

## A Experiment on the Cosine Similarity of SGA and DRA

We use ALBEF as the surrogate model and conduct experiments on the Flickr30K dataset. Specifically, we measure the similarity for both positive and negative image-text pairs over 10 iterations for SGA, DRA, and HQA-VLAttack. By “negative pairs”, we mean the non-matching image-text pairs. For all the mentioned methods SGA, DRA and HQA-VLAttack, the “negative pairs” are constructed by the perturbed images and non-matching perturbed texts drawn from other examples in the same batch. For each image, the positive similarity is calculated as the average similarity between the image and its 5 corresponding captions, while the negative similarity is computed by averaging the similarity between the image and 5 randomly selected negative captions. Finally, these values are averaged across 128 images. The results, shown in Figure 1, illustrate the evolution of both positive and negative similarities over iterations, providing strong evidence in support of our hypothesis.

## B Investigation on the Layer Importance

We randomly select 100 image-text pairs from the Flickr30K dataset and conduct experiments using the image encoder of the ALBEF model. To compute image feature similarity, we first extract the original image features using the full 12-layer encoder. Then, we systematically omit each intermediate layer, extract features using the remaining 11 layers, and calculate their similarity to the original features, averaging the results. For [CLS] token embedding similarity, we extract the [CLS] embeddings from each of the 12 encoder layers and compute their similarity to the embeddings from the final (12th) layer. The final score is obtained by averaging the similarities across all layers.

## C Algorithm of HQA-VLAttack

---

### Algorithm 1: HQA-VLAttack

---

**Input** : Image batch  $V = \{v_1, v_2, \dots, v_B\}$ , Caption batch  $T = \{t_{1,1}, t_{1,2}, \dots, t_{B,M}\}$ , iteration steps  $N$ , random perturbation  $\delta \sim U(-\epsilon, \epsilon)$ , step-size  $\alpha$ , maximum perturbation boundary of the image  $\epsilon_v$

**Output** : Adversarial image batch  $V'$ , Adversarial caption batch  $T'$

```

1 for  $t$  in  $T$  do
2   | Determining the substitute word set  $C(x)$  for each word  $x$  in  $t$  according to Eq. (3)
3   | Generating the candidate text set  $\mathbb{C}(t)$  for  $t$ 
4 end
5 Generating textual adversarial example  $T'$  by Eq. (4)
6 Initialization adversarial image batch  $V'_0$ 
7 Using  $V$  to determine layer importance  $w_{i,l}$  by Eq. (5)
8 while  $n \leq N$  do
9   | // Generate initial image adversarial example
10  | Using  $(V', V'_{n-1})$  to calculate  $\mathcal{L}_l$  by Eq. (6)
11  |  $V'_n \leftarrow \text{Clip}_{\epsilon_v}(V'_{n-1} + \alpha \cdot \text{sign}(\frac{\nabla \mathcal{L}_l}{\|\mathcal{L}_l\|}))$ 
12  | // Contrastive Learning Based Image Adversarial Example Optimization
13  | Using  $(V'_n, T')$  to calculate  $\mathcal{L}_c$  by Eq. (7)
14  |  $V'_{n+1} \leftarrow \text{Clip}_{\epsilon_v}(V'_n + \alpha \cdot \text{sign}(\frac{\nabla \mathcal{L}_c}{\|\mathcal{L}_c\|}))$ 
15 end
16 return the adversarial example  $T'$  and  $V'_N$ 

```

---

## D Dataset Description

The detailed dataset description is as follows.

- **Flickr30K** is a widely used multimodal dataset comprising 31,783 images, each annotated with five descriptive captions. Typically, the dataset is partitioned into 29,783 images for training, 1,000 images for validation, and 1,000 images for testing, providing a standardized split for evaluating various vision-language tasks.

