# Boosting Transferability in Vision-Language Attacks via Diversification along the Intersection Region of Adversarial Trajectory

Sensen Gao [1,3] ⋆, Xiaojun Jia [2*†], Xuhong Ren [2],
Ivor Tsang [2,3], and Qing Guo [3†]

[1] Nankai University, Tianjin, China
[2] Nanyang Technological University, Singapore
[3] CFAR and IHPC, Agency for Science, Technology and Research (A*STAR), Singapore

In this supplementary material, we present our visualizations of multimodal data perturbations and the formulation of attacks. Additionally, we delve into the intersection region of adversarial trajectories and the rationale behind the improvement in transferability. Moreover, we explore adversarial attacks on Large Language Models (LLMs) and compare the time efficiency. Furthermore, we provide more detailed results regarding cross-model transferability.

- In Section 1, we present visualizations of the perturbations caused by our multimodal adversarial attack methods on both image and text modalities.
- In Section 2, we present attack formulation, which outlines the step-by-step process of generating adversarial images and texts.
- In Section 3, we explore the proportions of clean images, last adversarial images, and adversarial images in the intersection regional optimal samples, as well as the changes in these proportions during the iteration process.
- In Section 4, we discuss the rationale behind the transferability improvement of our method compared to SGA.
- In Section 5, We employ the VLP models as surrogates to generate adversarial examples and explore the adversarial robustness of existing state-of-the-art LLMs.
- In Section 6, we compare the time costs between SGA and our method.
- In Section 7, we present more detailed results about cross-model transferability, incorporating the R@5 and R@10 metrics, as well as extending the analysis to the MSCOCO dataset and different perturbation settings.

## 1 Visualization on Multimodal Dataset

In Figure A1, we present the adversarial perturbations of our method for both image and text modalities. As mentioned in our experimental setup, we perturb only one token for text and use an $8/255$ perturbation for images. The first two lines represent the clean image and the adversarial image, respectively. The third line indicates the added adversarial perturbation. The last two lines represent

⋆ co-first authors; † co-corresponding authors

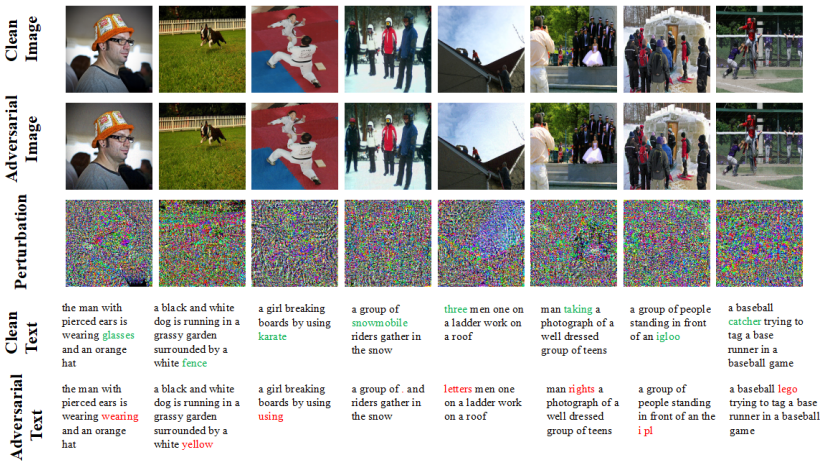


**Fig. A1: Adversarial perturbation visualization on Multimodal dataset.**

the clean text and the adversarial text, where green highlights the original words, while red highlights the words modified in the adversarial text.

## 2 Attack Formulation

In Algorithm 1, we provide a detailed demonstration of the process by which our method generates adversarial images and adversarial texts.

## 3 Intersection Regional Optimal Sample

The current state-of-the-art method, namely SGA (Set-level Guidance Attack) [6], focuses on improving the transferability of multimodal adversarial example transferability by exploiting the diversity of adversarial examples along the optimization path. Despite achieving notable success, this approach primarily considers the diversity surrounding adversarial examples within the optimization trajectory. It is crucial to note that adversarial examples generated iteratively by the surrogate model pose a potential risk of overfitting to the victim model, thereby impacting transferability.

To address the potential overfitting risk, we undertake a groundbreaking initiative to enhance the transferability of multimodal adversarial attacks by incorporating diversity not only around online adversarial examples (AEs) throughout the optimization process but also around clean inputs. Specifically, we introduce a novel concept termed the intersection region of adversarial trajectory, which encompasses the original image, the adversarial image from the previous step, and the current adversarial image during the iterative attack process. This pioneering approach aims to mitigate overfitting by strategic sampling within

**Algorithm 1:** Attack formulation.

**Data:** Encoder $F_I, F_T$, perturbation bound $\epsilon_v, \epsilon_t$, Image-caption $(v, t)$, Iteration steps $T$, Sample number $N$, Augmentation number $M$, Image scale sets $S = \{s_1, s_2, ..., s_M\}$

**Output:** adversarial image $v'$, adversarial caption $t'$

Follow SGA method and get adversarial image $v'_0, v'_1$

**for** $i = 2, 3, ..., T$ **do**

  /* Build sample set $e = \{e_1, e_2, ..., e_N\}$ */

  **for** *iter* $k = 1, 2, ..., N$ **do**

    $e_k = \beta \cdot v + \gamma \cdot v'_{i-1} + \eta \cdot v'_i, \beta + \gamma + \eta = 1.0$

  **end**

  /* Build perturbation set $P = \{p_1, p_2, ..., p_N\}$ */

  **for** *iter* $k = 1, 2, ..., N$ **do**

    $p_k = \alpha \cdot sign(\frac{\nabla_e J(F_I(e_k), F_T(t))}{\|\nabla_e J(F_I(e_k), F_T(t))\|})$.

  **end**

  /* Get optimal sample $e_m$ */

  $m = \underset{p_m \in P}{\arg\max} J(F_I(v'_i + p_m), F_T(t))$.

  /* Build augmentation set $E_m = \{e_{m1}, e_{m2}, ..., e_{mM}\}$ */

  **for** *iter* $j = 1, 2, ..., M$ **do**

    $e_{mj} = resize(e_m, s_j) + 0.05 \cdot \boldsymbol{N}(0, 1)$

  **end**

  /* Update adversarial image $v'_i$ */

  $v'_{i+1} = v'_i + \alpha \cdot sign(\frac{\nabla_e \sum_{j=1}^{M} J(F_I(e_{mj}), F_T(t))}{\|\nabla_e \sum_{j=1}^{M} J(F_I(e_{mj}, F_T(t))\|})$.

**end**

/* Generate adversarial text $t'$ */

$t' = \underset{t' \in B[t, \epsilon_t]}{\arg\max}(\lambda \cdot J(F_I(v), F_T(t')) + \mu \cdot J(F_I(v'_T), F_T(t'))) + \nu \cdot J(F_I(v'_{T-1}), F_T(t')))$.

this region, thereby avoiding an exclusive focus on adversarial example diversity solely around adversarial images. Upon acquiring multiple sample points, we compute gradients for each, allowing the determination of a perturbation direction away from the text. Subsequently, we incorporate these perturbations individually into the current adversarial image and select the perturbation that deviates the most from the text.

In the iterative process, the intersection region is defined by the original image $v$, the adversarial image from the previous step $v'_{i-1}$, and the current adversarial image $v'_i$. Each sample within this region can be represented as $e_k = \beta \cdot v + \gamma \cdot v'_{i-1} + \eta \cdot v'_i$, where $\beta + \gamma + \eta = 1$. Our objective is to delve into the specific locations within this region where optimal samples occur at each update. This involves a detailed examination of the precise magnitude and variation of the three numerical values, $\beta$, $\gamma$, and $\eta$.

