Vision-Language Pre-Training with Triple Contrastive Learning

Jinyu Yang[1], Jiali Duan[2], Son Tran[2], Yi Xu[2], Sampath Chanda[2], Liqun Chen[2], Belinda Zeng[2],
Trishul Chilimbi[2], and Junzhou Huang[1]
[1]University Of Texas at Arlington, [2]Amazon

CVPR 2022

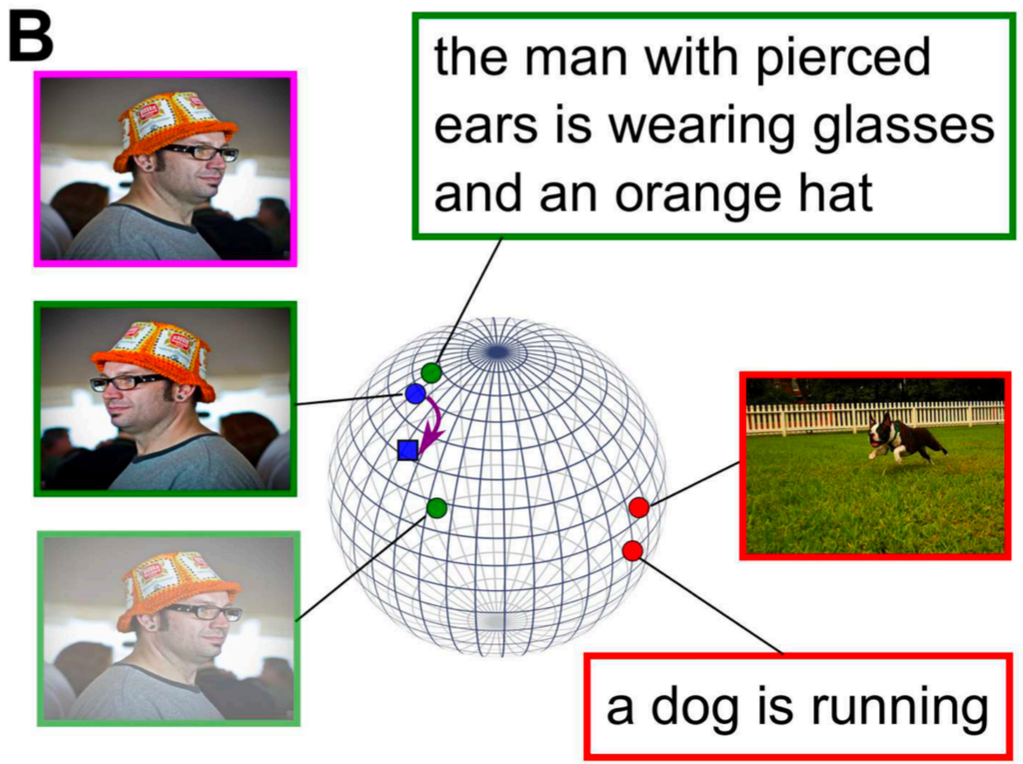
Vision-Language Representation Learning

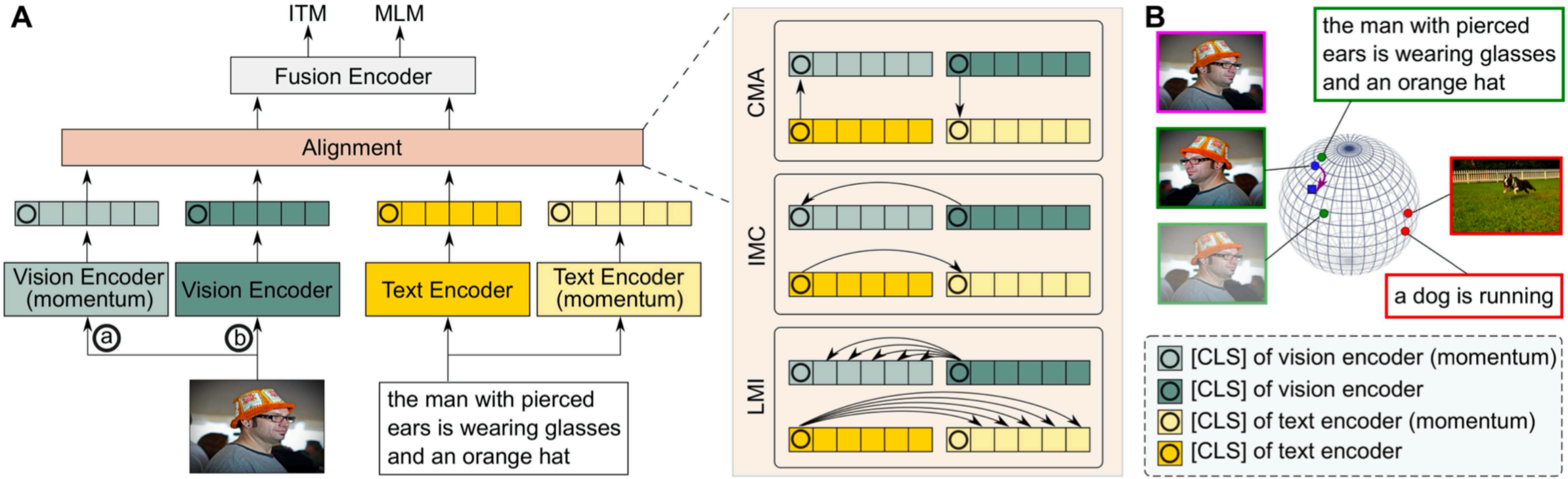
Pre-training visual and text models in a shared feature space

