ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-6874-1/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICASSP49660.2025.10890065

Imperceptible Transfer Attack on Large Vision-Language Models

Xiaowen Cai
School of Cyber Science and Engineering
Huazhong University of Science and Technology
Wuhan, China
xwcai@hust.edu.cn

Daizong Liu
Wangxuan Institute of Computer Technology
Peking University
Beijing, China
dzliu@stu.pku.edu.cn

Runwei Guan
Department of Electrical Engineering and Electronics
University of Liverpool
Liverpool, United Kingdom
runwei.guan@liverpool.ac.uk

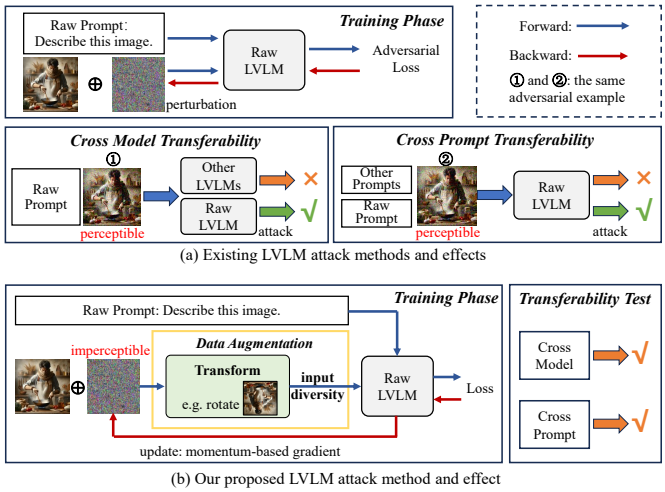
Pan Zhou*
School of Cyber Science and Engineering
Huazhong University of Science and Technology
Wuhan, China
panzhou@hust.edu.cn

***Abstract*—In spite of achieving significant progress in recent years, Large Vision-Language Models (LVLMs) are proven to be vulnerable to adversarial examples. Therefore, there is an urgent need for an effective adversarial attack to identify the deficiencies of LVLMs in security-sensitive applications. However, existing LVLM attackers generally optimize adversarial samples against a specific textual prompt with a certain LVLM model, tending to overfit the target prompt/network and hardly remain malicious once they are transferred to attack a different prompt/model. To this end, in this paper, we propose a novel Imperceptible Transfer Attack (ITA) against LVLMs to generate prompt/model-agnostic adversarial samples to enhance such adversarial transferability while further improving the imperceptibility. Specifically, we learn to apply appropriate visual transformations on image inputs to create diverse input patterns by selecting the optimal combination of operations from a pool of candidates, consequently improving adversarial transferability. We conceptualize the selection of optimal transformation combinations as an adversarial learning problem and employ a gradient approximation strategy with noise budget constraints to effectively generate imperceptible transferable samples. Extensive experiments on three LVLM models and two widely used datasets with three tasks demonstrate the superior performance of our ITA.**

***Index Terms*—Imperceptible transfer attack, LVLMs**

I. Introduction

Large Vision-Language Models (LVLMs) have achieved significant progress in multi-modal reasoning and understanding fields [1]–[19], due to the increase in the amount of data, computational resources, and number of model parameters. By further benefiting from the strong comprehension of large language models (LLMs), recent LVLMs [20]–[23] on top of LLMs show further superior performances in solving complex vision-language tasks by utilizing appropriate human-instructed prompts. However, existing studies [24]–[28] demonstrate that LVLMs are susceptible to adversarial examples, exposing LVLMs to various security threats and vulnerabilities in real-world applications. Therefore, there is an urgent need for an effective adversarial attack to identify the deficiencies of LVLMs in security-sensitive applications.



(a) Existing LVLM attack methods and effects

(b) Our proposed LVLM attack method and effect

Fig. 1. Illustation of our motivation. We propose to design the learnable transformation module with gradient approximation to improve both the transferability and imperceptibility of the LVLM attacks against different black-box models/prompts.

