# FRAME GENERATION IN HILBERT SPACE: GENERATIVE INTERPOLATION OF MEASUREMENT DATA FOR QUANTUM PARAMETER ADAPTATION*

**Chen-Yu Liu**[1] **Kuan-Cheng Chen**[2,3] **Samuel Yen-Chi Chen**[4] **Wei-Hao Huang**[5]
**Wei-Jia Huang**[6] **Yen-Jui Chang**[7,8]
[1] Graduate Institute of Applied Physics, National Taiwan University, Taipei, Taiwan
[2] Department of Electrical and Electronic Engineering, Imperial College London, London, UK
[3] Centre for Quantum Engineering, Science and Technology (QuEST), Imperial College London, London, UK
[4] Wells Fargo, New York, NY, USA
[5] Jij Inc., Tokyo, Japan
[6] Hon Hai (Foxconn) Research Institute, Taipei, Taiwan
[7] Quantum Information Center, Chung Yuan Christian University, Taoyuan City, Taiwan
[8] Master Program in Intelligent Computing and Big Data, Chung Yuan Christian University, Taoyuan City, Taiwan
d10245003@g.ntu.edu.tw, kuan-cheng.chen17@imperial.ac.uk,
yen-chi.chen@wellsfargo.com, w.huang@j-ij.com,
wei-jia.huang@foxconn.com, aceest@cycu.edu.tw

## ABSTRACT

Quantum Parameter Adaptation (QPA) has emerged as a promising approach for leveraging quantum neural networks (QNNs) to generate classical neural network (NN) parameters, enabling parameter-efficient fine-tuning of large language models (LLMs). However, the practical implementation of QPA is hindered by the need for an extremely large number of quantum measurement shots, posing a significant challenge in real-world quantum computing environments. To address this issue, this work introduces Generative Interpolation (GI), a method inspired by frame generation in video data, where missing measurement probabilities are interpolated using a neural network-based generative model. By treating quantum measurement probabilities as analogous to video frames, GI estimates unmeasured basis state probabilities, significantly reducing the required quantum measurements. Empirical results demonstrate that incorporating GI into QPA reduces the quantum measurement shot requirement to just 2.5% of the original count while achieving superior fine-tuning performance. This method not only enhances QPA efficiency but also establishes a broader connection between classical deep learning techniques and quantum measurement reconstruction. The proposed generative framework has the potential to extend to variational quantum algorithms, offering a pathway toward reducing quantum measurement overhead in hybrid quantum-classical computing paradigms.

## 1 INTRODUCTION

The recent surge in quantum-centric supercomputing (Alexeev et al., 2024) has paved the way for a new era of computational advancements, leveraging quantum mechanical principles to solve complex problems beyond the capabilities of classical systems (Nielsen & Chuang, 2010). One of the most promising applications of this paradigm is Quantum Machine Learning (QML), which explores quantum-enhanced models to accelerate learning tasks and improve efficiency in large-scale computations. Various quantum neural network (QNN) architectures have demonstrated theoretical advantages in certain domains (Abbas et al., 2021; Du et al., 2020; Caro et al., 2022) and have

*The views expressed in this article are those of the authors and do not represent the views of Wells Fargo. This article is for informational purposes only. Nothing contained in this article should be construed as investment advice. Wells Fargo makes no express or implied warranties and expressly disclaims all legal, tax, and accounting implications related to this article.

been shown empirically successful in various use cases related to optimization (Farhi et al., 2014), classification (Farhi & Neven, 2018; Mitarai et al., 2018; Mari et al., 2020; Chen et al., 2021; Qi et al., 2023; Chen et al., 2024b), generative modeling (Zoufal et al., 2019), sequence learning (Chen et al., 2022; Li et al., 2023; Yang et al., 2022; Di Sipio et al., 2022; Stein et al., 2023), reinforcement learning (Chen et al., 2020; Skolik et al., 2022; Jerbi et al., 2021; Yun et al., 2022) and quantum chemistry (Tilly et al., 2022; Chen et al., 2024c).

Despite the growing interest in QML, several fundamental challenges hinder its practical adoption. Two major obstacles are data encoding inefficiency and quantum hardware inference constraints. Data encoding remains a bottleneck, as classical-to-quantum data transformation often requires a large number of qubits and circuit depth, limiting scalability. Additionally, running inference directly on quantum hardware is costly and impractical due to decoherence and error rates in near-term quantum devices. To address these limitations, Quantum-Train (QT) (Liu et al., 2024a;b) has been proposed as an alternative paradigm. Instead of using quantum circuits for direct inference, QT employs QNNs to generate classical neural network (NN) parameters, effectively removing the need for quantum inference while still benefiting from quantum learning mechanisms (Lin et al., 2024; Liu et al., 2024c; Liu & Chen, 2024; Chen et al., 2024a).

Building on this approach, recent work has successfully scaled QT to fine-tune Large Language Models (LLMs) (Liu et al., 2025), introducing a novel method known as Quantum Parameter Adaptation (QPA). This approach utilizes a QNN to generate the trainable parameters for parameter-efficient fine-tuning (PEFT) methods, such as LoRA, reducing the overall number of parameters while preserving the performance. However, the practical implementation of QPA faces a major challenge—it requires an extremely large number of measurement shots (in the scale of Hilbert space size) from a quantum computer to estimate measurement probabilities accurately. The measurement process must reconstruct the probability distribution over basis states in a high-dimensional Hilbert space, leading to an exponential cost in quantum resources.

