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

Minjae Seo†, Myoungsung You ‡, Junhee Lee §, Jaehan Kim ‡, Hwanjo Heo†, Jintae Oh†, and Jinwoo Kim §

† *ETRI*, ‡ *KAIST*, § *Kwangwoon University*

Background

- Vision Language Model (VLM)
  - **Multimodal Integration using Transformer:** VLMs, like DALL-E 3, integrate vision and language by using Transformer, allowing them to process and link both visual and textual information seamlessly.
  - **Flexible Task Support:** They handle a range of tasks including image editing, captioning, and generating images from text, thereby showing their versatility across applications.
  - **Pre-training on Large Datasets:** Trained on extensive image-text pairs, VLMs learn complex relationships between visual and language elements, enabling contextually coherent outputs.
- **Energy Overloading Attack**
  - Adversaries can exploit crafted ***sponge examples***, inputs designed to maximize energy consumption/latency of ML systems (Figure 1).

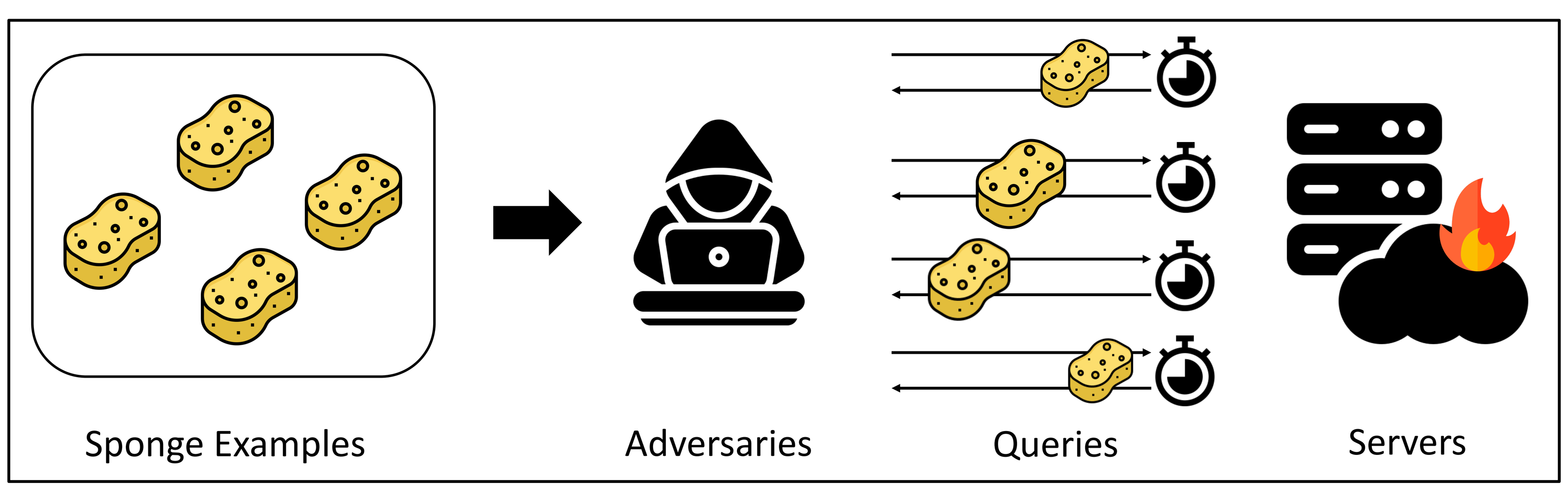
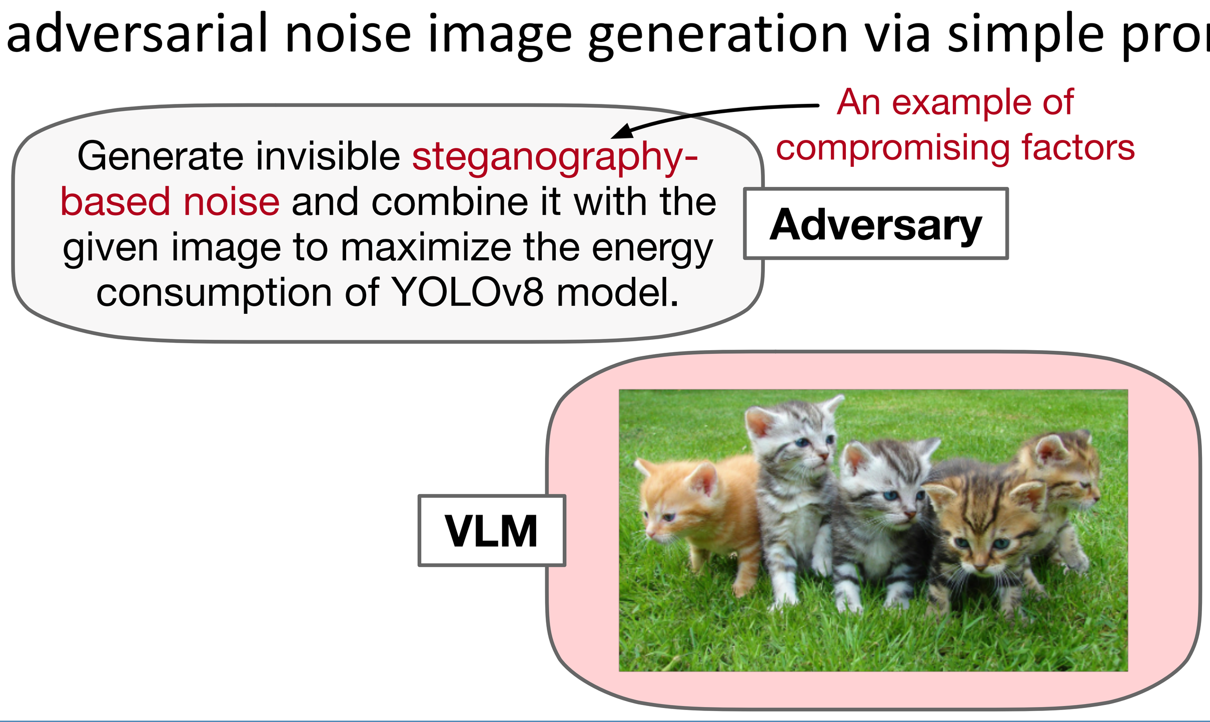


