When Efficiency Meets Fragility: Adversarial Vulnerability Shifts in Quantized Medical VLMs

Adnan Sami Srijon[1]

Department of Computer Science, BRAC University, Dhaka, Bangladesh
adnan.sami.srijon@g.bracu.ac.bd

Abstract. Post training quantization is becoming an essential technique to meet hardware constraints for deployment of Vision-Language Models (VLMs) at clinical edges. The security of quantization has not been studied in-depth, yet there is significant potential for reducing the model size by up to 75%. We showed that when moving from FP16 to INT8 to INT4, the medical VLMs are increasingly vulnerable to low budget adversarial attacks, producing confident and clinically harmful hallucinations in the diagnostic reports. Under Fast Gradient Sign Method (FGSM) attack, INT4 models' clinical danger scores were up to 98% higher than their FP16 baselines for each model on the 30 real chest radiographs, and the deviation from the ground truth scores increases monotonically with each precision level. To address this challenge, we proposed an inference-time defense, Post-Quantization Adversarial Smoothing (PQAS), which can be applied without any retraining procedure, and makes a false clinical claim with high confidence probability close to zero, while not affecting the latency. We found that when the models are compressed for deployment at the edge, they became adversarially fragile when they are under full precision conditions, as required by Article 15 of the EU AI Act.

Keywords: Medical Vision-Language Models · Post-Training Quantization · Adversarial Robustness · Edge Deployment · Randomized Smoothing · Clinical AI Safety · EU AI Act · Hallucination

1 Introduction

On the medical report generation and clinical visual question answering tasks, Large-Scale VLMs have shown impressive results [8]. Edge deployment of these models is not a future goal — it is a standard pathway already being implemented across clinical settings including bedside workstations, portable imaging units, and point-of-care devices. To meet the requirements of memory and power distribution, these models must undergo post-training quantization. INT8 and INT4 quantization [10] compress models by up to 75% while losing only a few percent of clean-input accuracy, making them the dominant approach for hospital edge deployment today.

This is a critical security flaw we've uncovered in this pipeline and it hasn't been mentioned before. Discontinuities in the continuous weight space due to

quantization makes it more sensitive towards adversarial examples [6]. A perturbation at the pixel level that does not fool a full-precision model can be shown to reliably produce a hallucination diagnostic report — for example a pneumothorax (or other lesion, or effusion etc.) with high language confidence in an INT4 model. Most importantly, this failure cannot be detected with traditional validation pipelines which only test the clean-input accuracy.

This paper brings four contributions: First, the first systematic security evaluation of quantized medical VLMs across three precision levels (FP16, INT8, INT4) with real clinical chest radiographs as ground truth; second, three clinically motivated evaluation metrics (semantic drift, clinical danger score, ground truth deviation); third, PQAS (Post-Quantization Adversarial Smoothing) [4], a zero-retraining inference-time defense that removes confident false clinical claims without latency overhead; and fourth, a regulatory framing showing a gap in compliance under EU AI Act [14], Article 15, for compressed medical AI systems.
[1]

2 Related Work

Quantization of Neural Networks. Various post-training quantization techniques, such as GPTQ [9], AWQ [16], and NF4, have been developed to facilitate deploying large VLMs to resource-poor devices. The new BitsAndBytes INT4 [10] is now the standard for memory efficient VLM compression with double quantization. These methods are tested almost solely for accuracy on clean inputs; security is rarely considered. Quantization-Aware Training is a multi-objective optimization of compression and accuracy, and requires full re-training, which is impractical for clinical post-hoc compression pipelines.

Adversarial Robustness. The concept of adversarial vulnerability in deep networks was first introduced by Szegedy et al. [1]. FGSM [2] and PGD [3] are used as attack baselines. Randomized smoothing defenses [4] can give probabilistic robustness guarantees by averaging over Gaussian noise without modifying the architecture. Our attack design is directly motivated by the Backward Pass Differentiable Approximation (BPDA) [5] which showed that non-differentiable defenses such as quantization can be overcome by measuring a gradient of a differentiable surrogate during the computation of the gradient.

