Benchmarking Robustness under Distribution Shift of Multimodal Image-Text Models

Jielin Qiu[◇,♣,*] **Yi Zhu**[◇], **Xingjian Shi**[◇], **Zhiqiang Tang**[◇], **Ding Zhao**[♣], **Bo Li**[◇,♦], **Mu Li**[◇]
[◇]Amazon Web Services, [♣]Carnegie Mellon University, [♦]University of Illinois Urbana-Champaign
{jielinq, yzaws, xjshi, zqtang, mli}@amazon.com, dingzhao@cmu.edu, lbo@illinois.edu

Abstract

Multimodal image-text models have shown remarkable performance in the past few years. However, the robustness of such foundation models against distribution shifts is crucial in downstream applications. In this paper, we investigate their robustness under image and text perturbations. We first build several multimodal benchmark datasets by applying 17 image perturbation and 16 text perturbation techniques. Then we extensively study the robustness of 6 widely adopted models on 3 downstream tasks (image-text retrieval, visual reasoning, and visual entailment). We observe that these powerful multimodal models are sensitive to image/text perturbations, especially to image perturbations. For text, character-level perturbations have shown higher adversarial impact than word-level and sentence-level perturbations. We also observe that models trained by generative objectives tend to be more robust. Our findings in terms of robustness study could facilitate the development of large image-text models, as well as their deployment for real-world applications.

1 Introduction and Related Work

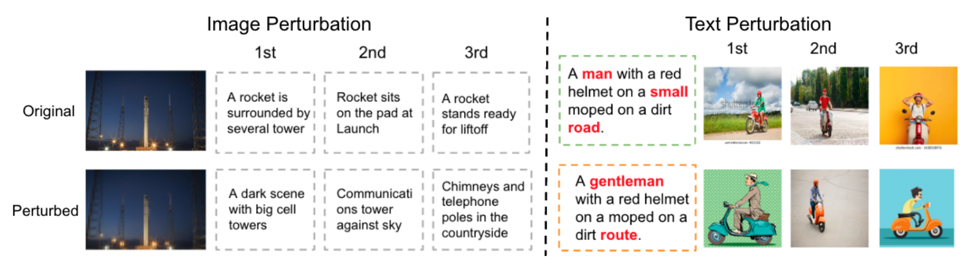

Figure 1: Multimodal models are sensitive to image/text perturbations. Take the image-text retrieval task as an example, perturbed image (i.e., adding pixelation) or perturbed text (i.e., synonym replacement) can both lead to inaccurate retrieval results.

Multimodal learning has drawn increasing attention in the past few years [9, 22, 47, 48, 94, 64, 41, 45, 44, 88, 15, 65, 79, 1, 64, 91]. Many datasets and models are collected and proposed to accelerate research in this field. However, despite the extraordinary performance and exciting potential, multimodal models might be vulnerable under distribution shifts. In Figure 1, we show an example of CLIP [64] model's performance on image-text retrieval under image or text perturbations. When pixelation is applied to the original image for image-to-text retrieval, the perturbed image retrieves less relevant or even wrong texts. For text perturbation, we replace words with their synonym and delete words for text-to-image retrieval. We find the retrieved images changed dramatically even

[*]Work done during internship at AWS

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

though the semantics of the sentence didn't change. Similar findings have also been observed in previous works [21, 20, 26, 62].

There is a sizable literature on robustness evaluation of unimodal vision models [89, 95, 16, 13, 27, 63, 3, 55, 57, 2, 96] or unimodal language models [81, 7, 80, 68, 24, 73, 14, 28, 56, 78]. However, robustness evaluation of multimodal image-text models under distribution shift has rarely been studied [25, 12]. (More related work can be found in Appendix 5.8). To our best knowledge, there is currently no benchmark dataset nor a comprehensive study of how the perturbed data can affect their performance. In this work, we would like to (1) build robustness evaluation benchmarks for multimodal image-text models, and (2) investigate these models' robustness under image or text perturbations in downstream applications. Our contributions can be summarized as follows:

- We build multimodal robustness evaluation benchmarks by leveraging existing datasets and tasks, e.g., image-text retrieval (Flicker30K, COCO), visual reasoning (NLVR2), and visual entailment (SNLI-VE). We design 17 image perturbation and 16 text perturbation strategies to extend them to multimodal evaluation settings.
- We observe that multimodal image-text models are more sensitive to image perturbations than text perturbations, while for text perturbations, character-level perturbations showed higher impact than word-level and sentence-level perturbations.
- We introduce a new metric, termed MMI (MultiModal Impact score), to account for the relative performance drop under distribution shift in downstream applications.

2 Multimodal Robustness Benchmark under Distribution Shift

To evaluate the robustness of large pretrained multimodal models under distribution shift, we start by building several evaluation benchmark datasets, by perturbing the original image-text pairs on either image side or text side.

Image Perturbation In this work, we adopt the perturbation strategies from ImageNet-C [33] and Stylize-ImageNet [23, 58]. The reason we include Stylize-ImageNet is because it is an effective method to perturb the original image by breaking its shape and texture [23]. The perturbations are drawn into five categories: Noise, Blur, Weather, Digital, and Stylize. Specifically, we have 17 image pertubation techniques **(1) Noise: Gaussian Noise, Shot Noise, Impulse Noise, Speckle Noise; (2) Blur: Defocus Blur, Frosted Glass Blur, Motion Blur, Zoom Blur; (3) Weather: Snow, Frost, Fog, Brightness; (4) Digital: Contrast, Elastic, Pixelate, JPEG Compression; and (5) Stylize**. Note that real-world corruptions can manifest themselves at varying intensities, we thus introduce variation for each corruption following [33, 23, 58]. In our evaluation setting, each category has five levels of severity $[1, 2, 3, 4, 5]$, resulting in 85 perturbation methods in total. More details can be found in the Appendix 5.2. These strategies are commonly considered as synthetic distribution shifts, and can serve as a good starting point since they are precisely defined and easy to apply. Examples of perturbed images from COCO dataset [50] are shown in the Appendix 5.4.

Text Perturbation To simulate the real-world distribution shift in language, we design the text perturbation into three categories: character-level, word-level and sentence-level. In detail, for character-level perturbation, we adopt 6 strategies from [54], including **Keyboard**, **OCR**, **Character Insert (CI)**, **Character Replace (CR)**, **Character Swap (CS)**, **Character Delete (CD)**. These perturbations can be considered as simulating real-world typos or mistakes during typing. For word-level perturbation, we adopt 5 strategies from EDA and AEDA [83, 40], including **Synonym Replacement (SR)**, **Word Insertion (WR)**, **Word Swap (WS)**, **Word Deletion (WD)**, and **Insert punctuation (IP)**. These perturbations aim to simulate different writing habits that people may replace, delete, or add words to express the same meaning. For sentence-level perturbation, (1) we first adopt the style transformation strategies from [42, 18, 70, 69], i.e., transferring text style into **formal**, **casual**, **passive**, and **active**; (2) we also adopt the **back translation** method from [54]. These perturbations will focus more on language semantics, due to the differences of speaking/writing styles, or translation error. For strategies within character-level and word-level, we apply 5 perturbation levels $[0.15, 0.20, 0.25, 0.30, 0.35]$, while for strategies within character-level, there is only one level. This leads to 60 text perturbation methods in total. Examples of the text perturbation of captions in Flickr30K dataset [90] and more details about each text perturbation strategy can be found in the Appendix 5.3 and Appendix 5.5.

Evaluation Tasks and Datasets We select three widely adopted downstream tasks for a comprehensive evaluation on the robustness of multimodal image-text models, including image-text retrieval, visual reasoning (VR), and visual entailment (VE). For each task, we perturbed the corresponding datasets, i.e., Flickr30K [90] and COCO [50] for image-text retrieval, and NLVR2 [74] for visual reasoning, SNLI-VE [86, 87] for visual entailment, using the image perturbation (IP) and text perturbation (TP) methods introduced above, which results in: (1) Flickr30K-IP, Flickr30K-TP, COCO-IP, and COCO-TP for image-text retrieval evaluation; (2) NLVR2-IP and NLVR2-TP for visual reasoning evaluation; and (3) SNLI-VE-IP and SNLI-VE-TP for visual entailment evaluation.

3 Experiments and Results

For evaluation, we select six representative large pretrained multimodal models which publicly released their pretrained models, including CLIP [64], ViLT [41], ALBEF [45], BLIP [44], TCL [88], and METER [15]. To qualitatively analyze the multimodal image-text models' robustness under perturbations, we propose a new impact score MMI (multimodal impact score) to calculate the averaged performance ("ave") drop compared with the non-perturbed performance ("clean"), which is defined as: MMI $= (s_c - s_p)/s_c$, where s_p is the perturbed score and s_c is the clean score. More experimental settings can be found in the Appendix 5.6.

Table 1: Image-Text Retrieval results of IP dataset (averaged RSUM), where the most effective perturbation results are marked bold and the least effective perturbation results are underlined.

			Noise				Blur				Weather				Digital				Stylize		
Dataset	Method	Clean	Gauss.	Shot	Impulse	Speckle	Defocus	Glass	Motion	Zoom	Snow	Frost	Fog	Bright	Contrast	Elastic	Pixel	JPEG	Stylize	**ave**	Impact
Flickr30K	CLIP ZS	533.7	501.7	504.2	481.2	515.5	502.1	530.1	509.7	457.8	470.7	495.6	519.7	530.1	515.4	510.4	469.5	524.6	**447.6**	499.2	↓ 6.5%
	CLIP FT	544.3	500.1	503.8	479.1	522.1	493.3	536.9	513.3	**444.4**	464.4	503.2	529.7	543.5	521.5	513.9	453.9	528.6	436.9	499.3	↓ 8.3%
	TCL ZS	563.8	464.9	467.0	458.4	498.0	429.8	506.6	388.5	**251.3**	407.3	449.5	434.2	509.1	473.2	434.4	247.2	502.2	343.4	427.4	↓ 24.2%
	TCL FT	573.4	529.9	532.6	527.7	551.6	504.5	566.0	513.9	397.3	521.7	551.0	554.1	568.0	557.1	421.0	**372.0**	555.4	448.7	516.2	↓ 10.0%
	ALBEF FT	577.7	533.8	538.3	532.0	557.8	528.8	569.2	516.0	**416.1**	532.0	558.1	560.4	572.0	550.6	538.7	435.9	559.8	464.1	527.3	↓ 8.7%
	BLIP FT	580.9	536.2	538.9	528.6	560.8	529.4	571.6	525.7	**412.1**	456.6	513.4	568.5	574.4	555.1	545.6	490.8	563.8	482.1	527.2	↓ 9.2%
COCO	CLIP ZS	394.5	363.0	361.2	330.2	368.7	358.7	391.6	362.2	294.6	**294.7**	329.0	371.8	391.9	356.4	369.7	308.2	388.0	314.9	350.3	↓ 11.2%
	CLIP FT	420.5	367.2	365.3	331.7	381.5	371.0	412.2	374.4	291.0	**289.3**	337.3	389.9	413.9	371.7	379.7	306.4	402.1	310.2	358.5	↓ 14,7%
	TCL ZS	477.2	419.8	418.4	418.4	439.0	400.0	450.8	357.5	**177.3**	316.5	372.0	400.6	452.2	416.1	369.0	190.3	442.7	280.1	371.8	↓ 22.1%
	TCL FT	497.2	454.3	454.4	453.9	468.1	447.8	491.9	433.8	**259.9**	408.9	443.2	470.1	489.1	467.8	438.2	309.1	474.9	360.9	430.9	↓ 13.3%
	ALBEF FT	504.6	460.0	460.6	460.3	376.4	447.1	493.0	436.5	**282.2**	408.8	449.8	472.6	493.8	452.1	455.0	347.0	480.9	475.8	438.3	↓ 13.1%
	BLIP FT	516.6	471.9	472.1	467.7	489.5	466.1	507.2	451.7	**291.6**	432.8	471.8	494.2	506.8	470.4	472.3	404.7	499.6	402.9	458.7	↓ 11.2%

Table 2: Image-Text Retrieval results of TP dataset (averaged RSUM), where the most effective perturbation results are marked bold and the least effective perturbation results are underlined.

