VL-BEIT

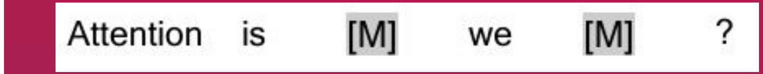
Parsa Sharifi

- vision-language foundation model
- bidirectional multimodal Transformer
- visual question answering, visual reasoning, and image-text retrieval

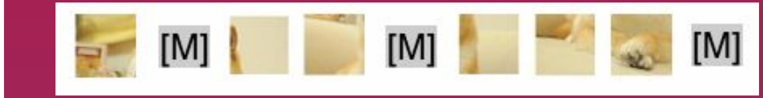
- Backbone model
- Pre-training

- Mixture Of-Modality-Experts (MOME) Transformer
- Multi-head self-attention layer
- A feed-forward expert layer(pool)
- Hard Routing Mechanism

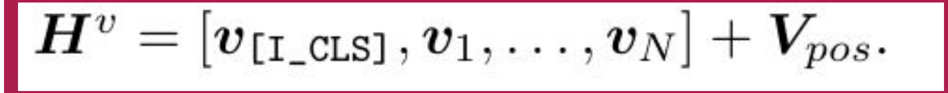
- Pretraining Tasks
  - Masked Language Modeling
  - Masked Image Modeling
  - Masked Vision-Language Modeling



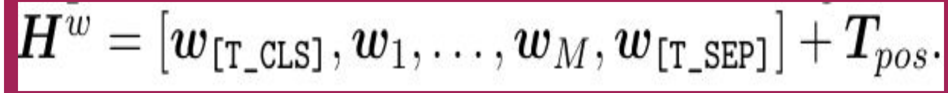
Attention is [M] we [M] ?



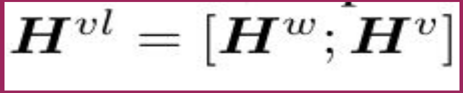
- Input Representations



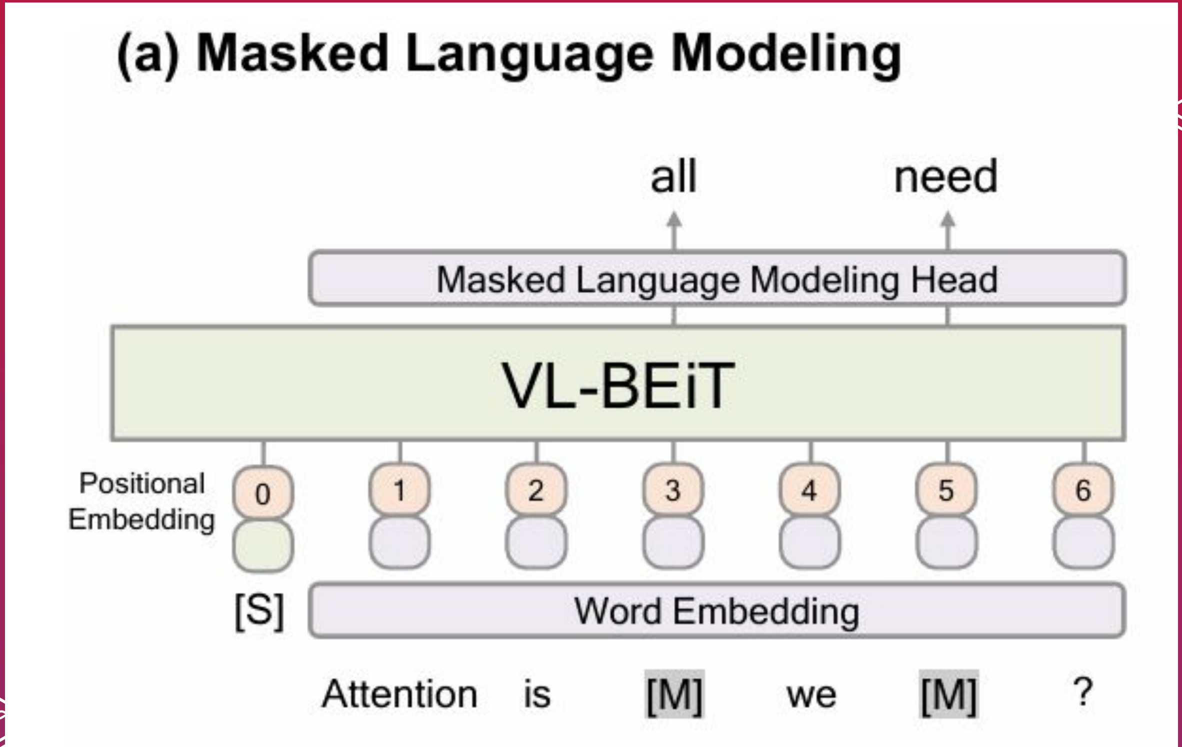
$$\boldsymbol{H}^{v} = \left[\boldsymbol{v}_{\texttt{[I\_CLS]}}, \boldsymbol{v}_{1}, \ldots, \boldsymbol{v}_{N}\right] + \boldsymbol{V}_{pos}.$$