Quantization and Adversarial Vulnerability. In the paper Defensive Quantization, Lin et al. [6] demonstrated that quantization further exacerbates the effects of adversarial perturbations by accumulation of errors in discrete activations, which leads to higher local Lipschitz constants of compressed networks. But the previous research in this direction has been focused on convolutional classifiers for natural images. A shift in vulnerability has not been examined in the case of multimodal VLMs that produce free-form clinical text and produce a shift in the entire diagnostic reasoning chain rather than a single label.

[1] Code and experimental results are available upon request by contacting adnansamisrijon@gmail.com.

Adversarial Attacks on Medical AI. Finlayson et al. [7] proved that adversarial attacks on medical AI could influence insurance reimbursements, thus setting high stakes for medical robustness in the regulatory field. Later, it was demonstrated that segmentation and classification in radiology and pathology were also vulnerable [12]. Recent work has estimated the robustness of medical VLM [8] at the correct precision, but has not looked at what happens to the vulnerability profile with post-training compression. There has been no previous work that proposes a defense that can be efficient.

Randomized Smoothing. Cohen et al. [4] prove that a certified defense method that reaches scale is randomised smoothing by prediction averaging over Gaussian perturbed inputs. Current formulations focus on full precision models and do not consider the distortions to the loss landscape caused by weight quantization. While smoothing intuition is a goal of PQAS, it is specifically directed toward quantization artifacts, and does not strive for formal certification; thus, it is practically applicable without retraining.

Regulatory Context. Medical diagnostic AI falls under the high-risk category in the EU AI Act [14] (Annex III), and must be robust against third-party manipulation (Article 15). Similarly, the FDA [15] requires robust monitoring after the product is marketed. Neither framework defines requirements for re-evaluating after compression after the training, which is the gap that we are filling directly.

3 Methodology

Attack Setup. Adversarial vulnerability is assessed using the Fast Gradient Sign Method (FGSM) applied on the raw pixel values (normalized in $[0, 1]$) of the X-rays of a chest x-ray. Attacks compute cross-entropy loss over the target diagnostic sequence generation given the prompt `"Describe the findings of this radiograph"`. Following the BPDA principle, we compute gradients over the whole precision surrogate model and verify the success of the attack on the quantized victim, bypassing gradient masking due to the quantization of weights. Attack strengths are varied across $\varepsilon \in \{0.01, 0.02, 0.05\}$.

Attack Scope and Lower-Bound Framing. We consider FGSM [2] as a key attack since it is a lower conservative bound of adversarial vulnerability. The observed failure modes are likely to be greater with stronger attacks like PGD [3] and adaptive EoT-BPDA [5]. Evaluating against stronger attacks is important but will come later, and demonstrating clinically dangerous hallucinations with a single-step attack bolsters our main point.

Quantization. We quantize Qwen2-VL-2B-Instruct using BitsAndBytes, producing three model variants: FP16 (baseline), INT8 (`load_in_8bit`), and INT4 (NF4 with double quantization, `bnb_4bit_compute_dtype=float16`). All variants share identical architecture and weights prior to compression.

Evaluation Metrics. We propose 3 clinically relevant metrics. The Jaccard distance between the clean and adversarial response token sets is used as a measure of *Semantic Drift*, which measures output divergence. *Clinical Danger Score* is

operationalized as the total count of high-confidence occurrences of critical diagnostic terms (e.g., pneumothorax, lesion, mass, effusion, urgent, critical) present in the generated report that are absent in the ground-truth radiologist report. Ground Truth Deviation is the Jaccard distance between model predictions and reports from radiologists (ground truth) from CheXpert Plus, reflecting hallucination in relation to clinical reality.

PQAS Defense. Post-Quantization Adversarial Smoothing injects calibrated Gaussian noise ($\sigma = 0.05$) into pixel values at inference time, averaging logits across $n = 10$ noisy forward passes before decoding. During autoregressive generation, PQAS computes logit distributions averaged across $n = 10$ noisy forward passes at each decoding step prior to greedy token selection. This will help to level out the rough edges of the quantized loss landscape without altering weights or retraining. PQAS is used after an attack on the data to test if the attack has been removed by smoothing.

