Set-level Guidance Attack: Boosting Adversarial Transferability of Vision-Language Pre-training Models

A. Motivation

The analysis and discussion presented in Section 3 shed light on the keys to improving adversarial transferability among VLP models: multimodal interaction and diverse data. To further figure out a practicable solution, we delve into the cases of transfer failure of existing attack methods. We find that half of the failure cases are raised by the existence of multiple paired captions.

Considering an image-text pair (v, t), the corresponding adversarial image v' generated on model f_{wb} in white-box manner (note that only (v, t) and f_{wb} are utilized in the process of crafting v'), a black-box model f_{bb} and another several matched captions $\mathbf{t} = \{t_1, ..., t_n\}$, we define two events:

- *Event A*: adversarial image v' cannot match anyone of the captions $t \cup \{t_1, ..., t_n\}$ in model f_{wb}.
- *Event B*: adversarial image v' can match one of captions $\{t_1, ..., t_n\}$ in model f_{bb}.

Event A indicates that the adversarial image successfully fools model f_{wb}, successful case of white-box attack. *Event B* indicates that the adversarial image cannot fully fool the target model f_{bb} in a transferring manner, failure case of transfer-based black-box attack. We present the statistic figures of $p($*Event A*$)$ and $p($*Event B* $\mid$ *Event A*$)$ in Table A. As shown in the table, even though the adversarial images have high attack ability in the white-box model (about 71% - 80% error rate), around half of them fail due to matching other paired captions when transferring to a black-box model (about 46%-57%).

In detail, existing attacks tend to restrict the generated adversarial image v' far away (Euclidean distance or cosine distance in the embedding space in most of the cases) from the original image v or the caption t. These methods only utilize the information of a single image-caption pair (v, t) in their processes of crafting adversarial examples. As a result, although in most of the cases v' is far away from t and other paired captions $\{t_1, ..., t_n\}$ in the embedding space of the white-box model, it is prone to approaching $\{t_1, ..., t_n\}$ when transferred to a black-box model and the embedding space changes, which means the failure of transfer-based black-box attack.

We attribute the failure of the transfer attack to the lack of cross-modal interaction (corresponding to the first two rows in Table A). The adversarial image v' generated merely based on image v or single image-text pair (v, t) can have strong attack ability to the caption t and always weak attack ability to $\{t_1, ..., t_n\}$. When transferred to a black-box model, the adversarial image v' may still maintain satisfactory attack ability to caption t but most likely to lose the attack ability to captions $\{t_1, ..., t_n\}$. Note that v' has the attack ability to t in model f means v' cannot successfully match t in the embedding space of model f. To validate the claim, in Figure A, we use the ranking to measure the adversarial image's attack ability to the caption. Higher ranking, stronger attack ability. Since an adversarial image has several paired captions in the gallery, we present the lowest, average, and highest ranking of these captions. As shown in Figure A, for the attack method with no cross-modal interaction (Sep-Attack) and the attack method with single-pair cross-modal interaction (Co-Attack), though the generated adversarial image can have a high attack ability to some captions, there always exists a caption that the adversarial image has weak attack ability to it (the lowest rankings of Sep-attack and Co-Attack are both around 600, which means weak attack ability compared the highest rankings of them, around 2,200 and 2,400).

An implicit assumption in the previous statement is that high attack ability in the white-box model means high adversarial transferability in the black-box model, which can also be verified among existing attack methods. Considering a image-caption pair (v, t), the corresponding adversarial image v' generated on the white-box model f_{wb} , a black-model f_{bb} and several matched captions of v, $\{t_1, ..., t_n\}$, we define two events:

- *Event C*: adversarial image v' cannot match t in white-box model f_{wb}.
- *Event D*: adversarial image v' cannot match t in black-box model f_{bb}.

We present the statistic figures of $p($*Event C*$)$ and $p($*Event D* $\mid$ *Event C*$)$ in Table B. If the adversarial image v' has a

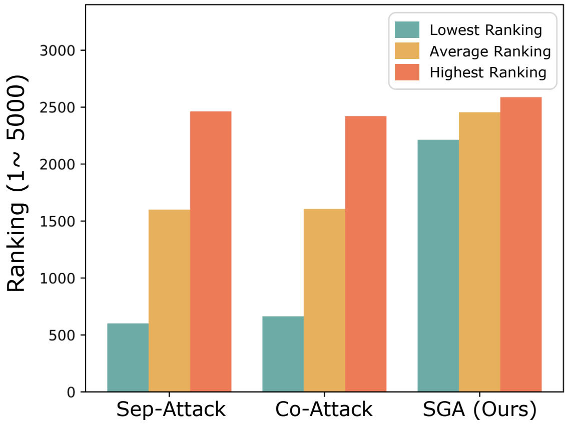

Figure A: The adversarial image may have weak attack ability to some paired captions. The experiment is conducted on model ALBEF, dataset Flickr30K.

Attack	Cross-modal Interaction	*p (Event A)* ↑	*p (Event B* | *Event A)* ↓		
		ALBEF	TCL	CLIP_{ViT}	CLIP_{CNN}
Sep-Attack	No	74.60%	46.78%	41.69%	41.82%
Co-Attack	Single-pair	80.60%	50.50%	40.82%	42.93%
SGA (Ours)	Set-level	**97.20%**	**28.91%**	**34.67%**	**38.58%**

Table A: ***Event A***: adversarial image v' cannot match anyone of $t \cup \{t_1, ..., t_n\}$ in white-box model f_{wb}. ***Event B***: adversarial image v' can match one of $\{t_1, ..., t_n\}$ in black-box model f_{bb}. The adversarial data is generated on model ALBEF, dataset Flickr30K.

high attack ability to caption t in the white-box model, it is very likely that it also maintains the attack ability towards caption t when transferred to a black-box model. For example, if the adversarial image v' generated on model ALBEF succeeds in attacking caption t in model ALBEF, there is a high probability that it can succeed in attacking caption t in model TCL, 40.51%, compared to the overall adversarial transferability from ALBEF to TCL, 15.21%.

According to the analysis above, to boost the adversarial transferability of the generated adversarial image, it is crucial to consider multiple paired captions and push the adversarial image away from all the paired captions, thus preserving the attack ability when transferring to other black-box models. Crafting adversarial captions for high transferability follows a similar approach, which can also benefit from more paired images.

