Transfer-Based Adversarial Attack Against Multimodal Models by Exploiting Perturbed Attention Region

Raffaele Disabato*, AprilPyone MaungMaung†, Huy H. Nguyen† and Isao Echizen†‡
* University of Bologna, Bologna, Italy
† National Institute of Informatics, Tokyo, Japan
‡ University of Tokyo, Tokyo, Japan

***Abstract*—Multimodal models such as GPT-4 and Claude 3 have shown remarkable performance with vision capabilities, enabling exciting new applications with multimodal interactions. With new opportunities, new security concerns come along. Previous studies showed that attackers can force multimodal models to generate unwanted output by manipulating image modality. However, images manipulated by previous attacks are noisy and do not work across different multimodal models. To this end, we propose a new adversarial attack against multimodal models that is stealthy and effective in attacking multiple models. Specifically, we reinforce the previous common weakness attack with multiple surrogate vision transformers to attain strength and utilize an attention mask to focus on essential areas in the image. By doing so, the proposed attack does not add noise to the whole area of the image while maintaining attack strength. Experiment results show that the proposed attack is stealthy (noise patterns in manipulated images by the proposed method are invisible) and can attack two open-source multimodal models, LLaVA 1.5 and MiniGPT-4, in gray-box settings assuming surrogate models are similar to the ones used by targeted victim multimodal models.**

I. Introduction

In recent years, vision-language models (VLMs) have garnered enormous attention and acclaim for their remarkable versatility across a spectrum of tasks, ranging from Visual Question Answering (VQA) [1], [2] to image captioning [3], [4] and beyond [5], [6]. This surge in popularity stems from their ability to seamlessly integrate information from diverse modalities, offering unprecedented performance in understanding and generating content. However, amidst the fervor surrounding their capabilities, a pressing concern emerges, that is, the vulnerability of these models to adversarial attacks.

In this paper, we delve into a concerning scenario where a user encounters an image online (*e.g.*, on social media) that has been manipulated by a malicious actor and unknowingly interacts with a VLM using the image. For example, a user can ask a VLM about a food recipe by showing a food image, and the VLM may output harmful/toxic content. The manipulated image is a carefully crafted adversarial example where imperceptible noise is injected to mislead VLMs like MiniGPT-4 [7] or LLaVA [8]. As the output of VLMs is compromised, VLMs may potentially generate harmful content that will lead to spreading misinformation or amplifying toxic narratives. Consequently, safeguarding the integrity and trustworthiness of VLMs against such threats is imperative, not only to preserve the reliability of VLMs but to also uphold the societal well-being reliant upon their outputs.

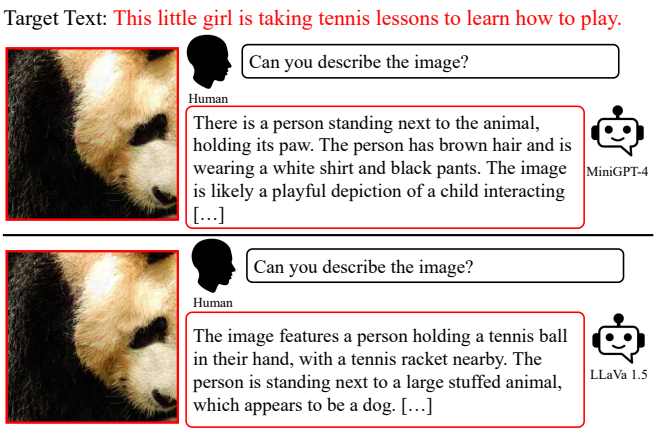


Fig. 1. Example of adversarial example by the proposed attack that reflects target text on LLaVa 1.5 and MiniGPT-4. The adversarial example is generated with perturbation budget ϵ of $4/255$ under maximum norm ℓ_∞.

