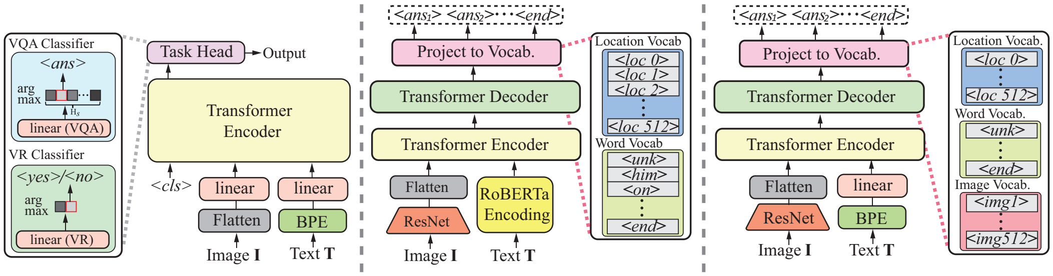
VLATTACK: Multimodal Adversarial Attacks on Vision-Language Tasks via Pre-trained Models (Supplementary Materials)

Anonymous Author(s)
Affiliation
Address
`email`

A. Details of VL Models



(a) ViLT (b) Unitab (c) OFA

Figure 1: An illustration of ViLT, Unitab, and OFA model structures.

This section gives the details of ViLT, Unitab, and OFA models, and their structures are illustrated in Figure 1

- **ViLT**. We select ViLT [1] as the **encoder-only** VL model because of its succinct structure and prominent performance on multiple downstream tasks. Given an input image $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$ and a sentence $\mathbf{T}$, ViLT yields M image tokens using a linear transformation on the flattened image patches, where each token is a 1D vector and $M = \frac{HW}{P^2}$ for a given patch resolution (P, P). Word tokens are encoded through a Byte-Pair Encoder (BPE) [2] and a word-vector linear projection. Then tokens of two modalities and a special learnable token $\langle cls \rangle$ are concatenated. By attending visual and text tokens and a special token $\langle cls \rangle$ in a Transformer encoder with twelve layers, the output feature from the $\langle cls \rangle$ token is fed into a task-specific classification head for the final output. Taking the VQA task as an example, the VQA classifier adopts a linear layer to output a vector with H_s elements, where H_s is the number of all possible choices in the closed answer set of the VQA task. The final output is obtained through the element with the highest response in the vector.
- **Unitab.** Unitab adopts an **encoder-decoder** framework. It first embeds text $\mathbf{T}$ via RoBERT$_a$ [3] and flats features after encoding image $\mathbf{I}$ through ResNet [4]. The attached visual and text token features are then fed into a standard Transformer network [5] with six encoder layers and six decoder layers. Finally, the sequence predictions $[\langle ans_1 \rangle, \langle ans_2 \rangle, \cdots, \langle end \rangle]$ are obtained auto-regressively through a projection head. The network stops regressing when an end token $\langle end \rangle$ appears. For different tasks, the output tokens may come from different pre-defined vocabularies. Given the REC task as an example, four tokens $[(\langle loc\ x_1 \rangle, \langle loc\ x_2 \rangle), (\langle loc\ x_3 \rangle, \langle loc\ x_4 \rangle)]$ will be selected from the location vocabulary, which forms the coordinate of a bounding box. As a result, these models can handle both text and grounding tasks.

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

Table 1: An illustration of all datasets and tasks evaluated in our paper.

Datasets	Task	Task description	Attack Model			Attack Modality	
			OFA	Unitab	ViLT	Image	Text
VQAv2	VQA	Scene Understanding QA	✓	✓	✓	✓	✓
SNLI-VE	VE	VL Entailment	✓			✓	✓
RefCOCO	REC	Bounding Box Localization	✓	✓		✓	✓
RefCOCOg	REC	Bounding Box Localization	✓	✓		✓	✓
RefCOCO+	REC	Bounding Box Localization	✓	✓		✓	✓
NLVR2	VR	Image-Text Pairs Matching			✓	✓	✓
MSCOCO	Captioning	Image Captioning	✓			✓	
ImageNet-1K	Classification	Object Classification	✓			✓	

- **OFA.** As shown in Figure 1 (c), OFA also adopts an **encoder-decoder** structure. Different from Unitab, it adopts the BPE to encode text and extends the linguistic vocabulary by adding image quantization tokens [6] $\langle img \rangle$ for synthesis tasks. ***Note that the main difference between OFA and Unitab lies in their pre-training and fine-tuning strategies rather than the model structure***. For example, in the pre-training process, Unitab focuses on learning alignments between predicted words and boxes through grounding tasks, while OFA captures more general representations through multi-task joint training that includes both single-modal and multi-modal tasks. Overall, OFA outperforms Unitab in terms of performance improvement.

