The Power of Decaying Steps: Enhancing Attack Stability and Transferability for Sign-based Optimizers

Supplementary Material

6. Convergence Analysis of MDCS-MI

Lemma 4. [28] *$\forall \boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{R}^d$, we have*

$$\|P_{\mathbf{Q},\boldsymbol{D}_t^{-1}}(\boldsymbol{x}_1) - P_{\mathbf{Q},\boldsymbol{D}_t^{-1}}(\boldsymbol{x}_2)\|_{\boldsymbol{D}_t^{-1}} \leq \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_{\boldsymbol{D}_t^{-1}}. \tag{11}$$

and

$$\boldsymbol{x}^* = P_{\mathbf{Q},\boldsymbol{D}_t^{-1}}(\boldsymbol{x}^*). \tag{12}$$

Lemma 5. *Let $1 > \lambda > 0$ and $\beta > 0$. Let $\beta_t = \beta\lambda^{t-1}$. Suppose $\{\boldsymbol{m}_{t+1}\}_{t=1}^{\infty}$ is generated by MDCS-MI. Then there exits a $M_2 > 1$ such that*

$$\|\boldsymbol{m}_{t+1}\| \leq M_2, \ \forall t > 0,$$

Proof Note

$$\|\nabla_{\boldsymbol{x}} J(\boldsymbol{x}^{adv})\| \leq \|\nabla_{\boldsymbol{x}} J(\boldsymbol{x}^{adv})\|_1.$$

Then

$$\|\boldsymbol{m}_{t+1}\| \leq \beta_t\|\boldsymbol{m}_t\|+1 \leq \beta_{t-1}\|\boldsymbol{m}_{t-1}\|+\beta_t+1 \leq \sum_{i=1}^{t}\beta_i+1.$$

Let $M_2 = \sum_{i=1}^{\infty}\beta_i + 1$, Lemma 5 follows from the convergence of $\sum_{i=1}^{t}\beta_i$.

Note

$$d_{t,i} = \min(\frac{1}{|\boldsymbol{m}_{t+1,i}|}, d_{t-1,i}).$$

According to Lemma 5, we can get

Lemma 6. *Let $1 > \lambda > 0$ and $\beta > 0$. Let $\beta_t = \beta\lambda^{t-1}$. Suppose $\{\boldsymbol{D}_t\}_{t=1}^{\infty}$ is generated by MDCS-MI. Then for each t and i,*

$$\frac{1}{M_2} \leq d_{t,i} \leq 1.$$

Proof of Theorem 3

From Lemma 4, we know

$$\begin{aligned}
&\|\boldsymbol{x}_{t+1}^{adv} - \boldsymbol{x}^*\|_{\boldsymbol{D}_t^{-1}}^2 \\
&= \|P_{\mathbf{Q},\boldsymbol{D}_t^{-1}}(\boldsymbol{x}_t^{adv} + \alpha_t\boldsymbol{D}_t\boldsymbol{m}_{t+1}) - P_{\mathbf{Q},\boldsymbol{D}_t^{-1}}(\boldsymbol{x}^*)\|_{\boldsymbol{D}_t^{-1}}^2 \\
&\leq \|\boldsymbol{x}_t^{adv} + \alpha_t\boldsymbol{D}_t\boldsymbol{m}_{t+1} - \boldsymbol{x}^*\|_{\boldsymbol{D}_t^{-1}}^2 \\
&= \|\boldsymbol{x}_t^{adv} - \boldsymbol{x}^*\|_{\boldsymbol{D}_t^{-1}}^2 + \|\alpha_t\boldsymbol{D}_t\boldsymbol{m}_{t+1}\|_{\boldsymbol{D}_t^{-1}}^2 \\
&\quad + 2\alpha_t\langle\boldsymbol{m}_{t+1}, \boldsymbol{x}_t^{adv} - \boldsymbol{x}^*\rangle \\
&= \|\boldsymbol{x}_t^{adv} - \boldsymbol{x}^*\|_{\boldsymbol{D}_t^{-1}}^2 + \|\alpha_t\boldsymbol{D}_t\boldsymbol{m}_{t+1}\|_{\boldsymbol{D}_t^{-1}}^2 \\
&\quad + 2\alpha_t\langle\beta_t\boldsymbol{m}_t + \frac{\nabla_{\boldsymbol{x}} J(\boldsymbol{x}_t^{adv})}{\|\nabla_{\boldsymbol{x}} J(\boldsymbol{x}_t^{adv})\|_1}, \boldsymbol{x}_t^{adv} - \boldsymbol{x}^*\rangle.
\end{aligned}$$

Rearrange the inequality, we have

$$\begin{aligned}
&\frac{2\alpha_t}{\|\nabla_{\boldsymbol{x}} J(\boldsymbol{x}_t^{adv})\|_1}\langle\nabla_{\boldsymbol{x}} J(\boldsymbol{x}_t^{adv}), \boldsymbol{x}_t^{adv} - \boldsymbol{x}^*\rangle \\
&\geq \|\boldsymbol{x}_{t+1}^{adv} - \boldsymbol{x}^*\|_{\boldsymbol{D}_t^{-1}}^2 - \|\boldsymbol{x}_t^{adv} - \boldsymbol{x}^*\|_{\boldsymbol{D}_t^{-1}}^2 \\
&\quad - \|\alpha_t\boldsymbol{D}_t\boldsymbol{m}_{t+1}\|_{\boldsymbol{D}_t^{-1}}^2 - 2\alpha_t\beta_t\langle\boldsymbol{m}_t, \boldsymbol{x}_t^{adv} - \boldsymbol{x}^*\rangle,
\end{aligned}$$

i.e.,

$$\begin{aligned}
&\frac{1}{\|\nabla_{\boldsymbol{x}} J(\boldsymbol{x}_t^{adv})\|_1}\langle\nabla_{\boldsymbol{x}} J(\boldsymbol{x}_t^{adv}), \boldsymbol{x}_t^{adv} - \boldsymbol{x}^*\rangle \\
&\geq \frac{\|\boldsymbol{x}_{t+1}^{adv} - \boldsymbol{x}^*\|_{\boldsymbol{D}_t^{-1}}^2 - \|\boldsymbol{x}_t^{adv} - \boldsymbol{x}^*\|_{\boldsymbol{D}_t^{-1}}^2}{2\alpha_t} - \frac{\alpha_t\|\boldsymbol{m}_{t+1}\|^2}{2} \\
&\quad - \beta_t\langle\boldsymbol{m}_t, \boldsymbol{x}_t^{adv} - \boldsymbol{x}^*\rangle \\
&\geq \frac{\|\boldsymbol{x}_{t+1}^{adv} - \boldsymbol{x}^*\|_{\boldsymbol{D}_t^{-1}}^2 - \|\boldsymbol{x}_t^{adv} - \boldsymbol{x}^*\|_{\boldsymbol{D}_t^{-1}}^2}{2\alpha_t} - \frac{\alpha_t\|\boldsymbol{m}_{t+1}\|^2}{2} \\
&\quad - \frac{\beta_t\alpha_t\|\boldsymbol{m}_t\|^2}{2} - \frac{\beta_t\|\boldsymbol{x}_t^{adv} - \boldsymbol{x}^*\|^2}{2\alpha_t}.
\end{aligned}$$

