SUPPLEMENTARY MATERIAL: A TWO-STAGE GLOBALLY-DIVERSE ADVERSARIAL ATTACK FOR VISION-LANGUAGE PRE-TRAINING MODELS

Wutao Chen[1], *Huaqin Zou*[1], *Chen Wan*[1*], *Lifeng Huang*[2]

[1]Department of Computer Science and Technology, Shantou University, Shantou, China
[2]College of Mathematics and Informatics, South China Agricultural University, Guangzhou, China

1. Further Details for Fig. 2

The experiments in Fig. 2 are conducted on the Flickr30K dataset. For the image modality, perturbations are generated using PGD with maximum perturbation $\varepsilon_v = 8/255$, step size $2/255$, and $T = 10$ iterations. For the text modality, the perturbation bound is set to $\varepsilon_t = 1$, with a candidate list size of $W = 10$. Adversarial examples are generated by attacking the ALBEF model at each stage (S1, S2, and S3) of the SGA framework. Fig. 2a presents the attack success rates (ASR) of these adversarial examples on four target models (ALBEF, TCL, CLIP_{ViT}, and CLIP_{CNN}). Furthermore, we compute the Euclidean distance and cosine similarity between adversarial examples (generated by S2 and S3) and the original inputs in the ALBEF feature space, and visualize their probability density distributions (PDD) in Figs. 2b and 2c.

2. Algorithm

The detailed procedure of 2S-GDA is presented in Algorithm 1.

Algorithm 1 2S-GDA

Input: Image encoder $\mathcal{F}_I$, Text encoder $\mathcal{F}_T$, Dataset D, Image-caption pair (v, t), Image scale sets $S = \{s_1, s_2, ..., s_N\}$, iteration steps T, number of paired captions M.
Output: adversarial image v', adversarial caption t'.
Build caption set $t = \{t_1, t_2, ..., t_M\}$ from D
/* Build adversarial caption set $t' = \{t'_1, t'_2, ..., t'_M\}$ by Bert-Attack */
for $i = 1$ to M **do**
  /* identify Top-k important words by MLM */
  $W_{\text{imp}} = \{w_1, w_2, \ldots, w_k\}$
  /* generate candidates by MLM and WordNet */
  $C(w) = \{c \in C_{\text{MLM}}(w) \cup C_{\text{WN}}(w) | c \neq R\}$
  $(w'_i \to \tilde{w}') = \underset{w_i \in W_{\text{imp}}, \tilde{w}}{\arg\max} \mathcal{J}(v, t : w_i \to \tilde{w} | \tilde{w} \in C(w_i))$
  $t'_i = \underset{t'_i \in \mathcal{B}[t_i, \varepsilon_t]}{\arg\max} \left(-\frac{\mathcal{F}_T(t'_i) \cdot \mathcal{F}_I(v)}{\|\mathcal{F}_T(t'_i)\| \cdot \|\mathcal{F}_I(v)\|} \right)$
end for
/* Generate adversarial image v' by PGD */
for $k = 1$ to T **do**
  /* Build image set $\mathcal{V}'_{\text{aug}}$ by applying BSR to v' */
  $v' = \underset{v' \in B[v, \varepsilon_v]}{\arg\max} -\sum_{i=1}^{M} \frac{\mathcal{F}_T(t'_i)}{\|\mathcal{F}_T(t'_i)\|} \sum_{s_i \in \mathcal{S}} \frac{\mathcal{F}_I(h(\mathcal{V}'_{\text{aug}}, s_i))}{\|\mathcal{F}_I(h(\mathcal{V}'_{\text{aug}}, s_i))\|}$
end for

3. Comparison with Prior Methods

Table A1 compares the attack success rates (ASR) of our proposed 2S-GDA method with four prior approaches, i.e., PGD, BERT-Attack, Sep-Attack, and Co-Attack. As observed, 2S-GDA consistently outperforms all four baselines in terms of ASR, demonstrating the effectiveness of our approach.

