonstrate that our method consistently preserves generation fidelity while enabling reliable watermark detection under various adversarial conditions.

In general, the contributions of this paper can be summarized in three main aspects.

• To the best of our knowledge, we are the first to focus on the pioneering demand of QCGMs copyright protection and propose an effective watermarking mechanism for potential IP risk in quantum cloud platforms.

• The proposed mechanism can sufficiently satisfy two properties of QCGMs watermarking: losslessness and robustness. Especially for robustness, we first investigate the robustness of four unique synchronization attacks in the circuit editing process.

• Extensive experiments show that the proposed method achieves superior performance in both losslessness and robustness. In particular, the watermark detection rate can reach 99.3% even under modification attacks.

2 Related work

2.1 Quantum circuit synthesis

Quantum computing is widely regarded as a transformative technology with immense potential and promising applications in a wide range of fields from basic research in cryptography and chemistry to machine learning and optimization [1, 8, 9, 21]. Despite the continuous advancements in quantum processor technology, the noisy intermediate-scale quantum (NISQ) [22] era still confronts numerous challenges, primarily stemming from the inherent limitations of the current devices. Therefore, efficient quantum circuits must be designed based on available resources and gate sets, which requires further research and expertise to explore optimal gate design and circuit construction strategies.

This task is often referred to as quantum circuit synthesis and mainly includes the following aspects: in order to generate the quantum state of the target output under the conditions of a given quantum processor, a series of basic elements (called quantum gates) need to be designed to generate the quantum state from the reference or standard input state [6]. Similarly, the goal may not be to generate only the desired quantum states but to implement quantum algorithms or more general unitary operations. In particular, researchers have recently demonstrated important advances in circuit synthesis by utilizing machine learning techniques, from reinforcement learning to generative models, such as in quantum state preparation, prediction, circuit optimization, or unified compilation [23–26].

2.2 Diffusion model

Diffusion models are a class of generative models designed to learn the diffusion process that generates a probability distribution for a given training data set. The denoising diffusion probability model (DDPM) introduced by Ho et al. [27] has attracted much attention. It consists of two Markov chains for adding and removing noise. Its training process is divided into two stages: the initial stage is to gradually add Gaussian noise to the image, which is called the forward process; subsequently, during the reverse process, the model parameters are trained so that it can learn the inversion of the noise. Subsequent studies [28, 29] all adopted this bi-directional chain framework. In order to reduce computational complexity and improve efficiency, the latent diffusion model (LDM) [30] came into being, in which the diffusion process takes place in the latent space.

With the emergence of potentially diffused LDM and its flexible guidance and regulation in generating capacity [31], diffusion models have become the preferred method for image generation [32], audio synthesis [33], video generation [34], and protein structure prediction [35]. Recently, a quantum circuit generation model [6] has been proposed to retrain the diffusion model by encoding a large quantum circuit data set. After training, a tensor of random noise is input into the model. The model iteratively de-noises under the guidance of selected conditions and generates high-quality quantum circuit samples after several steps.

2.3 Generative watermarking

Watermarking effectively solves copyright protection and content authentication by embedding copyright or traceable identification information in carrier data. Generative watermarking is a technology used to identify and protect the copyright of related content in generative tasks. Its concept comes from the new application of digital watermarking in generative tasks. Different from traditional watermarking and deep learning-based watermarking, on the one hand, generative watermarking can embed the watermark information in the original data set to prevent the data set from being used to train the generative model without authorization; on the other hand, it is also

possible to combine the watermark embedding with the generation process, so that the watermark is integrated into the generated content, thus protecting the copyright of the generated content. With the breakthrough of the diffusion model in the field of high-quality content generation, researchers have gradually turned their attention to the generated content protection strategy based on the diffusion model. To meet this demand, existing watermarking methods include post-processing [36–41], fine-tuning generation models [42–46], and implicit watermark embedding [47–50]. Most of the existing generation model watermarking techniques are designed with images as the carrier. Faced with the novel carrier of the quantum circuit, the existing watermarking methods cannot apply the correlation generation task well. Therefore, this paper deeply explores and effectively solves the problems faced by the model watermarking generation of quantum circuits.

2.4 Watermarking of quantum circuits

Digital watermarking has been proposed as a method to protect the copyright of quantum circuits while preserving their functionality. Existing methods for watermarking quantum circuits can be broadly categorized into two main types. The first type embeds watermarks during the deployment of the circuits, such as the decomposition, mapping, and scheduling stages of quantum circuit [12]. These methods rely on auxiliary information introduced during the deployment phase; consequently, they offer no copyright protection if the circuit is compromised or copied prior to deployment. The second type, in contrast, embeds watermarks in a circuit by adding additional quantum gates, modifying the structure while maintaining its functional integrity [19, 51]. These approaches typically embed watermarks by inserting redundant quantum gates, which incurs additional circuit overhead and fails to account for robustness against common post-generation transformations-such as optimization passes, noise-induced distortions, or adversarial tampering.

Drawing parallels from the classical integrated circuit (IC) domain, techniques such as logic locking and hardware obfuscation have long provided robust IP protection. However, the direct transferability of these classical paradigms to the quantum domain is hindered by fundamental physical constraints. In classical design, extra gates or logical modifications have negligible impact on functionality; in contrast, quantum circuits are highly sensitive to gate counts due to decoherence and accumulated physical noise. Recent attempts to adapt classical logic locking and obfuscation [52–55] have demonstrated that such structural interventions often necessitate compiler barriers or excessive gate overhead, which paradoxically compromise the circuit's original fidelity and expose the watermark to easy removal by compilers or adversarial analysis.

In summary, existing watermarking schemes are either deployment-dependent or structurally intrusive, and neither paradigm ensures reliable, robust, and intrinsic ownership protection in generative or pre-deployment settings.

3 Method

A complete overview of our proposed pipeline is provided in Figure 2, comprising three main stages: (a) watermark embedding phase, where a binary watermark is injected into the latent input of the QCGMs via an SSM; (b) watermark attacks phase, where the generated quantum circuit may be subject to structural manipulations intended to remove or disrupt the watermark; and (c) watermark extraction phase, where the watermark is recovered and verified from the possibly tampered circuit. During extraction, a synchronization restoration mechanism (SRM) is additionally applied to correct latent misalignments introduced by circuit-level edits prior to watermark decoding.

3.1 Watermark embedding

To embed a watermark, typically represented as a binary sequence, we first apply error-correcting codes (ECC) to introduce redundancy and enhance robustness against bit-level distortions. The encoded sequence is then spectrally transformed via the symmetric sampling mechanism, producing an initial latent vector that satisfies two essential criteria: (i) compatibility with the QCGMs input format, which statistical conformity to the standard Gaussian prior, and (ii) fidelity to the embedded watermark signal. This watermarked latent is subsequently processed by the QCGMs through iterative denoising steps, ultimately yielding a quantum circuit with imperceptibly embedded ownership information. The success of this embedding procedure hinges on synthesizing latent vectors that are both semantically meaningful and statistically indistinguishable from standard QCGMs inputs, while faithfully encoding the intended watermark. The detailed implementation procedures are as follows.