To tackle this issue, we draw inspiration from frame generation techniques in video data processing (Liu et al., 2019; Zhao et al., 2019; Kwon & Park, 2019). In classical computing, frame generation is used to enhance video quality and optimize performance, allowing a video recorded at a lower frame rate (e.g., 30 fps) to be displayed at a higher frame rate (e.g., 60 fps) by generating missing frames through deep learning models. This concept aligns with the problem faced in QPA, where certain basis states remain unmeasured due to limited quantum measurement shots, creating missing probability values. By treating missing quantum measurement probabilities as missing data, we propose the use of generative models to interpolate these values efficiently.

In this work, we introduce Generative Interpolation (GI), a novel approach to reconstructing missing measurement probabilities in QPA. Our contributions can be summarized as follows:

- **Generative Interpolation for Efficient Quantum Parameter Adaptation.** A generative model-based interpolation method is introduced to estimate unmeasured quantum measurement probabilities, significantly reducing the number of required quantum measurement shots. By integrating GI into QPA, the required quantum measurements are reduced to just $2.5\%$ of the original shot count, while achieving even better fine-tuning performance for LLMs with substantially fewer quantum resources.
- **Bridging Classical Deep Learning and Quantum Measurement Interpolation.** The conceptual relationship between frame generation in video data and quantum measurement interpolation is established, providing a novel perspective on how classical deep learning techniques can enhance QML. Empirical validation of the proposed GI method demonstrates its effectiveness in improving QPA, enhancing robustness, and reducing testing perplexity under limited measurement conditions.

## 2 Related works

**Fine-Tuning LLMs by Quantum Parameter Adaptation.** LLMs exhibit superior capability in processing complex tasks. However, fine-tuning these models to fit specific domain knowledge remains challenging due to their large parameter space and computational demands. PEFT methods (Hu et al., 2021; Liu et al., 2024d; Li & Liang, 2021; Yang et al., 2021; Houlsby et al., 2019) have been introduced to address this issue by reducing the number of trainable parameters while maintaining performance. Quantum Parameter Adaptation (QPA) (Liu et al., 2025) is a quantum circuit-based approach that further optimizes PEFT by leveraging QNNs to generate the trainable parameters, achieving even greater parameter efficiency. While QPA has demonstrated promising results, a major challenge persists: it requires an exponentially large number of quantum measurement shots, scaling with the size of the Hilbert space, to accurately estimate the measurement probabilities of basis states.

**Quantum State Reconstruction.** Reconstructing quantum states from measurement data is a fundamental challenge in quantum information science. Traditional approaches such as full quantum state tomography and shadow tomography have been developed to infer quantum states based on a series of measurement outcomes (Cramer et al., 2010; Christandl & Renner, 2012; Hsu et al., 2024). Full quantum tomography aims to reconstruct the entire density matrix of a quantum system but suffers from exponential scaling in resource requirements, making it impractical for large quantum systems. On the other hand, shadow tomography reduces the measurement burden by estimating specific properties of quantum states rather than reconstructing the entire state, leveraging randomized measurements and classical post-processing. While these methods improve efficiency, they still demand a significant number of measurements and computational resources, particularly for high-dimensional quantum systems.

**Generative Approaches for Quantum State Reconstruction.** To mitigate the high resource costs of traditional reconstruction techniques, machine learning-based approaches have been explored. These methods employ generative models to estimate unmeasured quantum basis states and reconstruct quantum states with fewer measurements. One approach utilizes pre-trained models to infer missing measurement probabilities, requiring high-precision data during training to generalize effectively (Ahmed et al., 2021; Zhu et al., 2022). Another direction, explored by (Carrasquilla et al., 2019), proposes a sampling-based generative model where a NN wave function ansatz is trained to reconstruct measurement data efficiently. These approaches leverage the ability of deep learning to approximate complex quantum probability distributions, offering a scalable alternative to traditional tomography. However, they still face challenges related to training stability, interpretability, and the generalization of learned quantum representations across different systems.

Although the approaches discussed above appear promising, they are unlikely to fully address the challenges faced by QPA. One major limitation is the requirement for pre-training a generative model, which raises critical questions about how such a model should be trained and how its generalization capability can be ensured across different QPA tasks. Without a robust training strategy and reliable generalization, the model may struggle to accurately reconstruct missing measurement probabilities, limiting its practical applicability. Another concern is the computational complexity shift introduced by sampling-based reconstruction methods. If the generative model requires an extensive number of samples to reconstruct the quantum state, this merely relocates the computational burden rather than alleviating it. Instead of reducing the number of required quantum measurement shots, such methods introduce inefficiencies elsewhere in the process, failing to provide a truly scalable solution.

An ideal approach should address these limitations by avoiding pre-training requirements and ensuring that the generative model learns dynamically alongside the QPA process. This would allow the model to adapt in real time, rather than relying on a fixed pre-trained network that may not generalize well. Additionally, the method should predict the **value** of missing measurement probabilities directly, eliminating the need for an additional sampling layer that could introduce unnecessary computational overhead. By adhering to these principles, a more efficient and practical integration of generative models into QPA can be achieved, significantly reducing the reliance on large-scale quantum measurements.