B. Dataset and Implementation

B.1 Tasks and Datasets

To verify the generalization ability of our proposed VLAttack, we evaluate a wide array of popular vision language tasks summarized in Table 1. Specifically, the selected tasks span from text understanding (visual reasoning, visual entailment, visual question answering) to image understanding (image classification, captioning) and localization (referring expression comprehension).

For each dataset, we randomly select 5K **correctly predicted samples** in the corresponding validation dataset to evaluate the ASR performance. All validation datasets follow the same split settings as adopted in the respective attack models. Because VQA is a multiclass classification task, we select a correct prediction only if the prediction result is the same as the label with the highest VQA score[1], and regard the label as the ground truth in Eq. (1). In the REC task, a correct prediction is considered when the Intersection-over-Union (IoU) score between the predicted and ground truth bounding box is larger than 0.5. We adopt the same IoU threshold as in Unitab [7] and OFA [8].

B.2 Implementation Details

For the perturbation parameters of images, we follow the setting in the common transferable image attacks [9, 10] and set the maximum perturbation σ_i of each pixel to 16/255 on all tasks except REC. Considering that even a single coordinate change can affect the final grounding results to a great extent, the σ_i of the REC task is 4/255 to better highlight the ASR differences among distinct methods. The total iteration number N and step size are set to 40 and 0.01 by following the projected gradient decent method [11], and N_s is 20. For the perturbation on the text, the semantic similarity constraint σ_s is set to 0.95, and the number of maximum modified words is set to 1 by following the previous text-attack work [12, 13] to ensure the semantic consistency and imperceptibility. All experiments are conducted on a single GTX A6000 GPU. The analysis of parameter selection can be found in Section B.3.

B.3 Parameter Sensitivity Analysis

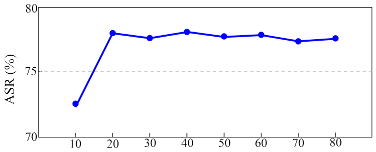
We discuss the effect of different iteration numbers of N and N_s in VLAttack. All experiments are conducted on the VQAv2 dataset and the ViLT model. The total iteration number N is set from

---

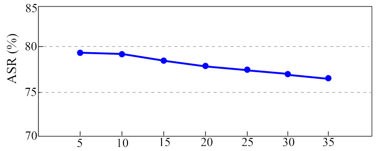
[1] The VQA score calculates the percentage of the predicted answer that appears in 10 reference ground truth answers. More details can be found via `https://visualqa.org/evaluation.html`

10 to 80, N_s is set to $\frac{N}{2}$. As depicted in Figure 2(a), the ASR performance is dramatically improved by increasing N from 10 to 20 steps and then achieves the best result when $N = 40$.

We next investigate the impact of different initial iteration numbers N_s. We test N_s from 5 to 40, but the total iteration number N is fixed to 40. As shown in Figure 2(b), the ASR score reaches the summit when N_s is 5, and it is smoothly decreased by continually enlarging N_s. Considering that the smaller initial iteration number N_s increases the ratio of text perturbations, we set N_s as 20 to obtain the best trade-off between attack performance and the naturalness of generated adversarial samples in our experiments.



(a) Iteration Number N



(b) Iteration Number N_s

Figure 2: Investigation of iteration number N and N_s. (a) Various total iteration number N, where N_s is set to $\frac{N}{2}$. (b) Various initial iteration numbers N_s, where N is set to 40.

C. More Ablation Results

In Section 5.4, we conduct an ablation study to show the effectiveness of each module in our model design on VQA, VE, and REC tasks. Here, we conduct additional ablation experiments for the remaining tasks, including visual reasoning, image captioning, and image classification.

