Model	MSCOCO (5K test set)						Flickr30K (1K test set)					
	Image → Text			Text → Image			Image → Text			Text → Image		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
FLAVA [1]	42.74	76.76	-	38.38	67.47	-	67.7	94.0	-	65.22	89.38	-
CLIP [2]	58.4	81.5	88.1	37.8	62.4	72.2	88.0	98.7	99.4	68.7	90.6	95.2
VLMo	74.8	93.1	96.9	57.2	82.6	89.8	92.3	99.4	99.9	79.3	95.7	97.8
BEiT-3	84.8	96.5	98.3	67.2	87.7	92.8	98	100	100	90.3	98.7	99.5
EMKUM (ours)	54.6	82.3	89.76	36.91	66.86	78.38	69.0	92.9	96.2	53.84	81.06	88.58

Table 1:

Bibliography

[1] A. Singh *et al.*, "FLAVA: A foundational language and vision alignment model," *CoRR*, 2021, [Online]. Available: https://arxiv.org/abs/2112.04482

[2] A. Radford *et al.*, "Learning transferable visual models from natural language supervision," in *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, M. Meila and T. Zhang, Eds., in Proceedings of Machine Learning Research, vol. 139. PMLR, 2021, pp. 8748–8763.