To embed a k-bit binary sequence $m \in \{0,1\}^k$, i.e., the watermark, the sequence is first encoded using an error correction coding mechanism **Enc** into an n-bit codeword $\mathcal{S}_{\text{en}} = \mathbf{Enc}(m)$, where $n = f_c \times f_h \times f_w$ and f_c, f_h,

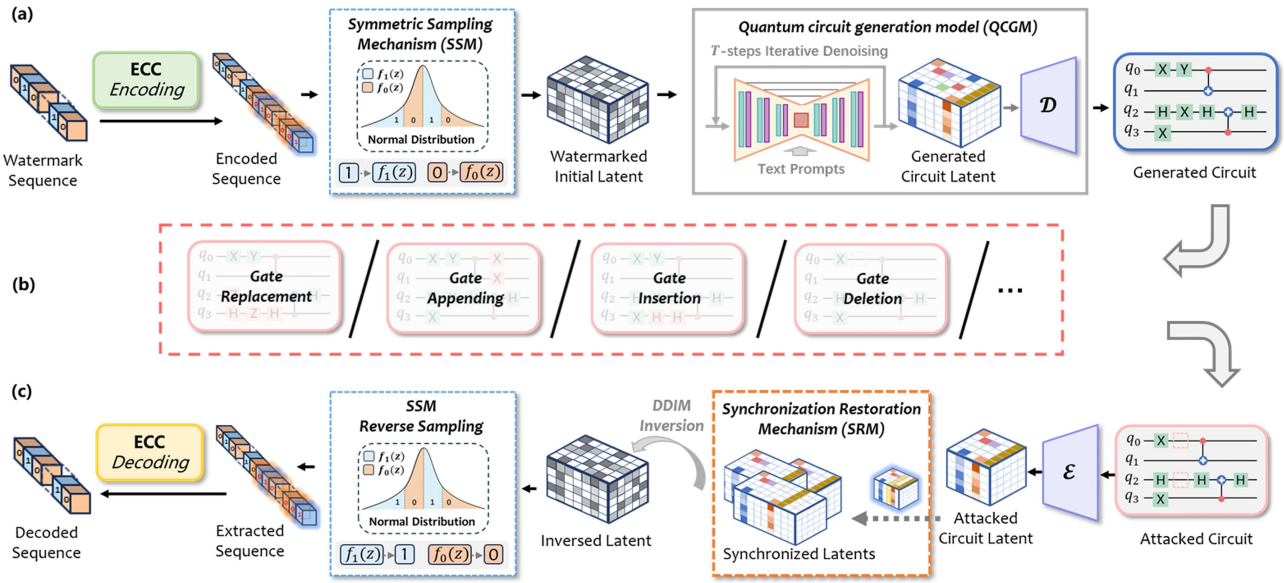

Figure 2 (Color online) Watermark (a) embedding, (b) attacks, and (c) extraction process for quantum circuits generative models (QCGMs). The watermark is embedded into the starting latent of QCGMs through SSM, then the watermarked QC is generated and potentially attacked. The SRM and SSM reverse sampling is applied in the extraction stage to extract the watermark.

and f_w denote the dimensions of the input latent. The encoded sequence $\mathcal{S}_{\text{en}} = \{s_{\text{en}}^1, s_{\text{en}}^2, \ldots, s_{\text{en}}^n\}$ satisfies two key properties: pseudorandomness, meaning that $\mathcal{S}_{\text{en}}$ is computationally indistinguishable from a uniformly distributed random sequence; and error-correctability, ensuring that the original message s can be reliably recovered even when $\mathcal{S}_{\text{en}}$ experiences a bounded number of bit errors. Next, SSM is applied to $\mathcal{S}_{\text{en}}$ to generate the watermarked starting latent $\mathcal{Z}_T$. This process yields the starting latent $\mathcal{Z}_T = \{z_T^1, z_T^2, \ldots, z_T^n\}$. Subsequently, $\mathcal{Z}_T$ is passed through QCGMs for T diffusion steps to produce the final latent $\mathcal{Z}_0$. Finally, the watermarked quantum circuit Q_m is generated via the trained decoder $\mathcal{D}$, following $Q_m = \mathcal{D}(\mathcal{Z}_0)$.

In this work, the encoding function **Enc** is instantiated using a lightweight combination of repetition coding and stream cipher encryption, designed to balance robustness and pseudorandomness. Given a k-bit binary message $m \in \{0,1\}^k$, we first apply repetition coding by duplicating each bit v times to obtain a length-n intermediate sequence $M \in \{0,1\}^n$, where $n = k \times v$. This repetition enhances error tolerance by introducing redundancy at the bit level.

Next, we generate a binary pseudorandom key $K \in \{0,1\}^n$ using a stream cipher, such as a seeded pseudorandom number generator. The final encoded sequence $\mathcal{S}_{\text{en}}$ is then computed by applying bitwise XOR between the repeated message and the key:

$$\mathcal{S}_{\text{en}} = M \oplus K.$$

This construction ensures that $\mathcal{S}_{\text{en}}$ exhibits pseudorandomness, rendering it statistically indistinguishable from a uniformly random sequence, while the underlying redundancy introduced by repetition supports error-correctability during watermark recovery. The encoded sequence $\mathcal{S}_{\text{en}}$ is subsequently used to guide the symmetric sampling process for watermark embedding. The whole embedding algorithm is shown in Algorithm 1.

3.2 Watermark attacks

In realistic deployment settings, it is crucial to account for an adversarial attacker who, upon identifying the existence of a watermark, may attempt to erase it while preserving circuit functionality. Here, we consider four representative attack operations: gate replacement, appending, insertion, and deletion; each of which subtly alters the structure of the quantum circuit without compromising its computational correctness (Figure 3).

Replacement. Functionally equivalent gate substitutions, such as replacing a Pauli-X gate with a Hadamard-Z-Hadamard sequence (Figure 3(a)), preserve circuit behavior but alter its internal structure, introducing mismatches with the expected watermark spectral signature.

Appending. Appending gates to the end of unmeasured qubit lines (Figure 3(b)) can leave final measurement outcomes unaffected, yet distort spatial encodings critical for watermark recovery.

Algorithm 1 Watermark embedding.

Input: Watermark sequence $m \in \{0,1\}^k$; QCGMs $\mathcal{G}$; decoder $\mathcal{D}$; ECC encoder **Enc**; number of diffusion steps T.
Output: Watermarked quantum circuit Q_m.
1. Encode watermark sequence:
$\mathcal{S}_{\text{en}} = \textbf{Enc}(m)$; // Apply error correction coding, obtain n-bit codeword
2. Symmetric sampling to generate starting latent: for $i = 1$ **to** n **do**
  Sample $z^i \sim \mathcal{N}(0, I)$; // Sample from standard Gaussian
  Partition $\mathcal{N}(0, I)$ into P_0 and P_1; // Symmetric partition
  if $s^i_{\text{en}} = 0$ **then**
    Adjust z^i to satisfy P_0; // Watermark embedding process
  else
    Adjust z^i to satisfy P_1;
  end
  Set $z^i_T = z^i$;
  Construct starting latent $\mathcal{Z}_T = \{z^1_T, z^2_T, \ldots, z^n_T\}$; // Watermarked latent
end
3. Diffusion process:
Apply T-step diffusion to $\mathcal{Z}_T$ using QCGMs $\mathcal{G}$ to obtain diffused latent $\mathcal{Z}_0$;
4. Decode quantum circuit:
Generate quantum circuit $Q_m = \mathcal{D}(\mathcal{Z}_0)$;
return Q_m.