# 3 Generative Interpolation as an Efficient Approach for Quantum Parameter Adaptation

The role of the parameterized quantum circuit (or quantum neural network) in the QPA is to generate measurement probabilities corresponding to the basis states in the Hilbert space. In classical simulations where exact probabilities can be computed, this method demonstrates promising results. However, when working with real quantum hardware, these probabilities must be estimated from a

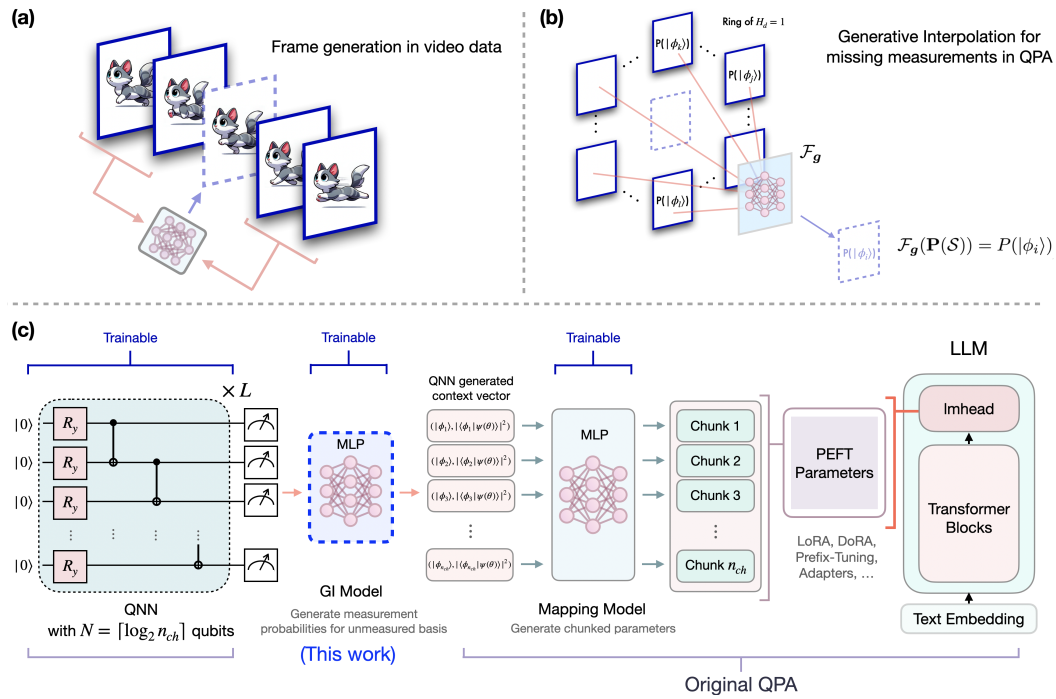


Figure 1: Overview of: (a) the general framework for frame generation in video data (Liu et al., 2019; Zhao et al., 2019; Kwon & Park, 2019), (b) the proposed Generative Interpolation method for addressing missing measurements in QPA, and (c) the integration of Generative Interpolation with QPA, highlighting its contribution in enhancing the input data for the mapping model.

finite number of measurement shots, introducing statistical noise. In this section, we describe the generative interpolation approach and explain why it is well-suited for the QPA framework.

### 3.1 Finite measurement issue of Quantum Parameter Adaptation

To understand the issue of QPA, we first need to introduce its formulation. Following the approach in (Liu et al., 2025), we consider a parameter generation process that deviates from conventional QML methods. Let $\boldsymbol{w} = (w_1, w_2, \ldots, w_m)$ represent the parameters of a target NN model, where $m$ denotes the number of parameters. To generate these parameters, we construct a parameterized quantum circuit (PQC) with $N = \lceil \log_2 m \rceil$ qubits and $L$ layers, using a predefined circuit ansatz:

$$|\psi(\boldsymbol{\theta})\rangle = \left( \prod_{i=1}^{N-1} \text{CNOT}^{i,i+1} \prod_{j=1}^{N} R_Y^j(\theta_j^{(L)}) \right)^L |0\rangle^{\otimes N}. \tag{1}$$

Here, each single-qubit rotation gate $R_Y^j$ is parameterized by $\theta_j^{(L)}$, where $j$ indexes the qubits, and $L$ represents the circuit depth. The controlled-NOT (CNOT) gates create entanglement between qubits. Since the Hilbert space size is $2^N$ and satisfies $2^N \geq m$, the PQC generates $2^N$ distinct measurement probabilities, given by $|\langle \phi_i | \psi(\boldsymbol{\theta}) \rangle|^2 \in [0, 1]$ for $i \in \{1, 2, \ldots, 2^N\}$. The number of parameters in $\boldsymbol{\theta}$ is determined by $N$ and $L$, where $L$ is a hyperparameter similar to depth settings in classical machine learning. Typically, $L$ scales as $O(N)$ or in some cases $O(N^2)$ (Cerezo et al., 2021; Sim et al., 2019), though a more general polynomial scaling, $O(\text{poly}(N))$, is also possible. With polynomial-depth layers, the PQC can generate $2^{\lceil \log_2 m \rceil} \geq m$ measurement probabilities using $O(\text{polylog}(m))$ quantum parameters. At this stage, the measurement probabilities lie within the range $[0, 1]$. To transform these into the target parameters $\boldsymbol{w} \in \mathbb{R}^m$, we introduce a mapping function $\mathcal{M}$, implemented as a multilayer perceptron (MLP) with tunable parameters $\boldsymbol{v}$. The input to $\mathcal{M}$ consists of the binary representation of each basis state (of length $N$) along with its corresponding measurement probability. This mapping is defined as:

$$\mathcal{M}_{\boldsymbol{v}}(|\phi_i\rangle, |\langle\phi_i|\psi(\boldsymbol{\theta})\rangle|^2) = w_i, \quad \forall i \in \{1, 2, \ldots, m\}. \tag{2}$$

