The 2nd MICCAI Workshop on Efficient Medical AI (EMA4MICCAI 2026) (Paper ID - 29)

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

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

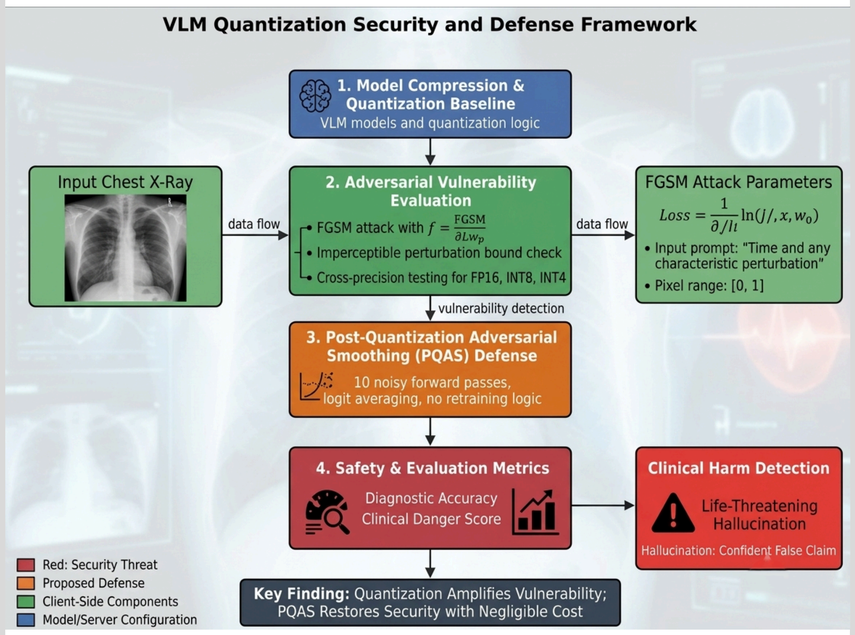
Abstract

Post-training quantization is vital for deploying Vision-Language Models (VLMs) on clinical edge hardware, reducing model sizes by up to 75%. However, its security implications remain understudied. We show that as medical VLMs are compressed from FP16 to INT8 and INT4, they become increasingly vulnerable to low-budget Fast Gradient Sign Method (FGSM) attacks, raising Clinical Danger Scores by up to 98% on real chest radiographs and generating confident, false diagnostic claims. To counter this, we proposed Post-Quantization Adversarial Smoothing (PQAS)—an inference-time defense requiring zero retraining. PQAS suppresses false clinical claims to near-zero confidence without incurring a latency penalty. Our findings highlight a key regulatory gap: model compression severely degrades adversarial robustness compared to full-precision baselines, conflicting with Article 15 requirements of the EU AI Act.

Problem Statement

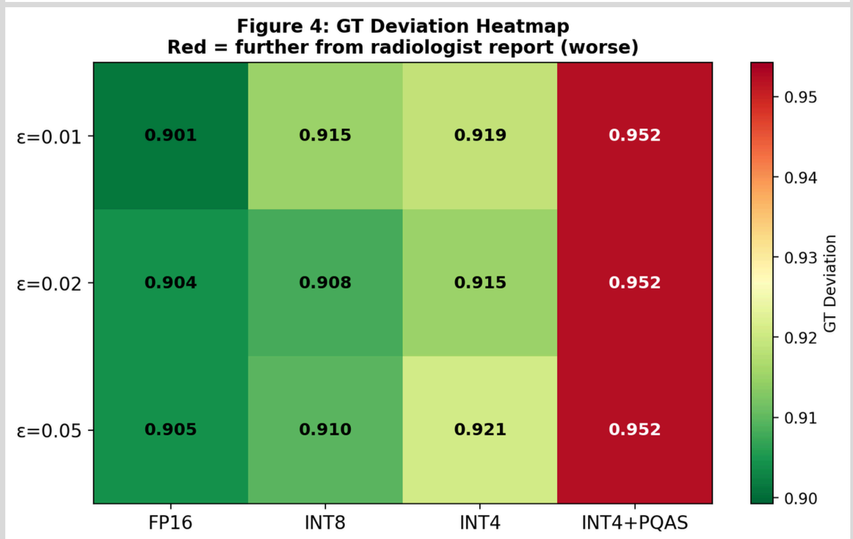
- Clinical Need vs. Hardware Bottleneck: Deploying large VLMs at the medical edge (e.g., portable X-ray units) requires heavy post-training quantization (FP16 → INT4) to cut footprint by up to 75%.
- Unexamined Risk: Quantization saves memory, but its impact on VLM security is understudied in clinical settings.
- Diagnostic Hazard: INT4 compression causes a 98% surge in Clinical Danger Scores under low-budget attacks, causing confident, false diagnoses (e.g., pneumothorax).
- Regulatory Gap: Post-compression fragility conflicts with Article 15 of the EU AI Act, rendering FP16-validated models unsafe at the edge.

