PA-Attack: Guiding Gray-Box Attacks on LVLM Vision Encoders with Prototypes and Attention

Supplementary Material

1. Out-of-distribution guidance dataset

To investigate the sensitivity of PA-Attack to the distribution of the guidance dataset, we conducted additional experiments using RVL-CDIP [4] (document images) and ScienceQA [7] (scientific diagrams) as guidance sources. These datasets represent a significant domain shift from the natural scenes typically found in COCO or Flickr30k. Specifically, we construct prototypes using 3,000 randomly sampled images from each dataset. As presented in Table 1, PA-Attack exhibits strong robustness to prototype sources, maintaining excellent performance even when the guidance data originates from vertical domains significantly different from the test images.

Table 1. Attack performance with different guidance datasets.

Guidance	Distribution	COCO	Flickr30k	TextVQA	SRR ↑
Clean	In	115.5	77.5	37.1	0.0
COCO	In	4.1	3.3	5.1	92.8
RVL-CDIP	Out	4.1	3.7	6.8	91.1
ScienceQA	Out	4.2	3.6	5.8	92.0

2. More recent and diverse LVLMs

To verify that our method remains effective against the latest advancements in multimodal learning, we extend our experiments to include Qwen3-VL-8B [1] and InternVL2-8B [2, 3] on RealWorldQA [10], ODinW-13 [5] and POPE [6] datasets. For a fair comparison, we set baseline iterations $S = 300$ and perturbation budget $\epsilon = 8/255$, while configuring PA-Attack with $S_1 = 100, S_2 = 200$. As shown in Table 2, our PA-Attack consistently demonstrates superior effectiveness across diverse LVLMs and tasks.

Table 2. Attack performance on Qwen3-VL and InternVL2.

Method	Qwen3-VL-8B		InternVL2-8B		SRR ↑
	RealWorldQA	ODinW-13	RealWorldQA	POPE	
Clean	73.3	35.5	62.5	88.6	0.0
VT-Attack	68.2	33.9	45.4	77.4	12.9
VEAttack	58.2	14.4	46.4	**63.8**	33.4
PA-Attack	**57.0**	**9.8**	**43.4**	64.8	**38.0**

3. Larger perturbation budget

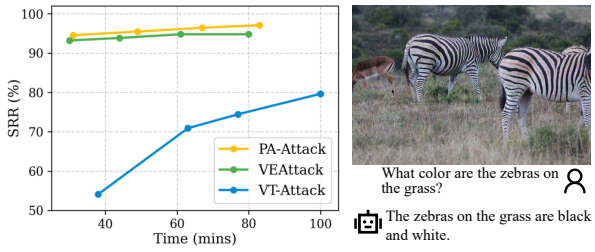
While standard gray-box attacks typically operate under strict imperceptibility constraints (e.g., $\epsilon = 2/255$ or $4/255$), it is crucial to evaluate attack scalability when these constraints are relaxed. We conducted additional experiments on LLaVA-1.5-7B with $\epsilon = 8/255$. Results in Table 3 demonstrate that PA-Attack maintains its superiority and scalability even under larger perturbation allowances.

Table 3. Attack performance with larger perturbation budget.

Method	COCO	Flickr30k	TextVQA	VQAv2	SRR ↑
Clean	115.5	77.5	37.1	74.5	0.0
VT-Attack	14.5	10.5	9.3	**25.0**	78.8
VEAttack	5.3	4.0	6.5	36.3	81.0
PA-Attack	**3.6**	**1.9**	**4.4**	31.9	**84.9**

4. Runtime comparison.

We analyze the trade-off between attack effectiveness (SRR) and computational cost (runtime) in Figure 1 (a). Results indicate that PA-Attack achieves superior SRR with only a marginal computational increase compared to VEAttack. Although our method introduces additional steps for prototype guidance and attention weight calculation, these operations are computationally lightweight compared to the gradient backpropagation through the vision encoder.



(a) Performance and efficiency of different methods. (b) Example of attacking failure cases.

Figure 1. Ablation of time and an example of failure cases.