Towards the robustness of VLMs, researchers have proposed attacks such as AttackVLM [9] and the common weakness attack (CWA) [10]. These previous attacks add adversarial noise to the whole image. Therefore, the noise produced by these attacks becomes visible as the strength of the attacks is increased. In addition, AttackVLM utilizes only one surrogate model and works only on one multimodal model. To this end, we propose a new adversarial attack reinforcing CWA [10] by drawing inspiration from the methodologies outlined for AttackVLM [9]. An example of adversarial example by the proposed attack on different VLMs are shown in Fig. 1. The proposed attack leverages the attention flow of vision transformers used in VLMs to achieve stealthiness while preserving the efficacy of the attack. We make the following contributions in this paper.

- We propose a novel adversarial attack against VLMs that exploits perturbed attention regions reinforced by

AttackVLM [9] and CWA [10] to achieve strength and stealthiness. Consequently, the adversarial examples that the proposed attack produces can fool multiple VLMs (MiniGPT-4 [7] and LLaVA 1.5 [8]) under gray-box settings.

- We conduct experiments to verify the effectiveness of the proposed attack.

This paper aims to elucidate the inherent vulnerabilities shared across multiple VLMs by synthesizing insights from diverse attack methodologies, underscoring the urgency of fortifying their defenses against adversarial manipulation.

II. Related Work

In this section, we briefly review vision language models (VLMs) and existing adversarial attacks on them.

A. Vison-Language Models (VLMs)

Multimodal models are current artificial intelligence (AI) frontiers. VLMs are multimodal models that can process and understand textual and visual information [1], [7], [8], [11], [12]. Most of these models utilize a pre-trained large language model (LLM) and a large vision encoder such as CLIP. During training, the vision encoder remains fixed. Only a projection layer or cross-attention that bridges vision and language is trained. Among the many VLMs, we focus on LLaVA 1.5 [8] and MiniGPT-4 [7] for their remarkable performance and open-source nature to evaluate the proposed attack in this paper.

B. Adversarial Attacks on VLMs

Generally, adversarial examples are carefully perturbed input data (usually imperceptible) to machine learning models that cause the models to make erroneous predictions [13]–[15]. Adversarial example generation (adversarial attacks) can be categorized into two major settings, white-box and black-box, according to the knowledge of the targeted victim model available to the attacker. The white-box threat model assumes the attacker has full knowledge of the model, including its weights and training data. In contrast, the attacker does not know the model in black-box settings. When partial knowledge of the model is available, adversarial example generation is often categorized as gray-box instead of black-box. Depending on the attacker's goal, adversarial attacks that generate adversarial examples can also be classified as targeted or non-targeted attacks. Targeted attacks force the model to output the attacker's intended output, and non-targeted attacks cause the model to make erroneous predictions without specifications.

In the multimodal domain, researchers have also investigated adversarial examples against VLMs. Nonetheless, compared with adversarial examples in the unimodal domain, adversarial examples against multimodal models are less explored. An early work showed that multimodal attacks are stronger than unimodal attacks in gray-box settings and demonstrated the attacks against visual question answering (VQA) and image captioning tasks [16]. Another work proposed an adversarial attack that forces a generative model to produce toxic and harmful outputs [17]. Similar to adversarial examples for image classification models, this study demonstrated that adversarial examples can be crafted against the OpenFlamingo model [18]. More recent works include AttackVLM [9] that introduces attacking strategies to VLMs, AttackBard [19] that investigates the adversarial robustness of Google's Bard, a multimodal language model with image processing capabilities, and the common weakness attack (CWA) [10], which is a powerful transfer-based attack for various models.

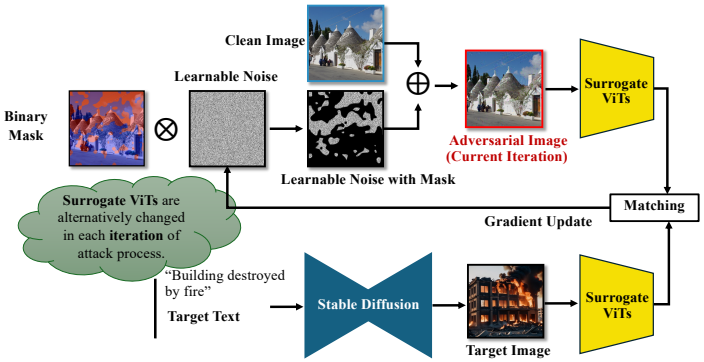


Fig. 2. Overview of proposed method.