Training data: image-text pair

Transferable to downstream vision-language tasks:

- Visual question answering
- Visual text retrieval



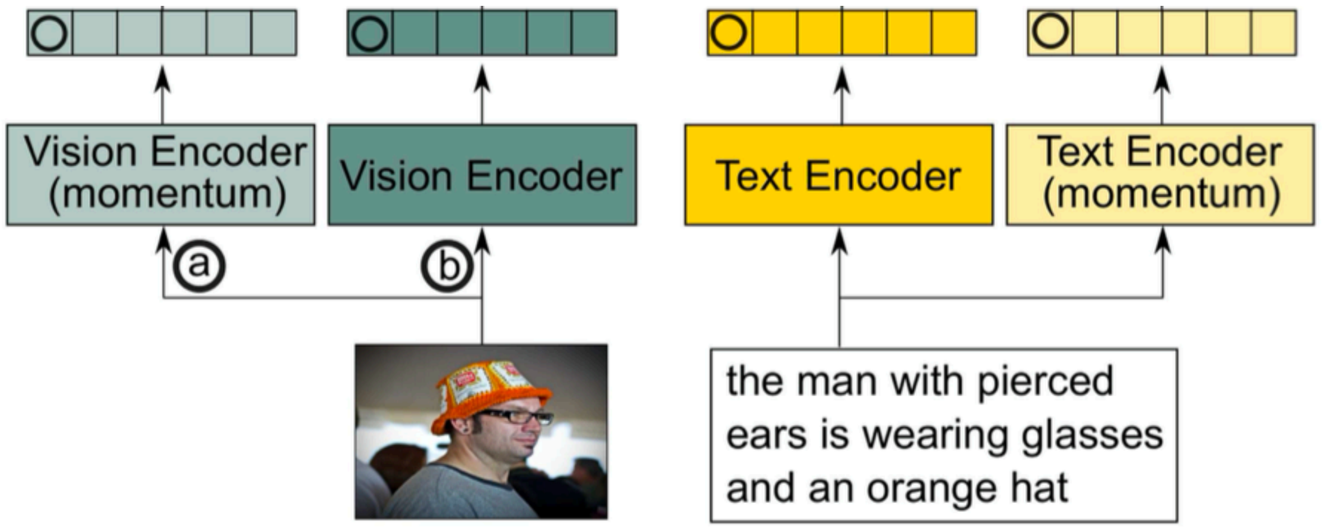


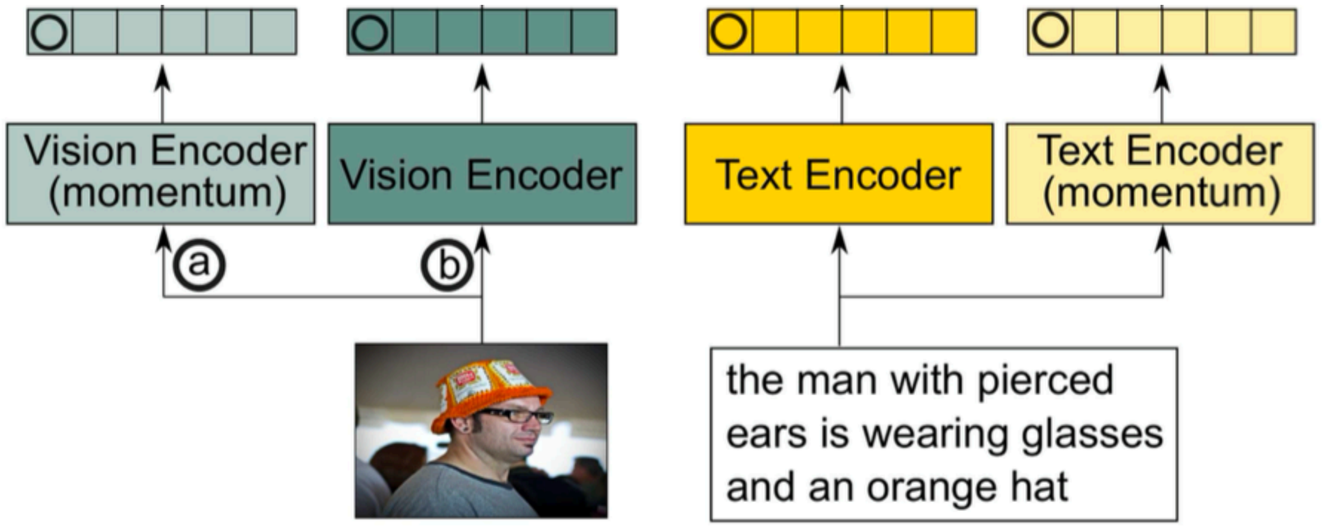
Contrastive learning:

Cross-modal alignment (CMA)
Intra modal contrastive (IMC)
Local MI maximization (LMI)

Predictive learning:

Image-text matching (ITM)
Masked language modeling (MLM)
Masked Image modeling (MLM)


Vision Encoder (momentum)
Vision Encoder
Text Encoder
Text Encoder (momentum)
a
b
the man with pierced ears is wearing glasses and an orange hat



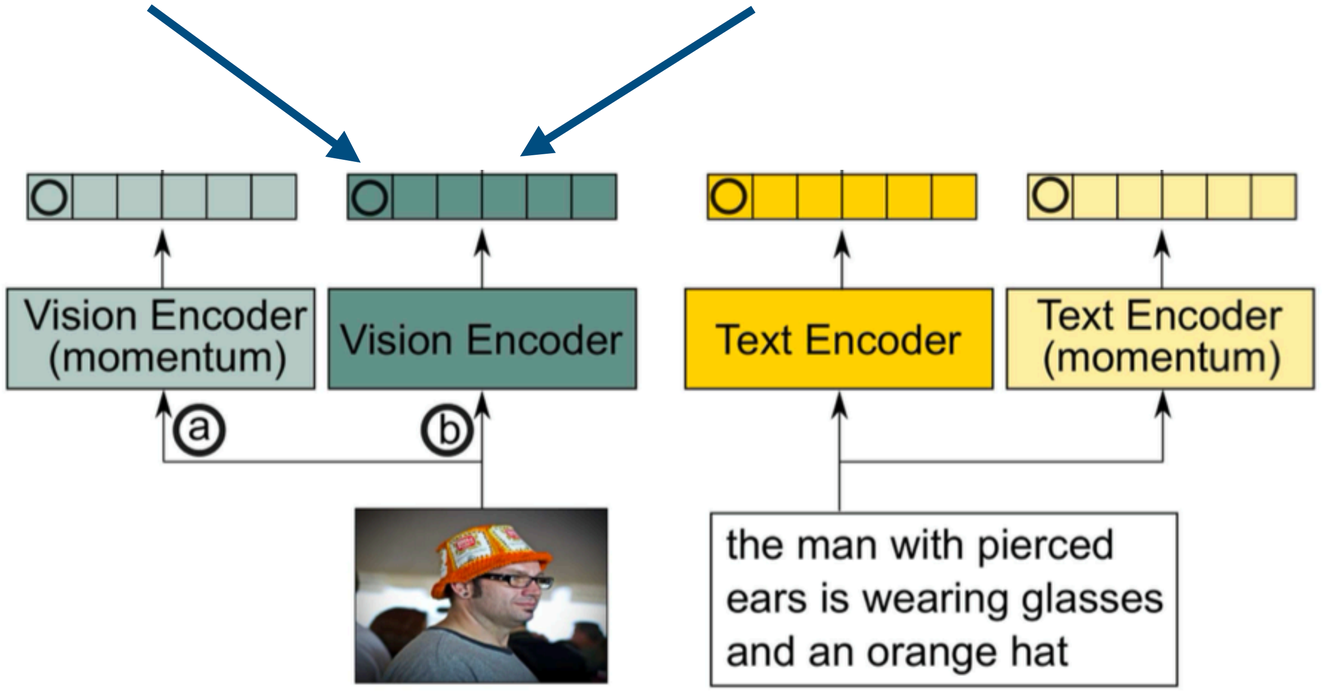
Augmentations