			Character-level						Word-level					Sentence-level						
Dataset	Method	Clean	Keyboard	OCR	CI	CR	CS	CD	SR	WI	WS	WD	IP	Formal	Casual	Passive	Active	Back_trans	**ave**	Impact
Flickr30K	CLIP ZS	533.7	**431.8**	478.2	450.5	435.2	444.6	451.3	497.1	509.6	503.3	514.1	519.4	531.7	529.3	524.8	531.4	524.2	492.3	↓ 7.8%
	CLIP FT	544.3	**458.4**	500.1	477.6	461.6	471.1	475.5	515.4	530.4	526.0	531.1	536.4	545.8	542.1	537.9	545.1	537.3	512.0	↓ 5.9%
	TCL ZS	563.8	433.3	499.9	443.3	**428.4**	444.4	448.9	511.9	523.8	519.1	528.8	548.6	544.4	542.4	530.1	547.1	535.8	501.9	↓ 11.0%
	TCL FT	573.4	494.3	545.0	504.9	**492.8**	501.9	502.4	554.7	566.4	560.0	564.2	573.4	571.5	569.6	562.8	572.1	566.5	543.9	↓ 5.1%
	ALBEF FT	577.7	506.2	552.0	516.2	**505.0**	511.7	513.0	561.9	571.6	568.6	570.0	577.7	576.2	575.0	569.5	576.4	572.5	551.5	↓ 4.5%
	BLIP FT	580.9	**518.0**	559.5	527.3	**518.0**	526.4	525.7	565.6	576.1	572.8	573.8	580.7	579.0	578.6	574.5	579.6	574.7	558.1	↓ 3.9%
COCO	CLIP ZS	394.5	285.5	286.4	286.1	**285.4**	285.6	285.8	347.5	363.8	355.5	368.6	374.2	393.0	391.6	379.6	393.5	381.2	341.5	↓ 13.4%
	CLIP FT	420.5	316.1	316.7	316.5	**316.4**	316.7	315.6	376.2	394.6	389.9	395.3	406.6	417.3	415.2	408.7	419.4	406.2	370.5	↓ 11.9%
	TCL ZS	477.2	**368.0**	428.4	381.3	368.4	382.0	383.4	439.3	453.4	445.7	450.9	477.2	474.4	471.8	464.7	475.7	462.0	432.9	↓ 9.3%
	TCL FT	497.2	**397.8**	455.1	412.0	398.5	408.8	410.5	463.7	481.3	471.8	477.7	497.1	494.6	493.0	487.3	496.0	483.5	458.0	↓ 7.9%
	ALBEF FT	504.6	**404.5**	461.7	418.9	406.1	414.7	415.5	471.4	488.9	483.3	486.3	504.5	503.1	502.0	496.4	503.7	491.3	465.8	↓ 7.7%
	BLIP FT	516.6	**429.1**	479.1	442.4	430.8	441.3	441.4	484.3	502.1	494.6	499.7	515.8	514.4	513.6	508.1	515.4	504.3	482.3	↓ 6.6%

Discussion To emphasize the important findings, we provide a summary of the experiments. (More discussion can be found in the Appendix 5.7.) According to our impact score, overall, both image and text perturbation methods can effectively attack the current multimodal image-text models, for image-text retrieval, visual reasoning, and visual entailment tasks. In general, models are more sensitive to image perturbations than text perturbations. We also observe that models trained by generative objectives tend to be more robust. In addition, different models' sensitivity to perturbation methods is also very different. To combine the similarities, we found that Zoom Blur shows a consistently high impact in three downstream tasks across different models as an effective image perturbation method. In contrast, Glass Blur and Brightness are less effective in attacking models. From text perturbation results, Keyboard and CR could be the two powerful perturbation methods, while sentence-level perturbation methods along with IP (Insert Punctuation) seem to be "soft" perturbation methods that rarely have a significant impact on models' performance.

Table 3: Visual reasoning evaluation results of NLVR2-IP dataset (averaged accuracy), where the most effective perturbation results are marked bold and the least effective ones are underlined.

			Noise				Blur				Weather				Digital				Stylize		
Dataset	Method	Clean	Gauss.	Shot	Impulse	Speckle	Defocus	Glass	Motion	Zoom	Snow	Frost	Fog	Bright	Contrast	Elastic	Pixel	JPEG	Stylize	**ave**	Impact
dev	ALBEF	82.55	<u>52.80</u>	52.46	52.61	52.63	52.22	52.44	51.78	50.79	**50.69**	52.05	52.58	52.09	51.98	52.45	50.99	52.37	51.80	52.04	↓ 37.0%
	ViLT	75.70	71.64	71.45	71.58	72.42	72.90	<u>74.71</u>	68.79	**63.97**	69.40	73.02	73.59	74.32	66.72	74.15	69.17	<u>74.71</u>	72.35	71.46	↓ 5.6%
	BLIP	82.48	<u>85.37</u>	78.54	72.68	76.59	80.00	73.66	78.54	**60.98**	73.66	76.59	83.90	76.10	77.07	81.46	74.63	82.93	71.71	77.42	↓ 6.1%
	TCL	80.54	78.20	77.63	78.21	78.60	77.04	<u>81.20</u>	77.37	**66.67**	75.96	79.47	79.65	80.76	74.04	78.92	73.92	81.01	75.05	77.28	↓ 4.0%
	METER	82.33	77.39	76.25	77.25	77.76	78.76	<u>82.01</u>	78.26	**69.31**	76.17	79.40	81.02	80.76	77.50	79.36	72.91	80.67	76.10	77.70	↓ 5.6%
test-P	ALBEF	83.14	53.17	52.85	53.22	53.50	52.68	53.09	52.39	**51.19**	51.60	52.98	<u>53.49</u>	52.78	53.13	53.12	51.72	53.10	52.95	52.76	↓ 36.5%
	ViLT	76.13	74.24	73.80	74.43	74.20	72.32	<u>76.70</u>	72.55	**62.34**	69.24	73.36	75.05	74.73	68.68	74.07	69.06	76.52	71.50	72.54	↓ 4.7%
	BLIP	83.08	75.39	75.39	85.10	72.31	<u>85.64</u>	79.49	76.92	**58.97**	80.51	75.90	81.54	76.92	81.03	77.95	73.333	78.97	73.85	77.01	↓ 7.3%
	TCL	81.33	78.10	77.87	78.25	78.91	78.00	<u>81.59</u>	78.17	**67.81**	75.74	79.62	80.64	81.52	74.35	79.76	74.61	81.28	75.85	77.77	↓ 4.4%
	METER	83.05	78.87	77.94	77.78	79.23	78.97	<u>82.10</u>	79.14	**68.89**	76.69	80.10	82.25	81.21	78.20	79.91	72.65	80.74	76.93	78.34	↓ 5.7%

Table 4: Visual reasoning evaluation results of NLVR2-TP dataset (averaged accuracy), where the most effective perturbation results are marked bold and the least effective ones are underlined.

			Character-level						Word-level					Sentence-level						
Dataset	Method	Clean	Keyboard	OCR	CI	CR	CS	CD	SR	WI	WS	WD	IP	Formal	Casual	Passive	Active	Back_trans	**ave**	Impact
dev	ALBEF	82.55	50.64	51.02	50.81	50.66	**50.53**	50.58	<u>51.96</u>	51.48	51.58	51.39	51.56	50.99	51.93	51.52	51.75	51.90	51.22	↓ 38.0%
	ViLT	75.70	66.23	69.16	65.47	**64.36**	64.76	64.96	67.11	72.71	70.77	71.75	73.42	73.22	73.40	71.83	74.47	<u>74.51</u>	69.88	↓ 7.7%
	TCL	80.54	71.15	75.89	71.84	**70.99**	72.01	71.58	74.96	78.89	77.84	78.05	<u>82.37</u>	81.56	80.33	79.47	81.46	80.67	71.77	↓ 10.9%
	BLIP	82.48	70.73	**70.24**	76.59	74.63	72.68	72.20	73.17	77.56	80.00	79.51	<u>87.81</u>	85.37	82.93	82.93	<u>87.81</u>	75.61	78.11	↓ 5.3%
	METER	82.33	**72.35**	75.83	74.10	72.71	73.89	73.30	75.16	79.36	75.41	77.64	81.68	<u>81.92</u>	81.55	78.69	81.01	82.25	77.30	↓ 6.1%
test-P	ALBEF	83.14	51.39	51.99	**51.04**	51.26	51.05	51.24	52.69	52.95	52.95	52.88	53.30	<u>53.39</u>	53.06	52.68	53.26	53.23	52.40	↓ 37.0%
	ViLT	76.13	64.85	69.66	66.76	65.64	65.56	**65.14**	68.96	73.36	71.35	72.53	75.14	75.86	74.27	72.58	<u>77.00</u>	75.70	70.90	↓ 6.9%
	TCL	81.33	**71.16**	76.31	72.35	71.56	71.90	72.07	75.49	80.03	78.80	78.78	<u>82.88</u>	82.46	81.52	80.25	82.28	81.53	72.37	↓ 11.0%
	BLIP	83.08	**67.69**	85.64	67.18	67.69	75.90	74.87	69.23	72.82	78.46	83.59	83.59	79.49	<u>87.18</u>	82.05	82.05	74.36	76.99	↓ 7.3%
	METER	83.05	73.10	77.63	74.05	72.49	**70.64**	74.27	76.10	79.62	75.96	78.55	<u>82.58</u>	81.87	80.42	79.52	82.34	81.45	77.54	↓ 6.6%

Table 5: Visual entailment evaluation results of SNLI-VE-IP dataset (averaged accuracy), where the most effective perturbation results are marked bold and the least effective ones are underlined.

			Noise				Blur				Weather				Digital				Stylize		
Dataset	Method	Clean	Gauss.	Shot	Impulse	Speckle	Defocus	Glass	Motion	Zoom	Snow	Frost	Fog	Bright	Contrast	Elastic	Pixel	JPEG	Stylize	**ave**	Impact
val	ALBEF	80.80	77.52	77.56	77.34	78.76	76.59	79.26	76.67	**71.70**	75.61	78.71	78.76	<u>79.83</u>	78.19	78.49	74.29	78.91	74.58	77.22	↓ 4.4%
	TCL	80.51	77.33	77.56	77.22	78.23	76.70	79.21	75.25	**70.98**	75.71	77.95	78.43	<u>79.31</u>	78.76	77.78	71.47	78.43	74.64	76.76	↓ 4.7%
	METER	80.86	77.05	77.19	76.76	78.37	77.14	79.72	77.04	**74.35**	77.18	79.38	80.10	<u>80.49</u>	79.12	78.78	73.08	78.93	75.88	77.68	↓ 3.9%
test	ALBEF	80.91	77.65	77.70	77.40	78.50	76.62	79.25	76.59	**71.70**	76.31	78.60	78.47	<u>79.77</u>	78.07	78.34	74.42	78.81	74.89	77.24	↓ 8.3%
	TCL	80.29	77.46	77.38	77.30	78.17	76.80	79.27	75.56	**71.07**	76.13	78.24	78.38	<u>79.19</u>	78.68	77.74	71.76	78.59	74.70	76.85	↓ 4.3%
	METER	81.19	77.16	77.09	76.90	78.58	77.14	80.13	77.39	**74.35**	77.79	79.84	80.18	<u>80.46</u>	79.18	78.91	72.67	79.32	76.08	77.79	↓ 4.2%

Table 6: Visual entailment evaluation results of SNLI-VE-TP dataset (averaged accuracy), where the most effective perturbation results are marked bold and the least effective ones are underlined.

			Character-level						Word-level					Sentence-level						
Dataset	Method	Clean	Keyboard	OCR	CI	CR	CS	CD	SR	WI	WS	WD	IP	Formal	Casual	Passive	Active	Back_trans	**ave**	Impact
val	ALBEF	80.80	65.35	71.97	66.54	**65.17**	67.22	67.46	74.63	74.15	74.88	78.62	<u>80.56</u>	<u>80.56</u>	<u>80.56</u>	<u>80.56</u>	<u>80.56</u>	76.94	74.11	↓ 8.3%
	TCL	80.51	65.24	71.63	65.58	**64.72**	67.67	67.16	74.32	74.04	74.52	77.84	<u>79.84</u>	<u>79.84</u>	<u>79.84</u>	<u>79.84</u>	<u>79.84</u>	75.79	73.61	↓ 8.6%
	METER	80.86	**66.70**	74.17	67.99	66.41	68.64	69.53	74.65	73.19	72.55	78.28	76.24	80.72	80.49	<u>80.76</u>	80.72	77.43	74.28	↓ 8.1%
test	ALBEF	80.91	**64.87**	71.90	65.99	65.03	66.91	67.27	74.77	74.93	74.90	78.44	<u>80.20</u>	<u>80.20</u>	<u>80.20</u>	<u>80.20</u>	<u>80.20</u>	77.31	73.96	↓ 8.6%
	TCL	80.29	**65.27**	71.83	65.81	64.66	67.69	67.25	74.59	73.70	74.49	78.01	79.77	79.77	79.77	<u>79.84</u>	<u>79.84</u>	76.62	73.67	↓ 8.2%
	METER	81.19	**66.09**	74.26	67.39	66.30	68.92	69.71	74.88	73.89	72.95	78.38	76.65	80.96	80.83	<u>81.21</u>	81.05	77.14	74.41	↓ 8.4%

4 Conclusion

In the study, we investigate the robustness of large multimodal pretrained image-text models. We introduce several evaluation benchmarks under distribution shift by applying 17 image perturbation and 16 text perturbation strategies. We select three downstream tasks, including image-text retrieval, visual reasoning, and visual entailment, to evaluate 6 popular models. Our developed multimodal perturbation datasets could serve as robustness evaluation benchmarks for image-text models. We hope our findings could provide inspiration on how to develop and deploy more robust models for real-world applications.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning. *ArXiv*, abs/2204.14198, 2022.

[2] Ahmed Aldahdooh, Wassim Hamidouche, and Olivier Déforges. Reveal of vision transformers robustness against adversarial attacks. *ArXiv*, abs/2106.03734, 2021.

[3] Srinadh Bhojanapalli, Ayan Chakrabarti, Daniel Glasner, Daliang Li, Thomas Unterthiner, and Andreas Veit. Understanding robustness of transformers for image classification. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10211–10221, 2021.

[4] Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Yue Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph P. Turian. Experience grounds language. In *EMNLP*, 2020.

[5] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *ArXiv*, abs/2005.14165, 2020.

[6] Min Cao, Shiping Li, Juntao Li, Liqiang Nie, and Min Zhang. Image-text retrieval: A survey on recent research and development. *ArXiv*, abs/2203.14713, 2022.

[7] Kai-Wei Chang, He He, Robin Jia, and Sameer Singh. Robustness and adversarial examples in natural language processing. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*, 2021.