B. Experiments & Analysis

B.1. Experimental Settings

Since fused VLP models contain both multimodal encoder and unimodal encoder, two types of embedding can be perturbed, *i.e.*, multimodal embedding, and unimodal embedding. The embeddings can be further divided into the full embedding (denoted as $\texttt{Multi}_{\text{full}}$ or $\texttt{Uni}_{\text{full}}$) and [CLS] of embedding (denoted as $\texttt{Multi}_{\text{CLS}}$ or $\texttt{Uni}_{\text{CLS}}$). For aligned VLP models (*e.g.*, CLIP), since the image encoder can be ViT or CNN, only [CLS] of embedding for CLIP_{ViT} is discussed and consider the embedding of CLIP_{CNN} as [CLS] of embedding.

Attack	Cross-modal Interaction	*p (Event C)* ↑	*p (Event D* | *Event C)* ↑		
		ALBEF	TCL	CLIP_{ViT}	CLIP_{CNN}
Sep-Attack	No	83.30%	33.37%	38.77%	45.38%
Co-Attack	Single-pair	89.60%	40.51%	39.62%	47.32%
SGA (Ours)	Set-level	**98.20%**	**54.38%**	**46.84%**	**55.61%**

Table B: ***Event C***: adversarial image v' cannot match caption t in white-box model. ***Event D***: adversarial image v' cannot match caption t in black-box model. The adversarial data is generated on model ALBEF, dataset Flickr30K.

B.2. Transferability Analysis

Table C, Table D, Table E, and Table F show adversarial transferability among different VLP models and configurations under Sep-Attack and Co-Attack. We report the attack success rates of the adversarial examples generated by the source model to attack the target models.

Some observations on adversarial transferability are summarized below:

- For all VLP models, attacking two modalities simultaneously shows better adversarial transferability than only attacking a single modality. This is consistent with the observation for the white-box setting.
- Even though models with exact same architectures but with different pretrain objectives (*e.g.*, ALBEF and TCL), the adversarial examples cannot directly pass through another model with a similar success attack rate.
- The adversarial transferability from fused VLP models to aligned VLP models is higher than that from backward (*e.g.*, from ALBEF or TCL to CLIP_{ViT} and CLIP_{CNN}).
- Although ALBEF, TCL, and CLIP_{ViT} are using ViT as image-encoders, the adversarial transferability from ALBEF or TCL to CLIP_{CNN} will be higher than that of CLIP_{ViT}; similarly, the adversarial transferability of CLIP_{ViT} to CLIP_{CNN} is higher than that of CLIP_{CNN} to CLIP_{ViT}.

B.3. Main Results

We present a thorough analysis of the performance of our proposed high transferable multimodal attack method,

Sep-Attack								
Source	Attack	Target	Image-to-Text			Text-to-Image		
			R@1	R@5	R@10	R@1	R@5	R@10
ALBEF	**Text**@Uni$_{\text{full}}$	ALBEF	**8.34***	**1.40***	**0.60***	**21.19***	**11.36***	**9.18***
		TCL	7.90	1.21	0.30	19.45	8.87	6.26
		CLIP$_{\text{ViT}}$	23.31	9.55	4.88	36.05	20.77	15.98
		CLIP$_{\text{CNN}}$	26.05	9.73	5.97	38.04	21.83	16.85
	Image@Uni$_{\text{full}}$	ALBEF	**62.46***	**50.70***	**45.00***	**68.73***	**57.38***	**52.12***
		TCL	5.48	1.21	0.80	10.43	3.33	1.89
		CLIP$_{\text{ViT}}$	7.36	1.66	0.61	13.18	5.21	3.10
		CLIP$_{\text{CNN}}$	10.09	2.85	1.24	15.54	6.28	3.61
	Bi@Uni$_{\text{full}}$	ALBEF	**68.93***	**55.21***	**49.40***	**76.33***	**65.46***	**59.66***
		TCL	16.86	3.32	1.70	27.07	13.27	8.79
		CLIP$_{\text{ViT}}$	25.40	9.55	4.88	36.15	20.93	15.57
		CLIP$_{\text{CNN}}$	26.82	9.73	6.49	38.80	22.34	16.90
	Text@Multi$_{\text{full}}$	ALBEF	**15.43***	**2.91***	**1.40***	**30.54***	**16.41***	**12.66***
		TCL	12.64	2.21	0.60	28.64	14.62	10.40
		CLIP$_{\text{ViT}}$	26.75	10.49	5.28	41.33	24.62	19.15
		CLIP$_{\text{CNN}}$	30.27	12.16	7.11	43.43	26.64	20.96
	Image@Multi$_{\text{full}}$	ALBEF	**35.97***	**25.35***	**21.40***	**50.54***	**40.57***	**37.24***
		TCL	1.79	0.50	0.20	6.50	1.88	1.10
		CLIP$_{\text{ViT}}$	7.12	1.56	0.30	13.02	5.05	3.03
		CLIP$_{\text{CNN}}$	9.83	2.85	1.34	14.75	5.56	3.23
	Bi@Multi$_{\text{full}}$	ALBEF	**51.09***	**36.97***	**31.90***	**64.17***	**52.87***	**48.73***
		TCL	16.86	4.02	1.30	32.57	16.77	12.08
		CLIP$_{\text{ViT}}$	27.48	10.70	5.69	41.78	25.02	18.88
		CLIP$_{\text{CNN}}$	31.16	12.16	6.90	43.77	26.56	21.12
	Text@Uni$_{\text{CLS}}$	ALBEF	**11.57***	**1.80***	**1.10***	**27.46***	**14.48***	**10.98***
		TCL	12.64	2.51	0.90	28.07	14.39	10.26
		CLIP$_{\text{ViT}}$	29.33	11.63	6.30	43.17	26.37	19.91
		CLIP$_{\text{CNN}}$	32.69	15.43	8.65	46.11	28.43	22.14
	Image@Uni$_{\text{CLS}}$	ALBEF	**52.45***	**36.57***	**30.00***	**58.65***	**44.85***	**38.98***
		TCL	3.06	0.40	0.10	6.79	2.21	1.20
		CLIP$_{\text{ViT}}$	8.96	1.66	0.41	13.21	5.19	3.05
		CLIP$_{\text{CNN}}$	10.34	2.96	1.85	14.65	5.60	3.39
	Bi@Uni$_{\text{CLS}}$	ALBEF	**65.69***	**47.60***	**42.10***	**73.95***	**59.50***	**53.70***
		TCL	17.60	3.72	1.90	32.95	17.10	11.90
		CLIP$_{\text{ViT}}$	31.17	12.05	7.01	45.23	25.93	19.95
		CLIP$_{\text{CNN}}$	32.82	15.86	9.06	45.49	28.43	22.32
	Text@Multi$_{\text{CLS}}$	ALBEF	**15.43***	**2.81***	**1.30***	**30.47***	**15.85***	**11.85***
		TCL	13.59	3.02	1.20	30.26	15.42	11.09
		CLIP$_{\text{ViT}}$	27.12	11.94	6.50	42.53	25.20	19.36
		CLIP$_{\text{CNN}}$	30.78	13.21	7.52	44.39	28.07	21.89
	Image@Multi$_{\text{CLS}}$	ALBEF	**30.76***	**21.24***	**17.10***	**43.85***	**34.84***	**31.44***
		TCL	2.53	0.20	0.00	6.74	1.98	1.20
		CLIP$_{\text{ViT}}$	7.98	1.35	0.30	12.85	5.00	3.16
		CLIP$_{\text{CNN}}$	9.96	2.64	1.75	14.92	5.65	3.37
	Bi@Multi$_{\text{CLS}}$	ALBEF	**42.13***	**26.95***	**22.20***	**57.76***	**44.91***	**39.95***
		TCL	16.65	3.92	1.90	34.02	17.16	12.10
		CLIP$_{\text{ViT}}$	28.71	11.42	6.30	42.01	24.90	18.99
		CLIP$_{\text{CNN}}$	31.03	14.16	8.96	43.98	27.17	21.30