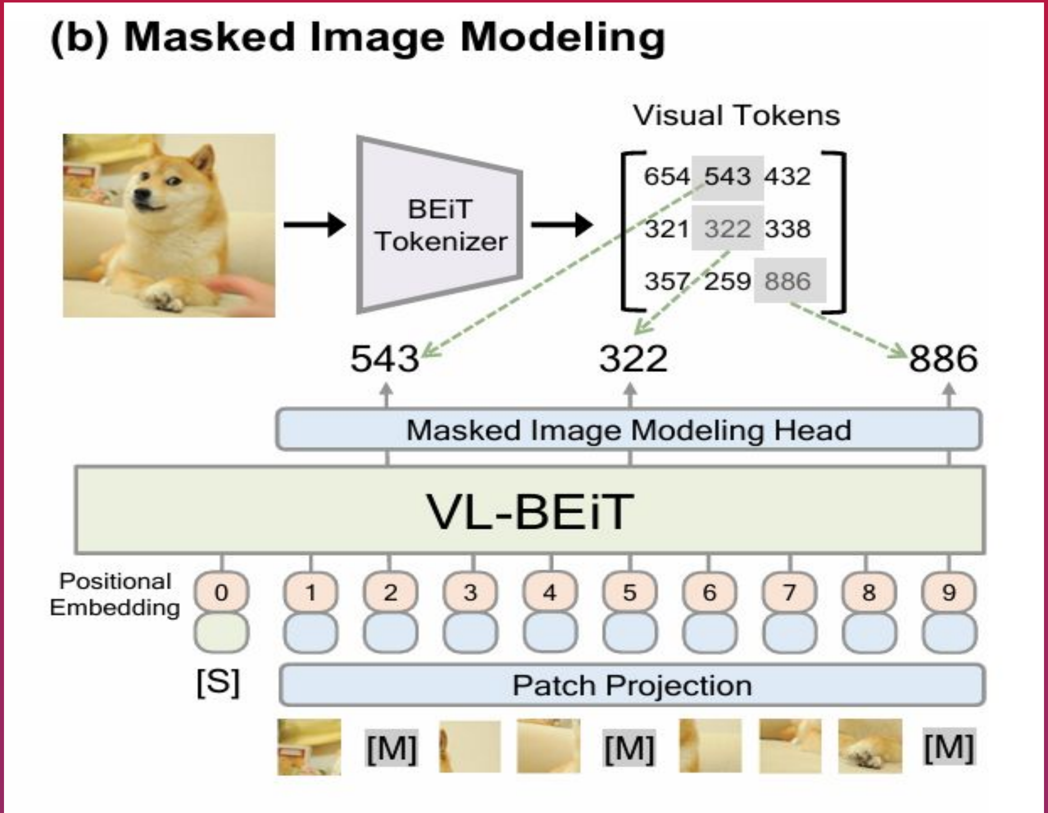
$$\boldsymbol{H}^{w} = \left[\boldsymbol{w}_{\texttt{[T\_CLS]}}, \boldsymbol{w}_{1}, \ldots, \boldsymbol{w}_{M}, \boldsymbol{w}_{\texttt{[T\_SEP]}}\right] + \boldsymbol{T}_{pos}.$$

