One Perturbation is Enough: On Generating Universal Adversarial Perturbations against Vision-Language Pre-training Models

Supplementary Material

A. Pseudocode of the Proposed Algorithm

We present the pseudocode of our proposed attack algorithm for image modality in Alg. 1. Note that the text attacks are completely symmetrical as illustrated in Sec. 3.

Algorithm 1 Pseudocode of universal image attacks

```
Require: $G_w(\cdot)$: the perturbation generator; $D_s$: the multimodal training set; $f_I, f_T$: image encoder and text encoder of the surrogate VLP model; $K$: the max iteration; $\epsilon_v$: the perturbation budget; $N$: the scaling times;
Ensure: Universal image perturbation $\delta_v$;
Initialize the fixed noise $z_v$ with Gaussian distribution;
for $i \leftarrow 0$ to $K$ do
    Randomly sample an image-text pair $(v, \mathbf{t}) \sim D_s$;
    $\delta_v = Clip_{\epsilon_v}(G_w(z_v; f_T(\mathbf{t})))$, $v_{adv} = v + \delta_v$;
    Augment $v$ and $v_{adv}$ into different scales and apply random Gaussian noises to obtain $\mathbf{v} = \{v_1 \ldots, v_N\}$ and $\mathbf{v}_{adv} = \{v_1^{adv} \ldots, v_N^{adv}\}$;
    Randomly sample a batch of text sets from $D_s$ and obtain $\mathbf{t}_{pos} = \{t'_1 \ldots, t'_K\}$ by selecting the one with the farthest feature distance from the clean image $v$;
    Compute $\mathcal{L}_{CL}$ with $\mathbf{v}_{adv}$, $\mathbf{t}$ and $\mathbf{t}_{pos}$ by Eq. (3);
    Compute $\mathcal{L}_{Dis}$ with $\mathbf{v}$ and $\mathbf{v}_{adv}$ by Eq. (4);
    Optimize the generator $G_w$ based on Eq. (5);
    Backward pass and update $G_w$;
end for
Return $\delta_v$
```

B. More Training Details

For Flickr30K and MSCOCO, we randomly sample 30,000 images and their captions from the training set to train our perturbation generator. For SNLI-VE and RefCOCO+, we learn the C-PGC directly using their training sets with 29,783 and 16,992 images, respectively. Since an image corresponds to multiple text descriptions in these datasets, we calculate the average of their textual embedding as the multimodal condition for the cross-attention modules.

We initialize the noise variable z_v as a 3×3 matrix. Meanwhile, the initial noise z_t's dimensions in the text modality depend on the size of the hidden layer within the specific VLP model. Concretely, we set its dimension to 1×3 for ALBEF, TCL, BLIP, and X-VLM, while 1×2 for the CLIP model. When computing the multimodal contrastive loss $\mathcal{L}_{CL}$, the temperature τ is set as 0.1. The generator is trained over 40 epochs with the Adam optimizer at a learning rate of 2^{-4}. Following previous works [5, 8], we employ the attack success rate (ASR) as our quantitative measurement in ITR tasks by computing the extent to which the adversarial perturbations result in victim models' performance deviations from the clean performance.

C. More Experimental Results

In this section, we provide more experimental results of our method in various tasks and scenarios.

Visual entailment tasks. Given an image and a textual description, visual entailment involves determining whether the textual description can be inferred from the semantic information of the image. We align with previous VLP attacks [8, 10] and conduct experiments on the SNLI-VE [9] dataset using the ALBEF and TCL models. Note that the Baseline represents the clean performance of the target model on the clean data and the orange and green indicate ALBEF and TCL as source models respectively. The results presented in Fig. 1 reveal that C-PGC obtains impressive attack effects by decreasing the average accuracy by nearly 20%.

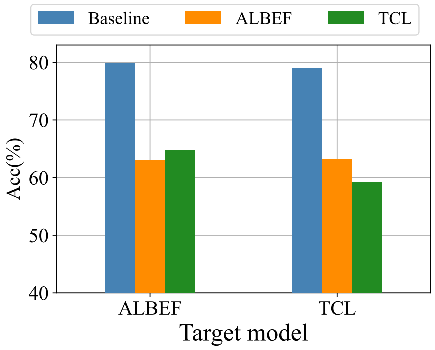

Fig. 1: Accuracy of VE tasks for different source and target models. The Baseline indicates the clean performance.

Notably, [2] has reported a large number of annotation errors in the labels of the SNLI-VE corpus used for VE tasks. Therefore, the presented results are only for experimental integrity and reference purposes.

Image-text retrievals on MSCOCO dataset. We then supplement the ASR of the ITR tasks on the MSCOCO dataset in Table 1. The results again reveal that C-PGC greatly enhances the attack. Particularly in the more realistic and challenging transferable scenarios, the proposed method achieves considerably better performance, *e.g.*, 82.49% and 76.24% increase in ASR of TR and IR tasks

Table 1. ASR (%) of different methods for image-text retrieval tasks on MSCOCO dataset. TR indicates text retrieval based on the input image, while IR is image retrieval using the input text.