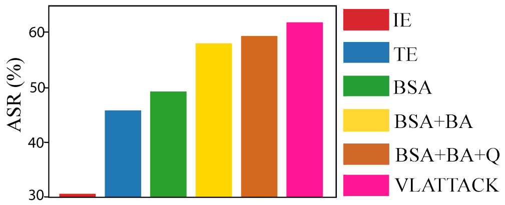


Figure 3: ViLT-VR.

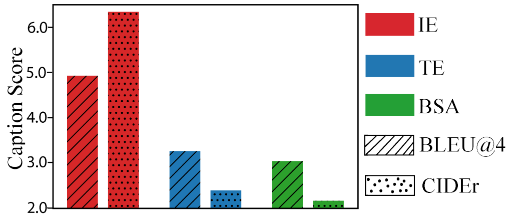


Figure 4: OFA-captioning.

Figure 3 shows the results of the ablation study on the VR task using the ViLT model. We can observe that only using the image encoder results in significantly low ASR. However, by combining it with the Transformer encoder (TE), BSA can achieve a high ASR. These results show the reasonableness of considering two encoders simultaneously when attacking the image modality. The result of BSA+BA demonstrates the usefulness of attacking the text modality. Although BSA+BA+Q outperforms other approaches, its performance is still lower than that of the proposed VLAttack. This comparison proves that the proposed iterative cross-search attack (ICSA) strategy is effective for the multimodal attack again.

Figure 4 shows the results of the image captioning task using the OFA model. Because the image captioning task only accepts a fixed text prompt for prediction, we only perturb the image and report the results on IE, TE, and BSA. For this task, we report BLEU@4 and CIDEr scores. **The lower, the better**. We can observe that the proposed BSA outperforms IE and TE, indicating our model design's effectiveness.

Figure 5 shows the results of the image classification task using the OFA model. Similar to the image captioning task, we only attack images. The evaluation metric for this task is ASR. **The higher, the better**. We can have the same observations with other ablation studies, where attacking both encoders outperforms attacking a single encoder.

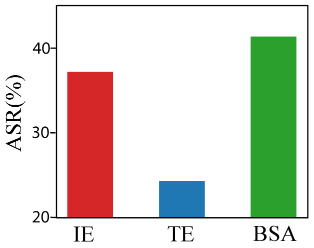


Figure 5: OFA-classification.

D. Different Optimization Methods

VLAttack can be easily adapted to various optimization methods in image attacks. To demonstrate the generalizability of our method, we replace the projected gradient decent [11] in VLAttack with Momentum Iterative method (MI) [14] and Diverse Input attack (DI) [15] since they

Table 2: Combining VLAttack with different gradient-based image attack schemes.

Method	ViLT		Unitab			
	VQAv2	NLVR2	VQAv2	RefCOCO	RefCOCO+	RefCOCOg
BSA$_{MI}$	65.40	52.32	50.38	86.00	89.20	87.39
VLAttack$_{MI}$	**78.77**	**67.16**	**63.02**	**92.46**	**93.10**	**94.34**
BSA$_{DI}$	65.94	52.30	42.74	90.30	91.56	91.00
VLAttack$_{DI}$	**78.07**	**67.50**	**61.22**	**93.98**	**94.04**	**94.76**

have shown better performance than traditional iterative attacks [11, 16]. The replaced methods are denoted by BSA$_{MI}$, and VLAttack$_{MI}$ using MI, BSA$_{DI}$ and VLAttack$_{DI}$ using DI, respectively. Experiments are developed on ViLT and Unitab. Results are shown in Table 2. Using MI and DI optimizations, BSA$_{MI}$ and BSA$_{DI}$ still outperform all baselines displayed in Table 1 in the main manuscript. Also, VLAttack$_{MI}$ and VLAttack$_{DI}$ outperform the image attack method BSA$_{MI}$ and BSA$_{DI}$ with an average ASR improvement of 9.70% and 9.29% among all datasets. The gain of performance demonstrates that the proposed VLAttack can be further improved by combining with stronger gradient-based optimization schemes.