To achieve the aforementioned objectives, we employ ALBEF [4], TCL [11], and CLIP [9] as surrogate models on the Flickr30K [8] dataset to generate multi-

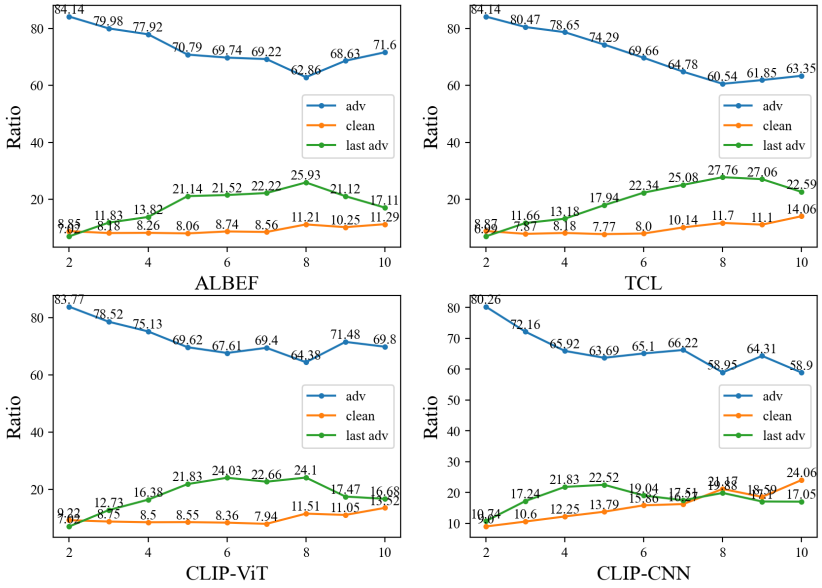


**Fig. A2: Intersection Regional Optimal Sample**. ALBEF, TCL, and CLIP are sequentially used as source models. During the iteration process, the optimal sample from the intersection region of adversarial trajectory is characterized by the proportions of the clean image, the last adversarial image, and the current adversarial image.

modal adversarial examples, respectively. The average value is computed across all examples. Given that we set the PGD [7] step size to 10 and utilize random noise initialization, there will be 9 iterations of sampling and selecting the optimal sample. The detailed results are illustrated in Figure A2.

First, let's analyze the numerical distribution of the three ratios: $\beta$, $\gamma$, and $\eta$. As depicted in Figure A2, it is evident that, for any VLP model and at each step, the proportion of adversarial images consistently remains the highest, typically surpassing 60%. Furthermore, in the optimal samples, the proportion of adversarial images $v_i'$ does not exceed 85%. In the case of SGA, it can be viewed as a special scenario within our method, wherein the proportion of adversarial images in the selected samples consistently reaches 100% in each iteration.

Due to an excessive focus on optimizing diversity around adversarial images and insufficient consideration of clean images in the optimization path, the SGA method tends to exhibit overfitting on the surrogate model, thereby limiting its transferability. In contrast, our approach introduces approximately 5-15% clean images at each iteration while selecting optimal samples, effectively mitigating this risk. Furthermore, within each optimal sample, there exists an additional proportion of 5-25% derived from the previous adversarial image. This inclusion of historical adversarial perturbations enriches the diversity of adversarial examples, thus effectively preventing the occurrence of overfitting.


**Fig. A3: Adversarial Transferability on GPT-4.** The images on the left display the responses generated by GPT-4 when presented with **clean images** and the query "Describe this image", while the images on the right show the responses to the **adversarial images**.

Next, we delve into an examination of the dynamic shifts in the optimal sample proportions occupied by adversarial images, historical adversarial images, and clean images throughout the iteration process. The specific outcomes are depicted in Figure A2. Across the four VLP models, a discernible diminishing trend is observed in the proportion occupied by adversarial images. In the PGD multi-step iterative attack method, a single-step attack undergoes refinement through the incorporation of multiple successive perturbations. In the initial iterations, owing to the relatively modest accumulation of perturbations in the adversarial images, these perturbations fall below the constraint limits. Consequently, during these early steps, the optimal samples tend to align closely with the adversarial images. As the perturbations in the adversarial images continue to accrue, they eventually surpass the imposed constraints. The portions exceeding these constraints are subsequently adjusted to the constraint value,


**Fig. A4: Adversarial Transferability on Claude-3.** The images on the left display the responses generated by Claude-3 when presented with **clean images** and the query "Describe this image", while the images on the right show the responses to the **adversarial images**.

*e.g.*, 8/255. At this juncture, an excessive reliance on adversarial images inevitably leads to a decline in the effectiveness of the attack. However, judicious consideration of the adversarial perturbations around clean images and the most recent adversarial image helps mitigate the risk of overfitting. Furthermore, our analysis extends to the changing trend in the proportion of clean images within the optimal samples. Notably, for the $\text{CLIP}_{\text{CNN}}$ model, there is a conspicuous upward pattern, while the other three VLP models also exhibit an upward trend, albeit at a very slow pace. Across all VLP models, the proportion of the last adversarial image shows a slow increase over the first 7 or 8 steps, followed by a decrease in the final two or three steps.

## 4 Rationale behind transferability improvement

The main reason for the low transferability of an adversarial example(AE) is that it heavily relies on the surrogate vision-language model, also known as the overfitting issue.

**Why is SGA effective?** SGA conducts augmentations around the online AEs (See Figure 1 (a)) to improve transferability "instead of increasing the over-

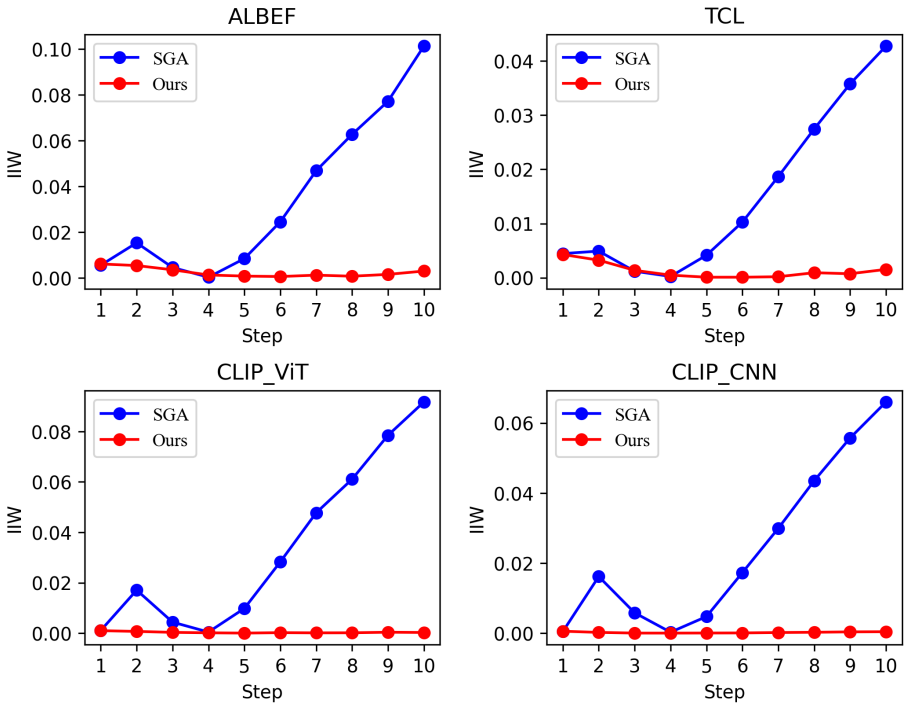


**Fig. A5: Overfitting analysis of SGA & our method via IIWs.**

fitting risk". However, the augmentations still highly rely on the examples along the adversarial trajectory, leading to overfitting risks.

**Why is our method effective?** Unlike SGA, our method performs augmentations around online AEs considering the raw clean example and the AE at the previous step (See Figure 1 (b)). We utilize the clean example since it is entirely independent of the surrogate model and offers the potential to mitigate the risk of overfitting.

**Overfitting analysis of SGA & our method via IIWs.** To quantify overfitting risk, we use the PAC-Bayes theorem to measure the amount of information stored in the network's weights (IIW) [10], a promising indicator of generalization ability. Smaller IIW means lower overfitting risks. Given an AE from SGA or our method during optimization iteration, we calculate its IIWs by feeding it to four VLMs, respectively. We count IIWs of 1000 AEs during optimization and show the averaging results in Figure A5. During optimization, the IIW of AEs from the SOTA baseline (*i.e.*, SGA) initially decreases before sharply increasing. In contrast, our method leads to a continuous low IIW value in generated AEs, demonstrating that our method can overcome the overfitting risks. As a result, the attack transferability can be enhanced.

| Attack | GPT-4 | | | Claude-3 | | | Qwen-VL | | |
|---|---|---|---|---|---|---|---|---|---|
| | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 |
| No Attack | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| SGA | 6.0 | 0.0 | 0.0 | 18.0 | 4.0 | 3.0 | 6.0 | 2.0 | 0.0 |
| Ours | 16.0 | 6.0 | 4.0 | 22.0 | 12.0 | 9.0 | 15.0 | 6.0 | 1.0 |

**Table A1: Attacking against open VLP large models.**

## 5 Adversarial Transferability on Large Language Models

Recently, Large Language Models (LLMs) have garnered widespread applications. Their utilization spans various domains, showcasing their versatility and efficacy in tackling complex tasks. To further explore the adversarial robustness of LLMs, we conduct specialized extension experiments. Specifically, we utilize ALBEF as a surrogate model, generating adversarial examples under settings with perturbation constraints of $16/255$ and single-step perturbation of $0.5/255$, with a step size of 500. Furthermore, we evaluate them on advanced LLMs (*e.g.*, GPT-4 [1], Claude-3) by prompting these systems with the query "Describe this image". The results, depicted in Figure A3 and A4, demonstrate that adversarial examples generated by our method can mislead today's state-of-the-art LLMs into producing incorrect answers.

