Vision Language Model for Interpreting Wigner Distributions in Quantum Optics

Anas M. Nabih, Brian E. Arfeto, Uman Khalid, and Hyundong Shin
Department of Electronics and Information Convergence Engineering, Kyung Hee University, Korea
Email: hshin@khu.ac.kr

***Abstract*—This paper presents a novel integration of Quantum Optics and Vision-Language Models (VLMs) to address challenges in Quantum State Tomography (QST). Quantum systems, inherently characterized by superposition and entanglement, pose significant difficulties in direct state measurement and reconstruction. To overcome these challenges, we propose a Quantum Optical Vision Language Model (QOVLM), leveraging the Qwen2.5-VL architecture, fine-tuned on quantum optical state datasets. The model employs visual analysis of Wigner functions and quantum images to classify and infer photon states, qubit counts, and coherence properties via chain-of-thought prompting. Evaluation demonstrates the model's capability to identify quantum states (e.g., Fock, cat, coherent, thermal) with visual reasoning, offering a new paradigm in quantum information processing using VLM.**

***Index Terms*—vision language model, quantum state tomography, qubit, quantum states**

I. Introduction

Large language models (LLMs) and vision-language models (VLMs) have recently achieved impressive breakthroughs, powered by transformer-based architectures and large-scale datasets. LLMs demonstrate strong performance in understanding, generating, and reasoning over natural language, while VLMs combine textual and visual modalities to enable tasks like image captioning and recognition. Researchers have begun leveraging the reasoning power of VLMs for quantum many-body problems and derivations; for instance, Pan et al. showed that GPT-4 could reproduce Hartree-Fock equations with high accuracy when guided by expert prompts [1]. In parallel, machine learning is being applied to Quantum State Tomography (QST), a traditionally resource-intensive process. Ahmed et al. demonstrated that convolutional neural networks can reconstruct optical quantum states even under noise and data scarcity, and subsequent work extended this with generative models embedding physical constraints for better fidelity [2].

Beyond QST, the integration of LLMs into quantum algorithm development and simulation is accelerating. Zhou et al. trained a transformer model to simulate 2–3 qubit circuits with minimal error, drastically reducing computational cost [3]. Specialized LLMs like GroverGPT further exemplify this trend—Wang et al.'s model achieved over 95% accuracy for quantum search problems with up to 20 qubits, outperforming GPT-4 [4]. Nakaji et al. proposed a GPT-based generative quantum eigensolver capable of autonomously designing quantum circuits for ground-state preparation [5]. Beyond these, machine learning has also been applied to other quantum tasks such as generative AI integration [6], beam sensing [7], and entanglement detection [8]. These works highlight how large-scale deep learning models can emulate quantum behavior and aid discovery. This paper builds upon this momentum by applying LLMs and VLMs—specifically, a fine-tuned Qwen2.5-VL model—to QST tasks using a novel Quantum Optical Dataset. The dataset includes visual Wigner function representations of cat, fock, coherent, and thermal states, annotated with parameters like photon count and alpha. Through chain-of-thought prompting and multimodal reasoning, the model classifies states, estimates quantum parameters, and learns qubit numbers.