Table C: **Attack success rates** (%) with different adversarial input modalities under Sep-Attack on image-text retrieval. The adversaries are crafted on ALBEF using Flickr30K. * indicates white-box attacks. A higher ASR indicates better adversarial transferability.

Sep-Attack								
Source	**Attack**	**Target**	**Image-to-Text**			**Text-to-Image**		
			R@1	R@5	R@10	R@1	R@5	R@10
TCL	`Text@Uni`$_{\text{full}}$	TCL	**9.48***	**1.51***	**0.60***	**23.50***	**11.83***	**8.53***
		ALBEF	9.91	1.80	0.70	23.64	12.80	10.07
		CLIP_{ViT}	25.89	8.41	4.57	39.79	24.22	18.62
		CLIP_{CNN}	28.35	11.21	7.11	41.96	26.30	20.42
	`Image@Uni`$_{\text{full}}$	TCL	**45.10***	**34.07***	**28.76***	**53.21***	**38.27***	**32.49***
		ALBEF	3.86	0.90	0.20	7.62	2.40	1.29
		CLIP_{ViT}	7.12	1.66	0.71	12.82	5.35	3.14
		CLIP_{CNN}	9.07	2.54	1.75	15.68	5.70	3.39
	`Bi@Uni`$_{\text{full}}$	TCL	**55.95***	**39.80***	**33.77***	**65.38***	**49.58***	**42.28***
		ALBEF	15.02	4.01	2.60	30.47	17.04	13.28
		CLIP_{ViT}	27.48	8.10	4.37	39.85	24.50	18.23
		CLIP_{CNN}	29.50	11.21	7.52	42.13	26.47	20.53
	`Text@Multi`$_{\text{full}}$	TCL	**12.86***	**2.81***	**1.00***	**30.33***	**15.32***	**10.89***
		ALBEF	13.24	2.61	1.20	27.13	15.16	11.28
		CLIP_{ViT}	26.75	9.24	4.57	40.85	24.57	18.77
		CLIP_{CNN}	28.35	11.31	6.49	42.95	26.13	20.71
	`Image@Multi`$_{\text{full}}$	TCL	**52.05***	**41.81***	**35.47***	**63.05***	**51.46***	**46.67***
		ALBEF	4.38	1.50	0.90	9.87	3.28	2.10
		CLIP_{ViT}	7.73	1.97	0.41	13.56	5.68	3.34
		CLIP_{CNN}	9.32	2.64	1.44	14.92	5.70	3.32
	`Bi@Multi`$_{\text{full}}$	TCL	**61.96***	**48.64***	**42.08***	**71.74***	**61.07***	**55.83***
		ALBEF	19.29	6.21	3.00	35.17	19.71	14.96
		CLIP_{ViT}	26.75	9.55	5.08	41.37	24.78	18.71
		CLIP_{CNN}	30.78	11.31	7.42	43.53	26.23	20.42
	`Text@Uni`$_{\text{CLS}}$	TCL	**14.54***	**2.31***	**0.60***	**29.17***	**15.03***	**10.91***
		ALBEF	11.89	2.20	0.70	26.82	14.09	10.80
		CLIP_{ViT}	29.69	12.77	7.62	44.49	27.47	21.00
		CLIP_{CNN}	33.46	14.38	9.37	46.07	29.28	22.59
	`Image@Uni`$_{\text{CLS}}$	TCL	**77.87***	**65.13***	**58.72***	**79.48***	**66.26***	**60.36***
		ALBEF	6.15	1.30	0.70	10.78	3.36	1.70
		CLIP_{ViT}	7.48	1.45	0.81	13.72	5.37	3.01
		CLIP_{CNN}	10.34	2.75	1.54	15.33	5.77	3.28
	`Bi@Uni`$_{\text{CLS}}$	TCL	**84.72***	**73.07***	**65.43***	**86.07***	**74.67***	**68.83***
		ALBEF	20.13	4.91	2.70	36.48	19.48	14.82
		CLIP_{ViT}	31.29	12.98	7.72	44.65	26.82	20.37
		CLIP_{CNN}	33.33	14.27	9.89	45.80	29.18	23.02
	`Text@Multi`$_{\text{CLS}}$	TCL	**18.34***	**4.02***	**1.90***	**33.90***	**16.68***	**11.78***
		ALBEF	13.66	2.30	0.90	27.90	14.11	10.31
		CLIP_{ViT}	27.85	11.32	6.71	42.01	24.95	18.88
		CLIP_{CNN}	30.27	13.95	8.34	44.32	27.58	21.23
	`Image@Multi`$_{\text{CLS}}$	TCL	**37.41***	**28.04***	**24.15***	**48.93***	**39.01***	**35.72***
		ALBEF	2.92	0.90	0.50	8.07	2.65	1.62
		CLIP_{ViT}	7.48	1.77	0.41	13.34	4.91	3.10
		CLIP_{CNN}	9.58	3.07	1.54	15.40	5.26	3.25
	`Bi@Multi`$_{\text{CLS}}$	TCL	**47.31***	**34.77***	**29.66***	**60.31***	**48.07***	**43.36***
		ALBEF	18.87	5.21	2.70	34.03	18.17	13.12
		CLIP_{ViT}	28.47	11.63	6.30	42.53	25.53	19.06
		CLIP_{CNN}	31.03	14.06	8.55	44.46	27.00	20.74

Table D: **Attack success rates** (%) with different adversarial input modalities under Sep-Attack on image-text retrieval. The adversaries are crafted on TCL using Flickr30K. * indicates white-box attacks. A higher ASR indicates better adversarial transferability.