Source	Method	ALBEF		TCL		X-VLM		CLIP$_{\text{ViT}}$		CLIP$_{\text{CNN}}$		BLIP	
		TR	IR	TR	IR	TR	IR	TR	IR	TR	IR	TR	IR
ALBEF	GAP	82.65	84.35	53.6	45.46	15.09	15.64	25.18	29.94	28.06	35.28	37.44	33.61
	ETU	83.6	88.98	27.43	24.47	20.39	19.94	28.54	35.05	37.01	44.72	22.25	22.03
	Ours	**96.18**	**95.09**	**82.49**	**76.24**	**39.97**	**48.58**	**59.71**	**67.05**	**61.27**	**70.8**	**59.18**	**63.89**
TCL	GAP	55.92	48.22	95.16	92.29	17.34	17.01	28.73	31.19	32.27	39.81	43.59	39.64
	ETU	60.09	50.62	93.28	89.28	27.19	25.41	33	37.86	45.83	52.24	39.59	36.31
	Ours	**76.62**	**71.17**	**96.72**	**93.88**	**42.99**	**48.4**	**70.32**	**79.08**	**74.1**	**82.97**	**62.35**	**66.97**
X-VLM	GAP	26.35	23.72	27.8	22.91	95.1	88.84	32.39	38.16	52	55.4	24.67	22.65
	ETU	22.94	21.63	22.01	19.65	96.23	92.97	28.81	34.26	48.53	52.74	20.52	19.3
	Ours	**51.46**	**65.71**	**52.8**	**64.99**	**98.89**	**95.79**	**67.42**	**75.45**	**75.49**	**82.58**	**55.74**	**66.7**
CLIP$_{\text{ViT}}$	GAP	35.96	31.91	37.33	32.56	33.42	29.25	97.71	96.04	74.63	74.67	33.47	31.99
	ETU	31.5	29.62	33.25	30.38	32.36	29.92	95.88	96.34	**82.07**	83.41	30.62	30.7
	Ours	**46.92**	**53.89**	**46.03**	**50.87**	**41.49**	**48.6**	**98.74**	**98.01**	81.58	**86.5**	**47.35**	**57.55**
CLIP$_{\text{CNN}}$	GAP	13.57	25.21	19.05	28.87	11.59	23.13	27.46	43.16	73.18	81.6	15.25	27.94
	ETU	21.71	21.92	22.33	22.8	24.77	23.93	34.99	40.3	**95.34**	95.14	20.06	22.26
	Ours	**33.41**	**47.96**	**38.81**	**50.78**	**36.59**	**48.83**	**66.04**	**72.59**	94.73	**95.21**	**42.39**	**57.84**
BLIP	GAP	12.23	23.94	14.49	25.44	6.91	17.81	20.32	37.00	26.81	43.59	47.21	73.33
	ETU	46.07	43.27	44.58	37.61	33.14	29.85	33.77	40.02	48.28	52.88	81.27	83.59
	Ours	**61.95**	**60.92**	**60.95**	**59.57**	**51.81**	**52.53**	**62.23**	**72.51**	**69.61**	**78.44**	**91.67**	**90.42**

Table 2. ASR (%) of Cross-domain attacks on six models from Flickr30k to MSCOCO and vice versa.

Setting	Source	ALBEF		TCL		X-VLM		CLIP$_{\text{ViT}}$		CLIP$_{\text{CNN}}$		BLIP	
		TR	IR	TR	IR	TR	IR	TR	IR	TR	IR	TR	IR
Flickr30K ↓ MSCOCO	ALBEF	**96.83**	**94.69**	81.46	74.87	44.79	51.64	63.68	73.06	69.77	78.09	68.88	70.61
	TCL	78.27	73.17	**97.83**	**95.03**	40.46	47.34	64.98	73.27	70.96	78.18	63.71	67.1
	X-VLM	50.63	65.91	53.23	65.65	**95.91**	**93.32**	65.51	74.72	75.69	81.93	57.69	67.28
	CLIP$_{\text{ViT}}$	49.88	53.39	49.47	52.21	47.77	48.52	**95.5**	**97.01**	83.05	85.38	50.97	57.93
	CLIP$_{\text{CNN}}$	43.05	54.19	43.04	54.39	43.73	53.94	67.3	74.39	**98.61**	**97.41**	47.22	59.11
	BLIP	54.45	55.51	55.63	53.02	41.07	46.93	61.69	69.24	65.52	75.23	**83.19**	**82.17**
MSCOCO ↓ Flickr30K	ALBEF	**88.08**	**87.28**	58.9	61.53	17.58	36.07	39.78	61.08	47.28	64.95	35.02	49.4
	TCL	47.58	53.7	**87.27**	**83.55**	18.6	34.45	51.85	72.22	59.46	76.09	37.75	53.08
	X-VLM	25.39	46.74	27.33	49.13	**79.98**	**81.72**	42.73	66.48	59.46	73.07	31.65	51.48
	CLIP$_{\text{ViT}}$	21.07	39.47	24.53	42.44	15.45	36.52	**93.97**	**95.53**	62.95	77.21	25.55	45.91
	CLIP$_{\text{CNN}}$	13.87	37.93	19.36	41.25	15.85	37.47	42.61	66.73	**85.75**	**88.76**	22.92	48.45
	BLIP	33.2	46.07	36.02	47.97	23.58	38.48	43.97	65.3	56.35	71.08	**71.91**	**73.62**

when transferring from ALBEF to TCL, confirming the superiority of our contrastive learning-based paradigm.

Cross-domain scenarios. We proceed to discuss the attack performance of the proposed algorithm in a more challenging scenario where there is an obvious distribution shift between the training dataset and the test samples. Specifically, we generate universal adversarial perturbations based on MSCOCO or Flickr30K and evaluate them accordingly on the other dataset. We present the attack success rates on the retrieval tasks across six models in Table 2. It can be observed that the domain gap indeed has a negative effect on attack performance. However, our method still maintains excellent ASR in most cases, unveiling its outstanding cross-domain transferability.