Here, we utilize only the first $m$ basis states to ensure complete coverage of all target NN parameters. Since the input dimension of $\mathcal{M}_{\boldsymbol{v}}$ is $N+1$, the size of $\boldsymbol{v}$ is constrained to a manageable scale of $O(\text{polylog}(m))$. Thus, the final parameter set $\boldsymbol{w}$ is generated through the combined process of PQC sampling and the mapping model $\mathcal{M}_{\boldsymbol{v}}$. By optimizing $\boldsymbol{\theta}$ and $\boldsymbol{v}$, we effectively control the loss function $\mathcal{L}$, which evaluates the performance of the target NN on a given task.

***Issue of*** $|\langle\phi_i|\psi(\boldsymbol{\theta})\rangle|^2$ ***in finite measurement shots setting.*** In an exact classical simulation of a quantum state $|\psi(\boldsymbol{\theta})\rangle$, the measurement probability of a given basis state, $|\langle\phi_i|\psi(\boldsymbol{\theta})\rangle|^2$, can be computed precisely. However, when implementing this process on an actual quantum computer, this probability must be estimated from a finite number of measurement shots, as discussed in Appendix G of (Liu et al., 2025).

When a basis state $|\phi_i\rangle$ is measured, the estimation error can be analyzed using Hoeffding's inequality:

$$P\left(|\hat{P}(|\phi_i\rangle) - \mathbb{E}[P(|\phi_i\rangle)]| \geq \epsilon\right) \leq 2\exp(-2\epsilon n'_{\text{shot}}), \tag{3}$$

where $\hat{P}(|\phi_i\rangle)$ represents the empirical estimate of $|\langle\phi_i|\psi(\boldsymbol{\theta})\rangle|^2$, and $\mathbb{E}[P(|\phi_i\rangle)]$ is the expected probability in the limit $n'_{\text{shot}} \to \infty$. This expression highlights that an inevitable estimation error ($\epsilon > 0$) exists when using a finite number of shots.

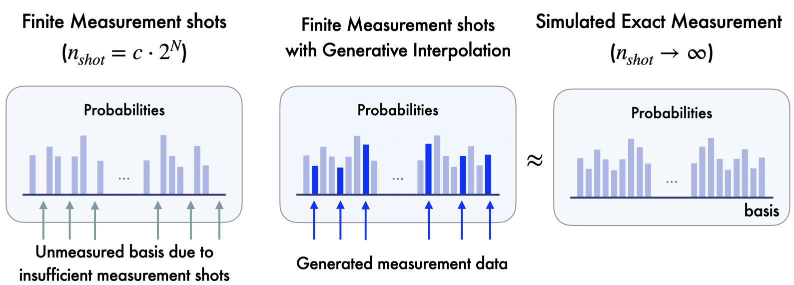


Figure 2: Graphical illustration of the finite measurement shot issue in QPA and how generative interpolation can be used to mitigate it.

Additionally, from the perspective of quantum fidelity tomography (Haah et al., 2017), the total number of measurement shots required to achieve an estimation accuracy of $\epsilon$ scales as:

$$n_{\text{shot}} = O\left(\frac{2^N}{\epsilon}\log\left(\frac{2^N}{\epsilon}\right)\right), \tag{4}$$

where $N$ denotes the number of qubits. This result illustrates the fundamental challenge of accurately estimating measurement probabilities when using finite measurement shots.

With an insufficient number of measurement shots, some basis states may remain unmeasured, leading to cases where $|\langle\phi_i|\psi(\boldsymbol{\theta})\rangle|^2 = 0$ in the input of Eq. 2. This significantly impacts the effectiveness of QPA, as missing probability values disrupt the overall learning process. The resulting degradation can be observed in Fig. 8 of (Liu et al., 2025).

Fig. 2 provides a graphical illustration of the finite measurement shot issue in QPA and demonstrates how the proposed method in the next section can effectively address it.

### 3.2 Frame generation in Hilbert space

To efficiently address the finite measurement issue discussed in the previous section, it is helpful to explore analogous scenarios in classical computing. One relevant example is the recent advancements in frame generation for video data (Liu et al., 2019; Zhao et al., 2019; Kwon & Park, 2019), which have been widely applied to enhance streaming video quality and optimize gaming performance, even on hardware with limited graphical capabilities. This technique enables, for instance, a video originally recorded at 30 frames per second (fps) to be displayed at 60 fps, with the additional frames being elegantly generated by a deep learning model. These methods typically generate frame data by using nearby frames as input, which is a reasonable approach since adjacent frames in a smoothly varying video sequence generally contain similar motion information.

In comparison, in the task of generate measurement probabilities, it is crucial to figure out what does it means to be "nearby" for the basis vectors in the Hilbert space. Since the basis representation of quantum system consists of qu"b"it is binary strings (i.e. $|\phi_i\rangle = |0100101\rangle$), we observe that it is possible to use Hamming distance as a measurement of "nearby" in the Hilbert space.