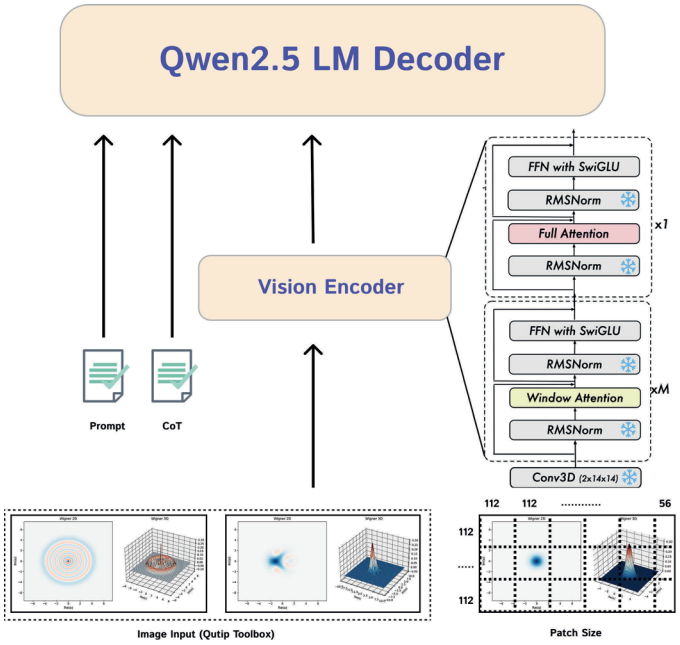

Fig. 1. Architecture of Quantum Vision Language Model

II. Methods

A. Dataset Construction

We generated our dataset using the open-source quantum computing library, `QuTiP`. For each quantum state, we varied parameters such as the number of qubits (ranging from 1 to 30), displacement amplitude (α), photon number, density, and line space resolution. Each configuration produced an

TABLE I
EVALUATION OF QWEN2.5VL BASE VS FINE-TUNED MODEL

Metric	Qwen2.5VL Base	Qwen2.5VL FT
Similarity Scores		
BERT(F1) Mean Score	0.9586	0.9680
BERT Mean Precision	0.9558	0.9696
BERT Mean Recall	0.9615	0.9664
BLEU1 Score	0.0088	0.0327
Error Metrics		
CER Score	1.2266	0.6106
MER Score	0.8771	0.7338
WER Score	1.5692	0.7502
Task Accuracy		
State classification	34.71%	94.01%
Parameter (α, density, photon)	16.67%	89.41%
Number of qubits	0.18%	99.26%
Linear space calculation	34.62%	65.75%
All correct	0.00%	54.33%

image with the corresponding quantum state's 2-dimensional and 3-dimensional Wigner function. These images were saved alongside metadata indicating the ground-truth state type (e.g., coherent, cat, Fock, thermal, random), qubit count, and specific parameters used in the simulation. Our objective is to perform reverse inference: predicting these original parameters solely from the visual representation using a vision-language model.

The Wigner function $W(q,p)$ is a quasiprobability distribution that provides a full description of the quantum state in phase space. For a quantum system described by a density operator $\hat{\rho}$, the Wigner function is defined as:

$$W(q,p) = \frac{1}{\pi\hbar}\int_{-\infty}^{\infty} dy\, \langle q-y|\hat{\rho}|q+y\rangle e^{2ipy/\hbar}, \qquad (1)$$

which for a pure state ψ reduces to:

$$W(q,p) = \frac{1}{\pi\hbar}\int_{-\infty}^{\infty} dy\, \psi^*(q+y)\,\psi(q-y)\,e^{2ipy/\hbar}. \qquad (2)$$

This function maps the quantum state into a two-dimensional phase space using the position q and momentum p coordinates, and can take on negative values, reflecting the non-classical features of quantum systems.

B. Model Architecture

We adopt Qwen2.5-VL as the backbone of our system. In particular, we leverage Qwen2.5-VL's vision encoder, which incorporates 2D rotary positional embeddings to capture spatial dependencies in visual data. This feature is critical for understanding Wigner function structures and spatial coherence patterns that distinguish quantum states. Given our dataset includes both 2D and pseudo-3D visualizations, we aim to evaluate the model's ability to generalize over diverse image encodings of quantum information. The whole architecture can be seen in Figure 1.

C. Fine-tuning Strategy

To adapt the pre-trained Qwen2.5-VL to the quantum domain, we fine-tune it on our custom dataset using parameter-efficient fine-tuning (PEFT) techniques. Specifically, we use LoRA parameter fine tuning. Training is conducted with a conservative learning rate of 0.1 to ensure stability, albeit requiring more epochs for convergence. We train the model over four epochs, monitoring accuracy and generalization performance across different quantum state types. PEFT layers are applied to the vision encoder and projection layers to minimize the number of trainable parameters while maximizing domain-specific adaptability.