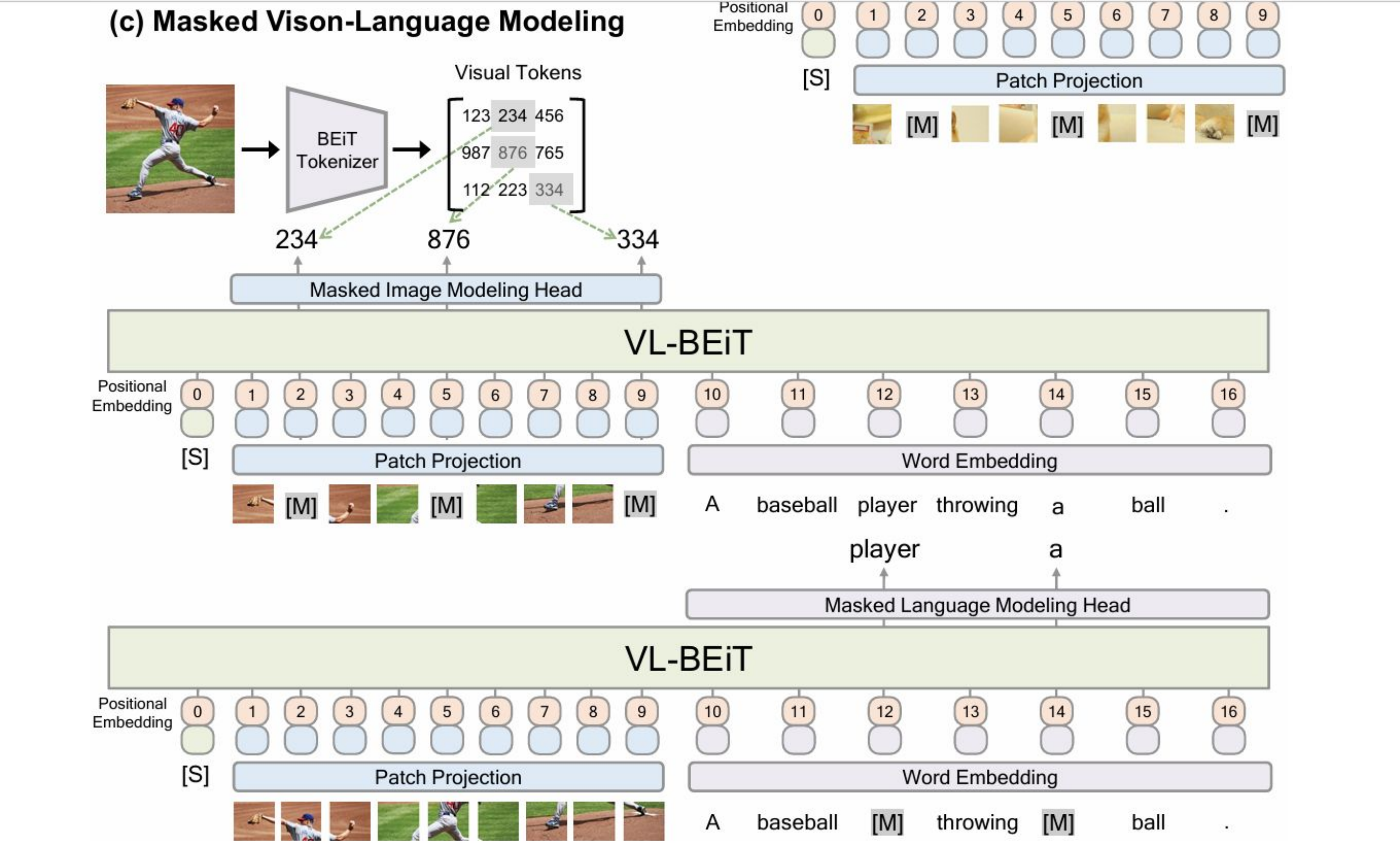


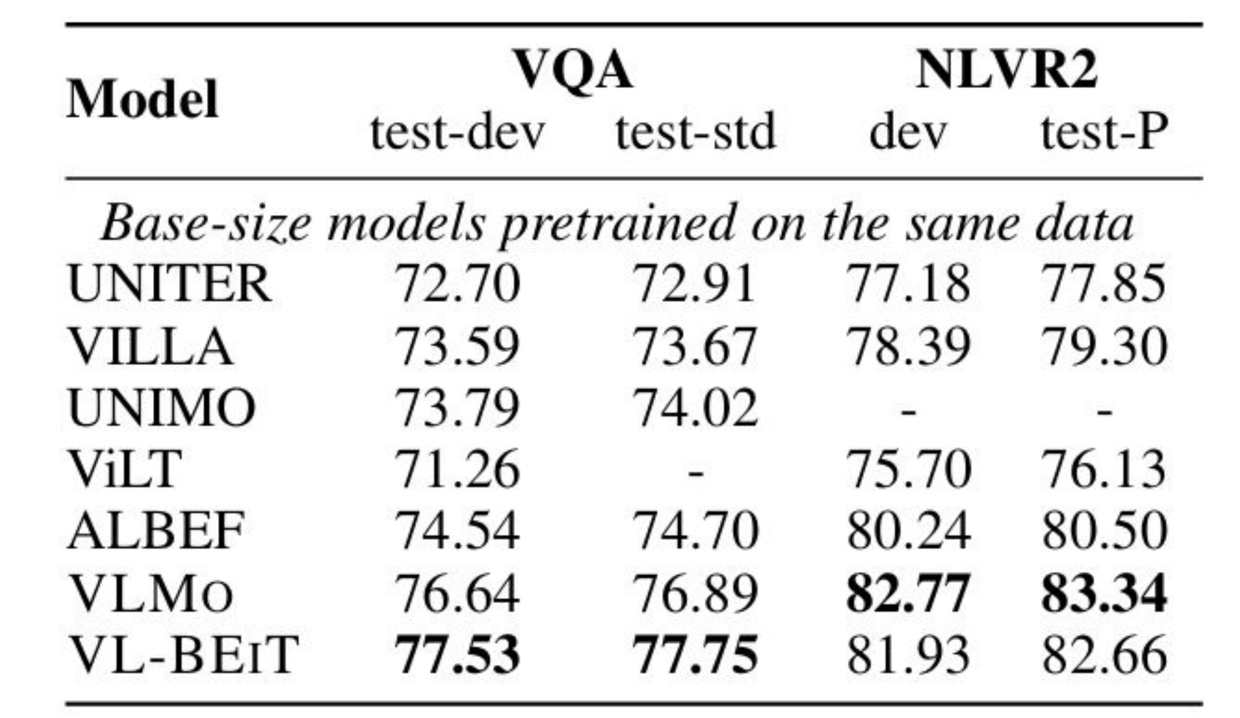
$$\boldsymbol{H}^{vl} = [\boldsymbol{H}^{w}; \boldsymbol{H}^{v}]$$



