eman ta zabal zazu

Universidad del País Vasco
Euskal Herriko Unibertsitatea

INFORMATIKA FAKULTATEA
FACULTAD DE INFORMÁTICA

Master Thesis

Master in Language Analysis and Processing

Grounding Language Models for Spatial Reasoning

Julen Etxaniz

Advisors
Oier Lopez de Lacalle
Aitor Soroa

2023

Acknowledgements

Abstract

Contents

List of Figures

List of Tables

List of algorithms

1 Introduction

2 Related Work

3 Datasets

This chapter introduces the datasets and metrics we used.

3.1 Winoground

3.1.1 Dataset

3.1.2 Metrics

3.1.2.1 Score

Performance on Winoground is computed according to three different metrics that evaluate different aspects of the models' visio-linguistic reasoning abilities.

The first metric is the **text score**, which measures whether a model can select the correct caption, given an image. Given images I_0 and I_1 and captions C_0 and C_1, the text score for an example (C_0, I_0, C_1, I_1) is computed according to:

$$f(C_0, I_0, C_1, I_1) = \begin{cases} 1 & \text{if } s(C_0, I_0) > s(C_1, I_0) \\ & \quad \text{and } s(C_1, I_1) > s(C_0, I_1) \\ 0 & \text{otherwise} \end{cases} \tag{3.1}$$

where $s(\cdot)$ is the model's score for the image/caption pair.

The second metric is the **image score**, which measures whether a model can select the correct image, given a caption. Given images I_0 and I_1 and captions C_0 and C_1, the image score for an example is computed according to:

$$g(C_0, I_0, C_1, I_1) = \begin{cases} 1 & \text{if } s(C_0, I_0) > s(C_0, I_1) \\ & \quad \text{and } s(C_1, I_1) > s(C_1, I_0) \\ 0 & \text{otherwise} \end{cases} \tag{3.2}$$

The group score in our framework is computed according to:

$$h(C_0, I_0, C_1, I_1) = \begin{cases} 1 & \text{if } f(C_0, I_0, C_1, I_1) \\ & \quad \text{and } g(C_0, I_0, C_1, I_1) \\ 0 & \text{otherwise} \end{cases} \tag{3.3}$$

3.1.2.2 Accuracy

Given images I_0 and I_1 and captions C_0 and C_1, the text accuracy for an example (C_0, I_0, C_1, I_1) is computed according to:

$$f(C_0, I_0, C_1, I_1) = \begin{cases} 1 & \text{if } s(C_0, I_0) > s(C_1, I_0) \\ & \quad \text{and } s(C_1, I_1) > s(C_0, I_1) \\ 0.5 & \text{if } s(C_0, I_0) > s(C_1, I_0) \\ & \quad \text{xor } s(C_1, I_1) > s(C_0, I_1) \\ 0 & \text{otherwise} \end{cases} \tag{3.4}$$

where $s(\cdot)$ is the model's score for the image/caption pair.

Given images I_0 and I_1 and captions C_0 and C_1, the image accuracy for an example is computed according to:

$$g(C_0, I_0, C_1, I_1) = \begin{cases} 1 & \text{if } s(C_0, I_0) > s(C_0, I_1) \\ & \quad \text{and } s(C_1, I_1) > s(C_1, I_0) \\ 0.5 & \text{if } s(C_0, I_0) > s(C_0, I_1) \\ & \quad \text{xor } s(C_1, I_1) > s(C_1, I_0) \\ 0 & \text{otherwise} \end{cases} \tag{3.5}$$

The group score in our framework is computed according to:

$$h(C_0, I_0, C_1, I_1) = (f(C_0, I_0, C_1, I_1) + g(C_0, I_0, C_1, I_1))/2 \tag{3.6}$$

4 Methods

5 Results

This chapter introduces baseline results and our results.