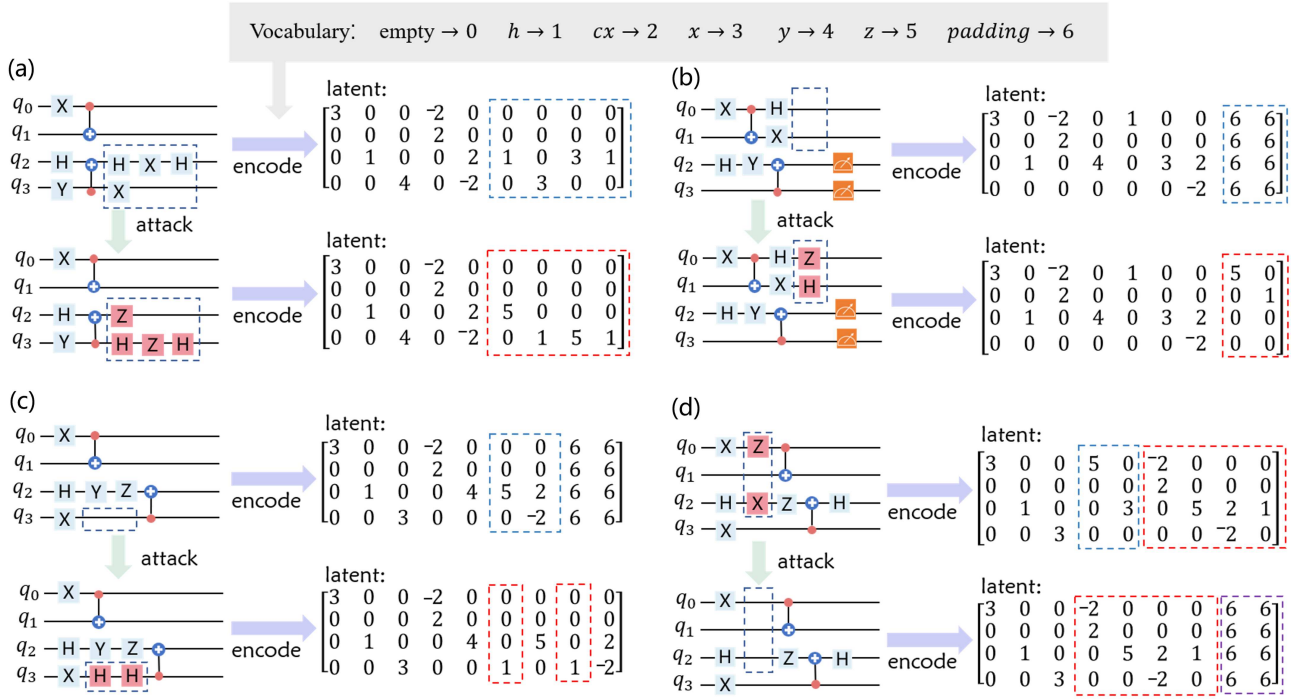

Figure 3 (Color online) Effects of circuit modifications on the encoded matrix. Edits to quantum circuits may alter the encoded matrix either by changing specific values or shifting its structural layout. Subfigures illustrate the impact of different edit operations: (a) replacement, (b) appending, (c) insertion, and (d) deletion. Each operation disrupts the latent of the quantum circuits, introducing distortions that may hinder accurate watermark extraction.

Insertion. Mid-circuit insertions of neutral gate pairs, such as two Hadamard gates satisfying $U^\dagger U = I$, and in particular $U^\dagger = U$ (Figure 3(c)), introduce logically redundant operations that can confound alignment with the embedded watermark structure.

Deletion. Removing non-critical gates (e.g., a Pauli-Z on q_0) can similarly preserve circuit output (Figure 3(d)), while degrading the positional coherence necessary for robust watermark decoding.

3.3 Watermark extraction

To recover the embedded watermark, we first encode the attacked circuit back into the latent domain. An SRM is then applied to the perturbed latent, generating a set of candidate latents that are better aligned with the original sampling structure. This step mitigates distortions introduced by structural edits and enhances alignment between the extracted latent and the expected watermark pattern. Subsequently, we apply denoising diffusion implicit

model (DDIM) inversion to estimate the original watermarked latent from the synchronized candidates. Finally, a reverse sampling procedure, mirroring the SSM used during embedding, is performed, followed by error-correcting code decoding to reconstruct the original watermark bit sequence. The whole extraction algorithm is shown in Algorithm 2. The detailed implementation procedures are as follows.

Algorithm 2 Watermark extraction.

Input: Distorted quantum circuit Q_d; encoder $\mathcal{E}$; ECC decoder **Dec**; threshold τ; number of diffusion steps T.
Output: Extracted watermark $\hat{m}$ or detection result.
1. Encode the circuit:
Obtain latent representation $\hat{\mathcal{Z}}_0 = \mathcal{E}(Q_d)$;
2. Apply DDIM inversion:
Perform T-step DDIM inversion on $\hat{\mathcal{Z}}_0$ to get inverted latent $\hat{\mathcal{Z}}_T$;
3. Standard extraction:
Perform reverse sampling on each element $\hat{z}_T^i$ to generate decoded sequence $\hat{\mathcal{S}}_{\text{de}}$;
Decode watermark $\hat{m} = \mathbf{Dec}(\hat{\mathcal{S}}_{\text{de}})$;
if $\mathcal{L}(m, \hat{m}) > \tau$ **then**
  return $\hat{m}$ (watermark detected); `// Watermark similarity judgment`
end
4. Synchronization and verification (SRM):
• Insert zero-matrix $\mathcal{O}$ of size $f_c \times f_h \times k$ at each column position $i \in [1, f_w - k]$ into $\hat{\mathcal{Z}}_0$ to generate synchronization sets $\bar{\mathbb{Z}}_0$; `// Synchronization process`
• **foreach** $\bar{\mathcal{Z}}_0^i \in \bar{\mathbb{Z}}_0$ **do**
  Perform T-step DDIM inversion on $\bar{\mathcal{Z}}_0^i$ to obtain $\bar{\mathcal{Z}}_T^i$; `// Verification process`
  Perform reverse sampling to generate $\hat{\mathcal{S}}_{\text{de}}^i$;
  Decode $\hat{m}^i = \mathbf{Dec}(\hat{\mathcal{S}}_{\text{de}}^i)$;
  if $\mathcal{L}(m, \hat{m}^i) > \tau$ **then**
    return $\hat{m}^i$ (watermark detected);
  end
end
return No valid watermark detected.

To extract the watermark from the generated quantum circuit Q_m, we first encode Q_m using the quantum circuit encoder $\mathcal{E}$ to obtain the latent representation: $\hat{\mathcal{Z}}_0 = \mathcal{E}(Q_m)$. Subsequently, diffusion model inversion [28] is applied to $\hat{\mathcal{Z}}_0$ over T steps to recover the inverted latent representation $\hat{\mathcal{Z}}_T$, which approximates the original watermarked latent $\mathcal{Z}_T$. For each element $\hat{z}_T^i \in \hat{\mathcal{Z}}_T$, a reverse sampling procedure is performed to obtain a binary sequence $\hat{\mathcal{S}}_{\text{de}}$. Specifically, if $\hat{z}_T^i \in P_0$, the i-th bit $\hat{s}_{\text{de}}^i$ is set to 0; otherwise, it is set to 1. The final extracted watermark $\hat{m}$ is obtained by decoding the sequence $\hat{\mathcal{S}}_{\text{de}} = \{\hat{s}_{\text{de}}^1, \hat{s}_{\text{de}}^2, \ldots, \hat{s}_{\text{de}}^n\}$ using the decoding mechanism **Dec** corresponding to the encoder **Enc**, such that $\hat{m} = \mathbf{Dec}(\hat{\mathcal{S}}_{\text{de}})$. The extracted watermark $\hat{m}$ is then compared with the original watermark m for verification. If the similarity $\mathcal{L}(m, \hat{m})$ exceeds a predefined threshold τ, the generated quantum circuit Q_m is identified as successfully watermarked with m. In this work, the **Dec** procedure can be described as a two-step process. First, the encrypted sequence $\hat{\mathcal{S}}_{\text{en}}$ is decrypted via bitwise XOR with the known pseudorandom key K. Then, majority voting is applied over every v-bit repetition block to reconstruct the original k-bit watermark $\hat{m}$.

We call the above extraction procedure standard extraction. Then to address the latent misalignment caused by structural perturbations, we propose a synchronization restoration mechanism, which aims to proactively correct positional desynchronization and enable reliable watermark extraction under adversarial conditions.

Synchronization restoration mechanism. Given a distorted quantum circuit Q_d, we first obtain its latent representation $\hat{\mathcal{Z}}_0 = \mathcal{E}(Q_d)$ with shape $f_c \times f_h \times f_w$. To compensate for potential misalignment, we construct a zero-initialized matrix $\mathcal{O} \in \mathbb{R}^{f_c \times f_h \times k}$ and systematically inject it at different column positions, where the parameter k $(1 \leqslant k < f_w)$ represents the maximum expected width of the shift to be compensated. For each candidate index $i \in [1, f_w - k]$, we insert $\mathcal{O}$ at the i-th column to generate a compensated latent representation $\bar{\mathcal{Z}}_0^i$, defined as

$$\bar{\mathcal{Z}}_0^i = [\hat{\mathcal{Z}}_0[1:i]; \mathcal{O}; \hat{\mathcal{Z}}_0[i:f_w - k]]. \tag{1}$$

This process yields a set of compensated latents $\bar{\mathbb{Z}}_0 = \bar{\mathcal{Z}}_0^i, i = 1, 2, \ldots, f_w - k$. We then apply the standard extraction procedure to each $\bar{\mathcal{Z}}_0^i \in \bar{\mathbb{Z}}_0$. If any candidate latent leads to a recovered watermark $\hat{m}$ that satisfies the similarity criterion $\mathcal{L}(m, \hat{m}) > \tau$, the circuit Q_d is deemed successfully watermarked.