5. Failure cases

To understand the boundaries of our attack, we analyze specific failure cases. As illustrated in Figure 1 (b), attacks tend to be less effective when the textual query triggers strong inherent knowledge priors within the Large Language Model (LLM). In such instances, the LVLM relies more heavily on its pre-trained linguistic knowledge than on the visual context provided by the vision encoder. For example, when asked about the color of a zebra, the LLM's strong association between "zebra" and "black and white" may override the perturbed visual embeddings. This suggests that while PA-Attack effectively disrupts visual representations, downstream hallucinations or strong language priors can sometimes bypass vision-targeted perturbations.

6. Effectiveness of Adversarial Training

To further validate the robustness of our method, we evaluate PA-Attack against state-of-the-art vision encoder-specific adversarial training (AT) defenses, specifically TeCoA [8] and FARE [9]. The results in Table 4 show that while the performance of baseline methods like VEAttack degrades sharply under these defenses, PA-Attack consistently maintains a higher Score Reduction Rate (SRR). This resilience indicates that the adversarial features generated by our prototype guidance and attention refinement mechanisms are semantically more robust and harder to mitigate than simple gradient-based noise. PA-Attack proves to be a more challenging threat model for current defense strategies.

Table 4. Attack performance with different defense methods.

Method	COCO		Flickr30k		TextVQA	
	TeCoA	FARE	TeCoA	FARE	TeCoA	FARE
VEAttack	43.9	51.1	24.2	34.1	16.5	22.2
PA-Attack	**25.8**	**47.5**	**15.1**	**30.7**	**10.6**	**14.1**

7. More visualizations of predictions

We provide qualitative comparisons in Fig. 4 of the paper and Fig. 2, 3, and 4 of the supplementary material to illustrate the different impacts of gray-box AttackVLM-ii, black-box M-Attack, and our gray-box PA-Attack. From a visual perspective, the perturbations generated by the black-box M-Attack are visibly more pronounced, while the gray-box methods generate much subtler and less perceptible noise, remaining visually closer to the original clean images. Moreover, the output-level comparisons highlight the task generalization of our method. Both the gray-box AttackVLM-ii and the black-box M-Attack, which lack prototype-anchored guidance and attention enhancement, often fail to fully attack core attributes in the image captioning task. This partial success leads to a cascading failure, as the VQA and POPE tasks that query these specific, unchanged attributes are not successfully attacked. For instance, in Fig. 2, both AttackVLM-ii and M-Attack still identify "a man is walking," which causes them to correctly answer the related VQA ("Is there someone crossing the street?") and POPE ("Is there a man in the street?") constructions. In contrast, our PA-Attack completely changes all attributes, which could lead to more general attack successes in other tasks. In Fig. 2, the subject is converted to "a clock", successfully deceiving all three tasks. These visualizations demonstrate that by leveraging prototype-anchored guidance and attention enhancement, PA-Attack achieves a more thorough and general attack by successfully altering core semantic attributes.

8. More visualizations of attentions

8.1. Attention maps

In Fig. 5, we provide a qualitative comparison by visualizing the attention maps of the vision encoder. We compare the attention on clean images against that on images perturbed by the AttackVLM-ii and both stages of our PA-Attack. In the "Clean" column, it can be seen that when processing clean images, the vision encoder distributes its focus across both the primary subject and the surrounding background, which allows it to extract comprehensive features necessary for downstream tasks. However, the attention maps for images in the second column attacked by AttackVLM-ii appear to force focus onto specific, high-vulnerability regions. For example, in the third row, it focuses on the plate itself, rather than the objects on it. PA-Attack demonstrates that the attention remains more generalized and semantically meaningful. In the third row, our method maintains a broader focus on both the plate and the diverse food items it contains. In addition, the attention enhancement module directs focus toward critical information. In the second row, PA-Attack achieves a much more comprehensive and complete attention coverage over the main subject. Finally, PA-Attack-Stage 1 in the third column and PA-Attack-Stage 2 in the fourth column reveal a shift in focus. This dynamic adjustment validates the effectiveness of our two-stage refinement, which could update and optimize the adversarial focus.