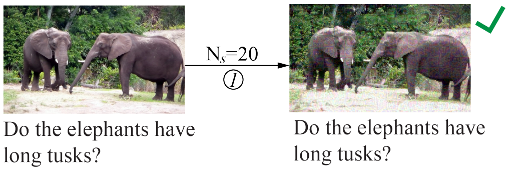

N_s=20
①
Do the elephants have long tusks?
Do the elephants have long tusks?

Figure 6: An adversarial image from BSA.

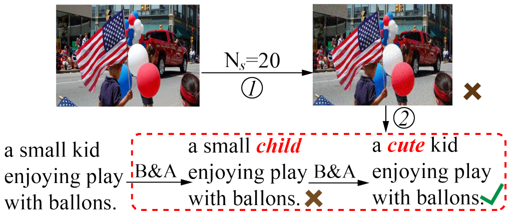

N_s=20
①
②
a small kid enjoying play with ballons.
B&A
a small *child* enjoying play with ballons.
B&A
a *cute* kid enjoying play with ballons.

Figure 7: An adversarial sentence from text attack.

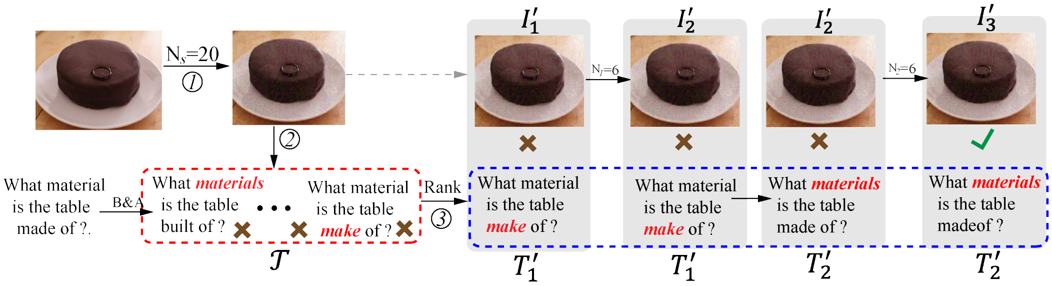

N_s=20
①
②
③
I'_1
I'_2
I'_2
I'_3
N_s=6
N_s=6
What material is the table made of ?.
B&A
What *materials* is the table built of ?
What material is the table *make* of ?
Rank
$\mathcal{T}$
What material is the table *make* of ?
What material is the table *make* of ?
What *materials* is the table made of ?
What *materials* is the table madeof ?
T'_1
T'_1
T'_2
T'_2

Figure 8: An adversarial image-text pair from multi-modal attack.

E. Case Study

E.1 How does VLAttack generate adversarial samples?

The proposed VLAttack aims to attack multimodal VL tasks starting by attacking single modalities. If they are failed, VLAttack uses the proposed interactive cross-search attack (ICSA) strategy to generate adversarial samples. In this experiment, we display the generated adversarial cases from different attack steps, including the image modality in Figure 6, the text modality in Figure 7, and the multimodal attack in Figure 8.

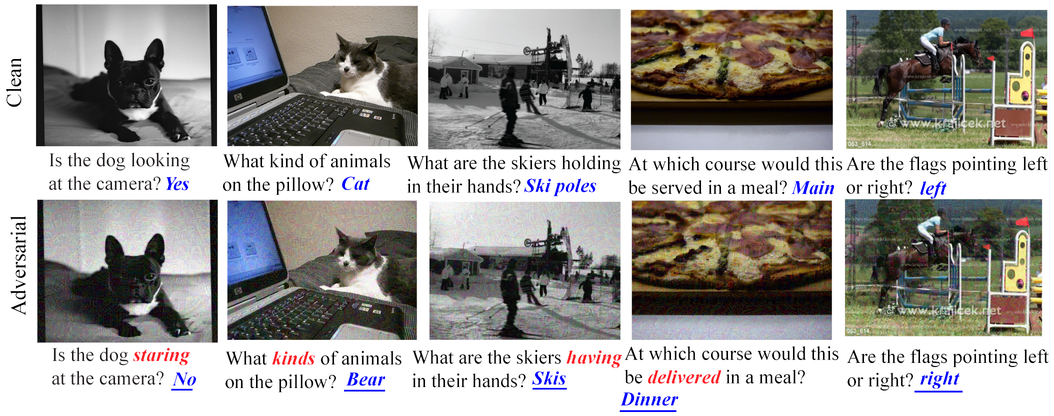
Single-modal Attacks (Section 4.1). VLAttack first perturbs the image modality using the proposed BSA and only outputs the adversarial image if the attack is successful (Algorithm 1 lines 1-5). As shown in Figure 6, only attacking the image modality, VLAttack can generate a successful adversarial sample to fool the downstream task. Then, VLAttack will stop. Otherwise, it will perturb the text through BERT-Attack (B&A) and use the clean image as the input, which is illustrated in Figure 7 (Algorithm 1, lines 6-15). During the text attack, B&A will generate multiple candidates by replacing the synonyms of a word. Since the length of text sentences is very short in the VL datasets, we only replace one word each time. From Figure 7, we can observe that B&A first replaces “kid” with its synonym “child”, but this is not an adversarial sample. B&A then moves to the next word “small” and uses its synonym “cute” as the perturbation. By querying the black-box downstream task model, VLAttack successes, and the algorithm will stop.