[8] Xi Chen, Xiao Wang, Soravit Changpinyo, A. J. Piergiovanni, Piotr Padlewski, Daniel M. Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish V. Thapliyal, James Bradbury, Weicheng Kuo, Mojtaba Seyedhosseini, Chao Jia, Burcu Karagol Ayan, Carlos Riquelme, Andreas Steiner, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. Pali: A jointly-scaled multilingual language-image model. *ArXiv*, abs/2209.06794, 2022.

[9] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *ECCV*, 2020.

[10] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek B Rao, Parker Barnes, Yi Tay, Noam M. Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Benton C. Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier García, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Oliveira Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Díaz, Orhan Firat, Michele Catasta, Jason Wei, Kathleen S. Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. Palm: Scaling language modeling with pathways. *ArXiv*, abs/2204.02311, 2022.

[11] Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Edoardo Debenedetti, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. Robustbench: a standardized adversarial robustness benchmark. *arXiv preprint arXiv:2010.09670*, 2020.

[12] Giannis Daras and Alexandros G Dimakis. Discovering the hidden vocabulary of dalle-2. *arXiv preprint arXiv:2206.00169*, 2022.

[13] Josip Djolonga, Jessica Yung, Michael Tschannen, Rob Romijnders, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Matthias Minderer, Alexander D'Amour, Dan I. Moldovan, Sylvan Gelly, Neil Houlsby, Xiaohua Zhai, and Mario Lucic. On robustness and transferability of convolutional neural networks. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16453–16463, 2021.

[14] Xinshuai Dong, Anh Tuan Luu, Rongrong Ji, and Hong Liu. Towards robustness against natural language word substitutions. *ArXiv*, abs/2107.13541, 2021.

[15] Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Nanyun Peng, Zicheng Liu, and Michael Zeng. An empirical study of training end-to-end vision-and-language transformers. *ArXiv*, abs/2111.02387, 2021.

[16] Nathan G. Drenkow, Numair Sani, Ilya Shpitser, and M. Unberath. Robustness in deep learning for computer vision: Mind the gap? *ArXiv*, abs/2112.00639, 2021.

[17] Nick Erickson, Xingjian Shi, James Sharpnack, and Alexander J. Smola. Multimodal automl for image, text and tabular data. *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2022.

[18] Isak Czeresnia Etinger and Alan W. Black. Formality style transfer for noisy, user-generated conversations: Extracting labeled, parallel data from unlabeled corpora. In *EMNLP*, 2019.

[19] Alexander W. Fang, Gabriel Ilharco, Mitchell Wortsman, Yuhao Wan, Vaishaal Shankar, Achal Dave, and Ludwig Schmidt. Data determines distributional robustness in contrastive language image pre-training (clip). 2022.

[20] Stanislav Fort. Pixels still beat text: Attacking the openai clip model with text patches and adversarial pixel perturbations. 2021.

[21] Yuri Galindo and Fabio A. Faria. Understanding clip robustness. 2021.

[22] Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. Large-scale adversarial training for vision-and-language representation learning. *ArXiv*, abs/2006.06195, 2020.

[23] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *ArXiv*, abs/1811.12231, 2019.

[24] Karan Goel, Nazneen Rajani, Jesse Vig, Samson Tan, Jason M. Wu, Stephan Zheng, Caiming Xiong, Mohit Bansal, and Christopher R'e. Robustness gym: Unifying the nlp evaluation landscape. In *NAACL*, 2021.

[25] Gabriel Goh, Nick Cammarata, Chelsea Voss, Shan Carter, Michael Petrov, Ludwig Schubert, Alec Radford, and Chris Olah. Multimodal neurons in artificial neural networks. *Distill*, 6(3):e30, 2021.

[26] Gabriel Goh, Nick Cammarata, Chelsea Voss, Shan Carter, Michael Petrov, Ludwig Schubert, Alec Radford, and Christopher Olah. Multimodal neurons in artificial neural networks. 2021.

[27] Priya Goyal, Quentin Duval, Isaac Seessel, Mathilde Caron, Ishan Misra, Levent Sagun, Armand Joulin, and Piotr Bojanowski. Vision models are more robust and fair when pretrained on uncurated images without supervision. *ArXiv*, abs/2202.08360, 2022.

[28] Tao Gui, Xiao Wang, Qi Zhang, Qin Liu, Yicheng Zou, Xin Zhou, Rui Zheng, Chong Zhang, Qinzhuo Wu, Jiacheng Ye, Zexiong Pang, Yongxin Zhang, Zhengyan Li, Ruotian Ma, Zichu Fei, Ruijian Cai, Jun Zhao, Xinwu Hu, Zhiheng Yan, Yiding Tan, Yuan Hu, Qiyuan Bian, Zhihua Liu, Bolin Zhu, Shan Qin, Xiaoyu Xing, Jinlan Fu, Yue Zhang, Minlong Peng, Xiaoqing Zheng, Yaqian Zhou, Zhongyu Wei, Xipeng Qiu, and Xuanjing Huang. Textflint: Unified multilingual robustness evaluation toolkit for natural language processing. In *ACL*, 2021.

[29] Tanmay Gupta, Ryan Marten, Aniruddha Kembhavi, and Derek Hoiem. Grit: General robust image task benchmark. *ArXiv*, abs/2204.13653, 2022.

[30] William Han, Jielin Qiu, Jiacheng Zhu, Mengdi Xu, Douglas Weber, Bo Li, and Ding Zhao. An empirical exploration of cross-domain alignment between language and electroencephalogram. *ArXiv*, abs/2208.06348, 2022.

[31] Xiaoshuai Hao, Yi Zhu, Srikar Appalaraju, Aston Zhang, Wanqian Zhang, Boyang Li, and Mu Li. Mixgen: A new multi-modal data augmentation. *ArXiv*, abs/2206.08358, 2022.

[32] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Lixuan Zhu, Samyak Parajuli, Mike Guo, Dawn Xiaodong Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8320–8329, 2021.

[33] Dan Hendrycks and Thomas G. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *ArXiv*, abs/1903.12261, 2019.

[34] Dan Hendrycks, Xiaoyuan Liu, Eric Wallace, Adam Dziedzic, Rishabh Krishnan, and Dawn Xiaodong Song. Pretrained transformers improve out-of-distribution robustness. In *ACL*, 2020.

[35] Dan Hendrycks, Norman Mu, Ekin Dogus Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple data processing method to improve robustness and uncertainty. *ArXiv*, abs/1912.02781, 2020.

[36] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Xiaodong Song. Natural adversarial examples. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15257–15266, 2021.

[37] Ronghang Hu, Daniel Fried, Anna Rohrbach, Dan Klein, Trevor Darrell, and Kate Saenko. Are you looking? grounding to multiple modalities in vision-and-language navigation. In *ACL*, 2019.

[38] Xun Huang and Serge J. Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1510–1519, 2017.

[39] Vidhi Jain, Yixin Lin, Eric Undersander, Yonatan Bisk, and Akshara Rai. Transformers are adaptable task planners. *ArXiv*, abs/2207.02442, 2022.

[40] Akbar Karimi, L. Rossi, and Andrea Prati. Aeda: An easier data augmentation technique for text classification. In *EMNLP*, 2021.

[41] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *ICML*, 2021.

[42] Juncen Li, Robin Jia, He He, and Percy Liang. Delete, retrieve, generate: a simple approach to sentiment and style transfer. In *NAACL*, 2018.

[43] Juncheng Billy Li, Shuhui Qu, Xinjian Li, Po-Yao Huang, and Florian Metze. On adversarial robustness of large-scale audio visual learning. In *ICASSP*, 2022.

[44] Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022.

[45] Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq R. Joty, Caiming Xiong, and Steven C. H. Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In *NeurIPS*, 2021.

[46] Linjie Li, Jie Lei, Zhe Gan, and Jingjing Liu. Adversarial vqa: A new benchmark for evaluating the robustness of vqa models. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2022–2031, 2021.

[47] Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. Unimo-2: End-to-end unified vision-language grounded learning. *arXiv preprint arXiv:2203.09067*, 2022.

[48] Xiujun Li, Xi Yin, Chunyuan Li, Xiaowei Hu, Pengchuan Zhang, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *ECCV*, 2020.

[49] Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Foundations and recent trends in multimodal machine learning: Principles, challenges, and open questions. 2022.

[50] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

[51] Wei Liu, Jie-Lin Qiu, Wei-Long Zheng, and Bao-Liang Lu. Comparing recognition performance and robustness of multimodal deep learning models for multimodal emotion recognition. *IEEE Transactions on Cognitive and Developmental Systems*, 14:715–729, 2022.

[52] Ziquan Liu, Yi Xu, Yuanhong Xu, Qi Qian, Hao Li, Rong Jin, Xiangyang Ji, and Antoni B. Chan. An empirical study on distribution shift robustness from the perspective of pre-training and data augmentation. *ArXiv*, abs/2205.12753, 2022.

[53] Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. Unified-io: A unified model for vision, language, and multi-modal tasks. *ArXiv*, abs/2206.08916, 2022.

[54] Edward Ma. Nlp augmentation. 2019.

[55] Kaleel Mahmood, Rigel Mahmood, and Marten van Dijk. On the robustness of vision transformers to adversarial examples. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7818–7827, 2021.

[56] Emanuele La Malfa and Marta Z. Kwiatkowska. The king is naked: on the notion of robustness for natural language processing. In *AAAI*, 2022.

[57] Xiaofeng Mao, Gege Qi, Yuefeng Chen, Xiaodan Li, Ranjie Duan, Shaokai Ye, Yuan He, and Hui Xue. Towards robust vision transformer. *ArXiv*, abs/2105.07926, 2021.

[58] Claudio Michaelis, Benjamin Mitzkus, Robert Geirhos, Evgenia Rusak, Oliver Bringmann, Alexander S. Ecker, Matthias Bethge, and Wieland Brendel. Benchmarking robustness in object detection: Autonomous driving when winter is coming. *arXiv preprint arXiv:1907.07484*, 2019.

[59] So Yeon Min, Devendra Singh Chaplot, Pradeep Ravikumar, Yonatan Bisk, and Ruslan Salakhutdinov. Film: Following instructions in language with modular methods. *ArXiv*, abs/2110.07342, 2022.

[60] Milad Moradi and Matthias Samwald. Evaluating the robustness of neural language models to input perturbations. In *EMNLP*, 2021.

[61] Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. Facebook fair's wmt19 news translation task submission. In *Proc. of WMT*, 2020.

[62] David Noever and Samantha E. Miller Noever. Reading isn't believing: Adversarial attacks on multi-modal neurons. *ArXiv*, abs/2103.10480, 2021.

[63] Sayak Paul and Pin-Yu Chen. Vision transformers are robust learners. In *AAAI*, 2022.

[64] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

[65] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *ArXiv*, abs/2204.06125, 2022.

[66] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *ICML*, 2019.

[67] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019.

[68] Barbara Rychalska, Dominika Basaj, Alicja Gosiewska, and P. Biecek. Models in the wild: On corruption robustness of neural nlp systems. In *ICONIP*, 2019.

[69] Madeline Chantry Schiappa, Yogesh Singh Rawat, Shruti Vyas, Vibhav Vineet, and Hamid Palangi. Multi-modal robustness analysis against language and visual perturbations. *ArXiv*, abs/2207.02159, 2022.

[70] Robert Schmidt. Generative text style transfer for improved language sophistication. 2020.

[71] Christoph Schuhmann, Romain Beaumont, Richard Vencu andCade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models. 2022.

[72] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *ArXiv*, abs/2111.02114, 2021.

[73] Rahul Singh, Karan Jindal, Yufei Yu, Hanyu Yang, Tarun Joshi, Matthew A. Campbell, and Wayne B. Shoumaker. Robustness tests of nlp machine learning models: Search and semantically replace. *ArXiv*, abs/2104.09978, 2021.

[74] Alane Suhr, Mike Lewis, James Yeh, and Yoav Artzi. A corpus of natural language for visual reasoning. In *ACL*, 2017.

[75] Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. *ArXiv*, abs/2007.00644, 2020.

[76] Yapeng Tian and Chenliang Xu. Can audio-visual integration strengthen robustness under multimodal attacks? *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5597–5607, 2021.

[77] Nishant Vishwamitra, Hongxin Hu, Ziming Zhao, Long Cheng, and Feng Luo. Understanding and measuring robustness of multimodal learning. *ArXiv*, abs/2112.12792, 2021.

[78] Boxin Wang, Chejian Xu, Shuohang Wang, Zhe Gan, Yu Cheng, Jianfeng Gao, Ahmed Hassan Awadallah, and Bo Li. Adversarial glue: A multi-task benchmark for robustness evaluation of language models. *ArXiv*, abs/2111.02840, 2021.

[79] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *ICML*, 2022.

[80] Tianlu Wang, Xuezhi Wang, Yao Qin, Ben Packer, Kang Li, Jilin Chen, Alex Beutel, and Ed H. Chi. Cat-gen: Improving robustness in nlp models via controlled adversarial text generation. In *EMNLP*, 2020.

[81] Xuezhi Wang, Haohan Wang, and Diyi Yang. Measure and improve robustness in nlp models: A survey. *ArXiv*, abs/2112.08313, 2022.

[82] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. *ArXiv*, abs/2108.10904, 2022.

[83] Jason Wei and Kai Zou. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. In *EMNLP*, 2019.