Using the property of concave functions,

$$\langle\nabla_{\boldsymbol{x}} J(\boldsymbol{x}_t^{adv}), \boldsymbol{x}_t^{adv} - \boldsymbol{x}^*\rangle \leq J(\boldsymbol{x}_t^{adv}) - J(\boldsymbol{x}^*).$$

Then

$$\begin{aligned}
&\frac{J(\boldsymbol{x}^*) - J(\boldsymbol{x}_t^{adv})}{M} \\
&\leq \frac{\|\boldsymbol{x}_t^{adv} - \boldsymbol{x}^*\|_{\boldsymbol{D}_t^{-1}}^2 - \|\boldsymbol{x}_{t+1}^{adv} - \boldsymbol{x}^*\|_{\boldsymbol{D}_t^{-1}}^2}{2\alpha_t} \\
&\quad + \frac{\alpha_t\|\boldsymbol{m}_{t+1}\|^2}{2} + \frac{\beta_t\alpha_t\|\boldsymbol{m}_t\|^2}{2} + \frac{\beta_t\|\boldsymbol{x}_t^{adv} - \boldsymbol{x}^*\|^2}{2\alpha_t}.
\end{aligned}$$

Summing this inequality from $t = 1$ to T, we obtain

$$\begin{aligned}
&\frac{1}{M}\sum_{t=1}^{T}\left[J(\boldsymbol{x}^*) - J(\boldsymbol{x}_t^{adv})\right] \\
&\leq \underbrace{\sum_{t=1}^{T}\frac{\|\boldsymbol{x}_t^{adv} - \boldsymbol{x}^*\|_{\boldsymbol{D}_t^{-1}}^2 - \|\boldsymbol{x}_{t+1}^{adv} - \boldsymbol{x}^*\|_{\boldsymbol{D}_t^{-1}}^2}{2\alpha_t}}_{P_1} \\
&\quad + \underbrace{\sum_{t=1}^{T}\frac{\beta_t\|\boldsymbol{x}_t^{adv} - \boldsymbol{x}^*\|^2}{2\alpha_t}}_{P_2} \\
&\quad + \underbrace{\sum_{t=1}^{T}\frac{\alpha_t\|\boldsymbol{m}_{t+1}\|^2}{2} + \sum_{t=1}^{T}\frac{\beta_t\alpha_t\|\boldsymbol{m}_t\|^2}{2}}_{P_3}.
\end{aligned}$$

To bound P_1, we have

$$\begin{aligned} P_1 =& \sum_{t=1}^{T} \frac{\|\boldsymbol{x}_t^{adv} - \boldsymbol{x}^*\|^2_{\boldsymbol{D}_t^{-1}} - \|\boldsymbol{x}_{t+1}^{adv} - \boldsymbol{x}^*\|^2_{\boldsymbol{D}_t^{-1}}}{2\alpha_t} \\ =& \sum_{t=2}^{T} \left(\frac{\|\boldsymbol{x}_t^{adv} - \boldsymbol{x}^*\|^2_{\boldsymbol{D}_t^{-1}}}{2\alpha_t} - \frac{\|\boldsymbol{x}_t^{adv} - \boldsymbol{x}^*\|^2_{\boldsymbol{D}_{t-1}^{-1}}}{2\alpha_{t-1}} \right) \\ &+ \frac{\|\boldsymbol{x}_1^{adv} - \boldsymbol{x}^*\|^2_{\boldsymbol{D}_1^{-1}}}{2\alpha_1} - \frac{\|\boldsymbol{x}_{T+1}^{adv} - \boldsymbol{x}^*\|^2_{\boldsymbol{D}_T^{-1}}}{2\alpha_T} \end{aligned} \tag{13}$$

According to Lemma 6,

$$\begin{aligned} P_1 \leq& \sum_{t=2}^{T} \sum_{i=1}^{d} \left(\frac{1}{2\alpha_t d_{t,i}} - \frac{1}{2\alpha_{t-1} d_{t-1,i}} \right) |\boldsymbol{x}_{t,i}^{adv} - \boldsymbol{x}_{\boldsymbol{i}}^*|^2 \\ &+ \frac{\|\boldsymbol{x}_1^{adv} - \boldsymbol{x}^*\|^2_{\boldsymbol{D}_1^{-1}}}{2\alpha_1} \\ \leq& \sum_{i=1}^{d} \frac{1}{2\alpha_T d_{T,i}} G^2 \\ \leq& \frac{dM_2 G^2 \sqrt{T}}{2\gamma}. \end{aligned} \tag{14}$$

To bound P_2, we have

$$\begin{aligned} P_2 =& \sum_{t=1}^{T} \frac{\beta_t \|\boldsymbol{x}_t^{adv} - \boldsymbol{x}^*\|^2}{2\alpha_t} \\ \leq& \frac{\beta G^2}{2\gamma} \sum_{t=1}^{T} \lambda^{t-1} \sqrt{t} \\ \leq& \frac{\beta G^2}{2\gamma} \sum_{t=1}^{T} \lambda^{t-1} t \\ \leq& \frac{\beta G^2}{2\gamma(1-\lambda)^2} \end{aligned} \tag{15}$$

To bound P_3, according to Lemma 4, we have

$$\begin{aligned} P_3 =& \sum_{t=1}^{T} \frac{\alpha_t \|\boldsymbol{m}_{t+1}\|^2}{2} + \sum_{t=1}^{T} \frac{\beta_t \alpha_t \|\boldsymbol{m}_t\|^2}{2} \\ \leq& \sum_{t=1}^{T} \frac{\alpha_t {M_2}^2}{2} + \sum_{t=1}^{T} \frac{\beta_t \alpha_t {M_2}^2}{2} \\ \leq& \frac{M_2^2}{2} \sum_{t=1}^{T} \frac{\gamma}{\sqrt{t}} + \frac{D_2^2}{2} \sum_{t=1}^{T} \frac{\beta_t \gamma}{\sqrt{t}} \\ \leq& 2\gamma {M_2}^2 \sqrt{T}. \end{aligned} \tag{16}$$

Combining (14), (15) and (16), we have

$$\begin{aligned} &\frac{1}{M} \sum_{t=1}^{T} \left(J(\boldsymbol{x}^*) - J(\boldsymbol{x}_t^{adv}) \right) \\ &\leq \frac{dM_2 G^2 \sqrt{T}}{2\gamma} + \frac{\beta G^2}{2\gamma(1-\lambda)^2} + 2\gamma {M_2}^2 \sqrt{T}. \end{aligned}$$

Thus

$$\begin{aligned} &\frac{1}{MT} \sum_{t=1}^{T} \left(J(\boldsymbol{x}^*) - J(\boldsymbol{x}_t^{adv}) \right) \\ &\leq \frac{dM_2 G^2}{2\gamma\sqrt{T}} + \frac{\beta G^2}{2\gamma(1-\lambda)^2} + \frac{2\gamma {M_2}^2}{\sqrt{T}}. \end{aligned}$$

By concavity of $J(\boldsymbol{x})$, we obtain

$$\begin{aligned} &J(\boldsymbol{x}^*) - J(\bar{\boldsymbol{x}}_T^{adv}) \\ &\leq M \left(\frac{dM_2 G^2}{2\gamma\sqrt{T}} + \frac{\beta G^2}{2\gamma(1-\lambda)^2 T} + \frac{2\gamma {M_2}^2}{\sqrt{T}} \right). \end{aligned} \tag{17}$$

This completes the proof of Theorem 3.