- **MSCOCO** is a large-scale benchmark extensively utilized in computer vision and vision-language research. Introduced in 2014, it comprises 123,287 images capturing complex scenes with multiple objects. Each image is annotated with five natural language captions and detailed object information—including bounding boxes, instance segmentation masks, and object labels—facilitating a wide range of tasks such as image captioning, object detection, and segmentation. For image captioning, the widely adopted Karpathy split is often employed. In this split, the dataset is partitioned into 82,783 training images, 5,000 validation images, and 5,000 test images. This standardized division ensures consistent evaluation across different models and tasks, further cementing MSCOCO’s role as an essential resource for advancing state-of-the-art methodologies.
- **RefCOCO+** is a benchmark dataset designed for referring expression comprehension and visual grounding. Derived from the MSCOCO images, RefCOCO+ provides natural language expressions that uniquely describe objects within complex scenes. Unlike its predecessor RefCOCO, the RefCOCO+ annotations deliberately avoid absolute spatial terms, focusing instead on appearance-based descriptions. This design choice makes the dataset particularly challenging, as models must rely on subtle visual cues rather than explicit spatial indicators. The dataset comprises 141,564 referring expressions for 50,000 objects across 19,992 MSCOCO images, offering rich linguistic diversity and detailed object annotations. As such, RefCOCO+ has become an essential resource for evaluating and advancing state-of-the-art methods in visual grounding and referring expression tasks.

## E Visualization

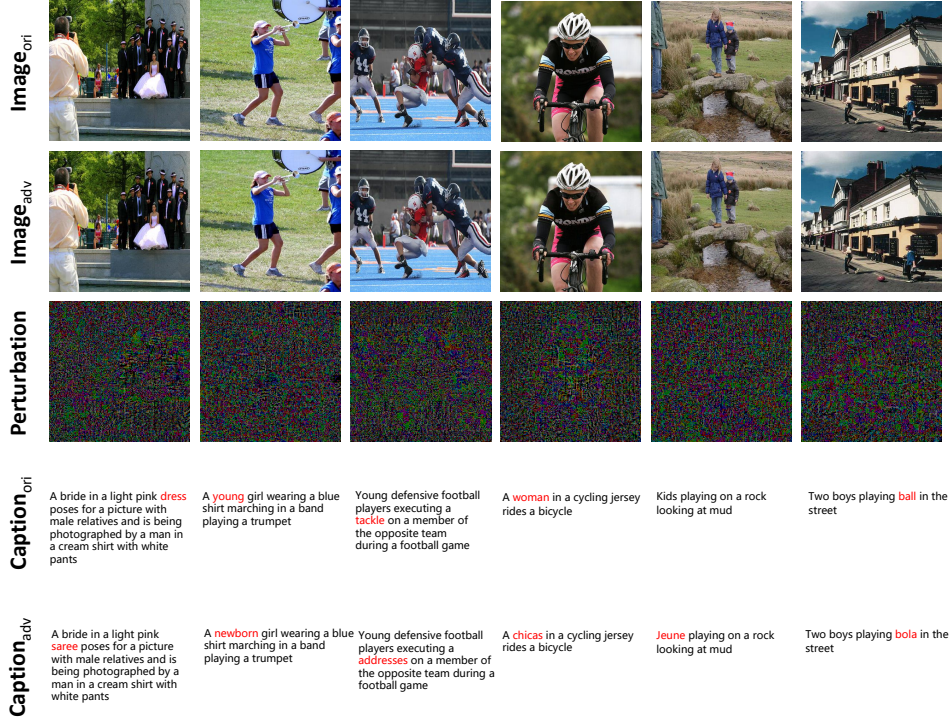


Figure 1: Adversarial Perturbation Visualization on Flickr30k Dataset with ALBEF as Surrogate Model.

In this section, we present some visualization results of HQA-VLAttack, as shown in Figure 1. The first row displays the original image, the second row shows the image with added perturbations, and the third row visualizes the perturbation effect magnified by a factor of 50. Even under the 50x magnification, the added perturbations remain barely noticeable, which demonstrates the high imperceptibility of our attack. The fourth row includes the original caption associated with the image,

while the fifth row shows the caption after the adversarial attack has been applied. This visualization demonstrates that HQA-VLAttack can generate adversarial examples that are imperceptible to humans but can successfully mislead VLP models, thereby proving the effectiveness of the adversarial examples generated by HQA-VLAttack.

## F Influence of Penalty Factor $\lambda$

Table 1: Impact of Penalty Factor  $\lambda$  on Results.