Multimodal Attack (Section 4.2). If the single-modal attack fails, VLAttack moves to the multimodal attack by iteratively cross-updating image and text perturbations, where image perturbations are added through BSA, and text perturbations are added according to the semantic similarity. The cross-updating process is repeated until an adversarial image-text pair is found (Algorithm 1, lines 16-24). Figure 8 shows an example. In Step ①, VLAttack fails to attack the image modality and outputs a perturbed image denoted as $\mathbf{I}'_1$. In Step ②, VLAttack also fails to attack the text modality and outputs a list of text perturbations $\mathcal{T}$. VLAttack has to use the multimodal attack to generate adversarial samples in Step ③. It first ranks the text perturbations in $\mathcal{T}$ according to the semantic similarity between the original text and each perturbation. The ranked list is denoted as $\{\hat{\mathbf{T}}'_1, \cdots, \hat{\mathbf{T}}'_K\}$. Then it equally allocates the iteration number of the image attack to generate the image perturbations iteratively. In Figure 8, this number is 6, which means we run BSA with the budget 6 to generate a new image perturbation.

VLAttack takes the pair $(\mathbf{I}'_1, \hat{\mathbf{T}}'_1)$ as the input to query the black-box downstream model, where $\hat{\mathbf{T}}'_1 =$ “*What material is the table make of?*”. If this pair is not an adversarial sample, then the proposed ICSA will adopt BSA to generate the new image perturbation $\mathbf{I}'_2$. The new pair $(\mathbf{I}'_2, \hat{\mathbf{T}}'_1)$ will be checked again. If it is still not an adversarial sample, VLAttack will use the next text perturbation $\hat{\mathbf{T}}'_2 =$ “*What materials is the table made of?*” and the newly generated image perturbation $\mathbf{I}'_2$ as the input and repeat the previous steps until finding a successful adversarial sample or using up all K text perturbations in $\mathcal{T}$. VLAttack employs a systematic strategy for adversarial attacks on VL models, sequentially targeting single-modal and multimodal perturbations to achieve successful adversarial attacks. *Note that we miss one line “**if** $S(\mathbf{I}'_{k+1}, \mathbf{T}'_k) \neq y$ **then return** $(\mathbf{I}'_{k+1}, \mathbf{T}'_k)$” between Lines 22 and 23 in Algorithm 1 of the main manuscript.*

E.2 Case Study on Different Tasks

We also provide additional qualitative results from Figure 9 to Figure 14 for experiments on all six tasks. For better visualization, we display the adversarial and clean samples side by side in a single column. By adding pixel and word perturbations, the fidelity of all samples is still preserved, but predictions are dramatically changed. For instance, in the image captioning task of Figure 13, all generated captions show no correlation with the input images. Some texts may even include replacement Unicode characters, such as “\ufffd”, resulting in incomplete sentence structures.



Is the dog looking at the camera? ***Yes***

What kind of animals on the pillow? ***Cat***

What are the skiers holding in their hands? ***Ski poles***

At which course would this be served in a meal? ***Main***

Are the flags pointing left or right? ***left***

Is the dog ***staring*** at the camera? ***No***

What ***kinds*** of animals on the pillow? ***Bear***

What are the skiers ***having*** in their hands? ***Skis***

At which course would this be ***delivered*** in a meal? ***Dinner***

Are the flags pointing left or right? ***right***

Figure 9: Additional quantitative results on visual question answering (VQA).