Visual encoder: ViT
Text encoder: BERT

Exponential moving average

$$\theta_{\hat{g}} = m\theta_{\hat{g}} + (1 - m)\theta_{g}$$

Global token [CLS] + Local tokens



Augmentations

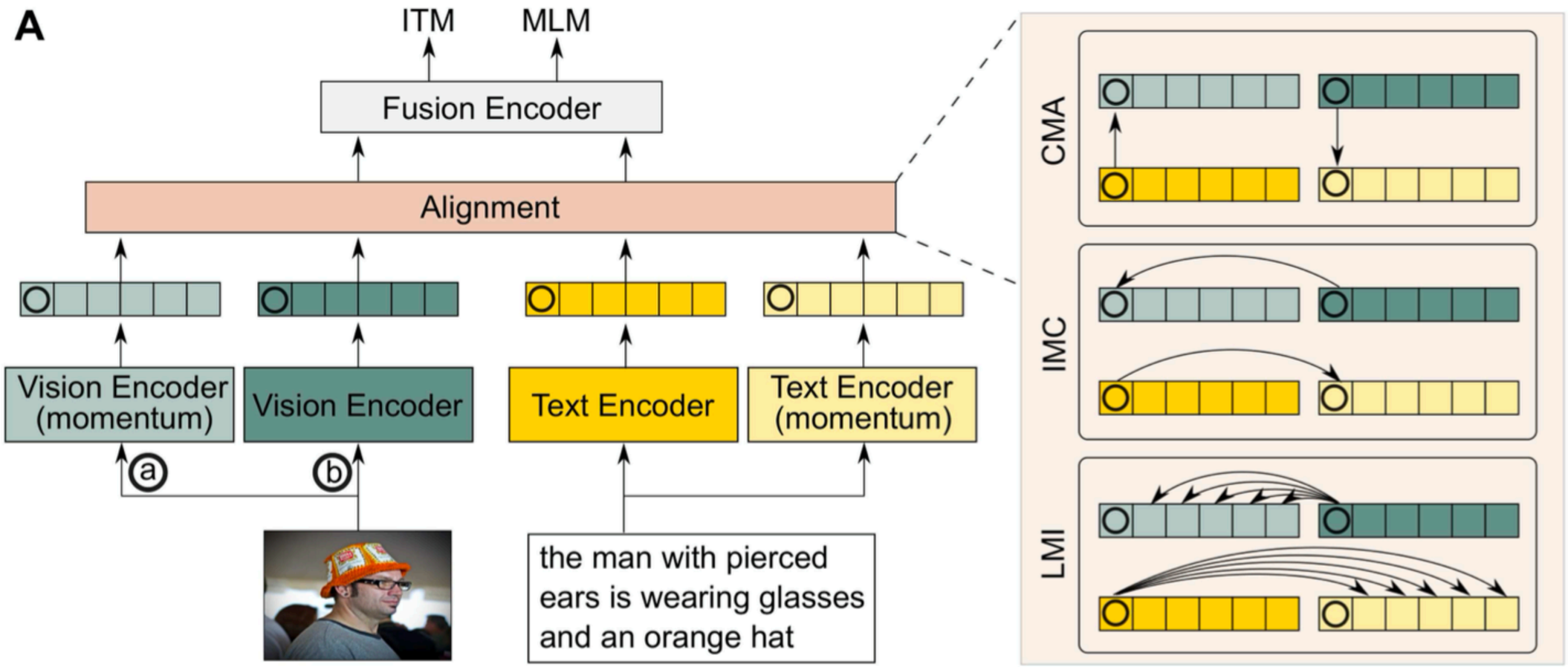
Visual encoder: ViT
Text encoder: BERT

Exponential moving average

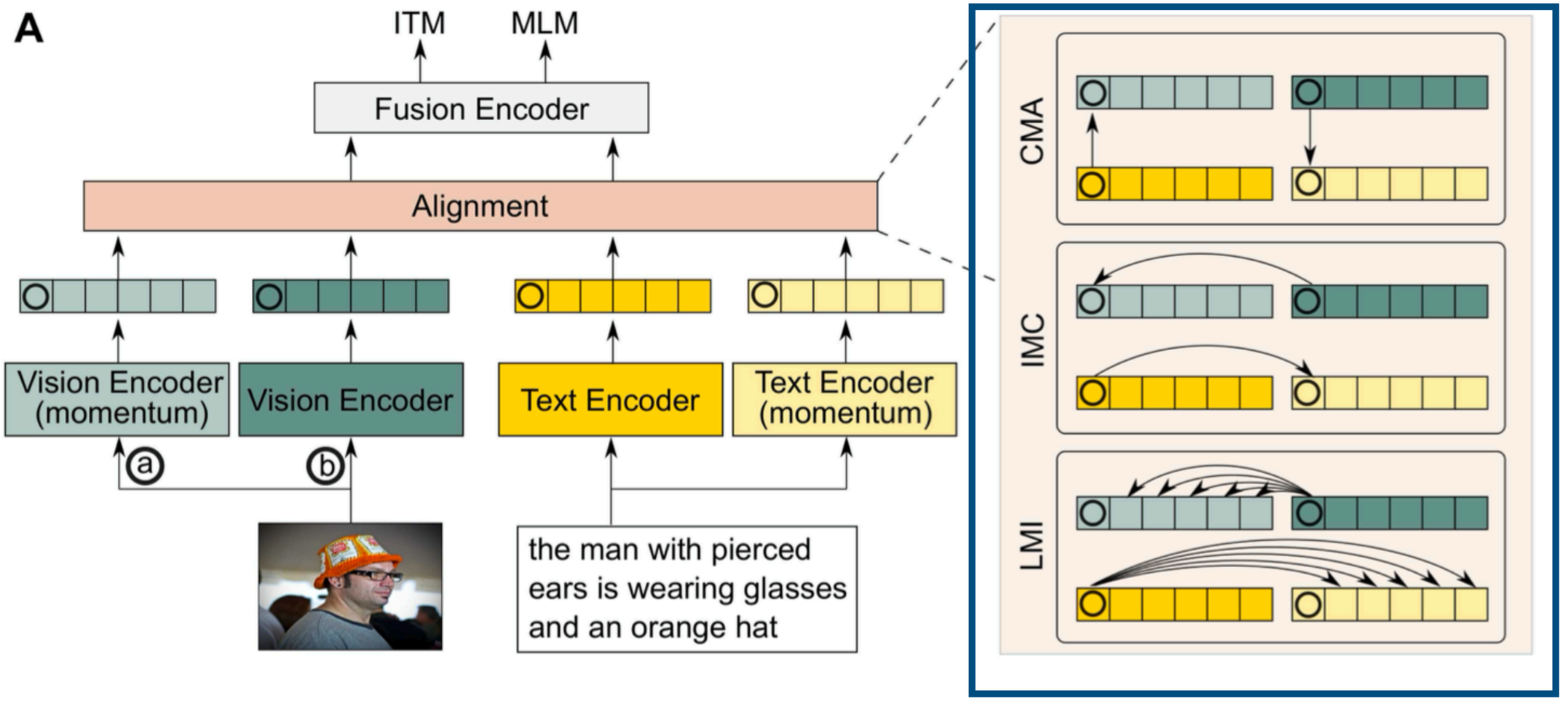
$$\theta_{\hat{g}} = m\theta_{\hat{g}} + (1 - m)\theta_{g}$$