In this paper, we focus on gray-box adversarial attacks assuming that surrogate models are highly similar to those used in victim multimodal models. Such gray-box settings are practical in real-world scenarios because open-source foundation models are publicly available online. Specifically, we focus on a transfer-based gray-box attack similar to CWA [10] against two victim VLMs: LLaVA 1.5 [8] and MiniGPT-4 [7].

III. Method

The proposed attack is a targeted attack against VLMs. Given a clean image $\mathbf{x}_{\text{cln}} \in [0,1]^{c\times h\times w}$ with channel c, height h, and width w and a targeted text P, our goal is to create an adversarial example $\mathbf{x}$ that forces a VLM p_θ parameterized by θ that takes $\mathbf{x}$ and a query q to output a response that reflects the targeted text P (*i.e.*, $p_\theta(\mathbf{x}, q) \approx P$). For example, q may be "*describe the image*" or "*what is unusual about the image*" or any content-related question, and p_θ outputs the response related to P instead of the actual content of $\mathbf{x}_{\text{cln}}$. To achieve this, by taking inspiration from AttackVLM [9], we first convert the targeted text P to a targeted image $\mathbf{x}_{\text{tgt}}$ by utilizing existing text-to-image models such as Stable Diffusion [20]. Then, we match the features of $\mathbf{x}_{\text{tgt}}$ with those of $\mathbf{x}$ as in AttackVLM, leveraging a common weakness attack (CWA) with an attention mask for stealthiness. Fig. 2 depicts an overview of the proposed attack.

There are three major components in the proposed attack: image-image feature matching from AttackVLM [9], the common weakness attack (CWA) [10], and targeting important regions in an image with an attention mask. Next, we provide some background on these components and present the proposed attack.

A. Image-Image Feature Matching

Inspired by previous works, the authors of the paper on AttackVLM [9] pointed out that VLMs might not be reliable for optimizing similarity across different modalities. Therefore, they introduced matching image-image similarity (same modalities) by utilizing a public text-to-image model (e.g., Stable Diffusion [20]) to generate a targeted image first. Then, they optimized the following objective,

$$\arg\max_{\mathbf{x}} \; s(\mathbf{x})^\top s(\mathbf{x}_{\text{tgt}}) \text{ s.t. } \|\mathbf{x}_{\text{cln}} - \mathbf{x}\|_\infty \leq \epsilon, \qquad (1)$$

where $s(\cdot)$ is an image encoder (surrogate model). By doing so, VLMs misinterpret adversarial images generated by AttackVLM as the targeted text. We adopt this objective as the loss function in the proposed attack.

B. Common Weakness Attack

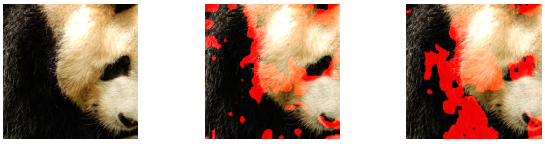
Informed by the previous work [21], the authors for the common weakness attack (CWA) [10] indicate that optimizing adversarial examples adheres to Empirical Risk Minimization (ERM) principles, and a restricted number of training models could result in significant generalization errors. Therefore, they proposed optimizing the loss landscape's flatness and the closeness between the local optima of different models to generate more transferable adversarial examples. CWA also incorporates the momentum iterative (MI) [22] attack and introduces MI-CWA to improve the attack success rate further. We base the proposed attack on the MI-CWA with two major modifications: (1) a different loss function (image-image feature matching) for targeting VLMs and (2) an attention mask for focusing on important regions only to reduce the perturbation.

C. Attention Masks

Although CWA [10] is an effective state-of-the-art transferrable adversarial attack, it perturbs all pixels in an image, resulting in visible noise artifacts. Therefore, we consider focusing only on important regions of an image (*i.e.*, attention) to reduce the impact of the noise added to adversarial images. We explore two attention masks: rollout attention mask (RAM) [23] and layer-wise relevance propagation attention mask (LRP-AM) [24]. An example of the different attention masks is shown in Fig. 3.