4. Visualization

Additional visualization results are provided to further validate the effectiveness of the proposed method. Fig. A1 presents a comparison between SGA-BSR and the proposed 2S-GDA on the image-text retrieval task. The proposed method leads to a more noticeable disruption of visual-text alignment. Figs. A2 and A3 present the experimental results of salient regions for the highlighted word in the caption, using ALBEF on both clean and adversarial images. In contrast to the baseline, the proposed method induces a clear shift in the model attention.

5. Additional Ablation Studies

Fig. A4 presents the impact of the number of influential tokens k in GAR on attack success rates against TCL and CLIP_{CNN}, where the adversarial examples are generated by ALBEF. As k increases from 1 to 20, the performance remains relatively stable in most cases, with the highest results generally achieved when $k = 3$. The baseline without GAR (w/o GAR) is included for comparison, showing that GAR consistently improves transferability across both image-to-text and text-to-image retrieval tasks.

Table A1: ASR (%) of our method and the baseline approaches on Flickr30k dataset. Diagonal entries indicate white-box attacks.

Source	Attack	ALBEF		TCL		CLIP_{ViT}		CLIP_{CNN}	
		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	PGD	90.09	92.00	18.86	24.07	10.06	14.24	12.13	17.08
	BERT-Attack	8.97	20.86	8.22	19.57	20.74	36.18	24.78	38.28
	Sep-Attack	94.16	94.20	26.13	39.02	22.82	36.34	27.33	39.90
	Co-Attack	94.26	97.47	33.19	48.12	25.28	39.95	28.22	42.78
	2S-GDA (ours)	**100.00**	**100.00**	**97.26**	**96.40**	**74.85**	**78.77**	**77.01**	**81.34**
TCL	PGD	31.91	38.40	98.95	99.33	9.08	15.30	13.03	18.97
	BERT-Attack	9.91	23.64	9.38	23.50	25.89	39.79	28.35	41.99
	Sep-Attack	46.51	58.21	99.58	99.48	27.36	41.17	30.40	44.29
	Co-Attack	39.21	54.58	95.36	98.10	27.73	39.98	30.78	44.01
	2S-GDA (ours)	**97.50**	**98.13**	**100.00**	**100.00**	**73.37**	**79.48**	**78.67**	**83.70**
CLIP_{ViT}	PGD	3.23	6.06	4.95	8.07	63.80	83.92	10.22	14.00
	BERT-Attack	7.30	18.71	8.11	18.74	22.45	30.67	22.48	30.98
	Sep-Attack	6.99	20.09	9.91	20.33	72.52	86.95	25.67	36.30
	Co-Attack	7.19	20.07	9.69	20.48	73.25	85.47	25.42	36.57
	2S-GDA (ours)	**54.54**	**65.18**	**55.43**	**65.45**	**100.00**	**99.97**	**90.04**	**89.98**
CLIP_{CNN}	PGD	2.29	5.43	4.74	7.71	3.31	10.24	86.97	92.42
	BERT-Attack	8.45	21.73	10.22	22.69	24.66	35.86	27.46	38.77
	Sep-Attack	9.59	22.71	12.43	24.55	24.29	38.92	93.87	95.03
	Co-Attack	10.11	22.38	12.43	24.83	25.89	38.82	92.72	96.26
	2S-GDA (ours)	**28.78**	**46.02**	**34.35**	**49.05**	**67.61**	**75.19**	**100.00**	**99.90**

Table A2: ASR (%) of our method and the baseline approaches on MSCOCO dataset, with the adversarial examples generated by attacking ALBEF.