***Generative Interpolation of quantum measurement probabilities.*** During the QPA process, given a finite set of measurement data obtained from a limited number of shots for an $N$-qubit QNN, certain basis states $|\phi_i\rangle$ may remain unmeasured, meaning:

$$|\langle\phi_i|\psi(\boldsymbol{\theta})\rangle|^2 = P(|\phi_i\rangle) = 0. \tag{5}$$

To address this, we construct a neighboring basis set $\mathcal{S} = \{|\phi_k\rangle, ..., |\phi_j\rangle\}$ , which consists of all basis states differing from $|\phi_i\rangle$ by a Hamming distance $H_d = 1$. Since each element in $\mathcal{S}$ can be transformed into $|\phi_i\rangle$ by applying an $X$ gate to a single qubit, the number of elements in $\mathcal{S}$ is exactly $N$. Next, we construct the corresponding measurement probability vector:

$$\mathbf{P}(\mathcal{S}) = \big(P(|\phi_k\rangle), ..., P(|\phi_j\rangle)\big) \in [0,1]^N. \tag{6}$$

To estimate the missing probability $P(|\phi_i\rangle)$ , we employ a NN-based generative model $\mathcal{F}_{\boldsymbol{g}}$ in MLP architecture, parameterized by $\boldsymbol{g}$, which learns the relationship between $\mathbf{P}(\mathcal{S})$ and the desired probability:

$$\mathcal{F}_{\boldsymbol{g}}(\mathbf{P}(\mathcal{S})) = P(|\phi_i\rangle). \tag{7}$$

In essence, this process interpolates the missing probability $P(|\phi_i\rangle)$ by leveraging information from its nearest neighbors (in terms of $H_d$), with the complex dependency captured by the NN model $\mathcal{F}_{\boldsymbol{g}}$. Thus, we refer to this technique as Generative Interpolation (GI).

This interpolation is performed for all unmeasured basis states in the Hilbert space, filling in all missing probabilities using the estimated values from $\mathcal{F}_{\boldsymbol{g}}$. It is important to note that at this stage, the probability values remain unnormalized. The parameters $\boldsymbol{g}$ of the generative model are then jointly optimized along with the QPA parameters $\boldsymbol{\theta}$ and $\boldsymbol{v}$ throughout the QPA fine-tuning process.

Fig. 1 presents a graphical illustration of three key components. In panel (a), the general framework for frame generation in video data is depicted, providing a baseline for comparison. Panel (b) illustrates the proposed GI method, where we arrange the set of basis states (measurement probabilities) with a Hamming distance $H_d = 1$ in a ring within a hypothesis space. This layout demonstrates how the NN-based generative model $\mathcal{F}_{\boldsymbol{g}}$ estimates the unmeasured data. Finally, panel (c) shows the integrated scheme combining GI with QPA. As discussed above, the QPA process now involves three trainable blocks corresponding to the QNN parameters $\boldsymbol{\theta}$, the GI parameters $\boldsymbol{g}$, and the mapping model parameters $\boldsymbol{v}$.

Notably, this approach aligns with the ideal solution outlined earlier. By integrating the generative model directly into the QPA process, it eliminates the need for pre-training, allowing the model to dynamically learn and adapt as QPA progresses. Additionally, this method circumvents the sampling overhead encountered in traditional quantum state reconstruction techniques by directly predicting measurement probability values, rather than relying on iterative sampling-based reconstruction. As a result, this framework provides an efficient and scalable means of addressing the finite measurement shot issue in QPA.

## 4 Empirical Experiments

To examine the effectiveness of GI when integrated with QPA, we use the task of fine-tuning GPT-2 model on the WikiText-2 dataset. Where the performance metric is text generation perplexity, with lower perplexity, the model is more certain of the generated output, leads to better performance. The experiment is conducted using quantum circuit simulation via PyTorch and TorchQuantum (Wang et al., 2022). At this stage, noise effects on the quantum system are ignored, where the interested reader could refer to the Appendix G in the original QPA paper (Liu et al., 2025) for the effect of noise without involving the GI method.

In the experiment setup, we use QPA to generate the parameters of LoRA method, where we also put the result of original LoRA method for comparison, with the hyperparameters shown in Table below.

Table 1: Hyperparameter configurations of LoRA and QPA generated LoRA for fine-tuning GPT-2 and Gemma-2 with WikiText-2 dataset.

| **Hyperparameters** | **LoRA GPT-2** | **QPA LoRA GPT-2** |
|---|---|---|
| $\alpha$ | 2r | 2r |
| Dropout | 0.05 | 0.05 |
| Optimizer | AdamW | AdamW |
| LR | 1e-5 | 1e-5 |
| LR Scheduler | Linear | Linear |
| Batch size | 1 | 1 |
| Warmup Steps | 0 | 0 |
| Epochs | 3 | 3 |

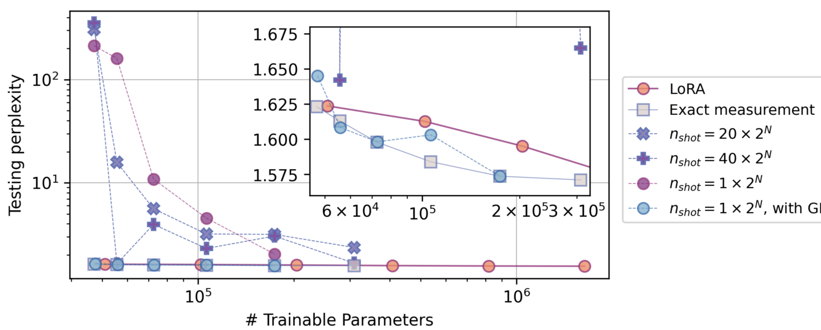


Figure 3: Testing perplexity of QPA on fine-tuning GPT-2 with different measurement shot settings.