Existing LVLM attackers [28]–[43] generally follow the adversarial perturbing strategy to craft and add learnable perturbations to benign image inputs for fooling the LVLM models. In particular, these works carefully design the adversarial objective goals to optimize the perturbations via backpropagated gradients with specific prompt and LVLM. Although they can achieve significant attack performance in both targeted and untargeted settings, as shown in Fig.1 (a), they are easily limited by their perturbation-specific design that can solely produce adversarial examples to deceive a particular prompt and model within a singular process. That is, to compromise different prompts and LVLMs, they must generate distinct adversarial perturbations, which incur significant time and resource expenditure. Therefore, how to design

*Corresponding author.

Authorized licensed use limited to: BEIHANG UNIVERSITY. Downloaded on April 09,2025 at 14:02:44 UTC from IEEE Xplore. Restrictions apply.

an effective LVLM attack that is transferable among different prompts and models is an emergency issue.

To develop an LVLM attack with high transferability, we propose to improve the generalization ability of adversarial examples by creating diverse input patterns with data augmentations, as shown in Fig.1 (b). Our work is inspired by image transformations [44]–[48] that can defend against adversarial examples under certain situations, which indicates adversarial examples cannot generalize well under different transformations. These transformed adversarial examples are known as hard examples [49]–[51] for attackers, which can then serve as good samples to produce more transferable adversarial examples. Hence, we can learn a more generalizable LVLM attack by resisting it against harmful transformations. Once the generated adversarial examples are able to resist such distortions during adversarial learning, they are transferable to attack other LVLM models with high success rates.

Based on the above observations, in this paper, we introduce a novel Imperceptible Transfer Attack (ITA) to enhance both the transferability and imperceptibility of adversarial samples against LVLM models. Instead of following the perturbation-specific design like previous LVLM attackers, we propose to generate more generalizable adversarial samples by resisting them against distortions of visual transformations. Specifically, we dynamically learn and apply the optimal input transformation with the adversarial learning strategy to investigate the impacts of all possible transformation operations and better utilize these transformations to improve the input diversity. Subsequently, we mimic their impacts via gradient approximation and impose them as perturbations on the raw images. In this manner, our method effectively learns the dynamics of optimal transformations in attacks, leading to a significant enhancement in adversarial transferability. Additionally, we also design perturbation budget constraints during the optimization process to improve the imperceptibility. Our main contributions are three-fold:

- We make an attempt to improve the transferability of adversarial samples against LVLMs. Compared to previous perturbation-specific LVLM attacks, our proposed ITA can efficiently and effectively attack different models and prompts in a single process.
- We utilize learnable visual transformation to improve the diversity of the same input. With further adversarial learning strategy, we utilize the gradient estimation of harmful transformations to impose their impacts to generate more generalizable adversarial samples.
- Experiments are conducted on three LVLM models and two widely used datasets, indicating the superior transfer-attack performance of our proposed ITA.

II. Methodology

A. Task Definition

Generally, an LVLM model F receives an image x_v and a prompt x_p, then returns a ground-truth answer y. In this paper, we mainly focus on untargeted adversarial attacks. Untargeted attack aims to craft the adversarial image x_v^{adv} that misguides the predicted answer from the ground-truth label y to the arbitrary answer. So the optimization goal is to maximize the loss as:

$$max\ J(F(x_v^{adv}, x_p), y),\ s.t. \|x_v^{adv} - x_v\|_\infty \leq \epsilon, \quad (1)$$

where $J(\cdot)$ is the loss function, and we use l_∞-norm to regularize the adversarial perturbation to the range ϵ.

