# EO-VLM: VLM-Guided Energy Overload Attacks on Vision Models

Minjae Seo
ETRI

Myoungsung You
KAIST

Junhee Lee
Kwangwoon University

Jaehan Kim
KAIST

Hwanjo Heo
ETRI

Jintae Oh
ETRI

Jinwoo Kim
Kwangwoon University

## Abstract

Vision models are increasingly deployed in critical applications such as autonomous driving and CCTV monitoring, yet they remain susceptible to resource-consuming attacks. In this paper, we introduce a novel energy-overloading attack that leverages vision language model (VLM) prompts to generate adversarial images targeting vision models. These images, though imperceptible to the human eye, significantly increase GPU energy consumption across various vision models, threatening the availability of these systems. Our framework, EO-VLM (Energy Overload via VLM), is model-agnostic, meaning it is not limited by the architecture or type of the target vision model. By exploiting the lack of safety filters in VLMs like DALL·E 3, we create adversarial noise images without requiring prior knowledge or internal structure of the target vision models. Our experiments demonstrate up to a 50% increase in energy consumption, revealing a critical vulnerability in current vision models.

**ACM Reference Format:**

Minjae Seo, Myoungsung You, Junhee Lee, Jaehan Kim, Hwanjo Heo, Jintae Oh, and Jinwoo Kim. 2024. EO-VLM: VLM-Guided Energy Overload Attacks on Vision Models. In *Proceedings of Annual Computer Security Applications Conference (ACSAC '24).* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

## 1 Introduction

The development of vision models has become increasingly prevalent, driving advancements in various industries. For instance, autonomous vehicles, such as Tesla's Autopilot, rely on sophisticated vision models for safe navigation, while CCTV systems use them for enhanced monitoring and threat detection. In these applications, agile and accurate perception is essential for real-time decision-making, ensuring both functionality and safety. However, recent resource-consuming attacks pose significant risks to the utilization of vision models. For example, Shumailov et al. [4] introduce sponge examples, which drain the energy consumed by a neural network, pushing the underlying hardware towards its worst-case performance. They adopt two approaches: (i) a gradient-based (white-box) attack that requires access to the DNN model's parameters, and (ii) a genetic algorithm (black-box) attack that evolves inputs based on energy or latency measurements, without access to the model.

Following this study, more advanced attacks have emerged, specifically targeting the vulnerabilities of vision models. For instance, Overload [1] creates adversarial images designed to attack object detection models. It exploits the Non-Maximum Suppression (NMS) process in object detection models to increase the number of predicted objects in an image, significantly increasing inference time. Similarly, SlowTrack [3] targets the entire processing pipeline of camera-based vision models by exploiting vulnerabilities in both the object detection and object tracking. It injects a number of fake objects, which creates excessive tracking boxes, and prevents the target model from removing the injected objects by periodically re-injecting them. This results in an accumulation of effective tracking boxes, increasing inference latency. Additionally, Gao et al. [2] propose generating verbose images that cause a vision-language model (VLM) to produce lengthy output sequences, thereby consuming substantial computational resources.

While these methods effectively increase inference time in vision models, they have two major limitations that restrict their broader applicability. First, both methods assume a white-box setting for the target vision models, where adversaries have full access to the model's architecture and internal parameters—an assumption that is overly strong and unrealistic in most real-world scenarios. Second, these methods are highly target-specific. For instance, Overload [1] is tailored specifically for object detection models, SlowTrack [3] is designed for multi-object tracking systems, and Gao et al. [2] target VLMs. Adversaries must manually learn and adapt these strategies for specific models within their test environments, a process that is both time-consuming and costly. This significantly impedes the rapid adoption of such methods across diverse vision models.

In this paper, we propose a novel framework called **EO-VLM** (Energy Overload via VLM) for automatically generating adversarial images that significantly increase energy consumption on a target model's GPUs, irrespective of the vision model architecture. Our key idea is to leverage a VLM with a series of carefully crafted prompts to: (i) identify factors that effectively raise the energy consumption of a target vision model and (ii) incorporate these factors into a given image to maximize GPU workload. This approach takes advantage of the fact that popular VLMs like DALL·E 3 lack robust safety filters, making them susceptible to generating adversarial noise images through simple prompts. Consequently, adversaries can automatically generate noise examples without incurring costs or requiring prior knowledge of the models' internal structures. In our experiments, adversarial examples generated by our framework showed up to a 50% increase in energy consumption of various vision models, highlighting the effectiveness and generalizability of such attacks.