[84] F. Wenzel, Andrea Dittadi, Peter V. Gehler, Carl-Johann Simon-Gabriel, Max Horn, Dominik Zietlow, David Kernert, Chris Russell, Thomas Brox, Bernt Schiele, Bernhard Scholkopf, and Francesco Locatello. Assaying out-of-distribution generalization in transfer learning. *ArXiv*, abs/2207.09239, 2022.

[85] Hao Wu, Jiayuan Mao, Yufeng Zhang, Yuning Jiang, Lei Li, Weiwei Sun, and Wei-Ying Ma. Unified visual-semantic embeddings: Bridging vision and language with structured meaning representations. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6602–6611, 2019.

[86] Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. Visual entailment task for visually-grounded language learning. *arXiv preprint arXiv:1811.10582*, 2018.

[87] 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.

[88] Jinyu Yang, Jiali Duan, S. Tran, Yi Xu, Sampath Chanda, Liqun Chen, Belinda Zeng, Trishul M. Chilimbi, and Junzhou Huang. Vision-language pre-training with triple contrastive learning. *ArXiv*, abs/2202.10401, 2022.

[89] Dong Yin, Raphael Gontijo Lopes, Jonathon Shlens, Ekin Dogus Cubuk, and Justin Gilmer. A fourier perspective on model robustness in computer vision. In *NeurIPS*, 2019.

[90] Peter Young, Alice Lai, Micah Hodosh, and J. Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.

[91] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *ArXiv*, abs/2205.01917, 2022.

[92] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel C. F. Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, Ce Liu, Mengchen Liu, Zicheng Liu, Yumao Lu, Yu Shi, Lijuan Wang, Jianfeng Wang, Bin Xiao, Zhen Xiao, Jianwei Yang, Michael Zeng, Luowei Zhou, and Pengchuan Zhang. Florence: A new foundation model for computer vision. *ArXiv*, abs/2111.11432, 2021.

[93] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1204–1213, 2022.

[94] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5575–5584, 2021.

[95] Stephan Zheng, Yang Song, Thomas Leung, and Ian J. Goodfellow. Improving the robustness of deep neural networks via stability training. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4480–4488, 2016.

[96] Daquan Zhou, Zhiding Yu, Enze Xie, Chaowei Xiao, Anima Anandkumar, Jiashi Feng, and José Manuel Álvarez. Understanding the robustness in vision transformers. In *ICML*, 2022.

[97] Mohammadreza Zolfaghari, Yi Zhu, Peter Gehler, and Thomas Brox. Crossclr: Cross-modal contrastive learning for multi-modal video representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1450–1459, October 2021.

5 Appendix

5.1 Multimodal Robustness Benchmark under Distribution Shift

Distribution shift is one of the significant problems of applying models in real-world scenarios [75, 52]. It is caused by scarcity of data, i.e., the models cannot be trained on all possible data in $p(\boldsymbol{x}, \boldsymbol{y})$, where $p(\boldsymbol{x}, \boldsymbol{y})$ is considered as the real-world data distribution. In other words, the training set contains the collected data that fits a certain distribution $p_{tr}(\boldsymbol{x} \mid \boldsymbol{y})$, but the test set usually has a different distribution $p_{te}(\boldsymbol{x} \mid \boldsymbol{y}) \neq p_{tr}(\boldsymbol{x} \mid \boldsymbol{y})$.

5.2 Image Perturbation

In Table 7 and Table 8, we show more details about the image and text perturbations, respectively.

Table 7: Image perturbations.

Category	Perturbation	Description	Severities
Noise	Gaussian Noise	Gaussian noise can appear in low-lighting conditions.	5
	Shot Noise	Shot noise, also called Poisson noise, is electronic noise caused by the discrete nature of light itself.	5
	Impulse Noise	Impulse noise is a color analogue of salt-and-pepper noise and can be caused by bit errors.	5
	Speckle Noise	Speckle noise is the noise added to a pixel tends to be larger if the original pixel intensity is larger.	5
Blur	Defocus Blur	Defocus blur occurs when an image is out of focus.	5
	Frosted Glass Blur	Frosted Glass Blur appears with “frosted glass” windows or panels.	5
	Motion Blur	Motion blur appears when a camera is moving quickly.	5
	Zoom Blur	Zoom blur occurs when a camera moves toward an object rapidly.	5
Weather	Snow	Snow is a visually obstructive form of precipitation.	5
	Frost	Frost forms when lenses or windows are coated with ice crystals.	5
	Fog	Fog shrouds objects and is rendered with the diamond-square algorithm.	5
	Brightness	Brightness varies with daylight intensity.	5
Digital	Contrast	Contrast can be high or low depending on lighting conditions and the photographed object’s color.	5
	Elastic	Elastic transformations stretch or contract small image regions.	5
	Pixelate	Pixelation occurs when upsampling a low-resolution image.	5
	JPEG Compression	JPEG is a lossy image compression format which introduces compression artifacts.	5
Stylized	Stylize	Stylized data is generated by transferring the style information to the content images by AdaIN style transfer [38].	5
Sum	**17**	—	**85**

5.3 Text Perturbation

Fidelity To build a convincing benchmark, we need to ensure the perturbed text remains the same semantics as the original one. Otherwise, for image-text pairs in multimodal learning, the perturbed text won’t be a matching pair to the original image. In this work, we use paraphrases from pretrained sentence-transformers [67] to evaluate the semantic similarity between the original and perturbed sentences. Specifically, “paraphrase-mpnet-base-v2” is used to extract the original and perturbed sentence embeddings for computing similarity score α_s. Given a predefined tolerance threshold α_0, a higher score $\alpha_s > \alpha_0$ means the perturbed text still has similar semantics. However, if $\alpha_s < \alpha_0$ means their semantics are different, we will perturb the sentence again until the semantic similarity score meets the requirement, in a reasonable looping time. For example, we set number of loops to

Table 8: Text perturbations.

Category	Perturbation	Description	Severities
Character-level	Keyboard	Substitute character by keyboard distance.	5
	OCR	Substitute character by pre-defined OCR error.	5
	Character Insert (CI)	Insert character randomly with probability p.	5
	Character Replace (CR)	Substitute character randomly with probability p.	5
	Character Swap (CS)	Swap character randomly with probability p.	5
	Character Delete (CD)	Delete character randomly with probability p.	5
Word-level	Synonym Replacement (SR)	Randomly choose n words from the sentence that are not stop words. Replace each of these words with one of its synonyms chosen at random.	5
	Word Insertion (WI)	Find a random synonym of a random word in the sentence that is not a stop word. Insert that synonym into a random position in the sentence. Do this n times.	5
	Word Swap (WS)	Randomly choose two words in the sentence and swap their positions. Do this n times.	5
	Word Deletion (WD)	Each word in the sentence can be randomly removed with probability p.	5
	Insert Punctuation (IP)	Random insert punctuation in the sentence with probability p.	5
Sentence-level	Formal	Transfer the text style to Formal.	1
	Casual	Transfer the text style to Casual.	1
	Passive	Transfer the text style to Passive.	1
	Active	Transfer the text style to Active.	1
	Back Translation	Translate source to German and translating it back to English via [61].	1
Sum	**16**	—	**60**

$N_{max} = 100$. Beyond N_{max}, we will just remove this text sample from our robustness benchmark. This procedure guarantees semantic closeness and ensures our benchmark is valid for evaluation.

5.4 Examples of Image Perturbation

Examples of perturbed images from COCO dataset [50] are shown in Figure 2.

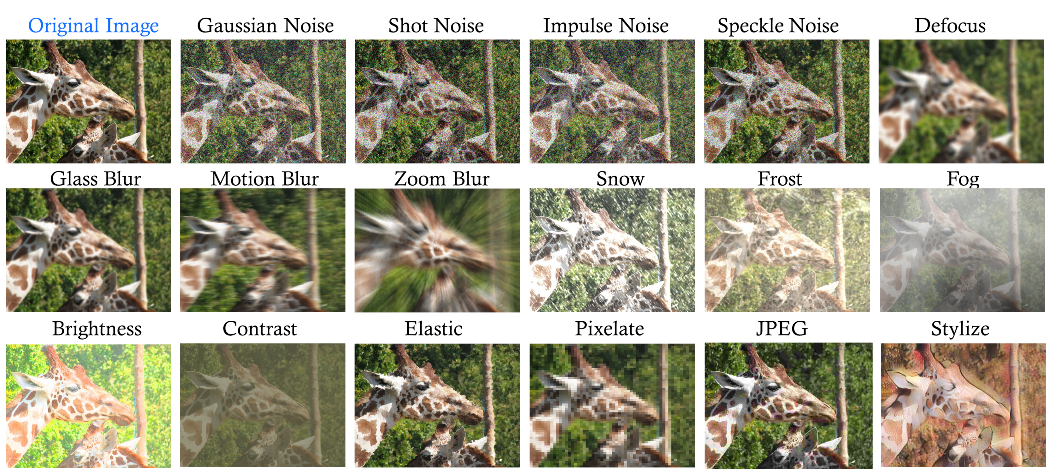

Figure 2: Examples of our 17 image perturbation strategies applied to the COCO dataset. The original image is shown on the top left.

5.5 Examples of Text Perturbation

Examples of the text perturbation of captions in Flickr30K dataset [90] are shown in Table 9.

Table 9: Example of text perturbation on Flickr30K text.

Category	Perturbation	Example
Original	Clean	An orange metal bowl strainer filled with apples.
Character	Keyboard	An orange metal bowk strainer filled witj apples.
	OCR	An 0range metal bowl strainer filled with app1es.
	CI	And orange metal bowl strainer filled with atpples.
	CR	An orange metal towl strainer fillet with apples.
	CS	An orange meatl bowl stariner filled with apples.
	CD	An orang[X] metal bowl strainer fil[X]ed with apples.
Word	SR	An orange alloy bowl strainer filled with apples.
	WI	An old orange metal bowl strainer filled with apples.
	WS	An orange metal strainer bowl filled with apples.
	WD	An orange metal bowl strainer [X] with apples.
	IP	An orange metal bowl ? strainer filled with apples.
Sentence	Formal	An orange metal bowl strainer contains apples.
	Casual	An orange metal bowl is filled with apples.
	Passive	Some apples are in an orange metal bowl strainer.
	Active	There are apples in an orange metal bowl strainer.
	Back trans	Apples are placed in an orange metal bowl strainer.

5.6 More Experimental Setting

By building the robustness benchmark datasets, we would like to answer the following questions: **(1)** How robust can multimodal pretrained image-text models be under distribution shift? **(2)** What is the sensitivity of each model under different perturbation methods? **(3)** Which model architecture or loss objectives might be more robust under image or text perturbations? **(4)** Are there any particular image/text perturbation methods that can consistently show significant influence?

Evaluation Tasks We select three widely adopted downstream tasks for a comprehensive evaluation on the robustness of multimodal image-text models, including image-text retrieval, visual reasoning (VR), and visual entailment (VE). Image-text retrieval includes two subtasks: (1) retrieve images with given text (Image Retrieval) and (2) retrieve text with given images (Text Retrieval) [6, 31]. Visual Reasoning (VR) requires the model to determine whether a textual statement describes a pair of images. [74]. Visual Entailment (VE) is a visual reasoning task to predict whether the relationship between an image and text is entailment, neutral, or contradictory [86, 87].

Evaluation Models We select six representative large pretrained multimodal models, which have publicly released their pretrained models[2], including CLIP [64], ViLT [41], ALBEF [45], BLIP [44], TCL [88], and METER [15]. In order to provide a fair comparison, we adopt the model weights

[2]We appreciate all the authors for making the models publicly available

provided by their official repositories[3] for either zero-shot prediction or fine-tuned results. We only perform the tasks of each model that have been studied in its original work, where their reported scores are marked as "clean" in our Tables.

Evaluation Metric We adopt standard evaluation metrics for each task. To be specific, for image-text retrieval, we use recall and RSUM [85]. Here, recall is defined as K (R@K) metric, where K = $\{1, 5, 10\}$, and RSUM is defined as the sum of recall metrics at K = $\{1, 5, 10\}$ of both image and text retrieval tasks. As for visual reasoning and visual entailment tasks, we use prediction accuracy as the evaluation metric.

However, there is no appropriate metric that could be used for robustness evaluation under distribution shift. Inspired by an example in [75], given a clean dataset d_1 and its perturbed dataset d_2, model m_1 should be considered more robust than model m_2 if m_1's performance drop is less significant than m_2 from d_1 to d_2, even though m_2's absolute accuracy/recall on d_2 is higher than m_1's. Thus we believe robustness should be evaluated relatively when there are distribution shifts. To qualitatively analyze the multimodal image-text models, we introduce a new evaluation metric, termed MultiModal Impact score (MMI). We compute MMI as the averaged performance drop compared with the non-perturbed performance ("clean"), i.e., MMI $= (s_c - s_p)/s_c$ where s_p is the perturbed score and s_c is the clean score. In the following experiments, we report both standard performance scores, i.e., recall, RSUM, accuracy, as well as our MMI.

Our Benchmark Datasets For each task, we perturb the corresponding datasets i.e., Flickr30K [90] and COCO [50] , NLVR2 [74], SNLI-VE [86, 87], using the image perturbation (IP) and text perturbation (TP) methods introduced in Section 2. This leads to our 8 benchmark datasets: (1) Flickr30K-IP, Flickr30K-TP, COCO-IP, and COCO-TP for image-text retrieval robustness evaluation; (2) NLVR2-IP and NLVR2-TP for visual reasoning robustness evaluation; and (3) SNLI-VE-IP and SNLI-VE-TP for visual entailment robustness evaluation.