This strategy hinges on the assumption that $\mathcal{L}(m, \hat{m})$ exceeds the threshold τ with high probability when alignment is correctly restored. Conversely, in cases of failed compensation or absence of a watermark, the likelihood of false detection remains negligibly low, thus ensuring robust and reliable verification.

3.4 Robustness evaluation metric

We evaluate robustness using TPR as the primary metric. The calculation of TPR can be described as follows: for the circuit with watermark s, we extract the potential watermark sequence $\hat{s}$ from the circuit. Then we calculate the similarity of s and $\hat{s}$, denoted as $\mathcal{L}(s, \hat{s})$. In our experiment, $\mathcal{L}$ is bitwise similarity, which is defined as the fraction of positions at which the corresponding bits are identical. Formally, it is given by

$$\mathcal{L}(s, \hat{s}) = \frac{1}{n} \sum_{i=1}^{n} (s_i = \hat{s}_i),$$

where the Boolean expression $(s_i = \hat{s}_i)$ evaluates to 1 if the two bits are equal and 0 otherwise, and n is the length of s and $\hat{s}$. A similarity score of 1 indicates perfect agreement, while a score approaching 0 reflects increasing dissimilarity. A watermarked circuit is considered successfully identified if its similarity score exceeds a predefined threshold τ, which we set to 0.7916. This threshold is derived via hypothesis testing to ensure a fixed false positive rate (FPR) of 10^{-3}. Since each circuit yields a single similarity score, it contributes exactly one instance to the TPR computation. To ensure statistical reliability and broader evaluation coverage, we conduct experiments over 1000 independently generated watermarked circuits. The final true positive rate is computed as

$$\mathrm{TPR} = \frac{N_w}{1000},$$

where N_w denotes the number of circuits for which $\mathcal{L}(s, \hat{s}) \geqslant \tau$.

4 Experiments

We assess the proposed method along two key dimensions: fidelity and robustness. Fidelity is evaluated through both theoretical analysis and empirical results, demonstrating that the watermark embedding process preserves the generation quality of quantum circuits. Robustness is measured using the true positive rate (TPR), the proportion of correctly identified watermarked circuits among all watermarked instances, which quantifies the reliability of watermark detection under various conditions. To further evaluate generalizability, we analyze how TPR varies with the number of inference and inversion steps. An ablation study is also conducted to isolate the impact of the SRM, highlighting its contribution to enhancing resilience against adversarial watermark attacks.

4.1 Implementation details

In this work, we adopt the quantum circuit generative model [6] as the foundation for our experiments. The model operates within a latent space of dimensionality $4 \times 8 \times 48$, corresponding to quantum circuits composed of 8 qubits with up to 48 quantum gates. During generation, we set the guidance scale to 7.5 and the denoising steps to 50. An empty prompt is used to initiate the generation process, ensuring unbiased latent initialization. For the inversion procedure, we similarly adopt 50 inversion steps as the default setting. The message to be embedded is a 24-bit binary sequence. All experiments are implemented using the PyTorch framework and executed on a single NVIDIA RTX 4090 GPU.

4.2 Parameter studies

4.2.1 *Threshold* τ

In this subsection, we will introduce how the threshold τ is determined according to the false positive rate. Recap the extraction process: for a specific watermark m, the proposed Q-Tag model starts with the latent sampled with $\mathcal{S}_{\mathrm{en}}$ to embed the watermark identity information into every generated quantum circuit. Then, assuming $\hat{m}$ is the extracted watermark, the similarity of the two watermark denoted $\mathcal{L}(m, \hat{m})$, is compared with a threshold value τ, when $\mathcal{L}(m, \hat{m}) \geqslant \tau$, the quantum circuit is identified as watermarked.

As discussed in [56], it is usually assumed that the extracted watermark bit $\hat{s}_1, \hat{s}_2, \ldots, \hat{s}_k$ is independent and equally distributed, where $\hat{s}_i$ follows a Bernoulli distribution with parameter 0.5. In this case, $\mathcal{L}(m, \hat{m})$ follows a binomial distribution with parameter $(k, 0.5)$.

Once the distribution of $\mathcal{L}(m, \hat{m})$ is established, the FPR is defined as the probability that the unwatermarked quantum circuit, $\mathcal{L}(m, \hat{m})$, surpasses the threshold τ. This probability can be computed using the regularized

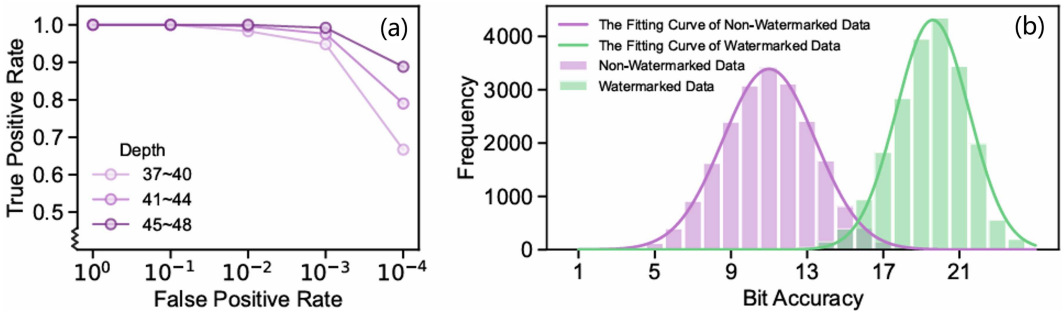

Figure 4 (Color online) Experiment on threshold selection. (a) TPR/FPR; (b) bit accuracy distribution.

incomplete beta function $B_x(a;b)$. Further elaboration of this expression can be found in [42].

$$\begin{aligned}\mathrm{FPR}(\tau) &= \Pr(\mathrm{Acc}(s,\hat{s}) > \tau) = \frac{1}{2^k}\sum_{i=\tau+1}^{k}\binom{k}{i} \\ &= B_{\frac{1}{2}}(\tau+1, k-\tau).\end{aligned} \tag{2}$$

For the threshold, we embed a random watermark m into QCGMs, generate 10000 quantum circuits, and use the test of $\mathcal{L}(m,\hat{m})$. We report the tradeoff between the TPR and the FPR, while varying $\tau \in \{0,\ldots,24\}$. The TPR is measured directly. The FPR is inferred from (2). The experiment is run on 10 random watermarks and we report averaged results. The experimental results are presented in Figure 4. A lower FPR inevitably leads to a reduction in TPR, as stricter detection criteria allow for fewer bit errors and result in a higher threshold for successful extraction. Nevertheless, Q-Tag consistently maintains a TPR above 90% even under the stringent setting of FPR $= 10^{-3}$, demonstrating reliable watermark detectability under conservative verification constraints.

4.2.2 *False positive of synchronization restoration mechanism*

In our extraction process, an SRM is employed to recover potential misalignments in the latent space. While this enhances robustness, it also increases the extraction payload, which may inadvertently raise FPR. To mitigate this, we introduce a similarity threshold th on the similarity score $\mathcal{L}(m,\hat{m})$, ensuring that only highly confident matches are accepted. This subsection describes how we determine an appropriate value for th.

To calibrate the threshold, we conduct a statistical analysis based on Neyman-Pearson theory. Specifically, we fix a watermark message m, and generate 20000 watermarked quantum circuits under various settings. In parallel, another 20000 circuits are generated using random latent initializations, without any embedded watermark. We apply the full extraction pipeline to both sets and record the similarity scores between m and the extracted message $\hat{m}$, producing two empirical distributions.

As shown in Figure 4(b), the resulting distributions of bit-level accuracy for the watermarked and non-watermarked cases approximately follow Gaussian profiles, but with distinct means and variances. We then apply hypothesis testing to determine the decision threshold th, defined under the following hypotheses.

- H_0: the circuit does not contain a watermark (null hypothesis).
- H_1: the circuit contains a valid watermark (alternative hypothesis).