III. RESULTS

Incorporating chain-of-thought prompting into the Qwen2.5-VL fine-tuning pipeline yields a pronounced uplift in both semantic reconstruction and discrete state classification. As detailed in Table 1, the fine-tuned model not only boosts its BERT-based F1 score and quadruples its BLEU-1 score, but it also slashes transcription error rates. Most critically for quantum state reconstruction, classification accuracy soars from 34.7 % to 94.0 %, parameter estimation (α, photon density) jumps from 16.7 % to 89.4 % qubit-count prediction from 0.18 % to 99.3 %, and linear-space calculation correctness from 34.6 % to 65.8 % resulting in fully correct end-to-end outputs in 54.3 % of cases versus 0 % previously. These gains demonstrate that explicitly guiding the model through intermediate reasoning steps dramatically improves its ability to parse complex visual-quantum inputs and reconstruct underlying physical parameters.

IV. CONCLUSION

This study demonstrates the potential of Vision-Language Models (VLMs) in analyzing Wigner function representations for quantum state classification and parameter reconstruction. Leveraging the Qwen2.5-VL vision encoder, the model successfully identifies key quantum state types and estimates associated parameters directly from visual input. While the results indicate promising capabilities, the current parameter classification accuracy remains below the threshold required for high-confidence applications. To address this, future work should explore hybrid architectures that combine VLM with specialized deep neural networks, enhancing both interpretability and precision. Additionally, scaling the vision-language model to larger parameter sizes may improve performance across diverse quantum configurations.

ACKNOWLEDGMENT

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) under RS-2025-00556064, by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2025-2021-0-02046) supervised by the IITP (Institute for Information Communications Technology Planning Evaluation), and by a grant from Kyung Hee University in 2023 (KHU-20233663).

REFERENCES

[1] H. Pan, N. Mudur, W. Taranto, M. Tikhanovskaya, S. Venugopalan, Y. Bahri, M. P. Brenner, and E.-A. Kim, "Quantum many-body physics calculations with large language models," *Communications Physics*, vol. 8, no. 1, p. 49, 2025.

[2] S. Ahmed, C. Sánchez Muñoz, F. Nori, and A. F. Kockum, "Classification and reconstruction of optical quantum states with deep neural networks," *Phys. Rev. Res.*, vol. 3, p. 033278, Sep 2021. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevResearch.3.033278

[3] S. Zhou, R. Chen, Z. An, C. Zhang, and S.-Y. Hou, "Application of large language models to quantum state simulation," *Science China Physics, Mechanics & Astronomy*, vol. 68, no. 4, p. 240313, 2025.

[4] H. Wang, P. Li, M. Chen, J. Cheng, J. Liu, and T. Chen, "Grovergpt: A large language model with 8 billion parameters for quantum searching," *arXiv preprint arXiv:2501.00135*, 2024.

[5] K. Nakaji, L. B. Kristensen, J. A. Campos-Gonzalez-Angulo, M. G. Vakili, H. Huang, M. Bagherimehrab, C. Gorgulla, F. Wong, A. McCaskey, J.-S. Kim *et al.*, "The generative quantum eigensolver (gqe) and its application for ground state search," *arXiv preprint arXiv:2401.09253*, 2024.

[6] B. E. Arfeto, S. Tariq, U. Khalid, T. Q. Duong, and H. Shin, "Gensc-6g: A prototype testbed for integrated generative ai, quantum, and semantic communication," 2025. [Online]. Available: https://arxiv.org/abs/2501.09918

[7] S. Tariq, B. E. Arfeto, U. Khalid, S. Kim, T. Q. Duong, and H. Shin, "Deep quantum-transformer networks for multimodal beam prediction in ISAC systems," vol. 11, no. 18, pp. 29 387–29 401, Sep. 2024.

[8] N. Asif, U. Khalid, A. Khan, T. Q. Duong, and H. Shin, "Entanglement detection with artificial neural networks," vol. 13, no. 1, p. 1562, Jan. 2023.