- For image-text retrieval, the Flickr30K dataset contains 1,000 images, and each of them has 5 corresponding captions, while the COCO dataset contains 5,000 images, and each of them also has 5 corresponding captions. We report the RSUM score averaged on five perturbation levels under each perturbation method to reveal the overall performance. More detailed results, including the recall at K (R@K) metric, K = $\{1, 5, 10\}$, can be found in the Appendix 5.9. For CLIP and TCL, we provide the evaluation results for both zero-shot (ZS) and fine-tuned (FT) settings, while for ALBEF and BLIP, we follow their original settings and report the fine-tuned (FT) results.
- For visual reasoning, the NLVR2 dev set contains 2,018 unique sentences and 6,982 samples, while the test-P set contains 1,995 unique sentences and 6,967 samples. We report the accuracy of both the dev set and test-P set of the NLVR2 dataset under image and text perturbations. We evaluate the robustness of ALBEF, ViLT, TCL, BLIP, and METER.
- For visual entailment, the SNLI-VE val set contains 1,000 images and 6,576 sentences, while the test set contains 1,000 images and 6,592 sentences. We evaluate the accuracy of both the dev set and test set of the SNLI-VE dataset under image and text perturbations. We report the results of ALBEF, TCL, and METER.

5.7 Experimental Results and Discussion

Image-text Retrieval Results and Observation The averaged RSUM results of different methods under five perturbation levels are shown in Table 1 and Table 2, for image perturbation and text perturbation, respectively. Through Table 1, we found that all the models' performance dropped under image perturbation. According to the impact score, overall, the CLIP model is, in general, more robust than other models, which we hypothesized might be due to the large datasets that CLIP was trained upon, where large data may lead to better performance, as also noted by [75]. Due to the generative loss objective, the BLIP model also shows good robustness performance. We think the generative loss objective can help to learn better data distribution, and we observe a recent paradigm change from using discriminative contrastive loss, i.e., CLIP, ALBEF [64, 45], to

[3] https://github.com/openai/CLIP, https://github.com/dandelin/ViLT, https://github.com/salesforce/ALBEF, https://github.com/salesforce/BLIP, https://github.com/uta-smile/TCL, https://github.com/zdou0830/meter

using generative loss, i.e., BLIP, CoCa, SimVLM, PaLI, Unified-IO, OFA [44, 91, 82, 8, 53, 79]. In detail, we also found that different image perturbation methods have different impact levels on the model's performance, and the methods that have the biggest impact also vary among different models and datasets. CLIP-FT, TCL-ZS, ALBEF, and BLIP seem to be more sensitive to Zoom Blur perturbation, while ViLT and TCL-FT are more sensitive to pixelation perturbation. Glass blur and brightness are the two "soft" perturbation methods, where all the models are very robust under these settings. Besides, fine-tuning may also help to improve robustness, i.e., TCL-FT shows robustness improvements compared with TCL-ZS on both Flickr30K and COCO datasets. As in Table 2, we found that all the models' performance dropped under text perturbation. BLIP overall shows the best robustness performance, which may bring the idea that the generative loss objective is useful. In addition, we found that character-level perturbations show much more influence than word-level and sentence-level perturbations, especially Keyboard and CR (Character Replace) methods consistently show high impact in attacking the model's performance. IP (Insert Punctuation), Formal, and Active are the three least effective text perturbation methods across different models.

Visual Reasoning Results and Observation The averaged accuracy results of different methods under five perturbation levels are shown in Table 3 and Table 4, for image perturbation and text perturbation, respectively. From Table 3, we can find that all the models' performance dropped under image perturbation, especially ALBEF. TCL shows better performance than ALBEF, where TCL introduced an intra-modal contrastive objective based on the ALBEF architecture, which may be helpful in improving the model's performance. In detail, Zoom Blur consistently shows the most effective impact on attacking all the models' performance for both the dev set and test-P set. In contrast, Glass Blur seems to be one of the least effective perturbation methods, while Gaussian Noise, Defocus Blur, Fog, and JPEG Compression can also be not effective in attacking the model's performance. Besides, as shown in Table 4, all the models' performance also drooped under text perturbation. In detail, character-level perturbation still shows a much stronger influence than word-level and sentence-level perturbations for the visual reasoning task, and different models seem to be sensitive to different character-level perturbations. The sensitivities of different models also vary, where Keyboard, OCR, CI, CR, CS, and CD show different impacts. However, IP (Insert Punctuation) seems to be one of the least effective ones in attacking in the visual reasoning task, while SR, Formal, Active, Back_trans can also be stable methods in different evaluation models.

Visual Entailment Results and Observation The averaged accuracy results of different methods under five perturbation levels are shown in Table 5 and Table 6, for image perturbation and text perturbation, respectively. Similar to the results in image-text retrieval and the visual reasoning tasks, the performance of all the models dropped under both image perturbation and text perturbation settings. In detail, Zoom Blur still serves as the most powerful image perturbation method, and Brightness is the least effective one, as shown in Table 5. In addition, as shown in Table 6, character-level perturbation also shows a much stronger influence than word-level and sentence-level perturbations for the visual entailment task, where IP, Formal, Casual, Passive, and Active can be stable perturbation strategies. Unlike the results in image-text retrieval and the visual reasoning tasks, the performance drop seems insignificant in the VE task, which may be due to VE being a relatively easy task, so different model variations are not shown explicitly.

Limitation Given our work is one of the early efforts in this direction, there are several limitations and promising future work. First, we only adopt synthetic image and text perturbation strategies in our benchmark. However, there are other perturbation methods that could be explored for further robustness evaluation, e.g., real distribution shift [75, 84]. Second, we only study three downstream tasks, while there are more interesting ones, such as visual question answering, image captioning and text-to-image generation. For those generation tasks, new evaluation metrics might be needed to properly evaluate the model's robustness. Third, we only evaluate these image-text models but the more important question is, how to improve their robustness. Data augmentation is a common technique to improve unimodal models' robustness [35, 32], which we could also explore for multimodal setting [31].

5.8 Related Work

Robustness of unimodal vision models is a longstanding and challenging goal of computer vision [89]. Stable training, adversarial robustness, out-of-distribution and transfer performance, and many other aspects have been studied by previous works in deep learning era [95, 16, 13, 27]. Recently, Vision Transformer (ViT) has shown improved robustness compared with previous models, i.e., the comparison between ViT and ResNet for robustness against common corruptions and perturbations [3], robustness under distribution shifts and natural adversarial setting [63], robustness against different Lp-based adversarial attacks [55], adversarial examples [57], and adaptive attacks [2]. In terms of benchmark, [33] proposed ImageNet-C and ImageNet-P benchmarks for classifier's robustness to common perturbations. [36] proposed ImageNet-A and ImageNet-O benchmarks for adversarial filtration and out-of-distribution detection. [66] proposed ImageNet-V2 for evaluating distribution shift. [23] proposed Stylized-ImageNet by removing local texture cues in ImageNet while retaining global shape information on natural images via AdaIN style transfer. Recently, [29] built the GRIT benchmark to evaluate the performance, robustness, and calibration of a vision system across a variety of image tasks, concepts, and data sources.

Robustness of unimodal language models under distribution shift or adversarial attack has also been explored by many previous works, i.e., [7, 81] provided reviews of how to define, measure and improve robustness of NLP systems, [80] proposed controlled adversarial text generation to improve robustness, [24] unified four standard evaluation paradigms, [73] proposed a search and semantically replace strategy, [14] studied robustness against word substitutions, [56] formalised the concept of semantic robustness, etc. In terms of benchmark, [34] systematically examined and measured the out-of-distribution (OOD) generalization for seven NLP datasets. [11] built a large benchmark and analyzed the impact of robustness on the performance of distribution shifts, calibration, out-of-distribution detection, fairness, privacy leakage, smoothness, and transferability. Recently, [60] presented empirical results achieved with a comprehensive set of non-adversarial perturbation methods for testing the robustness of NLP systems on non-synthetic text. [28] proposed a multilingual robustness evaluation platform that incorporates universal text transformation, task-specific transformation, adversarial attack, and subpopulation to provide comprehensive robustness analysis. [78] proposed a benchmark to evaluate the vulnerabilities of modern large-scale language models under adversarial attacks.

Multimodal learning has advanced quickly in recent years with appealing applications in different fields [49, 39, 59, 4, 37, 97, 17, 30, 51], i.e., embodied autonomous agents, image/video understanding, multimedia and healthcare. Thanks to the larger datasets [64, 92, 72, 71] and larger transformer models [93, 8, 5, 10], many powerful multimodal image-text models have been developed and shown great capability. However, unlike unimodal models, the robustness study of these multimodal models under distribution shift has rarely been explored.

Robustness of multimodal models is essential to study before deploying these amazing foundation models to real applications. Previous works [21, 20, 26, 62] have unsystematically tested some pre-trained models, i.e., CLIP [64], by attacking with text patches and adversarial pixel perturbations. [77] measured the robustness of multimodal learning by fusing the input modalities and adversarial attack. [19] found that diverse training distribution is the main cause for robustness gains. [76, 43] investigated the audio-visual model robustness under multimodal attacks. For benchmarks, [46] collected an Adversarial VQA dataset to evaluate the robustness of VQA models. A concurrent work [69] studied the robustness of video-text models under perturbations, but their models, tasks, and datasets are different from ours. In this work, we focus on studying robustness under distribution shifts for multimodal image-text models. We introduce new datasets and metrics, and extensively evaluate recent multimodal models.

5.9 Detailed Image-Text Retrieval Results

In the appendix, we provide the detailed robustness evaluation results for the image-text retrieval task, where the evaluation datasets are Flickr30K and COCO. In the following tables, we report the recall at K (R@K) metric, K = $\{1, 5, 10\}$, where $Mean$ is the averaged recall results for either text retrieval or image retrieval, RSUM is defined as the sum of recall metrics at K = $\{1, 5, 10\}$ of both image and text retrieval tasks.

5.9.1 Image Perturbations

Table 10: CLIP image perturbation performance comparison of Zero-Shot (ZS) image-text retrieval on Flickr30K and COCO datasets (results are averaged on five perturbation levels).

	Method	Flickr30K (1K)									MSCOCO (5K)								
		Text Retrieval				Image Retrieval					Text Retrieval				Image Retrieval				
		R@1	R@5	R@10	Mean	R@1	R@5	R@10	Mean	RSUM	R@1	R@5	R@10	Mean	R@1	R@5	R@10	Mean	RSUM
Noise	Gaussian	75.1	92.8	96.0	88.0	61.7	85.1	90.9	79.3	501.7	47.8	72.1	80.6	66.9	34.7	58.7	69.1	54.2	363.0
	Shot	75.6	93.4	96.6	88.5	61.7	85.5	91.4	79.5	504.2	47.6	71.6	80.3	66.5	34.2	58.5	69.1	53.9	361.2
	Impluse	68.2	90.2	94.3	84.2	57.4	82.1	88.9	76.2	481.2	40.1	65.6	75.4	60.4	30.1	54.1	64.8	49.7	330.2
	Speckle	80.2	95.8	98.0	91.3	62.9	86.4	92.2	80.5	515.5	49.5	73.9	82.0	68.5	34.6	59.1	69.6	54.4	368.7
Blue	Defocus	74.7	93.4	96.6	88.2	61.3	85.1	91.1	79.1	502.1	46.5	71.3	80.0	65.9	33.7	58.3	68.8	53.6	358.6
	Glass	85.5	97.8	99.0	94.1	66.1	88.4	93.4	82.6	530.1	55.6	78.9	86.4	73.6	37.3	61.7	71.7	56.9	391.6
	Motion	77.0	94.1	97.0	89.4	63.5	86.2	91.9	80.6	509.7	48.8	72.3	80.4	67.1	34.2	58.2	68.3	53.6	362.2
	Zoom	62.3	84.6	90.6	79.1	54.8	79.2	86.3	73.5	457.8	32.4	57.0	67.2	52.2	26.9	50.1	61.0	46.0	294.6
Weather	Snow	64.8	86.9	93.1	81.6	56.2	81.4	88.3	75.3	470.7	32.3	56.2	67.8	52.1	26.8	50.1	61.4	46.1	294.7
	Frost	72.8	92.6	96.5	87.3	59.4	84.0	90.4	77.9	495.6	41.1	65.6	75.6	60.8	29.4	53.2	64.1	48.9	329.0
	Fog	80.8	96.1	98.2	91.7	64.6	87.3	92.7	81.5	519.7	51.3	75.5	83.6	70.2	34.0	58.5	68.8	53.8	371.8
	Brightness	85.2	97.6	98.9	93.9	66.4	88.6	93.4	82.8	530.1	56.5	79.8	87.4	74.6	36.4	60.7	71.1	56.0	391.9
Digital	Contrast	80.7	95.9	98.0	91.5	62.7	86.2	91.9	80.3	515.4	48.0	71.5	80.1	66.5	32.5	56.9	67.4	52.2	356.4
	Elastic	79.5	94.9	97.3	90.6	61.6	85.8	91.4	79.6	510.4	50.6	74.7	83.1	69.5	33.8	58.5	69.1	53.8	369.7
	Pixelate	68.4	87.6	92.0	82.7	55.5	79.6	86.4	73.8	469.5	36.3	60.4	70.3	55.7	27.9	51.3	61.9	47.0	308.2
	JPEG	83.6	96.8	98.4	92.9	65.8	87.4	92.7	82.0	524.6	55.3	78.9	86.4	73.5	35.9	60.7	70.9	55.8	388.0
Stylize	Stylized	65.3	83.3	88.3	79.0	51.6	75.8	83.2	70.2	447.6	39.9	62.8	72.2	58.3	28.0	50.8	61.2	46.7	314.9

Table 11: CLIP image perturbation performance comparison of Fine-tuned (FT) image-text retrieval on Flickr30K and COCO datasets (results are averaged on five perturbation levels).

