

# MEDQUA: A NISQ-AWARE QUANTUM ADAPTER FOR MEDICAL VISION LANGUAGE MODELS

Yiwei Li<sup>1,\*</sup>, Yi Pan<sup>1,\*</sup>, Junhao Chen<sup>1</sup>, Yifan Zhou<sup>1</sup>, Hanqi Jiang<sup>1</sup>, Huaqin Zhao<sup>1</sup>, Yanjun Lyu<sup>2</sup>,  
Zhengliang Liu<sup>1</sup>, Lin Zhao<sup>4</sup>, Dajiang Zhu<sup>2</sup>, Xiang Li<sup>3</sup>, Tianming Liu<sup>1,†</sup>

<sup>1</sup>School of Computing, University of Georgia, Athens, GA, USA

<sup>2</sup>Department of Computer Science and Engineering, University of Texas at Arlington, TX, USA

<sup>3</sup>Department of Radiology, Massachusetts General Hospital and Harvard Medical School, MA, USA

<sup>4</sup>Department of Biomedical Engineering, New Jersey Institute of Technology, NJ, USA

## ABSTRACT

Vision–language models (VLMs) are promising for medical image classification but still face generalization limits from visual encoders and cross-site distribution shift. Fully quantum VLMs could offer richer representations, yet NISQ hardware makes end-to-end quantum training impractical. We introduce MEDQUA, a NISQ-aware quantum adapter that attaches to a pretrained VLM decoder. An entropy-driven router sparsely selects tokens for a shallow variational quantum bottleneck, while a lightweight LoRA-based classical path processes all tokens to ensure stability and low cost. On MIMIC-CXR and ChestMNIST, MEDQUA consistently improves accuracy and AUROC over classical VLM baselines (including SFT) with modest overhead, showing that adaptively integrated quantum modules already yield practical gains. As coherence, error rates, and compilation advance, the same adapter can scale to deeper circuits and larger qubit counts without redesigning the classical backbone, providing a pragmatic route to quantum-enhanced medical VLMs.

## 1. INTRODUCTION

VLMs have become central to modern medical-image understanding, supporting cross-modal retrieval, report generation, grounding, and decision support [1]. Their appeal lies in broad generalization from large-scale pretraining to diverse downstream tasks [2]. Yet persistent challenges remain in clinical practice: distribution shift across institutions and devices, scarce and noisy annotations, long-tail pathologies, and the need for calibrated, data-efficient adaptation [3]. Moreover, classical Transformers incur high compute and memory costs during fine-tuning and deployment, which limits scalability in resource-constrained clinical settings [4]. In contrast, quantum computing offers richer state representations via superposition and entanglement that could,

in principle, model complex, high-dimensional medical semantics more compactly [5]. However, today’s NISQ hardware—short coherence, nontrivial error rates, and shallow circuit budgets [6]—makes end-to-end, fully quantum VLMs impractical to train from scratch.

Motivated by this gap, we pursue a pragmatic route: *incrementally* combining strong classical VLM/LLM priors with *quantum adapters*. Rather than replacing proven Transformer backbones [7], we design lightweight quantum modules that attach to key attention pathways and are invoked only where they add value. This hybridization aims to retain the broad knowledge and inductive biases of pretrained LLMs/VLMs [8, 9] while selectively exploiting quantum feature mixing or similarity retrieval when complexity is high and quantum depth budgets permit. Crucially, the adapter design avoids deep quantum circuits, respects NISQ limits, and remains drop-in compatible with standard inference stacks. Existing fully quantum NLP pipelines [10, 11] and variational approaches [12] struggle with scalability and barren plateaus [13], while prior hybrids often partition computation statically [14, 15]. Our adapter-centric strategy instead *bridges* paradigms by learning when and how to offload subroutines to quantum hardware.

