

Live Demonstration: Hardware-Bound IP Protection for Edge-deployed Transformers

Peichun Hua¹, Hanxiu Zhang², Tuo Li³, Yue Zheng² and Wenye Liu⁴

¹School of Data Science, The Chinese University of Hong Kong, Shenzhen, Guangdong, 518172, P. R. China

²School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen, Guangdong, 518172, P. R. China

³Shandong Yunhai Guochuang Cloud Computing Equipment Industry Innovation Co., Ltd., China; ⁴Independent Researcher
{peichunhua, hanxiuzhang}@link.cuhk.edu.cn, lituo@inspur.com, zhengyue@cuhk.edu.cn, wenye.liu@ieee.org

Abstract—We demonstrate a hardware-bound Intellectual Property (IP) protection scheme for transformer models on edge devices. Our method binds the OWL-ViT [1], an open-vocabulary object detection model, to specific hardware using a Physical Unclonable Function (PUF). The model weights are reversibly obfuscated using PUF-derived cryptographic keys, catastrophically degrading the performance of unauthorized devices and thus preventing IP theft. The proposed method incurs only a moderate inference overhead on the authorized target for real time applications.

I. INTRODUCTION AND METHOD

Transformer models such as OWL-ViT are valuable assets due to the heavy investment in well-annotated training datasets, powerful computing infrastructures, and substantial domain expertise. Deploying these models to edge devices exposes them to the threat of IP theft. To counteract this, we propose an active defense that locks a model to authorized hardware, as shown in Fig. 1. Our framework [2] leverages a PUF to generate a unique device-specific master key K , which is never stored but regenerated on-the-fly. This key is used to reversibly obfuscate critical model weights through a *Dual Encryption* approach: Arnold’s Cat Map (ACM) [3] scrambles the square attention weights, while efficient row permutations are applied to the Feed-Forward Network (FFN) weights. At runtime, the model executes a per-batch *Decrypt → Infer → Re-encrypt* cycle, minimizing the time that plaintext weights reside in memory and mitigating software-based attacks. Without the correct PUF-derived key, the model’s accuracy collapses, rendering it useless to an adversary.

II. LIVE DEMONSTRATION SETUP

The demonstration showcases our protection scheme in a real-world edge AI scenario.

- **Hardware:** An NVIDIA Jetson Orin Nano performs inference for the OWL-ViT model. An Ultra96-V2 FPGA board with an integrated Arbiter PUF serves as the hardware root-of-trust, securely generating the cryptographic key. The two platforms communicate via WiFi using the Python Socket API.
- **Demonstration:** We demonstrate open-vocabulary object detection using the protected OWL-ViT model on images from the Pascal VOC dataset.
- **Visitor Experience:** Observers will first see the system running correctly on the authorized hardware, successfully

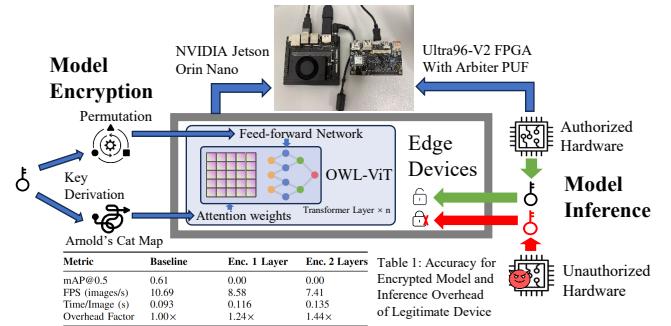


Fig. 1: Setup of our scheme and inference performance

identifying objects from the dataset. We will then simulate an IP theft scenario by running the model with an incorrect key, demonstrating an immediate and catastrophic failure of the detection task. This directly visualizes the effectiveness of the hardware-binding protection.

III. PERFORMANCE RESULTS

Our scheme is fully reversible, restoring the model to its original accuracy when the correct key is present. The primary trade-off is a moderate increase in inference latency due to the on-the-fly cryptographic operations. We benchmarked the OWL-ViT model (google/owlvit-base-patch32) on a batch of 8 images, detecting 20 Pascal VOC classes. As shown in Table 1, encrypting one or two critical transformer layers reduces the model’s mean Average Precision (mAP@0.5) from a functional 0.61 to 0.00 on an unauthorized device. This strong security guarantee comes at a modest cost, reducing the throughput by 19.7% to 30.7% (from 10.69 to 8.58 or 7.41). We show in [2] that encrypting more layers will further enhance the security and robustness of our scheme.

REFERENCES

- [1] M. Minderer *et al.*, “Simple open-vocabulary object detection,” in *Proc. European Conf. Computer Vision (ECCV)*, Tel Aviv, Israel, October 2022, pp. 728–755.
- [2] P. Hua, H. Zhang, T. Li, and Y. Zheng, “Securing on-device transformer with hardware binding and reversible obfuscation,” in *Proc. Annual Computer Security Applications Conference (ACSAC 2025)*, Honolulu, HI, USA, Dec. 2025. [Online]. Available: <https://drive.google.com/file/d/16H3I9SOxO9eW0zDMZY-EFUBGRdXdVmCo/view?usp=sharing>
- [3] J. Fridrich, “Symmetric ciphers based on two-dimensional chaotic maps,” *Int. J. Bifurcation and Chaos*, vol. 8, no. 06, pp. 1259–1284, 1998.