5.1 Compared To Humans

5.1.1 Baseline

	Score			Accuracy		
Model	Text	Image	Group	Text	Image	Group
MTurk Human	**89.50**	**88.50**	**85.50**	**93.75**	**93.88**	**93.81**
Random Chance	25.00	25.00	16.67	50.00	50.00	50.00
VinVL	**37.75**	17.75	14.50	**62.75**	**57.75**	**60.25**
UNITER$_{large}$	**38.00**	14.00	10.50	**63.25**	**55.75**	**59.50**
UNITER$_{base}$	**32.25**	13.25	10.00	**60.62**	**55.50**	**58.06**
ViLLA$_{large}$	**37.00**	13.25	11.00	**62.62**	**55.25**	**58.94**
ViLLA$_{base}$	**30.00**	12.00	8.00	**59.62**	**55.00**	**57.31**
VisualBERT$_{base}$	15.50	2.50	1.50	**50.50**	49.88	**50.19**
ViLT (ViT-B/32)	**34.75**	14.00	9.25	**60.50**	**55.38**	**57.94**
LXMERT	19.25	7.00	4.00	**52.12**	**51.88**	**52.00**
ViLBERT$_{base}$	23.75	7.25	4.75	**57.25**	**52.50**	**54.87**
UniT$_{ITMFinetuned}$	19.50	6.25	4.00	**50.25**	**50.75**	**50.50**
FLAVA$_{ITM}$	**32.25**	20.50	14.25	**62.75**	**59.13**	**60.94**
FLAVA$_{Contrastive}$	**25.25**	13.50	9.00	**59.25**	**55.12**	**57.19**
CLIP (ViT-B/32)	**30.75**	10.50	8.00	**60.38**	**53.25**	**56.81**
VSE++$_{COCO}$ (ResNet)	22.75	8.00	4.00	**51.38**	**50.88**	**51.12**
VSE++$_{COCO}$ (VGG)	18.75	5.50	3.50	**50.38**	49.75	**50.06**
VSE++$_{Flickr30k}$ (ResNet)	20.00	5.00	2.75	**51.50**	**50.25**	**50.88**
VSE++$_{Flickr30k}$ (VGG)	19.75	6.25	4.50	**52.75**	**51.00**	**51.88**
VSRN$_{COCO}$	17.50	7.00	3.75	**50.38**	**51.12**	**50.75**
VSRN$_{Flickr30k}$	20.00	5.00	3.50	**53.25**	**51.75**	**52.50**

Table 5.1: Results on the Winoground dataset across the text, image and group score and accuracy metrics. Results above random chance in **bold**.