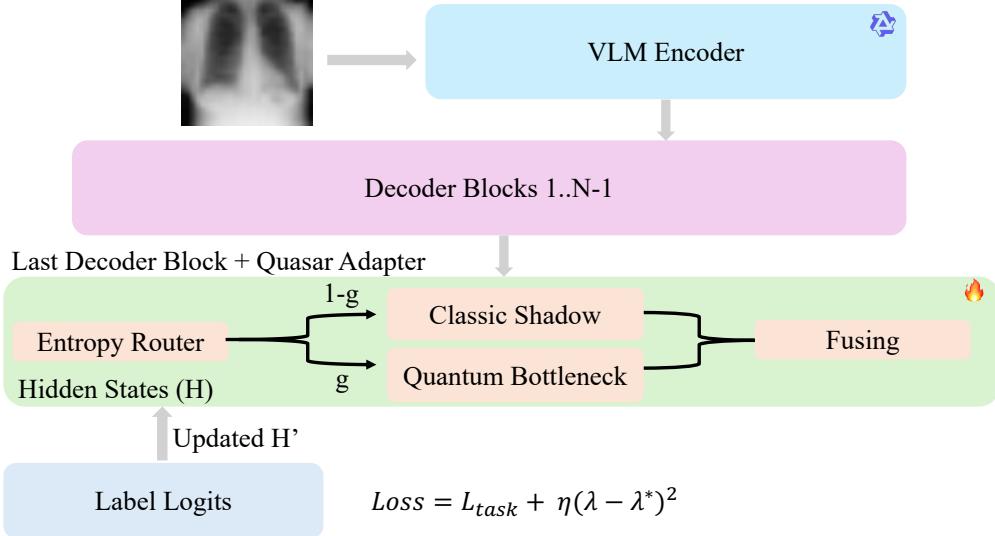
### Contributions.

- We propose a modular quantum adapter that interfaces with pretrained Transformer attention, leveraging classical VLM/LLM priors rather than replacing them.
- We introduce an activation policy that selectively routes high-complexity subproblems to shallow quantum circuits, respecting NISQ constraints [6] while maintaining classical fallback.
- On medical VLM benchmarks, our hybrid system achieves performance *comparable* to strong classical VLM baselines, with practical advantages in parameter efficiency and robustness in low-data settings.

This adapter-based bridge offers a realistic path from today’s clinical VLMs to future quantum-enhanced models without requiring full quantum training or deep circuits.

\*Equal Contribution.

†Corresponding Author.



**Fig. 1.** Overall Architecture of the proposed framework.

## 2. PRELIMINARIES

An  $n$ -qubit pure state lies in  $\mathcal{H} = (C^2)^{\otimes n}$  as  $\psi = \sum_{i=0}^{2^n-1} \alpha_i i$  with  $\sum_i |\alpha_i|^2 = 1$ . Quantum circuits realize unitaries  $U \in SU(2^n)$  compiled from single-qubit rotations and two-qubit entanglers (e.g., CNOT) [16]. On NISQ devices we employ variational quantum circuits (VQCs)

$$U(\theta) = \prod_{l=1}^L \left( W_l \prod_{i=1}^n R_i^{(l)}(\theta_i^{(l)}) \right), \quad (1)$$

where  $R_i^{(l)} \in \{R_x, R_y, R_z\}$  and  $W_l$  are entangling layers. Embeddings  $\mathbf{x} \in R^d$  are angle-encoded to  $\psi(\mathbf{x}) = \bigotimes_{i=1}^n R_y(\phi_i)0$  with  $\phi_i = \text{clip}(\alpha_i, -\pi, \pi)$ , enabling gradient training while entanglement captures cross-feature correlations.

To drive adaptive depth/routing, we compute a lightweight complexity signal

$$\mathbf{s}(\mathbf{x}) = [\mu(\mathbf{x}), \sigma^2(\mathbf{x}), H(\mathbf{x}), \kappa(\mathbf{x})], \quad (2)$$

where  $H(\mathbf{x}) = -\sum_j p_j \log p_j$  on  $p_j = \frac{|x_j|}{\sum_k |x_k|}$ . Higher entropy and heavy tails trigger deeper (yet shallow) subroutines and larger routing probabilities.

