Transform to Transfer: Boosting Adversarial Attack Transferability on Vision-Language Pre-training Models

Supplementary Material

A. Overview

Our main paper outlines the core concepts and techniques of the proposed method, and experimentally demonstrates its contributions as well as the effectiveness of the adopted configurations. In this supplementary material, we provide additional details that could not be included in the main paper due to space constraints. These include: more comprehensive experimental results regarding cross-model transferability, incorporating R@5 and R@10 metrics in Section B and visualizations illustrating the perturbations introduced by our multimodal adversarial attack on both image and text modalities in Section C.

B. Detailed Results

To comprehensively demonstrate the performance of our method, we follow the same scheme as in the main text by employing the image-text retrieval task to evaluate the cross-model transferability of adversarial examples. We adopt the key metrics—R@1, R@5 and R@10—to assess the effectiveness of our attacks. These metrics evaluate whether adversarial examples fail to appear in top-ranked positions during retrieval, serving as critical indicators of attack success. Due to space constraints in the main text, we focus primarily on the R@1 metric for cross-model transferability experiments. In Table B.1 and Table B.2, we provide the complete results for R@1, R@5, and R@10 on Flickr30K and MSCOCO, respectively. Table B.1 presents results on a Flickr30K test set comprising 1,000 images, each with approximately five captions. We also include evaluations on a larger MSCOCO test set, which contains 5,000 images, each accompanied by roughly five captions, as shown in Table B.2.

As demonstrated in Table B.1, our method achieves state-of-the-art performance in white-box scenarios on both ALBEF and TCL, achieving perfect scores of 100% across all metrics (R@1, R@5, R@10) for both text-to-image retrieval (IR) and image-to-text retrieval (TR). On CLIP_{ViT} and CLIP_{CNN}, while our results slightly trail the current best performance, they remain highly competitive. More importantly, in black-box scenarios, our method demonstrates substantial improvements.

When transferring to architecturally similar models, for fused models, while the improvement in R@1 appears modest due to limited headroom for enhancement, our method yields substantial gains on R@5 and R@10 metrics. Specifically, when transferring adversarial examples from ALBEF to TCL, our method elevates the R@5 metric from 92.96% to 99.40% and the R@10 metric from 89.78% to 99.20% in TR, significantly outperforming LSSA. For aligned models, our method achieves remarkable improvements across R@1, R@5 and R@10. When transferring adversarial examples from CLIP_{ViT} to CLIP_{CNN} for TR, our approach elevates R@1 from 78.54% to 95.15%, R@5 from 60.15% to 91.54%, and R@10 from 50.05% to 88.05%, compared with LSSA.

When adversarial examples are transferred to models with different architectures, the improvements on R@5 and R@10 become even more pronounced. Specifically, when transferring adversarial examples from ALBEF to CLIP_{ViT} in TR, our method elevates R@5 from 32.2% to 85.0%—an improvement of approximately 2.6-fold—compared to LSSA. Similarly, the R@10 metric increases from 25.8% to 78.66%, representing an approximate 3-fold enhancement. Similarly, in the reverse transfer scenario from CLIP_{ViT} to ALBEF, our method advances R@5 from 26.5% to 70.3% and R@10 from 20.9% to 64.00%, corresponding to improvements of approximately 2.6-fold and 3-fold, respectively.

Our method demonstrates consistent performance on MSCOCO. In white-box scenarios, it achieves perfect scores of 100% across both TR and IR on ALBEF and TCL. For the CLIP_{ViT} model, performance remains highly competitive, with only marginal gaps observed on TR R@5, TR R@10, and IR R@10 compared to the best reported results, while performance on CLIP_{CNN} is also slightly below the state-of-the-art. More notably, in black-box settings, our method significantly enhances the transferability of adversarial examples compared to existing approaches, establishing new state-of-the-art performance.