Predictive learning

Contrastive learning



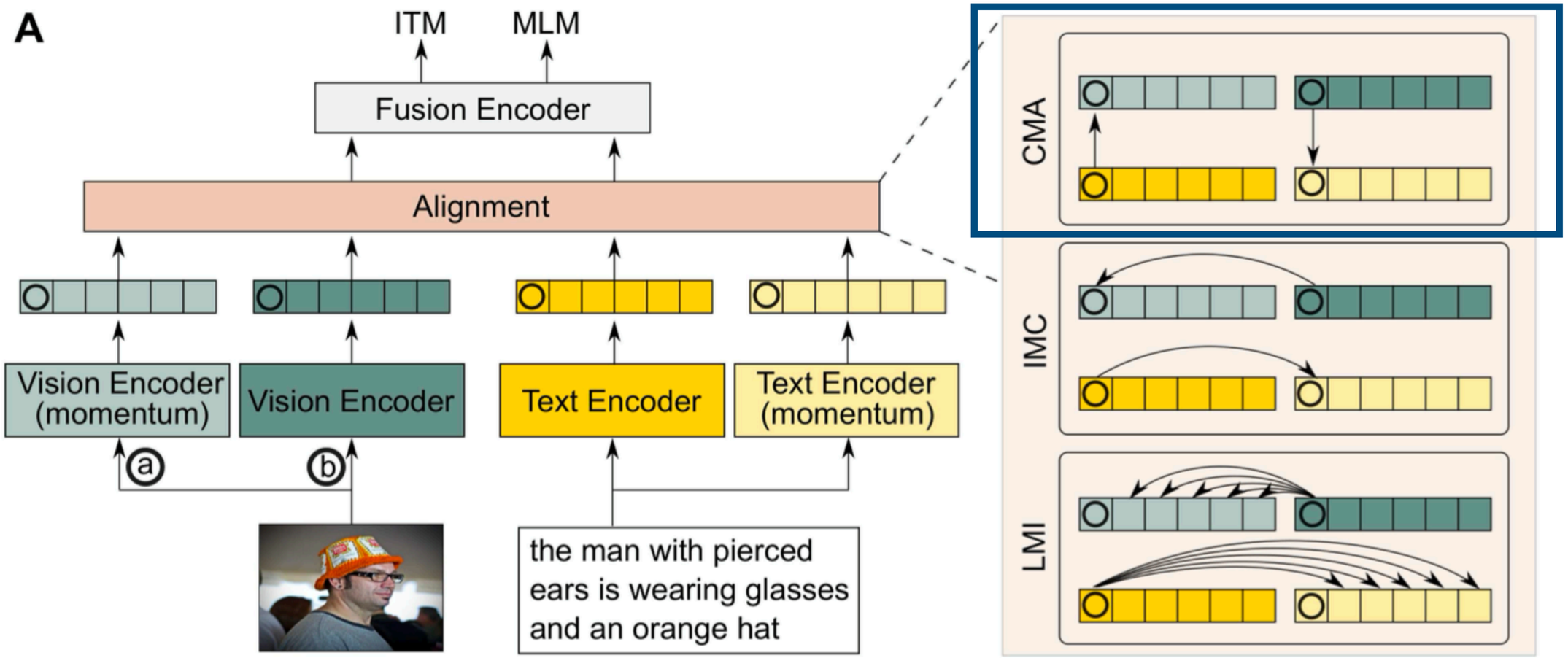
Contrastive learning



*Maximize feature similarities between **positive pairs***

*Minimize feature similarities between **negative pairs***

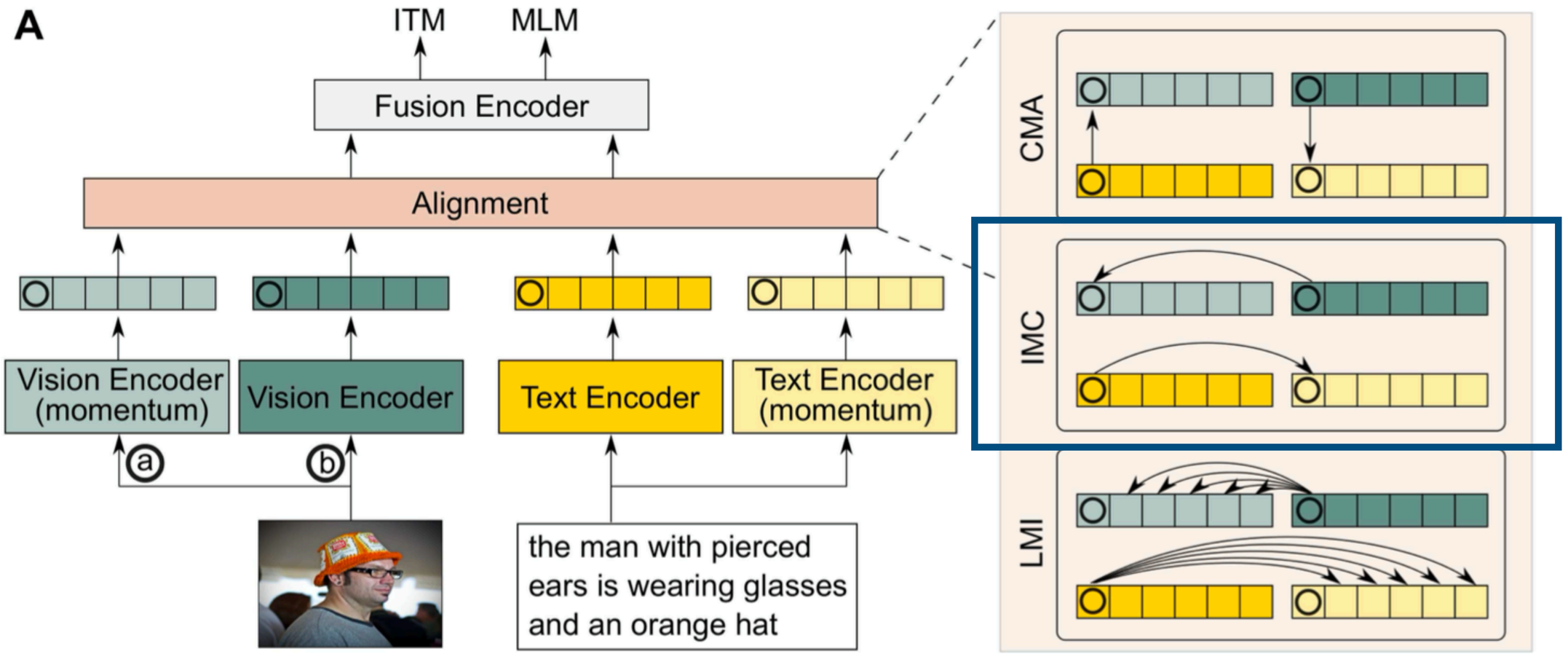