7. Pseudocode for MDCS-MEF and MDCS-OPS

For clarity, we give the detailed descriptions of MDCS-MEF (Algorithm 2) and MDCS-OPS (Algorithm 3) for image classification tasks.

Algorithm 2 MDCS-MEF

Input: Loss function J, a raw example $\boldsymbol{x}$ with ground-truth label y, the perturbation size ϵ, momentum parameter $0 < \beta_t \leq 1$, step-size $\alpha_t > 0$, maximum iteration T, the outer/inner decay factor μ_{outer}/μ_{inner}; the number of randomly sampled examples, N; neighborhood radius ξ, exploration radius γ.

Initialize $\boldsymbol{x}_0^{adv} = \boldsymbol{x}$, $d_{0,i} = 1$. $\boldsymbol{g}_0^{outer} = 0$; $\{g_{0,n}^{inner}\}_{n=1}^{N} = 0$; $\boldsymbol{x}_0^{adv} = \boldsymbol{x}$; $\alpha = \epsilon/T$
repeat
  $\{\boldsymbol{x}_{t,n}\}_{n=1}^{N} \in U_\gamma(\boldsymbol{x}_t^{adv})$
  $\{\boldsymbol{x}_{t,n}^{'}\}_{n=1}^{N} = \{\boldsymbol{x}_{t,n}\}_{n=1}^{N} + \xi \cdot \text{sign}(\, g_{t,n}^{inner})\}_{n=1}^{N})$
  $\{g_{t,n}\}_{n=1}^{N} = \nabla_{\boldsymbol{x}} J(\{\boldsymbol{x}_{t,i}^{'}\}_{i=1}^{N})$
  $\{g_{t,n}^{inner}\}_{n=1}^{N} = \frac{\{g_{t,n}\}_{n=1}^{N}}{\|\{g_{t,n}\}_{n=1}^{N}\|_1} - \mu_{inner}\{g_{t,n}^{inner}\}_{n=1}^{N}$
  $\boldsymbol{g}_t^{outer} = \mu_{outer}\, \boldsymbol{g}_t^{outer} + \frac{1}{N} \sum_{i=1}^{N} \frac{\{g_{t,n}\}_{n=1}^{N}}{\|\{g_{t,n}\}_{n=1}^{N}\|_1}$
  $d_{t,i} = \min(\frac{1}{|\boldsymbol{g}_{t,i}^{outer}|}, d_{t-1,i})$,
  $\boldsymbol{D}_t = \text{diag}(\boldsymbol{d}_t)$,
  $\boldsymbol{x}_{t+1}^{adv} = P_{\mathbf{Q},\boldsymbol{D}_t^{-1}}(\boldsymbol{x}_t^{adv} + \alpha_t \cdot \boldsymbol{D}_t \cdot g_t^{outer})$.
until $t = T$

Output: $\boldsymbol{x}_T^{adv}$.

Algorithm 3 MDCS-OPS

Input: Loss function J, a raw example $\boldsymbol{x}$ with ground-truth label y, the perturbation size ϵ, the maximum iteration T and decay factor μ; Level list K_{list}; Radius list R_{list}; Number of operator samples N_p; Number of perturbation samples N_e.

Initialize $\boldsymbol{x}_0^{adv} = \boldsymbol{x}$, $d_{0,i} = 1$, $\alpha = \epsilon/T, \boldsymbol{m}_0 = \mathbf{0}, \Delta_0 = 0$.
repeat
  $\bar{g} = \nabla_{\boldsymbol{x}} J(f(\boldsymbol{x} + \Delta_i), y)$.
  for δ in $S(D, N_e)$ **do**
    for op in $S(P, N_p)$ **do**
      $\bar{\boldsymbol{g}} = \bar{\boldsymbol{g}} + \nabla_{\boldsymbol{x}} J\left[f(op(\boldsymbol{x} + \Delta_i + \delta)), y\right]$
    end for
  end for
  $\bar{\boldsymbol{g}} = \frac{\bar{\boldsymbol{g}}}{N_e \times N_p + 1}$.
  $\boldsymbol{m}_{t+1} = \mu \cdot \boldsymbol{m}_t + \frac{\bar{\boldsymbol{g}}}{\|\bar{\boldsymbol{g}}\|_1}$.
  $d_{t,i} = \min(\frac{1}{|\boldsymbol{m}_{t,i}|}, d_{t-1,i})$,
  $\boldsymbol{D}_t = \text{diag}(\boldsymbol{d}_t)$,
  $\Delta_{i+1} = P_{\mathbf{Q}, \boldsymbol{D}_t^{-1}}(\Delta_i + \alpha_t \cdot \boldsymbol{D}_t \cdot \boldsymbol{m}_{t+1})$.
until $t = T$

Output: $\boldsymbol{x}_T^{adv}$.

8. Additional Experiments

This section presents more experimental results to provide a comprehensive empirical evaluation of the proposed methods on image classification, Visual Question Answering (VQA), and cross-modal retrieval tasks.

8.1. Image Classification Tasks

8.1.1. Transferability and Stability

In this subsection, we extend our evaluation to include additional source models, namely VGG16 and ViT-B/16. The success rates of these transfer attacks are quantified in Tab. 5. The results clearly demonstrate that our derived MDCS strategy consistently enhances adversarial transferability compared to the baseline attacks. Notably, the MDCS-OPS variant achieves the highest attack success rate in the black-box setting, establishing a new state-of-the-art. As shown in Fig. 7, integrating the MDCS strategy into TI and SI significantly enhances the iterative stability of adversarial attacks.

8.1.2. Convergence Behavior

To make a through comparison, we also investigate the white-box and black-box attack convergence behavior of loss function $J(\boldsymbol{x}_t^{adv}, y)$ with respect to the number of iterations. The relationship between the value of loss function $J(\boldsymbol{x}_t^{adv}, y)$ and the number of iterations is shown in Fig. 8. We summarize three key findings: (1) Overfitting is observed in I-FGSM and MI-FGSM. (2) The stability of black-box ASR ensures that whenever the iteration is stopped, relatively good and reliable transferability can be provided. (3) From an optimization perspective, Fig.8 clearly verify the overfitting of I-FGSM and MI-FGSM, and the stability of our MDCS. Naturally, if the concerned momentum is already stable like that in OPS, gains from MDCS are moderate.