The likelihood function of these two hypotheses can be written as

$$f(x|H_0) = \frac{1}{\sigma_0\sqrt{2\pi}}\cdot \mathrm{e}^{-\frac{(x-\mu_0)^2}{2\sigma_0^2}}, \tag{3}$$

$$f(x|H_1) = \frac{1}{\sigma_1\sqrt{2\pi}}\cdot \mathrm{e}^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}}. \tag{4}$$

In the case of the Neyman-Pearson approach in signal detection, the decision rule is defined as

$$\frac{f(x|H_1)}{f(x|H_0)} \underset{H_1}{\overset{H_0}{\gtrless}} \xi, \tag{5}$$

that is

$$x \underset{H_1}{\overset{H_0}{\gtrless}} g(\xi), \tag{6}$$

where ξ is the decision threshold and $g(\xi)$ is a function of ξ which is determined by (3), (4), and (6). Let $th = g(\xi)$, the decision rule is equivalent to

$$x \underset{H_1}{\overset{H_0}{\gtrless}} th. \tag{7}$$

Then, according to the given probability of false-alarm (denoted by α_0), which is defined as

$$\alpha_0 = \int_{th}^{\infty} f(x|H_0)\, dx. \tag{8}$$

Based on the empirical distributions, we compute the detection threshold th corresponding to a target false positive rate $\alpha_0 = 10^{-3}$. Using the fitted parameters: $\mu_0 = 9.00$, $\sigma_0 = 2.43$, $\mu_1 = 18.60$, and $\sigma_1 = 1.91$, we apply Neyman-Pearson analysis to determine the optimal threshold. According to (8), the resulting value is $th = 16.5$, meaning that circuits with fewer than 16.5 accurately extracted bits are classified as non-watermarked.

In our implementation, we conservatively set the final verification threshold $\tau = 0.7916$, which corresponds to 19 correctly extracted bits-well above the decision boundary of $th = 16.5$. This ensures that both the compensation process and watermark detection maintain a false positive rate below 10^{-3}, thus preserving detection reliability under practical deployment conditions.

4.3 Fidelity evaluation

Achieving high-fidelity watermark embedding is nontrivial, particularly in the context of QCGMs that rely on strict statistical assumptions about their inputs. In QCGMs, the generation process begins from a starting latent vector sampled from a standard Gaussian distribution. This latent serves as the origin of the diffusion trajectory and directly determines the structure of the resulting quantum circuit. Embedding watermark information into this latent is appealing, as it provides a direct and global influence over the output. However, naively modifying the latent to encode information risks violating the Gaussian prior, which may lead to distributional mismatch, degraded generation quality, or even failure modes in the denoising process.

Our motivation, therefore, is to enable watermark embedding at this sensitive entry point while preserving the statistical integrity of the latent space. To this end, we propose a symmetric sampling mechanism that injects watermark information through controlled perturbations that maintain exact adherence to the standard Gaussian distribution. This distribution-preserving design ensures compatibility with the QCGMs' generative assumptions, allowing the watermark to be imperceptibly embedded without compromising fidelity. By addressing the tension between expressiveness and distributional correctness, our approach enables robust and high-quality watermarking within diffusion-based generative pipelines.

In this subsection, we evaluate the effectiveness of the proposed SSM with respect to fidelity. We begin by outlining the design of SSM. Given a standard Gaussian distribution $f(z)$, we split it into two symmetric partition P_0 and P_1, which are mirror-symmetric about $z = 0$. For each bit in the binary watermark sequence to be embedded s^i_{en}, we sample a value $z^i \sim \mathcal{N}(0, I)$. If $z^i \in Ps^i_{\text{en}}$, we set $z^i_T = z^i$; otherwise, we flip its sign and set $z^i_T = -z^i$. For instance, if $s^i_{\text{en}} = 0$, then $P_{s^i_{\text{en}}} = P_0$, and similarly for $s^i_{\text{en}} = 1$.

4.3.1 *Theoretical proof*

We provide a theoretical proof that each element z^i_T generated via the SSM follows a standard Gaussian distribution, i.e., $z^i_T \sim \mathcal{N}(0, I)$.

Theorem 1 (Distribution preservation of SSM)**.** Let $z^i \sim \mathcal{N}(0, I)$ be a standard Gaussian variable, and let $s^i_{\text{en}} \in \{0, 1\}$ be a pseudorandom bit sampled uniformly at random. Let z^i_T be the sample generated by the SSM, defined as

$$z^i_T = \begin{cases} z^i, & \text{if } z^i \in P_{s^i_{\text{en}}}, \\ -z^i, & \text{otherwise,} \end{cases}$$

where P_0 and P_1 are symmetric, equiprobable partitions of the Gaussian distribution with respect to the origin (i.e., $\Pr(z^i \in P_0) = \Pr(z^i \in P_1) = 0.5$, and $P_1 = -P_0$). Then $z^i_T \sim \mathcal{N}(0, I)$.

Proof. To show that $z^i_T \sim \mathcal{N}(0, I)$, it suffices to prove that for any $x \in \mathbb{R}$, the cumulative distribution function satisfies

$$\Pr(z^i_T \leqslant x) = \Pr(y^i \leqslant x),$$

where $y^i \sim \mathcal{N}(0, I)$.

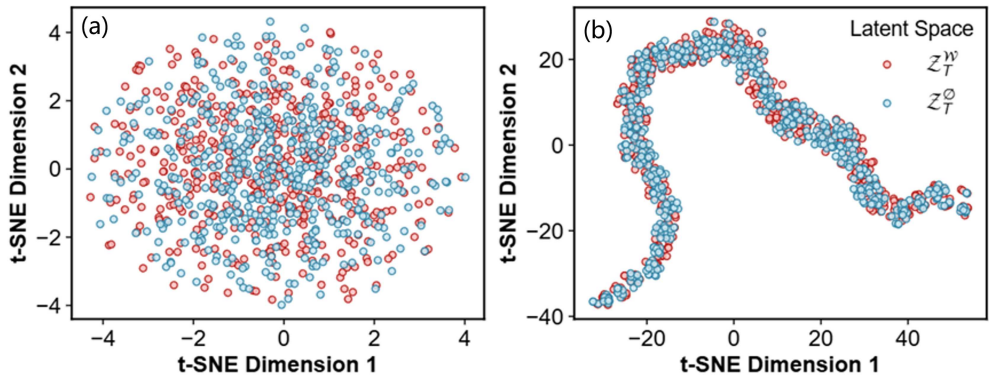

Figure 5 (Color online) Fidelity experiments of Q-Tag generation. (a) The initial starting latent with ($\mathcal{Z}_T^{\mathcal{W}}$) and without ($\mathcal{Z}_T^{\emptyset}$) watermark are indistinguishable through t-SNE; (b) the diffused latent with ($\mathcal{Z}_0^{\mathcal{W}}$) and without ($\mathcal{Z}_0^{\emptyset}$) watermark are also indistinguishable through t-SNE.

According to SSM, we have

$$\begin{aligned}\Pr(z_T^i \leqslant x) &= \Pr(z^i \leqslant x,\, z^i \in P_{s_{\text{en}}^i}) \\ &\quad + \Pr(-z^i \leqslant x,\, z^i \in P_{1-s_{\text{en}}^i}) \\ &= \Pr(z^i \leqslant x) \cdot \Pr(z^i \in P_{s_{\text{en}}^i}) \\ &\quad + \Pr(z^i \geqslant -x) \cdot \Pr(z^i \in P_{1-s_{\text{en}}^i}).\end{aligned} \tag{9}$$

By the symmetry of the Gaussian distribution and the equiprobability of partitions,

$$\begin{aligned}&\Pr(z^i \in P_{s_{\text{en}}^i}) = \Pr(z^i \in P_{1-s_{\text{en}}^i}) = 0.5, \\ &\Pr(z^i \geqslant -x) = \Pr(z^i \leqslant x).\end{aligned}$$

Substituting into (9) gives

$$\begin{aligned}\Pr(z_T^i \leqslant x) &= 0.5 \cdot \Pr(z^i \leqslant x) + 0.5 \cdot \Pr(z^i \leqslant x) \\ &= \Pr(z^i \leqslant x),\end{aligned}$$

which matches the CDF of $\mathcal{N}(0, I)$. Therefore, $z_T^i \sim \mathcal{N}(0, I)$.