Cross-modal alignment



Positive pair: *The global features of matched image-text pair*

Negative pair: *The global features of random image-text pair*

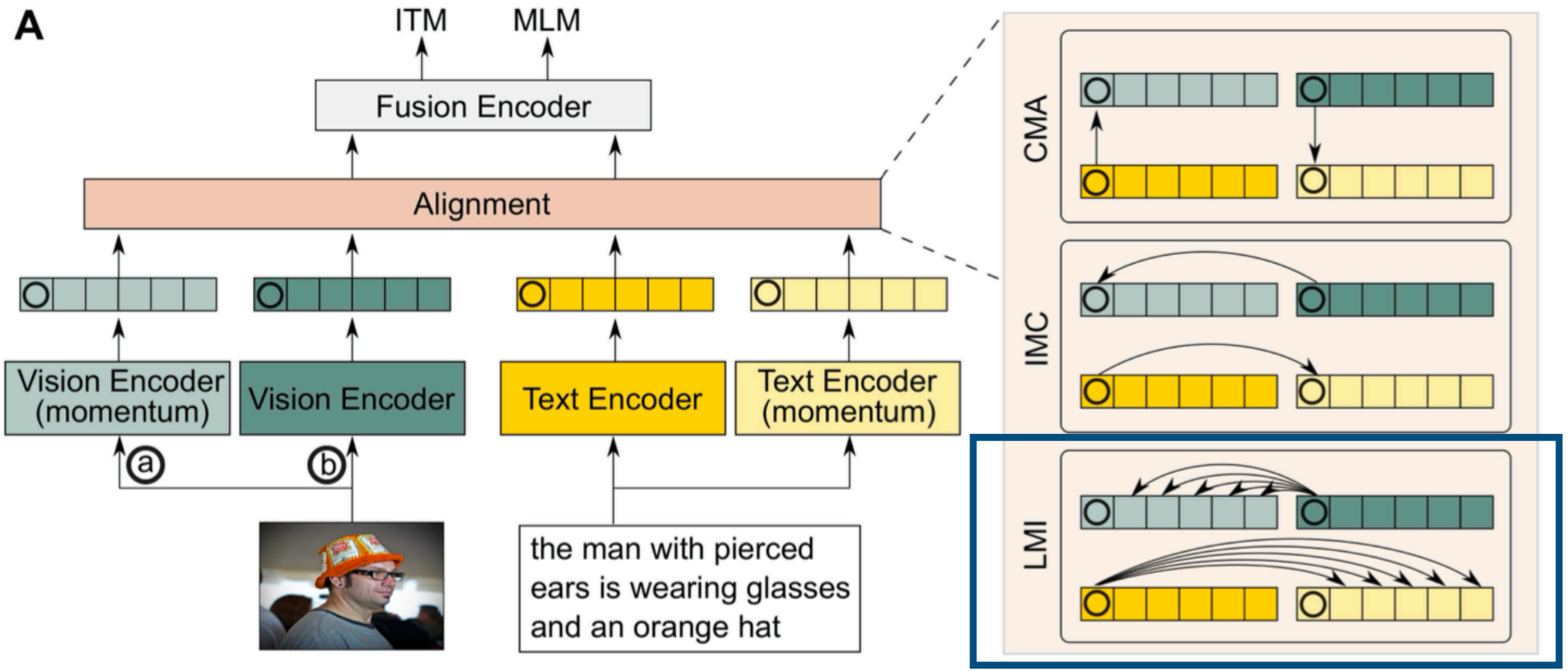
Intra modal contrastive (IMC)