Sep-Attack								
Source	Attack	Target	Image-to-Text			Text-to-Image		
			R@1	R@5	R@10	R@1	R@5	R@10
CLIP_{ViT}	`Text@Uni`	CLIP_{ViT}	**28.34***	**11.73***	**6.81***	**39.08***	**24.08***	**17.44***
		CLIP_{CNN}	30.40	11.63	5.97	37.43	24.96	18.66
		ALBEF	9.59	1.30	0.40	22.64	10.95	8.17
		TCL	11.80	1.91	0.70	25.07	12.92	8.90
	`Image@Uni`	CLIP_{ViT}	**70.92***	**50.05***	**42.28***	**78.61***	**60.78***	**51.50***
		CLIP_{CNN}	5.36	1.16	0.72	8.44	2.35	1.54
		ALBEF	2.50	0.40	0.10	4.93	1.44	1.01
		TCL	4.85	0.20	0.20	8.17	2.27	1.46
	`Bi@Uni`	CLIP_{ViT}	**79.75***	**63.03***	**53.76***	**86.79***	**75.24***	**67.84***
		CLIP_{CNN}	30.78	12.16	6.39	39.76	25.62	19.34
		ALBEF	9.59	1.30	0.50	23.25	11.22	8.01
		TCL	11.38	2.11	0.90	25.60	12.92	9.14
CLIP_{CNN}	`Text@Uni`	CLIP_{CNN}	**30.40***	**13.00***	**7.31***	**40.10***	**26.71***	**20.85***
		CLIP_{ViT}	27.12	11.21	6.81	37.44	23.48	17.66
		ALBEF	8.86	1.50	0.60	23.27	11.34	8.41
		TCL	12.33	2.01	0.90	25.48	13.25	8.81
	`Image@Uni`	CLIP_{CNN}	**86.46***	**69.13***	**61.17***	**92.25***	**81.00***	**75.04***
		CLIP_{ViT}	1.10	0.52	0.41	6.60	2.73	1.48
		ALBEF	2.09	0.30	0.10	4.82	1.29	0.87
		TCL	4.00	0.40	0.20	7.81	2.09	1.34
	`Bi@Uni`	CLIP_{CNN}	**91.44***	**78.54***	**71.58***	**95.44***	**88.48***	**82.88***
		CLIP_{ViT}	28.34	10.8	6.30	39.43	24.34	18.36
		ALBEF	8.55	1.50	0.60	23.41	11.38	8.23
		TCL	12.64	1.91	0.70	26.12	13.44	8.96

Table E: **Attack success rates** (%) with different adversarial input modalities under Sep-Attack on image-text retrieval. The adversaries are crafted on CLIP using Flickr30K. * indicates white-box attacks. A higher ASR indicates better adversarial transferability.

SGA, on the popular benchmark datasets Flickr30K and MSCOCO. The experimental results are summarized in Table G and Table H, providing a clear comparison between the performance of our SGA and existing multimodal attack methods across different attack scenarios. As we can see, our proposed SGA outperforms the state-of-the-art in all white-box and black-box settings. Moreover, as illustrated in Table I, we conduct extensive experiments on Flickr30K under a unimodal scenario, with perturbed input in either the image or text modality. Empirical evidence suggests that even in scenarios where only query data are accessible, the performance of SGA consistently surpasses that of existing methods.

Our results suggest that the proposed SGA can serve as a promising method for evaluating the robustness of multimodal models and improving their security in real-world applications.

B.4. Ablation Study

This section presents the ablation experiments on the augmented multimodal data and the iterative strategy sequence of SGA. To provide a thorough analysis, detailed experimental results are presented and discussed.

Iterative Strategy. In this study, we generate adversarial examples through cross-modal guidance. This allows for the disruption of multimodal interactions through the collaborative generation of perturbations. Notably, our process follows a "text-image-text" (t-i-t) pipeline.

We have conducted additional experiments to evaluate the effectiveness of our attack strategy. As shown in Table J, an interesting observation is that reversing the "t-i-t" pipeline does not significantly impact the results. Furthermore, although adding one iteration (t-i-t-i-t) slightly enhances performance, it doubles the computational cost. This suggests that our SGA is not sensitive to the exact order of the pipeline, but rather benefits from cross-modal guidance.

Multi-scale Image Set. In SGA, an augmented image set is used to generate adversarial data based on the scale-invariant property of deep learning models. To verify the effectiveness of the augmented image set, we choose different scale ranges to build the image sets and evaluate the adversarial