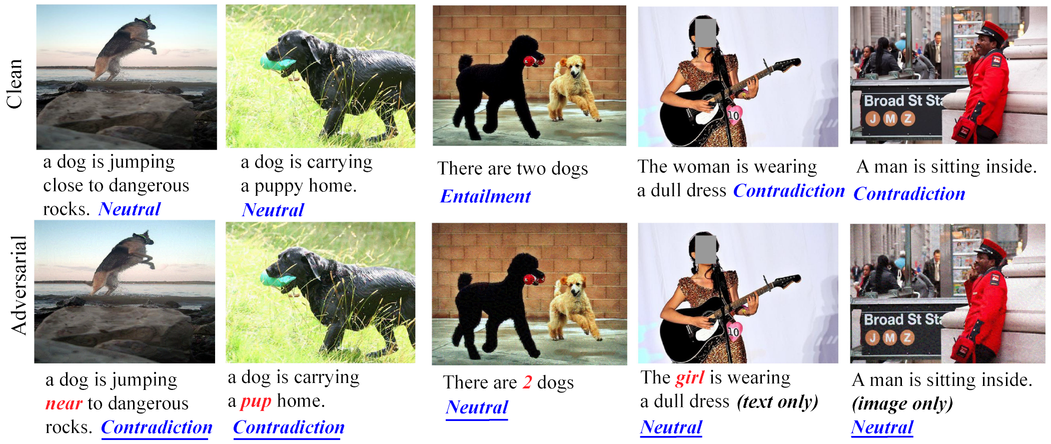

Clean
a dog is jumping close to dangerous rocks. Neutral
a dog is carrying a puppy home. Neutral
There are two dogs Entailment
The woman is wearing a dull dress Contradiction
A man is sitting inside. Contradiction
Adversarial
a dog is jumping near to dangerous rocks. Contradiction
a dog is carrying a pup home. Contradiction
There are 2 dogs Neutral
The girl is wearing a dull dress (text only) Neutral
A man is sitting inside. (image only) Neutral

Figure 10: Additional quantitative results on visual entailment (VE).

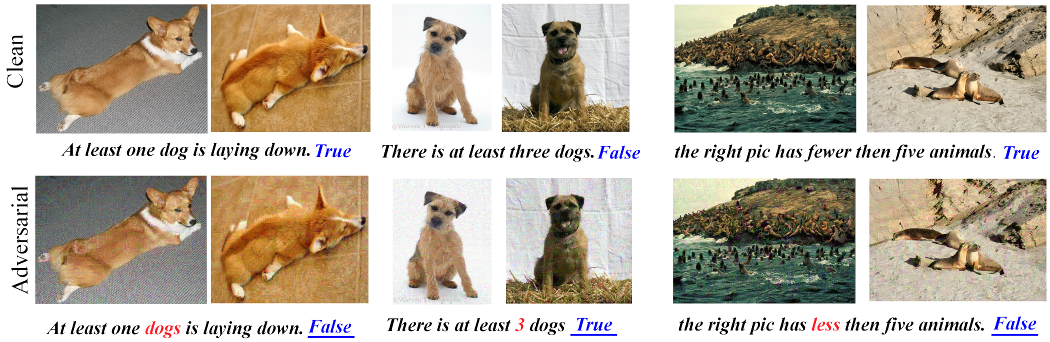

Clean
At least one dog is laying down. True
There is at least three dogs. False
the right pic has fewer then five animals. True
Adversarial
At least one dogs is laying down. False
There is at least 3 dogs True
the right pic has less then five animals. False

Figure 11: Additional quantitative results on visual reasoning (VR).