Finally, expressive random circuits face barren plateaus: the gradient variance decays exponentially,  $\text{Var}[\partial \mathcal{L}/\partial \theta_k] = \mathcal{O}(c^{-n})$ ,  $c > 1$  [13]. We therefore use shallow, structured ansätze and adapt depth/routing to remain trainable on NISQ hardware.

## 3. METHOD

### 3.1. Overview

We introduce a plug-and-play *Quasar Adapter* that inserts a lightweight quantum bottleneck into pretrained Transformer decoders while retaining a parallel classical (LoRA-like) path. A learned *Entropy Router* selects a sparse subset of tokens for quantum processing; all tokens pass through the classical path. The two paths are fused and added residually to the backbone hidden states. The design is NISQ-aware: (i) sparse quantum execution, (ii) shallow, adaptive quantum depth, and (iii) differentiable routing with straight-through (STE) or Gumbel relaxation. (Figure 1)

### 3.2. Entropy-based Router

Given hidden states  $H \in R^{B \times T \times D}$ , the router produces token-wise gates. A linear scorer and an entropy correction yield logits:

$$s = \text{Linear}(H) \in R^{B \times T}, \quad (3)$$

$$p = \frac{H}{\|H\|_2 + \varepsilon}, \quad \text{Ent}(H) = -\sum_{d=1}^D p_{:, :, d} \log(|p_{:, :, d}| + \varepsilon), \quad (4)$$

$$\ell = \frac{s + \text{Ent}(H)}{\tau_r}, \quad g = \sigma(\ell) \in (0, 1)^{B \times T}, \quad (5)$$

where  $\tau_r$  is a learnable temperature and  $\varepsilon > 0$ . We align any incoming attention mask to  $[B, T]$  (right-aligned, left-padded if needed) and zero out invalid positions:  $g \leftarrow g \odot \text{active}$ .

To enforce a target quantum activation ratio  $\lambda^* \in [0, 1]$ , we perform masked Top- $k$  selection over valid tokens, with

Method	MIMIC-CXR Dataset		ChestMNIST Dataset	
	Accuracy (%)	AUROC	Accuracy (%)	AUROC
<i>Classic VLM</i>				
Qwen2.5VL-3B	67.44	0.74	62.31	0.68
Qwen2.5VL-3B-SFT	89.17	0.88	82.24	0.78
<i>VLM + Quantum Adapter</i>				
MEDQUA	<b>91.88</b>	<b>0.90</b>	<b>82.30</b>	<b>0.81</b>

**Table 1.** Accuracy and AUROC on MIMIC-CXR and ChestMNIST classification.

$k = \lfloor \lambda^* \cdot N_{\text{active}} \rfloor$ . The minimum selected score defines a *dynamic threshold*  $\theta_t$ . We maintain an EMA of the threshold,

$$\theta_{t+1} \leftarrow \alpha \theta_t + (1 - \alpha) \text{clip}(\theta_{\text{dyn}}, \theta_{\min}, \theta_{\max}), \quad (6)$$

to stabilize the target sparsity over time. The binary route mask is  $m = 1[g \geq \theta_{t+1}]$  on active tokens. For differentiability, we use either (i) STE:  $g_{\text{hard}} = \text{stopgrad}(m - g) + g$ , or (ii) Gumbel–Sigmoid  $\tilde{g} = \sigma((\ell + \text{Gumbel})/\tau)$  with an exponential annealing schedule for  $\tau$ .

### 3.3. Quasar Adapter: Sparse Quantum & Classical Shadow

For every  $K$ -th decoder block, we attach a dual-path adapter: **Classical shadow (LoRA-like)**. A low-rank residual path

$$H_{\text{cls}} = \text{Drop}(UDH) \cdot \alpha/r, \quad (7)$$

with  $D \in R^{r \times D}$ ,  $U \in R^{D \times r}$  (no biases), provides a cheap universal improvement and a strong fallback when tokens are not routed to quantum.