Results of R@5 and R@10. As aforementioned, we supplement the ASR of the ITR tasks based on R@5 and R@10 metrics and provide the attack success rates in Table 3. Obviously, our proposed C-PGC still consistently attains better performance than the baseline method ETU, regardless of the evaluation measurements for retrieval results.

Table 3. Attack success rates (%) regarding R@5 and R@10 metrics of our C-PGC and ETU for image-text retrieval tasks.

Dataset	Source	Method	ALBEF		TCL		X-VLM		CLIP$_{\text{ViT}}$		CLIP$_{\text{CNN}}$		BLIP	
			TR	IR	TR	IR	TR	IR	TR	IR	TR	IR	TR	IR
Flickr30K (R@5)	ALBEF	ETU	68.54	77.68	14.41	14.43	4.1	6.41	6.11	16.97	11.57	23.11	9.36	13.3
		Ours	**83.67**	**80.02**	**41.84**	**42.18**	**6.9**	**17.19**	**18.34**	**41.03**	**26.22**	**49.42**	**24.25**	**34.59**
	TCL	ETU	14.43	16.23	84.88	80.43	2.4	5.8	5.6	17.37	13.27	24.27	7.24	10.98
		Ours	**29.76**	**35.62**	**90.89**	**84.18**	**3.2**	**13.65**	**20.93**	**42.06**	**25.27**	**49.1**	**16.5**	**30.32**
	X-VLM	ETU	3.81	5.85	3.9	6.51	89.2	84.57	5.18	16.9	14.33	24.52	3.12	7.81
		Ours	**7.62**	**25.1**	**8.71**	**26.63**	**89.2**	**85.84**	**19.38**	**42.48**	**30.89**	**50.7**	**13.68**	**29**
	CLIP$_{\text{ViT}}$	ETU	5.31	7.88	7.51	10.42	4.7	8.92	76.42	84.54	35.46	41.73	5.03	10.2
		Ours	**6.31**	**17.51**	**8.01**	**19.65**	**4.3**	**15.1**	**76.89**	**85.2**	**39.6**	**54.68**	**9.15**	**23.23**
	CLIP$_{\text{CNN}}$	ETU	1.7	5.02	3.4	7.16	1.4	6.02	6.22	17.61	84.39	87.85	2.31	8.26
		Ours	**4.41**	**19.8**	6.41	**25.19**	**4.8**	**23.49**	**18.76**	**43.82**	**90.34**	**88.12**	**8.95**	**26.89**
	BLIP	ETU	8.22	10.11	6.71	10.46	3.2	5.62	5.28	16.39	9.24	21.97	45.98	73.75
		Ours	**14.43**	**21.67**	**13.91**	**21.59**	**5.4**	**14.54**	**18.03**	**36.26**	**23.89**	**44.79**	**59.26**	**74.82**
MSCOCO (R@5)	ALBEF	ETU	81.73	88.76	13.45	11.51	9.01	9.32	16.85	20.74	22.6	28.86	10.96	12.02
		Ours	**93.36**	**91.56**	**70.76**	**62.31**	**19.97**	**30.46**	**41.58**	**51.23**	**44.14**	**55.98**	**41.08**	**49.22**
	TCL	ETU	44.94	36.45	90.21	84.54	13.46	13.1	19.52	23.03	30.81	36.37	25.63	25.11
		Ours	**60.62**	**56.21**	**94.89**	**90.33**	**22.08**	**30.38**	**53.14**	**64.98**	**58.85**	**70.77**	**45.28**	**53.55**
	X-VLM	ETU	11.03	11.11	10.22	9.24	94.36	90.02	17.56	20.8	33.45	38.14	10.08	10.12
		Ours	**31.59**	**48.69**	**32.1**	**48.11**	**96.7**	**91.66**	**49.53**	**60.82**	**59.83**	**69.59**	**37.4**	**52.5**
	CLIP$_{\text{ViT}}$	ETU	15.89	16.01	18.61	16.11	16.45	16.61	93.12	94.66	**72.97**	75.5	17.1	19.26
		Ours	**25.69**	**35.95**	**24.69**	**33.14**	**21.37**	**31.38**	**96.7**	**96.49**	70.76	**77.86**	**28.72**	**42.01**
	CLIP$_{\text{CNN}}$	ETU	9.55	10.34	9.9	10.95	10.98	11.96	22.63	26.84	90.7	**92.07**	9.69	12.25
		Ours	**16.83**	**31**	**19.86**	**34.54**	**18.84**	**34.02**	**50.21**	**59.2**	**90.94**	90.43	**25.5**	**44.89**
	BLIP	ETU	29.6	29.6	26.95	21.97	17.67	16.14	20.83	24.27	32.77	36.72	76.16	80.69
		Ours	**42.56**	**43.73**	**41.72**	**41.8**	**31.05**	**35.63**	**44.37**	**57.9**	**54.47**	**66.01**	**81.71**	**81.91**
Flickr30K (R@10)	ALBEF	ETU	65.8	74.89	10	9.36	2.8	3.79	2.43	11.04	6.95	15.79	6.62	8.57
		Ours	**80.5**	**75.17**	**34.8**	**34.28**	**4.2**	**11.72**	**9.83**	**31.14**	**16.87**	**39.40**	**18.76**	**27.08**
	TCL	ETU	11.4	11.35	82.2	77.33	1.6	3.38	3.04	10.63	7.06	17.14	4.91	7.34
		Ours	**24.2**	**27.32**	**89.2**	**80.73**	**2.1**	**9.33**	**12.77**	**32.4**	**16.97**	**38.63**	**12.54**	**24.1**
	X-VLM	ETU	1.9	3.48	2.4	3.48	**87.1**	81.87	3.14	10.91	8.59	16.66	2.01	4.72
		Ours	**4.1**	**17.79**	**4.6**	**19.27**	86.3	**82.94**	**11.14**	**31.49**	**21.06**	**40.3**	**7.32**	**21.81**
	CLIP$_{\text{ViT}}$	ETU	3.6	4.75	4.6	6.43	2.7	5.47	**68.17**	78.93	27.2	32.94	3.11	6.35
		Ours	**4.2**	**11.45**	**4.6**	**13.08**	**2.8**	**10.15**	67.98	**79.46**	**29.75**	**45.56**	**5.52**	**17.23**
	CLIP$_{\text{CNN}}$	ETU	0.6	2.64	1.6	3.86	0.7	3.48	3.65	11.32	80.37	**85.18**	1.2	5.31
		Ours	**2.4**	**14.59**	**3.5**	**18.36**	**2.3**	**17.95**	**11.75**	**34.23**	**86.4**	83.83	**5.52**	**20.86**
	BLIP	ETU	6.5	6.52	4.9	6.57	1.8	3.32	2.63	10.52	5.52	14.57	42.43	71.26
		Ours	**11.2**	**15.49**	**9.1**	**14.14**	**2.8**	**10.19**	**9.83**	**27.46**	**14.83**	**34.43**	**53.46**	**72**
MSCOCO (R@10)	ALBEF	ETU	81.14	87.58	9.08	7.92	5.44	6.33	12.62	16.15	17.71	23.05	7.63	9.15
		Ours	**91.58**	**89.62**	**64.5**	**55.3**	**13.81**	**23.34**	**33.3**	**43.75**	**35.82**	**48.38**	**33.77**	**43.11**
	TCL	ETU	38.6	30.39	88.56	82.6	8.92	9.3	14.61	18.43	25.24	30.06	20.37	21.41
		Ours	**52.59**	**49.09**	**93.63**	**88.53**	**15.04**	**23.25**	**44.22**	**58.02**	**50.16**	**63.95**	**37.77**	**47.26**
	X-VLM	ETU	7.41	7.37	6.76	6.27	93.17	88.66	12.93	16.43	28.32	32.47	6.73	8.01
		Ours	**23.01**	**40.39**	**23.15**	**40.07**	**94.97**	**88.95**	**40.24**	**53.74**	**52**	**62.7**	**30.43**	**45.67**
	CLIP$_{\text{ViT}}$	ETU	11.05	11.5	13.67	11.67	11.21	12.14	91.47	93.8	**67.7**	71.14	13.03	15.4
		Ours	**17.87**	**28.52**	**17.48**	**26.09**	**14**	**24.67**	**95.55**	**95.31**	64.04	**72.75**	**22.05**	**35.65**
	CLIP$_{\text{CNN}}$	ETU	6.05	7.02	6.54	7.54	6.97	8.36	17.65	21.66	88.13	**90.16**	6.73	9.45
		Ours	**10.77**	**24.11**	**13.34**	**27.53**	**12.31**	**28.03**	**41.33**	**52.02**	**88.28**	87.14	**20.26**	**39.43**
	BLIP	ETU	23.63	24.4	19.69	16.37	11.55	11.86	16.3	19.66	26.04	30.64	73.88	77.63
		Ours	**33.64**	**36.14**	**32.15**	**33.8**	**22.64**	**28.52**	**36.07**	**50.3**	**47**	**59.07**	**78.39**	**78.98**