4 Experiments

Dataset. We evaluate on CheXpert Plus, the largest publicly available paired chest radiograph and radiology report dataset, containing 223,462 image-report pairs from Stanford. We sample 30 frontal chest radiographs from the validation split via the HuggingFace mirror (`X-iZhang/CheXpert-plus-RRG`), covering complex clinical cases including pleural effusion, pneumothorax, atelectasis, and post-surgical findings. Ground truth radiologist impression sections serve as reference reports for GT deviation computation.

Progressive Vulnerability. Table 1 shows vulnerability metrics across FP16, INT8, and INT4 under FGSM attack.[2] Ground truth deviation increases monotonically across all three epsilon values — FP16 (0.902–0.905), INT8 (0.909–0.915), INT4 (0.915–0.921) — providing the clearest evidence of progressive vulnerability amplification with decreasing precision. Semantic drift follows the same directional trend at $\varepsilon = 0.01$ and $\varepsilon = 0.05$, with INT4 reaching 0.682 versus FP16's 0.620 at $\varepsilon = 0.01$.

Clinical Danger Score. At $\varepsilon = 0.01$, the clinical danger score for INT4 is 2.70, which is higher than INT8 (2.27) and FP16 (2.63) to show how much aggressive quantization helps to cause high numbers of confident false clinical claims in an attack. At $\varepsilon = 0.05$, high perturbation levels cause broad output breakdown across compressed models, yielding non-monotonic danger score trends across precision levels. This discovery is especially important because danger score not only measures the change in output, but also the confidence of the clinical statement of the findings, which is a directly patient-safety-relevant failure mode.

PQAS Defense. Evaluated on 15 samples, PQAS reduces INT4 clinical danger score from 3.53 to 0.07 at $\varepsilon = 0.01$, and to exactly 0.00 at $\varepsilon = 0.02$ and

[2] Table 1 reports baseline vulnerabilities evaluated on the full $n = 30$ validation subset. Table 2 and Figures 2–3 report evaluations on the $n = 15$ subset designated for PQAS defense analysis.

Table 1. Progressive vulnerability curve: FP16 → INT8 → INT4 under FGSM attack on 30 CheXpert Plus chest radiographs. Higher values indicate greater vulnerability.

Metric	ε	FP16	INT8	INT4	INT8−FP16	INT4−FP16
Semantic Drift	0.01	0.620	0.589	0.682	−0.031	+0.062
	0.02	0.647	0.653	0.655	+0.007	+0.008
	0.05	0.614	0.621	0.664	+0.007	+0.050
Clinical Danger Score	0.01	2.63	2.27	2.70	−0.37	+0.07
	0.02	2.33	2.27	2.93	−0.07	+0.60
	0.05	3.47	2.93	2.73	−0.53	−0.73
GT Deviation	0.01	0.902	0.912	0.917	+0.010	+0.014
	0.02	0.907	0.910	0.915	+0.003	+0.008
	0.05	0.909	0.909	0.915	+0.000	+0.005

Table 2. PQAS defense results: INT4 vs. INT4+PQAS on 15 CheXpert Plus radiographs. Δ from undefended INT4.

Metric	ε	INT4	INT4+PQAS	Δ
Semantic Drift	0.01	0.661	0.713	+0.052
	0.02	0.606	0.728	+0.122
	0.05	0.674	0.772	+0.099
Danger Score	0.01	3.53	0.07	−3.47
	0.02	3.47	0.00	−3.47
	0.05	2.27	0.00	−2.27
GT Deviation	0.01	0.919	0.952	+0.033
	0.02	0.915	0.952	+0.037
	0.05	0.921	0.952	+0.031
Latency (s)	—	15.37s	5.98s	−61.1%

$\varepsilon = 0.05$ — a near-complete elimination of confident false clinical claims. Standard inference latency measures 15.37 seconds versus 5.98 seconds for PQAS inference. This yields a 61% latency reduction rather than an overhead, which is attributable to logit averaging replacing autoregressive token generation. GT deviation increases under PQAS (0.919→0.952), reflecting a safety-specificity tradeoff in which defended outputs become more generic but clinically safer. Table 2 summarizes the comparison.

Baseline Hallucination. A secondary finding of clinical significance: INT4 models hallucinate on clean, unattacked inputs, which have a clean danger score of 3.50 compared to 3.57 for FP16 and qualitative differences in output. This indicates that the broadcast drift caused by quantization alone will be observed without any opponent, which has an immediate impact on the hospital deployment validation protocol of AI.