	Method	Flickr30K (1K)									MSCOCO (5K)								
		Text Retrieval				Image Retrieval					Text Retrieval				Image Retrieval				
		R@1	R@5	R@10	Mean	R@1	R@5	R@10	Mean	RSUM	R@1	R@5	R@10	Mean	R@1	R@5	R@10	Mean	RSUM
Noise	Gaussian	72.7	91.2	95.0	86.3	63.1	86.5	91.6	80.4	500.1	43.0	70.3	80.1	64.5	35.1	63.5	75.1	57.9	367.2
	Shot	73.0	91.9	95.8	86.9	63.9	87.1	92.1	81.0	503.8	42.4	69.9	79.9	64.1	34.9	63.3	74.9	57.7	365.3
	Impluse	65.1	87.9	92.5	81.8	59.2	84.3	90.1	77.9	479.2	35.6	63.0	74.3	57.6	29.8	58.3	70.7	53.0	331.7
	Speckle	78.1	95.0	97.8	90.3	66.9	89.9	94.4	83.7	522.1	36.5	65.7	77.1	59.8	36.5	65.7	77.1	59.8	381.5
Blue	Defocus	70.1	90.2	94.5	84.9	61.6	85.6	91.4	79.5	493.4	43.7	71.7	81.5	65.6	35.2	63.8	75.2	58.1	371.0
	Glass	82.3	97.1	99.1	92.9	70.6	91.9	95.8	86.1	536.9	52.3	80.1	88.5	73.7	40.8	69.9	80.6	63.8	412.2
	Motion	76.1	93.7	96.8	88.9	65.0	88.4	93.3	82.2	513.3	44.6	71.7	81.0	65.8	36.4	64.9	75.8	59.1	374.4
	Zoom	58.7	80.9	87.8	75.8	53.0	78.5	85.5	72.3	444.3	28.4	54.1	65.1	49.2	26.6	52.3	64.4	47.8	291.0
Weather	Snow	69.6	91.3	95.7	85.5	64.2	88.8	93.4	82.1	503.0	26.6	51.7	63.9	47.4	26.4	54.0	66.6	49.0	289.3
	Frost	81.7	97.0	98.9	92.5	69.1	90.9	95.0	85.0	532.5	37.3	65.2	75.8	59.4	30.3	58.4	70.4	53.0	337.3
	Fog	80.5	95.9	98.3	91.6	69.0	90.8	95.2	85.0	529.7	47.0	75.3	84.6	69.0	37.7	67.0	78.2	61.0	389.9
	Brightness	85.9	97.8	99.3	94.3	72.3	92.3	96.1	86.9	543.7	52.8	80.1	88.4	73.8	41.2	70.4	80.9	64.2	413.9
Digital	Contrast	78.1	94.9	97.5	90.2	66.9	89.8	94.3	83.6	521.5	43.4	71.6	81.5	65.5	35.6	64.1	75.5	58.4	371.7
	Elastic	76.9	93.8	96.9	89.2	65.4	88.0	92.9	82.1	513.9	45.8	73.6	82.8	67.4	36.2	65.0	76.3	59.1	379.7
	Pixelate	62.5	83.9	88.8	78.4	54.4	78.6	85.5	72.8	453.8	32.4	58.3	68.9	53.2	27.3	53.8	65.7	48.9	306.4
	JPEG	81.5	96.2	98.3	92.0	68.2	90.1	94.2	84.2	528.5	50.4	78.1	86.8	71.8	39.2	68.2	79.4	62.3	402.1
Stylize	Stylized	59.9	80.8	86.5	75.7	51.3	76.0	82.6	70.0	437.0	33.3	59.1	69.3	53.9	28.1	54.5	65.9	49.5	310.2

Table 12: BLIP image perturbation performance comparison of Fine-tuned (FT) image-text retrieval on Flickr30K and COCO datasets (results are averaged on five perturbation levels).

	Method	Flickr30K (1K)									MSCOCO (5K)								
		Text Retrieval				Image Retrieval					Text Retrieval				Image Retrieval				
		R@1	R@5	R@10	Mean	R@1	R@5	R@10	Mean	RSUM	R@1	R@5	R@10	Mean	R@1	R@5	R@10	Mean	RSUM
Noise	Gaussian	85.1	94.9	96.4	92.1	74.3	91.1	94.4	86.6	536.2	70.1	88.4	92.8	83.8	55.2	79.0	86.4	73.5	471.9
	Shot	85.4	95.0	96.8	92.4	75.1	91.6	95.0	87.3	538.9	70.1	88.2	92.8	83.7	55.2	79.2	86.5	73.7	472.1
	Impluse	83.3	93.4	95.7	90.8	72.9	89.9	93.5	85.4	528.6	68.7	87.6	92.3	82.9	54.5	78.6	86.1	73.1	467.7
	Speckle	91.3	98.2	99.1	96.2	80.2	94.8	97.2	90.7	560.8	74.4	91.5	95.0	87.0	58.4	81.6	88.5	76.2	489.5
Blue	Defocus	83.8	93.9	96.0	91.2	73.1	89.5	93.2	85.3	529.4	68.0	87.5	92.2	82.6	54.6	78.3	85.4	72.8	466.1
	Glass	94.6	99.6	99.8	98.0	83.4	96.1	98.0	92.5	571.6	79.1	94.3	97.2	90.2	62.0	84.3	90.3	78.9	507.2
	Motion	82.6	93.4	96.0	90.7	71.9	88.9	92.9	84.6	525.7	65.8	85.0	89.8	80.2	52.9	75.6	82.5	70.3	451.7
	Zoom	56.2	74.9	80.4	70.5	53.3	74.7	81.6	69.9	421.1	30.7	52.2	61.0	48.0	31.8	53.4	62.5	49.2	291.6
Weather	Snow	62.2	82.7	88.8	77.9	56.7	79.7	86.5	74.3	456.6	58.3	80.5	87.1	75.3	49.7	74.5	82.8	69.0	432.8
	Frost	79.1	93.0	96.1	89.4	66.4	86.8	91.9	81.7	513.4	69.2	88.0	92.7	83.3	55.7	79.5	86.7	74.0	471.8
	Fog	92.9	99.2	99.6	97.2	82.8	96.0	98.0	92.3	568.5	74.7	91.7	95.4	87.2	60.1	82.9	89.4	77.5	494.2
	Brightness	95.6	99.6	99.8	98.3	84.8	96.5	98.3	93.2	574.5	79.1	94.0	96.8	90.0	61.9	84.4	90.5	78.9	506.8
Digital	Contrast	90.2	97.5	98.4	95.4	79.4	93.5	96.1	89.7	555.1	69.5	87.6	92.1	83.1	56.1	79.1	86.1	73.8	470.4
	Elastic	87.3	95.4	96.8	93.2	77.5	92.8	95.7	88.7	545.6	70.4	87.9	92.4	83.6	55.9	79.3	86.4	73.9	472.3
	Pixelate	75.6	88.2	91.5	85.1	64.7	83.0	87.8	78.5	490.8	56.1	76.3	82.6	71.6	44.9	68.3	76.5	63.3	404.7
	JPEG	92.7	98.5	99.3	96.8	81.2	94.9	97.2	91.1	563.8	77.5	93.2	96.4	89.1	60.1	83.0	89.5	77.5	499.6
Stylize	Stylized	73.3	86.4	89.3	83.0	64.1	82.1	87.0	77.7	482.1	55.1	75.3	81.6	70.7	45.9	68.6	76.5	63.6	402.9

Table 13: ALBEF image perturbation performance comparison of Fine-tuned (FT) image-text retrieval on Flickr30K and COCO datasets (results are averaged on five perturbation levels).

	Method	Flickr30K (1K)									MSCOCO (5K)								
		Text Retrieval				Image Retrieval					Text Retrieval				Image Retrieval				
		R@1	R@5	R@10	Mean	R@1	R@5	R@10	Mean	RSUM	R@1	R@5	R@10	Mean	R@1	R@5	R@10	Mean	RSUM
Noise	Gaussian	83.9	94.6	96.5	91.7	73.4	90.9	94.5	86.3	533.8	66.1	86.5	92.0	81.5	52.1	77.6	85.7	71.8	460.0
	Shot	84.9	95.2	97.1	92.4	74.0	91.8	95.2	87.0	538.3	66.2	86.6	92.0	81.6	52.1	77.9	85.8	71.9	460.6
	Impluse	83.7	94.4	96.3	91.5	73.0	90.5	94.1	85.9	532.0	66.0	86.8	92.1	81.6	52.1	77.6	85.7	71.8	460.3
	Speckle	90.1	98.1	99.1	95.8	78.8	94.6	97.2	90.2	557.8	69.9	89.3	94.1	84.4	54.7	80.1	87.6	74.1	475.8
Blue	Defocus	82.6	94.0	96.5	91.1	71.8	90.2	93.6	85.2	528.8	62.6	84.1	90.1	79.0	50.6	75.7	83.9	70.1	447.1
	Glass	93.8	99.2	99.7	97.6	82.3	96.3	97.9	92.1	569.2	75.1	92.1	96.2	87.8	58.1	82.2	89.2	76.5	493.0
	Motion	80.0	92.0	94.2	88.7	69.3	88.2	92.3	83.3	516.0	61.6	82.4	87.9	77.3	49.3	73.8	81.5	68.2	436.5
	Zoom	56.0	73.8	79.4	69.7	52.6	73.8	80.4	69.0	416.1	29.4	51.1	60.2	46.9	29.2	51.3	60.9	47.1	282.2
Weather	Snow	81.7	94.4	96.8	91.0	73.2	91.2	94.7	86.4	532.0	51.3	76.8	84.8	71.0	44.9	71.0	79.9	65.3	408.8
	Frost	90.4	97.5	98.8	95.5	79.5	94.7	97.2	90.5	558.1	62.1	84.7	90.7	79.2	51.0	76.7	84.6	70.8	449.8
	Fog	90.2	98.1	99.1	95.8	80.5	95.1	97.4	91.0	560.4	68.3	89.1	94.2	83.9	54.6	79.6	86.9	73.7	472.6
	Brightness	94.5	99.4	99.7	97.8	83.7	96.6	98.2	92.8	572.0	74.6	92.7	96.2	87.8	58.1	82.7	89.5	76.8	493.8
Digital	Contrast	88.2	96.7	97.9	94.3	78.3	93.4	96.0	89.2	550.6	63.8	85.0	90.8	79.9	51.7	76.5	84.3	70.8	452.1
	Elastic	85.3	94.7	96.5	92.2	75.3	91.8	95.1	87.4	538.7	65.7	85.6	91.1	80.8	51.7	76.5	84.4	70.9	455.0
	Pixelate	63.8	78.2	82.4	74.8	55.4	75.3	80.7	70.5	435.9	45.9	65.7	72.7	61.4	36.3	58.9	67.5	54.2	347.0
	JPEG	91.7	98.2	99.1	96.3	79.1	94.6	97.1	90.3	559.8	71.7	91.1	95.4	86.1	55.3	80.0	87.4	74.2	480.9
Stylize	Stylized	70.0	83.7	86.9	80.2	60.0	79.0	84.5	74.5	464.1	50.6	71.9	78.6	67.0	40.3	63.2	71.7	58.4	376.4

Table 14: TCL image perturbation performance comparison of Zero-Shot (ZS) image-text retrieval on Flickr30K and COCO datasets (results are averaged on five perturbation levels).