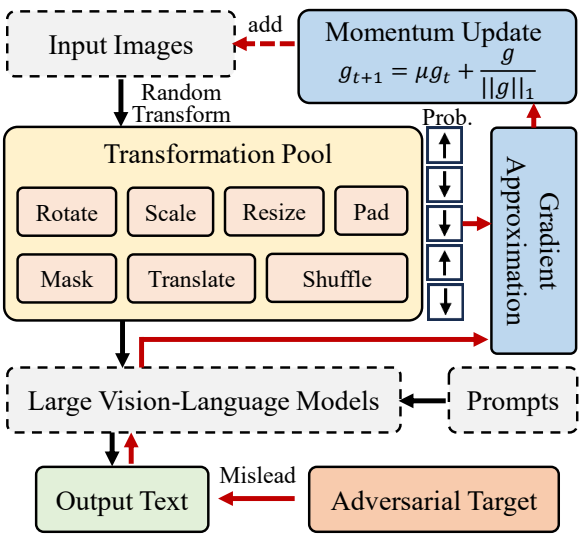


Fig. 2. Overview of our proposed imperceptible transfer attack against LVLM models. Our method learns to adaptively mimic the harmful impacts of transformation combination to update the noise for improving the transferability.

B. Overall Illustration of the Proposed Method

Existing adversarial attack methods against LVLM have shown strong attack capabilities in the white-box setting. However, they are limited to their model/prompt-specific perturbation designs, and they will achieve a lower success rate once they are transferred to attack other models/prompts. We assume the reason is: previous works greedily perturb the images in the direction of the sign of the loss gradient at each iteration, easily falling into the poor local maxima and overfitting to the specific network. This phenomenon of generated adversarial samples overfitting to specific models/prompts is similar to the phenomenon that neural networks perform well on the training set but poorly on the test set.

Therefore, we propose to generate model/prompt-agnostic adversarial samples against LVLMs from the perspective of data augmentation, as shown in Fig.2. Specifically, we develop an adaptive transformation module to perform a wide spectrum of transformations and infer the most harmful deformations to adversarial examples. Once the generated adversarial examples are able to resist such distortions introduced during adversarial learning, they are transferable to attack unknown models/prompts with higher generalizability. Besides, to improve the imperceptibility, we optimize stealth pixel-wise perturbations to mimic the harmful transformation impacts via gradient approximation and impose them on the raw images. In this manner, our method can effectively generate transferable

Authorized licensed use limited to: BEIHANG UNIVERSITY. Downloaded on April 09,2025 at 14:02:44 UTC from IEEE Xplore. Restrictions apply.

adversarial samples with transformations and make the perturbation imperceptible with gradient optimization algorithms. We will illustrate them in the following.

C. Improving Transferability with Diverse Transformations

Data augmentation has been proven to be an effective method to prevent deep neural networks from overfitting during training. Since prompts are discrete data, generating adversarial text attacks is simple but less effective, and more resource-intensive than generating adversarial image attacks. Therefore, we choose to impose the transformation impacts on the input visual images to improve the transferability of the adversarial examples. Extensive studies [44], [45] also show that adversarial samples lose their attack effect under image transformation, which indicates these transformed adversarial images can serve as good samples for better optimization. Therefore, we consider a variety of image operations and the number of generated transformed images to enrich the diversity of adversarial inputs for improving the transferability.

Specifically, during the adversarial optimization, we first formulate various random transformation operations $T(\cdot)$ as:

$$T(\cdot) \in \{\text{Rotate, Scale, Resize, Pad, Mask, Translate, Shuffle}\}. \tag{2}$$

Then, we apply the random transformation combinations on images and expect both adversarial examples and their transformed versions can fool the LVLMs by:

$$max\ J(F(x_v^{adv}, x_p), y) + J(F(T(x_v^{adv}), x_p), y). \tag{3}$$

Although these transformations help to improve the adversarial transferability, they will make the perturbation more noticeable to the human. Therefore, we propose to further improve the imperceptibility in the next subsection.