C. Visualization of Adversarial Perturbations

In Figure C.1, we present the adversarial perturbations of our method for both image and text modalities. As mentioned in our experimental setup, we perturb only one word for each sentence and use an 8/255 perturbation for each image. The first two rows represent the clean image and the adversarial image, respectively. In the third row, we visualize the adversarial perturbations of the adversarial images. Since the magnitude of the adversarial perturbation is very small and difficult to detect, we have amplified the adversarial perturbation by 20 times. The last two rows represent the clean text and the adversarial text, where green indicates the original word, while red indicates the word modified in

Image-To-Text Retrieval (TR)

Source	Method	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	PGD	93.74	85.97	82.50	24.30	10.15	7.62	10.67	3.12	1.20	14.05	3.70	2.47
	Co-Attack	97.08	94.59	92.60	39.52	20.40	14.53	29.82	11.73	6.61	31.29	11.73	5.77
	SGA	99.90	99.90	99.90	87.88	77.99	71.84	36.69	19.31	14.13	39.59	21.99	15.96
	DRA	99.90	99.80	99.80	91.57	81.91	76.35	46.26	24.30	17.17	49.55	28.75	20.60
	LSSA	99.90	99.80	99.80	96.21	92.96	89.78	53.25	32.29	25.81	56.45	35.41	28.22
	Ours	**100.0**	**100.0**	**100.0**	**99.89**	**99.40**	**99.20**	**92.27**	**85.05**	**78.66**	**93.36**	**85.41**	**80.23**
TCL	PGD	35.77	21.74	16.20	99.37	98.59	97.70	10.18	2.91	1.42	14.81	4.55	2.37
	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	92.49	87.37	83.50	**100.0**	**100.0**	**100.0**	36.81	19.00	12.20	41.89	23.68	16.89
	DRA	95.20	91.38	89.60	**100.0**	**100.0**	99.90	47.24	26.48	18.90	52.23	30.44	22.66
	LSSA	98.75	96.59	95.50	**100.0**	**100.0**	**100.0**	52.88	30.53	23.37	56.83	40.41	37.84
	Ours	**100.0**	**99.90**	**99.80**	**100.0**	**100.0**	**100.0**	**87.36**	**77.05**	**70.93**	**90.17**	**82.45**	**77.14**
$CLIP_{ViT}$	PGD	3.13	0.40	0.30	4.43	1.01	0.20	69.33	45.59	36.99	13.03	3.81	1.75
	Co-Attack	8.55	1.50	0.50	10.01	2.01	0.70	78.53	57.42	45.53	29.50	11.42	6.08
	SGA	22.42	8.92	5.60	25.08	10.15	4.91	**100.0**	**100.0**	**100.0**	53.26	33.93	24.20
	DRA	27.81	12.22	7.70	27.82	11.76	7.11	**100.0**	99.90	99.90	64.88	42.60	31.72
	LSSA	45.99	26.55	20.90	45.10	26.03	19.84	**100.0**	99.79	99.80	78.54	60.15	50.05
	Ours	**81.02**	**70.34**	**64.00**	**80.51**	**67.34**	**60.72**	99.88	99.79	99.59	**95.15**	**91.54**	**88.05**
$CLIP_{CNN}$	PGD	2.29	0.30	0.30	4.53	0.30	0.10	5.40	1.45	0.81	89.78	77.70	70.75
	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	15.64	5.51	3.00	18.02	6.43	2.91	39.02	19.21	13.01	99.87	99.58	99.38
	DRA	19.50	6.31	3.20	21.60	7.64	3.71	48.47	26.38	17.07	99.87	99.47	99.07
	LSSA	31.39	15.83	10.80	35.41	16.88	11.12	65.28	42.16	31.91	**100.0**	**100.0**	**99.79**
	Ours	**55.16**	**33.57**	**26.70**	**57.74**	**37.29**	**29.46**	**80.74**	**64.69**	**55.79**	99.87	99.79	99.49

Text-To-Image Retrieval (IR)