To sum up, stability refers to maintaining performance across iterations without degradation. The degradation in black-box ASR typically indicates overfitting. Therefore, stability contributes to improved transferability.

8.1.3. Ablation Study

Here, we investigate the impact of the step-size parameter γ on the performance of MDCS-MI, which governs the trade-off between transferability and imperceptibility. According to [49], there remains a notable lack of consensus and attention regarding the proper metrics for evaluating imperceptibility. A popular proxy measure of imperceptibility is the L_p-norm of the perturbation. It offers a good trade-off between simplicity, mathematical tractability, and practicality in applications [34]. Recent research [48, 49] suggests that it is problematic to solely constrain all attacks with the same L_∞ norm bound without more comprehensive comparisons. Thus, we employ 7 different metrics as the indicators of imperceptibility of the crafted AEs, including the Average L_∞ Distortion (ALD$_p$), Average L_2 Distortion (ALD$_p$) [23], Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM), Multi Scale Structural Similarity Index Measure (MS-SSIM), Frechet Inception Distance (FID) [16], Learned Perceptual Image Patch Similarity (LPIPS) [47]. As shown in Fig. 9, the attack successful rate of MDCS-MI increases with the growth of γ (varied within the range of 0.25 to 10). The rate plateaus at around 47% when γ is greater than 2. For imperceptibility, the increase of γ leads to a significant improvement in all metrics except for FID and ALD$_p$($p = \infty$). Therefore, a holistic assessment with diverse metrics is essential for evaluating the imperceptibility of AEs. By balancing the transferability and imperceptibility in our experiment, γ is determined by simple grid search over the range [2, 4].

8.2. VQA Tasks

VQA serves as a practical application of large multimodal models, in which the model is tasked with generating an open-ended answer from a given image and an associated question. In contrast to adversarial attacks on image classification, white-box attacks for VQA have not achieved satisfactory performance [5, 44]. Therefore, this subsection specifically investigates white-box attacks on VQA, utilizing image perturbation techniques.

We conduct experiments on the TextVQA dataset [33], targeting two representative VLMs: LLaVA-1.5 [25] and PrismVLM [18]. To validate the effect of our MDCS, we focus on attacking visual encoders to generate AEs

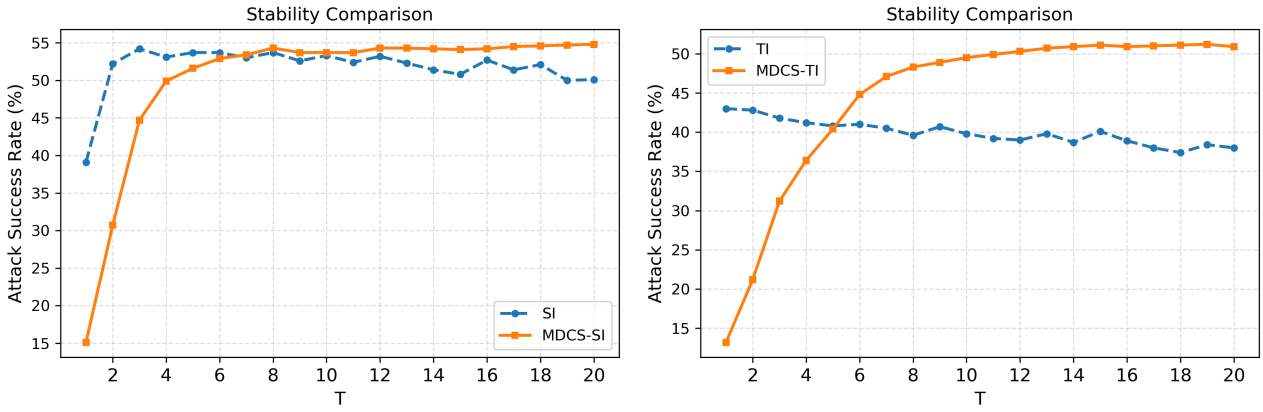

Figure 7. Stability comparison of SI and TI (based on MI-FGSM). We employ Res50 as surrogate model and Inc-v3 as target model with $\epsilon = 16/255$.

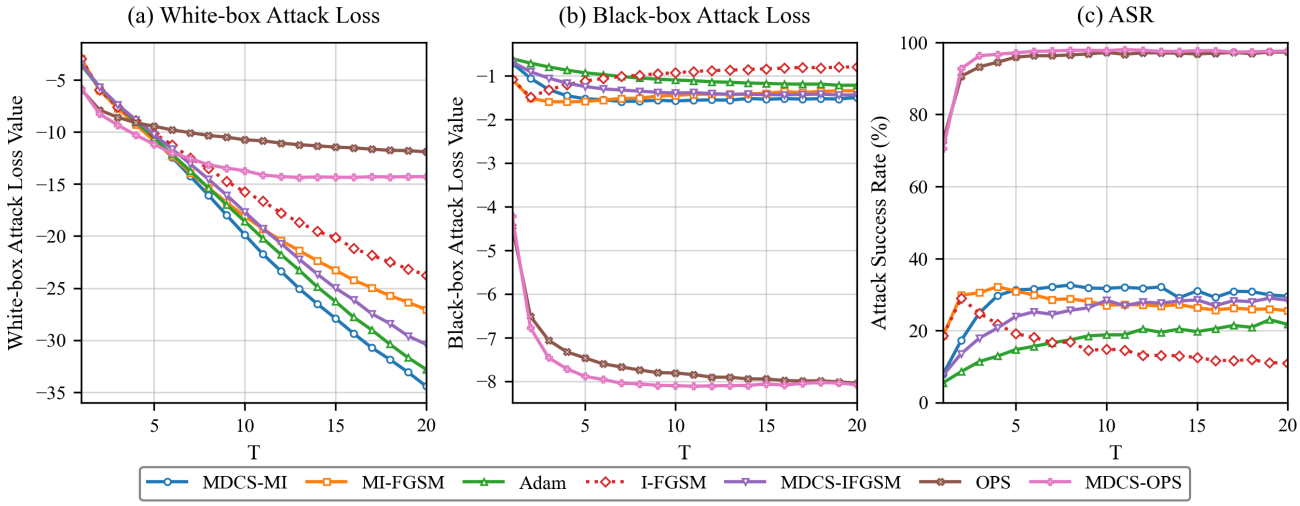

Figure 8. Dynamics of white-box and black-box losses. We employ Res50 as surrogate model and Vis-S as target model.

for LLaVA-1.5 and PrismVLM, respectively. The white-box attack success rates are reported for two input types: Pure (raw textual questions) and OCR (augmented prompts that incorporate both raw questions and OCR-extracted text from the images).