Co-Attack								
Source	**Attack**	**Target**	**Image-to-Text**			**Text-to-Image**		
			R@1	R@5	R@10	R@1	R@5	R@10
ALBEF	`Text`@Multi	ALBEF	**9.18***	**1.50***	**1.00***	**21.70***	**11.96***	**9.22***
		TCL	9.38	1.31	0.30	20.40	9.74	6.80
		CLIP_{ViT}	20.98	7.79	4.57	31.73	19.13	14.63
		CLIP_{CNN}	22.48	7.40	4.12	31.94	21.36	15.59
	`Image`@Multi	ALBEF	**75.50***	**59.22***	**53.30***	**83.63***	**75.14***	**70.32***
		TCL	4.64	1.21	0.50	11.33	3.72	2.25
		CLIP_{ViT}	7.24	1.97	0.51	13.53	5.23	3.01
		CLIP_{CNN}	10.09	2.85	1.65	15.27	6.11	3.52
	`Bi`@Multi	ALBEF	**77.16***	**64.60***	**58.37***	**83.86***	**74.63***	**70.13***
		TCL	15.21	4.19	1.47	29.49	14.97	10.55
		CLIP_{ViT}	23.60	7.82	3.93	36.48	21.09	15.76
		CLIP_{CNN}	25.12	8.42	5.39	38.89	22.38	17.49
TCL	`Text`@Multi	TCL	**12.86***	**2.81***	**1.0***	**30.33***	**15.32***	**10.89***
		ALBEF	13.24	2.61	1.2	27.13	15.16	11.28
		CLIP_{ViT}	25.28	9.87	5.79	37.11	22.85	17.25
		CLIP_{CNN}	26.18	11.21	5.25	37.84	24.65	18.71
	`Image`@Multi	TCL	**72.5***	**55.98***	**46.49***	**79.26***	**64.65***	**56.99***
		ALBEF	5.94	1.6	0.8	12.16	3.79	2.26
		CLIP_{ViT}	7.85	1.97	0.61	13.43	5.37	3.31
		CLIP_{CNN}	9.71	2.85	1.65	15.44	5.7	3.37
	`Bi`@Multi	TCL	**78.08***	**65.53***	**56.81***	**87.43***	**75.23***	**68.87***
		ALBEF	22.94	6.61	3.6	40.13	22.72	17.51
		CLIP_{ViT}	27.98	9.66	5.08	41.46	25.11	18.99
		CLIP_{CNN}	30.78	12.47	7.52	44.19	26.93	20.63
CLIP_{ViT}	`Text`@Uni	CLIP_{ViT}	**28.34***	**11.73***	**6.81***	**38.89***	**24.08***	**17.42***
		CLIP_{CNN}	29.89	11.52	5.87	37.36	24.97	18.62
		ALBEF	7.61	1.00	0.30	19.97	9.58	6.59
		TCL	8.43	0.90	0.30	20.90	9.96	7.03
	`Image`@Uni	CLIP_{ViT}	**87.73***	**78.09***	**72.05***	**91.72***	**83.32***	**78.67***
		CLIP_{CNN}	7.66	1.90	1.44	9.37	3.90	2.53
		ALBEF	2.50	0.60	0.20	5.80	1.78	1.11
		TCL	5.27	0.40	0.20	9.12	2.75	1.48
	`Bi`@Uni	CLIP_{ViT}	**93.25***	**84.88***	**78.96***	**95.86***	**90.83***	**87.36***
		CLIP_{CNN}	32.52	13.78	7.52	41.82	26.77	21.10
		ALBEF	10.57	1.87	0.63	24.33	11.74	8.41
		TCL	11.94	2.38	1.07	26.69	13.80	9.46
CLIP_{CNN}	`Text`@Uni	CLIP_{CNN}	**30.40***	**13.11***	**7.21***	**40.03***	**26.79***	**20.74***
		CLIP_{ViT}	26.99	11.11	6.81	37.37	23.48	17.64
		ALBEF	7.72	0.90	0.50	20.79	9.84	6.98
		TCL	9.69	1.31	0.30	21.67	10.73	7.49
	`Image`@Uni	CLIP_{CNN}	**88.12***	**79.70***	**74.87***	**93.69***	**87.66***	**83.03***
		CLIP_{ViT}	1.84	0.10	0.30	5.51	2.50	1.02
		ALBEF	1.98	0.30	0.20	5.12	1.42	0.91
		TCL	4.74	0.50	0.10	7.95	2.32	1.42
	`Bi`@Uni	CLIP_{CNN}	**94.76***	**87.03***	**82.08***	**96.89***	**92.87***	**89.25***
		CLIP_{ViT}	28.79	11.63	6.40	40.03	24.60	18.83
		ALBEF	8.79	1.53	0.60	23.74	11.75	8.42
		TCL	13.10	2.31	0.93	26.07	13.53	9.23

Table F: **Attack success rates** (%) with different adversarial input modalities under Co-Attack on image-text retrieval. The adversaries are crafted using Flickr30K. * indicates white-box attacks. A higher ASR indicates better adversarial transferability.

transferability. As presented in Table K, there exists a positive correlation between transferability and the scale range, with the highest transferability observed at a scale range of $[0.50, 1.50]$ with a step size of 0.25. The experimental results show that the augmented image set plays a crucial role in increasing the transferability of the generated adversarial data.

Multi-pair Caption Set. The proposed SGA involves augmenting the original caption into a caption set for the purpose of generating adversarial data. To determine the effectiveness of the augmented caption set, various numbers of captions are utilized to construct the caption sets, and the transferability of the resulting adversarial data is evaluated. As illustrated in Table L, the use of multiple captions in the process of crafting adversarial data is observed to have a significant positive impact on adversarial transferability. Experimental results demonstrate that the augmented caption set also helps enhance the transferability of the generated adversarial data.

B.5. Visualization

Figure B depicts randomly selected original clean images and the corresponding adversarial examples, and such small perturbations are hard to be perceived. We magnified the imperceptible perturbation by a factor of 50 for visualization.

C. Algorithm

Algorithm 1 Set-level Guidance Attack

Input: Image encoder f_I, Text encoder f_T, Dataset D, Image-caption pair (v, t), Image scale sets $S = \{s_1, s_2, ..., s_N\}$, iteration steps K, number of paired captions M

Output: adversarial image v', adversarial caption t'

Build caption set $\boldsymbol{t} = \{t_1, t_2, ..., t_M\} \leftarrow D$

$/*$ Build adversarial caption set $\boldsymbol{t'} = \{t_1^{'}, t_2^{'}, ..., t_M^{'}\}$ $*/$

for $iter\ i = 1, 2, ..., M$ **do**

$$t_i' = \underset{t_i' \in B[t_i, \epsilon_t]}{\arg\max} - \frac{f_T(t_i') \cdot f_I(v)}{\|f_T(t_i')\| \|f_I(v)\|}$$

end for

$/*$ Build image set $\boldsymbol{v} = \{v_1, v_2, ..., v_N\}$ $*/$

for $iter\ i = 1, 2, ..., N$ **do**

$$v_i = resize(v, s_i) + 0.05 \cdot \boldsymbol{N}(0, 1)$$

end for

$/*$ Generate adversarial image v' $*/$

for $iter\ k = 1, 2, ..., K$ **do**

$$v' = \underset{v' \in B[v, \epsilon_v]}{\arg\max} - \sum_{i=1}^{M} \frac{f_T(t_i')}{\|f_T(t_i')\|} \sum_{v_i \in \boldsymbol{v}} \frac{f_I(v_i)}{\|f_I(v_i)\|}$$

end for

$/*$ Generate adversarial caption t' $*/$

$$t' = \underset{t' \in B[t, \epsilon_t]}{\arg\max} - \frac{f_T(t') \cdot f_I(v')}{\|f_T(t')\| \|f_I(v')\|}$$

The detailed training process of our proposed SGA is described in Algorithm 1.

