

Supplementary Materials

Anonymous Author(s)

1 DATASETS AND MODELS

We evaluate Blacklight, PIHA, and our ACPT methods on 5 benchmark datasets: CIFAR-10, GTSRB, ImageNet, Flowers, and Pets. The training phase was performed with 3090 GPUs, utilizing PyTorch with the Adam optimizer for 100 epochs. Table 1 summarizes the 5 image classification tasks used in our experiments.

Table 1: A summary of the datasets and the corresponding models used in our experiments.

Dataset	Model	Dimension	Category	Top-1 Acc.
CIFAR-10	ResNet20	32×32×3	10	91.73%
GRSRB	ResNet34	32×32×3	43	94.96%
Flowers	ResNet101	224×224×3	102	85.80%
ImageNet	ResNet152	224×224×3	1000	78.33%
Pets	Vit-B/16	224×224×3	37	93.13%

2 DETECTING QUERY-BASED ATTACKS ON VISION-LANGUAGE MODELS

Our previous experiments have shown that the ACPT method is effective, efficient, and robust at detecting query-based attacks in image classification tasks. Here, we extend ACPT to detect query-based attacks on the image captioning task. Additionally, Figure 1 visualizes the process of query-based attacks, showcasing intermediate adversarial examples such as those generated by Boundary [2], HSJA [3], NESS [6], QEBA [7], Square [1], SurFree [11], ZOO [4], and AttackVLM [13]. While methods like Boundary, HSJA, NESS, QEBA, Square, SurFree, and ZOO are specifically designed for image classification, AttackVLM targets image captioning tasks. This visualization reveals that the underlying process of such attacks is consistent across different tasks.

Unlike query-based attacks on image classification, AttackVLM [13] first employs pre-trained CLIP [12] and BLIP [9] as surrogate models to generate attacks, either by matching image or textual embeddings, aiming to generate targeted responses. These adversarial examples are then transferred to other large Vision-Language Models (VLMs), including MiniGPT-4 [14], LLaVA [10], and BLIP-2 [8]. Furthermore, AttackVLM utilizes query-based attacks that incorporate transfer-based attacks as an initial step, significantly boosting the effectiveness of targeted evasion against such VLMs, aiming for the targeted response generation over large VLMs. Despite these advanced techniques, our experiments show that ACPT can effectively detect AttackVLM attacks within 3 attempts.

3 THE TRADE-OFF: DETECTION RATE VS. FALSE POSITIVES RATE

The trade-off is related to encoder E and the distribution of query data. Following the OARS work [5], we assume an isotropic Gaussian distribution for benign queries $\mathcal{N}(\mathbf{p}_x, I\sigma^2)$, and another Gaussian distribution $\delta \sim \mathcal{N}(0, I\beta^2)$ for adversarial perturbations. A false

