Model	# Pre-train Images	Use VG–QA For Fine-tuning	Test-Dev Overall	Test-Standard Yes/No	Number	Other	Overall
ViLT$_{\text{BASE}}$ [29]	4M	✗	71.26	-	-	-	-
UNITER$_{\text{BASE}}$ [5]	4M	✓	72.70	-	-	-	72.91
VILLA$_{\text{BASE}}$ [16]	4M	✓	73.59	-	-	-	73.67
UNIMO$_{\text{BASE}}$ [37]	4M	✗	73.79	-	-	-	74.02
ALBEF$_{\text{BASE}}$ [35]	4M	✓	74.54	-	-	-	74.70
ALBEF$_{\text{BASE}}$ [35]	14M	✓	75.84	-	-	-	76.04
VinVL$_{\text{BASE}}$ [91]	5.7M	✗	75.95	-	-	-	76.12
VLMO$_{\text{BASE}}$ [77]	4M	✗	76.64	-	-	-	76.89
BLIP$_{\text{BASE}}$ [34]	14M	✓	77.54	-	-	-	77.62
METER-CLIP-ViT$_{\text{BASE}}$ [13]	4M	✗	77.68	92.49	58.07	69.20	77.64
OFA$_{\text{BASE}}$ [75]	54M	✓	77.98	-	-	-	78.07
SimVLM$_{\text{BASE}}$ [78]	1.8B	✗	77.87	-	-	-	78.14
BLIP$_{\text{BASE}}$ [34]	129M	✓	78.24	-	-	-	78.17
BLIP$_{\text{BASE-CapFilt-L}}$ [34]	129M	✓	78.25	-	-	-	78.32
BRIDGE-TOWER$_{\text{BASE}}$ (Ours)	**4M**	✗	**78.66**	**92.92**	**60.69**	**70.51**	**78.73**
BRIDGE-TOWER$_{\text{BASE}}$ (Ours)	**4M**	✓	**79.10**	**93.06**	**62.19**	**70.69**	**79.04**
UNITER$_{\text{LARGE}}$ [5]	4M	✓	73.82	-	-	-	74.02
VILLA$_{\text{LARGE}}$ [16]	4M	✓	74.69	-	-	-	74.87
UNIMO$_{\text{LARGE}}$ [37]	4M	✗	75.06	-	-	-	75.27
VinVL$_{\text{LARGE}}$ [91]	5.7M	✗	76.52	92.04	61.50	66.68	76.63
SimVLM$_{\text{LARGE}}$ [78]	1.8B	✗	79.32	-	-	-	79.56
VLMO$_{\text{LARGE}}$ [77]	4M	✗	79.94	-	-	-	79.98
SimVLM$_{\text{HUGE}}$ [78]	1.8B	✗	80.03	93.29	66.54	72.23	80.34
METER-CoSwin$_{\text{HUGE}}$ [13]	14M	✗	80.33	94.25	64.37	72.30	80.54
OFA$_{\text{LARGE}}$ [75]	54M	✓	80.43	93.32	**67.31**	72.71	80.67
BRIDGE-TOWER$_{\text{LARGE}}$ (Ours)	**4M**	✗	**81.25**	**94.69**	64.58	**73.16**	**81.15**
BRIDGE-TOWER$_{\text{LARGE}}$ (Ours)	**4M**	✓	**81.52**	**94.80**	66.01	**73.45**	**81.49**