Model	# Pre-train Images	SNLI-VE		Flickr30K (1K test set)						
		dev	test	IR@1	IR@5	IR@10	TR@1	TR@5	TR@10	RSUM
Pre-trained on More Data										
ALIGN$_{\text{BASE}}$ [27]	1.8B	-	-	84.9	97.4	98.6	95.3	99.8	100.0	576.0
ALBEF$_{\text{BASE}}$ [35]	14M	80.80	80.91	85.6	97.5	98.9	95.9	99.8	100.0	577.7
Pre-trained on CC, SBU, MSCOCO and VG datasets										
ViLT$_{\text{BASE}}$ [29]	4M	-	-	64.4	88.7	93.8	83.5	96.7	98.6	525.7
UNITER$_{\text{LARGE}}$ [5]	4M	79.30	79.38	75.6	94.1	96.8	87.3	98.0	99.2	550.9
VILLA$_{\text{LARGE}}$ [16]	4M	80.18	80.02	76.3	94.2	96.8	87.9	97.5	98.8	551.5
UNIMO$_{\text{LARGE}}$ [37]	4M	**81.11**	80.63	78.0	94.2	97.1	89.4	98.9	99.8	557.5
ALBEF$_{\text{BASE}}$ [35]	4M	80.14	80.30	82.8	96.7	98.4	94.3	99.4	99.8	571.4
METER-CLIP-ViT$_{\text{BASE}}$ [13]	4M	80.86	**81.19**	82.2	96.3	98.4	94.3	99.60	99.9	570.7
BRIDGE-TOWER$_{\text{BASE}}$ (Ours)	**4M**	**81.11**	**81.19**	**85.8**	**97.6**	**98.9**	**94.7**	**99.61**	**100.0**	**576.6**