8.2. Difference of attentions with iterations

In Fig. 6, we visualize the evolution of token feature deviations during the optimization process on three datasets. The heatmaps illustrate the difference between token features at each attack iteration and the original clean features across all token indexes. Vision Encoder Attack consistently concentrates its perturbations on a few specific token indices throughout the attack. After introducing Prototype-anchored Guidance, influence is spread across a much broader range of tokens, indicating that the guidance prevents the optimization from collapsing onto a few specific features and promotes more attributes. Attention Enhancement adds a few more distinct vertical bands, representing its focus on specific important tokens. This reveals a key trade-off that prototype-anchored guidance succeeds in broadening the attack's scope but may reduce the relative focus on the most critical tokens. Our PA-Attack inherits the widespread, generalized deviation pattern from the prototype guidance, but it also exhibits stronger, more pronounced perturbations on the relatively important tokens identified by the attention enhancement module. This demonstrates that our method successfully achieves a trade-off, generalizing the attack vector while simultaneously prioritizing the most impactful tokens for optimization.

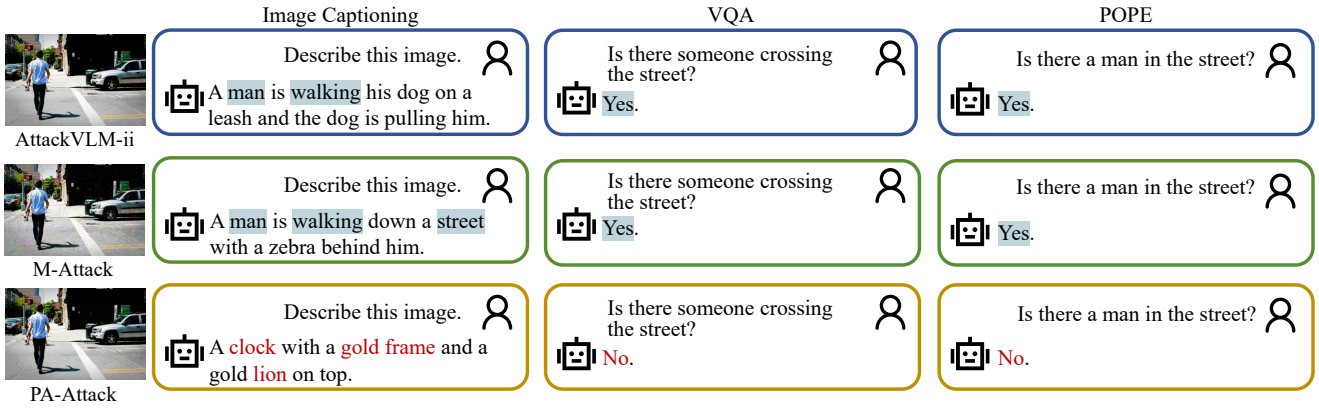


Figure 2. **More comparison of the responses of LLaVa1.5-7B with different attacks.** The attributes with a blue background remain unchanged, while the red texts indicate that the attributes have changed.