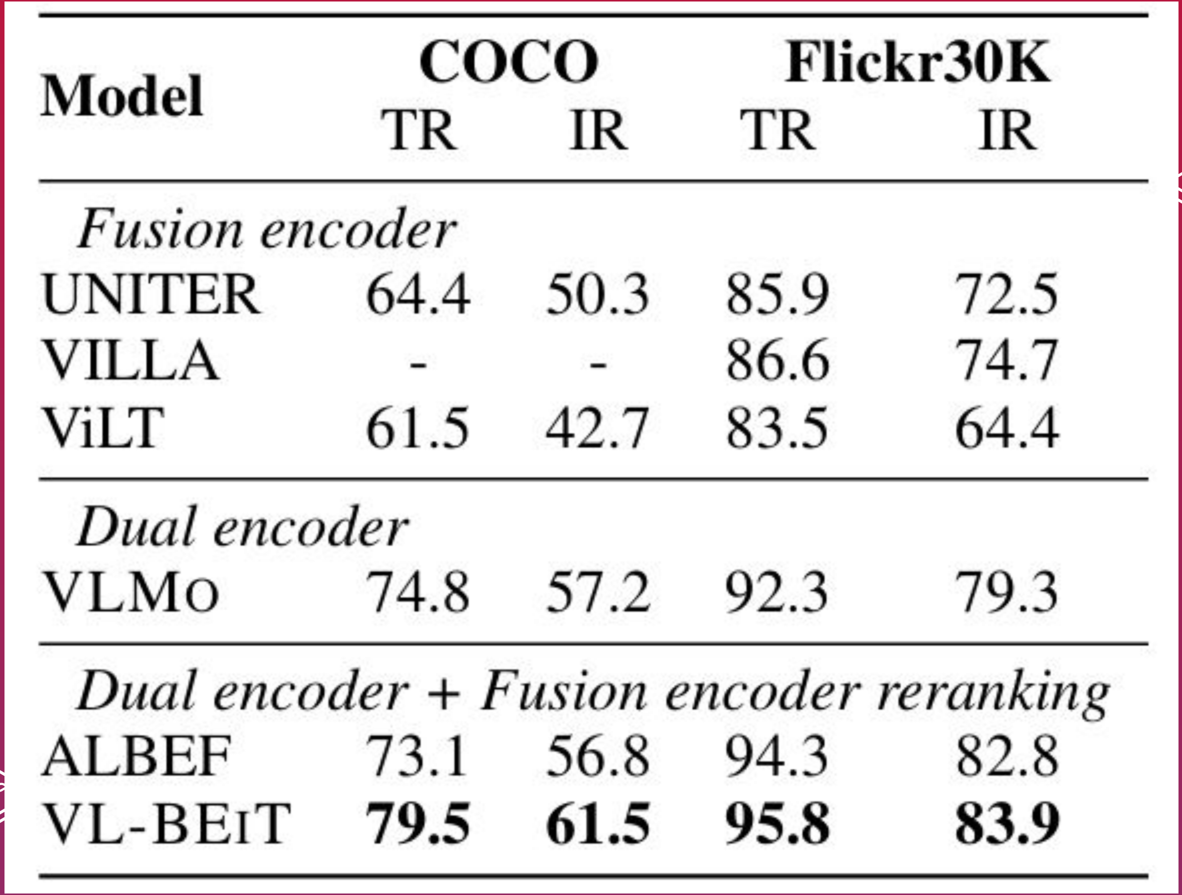
(a) Masked Language Modeling


(b) Masked Image Modeling
Visual Tokens
BEiT Tokenizer
654 543 432
321 322 338
357 259 886
543
322
886
Masked Image Modeling Head
VL-BEiT
Positional Embedding
0
1
2
3
4
5
6
7
8
9
[S]
Patch Projection
[M]
[M]
[M]


(c) Masked Vison-Language Modeling
Visual Tokens
BEiT Tokenizer
123 234 456
987 876 765
112 223 334
234
876
334
Masked Image Modeling Head
VL-BEiT
Positional Embedding
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
[S]
Patch Projection
Word Embedding
[M] [M] [M]
A baseball player throwing a ball .
player
a
Masked Language Modeling Head
VL-BEiT
Positional Embedding
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
[S]
Patch Projection
Word Embedding
A baseball [M] throwing [M] ball .
Positional Embedding
0 1 2 3 4 5 6 7 8 9
[S]
Patch Projection
[M] [M] [M]



Model	VQA		NLVR2	
	test-dev	test-std	dev	test-P
Base-size models pretrained on the same data				
UNITER	72.70	72.91	77.18	77.85
VILLA	73.59	73.67	78.39	79.30
UNIMO	73.79	74.02	-	-
ViLT	71.26	-	75.70	76.13
ALBEF	74.54	74.70	80.24	80.50
VLMo	76.64	76.89	**82.77**	**83.34**
VL-BEiT	**77.53**	**77.75**	81.93	82.66



Model	COCO		Flickr30K	
	TR	IR	TR	IR
Fusion encoder				
UNITER	64.4	50.3	85.9	72.5
VILLA	-	-	86.6	74.7
ViLT	61.5	42.7	83.5	64.4
Dual encoder				
VLMo	74.8	57.2	92.3	79.3
Dual encoder + Fusion encoder reranking				
ALBEF	73.1	56.8	94.3	82.8
VL-BEiT	**79.5**	**61.5**	**95.8**	**83.9**