RAM [23]. Rollout attention was originally proposed to quantify the attention flow in the Transformer for the natural language processing domain. However, this technique was also used for vision transformers (ViTs) to visualize attention in images. Therefore, we utilize this attention mask to focus on important regions of an image in the proposed attack. Specifically, we use the image encoder components of ViT-L/336 [25] and EVA-CLIP-giant/224 [26] to generate attention masks for clean images. Then, we merge all masks from all surrogate models by using the OR logic operation.

LRP-AM [24]. LRP is an explainability technique applicable to deep neural networks. It functions by propagating prediction backward through a neural network, using a predefined set of propagation rules designed for this purpose. As LRP provides pixel-level relevance and highlights important regions that influence VLMs, the mask produced by LRP (LRP-AM) offers superior performance in our experiments. Unlike RAM, we use ViT-L/336 [25] only to analyze each pixel's relevance in clean images directly.



(a) Clean (b) LRP-AM (c) RAM

Fig. 3. Attention mask visualization. LRP-AM shows less noisy attention.

D. Proposed Attack

We build the proposed attack on top of the MI-CWA [10]. Algorithm 1 details the procedure of the proposed attack. First, we need to generate a targeted image $\mathbf{x}_{\text{tgt}}$ from a given targeted text P using existing public text-to-image models. We utilized Stable Diffusion [20] in our experiments. Since we are attacking VLMs, we substitute the loss function in MI-CWA with image-image feature matching loss (Eq. 1), as described in Section III-A. We generate a binary mask based on the attention mask described in Section III-C to reduce the noise added to adversarial examples. We utilize two surrogate models in the proposed attack (*i.e.*, $\mathcal{S} = \{$ViT-L/336 [25], EVA-CLIP-giant/224 [26]$\}$). Given these two surrogate models with attention masks, we run momentum-based optimization as in MI-CWA [10]. The proposed attack can generate stealthier adversarial examples against LLaVA 1.5 [8] and MiniGPT-4 [7] in transfer-based gray-box settings.

IV. Experiments

A. Settings

Datasets. We used two well-known datasets: a development dataset used in a competition on adversarial attacks and defense for NeurIPS 2017 (hereafter referred to as NIPS17) [27] and ImageNet [28]. NIPS17 contains 1000 images that are labelled with ImageNet labels. We utilized 200 images from NIPS17. ImageNet contains 1.28 million images, and we used 1000 images from its validation set to represent a broad range of visual content. We randomly sampled MS-COCO captions [29] for targeted text descriptions, ensuring a diverse set of image concepts encompassing a wide range of objects, scenes, and activities.

Target Image Generation. We generated targeted images from targeted texts by using Stable Diffusion [20] 2.1. We used Pseudo Linear Multi-Step (PLMS) sampling for 50 steps with a classifier-free guidance scale value of 7.5. The targeted images had a dimension of 512×512.

Victim and Surrogate Models. Our experiments considered two open-source VLMs, LLaVA 1.5 [8] and MiniGPT-4 [7],

Algorithm 1 Modified MI-CWA algorithm (Proposed)

Require: clean image $\mathbf{x}_{\text{cln}}$, targeted image $\mathbf{x}_{\text{tgt}}$, perturbation budget ϵ, iterations T, loss function L, surrogate models $\mathcal{S} = \{s_i\}_{i=1}^{n}$, step sizes β and α