Fig. 3 presents a comparative analysis of testing perplexity as a function of the number of trainable parameters across various measurement strategies in the QPA framework. The figure provides insights into the impact of finite measurement shots and the effectiveness of the proposed GI method in mitigating performance degradation due to limited quantum measurements.

The performance of different approaches is visualized through multiple curves. The LoRA baseline (red circles, solid line) serves as a reference, demonstrating stable performance across varying parameter scales. The exact measurement method (light blue squares, dashed line) represents an idealized scenario where full measurement data is available without shot limitations. However, in practical quantum implementations, measurement probabilities are estimated from a finite number of shots, leading to noticeable performance degradation. This is particularly evident in the cases of finite-shot measurement schemes, where testing perplexity increases significantly for smaller shot counts. Specifically, models with $n_{\text{shot}} = 20 \times 2^N$ (blue crosses), $n_{\text{shot}} = 40 \times 2^N$ (blue plus signs), and $n_{\text{shot}} = 1 \times 2^N$ (purple circles) exhibit progressively worse perplexity as the number of trainable parameters decreases. This behavior underscores the fundamental limitation imposed by shot noise in quantum measurements.

To address this issue, the proposed GI method (light blue circles, dashed line) is introduced, demonstrating a substantial improvement over direct finite-shot measurement estimates, shows that even when using only $n_{\text{shot}} = 1 \times 2^N$, which is $2.5\%$ compared to $40 \times 2^N$, incorporating GI significantly reduces testing perplexity, bringing it closer to the exact measurement baseline. This suggests that GI effectively enhances the quality of QPA while minimizing quantum resource requirements.

Additionally, the inset graph provides a magnified view of the region where testing perplexity differences are most pronounced, particularly for lower parameter counts. The zoomed-in view highlights that the GI-enhanced approach consistently outperforms direct finite-shot measurements, reinforcing its efficacy in scenarios with constrained quantum resources.

Notably, the qubit usage for the GI result data points, ranging from fewer to more trainable parameters, follows the sequence $N = 10, 9, 8, 7, 6$. From a real quantum hardware perspective, particularly for N = 10, reducing the number of measurement shots from $40 \times 2^N$ to $1 \times 2^N$ decreases the required shots from 40960 to just 1024, while also improving testing perplexity.

Using IBM Quantum's usage estimation tool [1], we find that for a 10-qubit circuit with $L = 8$ in the circuit ansatz, each quantum circuit execution in QPA accelerates from 135 seconds to just 2.3 seconds. Considering the fine-tuning process with 36718 steps per epoch over three epochs, this optimization reduces the total runtime from approximately 172 days to just 2.93 days.

## 5 Conclusion

This work extends the concept of video frame generation to the domain of quantum measurement probability estimation, enabling a more shots-efficient fine-tuning process for QPA. By drawing parallels between missing frames in video interpolation and missing quantum measurement data, GI is introduced as a method to estimate unmeasured quantum probabilities, significantly reducing the number of required quantum measurement shots. This approach ensures that even with only $2.5\%$ of the original measurement shot count, the fine-tuning of LLMs can be performed with improved performance and reduced quantum resource overhead.

Looking ahead, this generative interpolation framework can be further explored in broader variational quantum algorithms, where quantum-generated parameters play a critical role. If successfully extended, this method has the potential to drastically reduce the number of measurement shots required for quantum computing applications, thereby accelerating the overall quantum-classical computational pipeline. By leveraging deep learning techniques for quantum measurement reconstruction, this approach bridges classical generative modeling with quantum machine learning, offering a scalable and resource-efficient solution for future quantum-enhanced AI applications.

[1] IBM Quantum's computing usage estimation tool. Estimates may vary depending on the quantum computing provider and architecture. https://docs.quantum.ibm.com/guides/estimate-job-run-time

### ACKNOWLEDGMENTS

Wei-Hao Huang (Jij inc) acknowledges this work was performed for the Council for Science, Technology and Innovation (CSTI), Cross-ministerial Strategic Innovation Promotion Program (SIP), "Promoting the application of advanced quantum technology platforms to social issues" (Funding agency: QST). This work was also supported by the Engineering and Physical Sciences Research Council (EPSRC) under grant number EP/W032643/1.

# References


Amira Abbas, David Sutter, Christa Zoufal, Aurélien Lucchi, Alessio Figalli, and Stefan Woerner. The power of quantum neural networks. *Nature Computational Science*, 1(6):403–409, 2021.

Shahnawaz Ahmed, Carlos Sánchez Muñoz, Franco Nori, and Anton Frisk Kockum. Quantum state tomography with conditional generative adversarial networks. *Physical review letters*, 127(14): 140502, 2021.

Yuri Alexeev, Maximilian Amsler, Marco Antonio Barroca, Sanzio Bassini, Torey Battelle, Daan Camps, David Casanova, Young Jay Choi, Frederic T Chong, Charles Chung, et al. Quantum-centric supercomputing for materials science: A perspective on challenges and future directions. *Future Generation Computer Systems*, 160:666–710, 2024.