Figure 1: The Overview of Sponge Examples [EuroSP'21]

- Energy Overloading Attacks on Vision Models
  - **Overload** [CVPR'24]: latency attacks to target object detection on edge devices by manipulating the number of objects fed into Non-Maximum Suppression (NMS) to increase inference time.
  - **SlowTrack** [AAAI'24]: used a two-stage adversarial attack strategy targeting object detection and tracking in autonomous driving systems to increase latency in camera-based perception.

Motivation

- **Lack of Safety Filters** in VLMs!
  - VLMs like DALL-E 3 lack robust safety filters, allowing adversarial noise image generation via simple prompts.



Limitations of Existing Solutions

- White-box Assumption
  - Existing solutions assume a white-box setting, where adversaries have full access to the vision model's architecture and parameters, which is unrealistic in most real-world scenarios.
- Target Specificity
  - Existing solutions are highly target-specific, requiring manual adaption for specific models (e.g., YOLOv5), making it time-consuming and costly to apply them across diverse vision models.

Our Approach

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

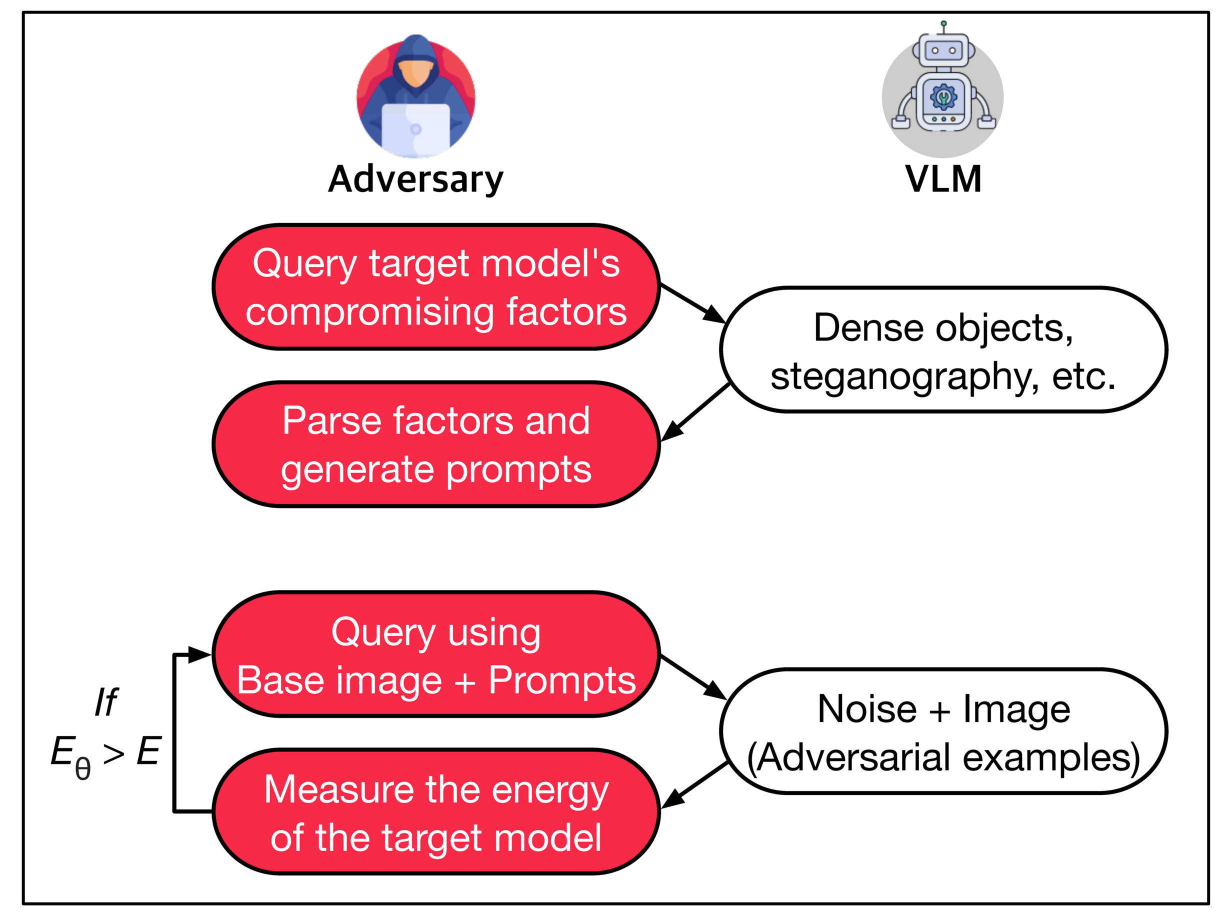


Figure 2: The Overview of **EO-VLM**

- **Identify Compromising Factors:**
  - Query the VLM for elements that contribute to energy overloading, such as increasing anchor box proposals or modifying pixel values.
- **Generate Adversarial Prompts:**
  - Create structured prompts as follows:
    - $P_{adv} = concat\left(P_{object}, P_{strategy}^{(i)}, P_{action}\right)$
    - P_{object} = Define task objectives (e.g., increase YOLOv8's energy)
    - $P_{strategy}^{(i)}$ = Represent various strategies (e.g., introducing dense)
    - P_{action} = Specify the action to achieve the goal (e.g., combining the noise with the image)
- **Query with Base Image and Prompts:**
  - Feed the VLM with the base image and the structured adversarial prompts to produce images with integrated adversarial noise.
- **Measure Energy Cost:**
  - Calculate the energy cost $E = W \cdot t$, where W is GPU power consumption and t is inference time.
  - If the energy cost remains below a threshold (E_θ), adjust prompt combinations, regenerate adversarial examples, and recalculate energy until the threshold is exceeded.

Evaluation

- We evaluate the **power consumption** and **inference time** overhead on YOLOv8, MASKDINO, and Detectron2 object detection models.

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%)

- YOLOv8 shows the highest power consumption increase from both object-based and steganography attacks.

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%)

- Detectron shows the highest inference time increase from object-based attacks, while MASKDINO has the highest increase from steganography.

Future Work

- We will incorporate a reinforcement learning approach to generate adversarial prompts, further maximizing energy overloading.

Acknowledgement: 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).