Initialize $\boldsymbol{m} \leftarrow 0$, inner momentum $\hat{\boldsymbol{m}} \leftarrow 0$, $\mathbf{x}_0 \leftarrow \mathbf{x}_{\text{cln}}$
Generate a binary mask bm
for all $i \leftarrow 1, \ldots, n$ **do**
  Calculate target features $F_{\text{tgt}}^{i} \leftarrow s_i(\mathbf{x}_{\text{tgt}})$
end for
for all $t \leftarrow 0, \ldots, T-1$ **do**
  $\mathbf{x}_t^0 \leftarrow \mathbf{x}_0$
  for all $i \leftarrow 1, \ldots, n$ **do**
    Calculate adversarial features $F_{\text{adv}}^{i-1} \leftarrow s_i(\mathbf{x}_t^{i-1})$
    Calculate $\boldsymbol{g} \leftarrow \nabla_{\boldsymbol{x}} L(F_{\text{tgt}}^{i-1}, F_{\text{adv}}^{i-1})$
    Update inner momentum by $\hat{\boldsymbol{m}} \leftarrow \mu \cdot \hat{\boldsymbol{m}} + \frac{\boldsymbol{g}}{\|\boldsymbol{g}\|_2}$
    Update $\boldsymbol{x}_t^i$ by $\boldsymbol{x}_t^i \leftarrow \text{clip}_{\boldsymbol{x}_{cln},\epsilon}(\boldsymbol{x}_t^{i-1} + \beta \cdot \boldsymbol{m} \cdot bm)$
  end for
  Calculate update $\boldsymbol{g} \leftarrow \boldsymbol{x}_t^n - \boldsymbol{x}_t$
  Update momentum $\boldsymbol{m} \leftarrow \mu \cdot \boldsymbol{m} + \boldsymbol{g}$
  Update $\boldsymbol{x}_{t+1}$ by $\boldsymbol{x}_{t+1} \leftarrow \text{clip}_{\boldsymbol{x}_{cln},\epsilon}(\boldsymbol{x}_t + \alpha \cdot \text{sign}(\boldsymbol{m}))$
end for
return $\boldsymbol{x}_T$

as victims for evaluation. We used two surrogate models, the image encoder components of ViT-L/336 [25] and EVA-CLIP-giant/224 [26], to generate attention masks and extract image embeddings in the proposed attack.

Evaluation Method. Sentence T5 Large [30], a state-of-the-art Sentence Transformer model, was chosen to evaluate the semantic content of the captions. This selection leverages Sentence T5 Large's ability to handle full-length descriptions, capturing the nuances of the semantic relationships compared with traditional text encoders with truncation limitations.

Implementation Details. We implemented the proposed attack on top of the publicly available MI-CWA code[1]. We utilized feature matching loss from the AttackVLM official code repository[2] and attention masks from the public code base[3]. The parameters were an inner step size value of 250 and an outer step size value of 2, and we ran the attack for 500 steps. The perturbation budget ϵ had a value of $4/255$ under maximum norm ℓ_∞. For comparison, we ran AttackVLM image-image feature matching (FM-ii) [9] with a step size of 1 and a perturbation budget ϵ of $4/255$ under ℓ_∞. We used the query *"Describe this image"* when querying VLMs for both clean and adversarial examples.

B. Results

To verify the effectiveness of the proposed attack, we calculated the Sentence BERT score between the targeted texts and generated texts mentioned in Section IV-A to measure the attack strength. In addition, to assess visual quality while maintaining attack strength, we utilized three metrics: the structural similarity index measure (SSIM), L2 distance, and a state-of-the-art perceptual quality metric, TOPIQ [31]. We also compared the performance of the proposed attack with two state-of-the-art attacks against VLMs: CWA [10] and AttackVLM (FM-ii) [9].

[1]https://github.com/huanranchen/AdversarialAttacks/
[2]https://github.com/yunqing-me/AttackVLM
[3]https://github.com/jacobgil/vit-explain