Matthias C Caro, Hsin-Yuan Huang, Marco Cerezo, Kunal Sharma, Andrew Sornborger, Lukasz Cincio, and Patrick J Coles. Generalization in quantum machine learning from few training data. *Nature communications*, 13(1):1–11, 2022.

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

Marco Cerezo, Andrew Arrasmith, Ryan Babbush, Simon C Benjamin, Suguru Endo, Keisuke Fujii, Jarrod R McClean, Kosuke Mitarai, Xiao Yuan, Lukasz Cincio, et al. Variational quantum algorithms. *Nature Reviews Physics*, 3(9):625–644, 2021.

Kuan-Cheng Chen, Samuel Yen-Chi Chen, Chen-Yu Liu, and Kin K Leung. Quantum-train-based distributed multi-agent reinforcement learning. *arXiv preprint arXiv:2412.08845*, 2024a.

Kuan-Cheng Chen, Tai-Yue Li, Yun-Yuan Wang, Simon See, Chun-Chieh Wang, Robert Wille, Nan-Yow Chen, An-Cheng Yang, and Chun-Yu Lin. Validating large-scale quantum machine learning: Efficient simulation of quantum support vector machines using tensor networks. *Machine Learning: Science and Technology*, 2024b.

Kuan-Cheng Chen, Xiaoren Li, Xiaotian Xu, Yun-Yuan Wang, and Chen-Yu Liu. Quantum-classical-quantum workflow in quantum-hpc middleware with gpu acceleration. In *2024 International Conference on Quantum Communications, Networking, and Computing (QCNC)*, pp. 304–311. IEEE, 2024c.

Samuel Yen-Chi Chen, Chao-Han Huck Yang, Jun Qi, Pin-Yu Chen, Xiaoli Ma, and Hsi-Sheng Goan. Variational quantum circuits for deep reinforcement learning. *IEEE access*, 8:141007–141024, 2020.

Samuel Yen-Chi Chen, Chih-Min Huang, Chia-Wei Hsing, and Ying-Jer Kao. An end-to-end trainable hybrid classical-quantum classifier. *Machine Learning: Science and Technology*, 2(4): 045021, 2021.

Samuel Yen-Chi Chen, Shinjae Yoo, and Yao-Lung L Fang. Quantum long short-term memory. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8622–8626. IEEE, 2022.

Matthias Christandl and Renato Renner. Reliable quantum state tomography. *Physical Review Letters*, 109(12):120403, 2012.

Marcus Cramer, Martin B Plenio, Steven T Flammia, Rolando Somma, David Gross, Stephen D Bartlett, Olivier Landon-Cardinal, David Poulin, and Yi-Kai Liu. Efficient quantum state tomography. *Nature communications*, 1(1):149, 2010.

Riccardo Di Sipio, Jia-Hong Huang, Samuel Yen-Chi Chen, Stefano Mangini, and Marcel Worring. The dawn of quantum natural language processing. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8612–8616. IEEE, 2022.

Yuxuan Du, Min-Hsiu Hsieh, Tongliang Liu, and Dacheng Tao. Expressive power of parametrized quantum circuits. *Physical Review Research*, 2(3):033125, 2020.

Edward Farhi and Hartmut Neven. Classification with quantum neural networks on near term processors. *arXiv preprint arXiv:1802.06002*, 2018.

Edward Farhi, Jeffrey Goldstone, and Sam Gutmann. A quantum approximate optimization algorithm. *arXiv preprint arXiv:1411.4028*, 2014.

Jeongwan Haah, Aram W. Harrow, Zhengfeng Ji, Xiaodi Wu, and Nengkun Yu. Sample-optimal tomography of quantum states. *IEEE Transactions on Information Theory*, pp. 1–1, 2017. ISSN 1557-9654. doi: 10.1109/tit.2017.2719044. URL http://dx.doi.org/10.1109/TIT.2017.2719044.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pp. 2790–2799. PMLR, 2019.

Ming-Chien Hsu, En-Jui Kuo, Wei-Hsuan Yu, Jian-Feng Cai, and Min-Hsiu Hsieh. Quantum state tomography via nonconvex riemannian gradient descent. *Physical Review Letters*, 132(24): 240804, 2024.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Sofiene Jerbi, Casper Gyurik, Simon Marshall, Hans Briegel, and Vedran Dunjko. Parametrized quantum policies for reinforcement learning. *Advances in Neural Information Processing Systems*, 34:28362–28375, 2021.

Yong-Hoon Kwon and Min-Gyu Park. Predicting future frames using retrospective cycle gan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1811–1820, 2019.

Shuyue Stella Li, Xiangyu Zhang, Shu Zhou, Hongchao Shu, Ruixing Liang, Hexin Liu, and Leibny Paola Garcia. Pqlm-multilingual decentralized portable quantum language model. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*, 2021.

Chu-Hsuan Abraham Lin, Chen-Yu Liu, and Kuan-Cheng Chen. Quantum-train long short-term memory: Application on flood prediction problem. In *2024 IEEE International Conference on Quantum Computing and Engineering (QCE)*, volume 2, pp. 268–273. IEEE, 2024.

Chen-Yu Liu and Samuel Yen-Chi Chen. Federated quantum-train with batched parameter generation. *arXiv preprint arXiv:2409.02763*, 2024.