4.3.2 *Empirical evidence*

In addition to the theoretical guarantee, we empirically assess fidelity by evaluating the indistinguishability between watermarked and non-watermarked latents. Specifically, we employ t-distributed stochastic neighbor embedding (t-SNE) [57] to visualize their distributions in a reduced dimensional space, allowing us to qualitatively examine whether watermark embedding preserves the underlying structure of the latent space. We generate 500 initial latent vectors embedded with watermarks $\mathcal{Z}_T^{\mathcal{W}}$ and 500 without $\mathcal{Z}_T^{\emptyset}$, and propagate both sets through the generative process to obtain the final latents prior to variational autoencoder (VAE) decoding ($\mathcal{Z}_0^{\mathcal{W}}$ and $\mathcal{Z}_0^{\emptyset}$). These latents are then projected into two-dimensional space using t-SNE, as shown in Figures 5(a) and (b). The resulting visualizations reveal no clear separation between the two distributions, either before or after diffusion. This empirical evidence supports the conclusion that the proposed watermarking strategy preserves the statistical properties of the latent space, introducing no observable distributional bias.

4.4 Robustness evaluation

Ensuring robustness against structural attacks poses a significant challenge in watermark recovery. In practice, watermarked quantum circuits may undergo various attacks, such as gate insertion, deletion, that do not alter their functionality but can induce misalignments in the corresponding latent representations. These latent shifts disrupt the spatial correspondence required for reliable watermark extraction, particularly when the watermark is embedded at the input stage of the generative process.

To address this, we introduce a synchronization restoration mechanism designed to mitigate the effects of such structural attacks. The core idea is to prepend a zero-initialized alignment prior to the inversion process, effectively

Table 1 Comparison of the WQCs. Bold entries indicate the best performance across all compared methods.

Method	Identity	Replacement	Appending	Insertion	Deletion
WQC [19]-RGW	**1.000**	0.398	0.402	0.405	0.201
WQC [19]-RGIW	**1.000**	0.403	0.397	0.503	0.195
Ours	**1.000**	**1.000**	**0.992**	**1.000**	**1.000**

re-anchoring the latent structure and restoring compatibility with the original watermark embedding pattern. This lightweight correction allows the inversion path to converge more reliably to the intended watermarked latent, even under adversarial or noisy conditions. Details can be found in the SRM within Subsection 3.3.

We evaluate the robustness of the proposed watermarking framework by measuring the TPR of watermark detection under both clean and adversarial conditions. Specifically, we first assess TPR across different circuit lengths without any attacks, followed by experiments simulating four categories of adversarial watermark erasure.

4.4.1 *TPR under clean conditions*

Detection performance without distortion is shown in Figure 6(a), with the number of quantum gates ranging from 33 to 48. Across all configurations, the proposed method achieves a consistently high TPR ($\geqslant 90\%$), demonstrating reliable watermark embedding and extraction within QCGMs-generated circuits. A clear upward trend is observed as the gate count increases, which can be attributed to the increased latent dimensionality. Specifically, the latent space width f_w scales with gate number, and the overall latent size is given by $f_c \times f_h \times f_w$, matching the codeword length used for watermark embedding. Given a fixed message size (24 bits), longer codewords offer more redundancy for error correction, thereby improving detection reliability. In practice, increasing the default number of circuit gates further enhances robustness.

4.4.2 *TPR under attacks*

We next evaluate the framework's robustness against four types of attacks intended to erase the embedded watermark.

Gate replacement. Figure 6(b) presents the TPR after randomly replacing 1–5 gates in watermarked circuits. Across all gate counts and perturbation levels, the TPR remains above 85%, and approaches 100% for circuits containing 45–48 gates, confirming the method's resilience to function-preserving substitutions.

Gate appending. Appending operations add 1–5 gates to the end of selected qubit lines. As shown in Figure 6(c), the TPR consistently exceeds 90%, indicating robustness against appending-based obfuscation strategies that preserve output semantics but alter structure.

Gate insertion. In this scenario, 1–2 pairs of logically neutral gates are inserted into random circuit positions (e.g., self-inverse pairs such as Hadamard-Hadamard). Detection performance remains strong across all settings, with TPR exceeding 95%, as shown in Figure 6(d), demonstrating insensitivity to insertion-based perturbations.

Gate deletion. We randomly remove 1–3 gates from each watermarked circuit and evaluate the resulting TPR. As shown in Figure 6(e), the TPR remains above 90% in all configurations, indicating that the watermark retains integrity even under structural deletions.

4.4.3 *Comparative studies*

To highlight the advantages of embedding watermarks during the circuit generation process, we compare our framework against representative watermarking approaches-specifically, the structural watermarking method of Roy et al. [19], which inserts redundant gate pairs into a generated circuit while preserving functionality. To ensure a fair and consistent comparison, all methods are evaluated under identical experimental conditions: circuits are generated with 8 qubits and the number of quantum gates ranging from 45 to 48. Additionally, we also evaluate the performance under the four types of attacks described above: gate replacement, gate appending, gate insertion, and gate deletion. The experimental results are presented in Table 1. In the absence of attacks, all methods achieve 100% accuracy in watermark extraction for copyright verification. However, under attack scenarios, the method of Roy et al. [19] often fails to retain its embedded watermarks due to structural perturbations, resulting in extraction failure. In contrast, our method demonstrates strong robustness, reliably preserving watermark integrity and enabling successful extraction even in the presence of such attacks.

The method of Roy et al. [19] typically requires structural modifications to encode watermarks. This dependency on structural changes renders them vulnerable to structural attacks, which can inadvertently remove or corrupt the watermark. In contrast, a critical advantage of our proposed method lies in its provably lossless nature and

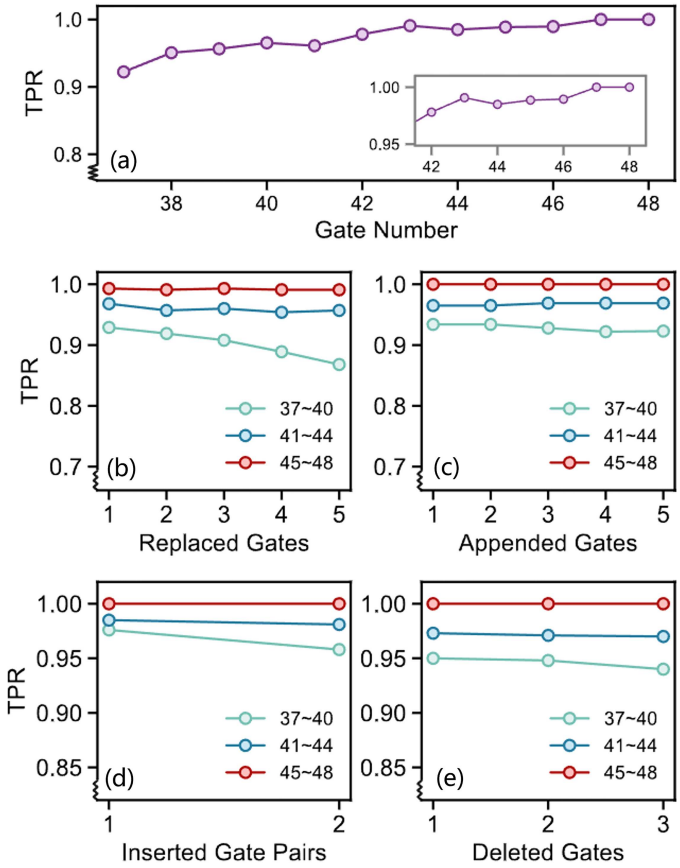

Figure 6 (Color online) True positive rate of watermark detection under different settings: (a) no-distortion, (b) replacement gates, (c) appending gates, (d) inserting gates, and (e) deleting gates.

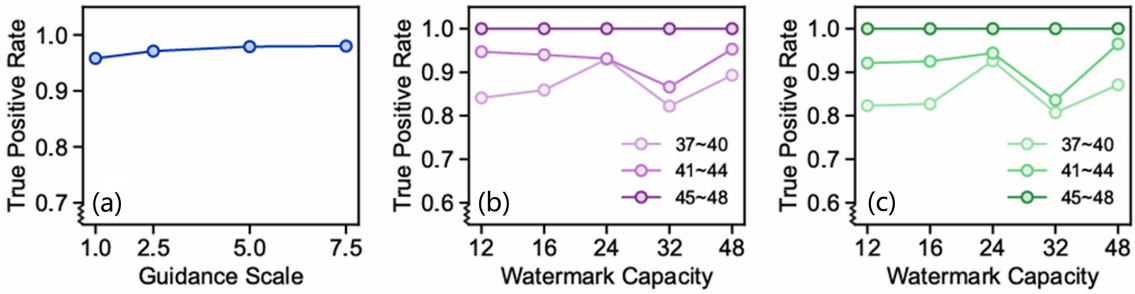

Figure 7 (Color online) (a) Comparison of the influence of different levels of text guidance on watermark extraction capability; (b) different capacity sizes on watermark extraction capability without attacks; (c) different capacity sizes on watermark extraction capability under attacks.

its ability to preserve the original quantum circuit structure, as theoretically demonstrated and experimentally validated in Subsection 4.3. Furthermore, combined with the SRM, our approach significantly mitigates the impact of structural attack on watermark extraction.