Model	Score			Accuracy		
	Text	Image	Group	Text	Image	Group
MTurk Human	**89.50**	**88.50**	**85.50**	**93.75**	**93.88**	**93.81**
Random Chance	25.00	25.00	16.67	50.00	50.00	50.00
ViLT (ViT-B/32)	**27.50**	8.75	6.00	**56.88**	**53.12**	**55.00**
ViLT$_{COCO}$ (ViT-B/32)	**32.75**	13.50	11.25	**61.88**	**56.00**	**58.94**
ViLT$_{Flickr30k}$ (ViT-B/32)	**35.00**	11.50	9.75	**61.62**	**54.50**	**58.06**
FLAVA$_{ITM}$	**32.25**	20.50	14.25	**62.75**	**59.13**	**60.94**
FLAVA$_{ITC}$	**25.25**	13.50	9.00	**59.25**	**55.12**	**57.19**
CLIP (ViT-B/32)	**30.75**	10.25	8.25	**60.38**	**53.12**	**56.75**
CLIP (ViT-B/16)	25.00	10.25	7.00	**57.88**	**53.75**	**55.81**
CLIP (ViT-L/14)	**28.50**	11.00	8.00	**60.38**	**54.62**	**57.50**
CLIP (ViT-L/14-336)	**27.50**	12.00	8.00	**59.38**	**55.12**	**57.25**
BLIP$_{ITM14M}$ (ViT-B/16)	**39.25**	19.00	15.00	**65.88**	**58.25**	**62.06**
BLIP$_{ITC14M}$ (ViT-B/16)	**32.25**	13.75	10.50	**62.25**	**56.50**	**59.38**
BLIP$_{ITM}$ (ViT-B/16)	**40.50**	20.50	16.50	**66.25**	**59.00**	**62.62**
BLIP$_{ITC}$ (ViT-B/16)	**29.75**	14.50	9.50	**59.88**	**56.12**	**58.00**
BLIP$_{ITM}$ (ViT-B/16) (CapFilt-L)	**37.50**	18.50	14.00	**65.00**	**59.13**	**62.06**
BLIP$_{ITC}$ (ViT-B/16) (CapFilt-L)	**31.50**	10.50	8.50	**61.38**	**53.62**	**57.50**
BLIP$_{ITM}$ (ViT-L/16)	**42.50**	18.25	15.50	**66.88**	**57.25**	**62.06**
BLIP$_{ITC}$ (ViT-L/16)	**33.25**	12.00	9.00	**61.75**	**55.00**	**58.38**
BLIP$_{ITMCOCO}$ (ViT-B/16)	**48.00**	24.50	**20.00**	**69.88**	**61.25**	**65.56**
BLIP$_{ITCCOCO}$ (ViT-B/16)	**37.75**	15.75	12.75	**65.00**	**56.88**	**60.94**
BLIP$_{ITMFlickr30k}$ (ViT-B/16)	**46.25**	24.25	**21.25**	**69.25**	**60.62**	**64.94**
BLIP$_{ITCFlickr30k}$ (ViT-B/16)	**38.25**	15.00	12.25	**65.38**	**56.12**	**60.75**
BLIP$_{ITMCOCO}$ (ViT-L/16)	**46.75**	24.00	**20.50**	**68.88**	**61.00**	**64.94**
BLIP$_{ITCCOCO}$ (ViT-L/16)	**37.75**	13.75	10.50	**64.88**	**55.75**	**60.31**
BLIP$_{ITMFlickr30k}$ (ViT-L/16)	**45.00**	24.75	**20.50**	**68.62**	**60.50**	**64.56**
BLIP$_{ITCFlickr30k}$ (ViT-L/16)	**36.00**	16.25	13.50	**63.38**	**56.75**	**60.06**

Table 5.2: Results on the Winoground dataset across the text, image and group score and accuracy metrics. Results above random chance in **bold**.