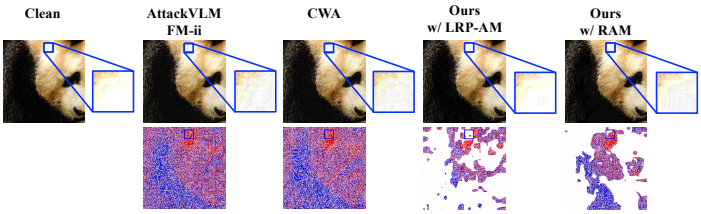


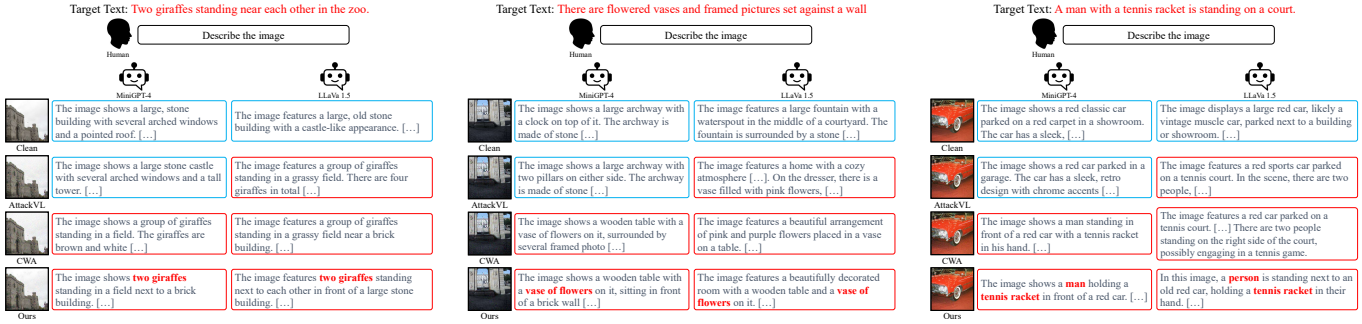
Fig. 4. Examples of adversarial examples generated by AttackVLM [9], CWA [10], and proposed attack with their corresponding noise maps.

TABLE I
SENTENCE BERT SCORE (↑) BETWEEN TARGETED TEXTS AND VICTIM MODEL GENERATED TEXTS. BEST RESULTS ARE IN BOLD, AND SECOND BEST RESULTS ARE UNDERLINED.

	NIPS17 (200 images)		ImageNet (1000 images)	
Attack	LLaVA 1.5	MiniGPT-4	LLaVA 1.5	MiniGPT-4
Clean	0.66	0.67	0.66	0.67
CWA [10]	<u>0.79</u>	**0.78**	**0.79**	**0.78**
AttackVLM (FM-ii) [9]	**0.80**	0.67	**0.79**	0.67
Ours (w/ RAM)	0.73	0.72	0.72	0.71
Ours (w/ LRP-AM)	0.75	<u>0.75</u>	<u>0.74</u>	<u>0.74</u>

Sentence BERT Score. Table I summarizes the Sentence BERT score for the two datasets: NIPS17 and ImageNet (random 1000 validation images). The generated adversarial examples were tested against the two VLMs: LLaVA 1.5 [8] and MiniGPT-4 [7]. AttackVLM achieved the highest score for LLaVA 1.5, but it did not work on MiniGPT-4. The reason is that AttackVLM is designed to use one surrogate model. In contrast, CWA and the proposed attack utilize multiple surrogate models. Consequently, CWA and the proposed attack could be applied to both VLMs. Although the proposed attack did not achieve the highest score due to noise reduction, the attack performance was comparable. Fig. 5 shows the output of adversarial examples by victim VLMs.

Visual Quality. Table II presents visual quality results for different metrics in comparison with the state-of-the-art methods. As expected, the proposed method achieved the highest visual quality performance across different metrics. The attention masks used in the proposed attack effectively reduced noise artifacts. Fig. 4 shows adversarial examples with their corresponding noise maps by different methods. From the figure, adversarial examples by the proposed attack contained less noise, which was confirmed in an objective visual quality evaluation. It is worth noting that the TOPIQ score for ImageNet (1000 images) was lower than that of NIPS17. One possibility is that the size of ImageNet images varies, and we did not consider the aspect ratio while resizing the images, which might have affected the perceptual visual quality.



(a) Example 1 (b) Example 2 (c) Example 3

Fig. 5. Responses of LLaVa 1.5 and MiniGPT-4 on clean images and adversarial examples by the proposed method, AttackVLM [9], and CWA [10].

TABLE II
VISUAL QUALITY EVALUATION OF GENERATED ADVERSARIAL EXAMPLES. BEST RESULTS ARE IN BOLD, AND SECOND BEST RESULTS ARE UNDERLINED.