We have shown two cases on open VLP large models above. Here, we further extend experiments by randomly selecting 100 images from the Flick30K and generating adversarial images using SGA and our method with ALBEF as the surrogate, under the same perturbation setting ($16/255$). Then, we feed adversarial images into LLMs and query "Describe this image", and rank descriptions with 100 irrelevant texts based on CLIP similarity. If the description doesn't rank top-N, consider a successful R@N attack. Table A1 illustrates ASR(%), showing our superior transferability to LLMs compared to SGA.

## 6 Time cost comparison.

Table A2 compares the time costs per AE between SGA and our method. $t_I$ and $t_T$ respectively represent the time for generating adversarial images and text. Our method takes 1.58-2.28 times longer overall compared to SGA, with generating adversarial images taking 2.33 to 2.59 times as long. This additional time primarily arises from the process of selecting optimal samples in the intersection region, involving a sample number of additional forward/backward passes compared to SGA. In the future, we will explore methods to select optimal samples to reduce computational costs.

## 7 Detailed Results about Cross Model Transferability

We conduct experiments with a focus on assessing the transferability of adversarial examples across two widely adopted VLP model architectures: fused

| Attack | ALBEF | | TCL | | $\text{CLIP}_{\text{ViT}}$ | | $\text{CLIP}_{\text{CNN}}$ | |
|---|---|---|---|---|---|---|---|---|
| | $t_I + t_T$ | $t_I$ | $t_I + t_T$ | $t_I$ | $t_I + t_T$ | $t_I$ | $t_I + t_T$ | $t_I$ |
| SGA | 2.02 | 1.66 | 1.99 | 1.64 | 1.02 | 0.56 | 0.78 | 0.36 |
| Ours | 4.60 | 4.27 | 4.53 | 4.21 | 1.85 | 1.45 | 1.23 | 0.84 |

**Table A2: Computational cost compared with SGA.**

and aligned. The selected VLP models for evaluation include ALBEF, TCL, $\text{CLIP}_{\text{ViT}}$, and $\text{CLIP}_{\text{CNN}}$ [2]. The chosen downstream vision-and-language (V+L) task for assessment is image-text retrieval [3]. Our methodology involves utilizing one of the four models to generate multimodal adversarial examples within the specified parameters of our proposed method. Subsequently, we aim to validate the effectiveness of the generated adversarial examples through a comprehensive set of experiments. These experiments encompass both self-attacks, where the model attacks itself, and attacks on the other three models. The evaluation framework includes a white-box attack and three black-box attacks.

In the image-text retrieval task, we employ key metrics—R@1, R@5, and R@10 to gauge the efficacy of our attacks. These metrics assess whether the adversarial examples fail to appear in the top-ranked positions during the retrieval process, serving as crucial indicators of attack success. Due to space constraints in the main text, we focus on presenting the R@1 metric for cross-model transferability experiments. In Table A3, we furnish a comprehensive overview of the R@1, R@5, and R@10 metrics specifically for the Flickr30k dataset. Our analysis delves into the performance of the source model and the transferability across models separately.

We can initially observe that our method maintains the highest performance in terms of the R@1 metric compared to SGA, regarding the success rate of attacks on the source model. However, for the R@5 and R@10 metrics, our method slightly underperforms compared to SGA. However, when transferring the adversarial examples generated on the source model to the other three VLP models, our method achieves a substantially higher attack success rate than SGA, reaching up to 9.96%. Whether for Image Retrieval or Text Retrieval, our method outperforms SGA significantly across all three metrics: R@1, R@5, and R@10. These findings collectively demonstrate that our method is capable of generating adversarial examples with superior transferability.

When validating the effectiveness of our method in improving the transferability of cross VLP models on Flickr30K [8], we utilize a test set consisting of 1,000 images, with approximately 5 captions for each image. To further validate the reliability of our method more extensively, we also employ a larger MSCOCO [5] test set for evaluation. This set comprises 5,000 test images, with each image having approximately 5 captions as well. The attack settings on MSCOCO are identical to those on Flickr30K, and detailed experimental results can be found in Table A4.

In our analysis of experimental results on the MSCOCO dataset, we observe that our method achieves optimal performance in terms of attack success rates at R@1, R@5, and R@10. This holds when transferring to different VLP models. Specifically, the transferability of our method, even across fused VLP models, reaches up to 95.89%. Furthermore, similar to the results on the Flickr30K dataset, in the white-box attack setting, the success rate of our method is slightly lower than that of SGA, but it still reaches a very high level, averaging over 99%. However, the transferability of the adversarial examples generated by our method far surpasses that of SGA.

In the experiments conducted on the Flickr30K and MSCOCO datasets, we set the perturbation constraint to $8/255$ and the single step size to $2/255$, with a total of 10 steps. To further explore the effectiveness of our method under different perturbation settings, we compare it with existing state-of-the-art methods using another perturbation configuration. We utilize a reduced perturbation configuration as outlined in the SGA paper, with a constraint of $2/255$, employing a single step size of $0.5/255$, and conducting a total of 10 steps. The results can be seen in Table A5 and A6. Similar to the limitation at $8/255$, our method achieves a slightly lower success rate on the source model compared to SGA, but still at a significantly high level. Furthermore, the transferability of adversarial examples generated by our method is much better than our baseline method, SGA. Our experimental results under different perturbations and datasets thoroughly demonstrate the effectiveness of our method in generating multimodal adversarial examples with higher transferability.

**Table A3: Comparison with state-of-the-art methods on Flickr30K dataset.** The source column shows the VLP models we use to generate multimodal adversarial examples. The gray area represents adversarial attacks under a white-box setting, the rest are black-box attacks. For both Image Retrieval and Text Retrieval, we provide R@1, R@5 and R@10 attack success rate(%). The constraint for adversarial perturbation is set to 8/255, with a per-step size of 2/255, and a total of 10 steps.