D. Improving Imperceptibility with Gradient Approximation

To further improve the imperceptibility of the adversarial examples, we propose to impose the harmful transformation impacts as gradient approximation to be added on the raw images to have the same adversarial impacts. These gradient-based perturbations not only preserve the original image contents for *improving the imperceptibility*, but also mimic the transformation impacts for *improving the transferability*. In particular, to estimate the gradients for optimization in the black-box attack setting, adopting momentum into attacks can stabilize the estimate of the gradient directions of potential harmful transformation as:

$$grad_i = \nabla_{x_{v,t}^{adv}} J(F(T_i(x_{v,t}^{adv}), x_p), y), \tag{4}$$

where $grad_i$ is the gradient calculated on the i-th image transformed by $T(\cdot)$ at the t-th iteration. However, not all transformations make positive contributions to the transferability improvement. Therefore, we update the momentum of these transformations with adaptive weights w ($N_{w=1}$ is the number of positive contributions) and accordingly optimize the adversarial examples with approximation strategy as:

$$g = \frac{1}{N_{w=1}} \sum_{i=1}^{N} w_i \cdot grad_i, \tag{5}$$

$$w_i = \begin{cases} 1, & J(F(T_i(x_{v,t}^{adv}), x_p), y) > J(F(x_{v,t}^{adv}, x_p), y) \\ 0, & J(F(T_i(x_{v,t}^{adv}), x_p), y) < J(F(x_{v,t}^{adv}, x_p), y) \end{cases}, \tag{6}$$

$$g_{t+1} = \mu \cdot g_t + \frac{g}{\|g\|_1}, \tag{7}$$

where g_t gathers the harmful gradient information up to the t-th iteration with a decay factor μ. The final adversarial transformed images can be obtained by:

$$x_{v,t+1}^{adv} = x_{v,t}^{adv} + \alpha \cdot \text{sign}(g_{t+1}),\ x_{v,0}^{adv} = x_v, \tag{8}$$

where α is step size and sign$(\cdot)$ is sign function.

E. Generating the Final Adversarial Samples

After optimizing the raw images to mimic the harmful transformation impacts via gradient approximation, we can obtain the final adversarial samples with high imperceptibility and transferability. These samples are model/prompt-agnostic and can be transferred to attack other LVLM models with different prompts. We believe that our attack method is more scalable and practical than existing model/prompt-specific LVLM attacks, and provides a promising direction for future research in the area of LVLM adversarial robustness.

III. Experiments

A. Implementation Details

Datasets. We utilize SVIT [52] and DALL-E [53] datasets with image captioning, image classification, and VQA tasks. Each dataset consists of images and prompts.

LVLM Models. We select three commonly utilized open-sourced LVLM models, *i.e*, BLIP-2 [2], and InstructBLIP [20] and MiniGPT-4 [23], to implement attacks.

Evaluation Metrics. To evaluate the semantic changes of the LVLM, we measure the semantic similarity between raw and adversarial answers by SentenceTransformer [54].

Experimental Settings. For attack setup, we utilize a maximum of 500 epochs to optimize the adversarial noise and set the perturbation budget $\epsilon = 16$, the decay factor $\mu = 0.9$, the step size $\alpha = \epsilon/epoch$. The number of transformed images is set as $N = 20$ for gradient approximation.

B. Main Performance

To investigate the adversarial performance of our attack, we compare our method with two LVLM attack baselines. One baseline is the traditional PGD-based attack without transformation module ("No Trans.") [42] and the other is the SOTA method CroPA [28]. As shown in Table I, our attack achieves more harmful performance than the other two attacks across different LVLM models, datasets, and downstream tasks, demonstrating the effectiveness of our proposed attack.

C. Evaluation on the Transferability

Transferability Across Different LVLM Models. We first investigate the transferability of the proposed attack across different LVLM models. As shown in Table II, our attack achieves better transfer-attack performance than the other

Authorized licensed use limited to: BEIHANG UNIVERSITY. Downloaded on April 09,2025 at 14:02:44 UTC from IEEE Xplore. Restrictions apply.