Source	Method	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	PGD	94.43	89.48	86.19	27.90	13.56	9.87	15.82	6.61	3.97	19.11	7.11	4.72
	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	99.98	99.94	99.94	88.05	77.53	70.56	46.78	26.66	21.96	49.87	32.27	24.96
	DRA	99.93	99.94	99.94	91.17	81.99	76.79	56.80	38.31	29.68	59.01	41.19	33.34
	LSSA	100.0	99.96	99.92	96.26	92.29	89.28	60.89	42.82	34.80	64.43	46.87	38.06
	Ours	**100.0**	**100.0**	**100.0**	**99.67**	**99.05**	**98.50**	**92.82**	**85.75**	**80.44**	**93.58**	**86.90**	**81.50**
TCL	PGD	41.67	25.80	19.35	99.33	98.24	97.08	16.30	6.61	4.25	21.10	8.56	5.26
	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	25.10
	DRA	95.58	91.78	88.90	99.98	99.98	99.96	57.28	39.31	31.88	62.23	44.06	35.78
	LSSA	98.57	96.74	95.79	**100.0**	**100.0**	**100.0**	60.70	43.19	35.15	67.38	48.33	40.41
	Ours	**99.70**	**99.51**	**99.35**	**100.0**	**100.0**	**100.0**	**89.66**	**81.01**	**75.80**	**91.63**	**84.40**	**79.49**
$CLIP_{ViT}$	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
	Co-Attack	20.18	9.54	7.12	21.29	9.51	6.79	87.50	77.95	73.40	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	61.10	43.18	35.37
	DRA	42.84	24.38	18.76	44.60	26.38	19.74	**100.0**	**100.0**	**100.0**	69.50	53.66	45.09
	LSSA	56.48	36.63	29.26	56.74	36.82	29.14	**100.0**	**100.0**	**100.0**	80.82	66.76	58.21
	Ours	**86.11**	**74.57**	**68.42**	**84.83**	**72.73**	**65.52**	**100.0**	99.95	99.91	**96.43**	**92.46**	**88.86**
$CLIP_{CNN}$	PGD	6.15	1.70	0.97	8.88	2.42	1.42	12.08	4.86	2.94	93.04	85.37	81.09
	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	15.26	10.82	33.07	17.22	12.33	51.45	33.29	25.77	99.90	99.81	99.64
	DRA	34.59	18.05	13.83	37.88	20.94	15.41	59.12	40.88	32.90	99.90	99.61	99.46
	LSSA	44.06	26.11	19.83	47.12	28.90	22.20	70.30	53.42	45.07	**100.0**	**100.0**	**99.93**
	Ours	**66.42**	**48.93**	**40.86**	**69.17**	**51.83**	**43.16**	**84.38**	**72.39**	**64.61**	99.97	99.93	99.89

Table B.1. The attack success rate (%) of multimodal adversarial examples against different VLP models compared with state-of-the-art methods on image-text retrieval task. The source column represents the VLP models used for crafting multimodal adversarial examples on the Flickr30K dataset. The constraint for adversarial perturbation is set to 8/255, with a per-step size of 2/255, and a total of 10 steps. The shaded area represents the white-box attacks.

Image-To-Text Retrieval (TR)