**Flickr30K (Image-Text Retrieval)**

| Source | Attack | ALBEF | | | TCL | | | $\text{CLIP}_{\text{ViT}}$ | | | $\text{CLIP}_{\text{CNN}}$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 |
| **ALBEF** | PGD | 93.74 | 85.97 | 82.5 | 24.03 | 10.15 | 7.62 | 10.67 | 3.12 | 1.02 | 14.05 | 3.7 | 2.47 |
| | BERT-Attack | 11.57 | 1.80 | 1.10 | 12.64 | 2.51 | 0.90 | 29.33 | 11.63 | 6.30 | 32.69 | 15.43 | 8.65 |
| | Sep-Attack | 95.72 | 90.18 | 87.2 | 39.3 | 18.49 | 12.32 | 34.11 | 14.12 | 8.54 | 35.76 | 16.7 | 10.5 |
| | Co-Attack | 97.08 | 94.59 | 92.6 | 39.52 | 20.4 | 14.53 | 29.82 | 11.73 | 6.61 | 31.29 | 11.73 | 5.77 |
| | SGA | **99.9** | **99.9** | **99.9** | 87.88 | 77.99 | 71.84 | 36.69 | 19.31 | 14.13 | 39.59 | 21.99 | 15.96 |
| | **Ours** | **99.9** | 99.8 | 99.8 | **91.57** | **81.91** | **76.35** | **46.26** | **24.3** | **17.17** | **49.55** | **28.75** | **20.6** |
| **TCL** | PGD | 35.77 | 21.74 | 16.2 | 99.37 | 98.59 | 97.7 | 10.18 | 2.91 | 1.42 | 14.81 | 4.55 | 2.37 |
| | BERT-Attack | 11.89 | 2.20 | 0.70 | 14.54 | 2.31 | 0.60 | 29.69 | 12.77 | 7.62 | 33.46 | 14.38 | 9.37 |
| | Sep-Attack | 52.45 | 32.26 | 24.7 | 99.58 | 98.69 | 98.0 | 37.06 | 16.1 | 9.45 | 37.42 | 15.86 | 11.12 |
| | Co-Attack | 49.84 | 27.45 | 60.36 | 91.68 | 85.23 | 80.96 | 32.64 | 13.4 | 6.81 | 32.06 | 14.27 | 8.14 |
| | SGA | 92.49 | 87.37 | 83.5 | **100.0** | **100.0** | **100.0** | 36.81 | 19.0 | 12.2 | 41.89 | 23.68 | 16.89 |
| | **Ours** | **95.2** | **91.38** | **89.6** | **100.0** | **100.0** | 99.9 | **47.24** | **26.48** | **18.9** | **52.23** | **30.44** | **22.66** |
| $\text{CLIP}_{\text{ViT}}$ | PGD | 3.13 | 0.4 | 0.3 | 4.43 | 1.01 | 0.2 | 69.33 | 45.59 | 36.99 | 13.03 | 3.81 | 1.75 |
| | BERT-Attack | 9.59 | 1.30 | 0.40 | 11.80 | 1.91 | 0.70 | 28.34 | 11.73 | 6.81 | 30.40 | 11.63 | 5.97 |
| | Sep-Attack | 7.61 | 1.3 | 0.5 | 10.12 | 1.81 | 0.8 | 76.93 | 55.35 | 44.61 | 29.89 | 11.52 | 5.87 |
| | Co-Attack | 8.55 | 1.5 | 0.5 | 10.01 | 2.01 | 0.7 | 78.53 | 57.42 | 45.53 | 29.5 | 11.42 | 6.08 |
| | SGA | 22.42 | 8.92 | 5.6 | 25.08 | 10.15 | 4.91 | **100.0** | **100.0** | **100.0** | 53.26 | 33.93 | 24.2 |
| | **Ours** | **27.84** | **12.22** | **7.7** | **27.82** | **11.76** | **7.11** | **100.0** | 99.9 | 99.9 | **64.88** | **42.6** | **31.72** |
| $\text{CLIP}_{\text{CNN}}$ | PGD | 2.29 | 0.3 | 0.3 | 4.53 | 0.3 | 0.1 | 5.4 | 1.45 | 0.81 | 89.78 | 77.7 | 70.75 |
| | BERT-Attack | 8.86 | 1.50 | 0.60 | 12.33 | 2.01 | 0.90 | 27.12 | 11.21 | 6.81 | 30.40 | 13.00 | 7.31 |
| | Sep-Attack | 9.38 | 1.2 | 0.7 | 11.28 | 1.91 | 0.7 | 26.13 | 11.63 | 6.3 | 93.61 | 84.36 | 78.17 |
| | Co-Attack | 10.53 | 1.6 | 0.4 | 12.54 | 2.01 | 0.7 | 27.24 | 12.05 | 6.5 | 95.91 | 89.75 | 85.99 |
| | SGA | 15.64 | 5.51 | 3.0 | 18.02 | 6.43 | 2.91 | 39.02 | 19.21 | 13.01 | **99.87** | **99.58** | **99.38** |
| | **Ours** | **19.5** | **6.31** | **3.2** | **21.6** | **7.64** | **3.71** | **48.47** | **26.38** | **17.07** | **99.87** | 99.47 | 99.07 |

**Flickr30K (Text-Image Retrieval)**

| Source | Attack | ALBEF | | | TCL | | | $\text{CLIP}_{\text{ViT}}$ | | | $\text{CLIP}_{\text{CNN}}$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 |
| **ALBEF** | PGD | 94.43 | 89.48 | 86.19 | 27.9 | 13.56 | 9.87 | 15.82 | 6.61 | 3.97 | 19.11 | 7.11 | 4.72 |
| | BERT-Attack | 27.46 | 14.48 | 10.98 | 28.07 | 14.39 | 10.26 | 43.17 | 26.37 | 19.91 | 46.11 | 28.43 | 22.14 |
| | Sep-Attack | 96.14 | 91.84 | 89.12 | 51.79 | 31.36 | 24.09 | 45.72 | 27.77 | 20.1 | 47.92 | 30.86 | 23.81 |
| | Co-Attack | 98.36 | 96.41 | 94.86 | 51.24 | 31.9 | 24.41 | 38.92 | 23.31 | 17.01 | 41.99 | 25.18 | 18.55 |
| | SGA | **99.98** | **99.94** | **99.94** | 88.05 | 77.53 | 70.96 | 46.78 | 28.66 | 21.96 | 49.78 | 32.27 | 24.96 |
| | **Ours** | 99.93 | 99.9 | 99.86 | **91.17** | **81.99** | **76.79** | **56.8** | **38.31** | **29.68** | **59.01** | **41.19** | **33.34** |
| **TCL** | PGD | 41.67 | 25.8 | 19.35 | 99.33 | 98.24 | 97.08 | 16.3 | 6.61 | 4.25 | 21.1 | 8.56 | 5.26 |
| | BERT-Attack | 26.82 | 14.09 | 10.80 | 29.17 | 15.03 | 10.91 | 44.49 | 27.47 | 21.00 | 46.07 | 29.28 | 22.59 |
| | Sep-Attack | 61.44 | 44.11 | 36.51 | 99.45 | 98.39 | 97.34 | 45.81 | 28.33 | 21.37 | 49.91 | 32.7 | 25.05 |
| | Co-Attack | 60.36 | 41.59 | 33.26 | 95.48 | 90.32 | 86.74 | 42.69 | 26.44 | 20.37 | 47.82 | 30.47 | 23.13 |
| | SGA | 92.77 | 86.94 | 83.64 | **100.0** | **100.0** | **99.98** | 46.97 | 29.41 | 22.81 | 51.53 | 33.21 | 25.1 |
| | **Ours** | **95.58** | **91.78** | **88.9** | 99.98 | 99.98 | 99.96 | **57.28** | **39.31** | **31.88** | **62.23** | **44.06** | **35.78** |
| $\text{CLIP}_{\text{ViT}}$ | PGD | 6.48 | 1.83 | 1.01 | 8.83 | 2.46 | 1.54 | 84.79 | 73.84 | 68.45 | 17.43 | 7.13 | 4.61 |
| | BERT-Attack | 22.64 | 10.95 | 8.17 | 25.07 | 12.92 | 8.90 | 39.08 | 24.08 | 17.44 | 37.43 | 24.96 | 18.66 |
| | Sep-Attack | 20.58 | 9.52 | 7.22 | 20.74 | 9.74 | 6.8 | 87.44 | 78.07 | 72.79 | 38.32 | 23.29 | 17.69 |
| | Co-Attack | 20.18 | 9.54 | 7.12 | 21.29 | 9.51 | 6.78 | 87.5 | 77.95 | 73.4 | 38.49 | 23.19 | 17.87 |
| | SGA | 34.59 | 18.25 | 13.61 | 36.45 | 18.79 | 13.87 | **100.0** | **100.0** | 99.98 | 61.1 | 43.18 | 35.37 |
| | **Ours** | **42.84** | **24.38** | **18.76** | **44.6** | **26.38** | **19.74** | **100.0** | **100.0** | **100.0** | **69.5** | **53.66** | **45.09** |
| $\text{CLIP}_{\text{CNN}}$ | PGD | 6.15 | 1.7 | 0.97 | 8.88 | 2.42 | 1.42 | 12.08 | 4.86 | 2.94 | 93.04 | 85.37 | 81.09 |
| | BERT-Attack | 23.27 | 11.34 | 8.41 | 25.48 | 13.25 | 8.81 | 37.44 | 23.48 | 17.66 | 40.10 | 26.71 | 20.85 |
| | Sep-Attack | 22.99 | 10.77 | 7.95 | 25.45 | 11.99 | 8.2 | 39.24 | 23.73 | 17.84 | 95.3 | 89.62 | 86.2 |
| | Co-Attack | 23.62 | 11.4 | 8.21 | 26.05 | 12.69 | 8.79 | 40.62 | 24.71 | 18.82 | 96.5 | 92.75 | 90.35 |
| | SGA | 28.6 | 15.26 | 10.82 | 33.07 | 17.22 | 12.33 | 51.45 | 33.29 | 25.77 | **99.9** | **99.81** | **99.64** |
| | **Ours** | **34.59** | **18.05** | **13.83** | **37.88** | **20.94** | **15.41** | **59.12** | **40.88** | **32.9** | **99.9** | 99.61 | 99.46 |

**Table A4: Comparison with state-of-the-art methods on MSCOCO dataset.** The source column shows the VLP models we use to generate multimodal adversarial examples. The gray area represents adversarial attacks under a white-box setting, the rest are black-box attacks. For both Image Retrieval and Text Retrieval, we provide R@1, R@5 and R@10 attack success rate(%). The constraint for adversarial perturbation is set to 8/255, with a per-step constraint of 2/255, and a total of 10 steps.