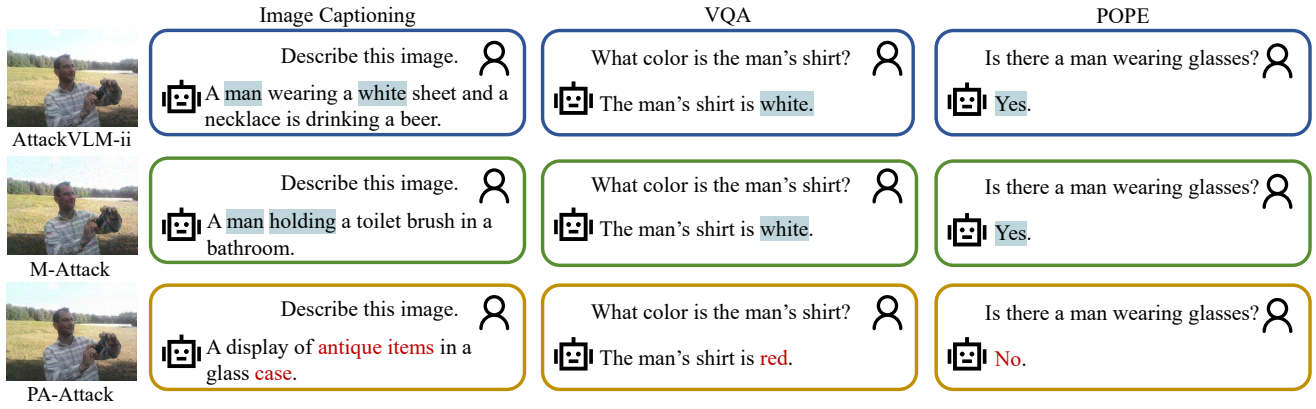


Figure 3. **More comparison of the responses of LLaVa1.5-7B with different attacks.** The attributes with a blue background remain unchanged, while the red texts indicate that the attributes have changed.

References

[1] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025. 1

[2] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024. 1

[3] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024. 1

[4] Adam W Harley, Alex Ufkes, and Konstantinos G Derpanis. Evaluation of deep convolutional nets for document image classification and retrieval. In *International Conference on Document Analysis and Recognition (ICDAR)*, 2015. 1

[5] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975, 2022. 1

[6] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023. 1

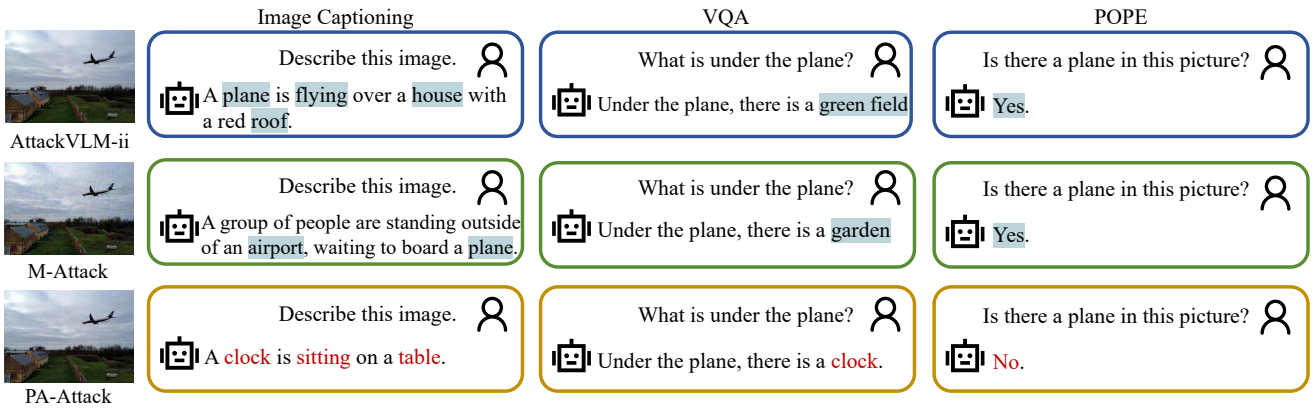


Figure 4. **More comparison of the responses of LLaVa1.5-7B with different attacks.** The attributes with a blue background remain unchanged, while the red texts indicate that the attributes have changed.

[7] Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*, 2022. 1

[8] Chengzhi Mao, Scott Geng, Junfeng Yang, Xin Wang, and Carl Vondrick. Understanding zero-shot adversarial robustness for large-scale models. In *International conference on learning representations*, 2023. 2

[9] Christian Schlarmann, Naman Deep Singh, Francesco Croce, and Matthias Hein. Robust clip: Unsupervised adversarial fine-tuning of vision embeddings for robust large vision-language models. In *International conference on machine learning*, 2024. 2

[10] xAI. Realworldqa: A benchmark for real-world spatial understanding. https://huggingface.co/datasets/xai-org/RealworldQA, 2024. Accessed: 2025-04-26. 1