Source	Method	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	PGD	94.35	87.80	84.50	34.15	18.63	13.05	21.71	10.94	7.34	23.83	11.75	8.78
	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	99.72	99.61	87.46	76.28	69.24	63.72	47.62	39.26	63.91	47.35	40.65
	DRA	99.90	99.77	99.51	88.81	78.63	71.93	69.25	52.01	43.90	68.53	52.03	44.14
	LSSA	**100.0**	**99.96**	99.92	96.01	90.97	87.40	72.99	57.35	49.92	74.21	58.17	49.36
	Ours	**100.0**	**99.96**	**99.94**	**99.44**	**98.69**	**97.68**	**95.35**	**91.18**	**88.37**	**94.56**	**91.54**	**88.40**
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
	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	92.70	87.12	84.01	**100.0**	99.98	99.98	59.79	43.80	36.70	60.52	46.15	38.19
	DRA	94.72	90.60	88.02	**100.0**	**100.0**	**100.0**	70.51	54.58	45.91	70.29	54.58	45.94
	LSSA	98.48	96.77	95.43	**100.0**	**100.0**	**100.0**	71.92	55.99	47.29	71.84	57.03	49.38
	Ours	**99.85**	**99.72**	**99.49**	**100.0**	**100.0**	**100.0**	**95.65**	**92.80**	**90.71**	**96.16**	**93.45**	**91.63**
$CLIP_{ViT}$	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
	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	43.75	25.67	17.93	44.05	25.31	17.91	**100.0**	**100.0**	**100.0**	70.66	56.43	48.92
	DRA	52.69	32.04	23.71	51.88	31.05	23.67	**100.0**	**100.0**	**100.0**	80.18	67.75	60.13
	LSSA	64.73	46.66	38.58	63.49	45.13	35.54	**100.0**	**100.0**	**100.0**	87.25	78.60	72.14
	Ours	**91.83**	**84.21**	**79.15**	**90.74**	**81.80**	**75.82**	**100.0**	**100.0**	99.98	**98.00**	**96.30**	**95.12**
$CLIP_{CNN}$	PGD	8.38	3.59	1.85	11.90	5.11	2.73	13.66	7.43	4.74	92.68	85.89	81.92
	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
	DRA	41.40	21.68	14.64	43.62	23.70	16.18	70.43	55.15	46.41	99.80	99.54	99.39
	LSSA	52.15	32.44	24.85	54.26	33.40	26.13	79.74	67.84	60.08	**100.0**	**100.0**	**100.0**
	Ours	**75.61**	**59.88**	**51.09**	**76.32**	**58.39**	**49.77**	**92.22**	**85.79**	**81.28**	99.96	99.95	99.95

Text-To-Image Retrieval (IR)

Source	Method	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	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
	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	99.94	99.81	99.74	88.17	77.33	70.64	69.71	55.21	47.73	70.78	56.34	48.74
	DRA	99.93	99.80	99.69	90.06	79.85	73.57	75.31	60.88	53.39	75.09	61.84	54.50
	LSSA	99.97	**99.92**	99.90	95.83	90.92	87.39	78.46	65.16	57.77	78.20	66.02	58.64
	Ours	**99.99**	**99.92**	**99.92**	**99.35**	**98.19**	**97.38**	**95.80**	**92.35**	**89.64**	**96.07**	**92.49**	**89.75**
TCL	PGD	44.09	27.75	21.25	98.20	95.46	93.65	26.92	14.91	11.64	32.17	18.42	13.89
	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	92.99	87.65	84.17	**100.00**	99.99	99.98	65.31	50.92	43.87	67.34	53.48	45.87
	DRA	95.89	91.87	89.29	**100.00**	**100.00**	99.99	74.95	61.05	54.25	76.99	63.73	56.49
	LSSA	98.22	96.43	95.17	**100.00**	99.99	99.99	76.01	62.89	56.09	78.30	66.10	58.96
	Ours	**99.91**	**99.68**	**99.52**	**100.00**	**100.00**	**100.00**	**96.42**	**92.80**	**90.71**	**97.05**	**94.63**	**92.55**
$CLIP_{ViT}$	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
	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	27.90	**100.00**	**100.00**	**100.00**	75.58	63.22	56.28
	DRA	61.50	43.50	36.08	61.06	43.89	36.28	**100.00**	**100.00**	99.99	84.11	74.54	67.85
	LSSA	69.24	52.94	44.94	67.69	51.17	43.26	**100.00**	**100.00**	99.99	89.74	81.52	76.14
	Ours	**92.66**	**86.10**	**81.85**	**91.68**	**83.93**	**79.40**	**100.00**	**100.00**	**100.00**	**98.29**	**97.16**	**96.11**
$CLIP_{CNN}$	PGD	12.73	5.43	3.61	15.68	7.38	5.05	20.62	11.76	8.62	94.71	89.30	86.32
	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	46.79	29.97	23.70	48.90	32.89	26.04	67.73	53.77	43.19	99.97	99.83	99.77
	DRA	52.25	35.85	28.75	54.15	38.03	30.45	74.14	62.21	55.07	99.92	99.73	99.66
	LSSA	60.68	42.59	35.18	61.40	44.19	36.63	83.02	72.41	66.52	**100.00**	**100.00**	**99.99**
	Ours	**79.63**	**67.05**	**60.36**	**80.14**	**67.21**	**60.89**	**93.07**	**88.21**	**84.28**	99.95	99.96	99.94