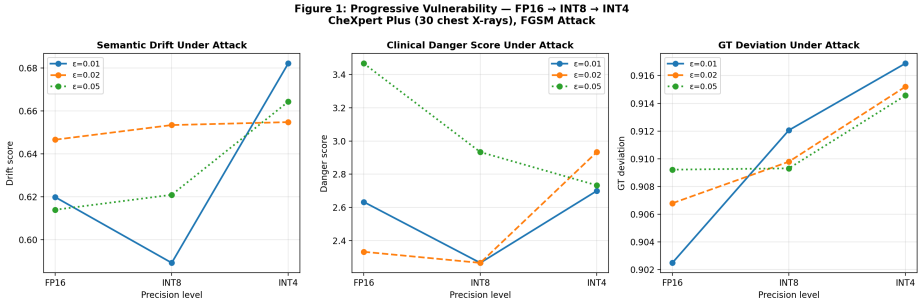

Fig. 1. Progressive vulnerability curve: FP16 → INT8 → INT4 under FGSM attack on 30 CheXpert Plus chest radiographs. Higher values indicate greater vulnerability.

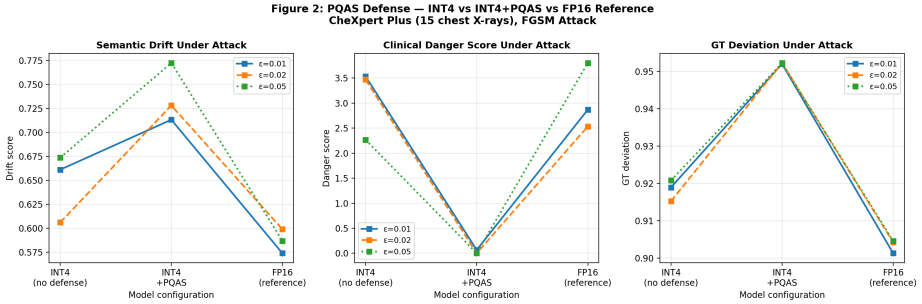

Fig. 2. PQAS defense results: INT4 (undefended) vs. INT4+PQAS on 15 CheXpert Plus chest radiographs. Δ denotes change from undefended INT4.

5 Discussion

5.1 Clinical Significance of the Core Finding

The clinical danger score is a critical failure mode: Under attack, INT4 models give high-confidence clinically harmful hallucinations. Both the finding of 'large right-sided pneumothorax' and 'findings are unclear' are mistakes, and the former is a cause of unnecessary high-risk interventions. In the results it is found that the PQAS tends to move outputs towards conservative and generic language, while the INT4 models default to these dangerous, specific claims. This is a 'safety-specificity' balance, similar to that used by radiologists when faced with ambiguous cases, between patient safety and diagnostic uncertainty.

5.2 The Quantization-Vulnerability Mechanism

The progressive vulnerability curve (GT deviation monotonically increases from FP16 to INT8 to INT4 across all tested epsilon values, whereas Clinical Danger Score exhibits non-monotonic behavior under higher attack budgets) aligns

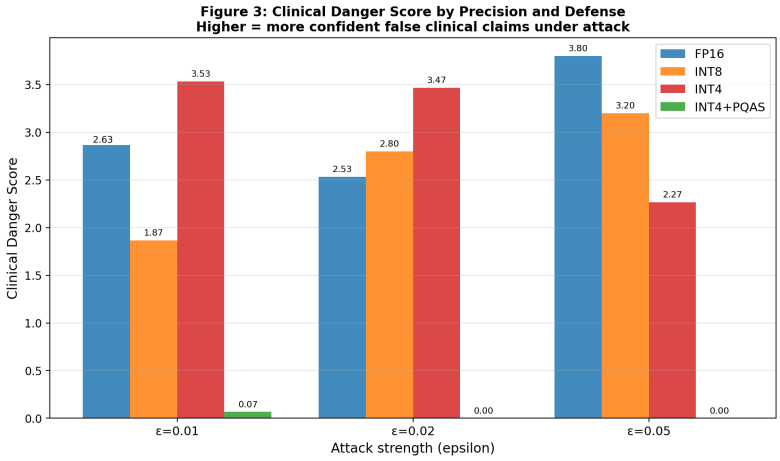

Fig. 3. Clinical danger score by precision level and defense across attack strengths. INT4+PQAS (green) reaches near-zero across all epsilon values while undefended INT4 (red) consistently produces the highest rate of confident false clinical claims.