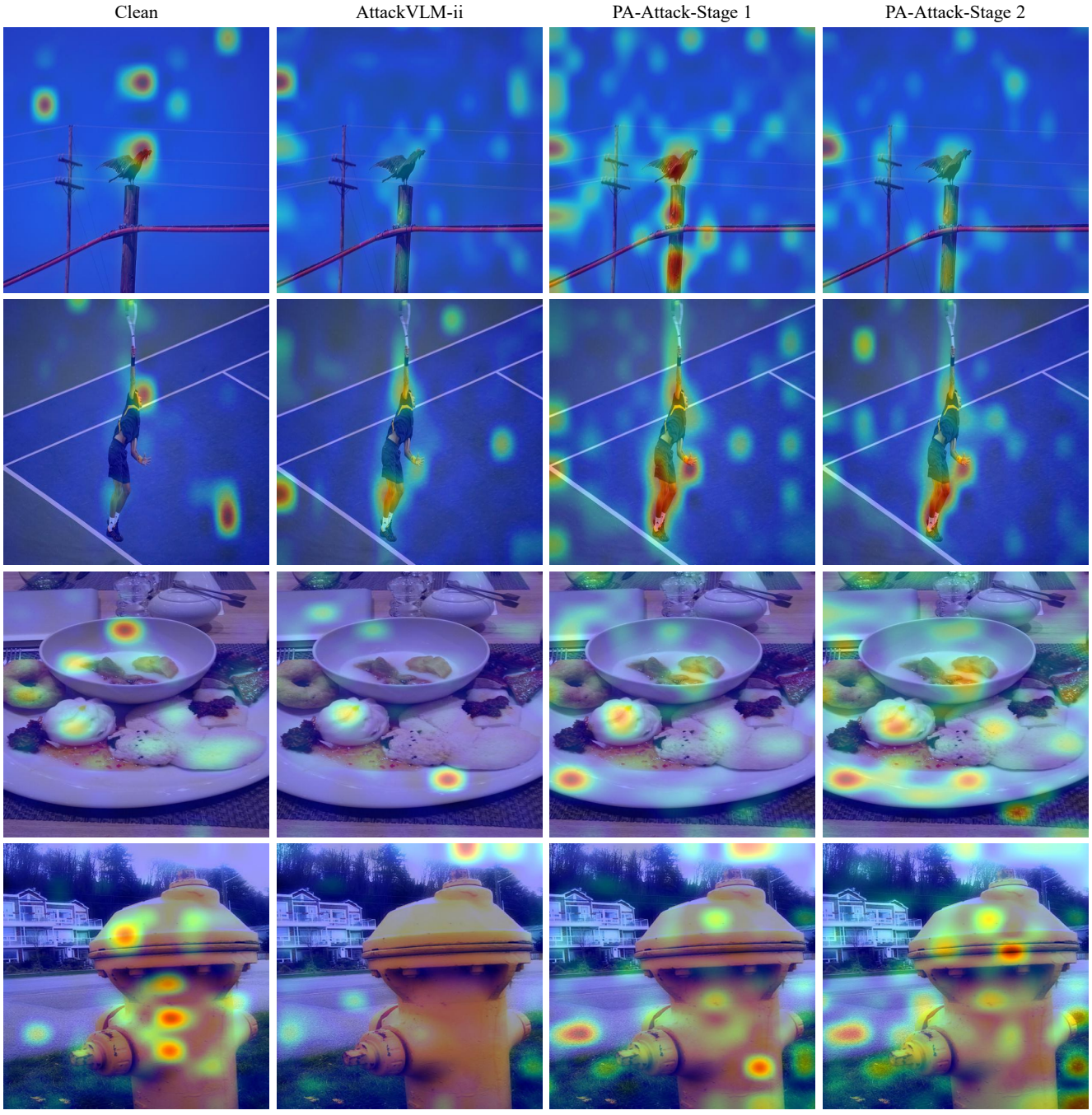


Figure 5. **Comparison of the attention maps of clean images and those after different attacks.** The iterations of AttackVLM-ii are 100. We visualize the attention map of the first stage of our PA-Attack with 50 iterations and the second stage with 100 iterations. The areas that are more reddish have higher attention values.