Table 4. ASR (%) of ITR tasks under defense strategies. The surrogate model is ALBEF and the dataset is Flick30K. LT denotes the LanguageTool that corrects adversarial words within the sentence.

Method	ALBEF		TCL		X-VLM		CLIP_{ViT}		CLIP_{CNN}		BLIP	
	TR	IR	TR	IR	TR	IR	TR	IR	TR	IR	TR	IR
Gaussian	37.92	49.49	32.4	47.04	19.31	37.79	42.49	65.61	50	**72.23**	29.65	48.77
Medium	53.13	61.6	39.54	51.96	20.43	39.69	46.31	66.92	57.9	74.51	33.75	52.68
Average	29.09	44.91	29.61	44.72	17.89	36.07	42.98	**65.42**	**49.74**	72.48	**27.55**	**46.9**
JPEG	59.3	63.7	42.34	52.52	21.65	41.58	**41.26**	65.77	53.5	72.62	37.01	55.04
DiffPure	64.34	74.63	65.22	74.8	66.06	75.19	78.08	86.7	82.25	88.03	70.45	79.09
NRP	32.33	40.63	**20.19**	39.23	**14.63**	32.62	48.4	69	59.72	74.09	30.28	52.2
NRP+LT	**29.05**	**35.23**	21.33	**37.41**	15.55	**29.63**	47.19	67.35	56.82	73.47	28.23	50.59

Table 5. ASR results of the proposed method with different loss functions on Flickr30 when the surrogate model is ALBEF.