4.4.4 *Generalizability*

Robustness to varying guidance scales. In real-world scenarios, quantum circuit generation may be performed under different guidance scales, making it essential for the watermarking algorithm to remain effective across a range of such settings. To evaluate this, we generate 1000 watermarked quantum circuits for each of four different guidance scales: 1.0, 2.5, 5.0, and 7.5. The corresponding watermark extraction results are summarized in Figure 7(a). As the figure shows, the TPR remains consistently high (>95%) across all tested scales, demonstrating that the proposed Q-Tag framework is robust to variations in guidance strength. This adaptability highlights the method's practical relevance for diverse generation configurations.

Impact of watermark capacity. We evaluate the TPR of watermark detection under varying watermark embedding

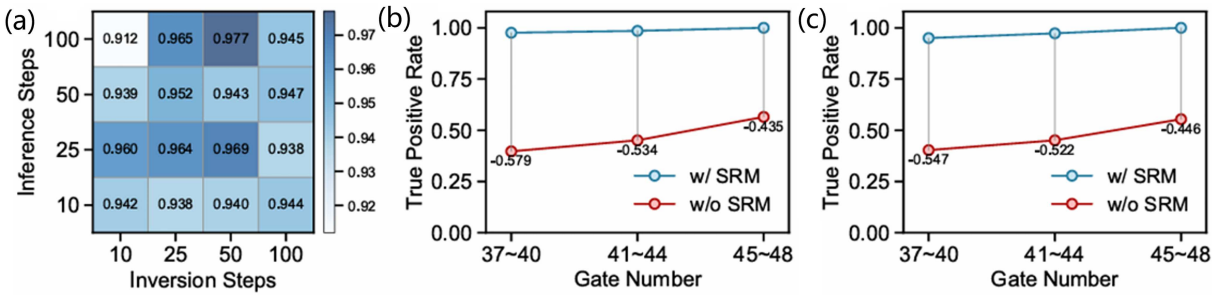

Figure 8 (Color online) Generalizability and ablation experiments. (a) The TPR of watermark detection with different inference and inversion steps; (b) the TPR of watermark detection with and without SRM under insertion attacks; (c) the TPR of watermark detection with and without SRM under deletion attacks.

capacities. Specifically, we test five capacity settings: 12, 16, 24, 32, and 48 bits. For each setting, we measure the TPR both without distortion and under distortion (including replacement and adversarial attacks), as shown in Figures 7(b) and (c), respectively. To ensure consistency, the FPR is fixed at 10^{-3}, and the corresponding detection thresholds are calibrated according to each bit length.

Watermark detection with different inference and inversion steps. In our experiments, the default configuration employs 50 generation steps during sampling and an identical 50-step inversion process for watermark extraction. However, real-world deployments may involve varied inference schedules depending on computational constraints or design preferences. To assess the generalization capability of the proposed framework under such conditions, we perform a series of cross-testing experiments by varying the number of generation steps (10, 25, 50, 100) and matching the inversion steps accordingly. The resulting TPR values across all tested combinations are reported in Figure 8(a). In all cases, the framework maintains a TPR exceeding 91%, demonstrating strong robustness and generalization across diverse generation-inversion step configurations.

The results show that increasing the embedding capacity from 12 to 24 bits improves TPR, owing to a larger tolerance for bit errors under a fixed FPR-thereby enhancing detection reliability. However, when the capacity increases beyond 24 bits, TPR begins to decline. This is attributed to reduced redundancy in the error correction code, which weakens its ability to recover from distortion-induced bit errors. It is worth noting that, independent of embedding capacity, increasing the number of gates in the quantum circuit consistently improves TPR, due to the extended latent space available for robust watermark embedding.

4.4.5 *Importance of SRM*

To improve robustness against watermark attacks, we introduce a synchronization restoration mechanism. To evaluate its effectiveness, we conduct an ablation study. A random watermark m is embedded into the model, and 10000 quantum circuits (QCs) are generated. Each circuit is then subjected to insertion and deletion attacks, after which watermark extraction is performed both with and without SRM. The results are summarized in Figures 8(b) and (c), where "w/ SRM" and "w/o SRM" denote extraction with and without synchronization restoration, respectively; negative values indicate the accuracy difference between "w/ SRM" and "w/o SRM". The TPR achieved with SRM is consistently and significantly higher than that without SRM, with improvements exceeding 40 percentage points. These findings underscore the critical role of SRM in mitigating desynchronization effects introduced by structural perturbations, and highlight its importance in maintaining detection reliability under adversarial conditions.

5 Conclusion

In conclusion, this work introduces the first watermarking framework explicitly designed for QCGMs, addressing a nontrivial challenge at the intersection of quantum design and model-level intellectual property protection. Our method departs from conventional circuit-level watermarking by embedding ownership signals directly into the generative process through a symmetric sampling mechanism, which non-invasively aligns watermark encoding with the model's latent Gaussian prior-preserving fidelity without disrupting the generative distribution. To ensure robustness, we develop a synchronization restoration mechanism capable of recovering watermark integrity under structural perturbations. Beyond its immediate application to quantum circuits, the architectural principle underlying SRM, resynchronization of misaligned generative trajectories, offers a general insight for improving robustness in watermarking across structurally dynamic domains such as image editing, neural program synthesis, and video

generation. This generalizability positions our framework as a foundational contribution to secure and accountable generation in both quantum and classical settings.

While this study establishes the first watermarking framework for quantum circuit generative models, several key challenges remain open. The current synchronization compensation mechanism relies on handcrafted heuristics, which may prove inadequate against sophisticated or adaptive adversaries. Furthermore, the watermark embedding capacity is limited, and the extraction process assumes a trusted environment. Moreover, our current method is constrained to generative models based on reversible diffusion mechanisms. Future work could explore adaptive, learning-based alignment strategies, scalable embedding techniques with higher payload efficiency, cryptographically verifiable extraction protocols, and a universal watermarking framework. These directions would not only improve the resilience and scalability of quantum watermarking but also extend its applicability to a broader class of generative models. As quantum circuit generation increasingly underpins quantum software development, our framework lays the groundwork for secure, verifiable, and accountable content creation in the age of AI-assisted quantum design.

Acknowledgements This work was supported in part by National Natural Science Foundation of China (Grant No. 62272003), Quantum Science and Technology-National Science and Technology Major Project (Grant No. 2021ZD0302300), Science and Technology Major Project of Anhui Province (Grant No. 202423s06050001), and National Key Research and Development Program of China (Grant Nos. 2023YFB4502500, 2024YFB4504100).