Positive pair: *The global features of augmentations of the same image (or text)*

Negative pair: *The global features of augmentations of different images (or texts)*

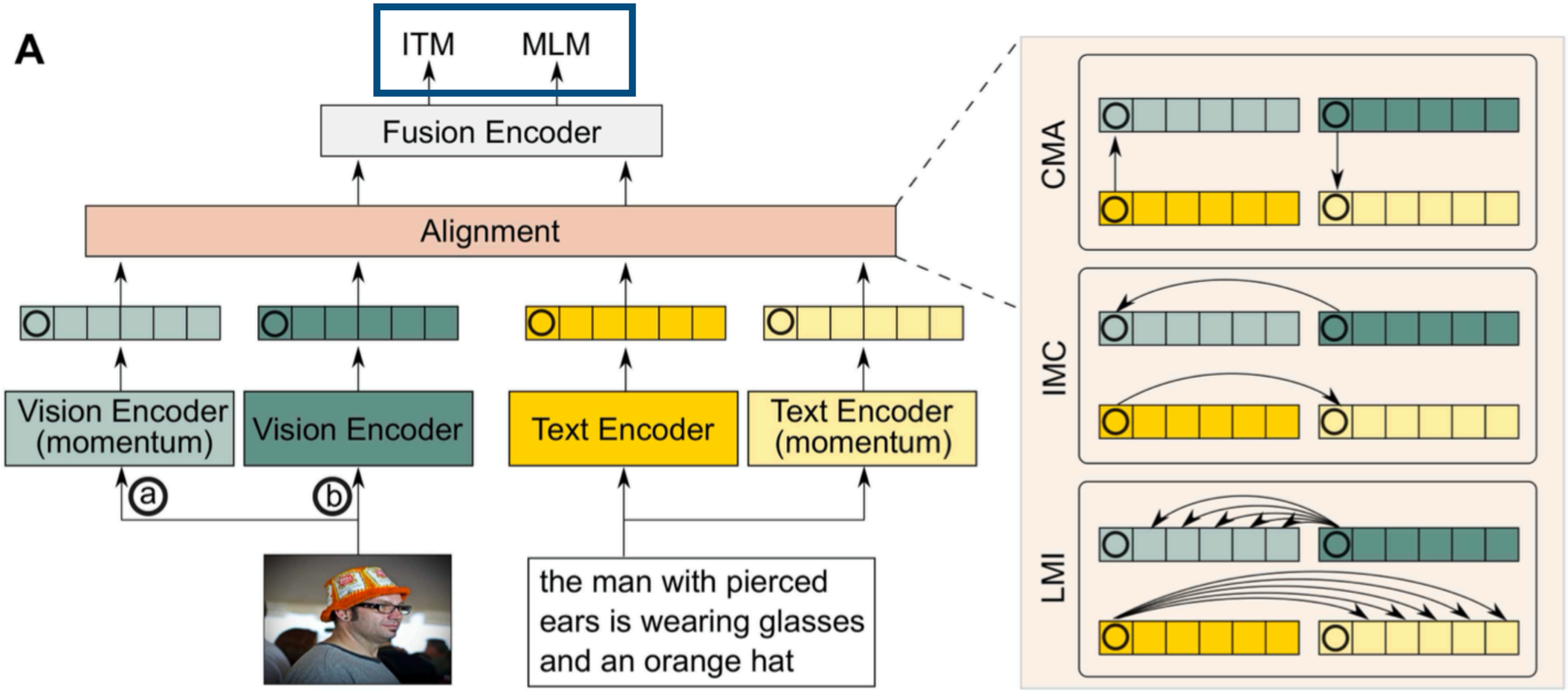
Local mutual information maximization (LMI)