**Sparse quantum path.** Tokens with  $m_{b,t} = 1$  are gathered and mapped to an *adaptive quantum bottleneck* (internal fp32) producing  $Q \in R^{N_{\text{sel}} \times n_q}$  with depth  $\leq L_{\max}$  decided by the bottleneck. A linear head projects back:

$$\tilde{H}_{\text{q}} = \text{Scatter}\left(\text{Proj}(Q), \{(b, t) : m_{b,t} = 1\}\right) \in R^{B \times T \times D}, \quad (8)$$

and zeros elsewhere. When no token is selected,  $\tilde{H}_{\text{q}} = \mathbf{0}$ .

**Fusion and residual.** We fuse paths with the (soft) gate and apply LayerNorm and dropout:

$$\hat{g} = (1-\epsilon)g + \epsilon, \quad F = \text{LN}\left(\text{Drop}(\hat{g} \odot \tilde{H}_{\text{q}} + (1 - \hat{g}) \odot H_{\text{cls}})\right) \quad (9)$$

and output  $H' = H + \gamma F$  (scale  $\gamma$ ). This keeps gradients through both routes and avoids degenerate all-0/1 gating.

### 3.4. Objective with Quantum Proportion Regularization

Let  $\mathcal{L}_{\text{task}}$  be the VLM loss. We regularize the realized quantum usage toward the target:

$$\hat{\lambda} = \frac{\sum(m \odot \text{active})}{\sum \text{active}}, \quad \mathcal{L}_{\lambda} = (\hat{\lambda} - \lambda^*)^2. \quad (10)$$

The final loss is

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \eta \mathcal{L}_{\lambda}, \quad (11)$$

with  $\eta$  controlling the trade-off. This encourages stable sparsity (and thus bounded quantum cost) while letting the model learn where quantum computation is most beneficial. Optionally, one may add a gating entropy term  $\xi H(g)$  for exploration in early training.

## 3.5. Integration

We discover decoder layers heuristically and attach adapters to every  $K$ -th block. The quantum module runs on the model’s device but keeps its internal arithmetic in fp32 for stability; all projections/norms match the backbone dtype. The router exposes per-layer diagnostics (*activation ratio*, *threshold*, *mean quantum depth*) for profiling and ablations.

## 4. EXPERIMENTS

### 4.1. Datasets and Task

**MIMIC-CXR.** We use the publicly available MIMIC-CXR-JPG collection of chest radiographs with paired free-text reports [17]. Following common practice for label classification, we derive multi-label targets from report-derived CheXpert-style observations (e.g., Atelectasis, Cardiomegaly). We use the official patient-level train/val/test split; uncertainty labels are mapped to “ignore” during loss computation.

**ChestMNIST (MedMNIST v2).** We use only the ChestMNIST subset from MedMNIST v2 [18], which consists of grayscale chest X-ray thumbnails derived from NIH ChestX-ray14 and annotated with 14 disease labels. We follow the official train/validation/test split and treat the task as multi-label classification, reporting accuracy and AUROC [19]. Images are upsampled to the model input size during preprocessing; no metadata or text is used.

### 4.2. Backbone and Our Modules

We adopt **Qwen2.5-VL** as the vision–language backbone (frozen or lightly tuned), and insert our *Quasar Adapter* every  $K$  decoder blocks (default  $K=6$ ). The classical path is

<b>Configuration</b>	<b>MIMIC-CXR Dataset</b>		<b>ChestMNIST Dataset</b>	
	<b>Acc. (%)</b>	<b>AUROC</b>	<b>Acc. (%)</b>	<b>AUROC</b>
<i>Proposed Method</i>				
MEDQUA	<b>91.88</b>	<b>0.90</b>	<b>82.30</b>	<b>0.81</b>
<i>Ablation Study</i>				
Fixed Quantum Ratio (20%)	79.59	0.79	71.24	0.68
Without Quantum Proportion Regularization	83.71	0.82	73.18	0.72

**Table 2.** Ablation study on the influence of each module.

a LoRA-style shadow; the quantum path is an adaptive variational bottleneck with at most  $L_{\max}$  layers and  $n_q$  qubits. The *Entropy Router* selects a sparse set of tokens for quantum processing with target activation  $\lambda^*$  and EMA-threshold control (Sec. 3). For classification, we pool visual tokens from Qwen2.5-VL [20] and attach a linear head; losses are: binary cross-entropy (multi-label) for MIMIC-CXR and cross-entropy (single/multi-class) for MedMNIST.