	NIPS17 (200 images)			ImageNet (1000 images)		
Attack	SSIM (↑)	L2 (↓)	TOPIQ (↑)	SSIM (↑)	L2 (↓)	TOPIQ (↑)
CWA [10]	0.79	17.22	0.65	0.79	19.29	0.35
AttackVLM (FM-ii) [9]	0.58	19.15	0.66	0.77	22.12	0.36
Ours (w/ RAM)	**0.93**	**9.5**	<u>0.71</u>	<u>0.91</u>	**10.39**	**0.38**
Ours (w/ LRP-AM)	**0.93**	<u>10.47</u>	**0.76**	**0.92**	<u>11.81</u>	<u>0.37</u>

V. DISCUSSION

We demonstrated a transfer-based adversarial attack against VLMs with comparable attack strength and less noise. Such stealthier adversarial examples may be proliferated and distributed online by malicious actors. Honest users may feed these adversarial examples to VLMs such as LLaVA 1.5 or MiniGPT-4. As a result, VLMs may be compromised, potentially generating harmful content, spreading misinformation, or amplifying toxic narratives.

Responsible Artificial Intelligence (AI) Development. As VLMs are still under development, paying attention to safety early in development is indispensable. The ability to generate high-quality adversarial examples emphasizes the importance of responsible AI development. VLMs are increasingly integrated into various applications, so it is crucial to consider the potential for misuse. Our work highlights the need for robust security measures and responsible development practices to ensure that VLMs are not weaponized to spread misinformation or manipulate public discourse.

Limitations. While the proposed attack can generate finer adversarial examples against VLMs, it still has limitations. This paper evaluated the proposed attack against LLaVA 1.5 and MiniGPT-4. We shall investigate other VLMs for further evaluation. In addition, the proposed attack in its current form only deploys two surrogate models, so we shall further investigate and explore more surrogate models to improve transferability in our future work. We shall also conduct an in-depth analysis of the relationship between surrogate models and VLMs.

VI. CONCLUSION

This paper proposes a transfer-based adversarial attack against vision-language models (VLMs) using multiple surrogate models. Unlike the previous methods, the proposed attack utilizes attention masks to reduce the noise added to adversarial examples. Experiments confirmed that the proposed attack is comparable to the previous attack methods in terms of attack strength and is superior in terms of visual quality. We hope that our work encourages model developers towards safe AI system development.

ACKNOWLEDGMENT

This work was partially supported by JSPS KAKENHI Grants JP21H04907, 23K19983, and JP24H00732, by JST CREST Grants JPMJCR18A6 and JPMJCR20D3 including AIP challenge program, by JST AIP Acceleration Grant JPMJCR24U3, by JST K Program Grant JPMJKP24C2 Japan, and by the project for the development and demonstration of countermeasures against disinformation and misinformation on the Internet with the Ministry of Internal Affairs and Communications of Japan.

REFERENCES

[1] J.-B. Alayrac, J. Donahue, P. Luc, *et al.*, "Flamingo: A visual language model for few-shot learning," *Advances in neural information processing systems*, vol. 35, pp. 23 716–23 736, 2022.

[2] J. Chen, H. Guo, K. Yi, B. Li, and M. Elhoseiny, "Visualgpt: Data-efficient adaptation of pretrained language models for image captioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 030–18 040.

[3] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *International conference on machine learning*, PMLR, 2023, pp. 19 730–19 742.

[4] M. Tsimpoukelli, J. L. Menick, S. Cabi, S. Eslami, O. Vinyals, and F. Hill, "Multimodal few-shot learning with frozen language models," *Advances in Neural Information Processing Systems*, vol. 34, pp. 200–212, 2021.

[5] F. Bao, S. Nie, K. Xue, *et al.*, "One transformer fits all distributions in multi-modal diffusion at scale," in *International Conference on Machine Learning*, PMLR, 2023, pp. 1692–1717.

[6] A. Q. Nichol, P. Dhariwal, A. Ramesh, *et al.*, "GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models," in *Proceedings of the 39th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 162, PMLR, 17–23 Jul 2022, pp. 16 784–16 804.