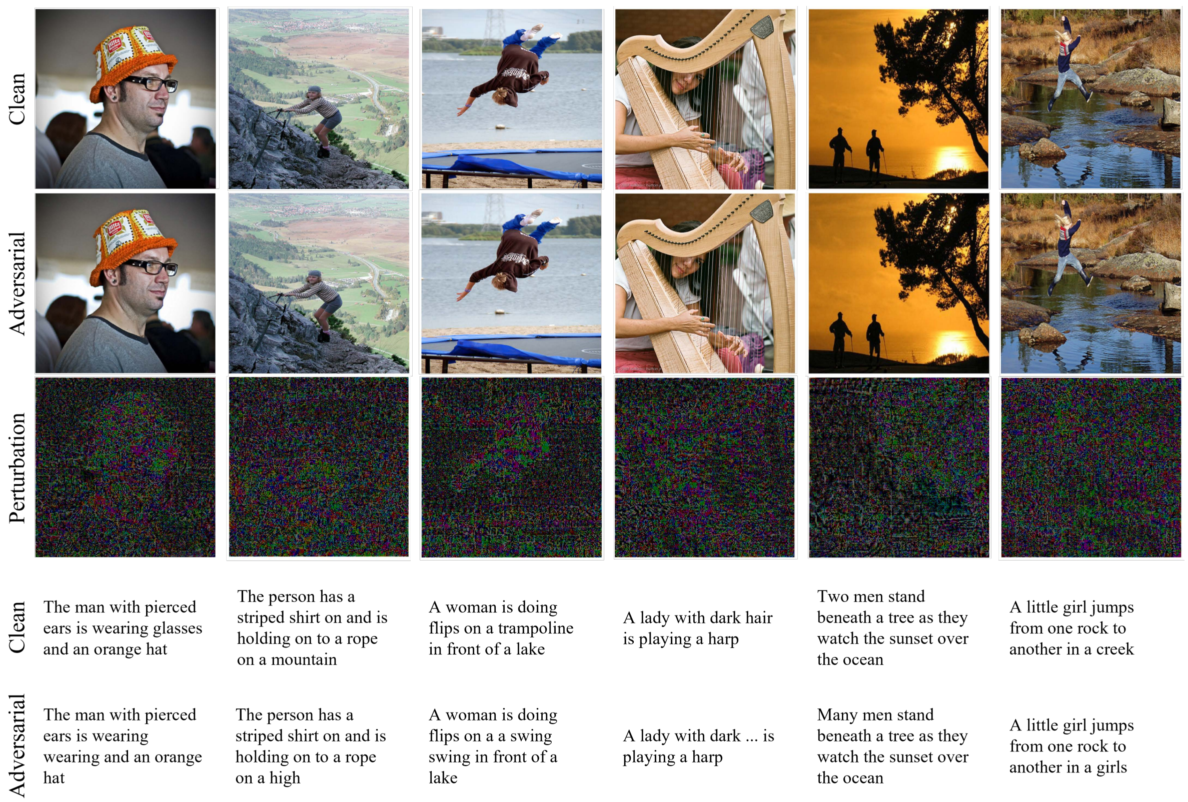

Figure B: **Visualization** of original images (**Upper**) and the corresponding adversarial examples (**Middle**) generated by our proposed SGA. Perturbations (**Lower**) are amplified by a factor of 50 for better illustration.

Source	Attack	ALBEF R@1	ALBEF R@5	ALBEF R@10	TCL R@1	TCL R@5	TCL R@10	CLIP$_{\text{ViT}}$ R@1	CLIP$_{\text{ViT}}$ R@5	CLIP$_{\text{ViT}}$ R@10	CLIP$_{\text{CNN}}$ R@1	CLIP$_{\text{CNN}}$ R@5	CLIP$_{\text{CNN}}$ R@10
Flickr30K (Image-Text Retrieval)													
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±0.22***	**94.09±0.42***	**92.30±0.28***	**45.42±0.60**	**24.93±0.15**	**16.48±0.49**	**33.38±0.35**	**13.50±0.30**	**9.04±0.15**	**34.93±0.99**	**17.07±0.23**	**10.45±0.95**
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±0.74**	**30.86±0.28**	**23.10±0.42**	**98.37±0.08***	**96.53±0.07***	**94.99±0.28***	**33.87±0.18**	**15.21±0.07**	**9.46±0.43**	**37.74±0.27**	**17.86±0.30**	**11.74±0.00**
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±0.07**	**2.46±0.08**	**1.35±0.07**	**16.23±0.45**	**3.77±0.21**	**1.10±0.14**	**99.08±0.08***	**97.25±0.07***	**95.22±0.15***	**38.76±0.27**	**19.45±0.00**	**11.95±0.44**
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±0.07**	**2.56±0.07**	**1.05±0.21**	**14.91±0.08**	**3.62±0.14**	**1.70±0.14**	**31.24±0.42**	**13.45±0.07**	**8.74±0.14**	**99.24±0.18***	**98.20±0.30***	**95.16±0.44***
Flickr30K (Text-Image Retrieval)													
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±0.15***	**94.27±0.04***	**92.58±0.03***	**55.25±0.06**	**36.01±0.03**	**27.25±0.13**	44.16±0.25	**27.35±0.30**	**20.84±0.04**	**46.57±0.13**	**29.16±0.17**	**22.68±0.00**
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±0.10**	**42.47±0.22**	**34.59±0.29**	**98.81±0.07***	**97.19±0.03***	**95.86±0.11***	**44.88±0.54**	**28.79±0.28**	**21.95±0.11**	**48.30±0.34**	**29.70±0.02**	**23.68±0.06**
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±0.06**	**13.21±0.00**	**9.76±0.11**	**30.76±0.07**	**16.36±0.26**	**12.08±0.06**	**98.94±0.00***	**97.53±0.16***	**96.03±0.08***	**47.79±0.58**	**30.36±0.36**	**24.50±0.37**
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±0.28**	**12.32±0.15**	**8.98±0.06**	**28.82±0.11**	**15.12±0.11**	**10.56±0.17**	**42.12±0.11**	**26.80±0.05**	**20.23±0.13**	**99.49±0.05***	**98.41±0.06***	**97.14±0.11***

Table G: **Attack success rate** (%) of four VLP models under existing adversarial attacks and SGA. The source column indicates the source models used to generate the adversarial data on Flickr30K. * indicates white-box attacks. A higher ASR indicates better adversarial transferability.