	Method	Flickr30K (1K)									MSCOCO (5K)								
		Text Retrieval				Image Retrieval					Text Retrieval				Image Retrieval				
		R@1	R@5	R@10	Mean	R@1	R@5	R@10	Mean	RSUM	R@1	R@5	R@10	Mean	R@1	R@5	R@10	Mean	RSUM
Noise	Gaussian	69.3	86.8	90.4	82.2	55.2	78.4	84.8	72.8	464.9	57.9	80.2	87.0	75.0	44.2	70.6	79.9	64.9	419.8
	Shot	70.1	87.0	91.2	82.8	55.5	78.4	84.7	72.9	467.0	57.2	79.9	86.9	74.7	44.0	70.5	79.9	64.8	418.4
	Impluse	67.3	85.9	90.3	81.2	53.7	77.4	83.8	71.6	458.4	57.2	80.2	87.0	74.8	43.8	70.4	79.8	64.7	418.4
	Speckle	78.1	92.9	96.4	89.1	60.3	82.3	88.2	76.9	498.0	62.0	84.2	90.5	78.9	46.7	73.3	82.4	67.5	439.0
Blue	Defocus	60.0	82.0	87.3	76.4	50.2	71.6	78.7	66.9	429.8	54.7	79.1	86.5	73.4	39.9	65.2	74.6	59.9	400.0
	Glass	78.2	94.0	97.2	89.8	63.8	84.1	89.4	79.1	506.6	66.7	88.7	94.7	83.4	46.5	72.6	81.6	66.9	450.8
	Motion	51.2	72.9	80.5	68.2	43.8	66.0	74.1	61.3	388.5	47.6	72.3	80.7	66.9	33.5	57.0	66.4	52.3	357.5
	Zoom	25.0	44.5	53.5	41.0	27.5	45.9	54.9	42.8	251.3	16.7	33.5	42.7	31.0	15.3	30.5	38.7	28.1	177.3
Weather	Snow	51.7	75.4	83.3	70.1	47.6	70.5	78.8	65.7	407.3	37.1	63.8	74.7	58.5	28.5	51.2	61.2	47.0	316.5
	Frost	62.8	85.5	91.3	79.9	52.0	75.2	82.8	70.0	449.5	48.9	75.1	83.9	69.3	34.5	59.7	69.8	54.7	372.0
	Fog	59.0	81.7	89.2	76.6	49.5	73.2	81.6	68.1	434.2	55.7	81.3	89.1	75.4	38.1	63.3	73.1	58.2	400.6
	Brightness	82.4	96.2	98.6	92.4	61.3	82.5	88.1	77.3	509.1	66.8	88.7	94.3	83.3	47.1	73.3	82.0	67.5	452.2
Digital	Contrast	69.8	89.9	94.0	84.6	56.3	78.3	85.0	73.2	473.2	58.5	82.9	89.7	77.0	41.2	67.2	76.6	61.7	416.1
	Elastic	62.4	80.6	85.9	76.3	52.0	73.3	80.3	68.5	434.4	50.6	73.3	80.7	68.2	35.6	59.6	69.2	54.8	369.0
	Pixelate	30.4	46.4	53.3	43.4	25.8	42.2	49.1	39.0	247.2	21.2	36.4	43.3	33.7	17.4	32.4	39.5	29.8	190.3
	JPEG	78.2	93.8	96.6	89.5	61.2	83.4	89.0	77.9	502.2	63.1	86.0	92.0	80.3	46.5	73.1	82.1	67.2	442.7
Stylize	Stylized	44.2	64.8	71.2	60.1	38.4	58.5	66.2	54.4	343.4	33.7	55.0	63.7	50.8	26.3	46.4	55.0	42.6	280.1

Table 15: TCL image perturbation performance comparison of Fine-tuned (FT) image-text retrieval on Flickr30K and COCO datasets (results are averaged on five perturbation levels).

	Method	Flickr30K (1K)									MSCOCO (5K)								
		Text Retrieval				Image Retrieval					Text Retrieval				Image Retrieval				
		R@1	R@5	R@10	Mean	R@1	R@5	R@10	Mean	RSUM	R@1	R@5	R@10	Mean	R@1	R@5	R@10	Mean	RSUM
Noise	Gaussian	83.1	94.3	96.7	91.4	71.4	90.3	94.1	85.3	529.9	64.8	85.8	91.3	80.6	50.8	76.6	84.9	70.8	454.3
	Shot	83.3	95.1	97.1	91.8	71.9	90.7	94.5	85.7	532.6	64.8	85.7	91.3	80.6	50.7	76.8	85.1	70.9	454.4
	Impluse	82.9	94.1	96.5	91.1	70.6	89.9	93.8	84.8	527.7	64.4	85.7	91.5	80.5	50.6	76.7	85.0	70.8	453.9
	Speckle	88.8	97.8	98.7	95.1	76.3	93.5	96.5	88.8	551.6	67.9	88.1	93.4	83.2	53.0	78.8	86.8	72.9	468.1
Blue	Defocus	77.0	90.6	93.5	87.1	66.6	86.1	90.7	81.1	504.5	62.8	84.6	90.7	79.4	50.1	75.8	83.8	69.9	447.8
	Glass	92.7	99.1	99.7	97.2	81.2	95.6	97.7	91.5	566.0	74.1	92.4	96.3	87.6	57.7	82.3	89.2	76.4	491.9
	Motion	78.9	92.2	94.9	88.7	68.1	87.6	92.2	82.6	513.9	60.5	81.9	87.8	76.7	48.4	73.4	81.7	67.8	433.8
	Zoom	51.8	70.5	76.4	66.2	48.4	71.3	78.9	66.2	397.3	24.5	45.2	54.6	41.5	27.2	49.3	59.1	45.2	259.9
Weather	Snow	78.8	93.3	95.9	89.3	70.0	89.9	93.8	84.6	521.7	51.5	76.4	84.7	70.9	44.6	71.2	80.5	65.4	408.9
	Frost	88.1	97.5	98.6	94.7	76.6	93.7	96.5	88.9	551.0	61.2	83.1	89.5	77.9	49.6	75.6	84.1	69.8	443.2
	Fog	88.1	98.0	99.1	95.1	77.9	94.2	96.7	89.6	554.1	67.7	88.3	93.5	83.2	53.9	79.5	87.3	73.5	470.1
	Brightness	93.7	99.0	99.6	97.4	81.9	95.9	97.9	91.9	568.0	73.4	91.6	95.9	87.0	57.1	82.0	89.1	76.1	489.1
Digital	Contrast	90.0	97.8	99.2	95.7	78.5	94.5	97.1	90.0	557.1	67.4	87.8	93.2	82.8	53.6	79.1	86.7	73.1	467.8
	Elastic	81.3	92.4	94.7	89.5	72.1	90.1	93.8	85.3	524.4	61.3	82.4	88.4	77.4	48.9	74.4	82.8	68.7	438.2
	Pixelate	50.1	66.2	72.0	62.8	45.7	65.4	72.5	61.2	372.0	37.7	57.1	65.0	53.3	32.0	54.1	63.1	49.8	309.1
	JPEG	90.2	98.3	99.3	95.9	77.1	93.9	96.7	89.2	555.4	69.9	89.3	94.3	84.5	54.1	79.8	87.4	73.8	474.9
Stylize	Stylized	65.0	80.7	85.0	76.9	57.4	77.5	83.2	72.7	448.7	45.3	67.5	75.3	62.7	38.8	62.6	71.3	57.6	360.9

5.9.2 Text Perturbations

Table 16: CLIP text perturbation performance comparison of Zero-Shot (ZS) image-text retrieval on Flickr30K and COCO datasets (results are averaged on five perturbation levels).

		Flickr30K (1K)									MSCOCO (5K)								
	Method	Text Retrieval				Image Retrieval					Text Retrieval				Image Retrieval				
		R@1	R@5	R@10	Mean	R@1	R@5	R@10	Mean	RSUM	R@1	R@5	R@10	Mean	R@1	R@5	R@10	Mean	RSUM
Character	Keyboard	62.4	86.9	93.1	80.8	43.5	68.8	77.0	63.1	431.8	36.8	62.1	72.8	57.2	21.0	41.2	51.6	37.9	285.5
	Ocr	73.4	93.2	96.7	87.8	52.9	77.3	84.6	71.6	478.2	37.2	62.2	72.6	57.4	21.1	41.5	51.8	38.1	286.4
	CI	66.4	89.6	94.7	83.6	47.3	72.3	80.2	66.6	450.5	37.0	62.1	72.8	57.3	21.2	41.4	51.6	38.1	286.1
	CR	63.0	88.4	93.8	81.7	44.1	68.7	77.2	63.3	435.2	36.6	62.1	72.7	57.1	21.0	41.4	51.7	38.0	285.4
	CS	65.5	89.3	94.9	83.2	45.7	70.4	78.7	65.0	444.6	36.5	62.2	72.6	57.1	21.1	41.4	51.8	38.1	285.6
	CD	66.3	90.4	95.4	84.0	47.2	71.9	80.1	66.4	451.3	36.6	62.2	73.0	57.3	21.1	41.4	51.6	38.0	285.8
Word	SR	76.0	95.1	98.0	89.7	58.0	81.7	88.2	76.0	497.1	47.0	72.8	81.8	67.2	29.2	53.0	63.6	48.6	347.5
	WI	78.3	95.7	98.3	90.8	61.6	84.9	90.9	79.1	509.6	49.9	74.9	83.5	69.4	32.1	56.5	66.9	51.8	363.8
	WS	77.2	95.1	98.0	90.1	59.7	83.6	89.8	77.7	503.3	48.9	73.6	82.3	68.3	30.6	54.7	65.3	50.2	355.5
	WD	80.9	96.8	98.5	92.1	61.4	85.4	91.1	79.3	514.1	51.7	76.4	84.6	70.9	32.3	56.5	67.1	51.9	368.6
	IP	81.8	97.1	98.8	92.6	63.8	86.1	91.6	80.5	519.4	52.4	76.6	84.5	71.2	34.1	58.2	68.4	53.6	374.2
Sentence	Formal	86.4	98.6	99.1	94.7	66.0	88.5	93.1	82.5	531.7	56.8	80.4	87.7	75.0	36.4	60.9	70.8	56.0	393.0
	Casual	84.9	97.9	99.2	94.0	66.1	88.4	92.8	82.4	529.3	57.1	79.6	87.7	74.8	35.9	60.6	70.7	55.7	391.6
	Passive	84.3	96.9	99.2	93.5	64.8	87.3	92.2	81.5	524.8	54.3	77.8	86.1	72.7	34.1	58.4	68.9	53.8	379.6
	Active	85.6	97.9	99.2	94.2	66.9	88.8	93.1	82.9	531.4	57.5	80.3	87.9	75.2	36.1	60.8	70.9	55.9	393.5
	Back_trans	83.9	97.0	98.5	93.1	65.5	87.2	92.2	81.6	524.2	55.1	78.2	85.7	73.0	34.3	58.9	69.1	54.1	381.2

Table 17: CLIP text perturbation performance comparison of Fine-tuned (FT) image-text retrieval on Flickr30K and COCO datasets (results are averaged on five perturbation levels).

		Flickr30K (1K)									MSCOCO (5K)								
	Method	Text Retrieval				Image Retrieval					Text Retrieval				Image Retrieval				
		R@1	R@5	R@10	Mean	R@1	R@5	R@10	Mean	RSUM	R@1	R@5	R@10	Mean	R@1	R@5	R@10	Mean	RSUM
Character	Keyboard	67.0	91.2	96.2	84.8	48.3	74.0	81.6	68.0	458.4	36.8	66.1	78.1	60.3	24.3	49.4	61.3	45.0	316.1
	Ocr	76.2	95.4	98.4	90.0	58.5	83.3	89.1	77.0	500.9	36.8	66.3	77.9	60.4	24.4	49.7	61.5	45.2	316.7
	CI	71.4	93.3	96.8	87.2	53.2	78.1	84.8	72.0	477.6	36.3	66.6	78.2	60.4	24.4	49.6	61.4	45.1	316.5
	CR	68.9	91.7	96.1	85.6	48.7	74.5	81.7	68.3	461.6	36.5	66.3	78.1	60.3	24.3	49.7	61.5	45.2	316.4
	CS	70.7	92.4	96.6	86.6	51.0	76.6	83.7	70.4	471.1	36.5	66.5	78.2	60.4	24.4	49.6	61.4	45.1	316.7
	CD	70.9	93.3	97.2	87.2	52.1	77.5	84.5	71.3	475.5	36.7	66.1	77.9	60.3	24.2	49.5	61.3	45.0	315.6
Word	SR	78.0	96.4	98.5	91.0	63.4	87.2	92.0	80.9	515.4	45.3	75.0	85.1	68.5	33.8	62.7	74.3	56.9	376.2
	WI	81.0	97.0	99.0	92.3	68.3	90.4	94.7	84.4	530.4	48.4	77.3	86.8	70.8	37.3	66.8	78.1	60.7	394.6
	WS	80.8	97.0	99.0	92.2	66.1	89.3	93.9	83.1	526.0	48.0	77.1	86.7	70.6	35.9	65.3	76.9	59.4	389.9
	WD	81.0	97.4	99.1	92.5	67.9	90.7	95.0	84.5	531.1	49.1	77.7	86.8	71.2	37.1	66.7	78.0	60.6	395.3
	IP	83.0	97.9	99.2	93.4	69.9	91.2	95.1	85.4	536.4	51.5	79.5	88.1	73.0	39.1	68.7	79.6	62.5	406.6
Sentence	Formal	85.2	98.4	99.5	94.4	73.3	92.9	96.4	87.6	545.8	53.5	81.0	88.9	74.5	41.7	70.8	81.3	64.6	417.3
	Casual	83.9	97.6	99.4	93.6	72.5	92.3	96.4	87.1	542.1	52.5	80.6	89.0	74.0	41.4	70.4	81.2	64.4	415.2
	Passive	82.9	97.7	99.1	93.2	71.3	91.3	95.6	86.1	537.9	51.9	80.0	88.3	73.4	39.6	68.9	80.0	62.8	408.7
	Active	85.0	97.6	99.4	94.0	73.5	92.9	96.6	87.7	545.1	54.1	81.4	89.0	74.8	42.2	71.1	81.7	65.0	419.4
	Back_trans	83.8	97.7	99.0	93.5	70.4	91.2	95.2	85.6	537.3	51.4	79.1	88.2	72.9	39.6	68.5	79.5	62.5	406.2

Table 18: BLIP text perturbation performance comparison of Fine-tuned (FT) image-text retrieval on Flickr30K and COCO datasets (results are averaged on five perturbation levels).