Method	ALBEF		TCL		X-VLM		CLIP_{ViT}		CLIP_{CNN}		BLIP	
	TR	IR	TR	IR	TR	IR	TR	IR	TR	IR	TR	IR
$\mathcal{L}_{MSE}$	12.02	30.75	14.39	35.08	11.41	30.79	37.32	56.05	40.17	56.39	19.66	37.33
$\mathcal{L}_{Cos}$	57.55	67.4	37.06	49.45	10.7	28.48	37.49	58.3	40.87	58.39	23.33	39.44
$\mathcal{L}_{CL}$	**76.46**	**82.46**	**56.52**	**62.61**	**14.33**	**33.61**	**42.98**	**62.81**	**46.11**	**65.58**	**27.13**	**46.44**
$\mathcal{L}_{MSE}$+$\mathcal{L}_{Dis}$	81.09	83.71	48.76	56.54	17.58	35.72	41.5	64.72	47.41	70.34	35.96	51.76
$\mathcal{L}_{Cos}$+$\mathcal{L}_{Dis}$	65.20	72.71	36.13	50.06	18.63	36.74	42.23	65.17	50.91	69.78	36.91	50.69
$\mathcal{L}_{CL}$+$\mathcal{L}_{Dis}$	**90.13**	**88.82**	**62.11**	**64.48**	**20.53**	**39.38**	**43.1**	**65.93**	**54.4**	**72.51**	**44.79**	**56.36**

Performance under Defenses. We next analyze several defense strategies to mitigate the potential harm from C-PGC. Concretely, we totally align with TMM [8] and consider several input preprocessing-based schemes, including image smoothing [1] (Gaussian, medium, average smoothing), JPEG compression [4], NRP [6], and the DiffPure [7], a powerful purification defense using diffusion models. For adversarial text correction, we choose the LanguageTool (LT) [8], which has been widely adopted in various scenarios due to its universality and effectiveness.

The attack results in Table 4 demonstrate that the proposed attack still attains great ASR against different powerful defenses. It also indicates that NRP+LT would be a decent choice to alleviate the threat brought by C-PGC. Another noteworthy finding is that, although DiffPure [7] exhibits remarkable performance in defending attacks in classification tasks, its ability is greatly reduced in V+L scenarios since the denoising process could also diminish some texture or semantic information that is critical for VLP models, thereby acquiring unsatisfactory defense effects.

D. Rationality behind the Loss Design

It is widely acknowledged that contrastive learning serves as a powerful and foundational tool for modality alignment in VLP models, establishing a nearly point-to-point relationship between image and text features. Since contrastive learning can establish robust and precise alignment, leveraging the same technique to disrupt the established alignments is also promising to yield effective performance.

Taking image attack as an example, the underlying principle behind our contrastive learning-based attack can be understood from two perspectives:

- Leverage the originally matched texts as negative samples to push aligned image-text pairs apart. This broadly corresponds to the objective of traditional untargeted adversarial attacks.
- Additionally, the proposed paradigm introduces dissimilar texts as positive samples to further pull the adversarial image out of its original subspace and relocate it to an incorrect feature area.

By simultaneously harnessing the collaborative effects of *push* (negative samples) and *pull* (positive samples), the proposed contrastive framework achieves exceptional attack performance, which has been validated by comprehensive experimental results. Besides, we also explore several potential alternative loss functions that more directly align with the common goal of untargeted attack in Table 5, including maximizing the cosine distance $\mathcal{L}_{Cos}$ or MSE distance $\mathcal{L}_{MSE}$ between features of matched image-text pairs.

Recall that $\mathcal{L}_{CL}$ and $\mathcal{L}_{Dis}$ denote our designed contrastive loss and unimodal loss terms respectively. As observed, the integration of $\mathcal{L}_{CL}$ consistently brings significant ASR improvements, verifying the rationality and superiority of the adopted contrastive loss.

Table 6. Comparison of BERTScore between clean and adversarial texts across different surrogate models.

Method	ALBEF			TCL			CLIP_{ViT}			CLIP_{CNN}		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Co-Attack [10]	0.8328	0.8589	0.8455	0.8325	0.8588	0.8453	0.8269	0.8526	0.8394	0.8271	0.8530	0.8397
SGA [5]	0.8389	**0.8654**	0.8518	0.8376	0.8646	0.8509	0.8416	**0.8697**	0.8553	0.8378	0.8650	0.8511
Ours	**0.8891**	0.8613	**0.8748**	**0.8924**	**0.8687**	**0.8802**	**0.8746**	0.8684	**0.8713**	**0.8948**	**0.8842**	**0.8893**

E. Semantic Similarity Analysis

The basic objective of untargeted adversarial attacks is to fool the victim model to output incorrect predictions [3], while the attacker is supposed to preserve semantic similarity between original and adversarial samples to ensure attack imperceptibility. In our implementation, we follow the rigorous setup in prior works [5, 8, 10] that modify only one single word to preserve the attack stealthiness. To quantitatively analyze the influence, we provide the BERT scores [11], which calculate the P (precision), R (recall), and F1 (F1 score), to measure the semantic distance between 5,000 clean and adversarial sentences in Table 6. Note that we provide existing well-acknowledged sample-specific algorithms Co-Attack [10] and SGA [5] as references.

As observed, C-PGC generally acquires higher similarity scores than existing sample-specific methods across various surrogate models, which validates that C-PGC achieves an eligible perturbation strategy in terms of text perturbation imperceptibility. Basically, the lower semantic similarity of sample-specific approaches stems from their word-selection mechanism, which maximizes the semantic distance tailored to every input sentence for attack enhancement. *I.e.*, these methods select the adversarial word that maximizes the distance between the original and perturbed texts for every input sentence, which inherently leads to relatively larger semantic deviations. This highlights that our method achieves a better balance between efficacy and stealthiness.