TABLE I

PERFORMANCE COMPARISON WITH EXISTING LVLM ATTACKS ON THREE LVLM MODELS ACROSS TWO DATASETS WITH THREE DOWNSTREAM TASKS. THE SMALLER SIMILARITY SCORE INDICATES A MORE HARMFUL ATTACK PERFORMANCE (↓).

Dataset	Methods	BLIP-2			InstructBLIP			MiniGPT-4		
		Captioning	Classification	VQA	Captioning	Classification	VQA	Captioning	Classification	VQA
SVIT	No Trans.	0.352	0.362	0.548	0.421	0.406	0.812	0.523	0.547	0.646
	CroPA	0.261	0.288	0.641	0.451	0.407	0.758	0.410	0.443	0.541
	Ours	**0.131**	**0.145**	**0.356**	**0.290**	**0.250**	**0.600**	**0.383**	**0.440**	**0.520**
DALL-E	No Trans.	0.402	0.400	0.510	0.443	0.407	0.876	0.583	0.552	0.678
	CroPA	0.287	0.282	0.503	0.510	0.437	0.884	0.443	0.451	0.581
	Ours	**0.100**	**0.182**	**0.303**	**0.254**	**0.254**	**0.770**	**0.429**	**0.435**	**0.545**

TABLE II

TRANSFER-ATTACK AVERAGED PERFORMANCE ACROSS LVLM MODELS OR PROMPTS. THE PERFORMANCE IS THE LOWER THE BETTER (↓).

Dataset	Methods	Across LVLM Models						Across Prompts		
		From BLIP-2 to		From InstructBLIP to		From MiniGPT-4 to		Num=1	Num=10	Num=50
		InstructBLIP	MiniGPT-4	BLIP-2	MiniGPT-4	BLIP-2	InstructBLIP			
SVIT	No Trans.	0.530	0.542	0.490	0.578	0.471	0.572	0.413	0.411	0.409
	CroPA	0.493	0.525	0.484	0.481	0.500	0.508	0.460	0.429	0.434
	Ours	**0.436**	**0.486**	**0.407**	**0.479**	**0.385**	**0.477**	**0.266**	**0.306**	**0.310**
DALL-E	No Trans.	0.599	0.617	0.424	0.636	0.545	0.679	0.407	0.386	0.389
	CroPA	0.541	0.559	0.438	**0.521**	0.506	0.605	0.350	0.382	0.382
	Ours	**0.516**	**0.542**	**0.326**	0.529	**0.429**	**0.583**	**0.277**	**0.273**	**0.275**

TABLE III

ADVERSARIAL ROBUSTNESS AGAINST DIFFERENT DEFENSES.

Dataset	Defense	No Trans.	CroPA	Ours
SVIT	No Defense	0.421	0.397	**0.211**
	RandomRotate	0.766	0.640	**0.445**
	RandomResize	0.743	0.696	**0.326**
DALL-E	No Defense	0.437	0.357	**0.195**
	RandomRotate	0.739	0.588	**0.419**
	RandomResize	0.717	0.598	**0.340**

two methods, demonstrating our transformation-aware design indeed help to improve the transferability.

Transferability Across Different Prompts. We then investigate the transferability across different prompts. Specifically, we feed different prompts with the same adversarial images to different models to assess the prompt-based transferability. As shown in Table II, we test the attacks on different numbers of prompts where our method still achieves the best performance.

D. Evaluation on Defense Methods

To investigate the adversarial robustness of our attack, we implement two transformation-based defenses following previous work [28], [45], [55], *i.e*, RandomRotate and RandomResize, to defend LVLM attacks. As shown in Table III, our attack not only achieves the lowest performance degeneration compared to other methods but also achieves the best attack performance against different defenses, demonstrating the effectiveness of our proposed attack method.

E. Visualization

We provide the visualization results of our adversarial effects in Fig. 3. We can find that: (1) The perturbed image is almost similar to the raw image, demonstrating our high imperceptibility. (2) Our attack can effectively guide the LVLM to output wrong semantic texts, demonstrating our effectiveness.