F. Limitations

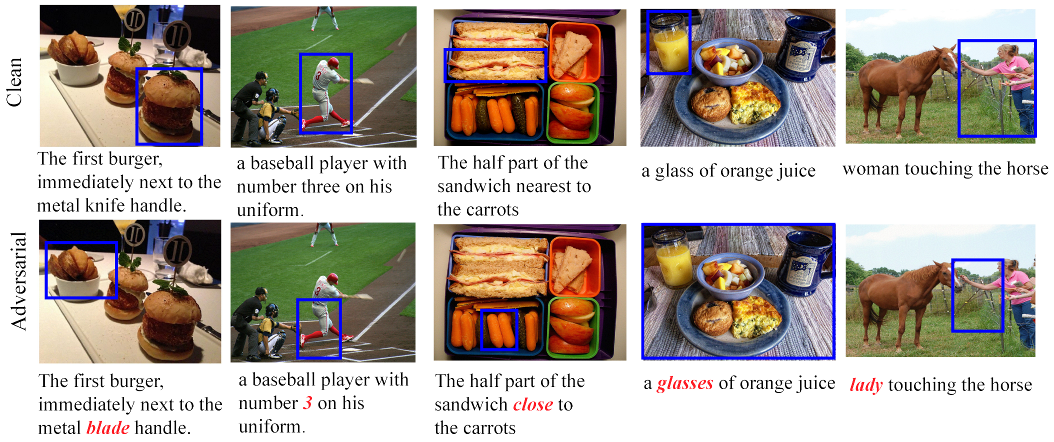
The limitations of our work can be summarized from the following two aspects. On the one hand, in our current model design, for the text modality, we directly apply the existing model instead of developing a new one. Therefore, there is no performance improvement on tasks that only accepts texts as input, such as text-to-image synthesis. On the other hand, our research problem is formulated by assuming the pre-trained and downstream models share similar structures. The adversarial transferability between different pre-trained and fine-tuned models is worth exploring, which we left to our future work.

G. Broad Impacts

Our research reveals substantial vulnerabilities in vision-language (VL) pretrained models, underlining the importance of adversarial robustness cross pre-trained and fine-tuned models. By exposing these vulnerabilities through the VLAttack strategy, we offer inspiration for developing more robust models. Furthermore, our findings underscore the ethical considerations of using VL models in real-world applications, especially those dealing with sensitive information and big data. Overall, our work emphasizes the necessity of balancing performance and robustness in VL models, with implications extending across computer vision, natural language processing, and broader artificial intelligence applications.

References

[1] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *ICML*, volume 139 of *Proceedings of Machine Learning*



The first burger, immediately next to the metal knife handle.

a baseball player with number three on his uniform.

The half part of the sandwich nearest to the carrots

a glass of orange juice

woman touching the horse

The first burger, immediately next to the metal ***blade*** handle.

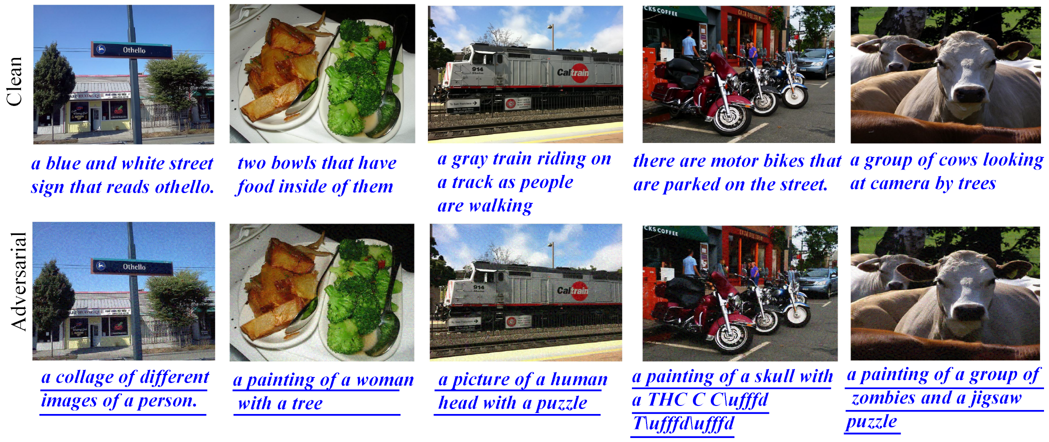
a baseball player with number ***3*** on his uniform.