F. Multimodal Alignment Destruction

To provide more intuitive evidence that our C-PGC successfully destroys the image-text alignment relationship, we compute the distance between the encoded image and text embeddings before and after applying the UAP. For an input pair (v, t), we calculate the relative distance d_{rel} by:

$$d_{rel} = \frac{||(f_I(v+\delta_v) - f_T(t \oplus \delta_t)||_2 - ||f_I(v) - f_T(t)||_2}{||f_I(v) - f_T(t)||_2}. \tag{1}$$

We provide the distances averaged on 5000 image-text pairs from Flickr30K in Table 7. Benefiting from our delicate designs, C-PGC achieves better disruption of the aligned multimodal relationship, thereby boosting the generalization ability and transferability of the produced UAP.

Table 7. Average relative cross-modal feature distances.

Source	Method	ALBEF	TCL	BLIP	X-VLM	CLIP_{ViT}	CLIP_{CNN}
ALBEF	GAP	7.18	6.54	0.91	1.74	0.31	0.98
	ETU	8.70	8.02	0.81	2.85	0.36	1.32
	C-PGC	**8.83**	**14.95**	**2.73**	**6.09**	**3.42**	**3.92**
TCL	GAP	4.02	24.27	0.91	0.87	0.12	0.07
	ETU	5.57	26.32	0.55	2.56	1.47	0.55
	C-PGC	**6.43**	**27.11**	**3.64**	**4.35**	**2.56**	**2.94**
BLIP	GAP	3.17	4.67	11.82	1.74	-1.71	-0.98
	ETU	5.26	6.03	13.14	2.90	0.25	1.14
	C-PGC	**6.41**	**12.15**	**13.64**	**4.35**	**1.71**	**1.96**

G. Detailed Architecture of the Generator

The architecture of our cross-modal knowledge conditioned perturbation generator is illustrated in Figure 2, which primarily consists of several deconvolution layers for upsampling and transformer layers for cross-modal information.

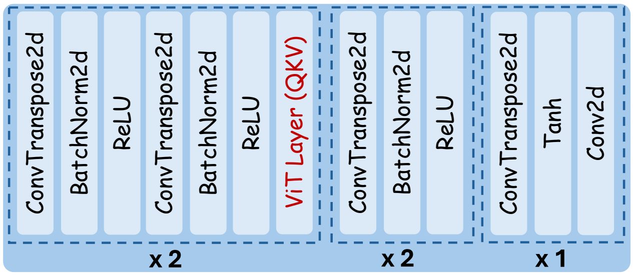

Fig. 2: Illustration of the designed generator's architecture.

H. More Visualization Results

This section presents a rich visual analysis of the proposed attack on a series of downstream tasks. Specifically, we first provide the visualization of the image retrieval task using the MSCOCO dataset in Fig. 3. Besides, we generate the UAP using the ALBEF model for the visual grounding (VG) task. As illustrated in Fig. 4, the prediction bounding boxes exhibit a notable deviation from the clean predictions, verifying that our generated adversarial samples significantly interfere with the multimodal alignment. In the visual entailment (VE) task, we employ BLIP as the victim model and present the results in Fig. 5. These qualitative visualizations again demonstrate the remarkable attack effects of our proposed method on various downstream tasks.

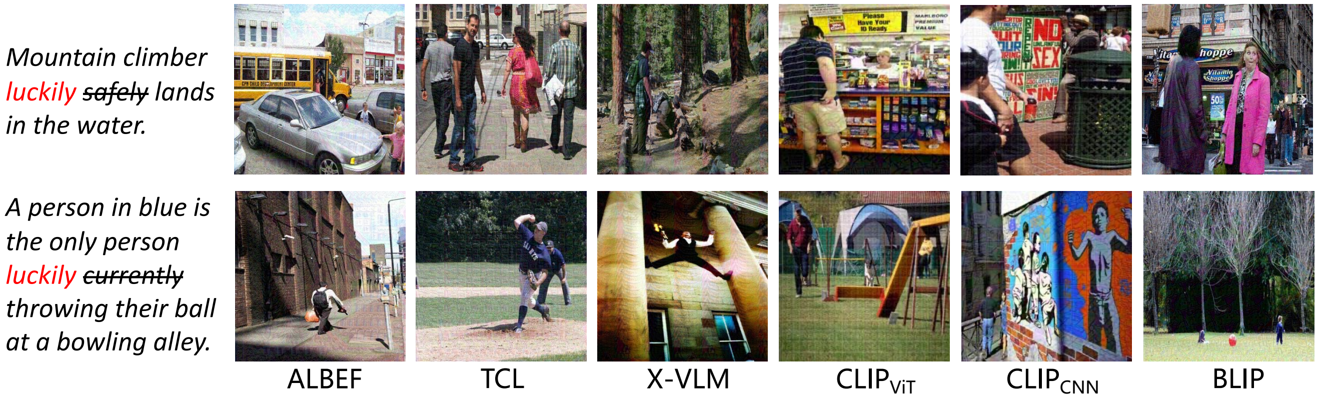
Mountain climber luckily ~~safely~~ lands in the water.
A person in blue is the only person luckily ~~currently~~ throwing their ball at a bowling alley.
ALBEF
TCL
X-VLM
CLIP$_{\text{ViT}}$
CLIP$_{\text{CNN}}$
BLIP

Fig. 3: Attacks results of C-PGC on the image retrieval task. The red indicates the UAP and the crossed-out word is the replaced one. We generate the word on ALBEF and test it on 6 target models. All retrieved images fail to correspond to the query text, validating the rationality of our contrastive learning-based attacks.