Fig. 3. Visualization results of our generated adversarial examples on three tasks on the InstructBLIP model, where our attack misleads the LVLM model to output wrong semantic texts.

IV. CONCLUSION

In this paper, we propose a novel imperceptible transfer attack to improve the transferability of the LVLM attacks. Thanks to our devised adversarial transformation learning with gradient approximation strategies, different from previous model/prompt-specific LVLM attackers, our attack is model/prompt-agnostic and can achieve high attack performance when being transferred to attack other LVLM models with diverse prompts. Extensive experiments demonstrate the effectiveness of the proposed method.

Acknowledgements. This work is supported by National Natural Science Foundation of China (NSFC) under grant No. 62476107.

Authorized licensed use limited to: BEIHANG UNIVERSITY. Downloaded on April 09,2025 at 14:02:44 UTC from IEEE Xplore. Restrictions apply.

REFERENCES

[1] D. Liu, X. Qu, and W. Hu, "Reducing the vision and language bias for temporal sentence grounding," in *ACM MM*, 2022.

[2] 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 *ICML*, 2023.

[3] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds *et al.*, "Flamingo: a visual language model for few-shot learning," *NeurIPS*, 2022.

[4] D. Liu, Y. Liu, W. Huang, and W. Hu, "A survey on text-guided 3d visual grounding: Elements, recent advances, and future directions," *arXiv preprint arXiv:2406.05785*, 2024.

[5] D. Liu, X. Ouyang, S. Xu, P. Zhou, K. He, and S. Wen, "Saanet: Siamese action-units attention network for improving dynamic facial expression recognition," *Neurocomputing*, 2020.

[6] D. Liu, S. Xu, X.-Y. Liu, Z. Xu, W. Wei, and P. Zhou, "Spatiotemporal graph neural network based mask reconstruction for video object segmentation," in *AAAI*, 2021.

[7] D. Liu, P. Zhou, Z. Xu, H. Wang, and R. Li, "Few-shot temporal sentence grounding via memory-guided semantic learning," *IEEE TCSVT*, 2022.

[8] J. Tang, W. Zhang, H. Liu, M. Yang, B. Jiang, G. Hu, and X. Bai, "Few could be better than all: Feature sampling and grouping for scene text detection," in *CVPR*, 2022.

[9] Y. Liu, J. Zhang, D. Peng, M. Huang, X. Wang, J. Tang, C. Huang, D. Lin, C. Shen, X. Bai *et al.*, "Spts v2: single-point scene text spotting," *IEEE TPAMI*, 2023.

[10] J. Tang, W. Qian, L. Song, X. Dong, L. Li, and X. Bai, "Optimal boxes: boosting end-to-end scene text recognition by adjusting annotated bounding boxes via reinforcement learning," in *ECCV*, 2022.

[11] H. Feng, Q. Liu, H. Liu, T. Jingqun, W. Zhou, H. Li, and C. Huang, "Docpedia: Unleashing the power of large multimodal model in the frequency domain for versatile document understanding," *SCIS*, 2024.

[12] Z. Zhao, J. Tang, C. Lin, B. Wu, C. Huang, H. Liu, X. Tan, Z. Zhang, and Y. Xie, "Multi-modal in-context learning makes an ego-evolving scene text recognizer," in *CVPR*, 2024.

[13] Z. Zhao, J. Tang, B. Wu, C. Lin, S. Wei, H. Liu, X. Tan, Z. Zhang, C. Huang, and Y. Xie, "Harmonizing visual text comprehension and generation," 2024.

[14] A.-L. Wang, B. Shan, W. Shi, K.-Y. Lin, X. Fei, G. Tang, L. Liao, J. Tang, C. Huang, and W.-S. Zheng, "Pargo: Bridging vision-language with partial and global views," in *AAAI*, 2025.

[15] J. Tang, W. Du, B. Wang, W. Zhou, S. Mei, T. Xue, X. Xu, and H. Zhang, "Character recognition competition for street view shop signs," *NSR*, 2023.