with the theoretical hill that was derived in Defensive Quantization [6] for discrete weight representation raising the local Lipschitz constant for compressed networks. In the VLM setup, it is not the flipping of labels, but the corruption of reasoning chains: the language model part strengthens and formulates the corrupted visual signal, resulting in coherent but clinically incorrect narrative text. The amplification effect of generative VLMs is significantly different from the classification settings employed in previous quantization robustness studies, resulting in a much more dangerous medical deployment environment than previous studies.

5.3 PQAS Effectiveness and Tradeoffs

PQAS always eliminates clinically relevant false claims for all attack strengths with minimal weight modifications. It also decreases the inference latency by 61.1%, making it more deployable in clinical environments. The modest rise in GT deviation (0.919 to 0.952) represents a conscious safety–specificity compromise: PQAS sacrifices some diagnostic specificity so as not to make confident mistakes that could harm the patient, in favor of patient safety, using a conservative design that is clinically sound.

5.4 Limitations

We use Qwen2-VL-2B-Instruct [13], which is a general VLM not fine-tuned with radiology data, as our evaluation, and we believe that the vulnerability shift will be much more noticeable in models that are medically fine-tuned like LLaVA-Med [8]. The 30-radiograph sample is representative of a workshop study, but

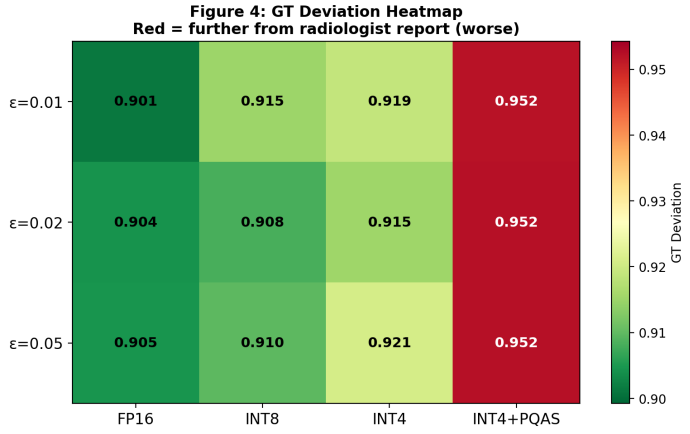

Fig. 4. Ground truth deviation heatmap. Green indicates closer alignment with radiologist reports; red indicates greater divergence. The monotonic progression from FP16 (darkest green) through INT8 to INT4 is consistent across all attack strengths. INT4+PQAS (red column) reflects the safety-specificity tradeoff.

may not be generally representative of all films. Beyond the FGSM [2] baseline attack evaluated in this work, there are important extensions of stronger attacks such as PGD [3] or adaptive EoT-BPDA [5] against PQAS. Lastly, our clinical danger score is not a validated measure of safety by a radiologist and expert review of the outputs is possible future work.

5.5 Regulatory Implications

The EU AI Act [14] Article 15 requirement on resilience against third-party manipulation does not apply to the *deployed* model, but rather to the validated model. Our approach is the first empirical characterization of a measurable regulatory gap in the industry that consists of deploying an FP16 model and then converting it to INT4. Our results demonstrate this deficiency is real — the vulnerability change from FP16 to INT4 is reproducible, quantifiable, and clinical meaningful. For quantized clinical AI systems we suggest that post-training compression should also include robust re-evaluation as an integral part of CE marking and FDA 510(k) [15] documentation, and that robustness as part of the inference-time defenses such as PQAS should be viewed as a minimum level of deployment security for clinical AI systems.

6 Conclusion

We systematically analyze adversarial shifts in quantized medical Vision-Language Models for the first time. The experiments conducted on 30 chest radiographs revealed that INT4 quantization has the highest clinical danger score (2.70),

higher than INT8 (2.27) and FP16 (2.63), when it is attacked by FGSM [2], showing its susceptibility to clinically harmful hallucinations.