$\lambda$	TCL		CLIP <sub>ViT</sub>		CLIP <sub>CNN</sub>	
	TR R@1	IR R@1	TR R@1	IR R@1	TR R@1	IR R@1
-6	71.13	<b>77.86</b>	51.17	62.15	58.75	65.15
-8	72.92	77.81	51.9	61.95	59.39	65.69
-10	<b>73.02</b>	77.60	<b>52.15</b>	62.05	<b>59.64</b>	65.59
-12	72.39	77.17	<b>52.15</b>	62.24	58.75	65.63
-14	72.71	76.79	51.9	<b>62.34</b>	59.13	<b>66.04</b>

**Penalty Factor  $\lambda$ .** To analyze the effect of the penalty factor on the results, we perform experiments on the Flickr30K dataset, using ALBEF as the surrogate model. Various values of  $\lambda$  are tested, as presented in Table 1. A penalty factor that is too small results in difficulties in reducing the similarity between positive image-text pairs, while a penalty factor that is too large makes it challenging to increase the similarity between negative pairs, ultimately leading to performance degradation. Overall, the highest attack success rate for transfer attacks is achieved when  $\lambda = -10$ . Our experiments are conducted on a single NVIDIA A800 80GB graphics card. Depending on the size of the dataset, the time for generating adversarial examples ranges from one hour to two hours.

## G Adversarial Transferability on Multimodal Large Language Models

Recently, Multimodal Large Language Models (MLLMs) have made exciting advancements, demonstrating exceptional capabilities in tackling a variety of complex downstream tasks by leveraging multiple modalities. To further investigate the adversarial robustness of MLLMs, we conduct extended experiments on state-of-the-art commercial MLLMs. Specifically, we adhere to the setup proposed in the DRA, utilizing ALBEF as a surrogate model to generate adversarial examples under the following conditions: an image perturbation of 16/255, a single-step perturbation of 0.5/255, and a step size of 500. These adversarial examples are then employed to challenge GPT-4o and Claude-3.7 Sonnet. As depicted in Figure 2 and Figure 3, the adversarial examples generated by our method successfully mislead these state-of-the-art MLLMs into producing erroneous responses.

We also conduct additional experiments to evaluate the transferability of HQA-VLAttack on Qwen2.5-VL (Qwen2.5-VL-32B-Instruct) and LLaVA-Next (llava-v1.6-34b-hf) under the image captioning task and visual question answering task. Specifically, we use ALBEF as the surrogate model to generate adversarial examples, with a perturbation budget of 16/255, 80 PGD steps, and a step size of 0.5/255. For the image captioning task, we randomly sample 128 image-text pairs from the Flickr30K dataset and generate adversarial images based on these clean pairs. Then we feed these adversarial images into Qwen2.5-VL and LLaVA-Next to generate captions. The generated captions are evaluated with the widely-used metrics BLEU-4, METEOR, ROUGE-L, CIDEr, and SPICE, where lower values indicate better attack performance. For the visual question answering task, we randomly sample 128 examples from the VQAv2 dataset and apply perturbations to both the input images and their corresponding questions. The adversarial images and adversarial questions are then jointly fed into Qwen2.5-VL and LLaVA-Next to generate answers. The generated answers are evaluated using the widely used metric accuracy, where a lower accuracy indicates better attack performance. The detailed results are shown in Table 2.

These results show that HQA-VLAttack substantially degrades the performance on both Qwen2.5-VL and LLaVA-Next, which consistently outperforms SGA and DRA across all evaluation metrics. This further demonstrates the strong transferability and effectiveness of our method against modern MLLMs.