[16] W. Zhao, H. Feng, Q. Liu, J. Tang, S. Wei, B. Wu, L. Liao, Y. Ye, H. Liu, W. Zhou, H. Li, and C. Huang, "Tabpedia: Towards comprehensive visual table understanding with concept synergy," in *NeurIPS*, 2024.

[17] J. Tang, C. Lin, Z. Zhao, S. Wei, B. Wu, Q. Liu, H. Feng, Y. Li, S. Wang, L. Liao *et al.*, "Textsquare: Scaling up text-centric visual instruction tuning," *arXiv*, 2024.

[18] J. Tang, S. Qiao, B. Cui, Y. Ma, S. Zhang, and D. Kanoulas, "You can even annotate text with voice: Transcription-only-supervised text spotting," in *ACM MM*, 2022.

[19] D. Liu, X. Fang, P. Zhou, X. Di, W. Lu, and Y. Cheng, "Hypotheses tree building for one-shot temporal sentence localization," in *AAAI*, 2023.

[20] W. Dai, J. Li, D. Li, A. M. H. Tiong, J. Zhao, W. Wang, B. Li, P. N. Fung, and S. Hoi, "Instructblip: Towards general-purpose vision-language models with instruction tuning," *NeurIPS*, 2024.

[21] J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, J. Lin, C. Zhou, and J. Zhou, "Qwen-vl: A frontier large vision-language model with versatile abilities," *arXiv preprint arXiv:2308.12966*, 2023.

[22] Y. Su, T. Lan, H. Li, J. Xu, Y. Wang, and D. Cai, "Pandagpt: One model to instruction-follow them all," *arXiv preprint arXiv:2305.16355*, 2023.

[23] 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.

[24] D. Liu, M. Yang, X. Qu, P. Zhou, W. Hu, and Y. Cheng, "A survey of attacks on large vision-language models: Resources, advances, and future trends," *arXiv preprint arXiv:2407.07403*, 2024.

[25] X. Liu, Y. Zhu, Y. Lan, C. Yang, and Y. Qiao, "Safety of multimodal large language models on images and text," *arXiv*, 2024.

[26] Y. Fan, Y. Cao, Z. Zhao, Z. Liu, and S. Li, "Unbridled icarus: A survey of the potential perils of image inputs in multimodal large language model security," *arXiv preprint arXiv:2404.05264*, 2024.

[27] Y. Zhao, T. Pang, C. Du, X. Yang, C. Li, N.-M. M. Cheung, and M. Lin, "On evaluating adversarial robustness of large vision-language models," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[28] H. Luo, J. Gu, F. Liu, and P. Torr, "An image is worth 1000 lies: Adversarial transferability across prompts on vision-language models," *arXiv preprint arXiv:2403.09766*, 2024.

[29] L. Bailey, E. Ong, S. Russell, and S. Emmons, "Image hijacks: Adversarial images can control generative models at runtime," *arXiv*, 2023.

[30] Y. Dong, H. Chen, J. Chen, Z. Fang, X. Yang, Y. Zhang, Y. Tian, H. Su, and J. Zhu, "How robust is google's bard to adversarial image attacks?" *arXiv preprint arXiv:2309.11751*, 2023.

[31] X. Wang, Z. Ji, P. Ma, Z. Li, and S. Wang, "Instructta: Instruction-tuned targeted attack for large vision-language models," *arXiv preprint arXiv:2312.01886*, 2023.

[32] Z. Wang, Z. Han, S. Chen, F. Xue, Z. Ding, X. Xiao, V. Tresp, P. Torr, and J. Gu, "Stop reasoning! when multimodal llms with chain-of-thought reasoning meets adversarial images," *arXiv*, 2024.

[33] H. Zhang, W. Shao, H. Liu, Y. Ma, P. Luo, Y. Qiao, and K. Zhang, "Avibench: Towards evaluating the robustness of large vision-language model on adversarial visual-instructions," *arXiv*, 2024.