### 4.3. Hardware

All experiments are conducted on a cluster with  $10 \times$  NVIDIA RTX A5000 (24 GB) GPUs. Wall-clock profiling includes forward-time per sample and effective quantum token ratio  $\hat{\lambda}$  logged from the router.

### 4.4. Classification Results

Table 1 summarizes accuracy and mean AUROC on MIMIC-CXR and ChestMNIST. On MIMIC-CXR, the vanilla Qwen2.5 VL-3B (zero-shot) attains 67.44% accuracy with 0.74 AUROC. Supervised fine-tuning (SFT) lifts performance to 89.17% / 0.88. Our MEDQUA (VLM + Quantum Adapter) further improves to 91.88% accuracy and 0.90 AUROC, a gain of +2.71 percentage points accuracy and +0.02 AUROC over SFT, and +24.44/+0.16 over zero-shot.

On ChestMNIST, the zero-shot VLM yields 62.31% / 0.68, while SFT reaches 82.24% / 0.78. MEDQUA achieves 82.30% / 0.81, essentially matching SFT in accuracy (+0.06) but improving AUROC by +0.03. The larger AUROC gain on ChestMNIST, despite similar accuracy, indicates better ranking/calibration under multi-label imbalance, whereas MIMIC-CXR shows consistent improvements in both top-1 accuracy and discriminative power.

Overall, integrating the quantum adapter on the last decoder block yields consistent benefits over classical baselines, with the largest improvements on the clinically realistic MIMIC-CXR benchmark while operating under NISQ-aware sparse routing.

### 4.5. Ablation Study

The ablations in Table 2 show two consistent trends. First, replacing the entropy-driven router with a fixed quantum ratio degrades both accuracy and AUROC, indicating that uniform allocation of quantum compute is suboptimal for medical VLM classification. Second, removing the quantum-proportion regularization also lowers performance and stability, reflecting the need to control the realized usage around a target level. These results support the central design choice: the quantum share should be dynamically adjusted at the token level, and the target proportion must participate in training via an explicit loss term to guide the router.

## 5. DISCUSSION AND CONCLUSION

We introduced a NISQ-aware, adapter-centric approach that integrates a shallow quantum bottleneck into pretrained VLMs via entropy-driven routing and a target usage proportion. On MIMIC-CXR and ChestMNIST, the method delivers consistent gains over classical baselines with modest overhead, indicating that adaptively integrated quantum modules can already provide practical benefits without redesigning the backbone.

Several limitations remain. First, most experiments rely on simulation; validating end-to-end on real hardware (with calibration, error mitigation, batching, and I/O constraints) is an immediate priority. Second, NISQ budgets on depth and qubits cap expressivity; even with sparse routing, token gathering and device transfers may dominate latency. Third, scaling from mid-size VLMs to billion-parameter backbones demands more efficient routing/communication and potentially asynchronous quantum queues. Finally, broader medical VLM tasks (e.g., grounding, retrieval, report generation) may require task-specific circuits and routing policies, and deeper circuits may reintroduce trainability risks.

As coherence times improve and error rates fall, the same adapter interface can exploit deeper circuits and more qubits while keeping classical priors intact. In short, dynamically allocating quantum compute and coupling it to a proportion regularizer offers a practical path from today’s clinical VLMs to scalable, quantum-enhanced systems.