Attack	ALBEF		TCL		CLIP_{ViT}		CLIP_{CNN}	
	TR R@1	IR R@1	TR R@1	IR R@1	TR R@1	IR R@1	TR R@1	IR R@1
PGD	91.36	92.04	28.49	31.79	17.78	24.86	19.70	27.37
Bert-Attack	25.29	36.94	24.76	34.66	47.00	55.54	47.81	56.49
Sep-Attack	94.72	95.02	45.74	53.90	50.10	57.20	50.02	59.49
Co-Attack	88.5	94.53	39.18	63.08	43.84	73.86	46.3	76.82
SGA	99.87	99.93	88.28	88.56	65.93	71.24	65.30	71.82
SGA-BSR	99.87	99.92	91.48	92.29	81.00	84.42	81.57	85.16
DRA	99.95	99.90	89.74	90.27	69.25	75.88	70.21	76.25
SA-AET	**100.0**	**100.0**	**96.83**	**96.88**	78.29	81.63	77.36	82.21
2S-GDA (ours)	99.92	99.97	95.48	95.34	**86.49**	**89.41**	**87.05**	**89.74**

6. Experiments on the MSCOCO Dataset

We further evaluate the proposed 2S-GDA on the MSCOCO dataset to demonstrate its effectiveness. ALBEF is adopted as the surrogate model, while TCL, CLIP_{ViT}, and CLIP_{CNN} serve as black-box targets. Our method is compared with several state-of-the-art baselines in the same experimental settings.

For the image modality, adversarial examples are crafted using PGD with an ℓ_∞ perturbation bound of $8/255$, 10 steps, and a step size of $2/255$. For the text modality, we set the candidate pool size to $W = 10$ and the perturbation budget to $\varepsilon_t = 1$.

The experimental results are summarized in Table A2. As seen, our method consistently achieves higher ASR than the existing baselines, demonstrating that 2S-GDA generalizes well to large-scale datasets and maintains strong transferability.

Fig. A1: Comparison between SGA-BSR and the proposed 2S-GDA on the image-text retrieval task. The adversarial examples are generated using ALBEF and evaluated on CLIP_{ViT}.

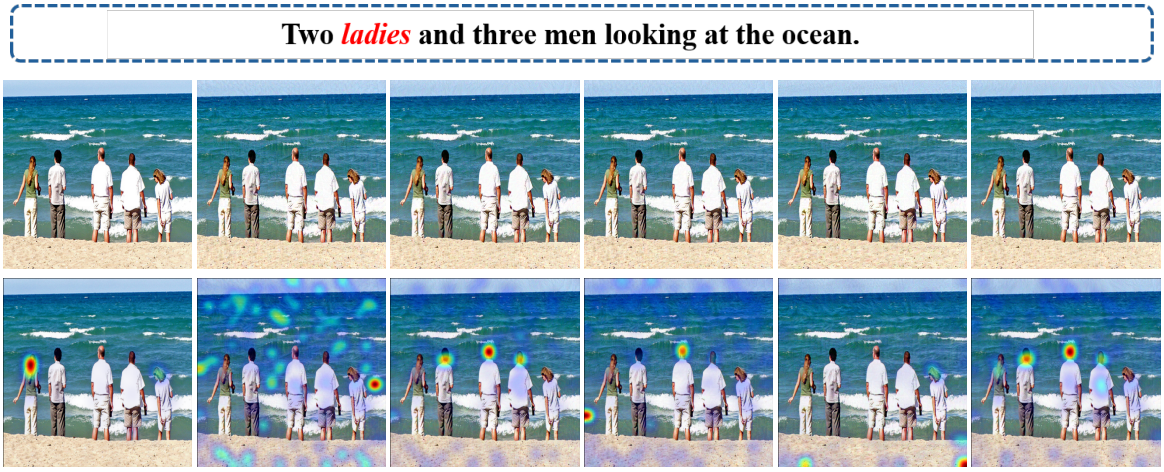


Fig. A2: Visualization of the original and adversarial examples. The first row shows the input captions, the second row displays the corresponding clean and adversarial images, and the third row highlights the salient regions for the word "ladies".