[34] D. Lu, T. Pang, C. Du, Q. Liu, X. Yang, and M. Lin, "Test-time backdoor attacks on multimodal large language models," *arXiv*, 2024.

[35] X. Tao, S. Zhong, L. Li, Q. Liu, and L. Kong, "Imgtrojan: Jailbreaking vision-language models with one image," *arXiv*, 2024.

[36] Y. Tao, D. Liu, P. Zhou, Y. Xie, W. Du, and W. Hu, "3dhacker: Spectrum-based decision boundary generation for hard-label 3d point cloud attack," in *ICCV*, 2023.

[37] M. Yang, D. Liu, K. Tang, P. Zhou, L. Chen, and J. Chen, "Hiding imperceptible noise in curvature-aware patches for 3d point cloud attack," in *ECCV*, 2025.

[38] D. Liu, Y. Tao, P. Zhou, and W. Hu, "Hard-label black-box attacks on 3d point clouds," *arXiv*, 2024.

[39] D. Liu, W. Hu, and X. Li, "Robust geometry-dependent attack for 3d point clouds," *IEEE TMM*, 2023.

[40] D. Liu and W. Hu, "Explicitly perceiving and preserving the local geometric structures for 3d point cloud attack," in *AAAI*, 2024.

[41] C. H. Wu, J. Y. Koh, R. Salakhutdinov, D. Fried, and A. Raghunathan, "Adversarial attacks on multimodal agents," *arXiv*, 2024.

[42] X. Cui, A. Aparcedo, Y. K. Jang, and S.-N. Lim, "On the robustness of large multimodal models against image adversarial attacks," in *CVPR*, 2024.

[43] D. Liu, M. Yang, X. Qu, P. Zhou, X. Fang, K. Tang, Y. Wan, and L. Sun, "Pandora's box: Towards building universal attackers against real-world large vision-language models," in *NeurIPS*, 2024.

[44] C. Guo, M. Rana, M. Cisse, and L. Van Der Maaten, "Countering adversarial images using input transformations," *arXiv*, 2017.

[45] C. Xie, J. Wang, Z. Zhang, Z. Ren, and A. Yuille, "Mitigating adversarial effects through randomization," *arXiv preprint arXiv:1711.01991*, 2017.

[46] D. Liu and W. Hu, "Imperceptible transfer attack and defense on 3d point cloud classification," *IEEE TPAMI*, 2022.

[47] Q. Hu, D. Liu, and W. Hu, "Exploring the devil in graph spectral domain for 3d point cloud attacks," in *ECCV*, 2022.

[48] D. Liu, W. Hu, and X. Li, "Point cloud attacks in graph spectral domain: When 3d geometry meets graph signal processing," *IEEE TPAMI*, 2023.

[49] A. Shrivastava, A. Gupta, and R. Girshick, "Training region-based object detectors with online hard example mining," in *CVPR*, 2016.

[50] E. Simo-Serra, E. Trulls, L. Ferraz, I. Kokkinos, P. Fua, and F. Moreno-Noguer, "Discriminative learning of deep convolutional feature point descriptors," in *ICCV*, 2015.

[51] X. Cai, Y. Tao, D. Liu, P. Zhou, X. Qu, J. Dong, K. Tang, and L. Sun, "Frequency-aware gan for imperceptible transfer attack on 3d point clouds," in *ACM MM*, 2024.

[52] B. Zhao, B. Wu, and T. Huang, "Svit: Scaling up visual instruction tuning," *arXiv preprint arXiv:2307.04087*, 2023.

[53] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," in *ICML*, 2021.

[54] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. ACL, 2019.

[55] I. Frosio and J. Kautz, "The best defense is a good offense: adversarial augmentation against adversarial attacks," in *CVPR*, 2023.

Authorized licensed use limited to: BEIHANG UNIVERSITY. Downloaded on April 09,2025 at 14:02:44 UTC from IEEE Xplore. Restrictions apply.