Source	Attack	ALBEF			TCL			CLIP$_{\mathrm{ViT}}$			CLIP$_{\mathrm{CNN}}$		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
MSCOCO (Image-Text Retrieval)													
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±0.11***	**92.83±0.13***	**90.37±0.03***	**58.56±0.06**	**39.00±0.40**	**30.68±0.22**	**57.06±0.51**	**39.38±0.22**	**31.55±0.06**	**58.95±0.19**	**42.49±0.13**	**34.84±0.28**
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±0.06**	**49.33±0.35**	**40.34±0.01**	**98.97±0.04***	**97.89±0.12***	**96.63±0.03***	**56.34±0.08**	**39.58±0.21**	**32.00±0.12**	**59.44±0.20**	**42.17±0.21**	**34.94±0.05**
CLIP$_{\mathrm{ViT}}$	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±0.22**	**16.73±0.04**	**10.98±0.25**	**37.54±0.30**	**19.09±0.04**	**12.92±0.31**	**99.79±0.03***	**99.37±0.07***	**98.89±0.04***	**58.93±0.11**	**44.60±0.11**	**37.53±0.74**
CLIP$_{\mathrm{CNN}}$	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±0.40**	**14.27±0.28**	**9.36±0.01**	**34.81±0.15**	**17.16±0.03**	**11.26±0.04**	**56.62±0.06**	**41.31±0.15**	**32.88±0.10**	**99.61±0.08***	**99.02±0.11***	**98.42±0.17***
MSCOCO (Text-Image Retrieval)													
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±0.08***	**93.44±0.04***	**91.00±0.06***	**65.38±0.08**	**47.61±0.07**	**38.96±0.07**	**65.25±0.09**	**50.42±0.08**	**43.47±0.12**	**66.52±0.18**	**52.44±0.28**	**45.05±0.07**
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±0.04**	**58.40±0.09**	**50.96±0.17**	**99.15±0.03***	**98.17±0.02***	**97.34±0.01***	**63.99±0.16**	**49.87±0.09**	**42.46±0.10**	**65.70±0.19**	**51.45±0.06**	**44.64±0.06**
CLIP$_{\mathrm{ViT}}$	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±0.00**	**28.66±0.13**	**22.64±0.09**	**47.76±0.25**	**32.30±0.04**	**25.70±0.04**	**99.79±0.00***	**99.37±0.01***	**98.94±0.07***	**65.83±0.35**	**53.58±0.25**	**46.84±0.16**
CLIP$_{\mathrm{CNN}}$	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±0.01**	**27.64±0.04**	**21.74±0.00**	**45.95±0.23**	**30.57±0.00**	**24.27±0.22**	**60.77±0.02**	**46.99±0.11**	**40.49±0.16**	**99.80±0.03***	**99.29±0.06***	**98.77±0.06***

Table H: **Attack success rate** (%) of four VLP models under existing adversarial attacks and SGA. The source column indicates the source models used to generate the adversarial data on MSCOCO. * indicates white-box attacks. A higher ASR indicates better adversarial transferability.