The notations Pre, Post_N in Tab. 6 refer to accuracy (%) for pre-attack, post-attack under normal setting, respectively. ASR represents the Attack Success Rate, which is derived from the values of Pre and Post_N. Tab. 6 presents the attack success rates of our method compared to several baselines. MDCS-MI consistently achieves the highest rates across all experimental settings. The performance gain is particularly significant on the PrismVLM model, achieving an improvement of 3.0% and 4.1% in the Pure and OCR settings, respectively. Fig. 10 presents a side-by-side comparison of the images and PrismVLM's answers on the textVQA dataset before and after being subjected to the MDCS-MI attack.

8.3. Cross-Modal Retrieval Tasks

In the main text, we have reported R@1 attack success rates for cross-model transferability experiments for conciseness. In this subsection, we present the complete evaluation results for both image-text and text-image retrieval tasks on the Flickr30K (Table 7) and MSCOCO (Table 8) datasets. These results include attack success rates at R@1, R@5, and R@10.

These results validate the general applicability of our MDCS module. Integrating MDCS with a spectrum of attack frameworks, from foundational methods like SGA and DRA to the state-of-the-art SA-AET, yields consistent performance gains. In each tested cross-model transferable scenario, MDCS-SGA, MDCS-DRA, and MDCS-SAAET demonstrably outperform their baseline counter-

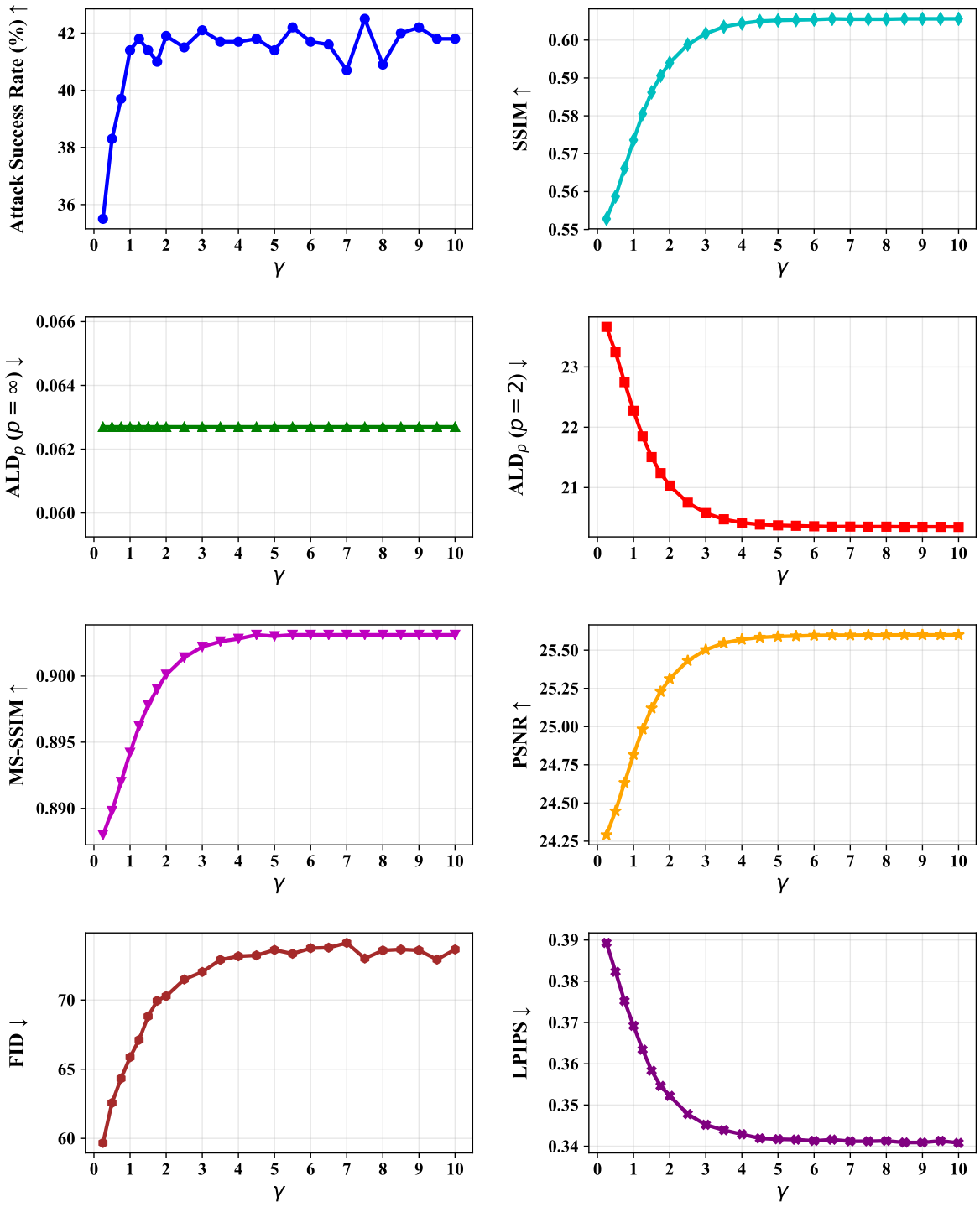

Figure 9. Ablation study on the value of γ for MDCS-MI. We employ Res50 as surrogate model and Inc-v3 as target model.

parts. This confirms that MDCS functions as a potent and model-agnostic component for amplifying the transferability of adversarial attacks across various VLP architectures and metrics.

Table 5. Transferability comparisons across different networks. AEs are crafted for VGG16 and ViT-B/16. The best results are marked in bold. The gray area represents adversarial attacks under a white-box setting, the rest are black-box attacks.