References

1 Biamonte J, Wittek P, Pancotti N, et al. Quantum machine learning. Nature, 2017, 549: 195–202
2 Li Q Y, Huang Y H, Jin S, et al. Quantum spectral clustering algorithm for unsupervised learning. Sci China Inf Sci, 2022, 65: 200504
3 Li G X, Zhao X Q, Wang X. Quantum self-attention neural networks for text classification. Sci China Inf Sci, 2024, 67: 142501
4 Tian J, Sun X, Du Y, et al. Recent advances for quantum neural networks in generative learning. IEEE Trans Pattern Anal Mach Intell, 2023, 45: 12321–12340
5 Ruiz F J R, Laakkonen T, Bausch J, et al. Quantum circuit optimization with AlphaTensor. Nat Mach Intell, 2025, 7: 374–385
6 Fürrutter F, Muñoz-Gil G, Briegel H J. Quantum circuit synthesis with diffusion models. Nature Mach Intell, 2024, 6: 515–524
7 Shi J, Li Z, Lai W, et al. Two end-to-end quantum-inspired deep neural networks for text classification. IEEE Trans Knowl Data Eng, 2021, 35: 4335–4345
8 Robledo-Moreno J, Motta M, Haas H, et al. Chemistry beyond the scale of exact diagonalization on a quantum-centric supercomputer. Sci Adv, 2025, 11: eadu9991
9 Wang M, Long G L. Lattice-based access authentication scheme for quantum communication networks. Sci China Inf Sci, 2024, 67: 222501
10 Shi J, Chen T, Lai W, et al. Pretrained quantum-inspired deep neural network for natural language processing. IEEE Trans Cybern, 2024, 54: 5973–5985
11 Mi X, Roushan P, Quintana C, et al. Information scrambling in quantum circuits. Science, 2021, 374: 1479–1483
12 Yang M, Guo X, Jiang L. Multi-stage watermarking for quantum circuits. In: Proceedings of 2024 IEEE International Conference on Quantum Computing and Engineering (QCE), 2024. 796–804
13 AbuGhanem M. IBM quantum computers: evolution, performance, and future directions. 2024. ArXiv:2410.00916
14 Gonzalez C. Cloud based QC with Amazon Braket. Digitale Welt, 2021, 5: 14–17
15 Wang S B, Wang P, Li G H, et al. Variational quantum eigensolver with linear depth problem-inspired ansatz for solving portfolio optimization in finance. Sci China Inf Sci, 2025, 68: 180504
16 Majumder S, Ray S, Dasgupta M, et al. QuanFraud: quantum state verification scheme for fraud detection in IoT-assisted quantum-blockchain networks. IEEE Trans Serv Comput, 2026, 19: 616–627
17 Xu J C, Fang H, Yang Y, et al. AI-generated image detection algorithm based on classical-quantum hybrid neural network. Sci China Inf Sci, 2026, 69: 112501
18 Coupel J, Farheen T. Security vulnerabilities in quantum cloud systems: a survey on emerging threats. 2025. ArXiv:2504.19064
19 Roy R, Ghosh S. Watermarking of Quantum Circuits. In: Proceedings of the 26th International Symposium on Quality Electronic Design (ISQED), 2025. 1–7
20 He X L, Xu G W, Han X S, et al. Artificial intelligence security and privacy: a survey. Sci China Inf Sci, 2025, 68: 181101
21 Farhi E, Goldstone J, Gutmann S. A quantum approximate optimization algorithm. 2014. ArXiv:1411.4028
22 Preskill J. Quantum computing in the NISQ era and beyond. Quantum, 2018, 2: 79
23 Bolens A, Heyl M. Reinforcement learning for digital quantum simulation. Phys Rev Lett, 2021, 127: 110502
24 Shen Y. Prepare ansatz for VQE with diffusion model. 2023. ArXiv:2310.02511
25 F?sel T, Niu M Y, Marquardt F, et al. Quantum circuit optimization with deep reinforcement learning. 2021. ArXiv:2103.07585
26 Preti F, Schilling M, Jerbi S, et al. Hybrid discrete-continuous compilation of trapped-ion quantum circuits with deep reinforcement learning. Quantum, 2024, 8: 1343
27 Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models. In: Proceedings of Advances in Neural Information Processing Systems, 2020. 6840–6851
28 Song J, Meng C, Ermon S. Denoising diffusion implicit models. 2020. ArXiv:2010.02502
29 Lu C, Zhou Y, Bao F, et al. Dpm-solver: a fast ODE solver for diffusion probabilistic model sampling in around 10 steps. In: Proceedings of Advances in Neural Information Processing Systems, 2022. 5775–5787
30 Rombach R, Blattmann A, Lorenz D, et al. High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022. 10684–10695
31 Ho J, Salimans T. Classifier-free diffusion guidance. 2022. ArXiv:2207.12598
32 Podell D, English Z, Lacey K, et al. SDXL: improving latent diffusion models for high-resolution image synthesis. 2023. ArXiv:2307.01952
33 Kong Z, Ping W, Huang J, et al. Diffwave: a versatile diffusion model for audio synthesis. 2020. ArXiv:2009.09761
34 Sener F, Saraf R, Yao A. Transferring knowledge from text to video: zero-shot anticipation for procedural actions. IEEE Trans Pattern Anal Mach Intell, 2022, 45: 7836–7852
35 Watson J L, Juergens D, Bennett N R, et al. De novo design of protein structure and function with RFdiffusion. Nature, 2023, 620: 1089–1100
36 Zhao Y, Pang T, Du C, et al. A recipe for watermarking diffusion models. 2023. ArXiv:2303.10137
37 Cui Y, Ren J, Xu H, et al. DiffusionShield: a watermark for data copyright protection against generative diffusion models. SIGKDD Explor Newsl, 2025, 26: 60–75

38 Cui Y, Ren J, Lin Y, et al. FT-Shield: a watermark against unauthorized fine-tuning in text-to-image diffusion models. SIGKDD Explor Newsl, 2025, 26: 76–88
39 Wang Z, Chen C, Lyu L, et al. Diagnosis: detecting unauthorized data usages in text-to-image diffusion models. 2023. ArXiv:2307.03108
40 Zhu P, Takahashi T, Kataoka H. Watermark-embedded adversarial examples for copyright protection against diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024. 24420–24430
41 Liu H, Sun Z, Mu Y. Countering personalized text-to-image generation with influence watermarks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024. 12257–12267
42 Fernandez P, Couairon G, Jégou H, et al. The stable signature: rooting watermarks in latent diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023. 22466–22477
43 Xiong C, Qin C, Feng G, et al. Flexible and secure watermarking for latent diffusion model. In: Proceedings of the 31st ACM International Conference on Multimedia, 2023. 1668–1676
44 Meng Z, Peng B, Dong J. Latent watermark: inject and detect watermarks in latent diffusion space. 2024. ArXiv:2404.00230
45 Kim C, Min K, Patel M, et al. Wouaf: weight modulation for user attribution and fingerprinting in text-to-image diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024. 8974–8983
46 Zhang X, Li R, Yu J, et al. Editguard: versatile image watermarking for tamper localization and copyright protection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024. 11964–11974
47 Wen Y, Kirchenbauer J, Geiping J, et al. Tree-ring watermarks: fingerprints for diffusion images that are invisible and robust. 2023. ArXiv:2305.20030
48 Zhang L, Liu X, Martin A V, et al. Robust image watermarking using stable diffusion. 2024. ArXiv:2401.04247
49 Liu G H, Chen T, Theodorou E, et al. Mirror diffusion models for constrained and watermarked generation. In: Proceedings of Advances in Neural Information Processing Systems, 2023. 42898–42917
50 Yang Z, Zeng K, Chen K, et al. Gaussian shading: provable performance-lossless image watermarking for diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024. 12162–12171
51 Saravanan V, Saeed S M. Decomposition-based watermarking of quantum circuits. In: Proceedings of the 22nd International Symposium on Quality Electronic Design (ISQED), 2021. 73–78
52 Liu Y, John J, Wang Q. E-loq: enhanced locking for quantum circuit IP protection. In: Proceedings of IEEE International Symposium on Hardware Oriented Security and Trust (HOST), 2025. 67–77
53 Rehman A, Langford V, John J, et al. Opaque: obfuscating phase in quantum circuit compilation for efficient IP protection. In: Proceedings of the 26th International Symposium on Quality Electronic Design (ISQED), 2025. 1–6
54 Wang Q, John J, Dong B, et al. Tetrislock: quantum circuit split compilation with interlocking patterns. In: Proceedings of the 62nd ACM/IEEE Design Automation Conference (DAC), 2025. 1–7
55 Das S, Ghosh S. Secure quantum circuit compilation methodology for untrusted compilers. In: Proceedings of IEEE International Conference on Quantum Computing and Engineering (QCE), 2025. 2191–2201
56 Yu N, Skripniuk V, Abdelnabi S, et al. Artificial fingerprinting for generative models: rooting deepfake attribution in training data. In: Proceedings of the IEEE/CVF International conference on computer vision, 2021. 14448–14457
57 Maaten L, Hinton G. Visualizing data using t-SNE. J Mach Learn Res, 2008, 9: 2579–2605