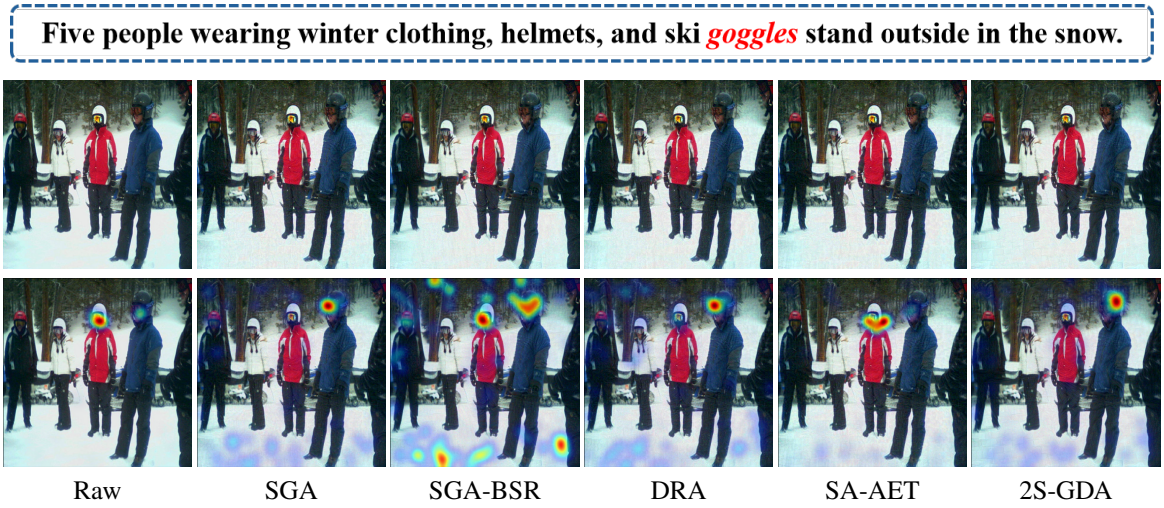
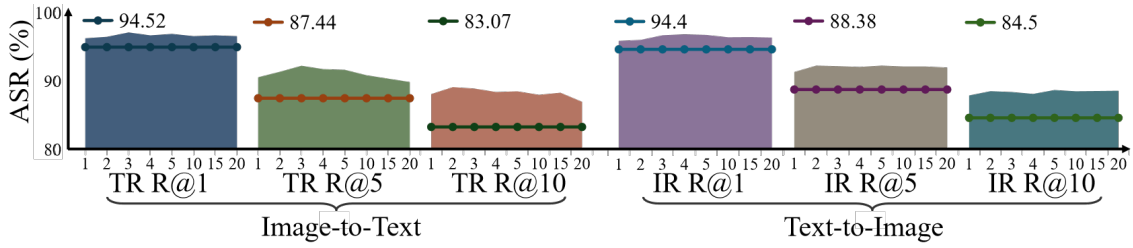
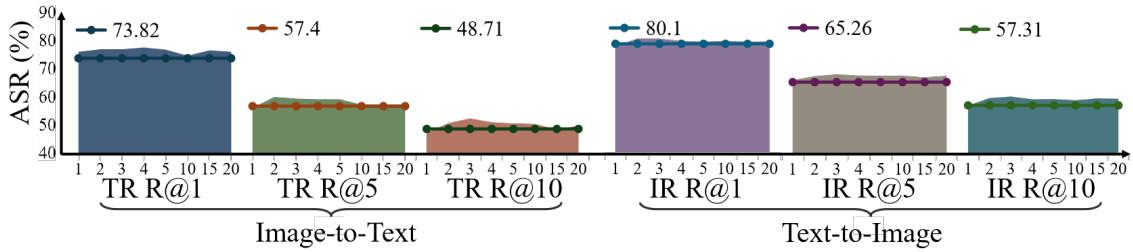


Fig. A3: Visualization of the original and adversarial examples. The first row shows the input captions, the second row displays the corresponding clean and adversarial images, and the third row highlights the salient regions for the word "goggles".



(a) ALBEF→TCL



(b) ALBEF→CLIP_{CNN}

Fig. A4: Comparison of attack success rates under different values of influential token count k in GAR, with the baseline (w/o GAR) included for reference. The notation "Model A → Model B" denotes that adversarial examples are generated by "Model A" and then evaluated on "Model B".