Model	Attack	Res50	VGG16	Mob-v2	Inc-v3	ViT-B	PiT-B	Vis-S
VGG16	I-FGSM	35.8	99.9	62.9	27.1	8.0	15.5	25.1
	MI	59.1	99.8	80.2	55.9	15.6	29.6	44.4
	MDCS-MI	59.5	99.8	81.0	59.9	17.5	30.3	45.9
	VMI	74.1	99.8	88.6	72.9	26.1	44.2	60.9
	EMI	74.6	100.0	90.5	69.8	21.5	38.9	57.9
	GRA	71.3	100.0	89.9	79.6	27.0	42.7	58.0
	MUMODIG	90.6	100.0	96.4	90.2	31.3	54.9	78.3
	PGN	72.1	100.0	89.4	80.8	28.0	44.9	58.6
	MEF	83.0	99.9	93.9	82.8	31.9	51.6	70.3
	MDCS-MEF	85.7	99.9	93.5	83.9	34.8	55.8	73.3
	SIA	92.9	100.0	98.1	86.9	35.5	61.9	84.8
	OPS	97.4	100.0	99.2	99.1	63.7	80.1	93.4
	MDCS-OPS	**98.1**	100.0	**99.6**	**99.3**	**66.7**	**82.3**	**93.9**
ViT-B/16	I-FGSM	18.0	37.8	35.5	25.4	99.9	29.0	28.6
	MI	41.2	63.9	58.7	47.3	100.0	50.2	51.0
	MDCS-MI	45.7	70.8	61.3	53.0	100.0	54.0	55.1
	VMI	54.6	68.9	64.7	56.2	100.0	67.5	67.3
	EMI	59.3	75.9	71.8	64.7	100.0	72.2	74.7
	GRA	66.8	76.1	74.3	71.4	99.6	80.3	80.4
	MUMODIG	73.2	78.6	77.2	75.0	99.4	83.7	82.8
	PGN	69.8	79.4	77.4	75.4	99.6	84.2	85.2
	MEF	75.0	81.1	82.8	78.1	99.3	88.2	88.1
	MDCS-MEF	74.8	84.1	84.6	80.1	99.8	88.0	88.0
	SIA	84.2	86.8	86.8	78.3	99.9	91.4	91.5
	OPS	91.2	93.8	93.9	94.4	99.3	95.9	96.0
	MDCS-OPS	**93.9**	**95.9**	**96.0**	**96.6**	99.9	**97.7**	**98.3**

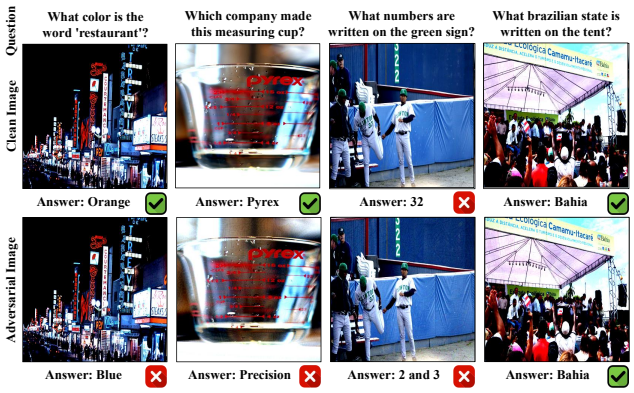

Figure 10. A comparison between responses from the clean image and the adversarial image generated by MDCS-MI against prismVLM. Image and questions source are from textVQA.

Table 6. The white-box adversarial attacks of Pure and OCR VQA tasks in VLMs.

Model	Type	Attack	Pre	Post$_N$	ASR
LLaVA-1.5	Pure	FGSM	47.3	37.2	10.1
		PGD	47.3	22.1	25.2
		MEF	47.3	22.1	25.2
		MI	47.3	20.9	**26.4**
		MDCS-MI	47.3	20.9	**26.4**
	OCR	FGSM	58.5	53.7	4.8
		PGD	58.5	37.5	21.0
		MEF	58.5	36.3	22.2
		MI	58.5	37.2	21.3
		MDCS-MI	58.5	35.4	**23.1**
PrismVLM	Pure	FGSM	56.9	41.3	15.6
		PGD	56.9	26.1	30.8
		MEF	56.9	25.8	31.1
		MI	56.9	26.0	30.9
		MDCS-MI	56.9	22.8	**34.1**
	OCR	FGSM	61.9	49.7	12.2
		PGD	61.9	31.4	30.5
		MEF	61.9	31.1	30.8
		MI	61.9	31.6	30.3
		MDCS-MI	61.9	27.0	**34.9**

Table 7. Detailed comparison of attack success rates on Flickr30K dataset. The gray-shaded cells represent attacks in a white-box setting, while the remaining cells show results for black-box transfer attacks. We report the attack success rates (%) for cross modal retrieval tasks using R@1, R@5, and R@10 metrics. The adversarial perturbations are constrained by an L_∞ norm of 8/255, generated over 10 iterations with a step-size of 2/255.

FLICKR30K (Image-Text Retrieval)

Source	Attack	ALBEF			TCL			CLIP_{ViT}			CLIP_{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	Co-Attack	97.1	94.59	92.60	39.52	20.40	14.53	29.82	11.73	6.61	31.29	11.73	5.77
	SGA	99.79	99.80	99.80	87.67	77.09	70.54	38.04	19.11	13.11	41.63	21.56	14.62
	MDCS-SGA	100.0	100.0	100.0	91.78	82.11	76.75	41.35	22.64	15.55	45.08	25.90	17.71
	DRA	99.79	99.80	99.80	89.78	80.90	75.75	46.63	24.30	18.60	50.32	27.91	20.19
	MDCS-DRA	99.9	99.80	99.80	93.26	85.63	79.26	49.94	28.35	20.43	55.56	32.66	23.38
	SA-AET	99.9	99.80	99.80	96.31	92.56	89.58	54.23	33.13	24.90	58.88	37.53	27.50
	MDCS-SAAET	99.9	99.90	99.90	**96.52**	**94.47**	**91.88**	**60.25**	**38.53**	**29.17**	**60.54**	**42.71**	**32.13**
TCL	Co-Attack	49.84	27.45	60.36	91.68	85.23	80.96	32.64	13.40	6.81	32.06	14.27	8.14
	SGA	93.33	87.17	83.70	100.0	100.0	100.0	37.42	18.28	12.20	42.02	22.73	15.65
	MDCS-SGA	95.10	90.08	87.10	100.0	100.0	100.0	42.45	21.18	15.75	46.87	26.74	20.29
	DRA	95.31	91.18	89.80	100.0	100.0	99.90	46.26	26.38	18.50	50.70	30.66	21.83
	MDCS-DRA	96.66	92.48	90.70	100.0	100.0	100.0	50.92	28.35	20.83	56.19	34.88	25.44
	SA-AET	98.85	96.69	95.40	100.0	100.0	100.0	53.99	33.02	26.42	59.64	39.96	30.38
	MDCS-SAAET	**99.17**	**96.69**	**96.10**	100.0	100.0	100.0	**57.79**	**39.25**	**29.17**	**62.71**	**44.19**	**34.29**
CLIP_{ViT}	Co-Attack	8.55	1.50	0.50	10.01	2.01	0.70	78.5	57.4	45.53	29.50	11.42	6.08
	SGA	21.58	8.02	5.10	24.66	9.45	4.91	100.0	100.0	99.9	52.49	34.04	25.23
	MDCS-SGA	29.30	13.23	8.60	30.56	13.07	8.52	100.0	100.0	100.0	57.60	37.32	28.73
	DRA	27.95	11.92	7.90	29.08	12.56	7.72	100.0	99.90	99.80	62.45	42.81	31.72
	MDCS-DRA	34.41	15.63	11.40	35.51	15.08	9.92	100.0	100.0	100.0	68.45	46.51	37.08
	SA-AET	35.97	19.74	15.30	37.93	19.90	13.83	100.0	100.0	100.0	69.09	50.11	42.12
	MDCS-SAAET	**42.34**	**23.35**	**18.30**	**43.52**	**25.03**	**17.84**	100.0	100.0	100.0	**73.18**	**55.29**	**46.34**
CLIP_{CNN}	Co-Attack	10.53	1.60	0.40	12.54	2.01	0.70	27.24	12.05	6.50	95.91	89.75	85.99
	SGA	15.02	4.71	2.50	18.34	6.33	3.21	39.51	19.83	12.70	100.0	99.58	99.38
	MDCS-SGA	19.19	8.62	5.00	23.92	9.25	6.01	45.89	24.30	17.68	99.87	99.68	99.59
	DRA	19.08	6.41	3.30	22.02	7.14	3.01	48.34	26.48	17.99	100.0	99.58	99.28
	MDCS-DRA	23.04	10.02	6.40	25.82	10.45	6.51	55.83	31.98	22.15	100.0	99.68	99.59
	SA-AET	23.88	9.12	6.00	25.18	10.05	6.91	54.60	31.26	23.78	100.0	100.0	99.79
	MDCS-SAAET	**30.76**	**13.93**	**10.00**	**33.40**	**15.68**	**10.52**	**61.72**	**41.33**	**31.50**	100.0	100.0	100.0