Source	Attack	Target	Method	Image-to-Text R@1	Image-to-Text R@5	Image-to-Text R@10	Text-to-Image R@1	Text-to-Image R@5	Text-to-Image R@10
ALBEF	**Text**@Multi	ALBEF	Co-Attack	9.18*	1.50*	1.00*	21.70*	11.96*	9.22*
			SGA	**13.03***	**2.71***	**1.40***	**26.17***	**14.17***	**10.76***
		TCL	Co-Attack	9.38	1.31	0.30	20.40	9.74	6.80
			SGA	**12.64**	**2.01**	**0.80**	**26.43**	**13.69**	**9.32**
		CLIP$_{\text{ViT}}$	Co-Attack	20.98	7.79	4.57	31.73	19.13	14.63
			SGA	**27.24**	**11.32**	**7.52**	**36.82**	**22.10**	**16.99**
		CLIP$_{\text{CNN}}$	Co-Attack	22.48	7.40	4.12	31.94	21.36	15.59
			SGA	**27.97**	**13.85**	**7.62**	**37.77**	**24.82**	**18.46**
	Image@Multi	ALBEF	Co-Attack	75.50*	59.22*	53.30*	83.63*	75.14*	70.32*
			SGA	**90.82***	**83.27***	**79.00***	**90.08***	**83.35***	**79.32***
		TCL	Co-Attack	4.64	1.21	0.50	11.33	3.72	2.25
			SGA	**21.18**	**9.15**	**5.91**	**28.00**	**13.50**	**9.16**
		CLIP$_{\text{ViT}}$	Co-Attack	7.24	1.97	0.51	13.53	5.23	3.01
			SGA	**10.92**	**3.53**	**1.52**	**16.72**	**6.70**	**4.34**
		CLIP$_{\text{CNN}}$	Co-Attack	10.09	2.85	1.65	15.27	6.11	3.52
			SGA	**12.52**	**3.91**	**2.47**	**17.77**	**7.44**	**4.65**
TCL	**Text**@Multi	ALBEF	Co-Attack	13.24	2.61	1.20	27.13	15.16	11.28
			SGA	**10.84**	**2.71**	**0.90**	**24.77**	**12.22**	**9.30**
		TCL	Co-Attack	12.86	2.81	1.00	30.33	15.32	10.89
			SGA	**13.38***	**3.72***	**1.00***	**27.17***	**14.06***	**10.07***
		CLIP$_{\text{ViT}}$	Co-Attack	25.28	9.87	5.79	37.11	22.85	17.25
			SGA	**27.98**	**12.05**	**7.32**	**37.69**	**22.73**	**17.31**
		CLIP$_{\text{CNN}}$	Co-Attack	26.18	11.21	5.25	37.84	24.65	18.71
			SGA	**30.40**	**13.85**	**8.14**	**37.77**	**25.14**	**19.41**
	Image@Multi	ALBEF	Co-Attack	5.94	1.60	0.80	12.16	3.79	2.26
			SGA	**27.11**	**13.93**	**9.80**	**34.49**	**19.24**	**13.67**
		TCL	Co-Attack	72.50	55.98	46.49	79.26	64.65	56.99
			SGA	**96.00***	**92.16***	**89.28***	**96.86***	**92.70***	**90.19***
		CLIP$_{\text{ViT}}$	Co-Attack	7.85	1.97	0.61	13.43	5.37	3.31
			SGA	**10.92**	**3.53**	**1.52**	**16.88**	**7.15**	**4.62**
		CLIP$_{\text{CNN}}$	Co-Attack	9.71	2.85	1.65	15.44	5.70	3.37
			SGA	**13.15**	**4.97**	**2.37**	**18.56**	**7.56**	**5.13**
CLIP$_{\text{ViT}}$	**Text**@Multi	ALBEF	Co-Attack	7.61	1.00	0.30	19.97	9.58	6.59
			SGA	**8.13**	**1.20**	**0.40**	**19.50**	**8.76**	**6.59**
		TCL	Co-Attack	8.43	0.90	0.30	20.90	9.96	7.03
			SGA	**8.96**	**1.01**	**0.30**	**21.64**	**10.59**	**7.88**
		CLIP$_{\text{ViT}}$	Co-Attack	28.34	11.73	6.81	38.89	24.08	17.42
			SGA	**31.78***	**15.16***	**8.43***	**39.43***	**25.58***	**19.30***
		CLIP$_{\text{CNN}}$	Co-Attack	29.89	11.52	5.87	37.36	24.97	18.62
			SGA	**29.89**	**12.37**	**6.90**	**36.40**	**23.13**	**18.48**
	Image@Multi	ALBEF	Co-Attack	2.50	0.60	0.20	5.80	1.78	1.11
			SGA	**3.86**	**0.70**	**0.30**	**7.69**	**2.73**	**1.52**
		TCL	Co-Attack	5.27	0.40	0.20	9.12	2.75	1.48
			SGA	**6.43**	**0.60**	**0.20**	**10.93**	**3.47**	**2.05**
		CLIP$_{\text{ViT}}$	Co-Attack	87.73	78.09	72.05	91.72	83.32	78.67
			SGA	**94.11***	**88.89***	**83.64***	**95.91***	**90.10***	**85.98***
		CLIP$_{\text{CNN}}$	Co-Attack	7.66	1.90	1.44	9.37	3.90	2.53
			SGA	**11.24**	**5.39**	**2.68**	**15.68**	**6.88**	**5.08**
CLIP$_{\text{CNN}}$	**Text**@Multi	ALBEF	Co-Attack	7.72	0.90	0.50	20.79	9.84	6.98
			SGA	**7.82**	**1.30**	**0.60**	**19.93**	**9.74**	**7.16**
		TCL	Co-Attack	9.69	1.31	0.30	21.67	10.73	7.49
			SGA	**9.59**	**1.91**	**0.60**	**21.88**	**10.96**	**7.74**
		CLIP$_{\text{ViT}}$	Co-Attack	26.99	11.11	6.81	37.37	23.48	17.64
			SGA	**26.50**	**11.63**	**6.40**	**37.66**	**22.89**	**17.01**
		CLIP$_{\text{CNN}}$	Co-Attack	30.40	13.11	7.21	40.03	26.79	20.74
			SGA	**36.27***	**17.34***	**11.02***	**44.29***	**29.16***	**22.82***
	Image@Multi	ALBEF	Co-Attack	1.98	0.30	0.20	5.12	1.42	0.91
			SGA	**2.09**	**0.60**	**0.20**	**6.20**	**1.70**	**1.19**
		TCL	Co-Attack	4.74	0.50	0.10	7.95	2.32	1.42
			SGA	**4.85**	**0.70**	**0.30**	**9.19**	**2.63**	**1.73**
		CLIP$_{\text{ViT}}$	Co-Attack	1.84	0.10	0.30	5.51	2.50	1.02
			SGA	**3.19**	**1.77**	**0.81**	**9.34**	**4.56**	**2.35**
		CLIP$_{\text{CNN}}$	Co-Attack	88.12	79.70	74.87	93.69	87.66	83.03
			SGA	**92.46***	**86.68***	**81.98***	**96.64***	**91.78***	**87.87***

Table I: **Attack success rates** (%) on four VLP models under Co-Attack and SGA with different single adversarial input modalities. The adversaries are crafted on Flickr30K. * indicates white-box attacks.

Iterative Strategy	Image-to-Text			Text-to-Image		
	R@1	R@5	R@10	R@1	R@5	R@10
t-i-t	45.42	24.93	16.48	55.25	36.01	27.25
i-t-i	45.84	26.43	18.24	56.45	36.39	27.60
t-i-t-i-t	48.37	26.63	19.44	57.19	38.08	28.72

Table J: **Ablation experiment on different iterative strategies.** The dataset is Flickr30K. The source model is ALBEF and the target model is TCL. Attack success rates (%) are utilized to measure the adversarial transferability.

Scales	Image-to-Text			Text-to-Image		
	R@1	R@5	R@10	R@1	R@5	R@10
$[1.00]$	34.04	13.17	8.62	44.12	25.95	19.25
$[0.75, 1.00, 1.25]$	44.57	22.70	14.63	54.55	34.36	26.22
$[0.50, 0.75, 1.00, 1.25, 1.50]$	45.94	24.82	16.13	55.21	35.99	27.15
$[0.25, 0.50, 0.75, 1.00, 1.25, 1.50, 1.75]$	44.15	24.22	16.13	55.10	35.35	26.81

Table K: **Ablation experiment on the image set.** The dataset is Flickr30K. The source model is ALBEF and the target model is TCL. Attack success rates (%) are utilized to measure the adversarial transferability.

Number of Captions	Image-to-Text			Text-to-Image		
	R@1	R@5	R@10	R@1	R@5	R@10
1	40.04	18.99	12.53	51.14	30.93	23.17
2	45.52	22.51	15.13	54.69	33.45	25.28
3	45.84	23.82	15.43	54.67	34.69	26.58
4	46.05	25.03	16.23	55.16	35.66	27.13
5	45.94	24.82	16.13	55.21	35.99	27.15

Table L: **Ablation experiment on the caption set.** The dataset is Flickr30K. The source model is ALBEF and the target model is TCL. Attack success rates (%) are utilized to measure the adversarial transferability.