**MSCOCO (Image-Text Retrieval)**

| Source | Attack | ALBEF | | | TCL | | | $\text{CLIP}_{\text{ViT}}$ | | | $\text{CLIP}_{\text{CNN}}$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 |
| **ALBEF** | PGD | 94.35 | 87.8 | 84.05 | 34.15 | 18.63 | 13.05 | 21.71 | 10.94 | 7.34 | 23.83 | 11.75 | 8.78 |
| | BERT-Attack | 11.57 | 1.80 | 1.10 | 12.64 | 2.51 | 0.90 | 29.33 | 11.63 | 6.30 | 32.69 | 15.43 | 8.65 |
| | Sep-Attack | 97.01 | 93.44 | 91.0 | 61.06 | 37.75 | 29.04 | 57.19 | 39.93 | 31.41 | 58.81 | 41.45 | 33.22 |
| | Co-Attack | 96.65 | 94.74 | 93.12 | 57.33 | 37.24 | 28.53 | 50.71 | 33.1 | 26.44 | 52.06 | 33.89 | 27.47 |
| | SGA | **99.95** | 99.72 | **99.61** | 87.46 | 76.28 | 69.24 | 63.72 | 47.62 | 39.26 | 63.91 | 47.35 | 40.65 |
| | **Ours** | 99.9 | **99.77** | 99.51 | **88.81** | **78.63** | **71.93** | **69.25** | **52.01** | **43.9** | **68.53** | **52.03** | **44.14** |
| **TCL** | PGD | 40.81 | 25.16 | 19.02 | 98.54 | 96.42 | 94.77 | 21.79 | 11.41 | 7.89 | 24.97 | 12.65 | 9.56 |
| | BERT-Attack | 11.89 | 2.20 | 0.70 | 14.54 | 2.31 | 0.60 | 29.69 | 12.77 | 7.62 | 33.46 | 14.38 | 9.37 |
| | Sep-Attack | 66.38 | 48.4 | 38.53 | 99.15 | 97.69 | 96.38 | 59.94 | 42.33 | 34.26 | 60.77 | 44.11 | 35.04 |
| | Co-Attack | 65.22 | 45.39 | 36.43 | 94.95 | 91.18 | 89.08 | 55.28 | 38.04 | 29.73 | 56.68 | 37.31 | 29.82 |
| | SGA | 92.7 | 87.12 | 84.01 | **100.0** | 99.98 | 99.98 | 59.79 | 43.8 | 36.7 | 60.52 | 46.15 | 38.19 |
| | **Ours** | **94.72** | **90.6** | **88.02** | **100.0** | **100.0** | **100.0** | **70.51** | **54.58** | **45.91** | **70.29** | **54.58** | **45.94** |
| $\text{CLIP}_{\text{ViT}}$ | PGD | 10.26 | 4.29 | 2.43 | 12.72 | 5.48 | 3.06 | 82.91 | 68.34 | 60.44 | 21.62 | 11.31 | 8.27 |
| | BERT-Attack | 9.59 | 1.30 | 0.40 | 11.80 | 1.91 | 0.70 | 28.34 | 11.73 | 6.81 | 30.40 | 11.63 | 5.97 |
| | Sep-Attack | 25.91 | 11.35 | 7.2 | 28.2 | 13.3 | 8.32 | 88.36 | 77.6 | 69.98 | 47.57 | 31.28 | 24.63 |
| | Co-Attack | 26.35 | 11.97 | 7.2 | 28.23 | 12.89 | 8.19 | 88.78 | 78.21 | 70.98 | 47.36 | 31.49 | 25.29 |
| | SGA | 43.75 | 25.67 | 17.93 | 44.05 | 25.31 | 17.91 | **100.0** | **100.0** | **100.0** | 70.66 | 56.43 | 48.92 |
| | **Ours** | **52.69** | **32.04** | **23.71** | **51.88** | **31.05** | **23.67** | **100.0** | **100.0** | **100.0** | **80.18** | **67.75** | **60.13** |
| $\text{CLIP}_{\text{CNN}}$ | PGD | 8.38 | 3.59 | 1.85 | 11.9 | 5.11 | 2.73 | 13.66 | 7.43 | 4.74 | 92.68 | 85.89 | 81.92 |
| | BERT-Attack | 8.86 | 1.50 | 0.60 | 12.33 | 2.01 | 0.90 | 27.12 | 11.21 | 6.81 | 30.40 | 13.00 | 7.31 |
| | Sep-Attack | 29.13 | 12.99 | 8.01 | 31.4 | 14.58 | 9.31 | 52.23 | 34.82 | 26.89 | 96.16 | 91.9 | 89.03 |
| | Co-Attack | 29.49 | 13.26 | 8.28 | 31.83 | 15.11 | 9.81 | 53.15 | 36.11 | 28.78 | 97.79 | 94.29 | 92.26 |
| | SGA | 36.94 | 18.31 | 11.86 | 38.81 | 20.09 | 13.65 | 62.19 | 47.96 | 38.78 | **99.92** | **99.86** | **99.66** |
| | **Ours** | **41.4** | **21.68** | **14.64** | **43.62** | **23.7** | **16.18** | **70.43** | **55.15** | **46.41** | 99.8 | 99.54 | 99.39 |

**MSCOCO (Text-Image Retrieval)**

| Source | Attack | ALBEF | | | TCL | | | $\text{CLIP}_{\text{ViT}}$ | | | $\text{CLIP}_{\text{CNN}}$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 |
| **ALBEF** | PGD | 93.26 | 88.46 | 85.84 | 36.86 | 20.52 | 14.86 | 27.06 | 14.31 | 11.19 | 30.96 | 17.61 | 13.43 |
| | BERT-Attack | 27.46 | 14.48 | 10.98 | 28.07 | 14.39 | 10.26 | 43.17 | 26.37 | 19.91 | 46.11 | 28.43 | 22.14 |
| | Sep-Attack | 96.59 | 93.28 | 90.91 | 66.13 | 47.31 | 38.28 | 65.82 | 49.08 | 42.29 | 68.61 | 53.2 | 45.45 |
| | Co-Attack | 98.33 | 96.6 | 95.3 | 64.19 | 46.17 | 37.83 | 57.36 | 42.19 | 35.53 | 60.74 | 45.9 | 38.77 |
| | SGA | **99.94** | **99.81** | **99.74** | 88.17 | 77.33 | 70.64 | 69.71 | 55.21 | 47.73 | 70.78 | 56.34 | 48.74 |
| | **Ours** | 99.93 | 99.8 | 99.69 | **90.06** | **79.85** | **73.57** | **75.31** | **60.88** | **53.39** | **75.09** | **61.84** | **54.5** |
| **TCL** | PGD | 44.09 | 27.75 | 21.25 | 98.2 | 95.46 | 93.65 | 26.92 | 14.91 | 11.64 | 32.17 | 18.42 | 13.89 |
| | BERT-Attack | 26.82 | 14.09 | 10.80 | 29.17 | 15.03 | 10.91 | 44.49 | 27.47 | 21.00 | 46.07 | 29.28 | 22.59 |
| | Sep-Attack | 72.19 | 55.74 | 47.8 | 98.98 | 97.18 | 95.66 | 65.95 | 50.63 | 42.78 | 69.37 | 54.56 | 46.65 |
| | Co-Attack | 72.41 | 56.37 | 48.16 | 97.87 | 95.32 | 93.42 | 62.33 | 46.9 | 39.68 | 66.45 | 49.95 | 42.72 |
| | SGA | 92.99 | 87.65 | 84.17 | **100.0** | 99.99 | 99.98 | 65.31 | 50.92 | 43.87 | 67.34 | 53.48 | 45.87 |
| | **Ours** | **95.89** | **91.87** | **89.29** | **100.0** | **100.0** | **99.99** | **74.95** | **61.05** | **54.25** | **76.99** | **63.73** | **56.49** |
| $\text{CLIP}_{\text{ViT}}$ | PGD | 13.69 | 5.76 | 3.92 | 15.81 | 7.56 | 5.13 | 90.51 | 81.79 | 76.42 | 28.78 | 16.98 | 12.64 |
| | BERT-Attack | 22.64 | 10.95 | 8.17 | 25.07 | 12.92 | 8.90 | 39.08 | 24.08 | 17.44 | 37.43 | 24.96 | 18.66 |
| | Sep-Attack | 36.84 | 22.84 | 17.58 | 38.47 | 23.08 | 17.87 | 97.09 | 91.59 | 85.1 | 57.79 | 43.14 | 36.37 |
| | Co-Attack | 36.69 | 22.86 | 17.71 | 38.42 | 23.51 | 17.88 | 96.72 | 91.28 | 85.46 | 58.45 | 43.78 | 36.77 |
| | SGA | 51.08 | 33.62 | 26.91 | 51.02 | 34.43 | 27.9 | **100.0** | **100.0** | **100.0** | 75.58 | 63.22 | 56.28 |
| | **Ours** | **61.5** | **43.5** | **36.08** | **61.06** | **43.89** | **36.28** | **100.0** | **100.0** | 99.99 | **84.11** | **74.54** | **67.85** |
| $\text{CLIP}_{\text{CNN}}$ | PGD | 12.73 | 5.43 | 3.61 | 15.68 | 7.38 | 5.05 | 20.62 | 11.76 | 8.62 | 94.71 | 89.3 | 86.32 |
| | BERT-Attack | 23.27 | 11.34 | 8.41 | 25.48 | 13.25 | 8.81 | 37.44 | 23.48 | 17.66 | 40.10 | 26.71 | 20.85 |
| | Sep-Attack | 40.64 | 25.75 | 20.23 | 42.99 | 27.25 | 21.12 | 59.73 | 44.52 | 37.51 | 97.54 | 94.51 | 92.39 |
| | Co-Attack | 41.5 | 26.14 | 20.51 | 43.44 | 27.92 | 21.61 | 60.15 | 45.53 | 38.56 | 98.54 | 96.16 | 94.71 |
| | SGA | 46.79 | 29.97 | 23.7 | 48.9 | 32.89 | 26.04 | 67.73 | 53.77 | 47.19 | **99.97** | **99.83** | **99.77** |
| | **Ours** | **52.25** | **35.85** | **28.75** | **54.15** | **38.03** | **30.45** | **74.14** | **62.21** | **55.07** | 99.92 | 99.73 | 99.66 |