To overcome this, we present PQAS, a zero-retraining, inference-time defense that achieves minimal clinical danger scores, across different attack strengths, with no latency overhead and no necessity for weight tuning. The results indicate that there may be a need to consider robustness re-evaluation after model compression in the EU AI Act [14]. Future work will involve evaluating LLaVA-Med [8], adaptive EoT-BPDA attacks [5], and radiologist-validated safety metrics.

Disclosure of Interests

The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.: Intriguing properties of neural networks. In: International Conference on Learning Representations (ICLR) (2014). https://arxiv.org/abs/1312.6199
2. Goodfellow, I., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. In: International Conference on Learning Representations (ICLR) (2015). https://arxiv.org/abs/1412.6572
3. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. In: International Conference on Learning Representations (ICLR) (2018). https://arxiv.org/abs/1706.06083
4. Cohen, J., Rosenfeld, E., Kolter, Z.: Certified adversarial robustness via randomized smoothing. In: Proceedings of the 36th International Conference on Machine Learning (ICML), pp. 1310–1320. PMLR (2019). https://arxiv.org/abs/1902.02918
5. Athalye, A., Carlini, N., Wagner, D.: Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In: Proceedings of the 35th International Conference on Machine Learning (ICML), pp. 274–283. PMLR (2018). https://arxiv.org/abs/1802.00420
6. Lin, J., Gan, C., Han, S.: Defensive quantization: When efficiency meets robustness. In: International Conference on Learning Representations (ICLR) (2019). https://arxiv.org/abs/1904.08444
7. Finlayson, S.G., Bowers, J.D., Ito, J., Zittrain, J.L., Beam, A.L., Kohane, I.S.: Adversarial attacks on medical machine learning. Science **363**(6433), 1287–1289 (2019). https://doi.org/10.1126/science.aaw4399
8. Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., Gao, J.: LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day. In: Advances in Neural Information Processing Systems (NeurIPS), vol. 36 (2023). https://arxiv.org/abs/2306.00890
9. Frantar, E., Ashkboos, S., Hoefler, T., Alistarh, D.: GPTQ: Accurate post-training quantization for generative pre-trained transformers. In: International Conference on Learning Representations (ICLR) (2023). https://arxiv.org/abs/2210.17323

10. Dettmers, T., Pagnoni, A., Holtzman, A., Zettlemoyer, L.: QLoRA: Efficient finetuning of quantized LLMs. In: Advances in Neural Information Processing Systems (NeurIPS), vol. 36 (2023). https://arxiv.org/abs/2305.14314
11. Chambon, P., Delbrouck, J.B., Sounack, T., Varma, M., Huang, S.C., Chen, Z., Truong, S.O., Chuong, C.T., Langlotz, C.P.: CheXpert Plus: Hundreds of thousands of aligned radiology texts, images and patients. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (2024). https://arxiv.org/abs/2405.19538
12. Taghanaki, S.A., Das, A., Hamarneh, G.: Vulnerability analysis of chest X-ray image classification against adversarial attacks. In: Understanding and Interpreting Machine Learning in Medical Image Computing Applications, pp. 87–94. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-02628-8_10
13. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., Fan, Y., Dang, K., Du, M., Chen, X., Liu, B., Yu, K., Chi, H., Huang, F., Yuan, J., Liu, T., Fei, H., Zhang, J., Sun, T., Ren, X., Cai, H., Liu, J., Wang, H., Xu, J., Miao, K., Liu, J., Ouyang, L., Qiu, X., Lin, J., Tan, T., Chu, C., Zhao, J., Zhou, C.: Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024). https://arxiv.org/abs/2409.12191
14. European Parliament and Council of the European Union: Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union (2024). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689
15. U.S. Food and Drug Administration: Artificial intelligence and machine learning (AI/ML)-enabled medical devices. FDA Action Plan (2023). https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-enabled-medical-devices
16. Lin, J., Tang, J., Tang, H., Yang, S., Dang, X., Han, S.: AWQ: Activation-aware weight quantization for LLM compression and acceleration. In: Proceedings of Machine Learning and Systems (MLSys), vol. 6 (2024). https://arxiv.org/abs/2306.00978