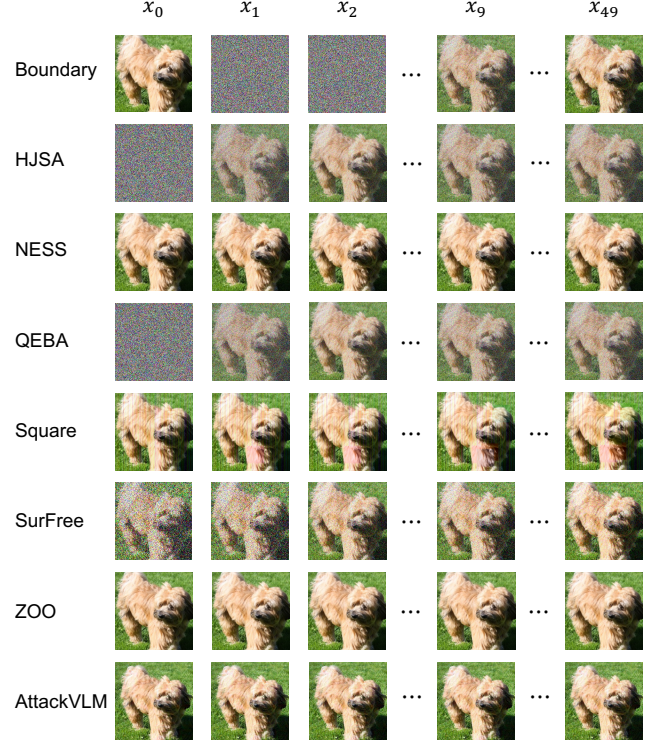


Figure 1: Malicious queries (x_0, x_1, \dots, x_{49}) generated by 8 query-based attacks (Boundary, HSJA, NESS, QEBA, Square, SurFree, ZOO, and AttackVLM) exhibit notable differences during the generation process. However, the sequences of query images produced by these attacks are highly similar to one another.

negative occurs when encoder E fails to identify the malicious query $\mathbf{x} + \delta$, especially if the embeddings of \mathbf{x} and $\mathbf{x} + \delta$ are significantly different, meaning $\text{sim}(E(\mathbf{x}), E(\mathbf{x} + \delta)) \leq \mu$. Consequently, we define the detection rate as $\alpha^{\text{det}} = \mathbb{P}[\text{sim}(E(\mathbf{x}), E(\mathbf{x} + \delta)) \geq \mu]$, while the false positive rate can be expressed as $\alpha^{\text{fp}} = \mathbb{P}[\text{sim}(E(\mathbf{x}_1), E(\mathbf{x}_2)) \geq \mu]$. Furthermore, the trade-off between the detection rate α^{det} and the false positive rate α^{fp} , is influenced by the standard deviation β of the perturbation distribution and the expected spread σ of natural queries. Hence, our observations find that natural images are sufficiently spread out, while adversarial examples generated by the query-based attacks tend to cluster more centrally. This suggests that a stronger encoder can achieve a high detection rate while maintaining a low false positive rate. Additionally, by implementing an effective defense action, such as returning cache predictions, our approach is designed to minimize the impact of false positives on benign users.

REFERENCES

- [1] Maksym Andriushchenko, Francesco Croce, Nicolas Flammarion, and Matthias Hein. 2020. Square attack: a query-efficient black-box adversarial attack via random search. In *European conference on computer vision*. Springer, 484–501.
- [2] Wieland Brendel, Jonas Rauber, and Matthias Bethge. 2018. Decision-Based Adversarial Attacks: Reliable Attacks Against Black-Box Machine Learning Models. In *International Conference on Learning Representations*.
- [3] Jianbo Chen, Michael I Jordan, and Martin J Wainwright. 2020. Hopskipjumpattack: A query-efficient decision-based attack. In *2020 IEEE Symposium on Security and Privacy (SP)*. IEEE, 1277–1294.
- [4] Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. 2017. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM workshop on artificial intelligence and security*. 15–26.
- [5] Ashish Hooda, Neal Mangaokar, Ryan Feng, Kassem Fawaz, Somesh Jha, and Atul Prakash. 2023. Theoretically Principled Trade-off for Stateful Defenses against Query-Based Black-Box Attacks. *arXiv preprint arXiv:2307.16331* (2023).
- [6] Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. 2018. Black-box adversarial attacks with limited queries and information. In *International conference on machine learning*. PMLR, 2137–2146.
- [7] Huichen Li, Xiaojun Xu, Xiaolu Zhang, Shuang Yang, and Bo Li. 2020. Qeba: Query-efficient boundary-based blackbox attack. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 1221–1230.
- [8] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*. PMLR, 19730–19742.
- [9] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*. PMLR, 12888–12900.
- [10] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. *Advances in neural information processing systems* 36 (2024).
- [11] Thibault Maho, Teddy Furon, and Erwan Le Merrer. 2021. Surfree: a fast surrogate-free black-box attack. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10430–10439.
- [12] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*.
- [13] Yunqing Zhao, Tianyu Pang, Chao Du, Xiao Yang, Chongxuan Li, Ngai-Man Man Cheung, and Min Lin. 2024. On evaluating adversarial robustness of large vision-language models. *Advances in Neural Information Processing Systems* 36 (2024).
- [14] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592* (2023).