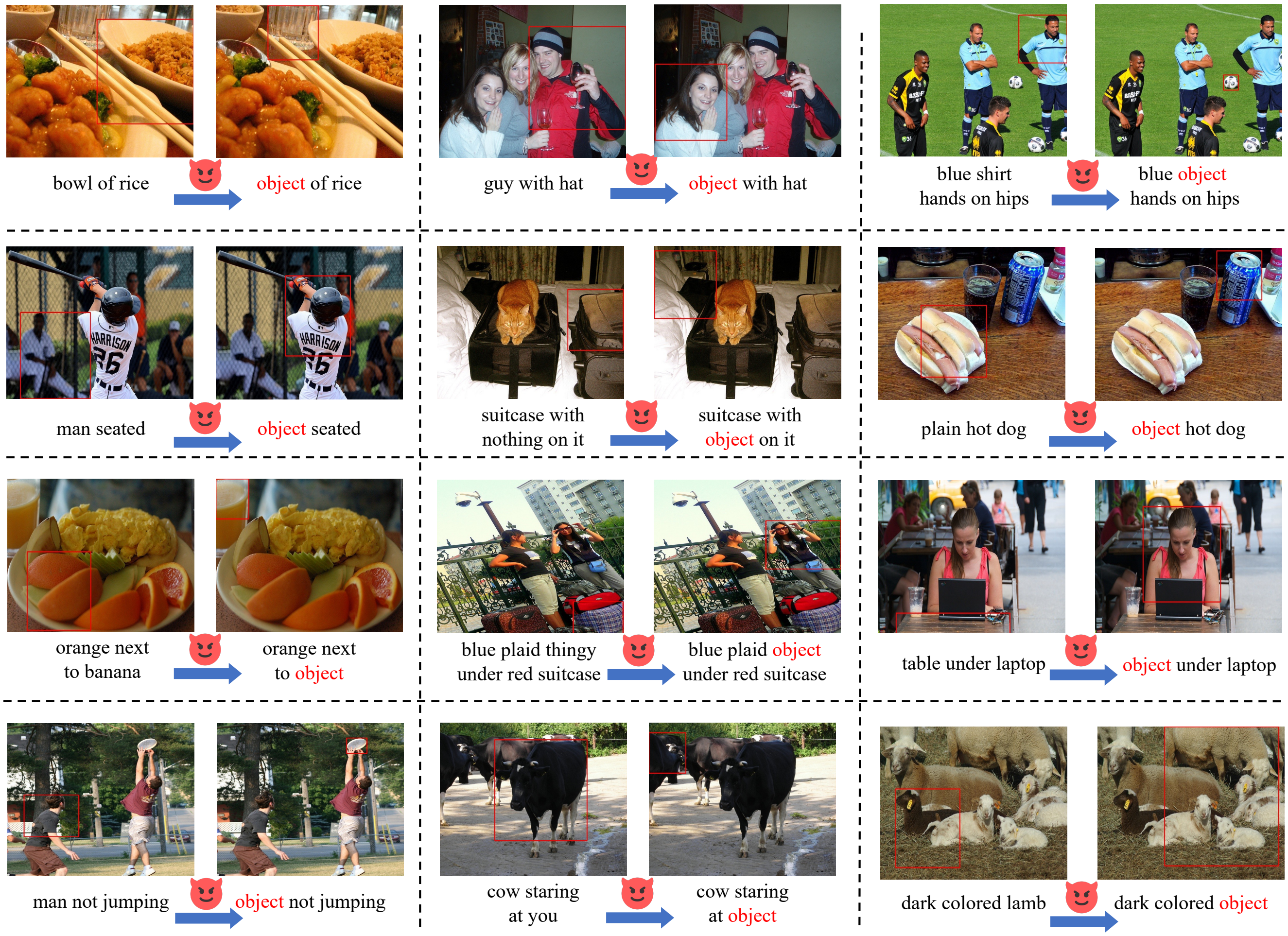

Fig. 4: Illustration of visual grounding. The predictions of clean pairs are on the left while the predictions of adversarial samples are on the right. The red word is the adversarial word perturbation.