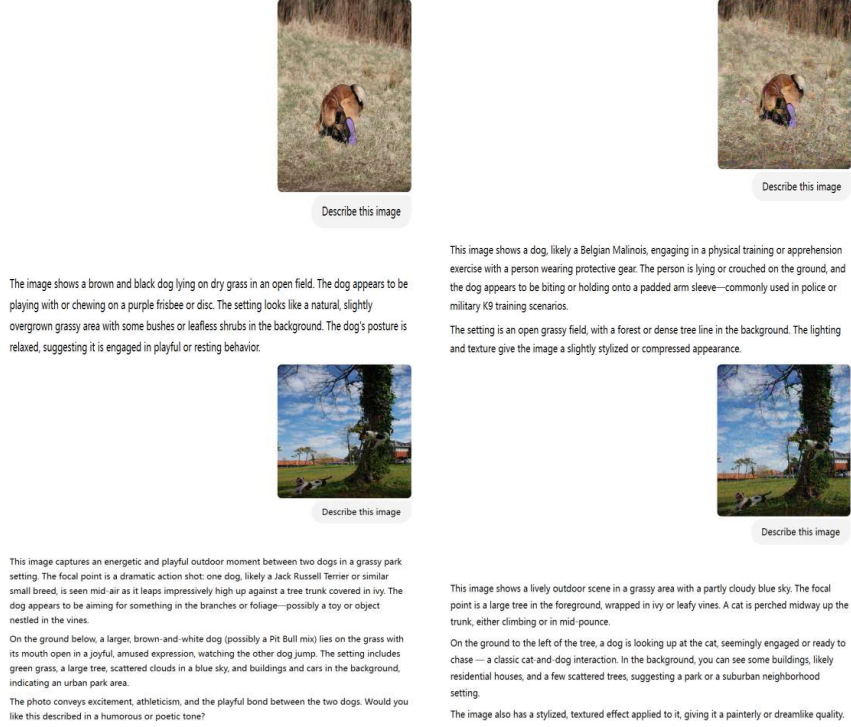


Figure 2: Adversarial Transferability on GPT-4o. The images on the left illustrate the responses produced by GPT-4o when it is provided with unaltered images and the prompt “Describe this image”. In contrast, the images on the right depict the responses generated in response to images that have been modified with adversarial perturbations.

Table 2: The Original represents the performance on clean data. We utilize ALBEF to generate multi-modal adversarial examples for attacking Image Captioning (IC) and Visual Question Answering (VQA).

Target Model	Method	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE	Accuracy
Qwen2.5-VL	Original	13.1	20.9	39.9	49.9	14.3	59.64
	SGA	12.3	18.6	37.6	42.5	12.6	45.31
	DRA	10.3	18.2	36.0	36.7	11.6	47.40
	HQA-VLAttack	<b>7.4</b>	<b>18.1</b>	<b>33.9</b>	<b>25.3</b>	<b>11.4</b>	<b>39.58</b>
LLaVA-Next	Original	14.1	15.7	33.2	42.5	10.6	75.78
	SGA	5.8	13.8	30.1	30.8	9.9	46.35
	DRA	5.8	13.3	28.8	28.7	9.5	50.00
	HQA-VLAttack	<b>4.8</b>	<b>11.9</b>	<b>27.8</b>	<b>23.7</b>	<b>7.8</b>	<b>41.93</b>

## H Time Cost Comparison

Table 3 provides a detailed comparison of the computational costs associated with our method, SGA, and DRA. Here,  $t_I$  and  $t_T$  denote the time required to generate adversarial images and texts, respectively. Our approach incurs an overall computational overhead of 1.58 to 2.32 times that of SGA, with the image generation component taking 0.79 to 1.18 times as long. In contrast, when compared to DRA, our method’s total computational time ranges from 0.73 to 1.78 times, with the adversarial image generation phase being particularly efficient at 0.43 to 0.48 times the duration of DRA. It is evident that our method does not significantly increase the time consumption for generating adversarial images. However, the generation of adversarial text is considerably more time-intensive. This additional time is primarily attributed to the process of selecting the optimal samples from a larger pool of candidate adversarial texts. In future work, we aim to explore more