Table B.2. The attack success rate (%) of multimodal adversarial examples against different VLP models compared with state-of-the-art methods on image-text retrieval task. The source column represents the VLP models used for crafting multimodal adversarial examples on the MSCOCO dataset. The constraint for adversarial perturbation is set to 8/255, with a per-step size of 2/255, and a total of 10 steps. The shaded area represents the white-box attacks.

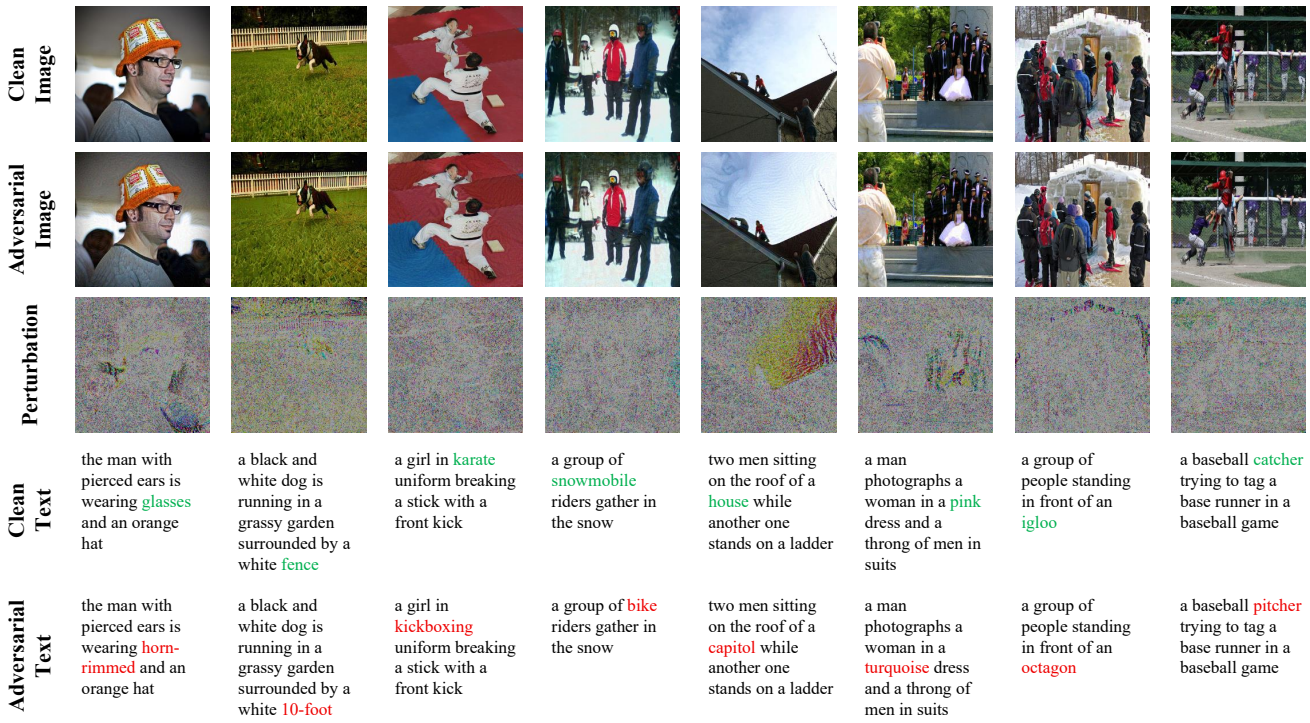

Figure C.1. Multimodal adversarial perturbation visualization on Flickr30K. The rows show the original images, adversarial images, adversarial perturbations, original texts, and corresponding adversarial texts. The perturbation magnitude ensures that the perturbations in the image are imperceptible, and only one word in each text has changed.

the adversarial text.