**Table A5: Comparison with state-of-the-art methods on Flickr30K dataset.** The source column shows the VLP models we use to generate multimodal adversarial examples. The gray area represents adversarial attacks under a white-box setting, the rest are black-box attacks. For both Image Retrieval and Text Retrieval, we provide R@1, R@5 and R@10 attack success rate(%). The constraint for adversarial perturbation is set to 2/255, with a per-step constraint of 0.5/255, and a total of 10 steps.

**Flickr30K (Image-Text Retrieval)**

| Source | Attack | ALBEF | | | TCL | | | $\text{CLIP}_{\text{ViT}}$ | | | $\text{CLIP}_{\text{CNN}}$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 |
| **ALBEF** | PGD | 52.45 | 36.57 | 30.00 | 3.06 | 0.40 | 0.10 | 8.96 | 1.66 | 0.41 | 10.34 | 2.96 | 1.85 |
| | BERT-Attack | 11.57 | 1.80 | 1.10 | 12.64 | 2.51 | 0.90 | 29.33 | 11.63 | 6.30 | 32.69 | 15.43 | 8.65 |
| | Sep-Attack | 65.69 | 47.60 | 42.10 | 17.60 | 3.72 | 1.90 | 31.17 | 12.05 | 7.01 | 32.82 | 15.86 | 9.06 |
| | Co-Attack | 77.16 | 64.60 | 58.37 | 15.21 | 4.19 | 1.47 | 23.60 | 7.82 | 3.93 | 25.12 | 8.42 | 5.39 |
| | SGA | **97.24** | **94.09** | **92.30** | 45.42 | 24.93 | 16.48 | 33.38 | 13.50 | 9.04 | 34.93 | 17.07 | 10.45 |
| | **Ours** | 96.14 | 92.48 | 89.8 | **49.74** | **26.83** | **19.24** | **39.14** | **18.9** | **12.4** | **41.38** | **20.08** | **12.98** |
| **TCL** | PGD | 6.15 | 1.30 | 0.70 | 77.87 | 65.13 | 58.72 | 7.48 | 1.45 | 0.81 | 10.34 | 2.75 | 1.54 |
| | BERT-Attack | 11.89 | 2.20 | 0.70 | 14.54 | 2.31 | 0.60 | 29.69 | 12.77 | 7.62 | 33.46 | 14.38 | 9.37 |
| | Sep-Attack | 20.13 | 4.91 | 2.70 | 84.72 | 73.07 | 65.43 | 31.29 | 12.98 | 7.72 | 33.33 | 14.27 | 9.89 |
| | Co-Attack | 23.15 | 6.98 | 3.63 | 77.94 | 64.26 | 56.18 | 27.85 | 9.80 | 5.22 | 30.74 | 12.09 | 7.28 |
| | SGA | 48.91 | 30.86 | 23.10 | **98.37** | **96.53** | **94.99** | 33.87 | 15.21 | 9.46 | 37.74 | 17.86 | 11.74 |
| | **Ours** | **51.09** | **31.86** | **24.8** | 98.21 | 94.87 | 93.49 | **40.25** | **19.52** | **12.4** | **42.91** | **21.99** | **14.32** |
| **$\text{CLIP}_{\text{ViT}}$** | PGD | 2.50 | 0.40 | 0.10 | 4.85 | 0.20 | 0.20 | 70.92 | 50.05 | 42.28 | 5.36 | 1.16 | 0.72 |
| | BERT-Attack | 9.59 | 1.30 | 0.40 | 11.80 | 1.91 | 0.70 | 28.34 | 11.73 | 6.81 | 30.40 | 11.63 | 5.97 |
| | Sep-Attack | 9.59 | 1.30 | 0.50 | 11.38 | 2.11 | 0.90 | 79.75 | 63.03 | 53.76 | 30.78 | 12.16 | 6.39 |
| | Co-Attack | 10.57 | 1.87 | 0.63 | 11.94 | 2.38 | 1.07 | 93.25 | 84.88 | 78.96 | 32.52 | 13.78 | 7.52 |
| | SGA | **13.40** | **2.46** | **1.35** | **16.23** | 3.77 | 1.10 | **99.08** | **97.25** | **95.22** | 38.76 | 19.45 | 11.95 |
| | **Ours** | 12.51 | 2.4 | 1.3 | 14.65 | **3.82** | **1.4** | 98.77 | 96.26 | 94.0 | **45.47** | **22.2** | **13.59** |
| **$\text{CLIP}_{\text{CNN}}$** | PGD | 2.09 | 0.30 | 0.10 | 4.00 | 0.40 | 0.20 | 1.10 | 0.52 | 0.41 | 86.46 | 69.13 | 61.17 |
| | BERT-Attack | 8.86 | 1.50 | 0.60 | 12.33 | 2.01 | 0.90 | 27.12 | 11.21 | 6.81 | 30.40 | 13.00 | 7.31 |
| | Sep-Attack | 8.55 | 1.50 | 0.60 | 12.64 | 1.91 | 0.70 | 28.34 | 10.8 | 6.30 | 91.44 | 78.54 | 71.58 |
| | Co-Attack | 8.79 | 1.53 | 0.60 | 13.10 | 2.31 | 0.93 | 28.79 | 11.63 | 6.40 | 94.76 | 87.03 | 82.08 |
| | SGA | 11.42 | **2.56** | 1.05 | **14.91** | **3.62** | **1.70** | 31.24 | 13.45 | 8.74 | **99.24** | **98.20** | **95.16** |
| | **Ours** | **12.2** | 2.51 | **1.4** | 14.33 | 3.32 | 1.4 | **35.21** | **15.58** | **9.65** | 99.11 | 97.15 | 94.54 |