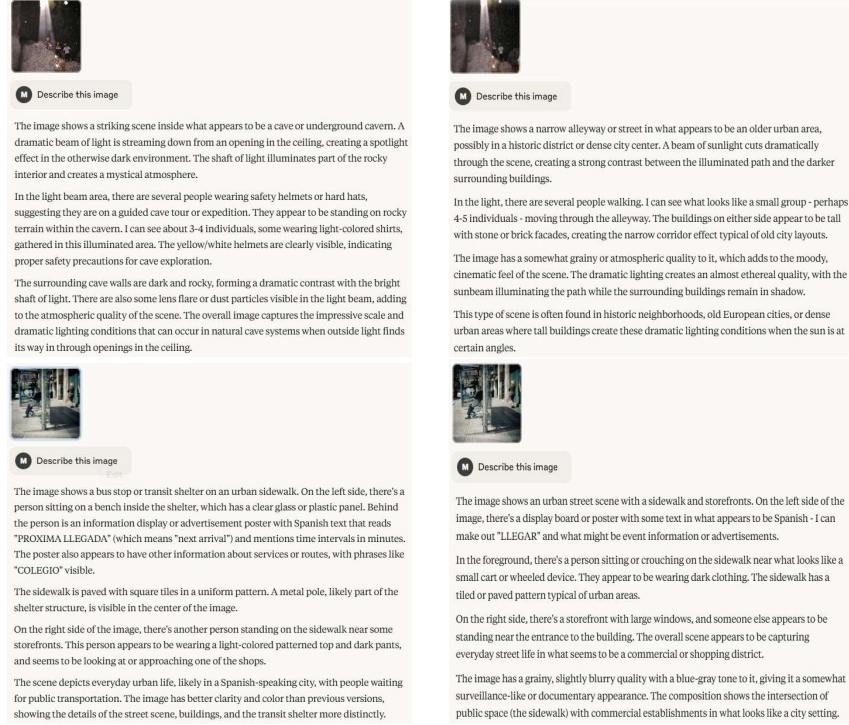


Figure 3: Adversarial Transferability on Claude-3.7 Sonnet. The images on the left illustrate the responses produced by Claude-3.7 Sonnet when it is provided with unaltered images and the prompt “Describe this image”. In contrast, the images on the right depict the responses generated in response to images that have been modified with adversarial perturbations.

Table 3: Computational cost comparison with SGA and DRA.

Attack	ALBEF		TCL		CLIP <sub>ViT</sub>		CLIP <sub>CNN</sub>	
	$t_I + t_T$	$t_I$	$t_I + t_T$	$t_I$	$t_I + t_T$	$t_I$	$t_I + t_T$	$t_I$
SGA	5.62	4.55	5.52	4.48	4.54	2.31	3.47	1.49
DRA	12.16	11.11	12.13	11.15	4.99	3.85	4.52	3.31
HQA-VLAttack	8.88	5.39	8.86	5.31	8.33	1.83	8.07	1.44

efficient methodologies for selecting the best adversarial texts, thereby reducing the computational overhead.

Beyond the time consumption for generating adversarial examples, we now present the computational complexity analysis as follows.

Assume that  $D$  is the feature dimension,  $L$  is the text length,  $C_{\text{mlm}}$  is the cost of one forward pass through a masked language model,  $C_I$  and  $C_T$  are the forward pass costs for the image encoder and text encoder,  $G_I$  is the backward pass cost through the image encoder,  $L_p$  is the number of image encoder layers,  $D_p$  is the number of tokens per layer,  $k$  is the number of substitute candidates per word,  $T_p$  and  $T_n$  are the number of positive and negative captions per batch,  $S = |\text{Trans}(v'_i)|$  is the number of image scales used during contrastive learning, and  $N$  is the number of attack iterations.

**Text attack.** For each word, we retrieve substitutes using either a counter-fitting synonym dictionary (constant-time) or an MLM ( $C_{\text{mlm}}$ ). With the hit ratio  $\rho$ , the complexity of this step is  $O(LC_{\text{mlm}}(1 - \rho))$ . We then evaluate  $kL$  candidate texts by computing their similarity with a fixed image feature, yielding a total cost of  $O(kLC_T)$ . Therefore, the overall time complexity of the text attack is  $O(LC_{\text{mlm}}(1 - \rho) + kLC_T)$ .

**Image attack.** Section 4.2.1 consists of determining layer importance and generating initial adversarial images. The first step requires one forward pass and  $L_p$  vector comparisons. The second step

needs to be repeated over  $N$  iterations, including a forward pass, a backward pass, and layer-wise feature comparisons. Ignoring minor terms, the total complexity is  $O(N(G_I + C_I + L_p D_p D))$ .

Section 4.2.2 applies contrastive learning over multiple image scales. At each iteration, the adversarial image is compared with  $T_p$  positive and  $T_n$  negative texts across  $S$  scales. This involves  $S$  forward passes through the image encoder and  $T_p + T_n$  forward passes through the text encoder, plus one backward pass. Ignoring minor terms, the total complexity is  $O(N(SC_I + (T_p + T_n)C_T + G_I))$ .

Therefore, the total time complexity of the image attack is  $O(N(G_I + C_I + L_p D_p D + SC_I + (T_p + T_n)C_T))$ .