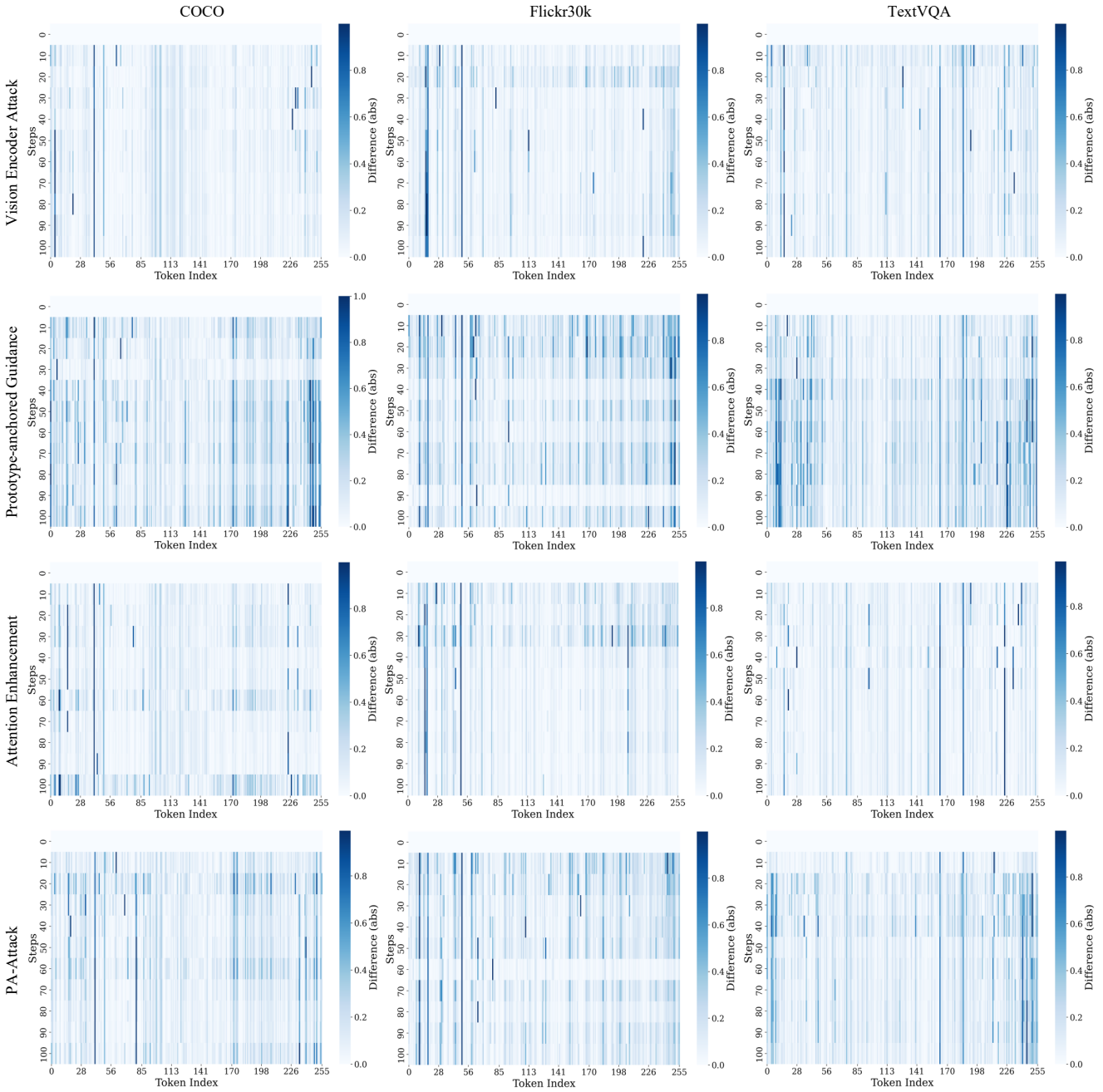


Figure 6. **Evolution of token-wise attention values during the attack process.** The heatmaps visualize the deviation in token attention compared to the token of clean images with step 0. The x-axis represents token indices, and the y-axis represents the optimization steps. A darker shade indicates a greater difference in attention value from the clean image. The rows correspond to different modules, and the columns show results for three datasets.