Model	Object			Relation			Both			1 Main Pred			2 Main Preds		
	Text	Image	Group	Text	Image	Group	Text	Image	Group	Text	Image	Group	Text	Image	Group
MTurk Human	**92.20**	**90.78**	**88.65**	**89.27**	**90.56**	**86.70**	**76.92**	**57.69**	**57.69**	**87.33**	**85.62**	**82.53**	**95.37**	**96.30**	**93.52**
VinVL	**36.88**	17.73	14.18	**37.77**	17.60	14.16	**42.31**	19.23	**19.23**	**39.38**	21.23	**17.47**	**33.33**	8.33	6.48
UNITER$_{large}$	**39.01**	12.77	9.93	**36.05**	14.16	9.87	**50.00**	19.23	**19.23**	**40.07**	16.44	13.36	**32.41**	7.41	2.78
UNITER$_{base}$	**34.04**	11.35	9.22	**30.04**	14.16	10.30	**42.31**	15.38	11.54	**35.27**	14.73	11.99	24.07	9.26	4.63
ViLLA$_{large}$	**36.88**	14.89	11.35	**37.34**	12.88	11.16	**34.62**	7.69	7.69	**39.73**	17.12	14.38	**29.63**	2.78	1.85
ViLLA$_{base}$	**33.33**	15.60	9.93	**27.04**	9.01	6.01	**38.46**	19.23	15.38	**33.22**	14.04	10.27	21.30	6.48	1.85
VisualBERT$_{base}$	19.15	2.13	0.71	12.88	2.15	1.72	19.23	7.69	3.85	16.44	2.74	1.71	12.96	1.85	0.93
ViLT (ViT-B/32)	**31.91**	15.60	9.22	**36.91**	11.59	8.15	**30.77**	**26.92**	**19.23**	**35.27**	17.12	11.64	**33.33**	5.56	2.78
LXMERT	22.70	9.22	6.38	17.60	5.58	2.58	15.38	7.69	3.85	19.18	8.56	5.14	19.44	2.78	0.93
ViLBERT$_{base}$	**29.08**	10.64	7.09	19.31	3.00	1.72	**34.62**	**26.92**	**19.23**	23.97	8.90	5.82	23.15	2.78	1.85
UniT$_{ITMfinetuned}$	17.73	5.67	2.13	18.03	4.72	3.43	**42.31**	23.08	**19.23**	21.58	6.85	4.11	13.89	4.63	3.70
FLAVA$_{ITM}$	**31.91**	23.40	14.89	**30.04**	16.31	12.02	**53.85**	**42.31**	**30.77**	**36.30**	24.66	**17.81**	21.30	9.26	4.63
FLAVA$_{Contrastive}$	23.40	19.15	11.35	23.61	8.58	5.58	**50.00**	**26.92**	**26.92**	**26.37**	16.44	10.62	22.22	5.56	4.63
CLIP (ViT-B/32)	**34.75**	7.80	6.38	22.75	8.58	5.58	**80.77**	**42.31**	**38.46**	**35.27**	13.01	10.27	18.52	3.70	1.85
VSE++$_{COCO}$ (ResNet)	21.99	6.38	1.42	23.61	9.01	5.58	19.23	7.69	3.85	25.00	9.59	4.79	16.67	3.70	1.85
VSE++$_{COCO}$ (VGG)	17.73	2.13	2.13	18.45	7.30	3.86	**26.92**	7.69	7.69	18.49	4.79	2.74	19.44	7.41	5.56
VSE++$_{Flickr30k}$ (ResNet)	20.57	6.38	3.55	18.88	4.29	2.15	**26.92**	3.85	3.85	21.58	6.51	3.42	15.74	0.93	0.93
VSE++$_{Flickr30k}$ (VGG)	17.73	4.96	2.84	19.74	6.87	5.15	**30.77**	7.69	7.69	20.55	6.16	4.79	17.59	6.48	3.70
VSRN$_{COCO}$	15.60	4.96	2.13	18.88	7.73	4.72	15.38	11.54	3.85	17.12	7.19	3.77	18.52	6.48	3.70
VSRN$_{Flickr30k}$	16.31	4.96	2.13	21.03	4.29	3.86	**30.77**	11.54	7.69	20.89	5.82	3.77	17.59	2.78	2.78

Table 5.3: The results by linguistic tag. Results above chance are in **bold**.