FLICKR30K (Text-Image Retrieval)

Source	Attack	ALBEF			TCL			CLIP_{ViT}			CLIP_{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	Co-Attack	98.36	96.41	94.86	51.24	31.90	24.41	38.92	23.31	17.01	41.99	25.18	18.55
	SGA	99.95	99.88	99.88	87.88	77.07	70.94	46.17	28.40	21.96	50.36	32.24	24.67
	MDCS-SGA	99.98	99.96	99.96	91.24	81.87	75.55	49.71	30.88	23.68	53.93	35.15	27.69
	DRA	99.91	99.92	99.86	90.52	82.05	76.42	57.28	38.15	29.87	59.11	41.41	33.59
	MDCS-DRA	99.98	99.92	99.92	92.98	85.67	80.63	59.31	40.74	31.95	62.44	45.20	36.05
	SA-AET	100.0	99.90	99.90	96.19	92.21	89.32	63.50	45.95	36.81	65.18	48.37	39.87
	MDCS-SAAET	100.0	99.94	99.92	**96.71**	**93.24**	**90.48**	**67.01**	**49.19**	**40.86**	**67.89**	**51.80**	**43.51**
TCL	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	92.84	87.31	83.46	100.0	100.0	100.0	46.39	29.57	22.50	51.36	33.38	25.23
	MDCS-SGA	94.69	89.54	86.78	100.0	100.0	100.0	49.19	31.58	25.05	55.71	37.17	28.33
	DRA	95.35	91.57	89.04	100.0	99.98	99.94	56.80	39.31	31.62	61.54	43.57	35.01
	MDCS-DRA	96.42	92.92	90.96	100.0	100.0	100.0	59.76	41.91	35.02	65.01	46.89	38.60
	SA-AET	98.48	96.82	95.43	99.98	99.96	99.96	63.27	46.13	38.42	68.44	50.68	42.04
	MDCS-SAAET	**98.88**	**97.13**	**96.20**	100.0	100.0	99.98	**65.98**	**49.71**	**42.02**	**71.01**	**54.20**	**45.92**
CLIP_{ViT}	Co-Attack	20.18	9.54	7.12	21.29	9.51	6.78	87.50	77.95	73.40	38.49	23.19	17.87
	SGA	34.89	18.07	13.69	35.83	18.79	14.11	100.0	100.0	100.0	60.38	42.58	35.19
	MDCS-SGA	40.69	21.90	16.28	39.81	22.45	16.02	100.0	100.0	100.0	66.00	48.20	39.71
	DRA	43.29	24.63	18.82	44.83	25.76	19.39	100.0	100.0	99.96	69.47	53.88	45.06
	MDCS-DRA	48.15	28.63	21.69	48.40	28.65	22.24	100.0	100.0	100.0	73.41	57.79	49.24
	SA-AET	50.28	32.08	25.76	51.36	32.77	26.18	100.0	99.98	99.98	74.00	59.22	51.62
	MDCS-SAAET	**54.86**	**36.89**	**29.32**	**55.45**	**36.49**	**29.16**	100.0	99.98	99.96	**77.53**	**63.15**	**55.61**
CLIP_{CNN}	Co-Attack	23.62	11.40	8.21	26.05	12.69	8.79	40.62	24.71	18.82	96.50	92.75	90.35
	SGA	28.60	14.99	10.92	32.26	17.32	12.67	51.16	33.45	25.56	100.0	99.90	99.73
	MDCS-SGA	33.44	17.86	12.90	36.48	19.02	14.09	56.22	38.64	30.14	100.0	99.95	99.86
	DRA	33.96	18.50	13.49	37.45	20.74	15.13	59.02	40.55	32.75	99.93	99.59	99.28
	MDCS-DRA	39.43	22.29	16.60	41.05	23.53	17.83	63.89	45.60	36.89	100.0	99.93	99.82
	SA-AET	37.89	21.68	16.52	41.69	24.33	18.14	63.14	44.45	35.52	99.97	99.78	99.62
	MDCS-SAAET	**45.02**	**27.71**	**21.17**	**47.57**	**30.10**	**22.68**	**69.56**	**52.18**	**43.76**	99.97	99.88	99.82

Table 8. Detailed comparison of attack success rates on MSCOCO dataset. We report the attack success rates (%) for both Image-Text Retrieval and Text-Image Retrieval tasks using R@1, R@5, and R@10 metrics. The adversarial perturbations are constrained by an L_∞ norm of 8/255, generated over 10 iterations with a step size of 2/255.

MSCOCO (Image-Text Retrieval)