The half part of the sandwich ***close*** to the carrots

a ***glasses*** of orange juice

lady touching the horse

Figure 12: Additional quantitative results on referring expression comprehension (REC).



a blue and white street sign that reads othello.

two bowls that have food inside of them

a gray train riding on a track as people are walking

there are motor bikes that are parked on the street.

a group of cows looking at camera by trees

a collage of different images of a person.

a painting of a woman with a tree

a picture of a human head with a puzzle

a painting of a skull with a THC C C\ufffd T\ufffd\ufffd

a painting of a group of zombies and a jigsaw puzzle

Figure 13: Additional quantitative results on the image captioning task.

Research, pages 5583–5594. PMLR, 2021.

[2] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *ACL (1)*. The Association for Computer Linguistics, 2016.

[3] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778. IEEE Computer Society, 2016.

[5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017.

[6] Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, pages 12873–12883. Computer Vision Foundation / IEEE, 2021.

[7] Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Faisal Ahmed, Zicheng Liu, Yumao Lu, and Lijuan Wang. Unitab: Unifying text and box outputs for grounded vision-language modeling. In *ECCV (36)*, volume 13696 of *Lecture Notes in Computer Science*, pages 521–539. Springer, 2022.

[8] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. OFA: unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *ICML*, volume 162 of *Proceedings of Machine Learning Research*, pages 23318–23340. PMLR, 2022.

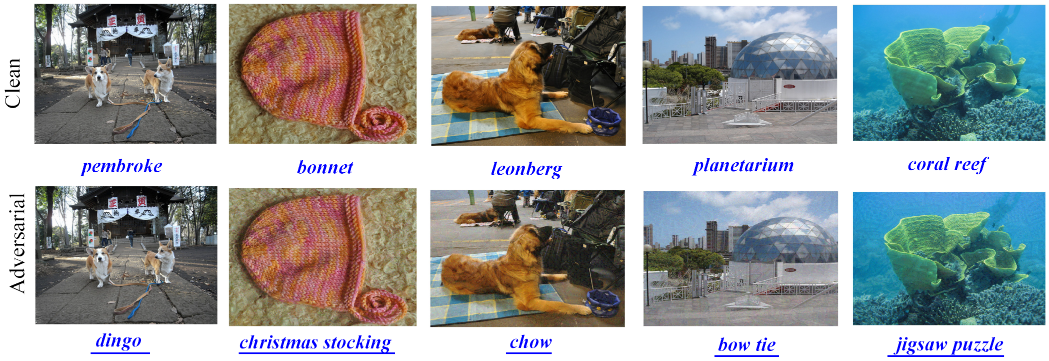


Figure 14: Additional quantitative results on the image classification task.

[9] Dongxian Wu, Yisen Wang, Shu-Tao Xia, James Bailey, and Xingjun Ma. Skip connections matter: On the transferability of adversarial examples generated with resnets. In *ICLR*. OpenReview.net, 2020.

[10] Zhipeng Wei, Jingjing Chen, Micah Goldblum, Zuxuan Wu, Tom Goldstein, and Yu-Gang Jiang. Towards transferable adversarial attacks on vision transformers. In *AAAI*, pages 2668–2676. AAAI Press, 2022.

[11] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *ICLR (Poster)*. OpenReview.net, 2018.

[12] Lei Xu, Alfredo Cuesta-Infante, Laure Berti-Équille, and Kalyan Veeramachaneni. R&r: Metric-guided adversarial sentence generation. In *AACL/IJCNLP (Findings)*, pages 438–452. Association for Computational Linguistics, 2022.

[13] Jiaming Zhang, Qi Yi, and Jitao Sang. Towards adversarial attack on vision-language pretraining models. In *ACM Multimedia*, pages 5005–5013. ACM, 2022.

[14] Jiafeng Wang, Zhaoyu Chen, Kaixun Jiang, Dingkang Yang, Lingyi Hong, Yan Wang, and Wenqiang Zhang. Boosting the transferability of adversarial attacks with global momentum initialization. *CoRR*, abs/2211.11236, 2022.

[15] Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L. Yuille. Improving transferability of adversarial examples with input diversity. In *CVPR*, pages 2730–2739. Computer Vision Foundation / IEEE, 2019.

[16] Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. Adversarial machine learning at scale. In *ICLR (Poster)*. OpenReview.net, 2017.