	Object			Relation			Both			1 Main Pred			2 Main Preds		
Model	Text	Image	Group	Text	Image	Group	Text	Image	Group	Text	Image	Group	Text	Image	Group
MTurk Human	**92.20**	**90.78**	**88.65**	**89.27**	**90.56**	**86.70**	**76.92**	**57.69**	**57.69**	**87.33**	**85.62**	**82.53**	**95.37**	**96.30**	**93.52**
ViLT (ViT-B/32)	**29.08**	10.64	4.96	**26.18**	7.73	6.44	**30.77**	7.69	7.69	**30.14**	10.62	7.53	20.37	3.70	1.85
ViLT$_{COCO}$ (ViT-B/32)	**33.33**	15.60	12.77	**30.90**	10.73	9.01	**46.15**	**26.92**	**23.08**	**36.64**	15.75	14.04	22.22	7.41	3.70
ViLT$_{Flickr30k}$ (ViT-B/32)	**32.62**	14.89	11.35	**35.62**	8.15	7.73	**42.31**	23.08	**19.23**	**36.99**	14.38	11.99	**29.63**	3.70	3.70
FLAVA$_{ITM}$	**31.91**	23.40	14.89	**30.04**	16.31	12.02	**53.85**	**42.31**	**30.77**	**36.30**	24.66	**17.81**	21.30	9.26	4.63
FLAVA$_{ITC}$	23.40	19.15	11.35	23.61	8.58	5.58	**50.00**	**26.92**	**26.92**	**26.37**	16.44	10.62	22.22	5.56	4.63
CLIP (ViT-B/32)	**35.46**	7.80	6.38	22.32	7.73	5.58	**80.77**	**46.15**	**42.31**	**35.62**	13.01	10.62	17.59	2.78	1.85
CLIP (ViT-B/16)	**27.66**	10.64	5.67	19.31	6.44	4.29	**61.54**	**42.31**	**38.46**	**30.14**	11.99	8.90	11.11	5.56	1.85
CLIP (ViT-L/14)	**27.66**	8.51	5.67	**25.75**	9.87	6.44	**57.69**	**34.62**	**34.62**	**30.14**	13.01	9.93	24.07	5.56	2.78
CLIP (ViT-L/14-336)	**32.62**	12.77	9.22	21.03	8.15	4.29	**57.69**	**42.31**	**34.62**	**30.48**	14.04	10.62	19.44	6.48	0.93
BLIP$_{ITM14M}$ (ViT-B/16)	**41.84**	23.40	**17.73**	**36.05**	14.59	11.59	**53.85**	**34.62**	**30.77**	**43.84**	23.63	**18.49**	**26.85**	6.48	5.56
BLIP$_{ITC14M}$ (ViT-B/16)	**34.04**	13.48	9.93	**28.33**	12.02	9.44	**57.69**	**30.77**	**23.08**	**37.67**	16.44	13.01	17.59	6.48	3.70
BLIP$_{ITM}$ (ViT-B/16)	**46.10**	22.70	**17.73**	**35.62**	17.60	14.16	**53.85**	**34.62**	**30.77**	**45.89**	**25.34**	**20.55**	**25.93**	7.41	5.56
BLIP$_{ITC}$ (ViT-B/16)	**34.75**	14.18	9.22	**25.32**	13.73	8.58	**42.31**	23.08	**19.23**	**33.56**	16.10	10.62	19.44	10.19	6.48
BLIP$_{ITM}$ (ViT-B/16) (CapFilt-L)	**39.01**	19.86	12.77	**34.76**	15.88	12.45	**53.85**	**34.62**	**34.62**	**41.10**	22.60	**17.12**	**27.78**	7.41	5.56
BLIP$_{ITC}$ (ViT-B/16) (CapFilt-L)	**36.88**	12.77	9.22	**26.18**	8.58	7.30	**50.00**	15.38	15.38	**35.96**	13.36	10.96	19.44	2.78	1.85
BLIP$_{ITM}$ (ViT-L/16)	**41.84**	19.86	**17.02**	**40.77**	16.31	13.73	**61.54**	**26.92**	**23.08**	**45.55**	23.29	**20.21**	**34.26**	4.63	2.78
BLIP$_{ITC}$ (ViT-L/16)	**34.04**	14.18	11.35	**30.90**	9.01	6.01	**50.00**	**26.92**	**23.08**	**36.99**	14.04	10.96	23.15	6.48	3.70
BLIP$_{ITMCOCO}$ (ViT-B/16)	**42.55**	**26.95**	**19.15**	**49.79**	21.89	**19.31**	**61.54**	**34.62**	**30.77**	**48.97**	**29.79**	**24.66**	**45.37**	10.19	7.41
BLIP$_{ITCCOCO}$ (ViT-B/16)	**36.88**	19.15	14.18	**36.05**	11.59	10.30	**57.69**	**34.62**	**26.92**	**41.78**	18.84	15.07	**26.85**	7.41	6.48
BLIP$_{ITMFlickr30k}$ (ViT-B/16)	**49.65**	**28.37**	**22.70**	**42.49**	19.74	**18.45**	**61.54**	**42.31**	**38.46**	**51.03**	**28.42**	**26.03**	**33.33**	12.96	8.33
BLIP$_{ITCFlickr30k}$ (ViT-B/16)	**36.88**	17.02	10.64	**36.48**	12.02	11.16	**61.54**	**30.77**	**30.77**	**40.75**	17.12	13.70	**31.48**	9.26	8.33
BLIP$_{ITMCOCO}$ (ViT-L/16)	**48.94**	**25.53**	**20.57**	**44.64**	22.32	**20.60**	**53.85**	**30.77**	**19.23**	**51.03**	**28.42**	**23.97**	**35.19**	12.04	11.11
BLIP$_{ITCCOCO}$ (ViT-L/16)	**36.88**	14.18	11.35	**36.05**	11.16	7.30	**57.69**	**34.62**	**34.62**	**41.10**	16.44	13.36	**28.70**	6.48	2.78
BLIP$_{ITMFlickr30k}$ (ViT-L/16)	**46.10**	22.70	16.31	**42.06**	24.89	**21.46**	**65.38**	**34.62**	**34.62**	**50.34**	**29.11**	**24.66**	**30.56**	12.96	9.26
BLIP$_{ITCFlickr30k}$ (ViT-L/16)	**39.01**	19.86	15.60	**30.47**	11.59	9.44	**69.23**	**38.46**	**38.46**	**39.38**	20.55	**17.12**	**26.85**	4.63	3.70