Chen-Yu Liu, En-Jui Kuo, Chu-Hsuan Abraham Lin, Sean Chen, Jason Gemsun Young, Yeong-Jar Chang, and Min-Hsiu Hsieh. Training classical neural networks by quantum machine learning. *arXiv preprint arXiv:2402.16465*, 2024a.

Chen-Yu Liu, En-Jui Kuo, Chu-Hsuan Abraham Lin, Jason Gemsun Young, Yeong-Jar Chang, Min-Hsiu Hsieh, and Hsi-Sheng Goan. Quantum-train: Rethinking hybrid quantum-classical machine learning in the model compression perspective. *arXiv preprint arXiv:2405.11304*, 2024b.

Chen-Yu Liu, Chu-Hsuan Abraham Lin, Chao-Han Huck Yang, Kuan-Cheng Chen, and Min-Hsiu Hsieh. Qtrl: Toward practical quantum reinforcement learning via quantum-train. In *2024 IEEE International Conference on Quantum Computing and Engineering (QCE)*, volume 2, pp. 317–322. IEEE, 2024c.

Chen-Yu Liu, Chao-Han Huck Yang, Min-Hsiu Hsieh, and Hsi-Sheng Goan. A quantum circuit-based compression perspective for parameter-efficient learning. In *The Thirteenth International Conference on Learning Representations*, 2025.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. Dora: Weight-decomposed low-rank adaptation. *arXiv preprint arXiv:2402.09353*, 2024d.

Yu-Lun Liu, Yi-Tung Liao, Yen-Yu Lin, and Yung-Yu Chuang. Deep video frame interpolation using cyclic frame generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 8794–8802, 2019.

Andrea Mari, Thomas R Bromley, Josh Izaac, Maria Schuld, and Nathan Killoran. Transfer learning in hybrid classical-quantum neural networks. *Quantum*, 4:340, 2020.

Kosuke Mitarai, Makoto Negoro, Masahiro Kitagawa, and Keisuke Fujii. Quantum circuit learning. *Physical Review A*, 98(3):032309, 2018.

Michael A Nielsen and Isaac L Chuang. *Quantum computation and quantum information*. Cambridge university press, 2010.

Jun Qi, C.-H. Huck Yang, and Pin-Yu Chen. Qtn-vqc: An end-to-end learning framework for quantum neural networks. *Physica Scripta*, 99, 12 2023. doi: 10.1088/1402-4896/ad14d6.

Sukin Sim, Peter D Johnson, and Alán Aspuru-Guzik. Expressibility and entangling capability of parameterized quantum circuits for hybrid quantum-classical algorithms. *Advanced Quantum Technologies*, 2(12):1900070, 2019.

Andrea Skolik, Sofiene Jerbi, and Vedran Dunjko. Quantum agents in the gym: a variational quantum algorithm for deep q-learning. *Quantum*, 6:720, 2022.

Jonas Stein, Ivo Christ, Nicolas Kraus, Maximilian Balthasar Mansky, Robert Müller, and Claudia Linnhoff-Popien. Applying qnlp to sentiment analysis in finance. In *2023 IEEE International Conference on Quantum Computing and Engineering (QCE)*, volume 2, pp. 20–25. IEEE, 2023.

Jules Tilly, Hongxiang Chen, Shuxiang Cao, Dario Picozzi, Kanav Setia, Ying Li, Edward Grant, Leonard Wossnig, Ivan Rungger, George H Booth, et al. The variational quantum eigensolver: a review of methods and best practices. *Physics Reports*, 986:1–128, 2022.

Hanrui Wang, Yongshan Ding, Jiaqi Gu, Zirui Li, Yujun Lin, David Z Pan, Frederic T Chong, and Song Han. Quantumnas: Noise-adaptive search for robust quantum circuits. In *The 28th IEEE International Symposium on High-Performance Computer Architecture (HPCA-28)*, 2022.

Chao-Han Huck Yang, Yun-Yun Tsai, and Pin-Yu Chen. Voice2series: Reprogramming acoustic models for time series classification. In *International conference on machine learning*, pp. 11808–11819. PMLR, 2021.

Chao-Han Huck Yang, Jun Qi, Samuel Yen-Chi Chen, Yu Tsao, and Pin-Yu Chen. When bert meets quantum temporal convolution learning for text classification in heterogeneous computing. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8602–8606. IEEE, 2022.

Won Joon Yun, Yunseok Kwak, Jae Pyoung Kim, Hyunhee Cho, Soyi Jung, Jihong Park, and Joongheon Kim. Quantum multi-agent reinforcement learning via variational quantum circuit design. In *2022 IEEE 42nd International Conference on Distributed Computing Systems (ICDCS)*, pp. 1332–1335. IEEE, 2022.

Lei Zhao, Shiqi Wang, Xinfeng Zhang, Shanshe Wang, Siwei Ma, and Wen Gao. Enhanced motion-compensated video coding with deep virtual reference frame generation. *IEEE Transactions on Image Processing*, 28(10):4832–4844, 2019.

Yan Zhu, Ya-Dong Wu, Ge Bai, Dong-Sheng Wang, Yuexuan Wang, and Giulio Chiribella. Flexible learning of quantum states with generative query neural networks. *Nature communications*, 13 (1):6222, 2022.

Christa Zoufal, Aurélien Lucchi, and Stefan Woerner. Quantum generative adversarial networks for learning and loading random distributions. *npj Quantum Information*, 5(1):103, 2019.