*ACSAC '24, December 09–13, 2024, Waikiki, Hawaii, USA*

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

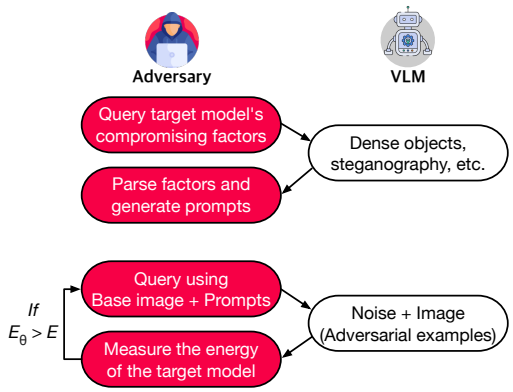


**Figure 1: EO-VLM framework overview.**

## 2 EO-VLM Framework

To generate adversarial examples while maintaining a model-agnostic approach, we follow the process outlined in Figure 1. First, we query the target model's compromising factors—elements that contribute to energy overloading—using a VLM. For instance, DALL·E 3 suggests compromising factors such as increasing anchor box proposals by subtly modifying pixel values or incorporating steganography to complicate feature extraction (see Figure 2). These factors are analyzed, and adversarial prompts are generated using the structured approach: $P_{\text{adv}} = \texttt{concat}(P_{\text{object}}, P^{(i)}_{\text{strategy}}, P_{\text{action}})$, where $P_{\text{object}}$ defines the task objective (e.g., "My objective is to increase resource consumption of YOLOv8"), $P^{(i)}_{\text{strategy}}$ represents the $i$-th strategy utilizing various compromising factors (e.g., "Introduce invisible noise with dense objects"), and $P_{\text{action}}$ specifies the action to achieve the goal (e.g., "Would you combine the noise with the image to maximize energy usage?").

We then query the VLM with the base images and the generated adversarial prompts, producing adversarial examples that integrate the noise into the image. The energy cost of these adversarial images is measured using $E = W \cdot t$, where $W$ is the total power consumption of the GPU, and $t$ is the time taken for the inference [4]. If the energy cost does not exceed a predefined threshold ($E_\theta$) the framework iteratively selects new prompt combinations, regenerates adversarial examples, and recalculates energy consumption until the threshold is surpassed.

## 3 Preliminary Results

In our experiments, we evaluate the power consumption and inference time overhead on YOLOv8, MASKDINO, and Detectron2 models, running on a server equipped with an RTX 4090 GPU. The results demonstrate significant resource overhead induced by adversarial images. As shown in Table 1, YOLOv8 exhibits a 44.4% increase in power consumption with object-based adversarial images and a 44.5% increase with steganography-based images. Similarly, MASKDINO shows a 13.1% and 14% increase, while Detectron2 records a 10.9% and 18.4% increase for object-based and steganography-based adversarial images, respectively. In terms of inference time, as presented in Table 2, YOLOv8 experiences a 21.3% and 23.3% increase for object-based and steganography-based images, respectively. MASKDINO shows a 29.7% and 40.6% increase,

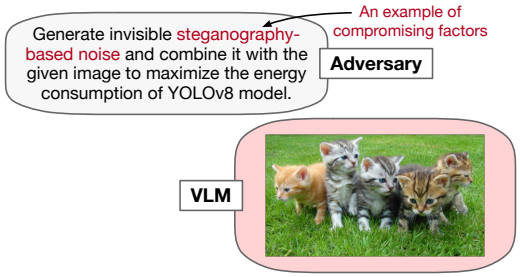


**Figure 2: A generated adversarial image using DALL·E 3.**

**Table 1: Power consumption overhead.**

| Model | YOLOv8 | MASKDINO | Detectron2 |
|---|---|---|---|
| Base image | 46.96 W | 61.44 W | 54.53 W |
| Object-based | 67.83 W (+ 44.4%) | 69.45 W (+ 13.1%) | 60.45 W (+ 10.9%) |
| Steganography | 67.86 W (+ 44.5 %) | 70.02 W (+ 14%) | 64.54 W (+ 18.4%) |

**Table 2: Inference time overhead.**

| Model | YOLOv8 | MASKDINO | Detectron2 |
|---|---|---|---|
| Base image | 0.30 ms | 2.56 ms | 0.20 ms |
| Object-based | 0.36 ms (+ 21.3%) | 3.32 ms (+ 29.7%) | 0.30 ms (+ 50%) |
| Steganography | 0.37 ms (+ 23.3%) | 3.60 ms (+ 40.6%) | 0.28 ms (+ 40%) |

while Detectron2 demonstrates a 50% and 40% increase for the same types of adversarial images. Notably, these substantial increases are achieved using a single adversarial image, even though the file size of the adversarial image is smaller than the base image.

## 4 Conclusion and Future Work

We introduce EO-VLM, a novel framework for conducting energy-overloading attacks. By exploiting the absence of safety filters in the VLM, we generate adversarial images that increase energy consumption by up to 50% with a single image, despite the smaller file size compared to the original. For future work, we aim to incorporate a reinforcement learning approach to systematically generate adversarial prompts, further maximizing energy overloading.

## Acknowledgments

This work was partly supported by Institute of Information & communications Technology Planning & Evaluation (IITP) and by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2021-0-00118, RS-2024-00457937).

## References

[1] Erh-Chung Chen, Pin-Yu Chen, I Chung, Che-Rung Lee, et al. 2024. Overload: Latency attacks on object detection for edge devices. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 24716–24725.

[2] Kuofeng Gao, Yang Bai, Jindong Gu, Shu-Tao Xia, Philip Torr, Zhifeng Li, and Wei Liu. 2024. Inducing high energy-latency of large vision-language models with verbose images. *arXiv preprint arXiv:2401.11170* (2024).

[3] Chen Ma, Ningfei Wang, Qi Alfred Chen, and Chao Shen. 2024. Slowtrack: Increasing the latency of camera-based perception in autonomous driving using adversarial examples. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 4062–4070.

[4] Ilia Shumailov, Yiren Zhao, Daniel Bates, Nicolas Papernot, Robert Mullins, and Ross Anderson. 2021. Sponge examples: Energy-latency attacks on neural networks. In *2021 IEEE European symposium on security and privacy (EuroS&P)*. IEEE, 212–231.