Based on the above analysis, the total complexity for each image-text pair is  $O(LC_{\text{mlm}}(1 - \rho) + kLC_T + N(G_I + C_I + L_p D_p D + SC_I + (T_p + T_n)C_T))$ .

## I Analysis of an Alternative Negative Sampling Strategy

To further verify that our method does not rely on a specific dataset distribution, we conduct the following experiments. Specifically, we pair each adversarial image from the Flickr30K dataset with randomly sampled captions from the MSCOCO dataset. These captions are irrelevant to the source images and come from a totally different dataset. We consider both image-to-text retrieval (TR) and text-to-image retrieval (IR) tasks. We use the attack success rate of R@1 as the evaluation metric for both TR and IR tasks, where a higher attack success rate indicates better attack quality. For reference, the original results of SGA and DRA under the same evaluation setting are also included in the table below.

Table 4: **Attack success rate (%) in image-text retrieval on the Flickr30K dataset.** The negative captions are randomly sampled from MSCOCO dataset.

Flickr30K Dataset									
Surrogate Model	Victim Model	ALBEF		TCL		CLIP <sub>ViT</sub>		CLIP <sub>CNN</sub>	
	Attack Method	TR R@1	IR R@1	TR R@1	IR R@1	TR R@1	IR R@1	TR R@1	IR R@1
ALBEF	SGA	-	-	45.42	55.25	33.38	44.16	34.93	46.57
	DRA	-	-	49.74	58.83	39.14	48.39	41.38	51.66
	HQA-VLAttack	-	-	<b>71.44</b>	<b>77.02</b>	<b>51.90</b>	<b>62.34</b>	<b>59.49</b>	<b>65.87</b>
TCL	SGA	48.91	60.34	-	-	33.87	44.88	37.74	48.30
	DRA	51.09	61.79	-	-	40.25	48.94	42.91	52.49
	HQA-VLAttack	<b>62.93</b>	<b>72.20</b>	-	-	<b>52.12</b>	<b>59.38</b>	<b>57.09</b>	<b>62.34</b>
CLIP <sub>ViT</sub>	SGA	13.40	27.22	16.23	30.76	-	-	38.76	47.79
	DRA	12.51	30.00	14.65	30.62	-	-	45.47	50.74
	HQA-VLAttack	<b>25.23</b>	<b>42.04</b>	<b>24.24</b>	<b>43.45</b>	-	-	<b>74.46</b>	<b>77.23</b>
CLIP <sub>CNN</sub>	SGA	11.42	24.80	14.91	28.82	31.24	42.12	-	-
	DRA	12.20	26.59	14.33	29.29	35.21	45.94	-	-
	HQA-VLAttack	<b>21.07</b>	<b>38.79</b>	<b>21.92</b>	<b>41.88</b>	<b>62.82</b>	<b>69.62</b>	-	-

As shown in Table 4, we can see that HQA-VLAttack consistently outperforms other strong baselines even when negative captions are randomly sampled from a totally different dataset.

## J Analysis of Iteration Steps for Baselines

We investigate the influence of iteration steps on the attack performance of two strong baselines, SGA and DRA, by varying the number of iteration steps to 15, 20, and 25 on the Flickr30K dataset. Specifically, we use ALBEF as the surrogate model to attack  $CLIP_{ViT}$  on both image-to-text retrieval (TR) and text-to-image retrieval (IR) tasks, and vice versa (i.e., using  $CLIP_{ViT}$  as the surrogate model to attack ALBEF). We use the attack success rate of R@1 as the evaluation metric for both TR and IR tasks, where a higher attack success rate indicates better attack quality. The detailed results are shown in Table 5.

Based on the above results, we can observe that increasing the number of steps does not lead to consistent performance improvement for SGA and DRA. And in some cases, the performance even slightly drops, which is probably due to the overfitting to the surrogate model. Moreover, HQA-VLAttack with only 10 steps outperforms SGA and DRA even when their iteration steps are increased to 25. This indicates that our chosen step setting does not disadvantage the baseline methods and highlights the superior efficiency and effectiveness of HQA-VLAttack.



Table 5: Attack success rate (%) of SGA and DRA with varying iteration steps on the Flickr30K dataset.