Source	Attack	ALBEF			TCL			CLIP_{ViT}			CLIP_{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	Co-Attack	96.65	94.74	93.12	57.33	37.24	28.53	50.71	33.10	26.44	52.06	33.89	27.47
	SGA	99.95	99.70	99.65	87.22	76.39	69.34	64.06	46.76	39.02	63.34	47.67	39.72
	MDCS-SGA	100.0	99.96	99.84	88.23	77.94	70.73	64.94	49.08	40.44	66.12	49.63	41.83
	DRA	99.92	99.68	99.44	88.78	78.81	72.24	69.21	52.56	43.97	69.06	52.13	43.87
	MDCS-DRA	100.0	99.85	99.77	90.29	80.75	74.00	71.77	54.26	45.61	71.19	55.04	46.49
	SA-AET	100.0	99.96	99.96	96.98	93.45	90.11	76.61	61.51	53.76	75.32	61.03	53.19
	MDCS-SAAET	100.0	99.96	99.94	**97.54**	**94.03**	**91.04**	**79.28**	**65.25**	**57.33**	**78.05**	**64.16**	**56.44**
TCL	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	92.78	86.95	83.68	100.0	99.98	99.96	58.68	44.11	36.70	61.05	45.91	38.36
	MDCS-SGA	93.50	88.63	85.36	100.0	100.0	100.0	62.15	46.39	39.56	62.44	48.68	40.94
	DRA	94.61	90.49	87.90	100.0	100.0	99.96	71.08	54.81	46.53	70.29	54.94	45.89
	MDCS-DRA	95.64	91.81	89.05	100.0	100.0	99.98	72.30	56.36	48.76	72.82	57.71	48.58
	SA-AET	97.96	95.78	93.95	100.0	99.98	99.96	76.00	61.54	54.63	75.64	61.00	53.51
	MDCS-SAAET	**98.12**	**95.97**	**94.55**	100.0	99.96	99.90	**78.33**	**64.84**	**57.29**	**79.12**	**65.79**	**56.86**
CLIP_{ViT}	Co-Attack	26.35	11.97	7.20	28.23	12.89	8.19	88.78	78.21	70.98	47.36	31.49	25.29
	SGA	43.88	25.53	17.99	43.73	25.44	18.23	100.0	100.0	100.0	71.60	56.46	48.63
	MDCS-SGA	48.47	28.96	20.96	47.75	28.29	20.71	100.0	100.0	100.0	74.09	59.94	52.22
	DRA	52.08	32.42	23.88	51.67	31.65	23.73	100.0	99.95	99.95	80.79	67.72	60.64
	MDCS-DRA	56.30	36.41	28.00	55.32	34.58	26.21	100.0	99.97	99.98	82.63	71.04	63.75
	SA-AET	57.90	38.93	29.81	57.41	38.14	30.06	99.96	99.97	99.93	84.51	73.48	66.63
	MDCS-SAAET	**62.80**	**44.26**	**34.89**	**62.43**	**42.56**	**34.53**	100.0	99.95	99.93	**85.70**	**76.12**	**70.44**
CLIP_{CNN}	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	36.58	19.14	11.90	38.47	20.38	14.19	62.69	47.46	38.64	99.96	99.84	99.68
	MDCS-SGA	40.53	21.90	14.64	43.28	24.11	17.05	66.58	52.15	44.23	99.96	99.97	99.95
	DRA	41.09	21.58	14.41	43.25	23.59	16.36	70.89	55.21	46.46	99.75	99.70	99.32
	MDCS-DRA	45.60	25.67	18.05	48.17	27.71	20.31	75.70	61.02	52.57	99.96	99.92	99.78
	SA-AET	43.57	24.85	17.35	47.04	26.65	19.24	72.61	57.95	50.20	99.92	99.84	99.76
	MDCS-SAAET	**50.71**	**31.17**	**22.87**	**53.04**	**33.03**	**24.24**	**79.13**	**66.90**	**59.04**	100.0	99.95	99.90

MSCOCO (Text-Image Retrieval)

Source	Attack	ALBEF			TCL			CLIP_{ViT}			CLIP_{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	Co-Attack	98.33	96.60	95.30	64.19	46.17	37.83	57.36	42.19	35.53	60.74	45.90	38.77
	SGA	99.95	99.81	99.73	87.96	76.98	70.35	70.01	55.43	48.06	70.95	56.31	49.00
	MDCS-SGA	99.99	99.95	99.92	88.28	77.93	71.47	70.88	56.85	49.08	72.08	57.69	50.76
	DRA	99.96	99.79	99.66	90.01	80.25	73.80	74.89	61.34	53.78	75.11	61.74	54.39
	MDCS-DRA	99.97	99.89	99.82	90.92	81.56	75.38	76.28	62.76	55.42	77.42	63.53	56.37
	SA-AET	99.99	99.96	99.92	96.87	93.16	90.34	79.99	68.16	61.23	80.86	68.58	61.39
	MDCS-SAAET	99.99	99.96	99.93	**97.25**	**94.02**	**91.46**	**81.91**	**70.70**	**63.88**	**82.51**	**70.60**	**63.92**
TCL	Co-Attack	72.41	56.37	48.16	97.87	95.32	93.42	62.33	46.90	39.68	66.45	49.95	42.72
	SGA	93.07	87.82	84.17	100.0	100.0	100.0	64.89	50.93	43.85	67.46	53.56	46.29
	MDCS-SGA	94.29	89.04	85.66	100.0	100.0	100.0	67.56	52.76	45.77	69.72	56.45	49.05
	DRA	95.74	91.87	89.31	99.99	99.99	99.98	75.14	61.13	53.92	76.87	63.69	56.22
	MDCS-DRA	96.49	92.67	90.14	100.0	99.99	99.99	76.68	63.53	56.24	78.68	65.78	58.75
	SA-AET	98.06	95.93	94.57	99.99	99.95	99.94	79.96	67.85	61.25	81.09	69.43	62.51
	MDCS-SAAET	**98.17**	**96.16**	**94.94**	99.99	99.97	99.93	**81.55**	**70.18**	**63.87**	**83.07**	**71.90**	**65.28**
CLIP_{ViT}	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	51.04	33.62	27.06	50.83	34.42	27.92	100.0	100.0	100.0	75.59	63.21	56.28
	MDCS-SGA	54.76	36.70	29.42	53.25	36.42	29.66	100.0	100.0	100.0	78.42	66.56	59.98
	DRA	61.51	43.47	36.22	60.81	43.89	36.16	100.0	100.0	99.99	84.61	74.06	67.92
	MDCS-DRA	64.52	46.76	38.89	62.95	46.17	38.31	100.0	100.0	99.99	86.00	77.26	71.28
	SA-AET	66.21	49.54	42.24	65.50	49.36	41.87	99.99	99.97	99.95	87.32	77.99	72.27
	MDCS-SAAET	**70.09**	**53.98**	**46.53**	**68.75**	**52.97**	**45.26**	100.0	99.98	99.98	**89.18**	**80.13**	**74.96**
CLIP_{CNN}	Co-Attack	41.50	26.14	20.51	43.44	27.92	21.61	60.15	45.53	38.56	98.54	96.16	94.71
	SGA	46.29	30.00	23.70	48.75	32.86	26.01	67.89	53.23	46.93	99.95	99.85	99.76
	MDCS-SGA	50.04	32.89	26.19	51.72	35.26	27.89	72.24	58.48	51.84	99.97	99.96	99.94
	DRA	52.32	35.71	28.85	54.26	37.89	30.60	74.83	62.53	55.45	99.93	99.76	99.64
	MDCS-DRA	56.64	39.42	32.21	57.84	41.30	33.90	78.85	66.94	60.37	99.97	99.91	99.84
	SA-AET	54.86	38.13	31.21	57.84	40.16	32.96	76.95	64.59	57.86	99.91	99.81	99.65
	MDCS-SAAET	**60.76**	**44.01**	**36.78**	**62.13**	**45.37**	**37.92**	**82.75**	**71.77**	**65.13**	99.97	99.91	99.80