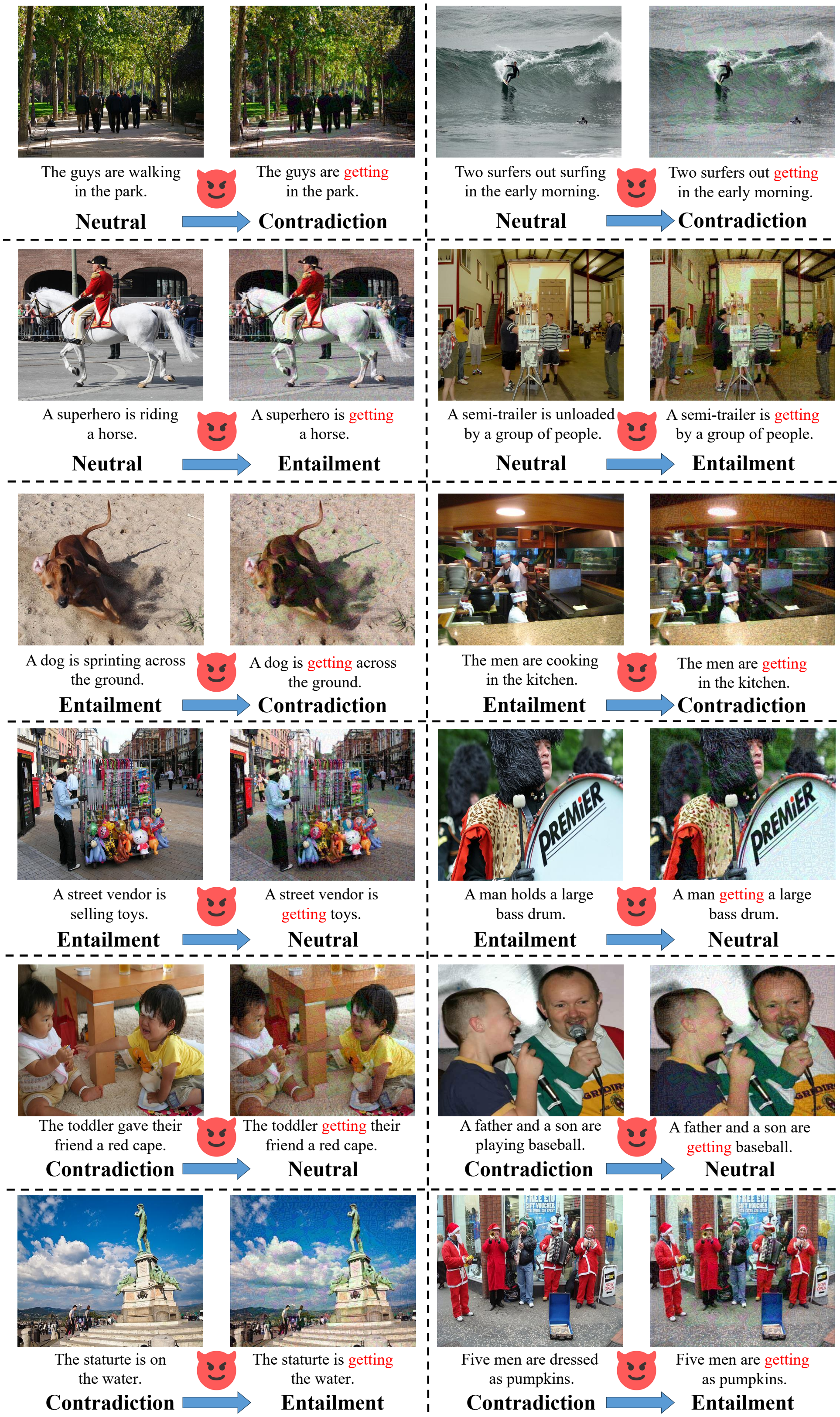
The guys are walking in the park.
Neutral
The guys are getting in the park.
Contradiction
Two surfers out surfing in the early morning.
Neutral
Two surfers out getting in the early morning.
Contradiction
A superhero is riding a horse.
Neutral
A superhero is getting a horse.
Entailment
A semi-trailer is unloaded by a group of people.
Neutral
A semi-trailer is getting by a group of people.
Entailment
A dog is sprinting across the ground.
Entailment
A dog is getting across the ground.
Contradiction
The men are cooking in the kitchen.
Entailment
The men are getting in the kitchen.
Contradiction
A street vendor is selling toys.
Entailment
A street vendor is getting toys.
Neutral
A man holds a large bass drum.
Entailment
A man getting a large bass drum.
Neutral
The toddler gave their friend a red cape.
Contradiction
The toddler getting their friend a red cape.
Neutral
A father and a son are playing baseball.
Contradiction
A father and a son are getting baseball.
Neutral
The staturte is on the water.
Contradiction
The staturte is getting the water.
Entailment
Five men are dressed as pumpkins.
Contradiction
Five men are getting as pumpkins.
Entailment

Fig. 5: Illustration of the visual entailment task. The red indicates the universal adversarial word. It can be observed that all predictions do not match with the ground truth.

References

[1] Gavin Weiguang Ding, Luyu Wang, and Xiaomeng Jin. Advertorch v0. 1: An adversarial robustness toolbox based on pytorch. *arXiv preprint arXiv:1902.07623*, 2019. 4

[2] Virginie Do, Oana-Maria Camburu, Zeynep Akata, and Thomas Lukasiewicz. e-snli-ve: Corrected visual-textual entailment with natural language explanations. *arXiv preprint arXiv:2004.03744*, 2020. 1

[3] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9185–9193, 2018. 5

[4] Gintare Karolina Dziugaite, Zoubin Ghahramani, and Daniel M Roy. A study of the effect of jpg compression on adversarial images. *arXiv preprint arXiv:1608.00853*, 2016. 4

[5] Dong Lu, Zhiqiang Wang, Teng Wang, Weili Guan, Hongchang Gao, and Feng Zheng. Set-level guidance attack: Boosting adversarial transferability of vision-language pre-training models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 102–111, 2023. 1, 5

[6] Muzammal Naseer, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Fatih Porikli. A self-supervised approach for adversarial robustness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 262–271, 2020. 4

[7] Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, and Anima Anandkumar. Diffusion models for adversarial purification. *arXiv preprint arXiv:2205.07460*, 2022. 4

[8] Haodi Wang, Kai Dong, Zhilei Zhu, Haotong Qin, Aishan Liu, Xiaolin Fang, Jiakai Wang, and Xianglong Liu. Transferable multimodal attack on vision-language pre-training models. In *2024 IEEE Symposium on Security and Privacy (SP)*, pages 102–102. IEEE Computer Society, 2024. 1, 4, 5

[9] Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. Visual entailment: A novel task for fine-grained image understanding. *arXiv preprint arXiv:1901.06706*, 2019. 1

[10] Jiaming Zhang, Qi Yi, and Jitao Sang. Towards adversarial attack on vision-language pre-training models. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 5005–5013, 2022. 1, 5

[11] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*. 5