## 6. REFERENCES

- [1] Yiwei Li, Yikang Liu, Jiaqi Guo, Lin Zhao, Zheyuan Zhang, Xiao Chen, Boris Mailhe, Ankush Mukherjee, Terrence Chen, and Shanhui Sun, “Rau: Reference-based anatomical understanding with vision language models,” *arXiv preprint arXiv:2509.22404*, 2025.
- [2] Tianyang Zhong, Wei Zhao, Yutong Zhang, Yi Pan, Peixin Dong, Zuowei Jiang, Xiaoyan Kui, Youlan Shang, Li Yang, Yaonai Wei, et al., “Chatradio-valuer: A chat large language model for generalizable radiology report generation based on multi-institution and multi-system data,” *arXiv preprint arXiv:2310.05242*, 2023.
- [3] Zhengliang Liu, Yiwei Li, Peng Shu, Aoxiao Zhong, Hanqi Jiang, Yi Pan, Longtao Yang, Chao Ju, Zihao Wu, Chong Ma, et al., “Radiology-gpt: a large language model for radiology,” *Meta-Radiology*, p. 100153, 2025.
- [4] Yiwei Li, Sekeun Kim, Zihao Wu, Hanqi Jiang, Yi Pan, Pengfei Jin, Sifan Song, Yucheng Shi, Tianming Liu, Quanzheng Li, et al., “Echopulse: Ecg controlled echocardio-grams video generation,” *arXiv preprint arXiv:2410.03143*, 2024.
- [5] Sitan Chen, Jordan Cotler, Hsin-Yuan Huang, and Jerry Li, “The complexity of nisq,” *Nature Communications*, vol. 14, no. 1, pp. 6001, 2023.
- [6] John Preskill, “Quantum computing in the NISQ era and beyond,” *Quantum*, vol. 2, pp. 79, 2018.
- [7] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin, “Attention is all you need,” in *Advances in neural information processing systems*, 2017, vol. 30.
- [8] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al., “Language models are few-shot learners,” *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.
- [9] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambo, Faisal Azhar, et al., “Llama: Open and efficient foundation language models,” *arXiv preprint arXiv:2302.13971*, 2023.
- [10] Konstantinos Meichanetzidis, Stefano Gogioso, Giovanni De Felice, Nicolò Chiappori, Alexis Toumi, and Bob Coecke, “Quantum natural language processing on near-term quantum computers,” *arXiv preprint arXiv:2005.04147*, 2020.
- [11] Robin Lorenz, Anna Pearson, Konstantinos Meichanetzidis, Dimitri Kartsaklis, and Bob Coecke, “Qnlp in practice: Running compositional models of meaning on a quantum computer,” *Journal of Artificial Intelligence Research*, vol. 76, pp. 1305–1342, 2023.
- [12] 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)*. IEEE, 2022, pp. 8602–8606.
- [13] Jarrod R McClean, Sergio Boixo, Vadim N Smelyanskiy, Ryan Babbush, and Hartmut Neven, “Barren plateaus in quantum neural network training landscapes,” *Nature Communications*, vol. 9, no. 1, pp. 4812, 2018.
- [14] Anthony M Smaldone, Yu Shee, Gregory W Kyro, Marwa H Farag, Zohim Chandani, Elica Kyoseva, and Victor S Batista, “A hybrid transformer architecture with a quantized self-attention mechanism applied to molecular generation,” *Journal of Chemical Theory and Computation*, vol. 21, no. 10, pp. 5143–5154, 2025.
- [15] Ren-Xin Zhao, Jinjing Shi, and Xuelong Li, “Qksan: A quantum kernel self-attention network,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.
- [16] Michael A Nielsen and Isaac L Chuang, *Quantum computation and quantum information*, Cambridge University Press, 10th anniversary edition, 2010.
- [17] Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chihying Deng, Roger G Mark, and Steven Horng, “Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports,” *Scientific data*, vol. 6, no. 1, pp. 317, 2019.
- [18] Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni, “Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification,” *Scientific Data*, vol. 10, no. 1, pp. 41, 2023.
- [19] Galadrielle Humblot-Renaux, Sergio Escalera, and Thomas B Moeslund, “Beyond auroc & co. for evaluating out-of-distribution detection performance,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 3881–3890.
- [20] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al., “Qwen2. 5-vl technical report,” *arXiv preprint arXiv:2502.13923*, 2025.