**Flickr30K (Text-Image Retrieval)**

| Source | Attack | ALBEF | | | TCL | | | $\text{CLIP}_{\text{ViT}}$ | | | $\text{CLIP}_{\text{CNN}}$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 |
| **ALBEF** | PGD | 58.65 | 44.85 | 38.98 | 6.79 | 2.21 | 1.20 | 13.21 | 5.19 | 3.05 | 14.65 | 5.60 | 3.39 |
| | BERT-Attack | 27.46 | 14.48 | 10.98 | 28.07 | 14.39 | 10.26 | 43.17 | 26.37 | 19.91 | 46.11 | 28.43 | 22.14 |
| | Sep-Attack | 73.95 | 59.50 | 53.70 | 32.95 | 17.10 | 11.90 | 45.23 | 25.93 | 19.95 | 45.49 | 28.43 | 22.32 |
| | Co-Attack | 83.86 | 74.63 | 70.13 | 29.49 | 14.97 | 10.55 | 36.48 | 21.09 | 15.76 | 38.89 | 22.38 | 17.49 |
| | SGA | **97.28** | **94.27** | **92.58** | 55.25 | 36.01 | 27.25 | 44.16 | 27.35 | 20.84 | 46.57 | 29.16 | 22.68 |
| | **Ours** | 96.63 | 93.44 | 91.49 | **58.83** | **39.51** | **30.61** | **48.39** | **30.93** | **24.07** | **51.66** | **33.92** | **27.17** |
| **TCL** | PGD | 10.78 | 3.36 | 1.70 | 79.48 | 66.26 | 60.36 | 13.72 | 5.37 | 3.01 | 15.33 | 5.77 | 3.28 |
| | BERT-Attack | 26.82 | 14.09 | 10.80 | 29.17 | 15.03 | 10.91 | 44.49 | 27.47 | 21.00 | 46.07 | 29.28 | 22.59 |
| | Sep-Attack | 36.48 | 19.48 | 14.82 | 86.07 | 74.67 | 68.83 | 44.65 | 26.82 | 20.37 | 45.80 | 29.18 | 23.02 |
| | Co-Attack | 40.04 | 22.66 | 17.23 | 85.59 | 74.19 | 68.25 | 41.19 | 25.22 | 19.01 | 44.11 | 26.67 | 20.66 |
| | SGA | 60.34 | 42.47 | 34.59 | **98.81** | **97.19** | **95.86** | 44.88 | 28.79 | 21.95 | 48.30 | 29.70 | 23.68 |
| | **Ours** | **61.79** | **45.24** | **37.34** | 98.33 | 96.38 | 95.04 | **48.94** | **32.14** | **24.77** | **52.49** | **34.64** | **27.35** |
| **$\text{CLIP}_{\text{ViT}}$** | PGD | 4.93 | 1.44 | 1.01 | 8.17 | 2.27 | 1.46 | 78.61 | 60.78 | 51.50 | 8.44 | 2.35 | 1.54 |
| | BERT-Attack | 22.64 | 10.95 | 8.17 | 25.07 | 12.92 | 8.90 | 39.08 | 24.08 | 17.44 | 37.43 | 24.96 | 18.66 |
| | Sep-Attack | 23.25 | 11.22 | 8.01 | 25.60 | 12.92 | 9.14 | 86.79 | 75.24 | 67.84 | 39.76 | 25.62 | 19.34 |
| | Co-Attack | 24.33 | 11.74 | 8.41 | 26.69 | 13.80 | 9.46 | 95.86 | 90.83 | 87.36 | 41.82 | 26.77 | 21.10 |
| | SGA | 27.22 | 13.21 | 9.76 | **30.76** | 16.36 | **12.08** | 98.94 | **97.53** | **96.03** | 47.79 | 30.36 | 24.50 |
| | **Ours** | **30.0** | **14.64** | **10.53** | 30.62 | **16.42** | 11.56 | **99.0** | 97.27 | 94.96 | **50.74** | **34.47** | **27.69** |
| **$\text{CLIP}_{\text{CNN}}$** | PGD | 4.82 | 1.29 | 0.87 | 7.81 | 2.09 | 1.34 | 6.60 | 2.73 | 1.48 | 92.25 | 81.00 | 75.04 |
| | BERT-Attack | 23.27 | 11.34 | 8.41 | 25.48 | 13.25 | 8.81 | 37.44 | 23.48 | 17.66 | 40.10 | 26.71 | 20.85 |
| | Sep-Attack | 23.41 | 11.38 | 8.23 | 26.12 | 13.44 | 8.96 | 39.43 | 24.34 | 18.36 | 95.44 | 88.48 | 82.88 |
| | Co-Attack | 23.74 | 11.75 | 8.42 | 26.07 | 13.53 | 9.23 | 40.03 | 24.60 | 18.83 | 96.89 | 92.87 | 89.25 |
| | SGA | 24.80 | 12.32 | 8.98 | 28.82 | 15.12 | 10.56 | 42.12 | 26.80 | 20.23 | **99.49** | 98.41 | **97.14** |
| | **Ours** | **26.59** | **13.37** | **10.01** | **29.29** | **15.46** | **10.68** | **45.94** | **30.32** | **23.24** | **99.49** | **98.2** | 96.59 |

**Table A6: Comparison with state-of-the-art methods on MSCOCO dataset.** The source column shows the VLP models we use to generate multimodal adversarial examples. The gray area represents adversarial attacks under a white-box setting, the rest are black-box attacks. For both Image Retrieval and Text Retrieval, we provide R@1, R@5 and R@10 attack success rate(%). The constraint for adversarial perturbation is set to 2/255, with a per-step constraint of 0.5/255, and a total of 10 steps.

**MSCOCO (Image-Text Retrieval)**

| Source | Attack | ALBEF | | | TCL | | | $\text{CLIP}_{\text{ViT}}$ | | | $\text{CLIP}_{\text{CNN}}$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 |
| **ALBEF** | PGD | 76.70 | 67.49 | 62.47 | 12.46 | 5.00 | 3.14 | 13.96 | 7.33 | 5.21 | 17.45 | 9.08 | 6.45 |
| | BERT-Attack | 24.39 | 10.67 | 6.75 | 24.34 | 9.92 | 6.25 | 44.94 | 27.97 | 22.55 | 47.73 | 29.56 | 23.10 |
| | Sep-Attack | 82.60 | 73.20 | 67.58 | 32.83 | 15.52 | 10.10 | 44.03 | 27.60 | 21.84 | 46.96 | 29.83 | 23.15 |
| | Co-Attack | 79.87 | 68.62 | 62.88 | 32.62 | 15.36 | 9.67 | 44.89 | 28.33 | 21.89 | 47.30 | 29.89 | 23.29 |
| | SGA | **96.75** | **92.83** | **90.37** | 58.56 | 39.00 | 30.68 | 57.06 | 39.38 | 31.55 | 58.95 | 42.49 | 34.84 |
| | **Ours** | 96.57 | 91.92 | 89.07 | **60.69** | **41.31** | **32.17** | **61.69** | **43.72** | **35.7** | **62.32** | **45.42** | **37.61** |
| **TCL** | PGD | 10.83 | 5.28 | 3.21 | 59.58 | 51.25 | 47.89 | 14.23 | 7.40 | 4.93 | 17.25 | 8.51 | 6.45 |
| | BERT-Attack | 35.32 | 15.89 | 10.25 | 38.54 | 19.08 | 12.10 | 51.09 | 31.71 | 25.40 | 52.23 | 33.75 | 27.06 |
| | Sep-Attack | 41.71 | 21.37 | 14.99 | 70.32 | 59.64 | 55.09 | 50.74 | 31.34 | 24.43 | 51.90 | 34.02 | 26.79 |
| | Co-Attack | 46.08 | 24.87 | 17.11 | 85.38 | 74.73 | 68.23 | 51.62 | 31.92 | 24.87 | 52.13 | 33.80 | 27.09 |
| | SGA | 65.93 | 49.33 | 40.34 | 98.97 | **97.89** | **96.63** | 56.34 | 39.58 | 32.00 | 59.44 | 42.17 | 34.94 |
| | **Ours** | **68.06** | **51.73** | **42.57** | **98.99** | 97.46 | 96.21 | **63.3** | **46.13** | **37.34** | **64.24** | **46.86** | **39.48** |
| $\text{CLIP}_{\text{ViT}}$ | PGD | 7.24 | 3.10 | 1.65 | 10.19 | 4.23 | 2.50 | 54.79 | 36.21 | 28.57 | 7.32 | 3.64 | 2.79 |
| | BERT-Attack | 20.34 | 8.53 | 4.73 | 21.08 | 7.96 | 4.65 | 45.06 | 28.62 | 22.67 | 44.54 | 29.37 | 23.97 |
| | Sep-Attack | 23.41 | 10.33 | 6.15 | 25.77 | 11.60 | 7.45 | 68.52 | 52.30 | 43.88 | 43.11 | 27.22 | 21.77 |
| | Co-Attack | 30.28 | 13.64 | 8.83 | 32.84 | 15.27 | 10.27 | 97.98 | 94.94 | 93.00 | 55.08 | 38.64 | 31.42 |
| | SGA | 33.41 | 16.73 | 10.98 | 37.54 | 19.09 | 12.92 | **99.79** | **99.37** | **98.89** | 58.93 | 44.60 | 37.53 |
| | **Ours** | **35.96** | **18.18** | **12.04** | **36.32** | **18.05** | **12.23** | 99.66 | 99.14 | 98.44 | **64.41** | **50.04** | **41.59** |
| $\text{CLIP}_{\text{CNN}}$ | PGD | 7.01 | 3.03 | 1.77 | 10.08 | 4.20 | 2.38 | 4.88 | 2.96 | 1.71 | 76.99 | 63.80 | 56.76 |
| | BERT-Attack | 23.38 | 10.16 | 5.70 | 24.58 | 9.70 | 5.96 | 51.28 | 33.23 | 26.63 | 54.43 | 38.26 | 30.74 |
| | Sep-Attack | 26.53 | 11.78 | 6.88 | 30.26 | 13.00 | 8.61 | 50.44 | 32.71 | 25.92 | 88.72 | 78.71 | 72.77 |
| | Co-Attack | 29.83 | 13.13 | 8.35 | 32.97 | 15.11 | 9.76 | 53.10 | 35.91 | 28.53 | 96.72 | 94.02 | 91.57 |
| | SGA | 31.61 | 14.27 | **9.36** | **34.81** | **17.16** | **11.26** | 56.62 | 41.31 | 32.88 | **99.61** | **99.02** | **98.42** |
| | **Ours** | **33.26** | **15.85** | **9.96** | 33.89 | 16.19 | 10.57 | **59.6** | **43.12** | **35.23** | 99.51 | 98.69 | 97.91 |