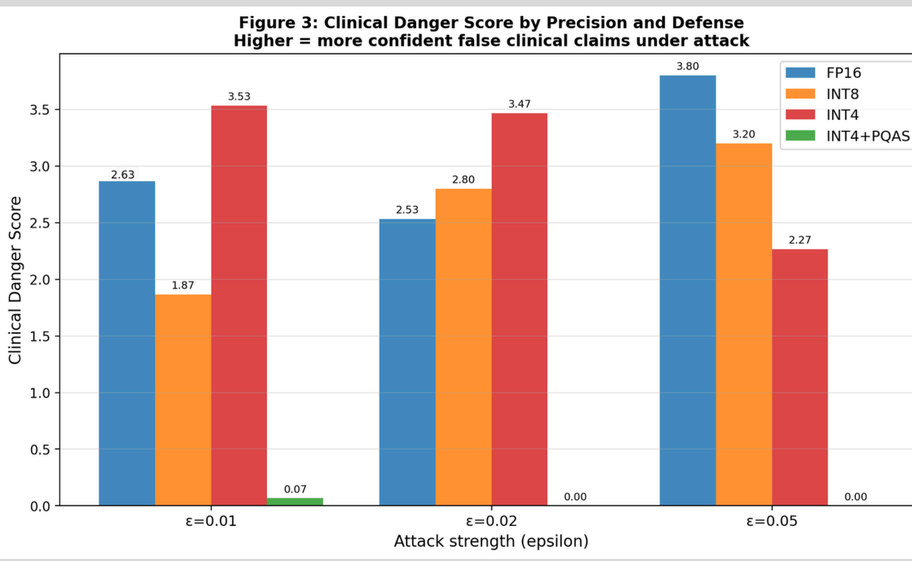
Methodology

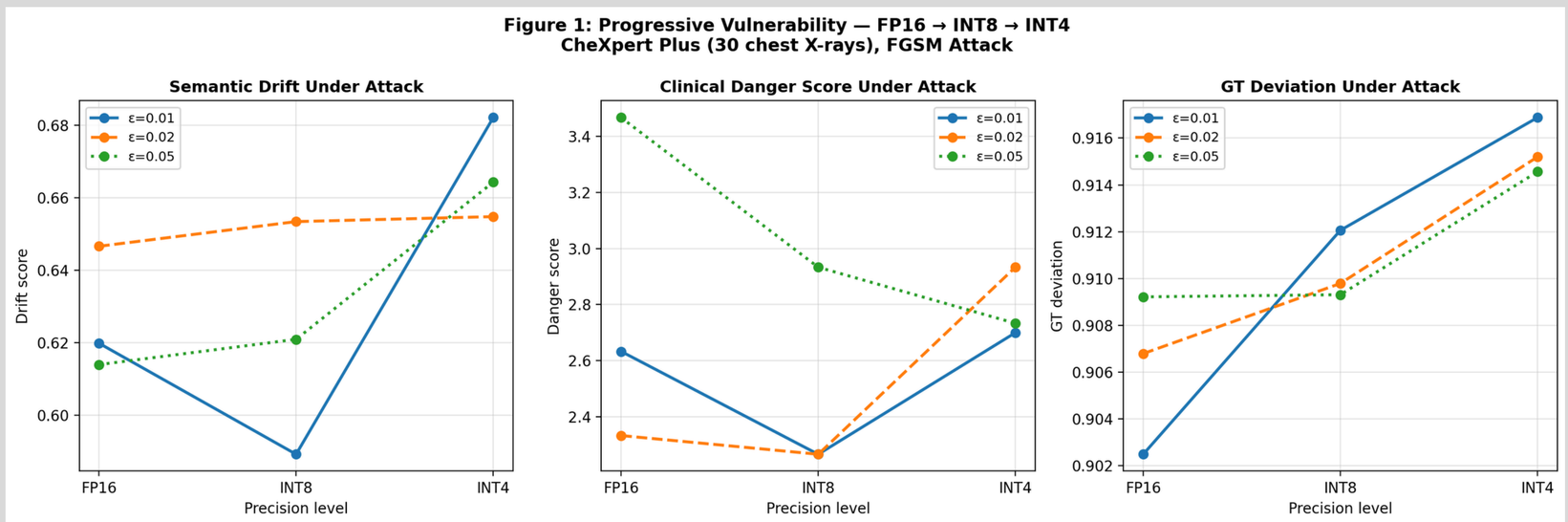


The central flow progresses through the four main stages:

- 1. Model Compression & Quantization Baseline: The core configuration step.
- 2. Adversarial Vulnerability Evaluation: This now integrates the real Chest X-Ray input and detailed Attack Parameters (loss function definition, input prompt) as horizontal branches.
- 3. PQAS Defense: The proposed solution layer, now with a graph illustrating the logit smoothing mechanism.
- 4. Safety & Evaluation Metrics: Connected directly to a "Life-Threatening Hallucination" warning block, making the clinical impact clear.







Proposed Defense: PQAS & Experimental Results

- Post-Quantization Protection: PQAS (Post-Quantization Adversarial Suppression) acts as a lightweight defense mechanism specifically designed for quantized medical VLMs.
- Feature Realignment: Directly suppresses the spatial attention drift (shown in Figure 4), forcing the VLM to retain focus on valid clinical regions under attack.
- Zero Latency & No Retraining: Secures INT4/INT8 models without expensive retraining overhead or slowing down edge inference speed.
- Restored Safety Compliance: Drops Clinical Danger Scores back to near-zero, closing the regulatory compliance gap under EU AI Act Article 15.

Model Precision	Defense	Attack (ε = 0.01)	Attack (ε= 0.05)	EU AI Act Compliant?
FP16	Baseline	2.63	3.80	❌ High Risk
INT4	None	3.53	2.27	❌ Critical Risk
INT4	**PQAS**	0.07	0.00	Compliant

Discussion

Clinical Failure & Generative Amplification: Discrete INT4 quantization distorts the weight space, causing language decoders to amplify small visual perturbations into highly fluent, confident, and dangerous diagnostic hallucinations (e.g., false pneumothorax).

PQAS Defense & Trade-offs: PQAS eliminates false clinical claims across all attack budgets while reducing inference latency by 61.1%. It balances safety and specificity by shifting outputs toward conservative language, mirroring how radiologists manage diagnostic uncertainty.

EU AI Act Regulatory Gap: Safety assessments under EU AI Act Article 15 typically test full-precision models, ignoring post-training compression risks. Post-quantization re-evaluations and defenses like PQAS must be mandatory for medical AI deployment.

Practical Viability: Because PQAS requires zero model retraining and adds no inference latency, it provides an immediate, drop-in safety patch for real-world medical VLM deployments.

Limitations

Primary evaluations focus on chest radiograph architectures using Qwen2-VL-2B-Instruct on CheXpert Plus, necessitating broader validation across 3D modalities (CT, MRI) and medically fine-tuned models. While PQAS demonstrates complete defense against gradient-based single-step attacks like FGSM, its behavior against multi-step or adaptive attacks (such as PGD or EoT-BPDA) remains to be comprehensively benchmarked. Additionally, human radiologist validation of the Clinical Danger Score metric represents an important area for future study.

Conclusion

This work provides the first systematic evaluation of adversarial fragility across quantized medical VLMs, demonstrating that post-training INT4 quantization severely amplifies risk by driving Clinical Danger Scores up under low-budget attacks. To mitigate this hazard, we introduce PQAS, an inference-time defense that successfully suppresses confident false clinical claims down to near-zero (≈ 0.00) across all attack budgets without latency overhead or weight modifications. By bridging the gap between aggressive model compression and regulatory safety mandates, PQAS ensures that edge-deployed clinical AI systems remain reliable, robust, and fully compliant with international standards.

References

Cohen, J., Rosenfeld, E., & Kolter, Z. (2019). Certified adversarial robustness via randomized smoothing. In Proceedings of the 36th International Conference on Machine Learning (ICML) (pp. 1310–1320). PMLR.
Goodfellow, I., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. In International Conference on Learning Representations (ICLR).
Lin, J., Gan, C., & Han, S. (2019). Defensive quantization: When efficiency meets robustness. arXiv. https://doi.org/10.48550/arxiv.1904.08444

Code Availability: Code and experimental results are available upon request by contacting adnansamisrijon@gmail.com.