[7] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny, "Minigpt-4: Enhancing vision-language understanding with advanced large language models," *arXiv preprint arXiv:2304.10592*, 2023.

[8] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *Advances in neural information processing systems*, vol. 36, 2024.

[9] Y. Zhao, T. Pang, C. Du, *et al.*, "On evaluating adversarial robustness of large vision-language models," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[10] H. Chen, Y. Zhang, Y. Dong, X. Yang, H. Su, and J. Zhu, "Rethinking model ensemble in transfer-based adversarial attacks," *arXiv preprint arXiv:2303.09105*, 2023.

[11] A. Awadalla, I. Gao, J. Gardner, *et al.*, "Openflamingo: An open-source framework for training large autoregressive vision-language models," *arXiv preprint arXiv:2308.01390*, 2023.

[12] J. Y. Koh, R. Salakhutdinov, and D. Fried, "Grounding language models to images for multimodal inputs and outputs," in *International Conference on Machine Learning*, PMLR, 2023, pp. 17 283–17 300.

[13] C. Szegedy, W. Zaremba, I. Sutskever, *et al.*, "Intriguing properties of neural networks," in *International Conference on Learning Representations*, 2014.

[14] B. Biggio, I. Corona, D. Maiorca, *et al.*, "Evasion attacks against machine learning at test time," in *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2013, Prague, Czech Republic, September 23-27, 2013, Proceedings, Part III 13*, Springer, 2013, pp. 387–402.

[15] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *International Conference on Learning Representations*, 2015.

[16] I. Evtimov, R. Howes, B. Dolhansky, H. Firooz, and C. C. Ferrer, "Adversarial evaluation of multimodal models under realistic gray box assumption," *arXiv preprint arXiv:2011.12902*, 2020.

[17] X. Qi, K. Huang, A. Panda, M. Wang, and P. Mittal, "Visual adversarial examples jailbreak large language models," *arXiv preprint arXiv:2306.13213*, 2023.

[18] C. Schlarmann and M. Hein, "On the adversarial robustness of multi-modal foundation models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 3677–3685.

[19] Y. Dong, H. Chen, J. Chen, *et al.*, "How robust is google's bard to adversarial image attacks?" *arXiv preprint arXiv:2309.11751*, 2023.

[20] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[21] H. Huang, Z. Chen, H. Chen, Y. Wang, and K. Zhang, "T-sea: Transfer-based self-ensemble attack on object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 20 514–20 523.

[22] Y. Dong, F. Liao, T. Pang, *et al.*, "Boosting adversarial attacks with momentum," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2018, pp. 9185–9193.

[23] S. Abnar and W. Zuidema, "Quantifying attention flow in transformers," *arXiv preprint arXiv:2005.00928*, 2020.

[24] G. Montavon, A. Binder, S. Lapuschkin, W. Samek, and K.-R. Müller, "Layer-wise relevance propagation: An overview," *Explainable AI: interpreting, explaining and visualizing deep learning*, pp. 193–209, 2019.

[25] A. Dosovitskiy, L. Beyer, A. Kolesnikov, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[26] Q. Sun, Y. Fang, L. Wu, X. Wang, and Y. Cao, "Eva-clip: Improved training techniques for clip at scale," *arXiv preprint arXiv:2303.15389*, 2023.

[27] A. Kurakin, I. Goodfellow, S. Bengio, *et al.*, "Adversarial attacks and defences competition," in *The NIPS'17 Competition: Building Intelligent Systems*, Springer, 2018, pp. 195–231.

[28] O. Russakovsky, J. Deng, H. Su, *et al.*, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.

[29] T.-Y. Lin, M. Maire, S. Belongie, *et al.*, "Microsoft coco: Common objects in context," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, Springer, 2014, pp. 740–755.

[30] J. Ni, G. H. Abrego, N. Constant, *et al.*, "Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models," *arXiv preprint arXiv:2108.08877*, 2021.

[31] C. Chen, J. Mo, J. Hou, *et al.*, "Topiq: A top-down approach from semantics to distortions for image quality assessment," *arXiv preprint arXiv:2308.03060*, 2023.