**MSCOCO (Text-Image Retrieval)**

| Source | Attack | ALBEF | | | TCL | | | $\text{CLIP}_{\text{ViT}}$ | | | $\text{CLIP}_{\text{CNN}}$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 |
| **ALBEF** | PGD | 86.30 | 78.49 | 73.94 | 17.77 | 8.36 | 5.32 | 23.10 | 12.74 | 9.43 | 23.54 | 13.26 | 9.61 |
| | BERT-Attack | 36.13 | 23.71 | 18.94 | 33.39 | 20.21 | 15.56 | 52.28 | 38.06 | 32.04 | 54.75 | 41.39 | 35.11 |
| | Sep-Attack | 89.88 | 82.60 | 78.82 | 42.92 | 27.04 | 20.65 | 54.46 | 40.12 | 33.46 | 55.88 | 41.30 | 35.18 |
| | Co-Attack | 87.83 | 80.16 | 75.98 | 43.09 | 27.32 | 21.35 | 54.75 | 40.00 | 33.81 | 55.64 | 41.48 | 35.28 |
| | SGA | **96.95** | **93.44** | **91.00** | 65.38 | 47.61 | 38.96 | 65.25 | 50.42 | 43.47 | 66.52 | 52.44 | 45.05 |
| | **Ours** | 96.47 | 92.86 | 90.26 | **67.46** | **49.9** | **41.24** | **67.43** | **53.08** | **45.66** | **69.22** | **55.21** | **48.08** |
| **TCL** | PGD | 16.52 | 8.40 | 5.61 | 69.53 | 60.88 | 57.56 | 22.28 | 12.20 | 9.10 | 23.12 | 12.77 | 9.49 |
| | BERT-Attack | 45.92 | 30.40 | 23.89 | 48.48 | 31.48 | 24.47 | 58.80 | 43.10 | 36.68 | 61.26 | 46.14 | 39.54 |
| | Sep-Attack | 52.97 | 36.33 | 28.97 | 78.97 | 69.79 | 65.71 | 60.13 | 44.13 | 37.32 | 61.26 | 45.99 | 38.97 |
| | Co-Attack | 57.09 | 39.85 | 32.00 | 91.39 | 83.16 | 78.05 | 60.46 | 45.16 | 37.73 | 62.49 | 46.61 | 39.74 |
| | SGA | 73.30 | 58.40 | 50.96 | **99.15** | **98.17** | **97.34** | 63.99 | 49.87 | 42.46 | 65.70 | 51.45 | 44.64 |
| | **Ours** | **75.86** | **61.6** | **54.02** | 99.02 | 97.97 | 97.16 | **68.59** | **54.33** | **47.1** | **70.58** | **56.55** | **49.69** |
| $\text{CLIP}_{\text{ViT}}$ | PGD | 10.75 | 4.64 | 2.91 | 13.74 | 6.77 | 4.32 | 66.85 | 51.80 | 46.02 | 11.34 | 6.50 | 4.66 |
| | BERT-Attack | 29.74 | 18.13 | 13.73 | 29.61 | 16.91 | 12.66 | 51.68 | 37.12 | 31.02 | 53.72 | 40.13 | 34.32 |
| | Sep-Attack | 34.61 | 21.00 | 16.15 | 36.84 | 22.63 | 17.03 | 77.94 | 66.77 | 60.69 | 49.76 | 37.51 | 31.74 |
| | Co-Attack | 42.67 | 27.20 | 21.46 | 44.69 | 29.42 | 22.85 | 98.80 | 96.83 | 95.33 | 62.51 | 49.48 | 42.63 |
| | SGA | 44.64 | 28.66 | 22.64 | 47.76 | 32.30 | 25.70 | **99.79** | **99.37** | **98.94** | 65.83 | 53.58 | 46.84 |
| | **Ours** | **48.0** | **31.87** | **25.37** | **48.56** | **32.45** | **25.92** | 99.7 | 99.35 | 98.84 | **69.99** | **58.04** | **50.66** |
| $\text{CLIP}_{\text{CNN}}$ | PGD | 10.62 | 4.51 | 2.76 | 13.65 | 6.39 | 4.32 | 10.70 | 6.20 | 4.52 | 84.20 | 73.64 | 67.86 |
| | BERT-Attack | 34.64 | 21.13 | 16.25 | 29.61 | 16.91 | 12.66 | 57.49 | 42.73 | 36.23 | 62.17 | 47.80 | 40.79 |
| | Sep-Attack | 39.29 | 24.04 | 18.83 | 41.51 | 26.13 | 20.17 | 57.11 | 41.89 | 35.55 | 92.49 | 85.84 | 81.66 |
| | Co-Attack | 41.97 | 26.62 | 20.91 | 43.72 | 28.62 | 22.35 | 58.90 | 45.22 | 38.72 | 98.56 | 96.86 | 95.55 |
| | SGA | 43.00 | 27.64 | 21.74 | 45.95 | 30.57 | **24.27** | 60.77 | 46.99 | 40.49 | **99.80** | **99.29** | **98.77** |
| | **Ours** | **45.15** | **29.81** | **23.6** | **46.49** | **30.6** | 24.0 | **64.87** | **50.88** | **44.32** | 99.7 | 99.04 | 98.44 |

## References


1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016)
3. Khan, Z., Fu, Y.: Exploiting bert for multimodal target sentiment classification through input space translation. In: Proceedings of the 29th ACM international conference on multimedia. pp. 3034–3042 (2021)
4. Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., Hoi, S.C.H.: Align before fuse: Vision and language representation learning with momentum distillation. Advances in neural information processing systems **34**, 9694–9705 (2021)
5. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014)
6. Lu, D., Wang, Z., Wang, T., Guan, W., Gao, H., Zheng, F.: Set-level guidance attack: Boosting adversarial transferability of vision-language pre-training models (2023)
7. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083 (2017)
8. Plummer, B.A., Wang, L., Cervantes, C.M., Caicedo, J.C., Hockenmaier, J., Lazebnik, S.: Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In: Proceedings of the IEEE international conference on computer vision. pp. 2641–2649 (2015)
9. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
10. Wang, Z., Huang, S.L., Kuruoglu, E.E., Sun, J., Chen, X., Zheng, Y.: Pac-bayes information bottleneck. arXiv preprint arXiv:2109.14509 (2021)
11. Yang, J., Duan, J., Tran, S., Xu, Y., Chanda, S., Chen, L., Zeng, B., Chilimbi, T., Huang, J.: Vision-language pre-training with triple contrastive learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15671–15680 (2022)