Table 5.4: The results by linguistic tag. Results above chance are in **bold**.

	Symbolic			Pragmatics			Same Image Series		
Model	Text	Image	Group	Text	Image	Group	Text	Image	Group
MTurk Human	**96.43**	**92.86**	**92.86**	**58.82**	**41.18**	**41.18**	**95.65**	**91.30**	**91.30**
VinVL	25.00	17.86	14.29	**29.41**	5.88	5.88	**34.78**	17.39	13.04
UNITER$_{large}$	**39.29**	**28.57**	**17.86**	**35.29**	0.00	0.00	4.35	8.70	0.00
UNITER$_{base}$	**46.43**	14.29	14.29	**29.41**	17.65	11.76	8.70	8.70	0.00
ViLLA$_{large}$	**39.29**	14.29	10.71	17.65	0.00	0.00	17.39	4.35	0.00
ViLLA$_{base}$	**42.86**	17.86	14.29	**29.41**	5.88	5.88	13.04	8.70	4.35
VisualBERT$_{base}$	**28.57**	0.00	0.00	5.88	0.00	0.00	13.04	0.00	0.00
ViLT (ViT-B/32)	**28.57**	17.86	10.71	**35.29**	0.00	0.00	**26.09**	0.00	0.00
LXMERT	**28.57**	3.57	3.57	17.65	5.88	0.00	8.70	4.35	0.00
ViLBERT$_{base}$	**28.57**	10.71	7.14	**29.41**	5.88	5.88	13.04	0.00	0.00
UniT$_{ITMfinetuned}$	14.29	10.71	7.14	17.65	5.88	5.88	21.74	4.35	4.35
FLAVA$_{ITM}$	25.00	**28.57**	**17.86**	17.65	**29.41**	11.76	17.39	8.70	0.00
FLAVA$_{Contrastive}$	17.86	10.71	10.71	11.76	23.53	5.88	17.39	4.35	4.35
CLIP (ViT-B/32)	**39.29**	3.57	3.57	**35.29**	5.88	5.88	8.70	0.00	0.00
VSE++$_{COCO}$ (ResNet)	**32.14**	10.71	10.71	23.53	11.76	0.00	13.04	4.35	4.35
VSE++$_{COCO}$ (VGG)	17.86	14.29	7.14	17.65	0.00	0.00	13.04	4.35	4.35
VSE++$_{Flickr30k}$ (ResNet)	21.43	3.57	0.00	23.53	0.00	0.00	17.39	4.35	0.00
VSE++$_{Flickr30k}$ (VGG)	**28.57**	10.71	10.71	11.76	0.00	0.00	13.04	4.35	0.00
VSRN$_{COCO}$	7.14	3.57	0.00	11.76	0.00	0.00	13.04	0.00	0.00
VSRN$_{Flickr30k}$	21.43	3.57	3.57	**35.29**	11.76	5.88	8.70	4.35	4.35

Table 5.5: The results by visual tag. Results above chance are in **bold**.