		Flickr30K (1K)									MSCOCO (5K)								
	Method	Text Retrieval				Image Retrieval					Text Retrieval				Image Retrieval				
		R@1	R@5	R@10	Mean	R@1	R@5	R@10	Mean	RSUM	R@1	R@5	R@10	Mean	R@1	R@5	R@10	Mean	RSUM
Character	Keyboard	84.5	97.3	98.9	93.6	63.8	84.1	89.4	79.1	518.0	64.1	86.4	91.9	80.8	42.7	67.5	76.6	62.2	429.1
	Ocr	93.6	99.5	99.8	97.6	77.5	93.1	96.0	88.9	559.5	74.3	92.2	96.0	87.5	53.6	77.7	85.3	72.2	479.1
	CI	86.6	98.0	99.3	94.7	66.3	86.1	90.9	81.1	527.3	66.7	88.1	93.4	82.7	45.0	70.2	79.0	64.7	442.4
	CR	84.6	97.5	99.0	93.7	63.9	83.8	89.2	79.0	518.0	64.5	86.7	92.1	81.1	42.9	67.7	76.9	62.5	430.8
	CS	87.4	97.9	99.3	94.9	65.9	85.4	90.5	80.6	526.4	67.0	88.1	93.2	82.8	44.6	69.7	78.6	64.3	441.3
	CD	86.8	97.7	99.2	94.6	65.9	85.7	90.4	80.7	525.7	67.0	88.1	93.3	82.8	44.8	69.7	78.6	64.4	441.4
Word	SR	93.8	99.6	99.9	97.8	80.6	94.7	97.0	90.7	565.6	74.2	92.4	96.1	87.6	55.5	79.5	86.7	73.9	484.3
	WI	96.0	99.8	99.9	98.6	85.0	96.9	98.5	93.4	576.1	78.1	94.0	97.1	89.7	60.1	83.2	89.6	77.6	502.1
	WS	94.8	99.6	100.0	98.1	83.6	96.5	98.4	92.8	572.9	75.9	93.2	96.6	88.6	58.1	82.0	88.9	76.3	494.6
	WD	95.1	99.8	100.0	98.3	83.8	96.7	98.5	93.0	573.8	77.3	93.9	97.0	89.4	59.2	82.7	89.5	77.1	499.7
	IP	97.3	99.9	100.0	99.0	87.2	97.5	98.9	94.5	580.7	81.8	95.4	97.8	91.7	63.9	85.6	91.3	80.3	515.8
Sentence	Formal	96.5	99.9	100.0	98.8	86.7	97.1	98.8	94.2	579.0	81.7	95.2	97.6	91.5	63.5	85.3	91.2	80.0	514.4
	Casual	96.8	100.0	100.0	98.9	86.0	97.1	98.7	93.9	578.6	81.3	95.0	97.7	91.3	63.4	85.1	91.1	79.8	513.6
	Passive	96.8	99.8	99.9	98.8	83.3	96.5	98.2	92.7	574.5	80.5	94.7	97.3	90.8	61.7	83.8	90.2	78.6	508.1
	Active	97.1	99.9	100.0	99.0	86.6	97.2	98.7	94.2	579.6	81.6	95.2	97.7	91.5	64.0	85.5	91.3	80.3	515.4
	Back_trans	96.0	99.9	100.0	98.6	84.5	96.1	98.2	92.9	574.7	79.9	94.2	97.0	90.4	61.0	82.9	89.3	77.8	504.3

Table 19: ALBEF text perturbation performance comparison of Fine-tuned (FT) image-text retrieval on Flickr30K and COCO datasets (results are averaged on five perturbation levels).

	Method	Flickr30K (1K)									MSCOCO (5K)								
		Text Retrieval				Image Retrieval					Text Retrieval				Image Retrieval				
		R@1	R@5	R@10	Mean	R@1	R@5	R@10	Mean	RSUM	R@1	R@5	R@10	Mean	R@1	R@5	R@10	Mean	RSUM
Character	Keyboard	82.1	96.0	98.5	92.2	59.7	82.1	87.7	76.5	506.2	57.9	82.6	89.6	76.7	38.0	63.4	73.0	58.1	404.5
	Ocr	91.3	99.2	99.6	96.7	74.6	92.1	95.1	87.3	552.0	69.3	89.9	94.8	84.7	49.5	74.9	83.3	69.2	461.7
	CI	84.4	97.2	98.6	93.4	62.5	84.2	89.2	78.6	516.2	60.8	84.7	91.0	78.8	40.6	66.2	75.6	60.8	418.9
	CR	82.1	95.9	98.4	92.1	59.9	81.6	87.2	76.2	505.0	58.3	82.9	89.9	77.0	38.3	63.6	73.1	58.3	406.1
	CS	82.9	96.8	98.8	92.8	61.6	83.2	88.4	77.7	511.7	59.9	84.1	90.8	78.3	39.8	65.3	74.8	60.0	414.7
	CD	83.6	96.7	98.5	92.9	61.9	83.6	88.7	78.1	513.0	60.0	84.1	90.8	78.3	39.9	65.7	75.1	60.2	415.5
Word	SR	92.9	99.2	99.8	97.3	78.7	94.5	96.8	90.0	561.9	70.1	90.6	95.1	85.3	52.4	77.7	85.5	71.9	471.4
	WI	94.3	99.6	99.9	97.9	82.9	96.6	98.3	92.6	571.6	73.2	92.4	96.3	87.3	56.8	81.6	88.7	75.7	488.9
	WS	93.3	99.4	99.9	97.6	81.5	96.3	98.1	92.0	568.6	72.0	91.8	96.1	86.6	55.1	80.6	88.2	74.6	483.7
	WD	93.4	99.5	99.9	97.6	82.2	96.5	98.3	92.4	570.0	72.9	92.1	96.1	87.0	55.7	81.1	88.5	75.1	486.3
	IP	95.9	99.8	100.0	98.6	85.5	97.5	98.9	94.0	577.7	77.6	94.3	97.2	89.7	60.7	84.3	90.5	78.5	504.5
Sentence	Formal	95.4	99.7	99.9	98.3	85.2	97.3	98.7	93.7	576.2	77.6	94.1	97.0	89.6	60.2	83.9	90.3	78.1	503.1
	Casual	95.1	99.7	100.0	98.3	84.6	97.1	98.5	93.4	575.0	77.1	94.1	97.4	89.5	59.7	83.6	90.1	77.8	502.0
	Passive	94.6	99.4	100.0	98.0	81.5	96.1	98.0	91.8	569.5	76.1	93.4	96.7	88.7	58.4	82.6	89.2	76.7	496.4
	Active	95.6	99.8	100.0	98.5	85.0	97.3	98.7	93.7	576.4	77.5	94.2	97.1	89.6	60.4	84.2	90.3	78.3	503.7
	Back_trans	95.9	99.7	99.9	98.5	83.0	96.1	98.0	92.3	572.5	75.2	93.0	96.4	88.2	57.4	81.0	88.3	75.6	491.3

Table 20: TCL text perturbation performance comparison of Zero-Shot (ZS) image-text retrieval on Flickr30K and COCO datasets (results are averaged on five perturbation levels).

	Method	Flickr30K (1K)									MSCOCO (5K)								
		Text Retrieval				Image Retrieval					Text Retrieval				Image Retrieval				
		R@1	R@5	R@10	Mean	R@1	R@5	R@10	Mean	RSUM	R@1	R@5	R@10	Mean	R@1	R@5	R@10	Mean	RSUM
Character	Keyboard	63.8	87.2	92.7	81.2	44.1	68.8	76.7	63.2	433.3	49.6	76.1	84.9	70.2	32.3	57.2	67.8	52.4	368.0
	Ocr	78.2	94.8	97.9	90.3	58.8	82.1	88.1	76.3	499.9	61.4	85.1	91.6	79.4	42.6	69.0	78.7	63.4	428.4
	CI	67.3	88.0	93.4	82.9	45.9	70.5	78.3	64.9	443.3	51.9	78.5	86.7	72.4	34.1	59.8	70.3	54.7	381.3
	CR	63.1	85.9	91.4	80.1	43.8	68.1	76.1	62.7	428.4	49.7	76.1	85.1	70.3	32.2	57.4	67.9	52.5	368.4
	CS	66.5	88.6	93.8	83.0	46.3	70.8	78.5	65.2	444.4	52.6	78.5	87.0	72.7	34.0	59.7	70.1	54.6	382.0
	CD	66.7	89.4	94.2	83.4	47.2	71.9	79.4	66.2	448.9	52.6	78.8	86.9	72.8	34.3	60.2	70.6	55.0	383.4
Word	SR	78.3	95.3	97.9	90.5	63.2	86.0	91.1	80.1	511.9	62.1	85.7	91.9	79.9	45.8	72.3	81.5	66.5	439.3
	WI	80.0	96.3	98.5	91.6	67.0	88.6	93.4	83.0	523.8	63.3	86.8	93.0	81.0	49.5	76.1	84.7	70.1	453.4
	WS	80.4	95.9	98.4	91.6	64.8	87.2	92.4	81.5	519.1	63.2	86.5	92.7	80.8	46.5	73.8	83.0	67.8	445.7
	WD	83.6	97.1	98.8	93.1	67.0	89.0	93.4	83.1	528.8	65.3	87.2	93.1	81.9	47.6	74.4	83.3	68.4	450.9
	IP	89.4	98.6	99.6	95.9	73.4	92.2	95.5	87.0	548.6	71.4	90.8	95.4	85.9	53.5	79.0	87.1	73.2	477.2
Sentence	Formal	88.0	98.0	99.8	95.3	72.0	91.6	95.1	86.2	544.4	70.8	90.6	95.2	85.5	52.9	78.4	86.5	72.6	474.4
	Casual	87.2	98.3	99.5	95.0	71.4	91.2	94.8	85.8	542.4	69.9	90.2	94.9	85.0	52.3	78.1	86.4	72.3	471.8
	Passive	84.5	97.1	99.4	93.7	67.6	88.6	92.9	83.0	530.1	68.6	89.1	94.4	84.0	50.5	76.9	85.2	70.9	464.7
	Active	89.3	98.3	99.9	95.8	72.9	91.5	95.1	86.5	547.1	70.9	90.6	95.3	85.6	53.1	78.9	86.9	73.0	475.7
	Back_trans	86.0	97.6	99.4	94.3	69.4	89.8	93.6	84.3	535.8	68.5	89.2	94.2	83.9	50.3	75.9	84.1	70.1	462.0

Table 21: TCL text perturbation performance comparison of Fine-tuned (FT) image-text retrieval on Flickr30K and COCO datasets (results are averaged on five perturbation levels).

	Method	Flickr30K (1K)									MSCOCO (5K)								
		Text Retrieval				Image Retrieval					Text Retrieval				Image Retrieval				
		R@1	R@5	R@10	Mean	R@1	R@5	R@10	Mean	RSUM	R@1	R@5	R@10	Mean	R@1	R@5	R@10	Mean	RSUM
Character	Keyboard	79.7	95.2	97.9	90.9	57.0	79.1	85.4	73.8	494.3	55.8	81.3	88.8	75.3	36.9	62.5	72.4	57.3	397.8
	Ocr	90.0	99.1	99.7	96.3	71.7	90.4	94.0	85.4	545.0	67.6	88.9	94.0	83.5	48.0	73.9	82.6	68.2	455.1
	CI	82.2	96.2	98.3	92.2	59.6	81.4	87.2	76.1	504.9	58.5	83.5	90.4	77.5	39.3	65.3	75.0	59.8	412.0
	CR	79.3	94.8	97.8	90.7	56.7	79.1	85.0	73.6	492.8	55.6	81.5	89.0	75.4	37.2	62.7	72.5	57.5	398.5
	CS	80.7	96.0	98.2	91.6	59.0	81.2	86.8	75.7	501.9	57.6	82.9	90.2	76.9	38.7	64.8	74.6	59.4	408.8
	CD	81.4	95.7	98.3	91.8	59.1	81.2	86.7	75.7	502.4	58.1	83.0	90.0	77.0	39.2	65.3	75.0	59.8	410.5
Word	SR	91.0	99.1	99.7	96.6	76.1	93.0	95.8	88.3	554.7	67.8	89.1	94.2	83.7	51.0	76.8	84.8	70.8	463.7
	WI	93.4	99.4	99.8	97.5	80.5	95.5	97.7	91.2	566.4	70.8	91.0	95.6	85.8	55.3	80.6	88.0	74.6	481.3
	WS	91.0	99.1	99.6	96.6	78.2	94.7	97.4	90.1	560.0	69.2	90.3	94.9	84.8	52.3	78.5	86.6	72.5	471.8
	WD	92.6	99.4	99.8	97.3	79.5	95.3	97.6	90.8	564.2	70.8	90.7	95.5	85.7	53.7	79.7	87.3	73.6	477.7
	IP	94.9	99.5	99.8	98.1	84.0	96.7	98.5	93.1	573.4	75.6	92.8	96.7	88.3	59.0	83.2	89.9	77.3	497.1
Sentence	Formal	94.4	99.4	99.8	97.9	83.2	96.5	98.3	92.6	571.5	75.3	92.4	96.7	88.1	58.2	82.7	89.5	76.8	494.6
	Casual	94.0	99.5	99.9	97.8	82.1	96.0	98.0	92.1	569.6	74.6	92.1	96.5	87.8	57.9	82.5	89.4	76.6	493.0
	Passive	92.7	99.1	99.8	97.2	79.5	94.5	97.1	90.4	562.8	73.5	91.9	96.1	87.2	56.3	81.3	88.3	75.3	487.3
	Active	94.8	99.5	99.8	98.0	83.5	96.4	98.2	92.7	572.1	75.4	92.7	96.6	88.2	58.7	83.0	89.7	77.1	496.0
	Back_trans	93.9	99.5	99.9	97.8	80.6	95.3	97.3	91.1	566.5	72.7	91.6	96.0	86.8	55.5	80.3	87.3	74.4	483.5