Method	Steps	ALBEF $\rightarrow$ CLIP <sub>VIT</sub>		CLIP <sub>VIT</sub> $\rightarrow$ ALBEF	
		TR R@1	IR R@1	TR R@1	IR R@1
SGA	10	33.38	44.16	13.40	27.22
	15	36.20	44.17	12.10	27.06
	20	36.20	44.59	11.68	25.98
	25	36.32	43.07	12.41	26.01
DRA	10	39.14	48.39	12.51	30.00
	15	39.14	48.74	13.45	29.51
	20	38.40	48.32	14.70	29.98
	25	38.90	48.74	14.08	29.44
HQA-VLAttack	10	<b>52.15</b>	<b>62.05</b>	<b>25.13</b>	<b>41.98</b>

## K Explanation and Analysis of Input Transformation

The input transformation aims to improve the generalization ability of the adversarial examples, and it has been used in previous baselines SGA and DRA, which ensures that the performance gains come from our method itself. The details of the scale transformation function are as follows. For a given image  $I \in \mathbb{R}^{C \times H \times W}$ , we perform the following steps:

Scaling: Resize the image to  $rH \times rW$  using predefined ratios  $r \in \{0.50, 0.75, 1.00, 1.25, 1.50\}$ .

Noise injection: Gaussian noise  $\mathcal{N}(0, \sigma^2)$  can be added before resizing. In our experiments, we set  $\sigma = 0.05$ .

Resizing: Each scaled image is resized back to the original resolution via bicubic interpolation.

Contrastive integration: The resulting set  $\text{Trans}(v'_i)$  is used in the contrastive learning loss to promote feature invariance across scales.

We include the ablation study for the input transformation and the contrastive learning. Specifically, we use ALBEF as the surrogate model on the Flickr30K dataset. Adversarial examples are generated with a perturbation budget of 2/255, using 10 PGD steps and a step size of 0.5/255. HQA-VLAttack (o, o) means the method without the input transformation and contrastive learning. HQA-VLAttack (o, w) means the method without the input transformation, but with contrastive learning. HQA-VLAttack (w, w) means the method with input transformation and contrastive learning, i.e., the original method. The results are shown in Table 6, where higher values indicate better performance.

Table 6: Ablation study of input transformation and contrastive learning for HQA-VLAttack.

Method	ALBEF $\rightarrow$ CLIP <sub>VIT</sub>		ALBEF $\rightarrow$ CLIP <sub>CNN</sub>	
	TR R@1	IR R@1	TR R@1	IR R@1
HQA-VLAttack (o,o)	41.32	52.80	46.49	57.63
HQA-VLAttack (o,w)	49.08	59.50	55.94	63.57
HQA-VLAttack (w,w)	<b>52.15</b>	<b>62.05</b>	<b>59.64</b>	<b>65.59</b>

The results show that both the input transformation and the contrastive learning can affect the performance to some extent.

## L Limitations

While our method demonstrates notable advancements, several aspects offer opportunities for further enhancement. First, although the proposed approach consistently achieves superior transfer attack success rates compared to existing strong baselines, the performance gap remains more evident when transferring adversarial examples from weaker to significantly stronger models. This suggests potential for future refinement in boosting cross-model generalization. Second, the current theoretical

analysis, though providing initial insights into the mechanism of our method, is still at an early stage. We envision a more comprehensive theoretical framework in future work to deepen the understanding of the underlying principles and guide further methodological improvements.

## **M Broad Impact**

This work highlights the susceptibility of Vision-Language Pre-training (VLP) models to black-box adversarial attacks, raising important considerations regarding their deployment in safety-critical or real-world applications. Although HQA-VLAttack could, in theory, be exploited to target multimodal systems, its primary intent is to raise awareness of potential vulnerabilities and to promote proactive development of more secure and resilient VLP architectures. By contributing to the understanding of adversarial robustness in multimodal settings, we aim to support the broader goal of building trustworthy and dependable AI systems.

**Safeguard Statement** We recognize the potential for misuse associated with the proposed HQA-VLAttack and acknowledge the ethical implications such vulnerabilities may pose to the credibility and security of multimodal AI systems. To address these concerns, we are committed to investigating and advancing effective multimodal adversarial defense strategies. Our long-term objective is to encourage responsible research practices and to support the development of defense techniques that reinforce the safety and robustness of future AI deployments.