	Symbolic			Pragmatics			Same Image Series		
Model	Text	Image	Group	Text	Image	Group	Text	Image	Group
MTurk Human	**96.43**	**92.86**	**92.86**	**58.82**	**41.18**	**41.18**	**95.65**	**91.30**	**91.30**
ViLT (ViT-B/32)	21.43	7.14	3.57	17.65	5.88	5.88	17.39	8.70	4.35
ViLT$_{COCO}$ (ViT-B/32)	21.43	10.71	10.71	**29.41**	17.65	5.88	21.74	8.70	4.35
ViLT$_{Flickr30k}$ (ViT-B/32)	**28.57**	7.14	7.14	23.53	0.00	0.00	**26.09**	4.35	4.35
FLAVA$_{ITM}$	25.00	**28.57**	**17.86**	17.65	**29.41**	11.76	17.39	8.70	0.00
FLAVA$_{ITC}$	17.86	10.71	10.71	11.76	23.53	5.88	17.39	4.35	4.35
CLIP (ViT-B/32)	**35.71**	3.57	3.57	**35.29**	5.88	5.88	13.04	0.00	0.00
CLIP (ViT-B/16)	21.43	3.57	3.57	**29.41**	11.76	11.76	4.35	4.35	0.00
CLIP (ViT-L/14)	**28.57**	10.71	3.57	23.53	17.65	11.76	13.04	8.70	4.35
CLIP (ViT-L/14-336)	**28.57**	14.29	7.14	17.65	17.65	5.88	13.04	4.35	0.00
BLIP$_{ITM14M}$ (ViT-B/16)	**46.43**	17.86	**17.86**	**35.29**	11.76	11.76	17.39	4.35	0.00
BLIP$_{ITC14M}$ (ViT-B/16)	**32.14**	14.29	10.71	**29.41**	0.00	0.00	13.04	0.00	0.00
BLIP$_{ITM}$ (ViT-B/16)	**50.00**	17.86	**17.86**	**29.41**	5.88	5.88	13.04	4.35	0.00
BLIP$_{ITC}$ (ViT-B/16)	**39.29**	10.71	7.14	5.88	11.76	0.00	4.35	8.70	0.00
BLIP$_{ITM}$ (ViT-B/16) (CapFilt-L)	**42.86**	17.86	14.29	23.53	17.65	**17.65**	17.39	4.35	0.00
BLIP$_{ITC}$ (ViT-B/16) (CapFilt-L)	**42.86**	0.00	0.00	17.65	0.00	0.00	4.35	0.00	0.00
BLIP$_{ITM}$ (ViT-L/16)	**53.57**	25.00	**25.00**	**29.41**	5.88	0.00	**26.09**	4.35	0.00
BLIP$_{ITC}$ (ViT-L/16)	**39.29**	17.86	14.29	**41.18**	11.76	11.76	8.70	4.35	4.35
BLIP$_{ITMCOCO}$ (ViT-B/16)	**53.57**	17.86	**17.86**	**58.82**	17.65	**17.65**	**39.13**	8.70	0.00
BLIP$_{ITCCOCO}$ (ViT-B/16)	25.00	10.71	7.14	**35.29**	5.88	5.88	17.39	8.70	4.35
BLIP$_{ITMFlickr30k}$ (ViT-B/16)	**53.57**	21.43	**21.43**	**35.29**	11.76	11.76	**26.09**	4.35	4.35
BLIP$_{ITCFlickr30k}$ (ViT-B/16)	**35.71**	10.71	10.71	23.53	17.65	11.76	17.39	4.35	0.00
BLIP$_{ITMCOCO}$ (ViT-L/16)	**39.29**	**35.71**	**25.00**	**58.82**	23.53	**17.65**	**26.09**	4.35	0.00
BLIP$_{ITCCOCO}$ (ViT-L/16)	**46.43**	14.29	14.29	17.65	5.88	5.88	13.04	0.00	0.00
BLIP$_{ITMFlickr30k}$ (ViT-L/16)	**39.29**	**28.57**	**25.00**	**47.06**	11.76	5.88	**30.43**	8.70	4.35
BLIP$_{ITCFlickr30k}$ (ViT-L/16)	**39.29**	14.29	14.29	**47.06**	5.88	5.88	21.74	13.04	13.04

Table 5.6: The results by visual tag. Results above chance are in **bold**.

5.1.2 Ours

5.2 Results By Linguistic Tag

5.2.1 Baseline

5.2.2 Ours

5.3 Results By Visual Tag

5.3.1 Baseline

5.3.2 Ours

6 Discussion

7 Conclusions

Appendix

Bibliography