Positive pair: *The global and local features of the same image (or texts)*

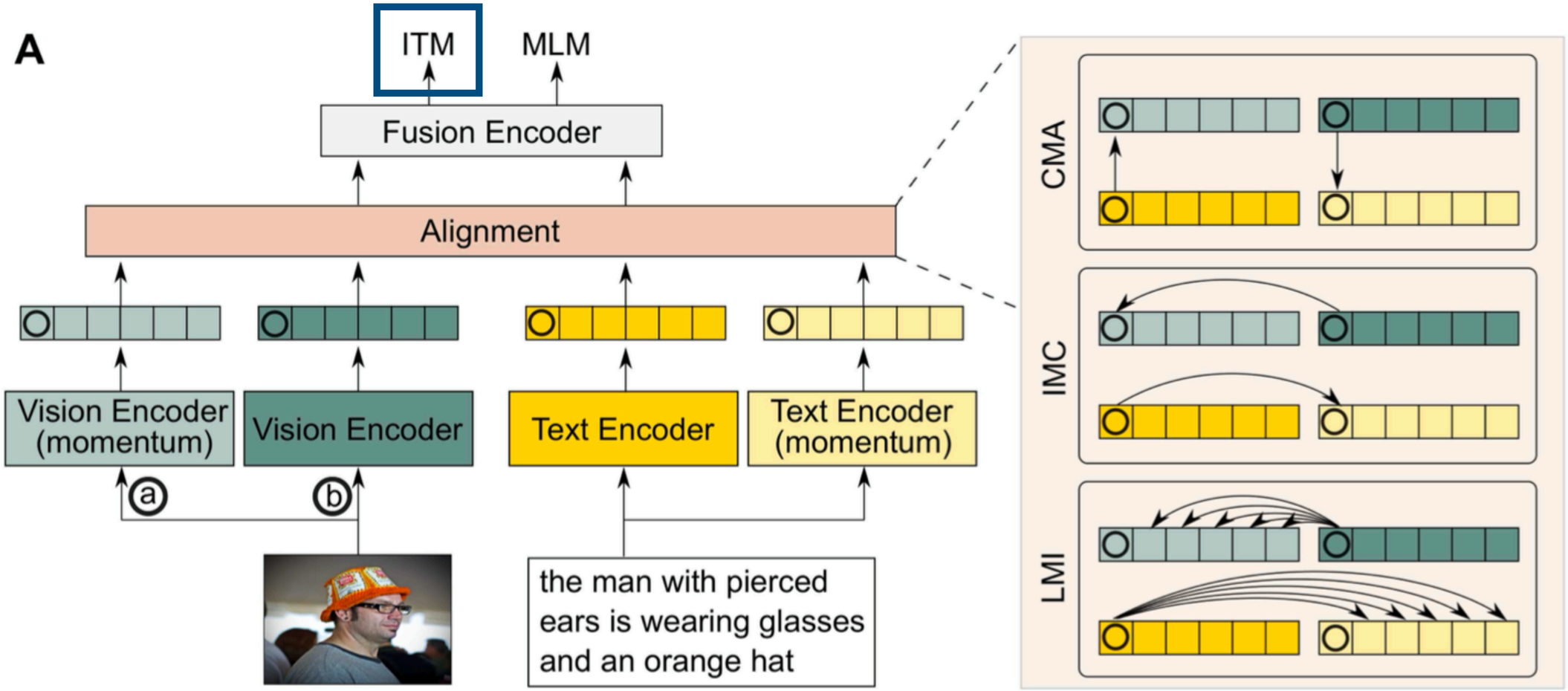
Negative pair: *The global and local features of different image (or texts)*

Predictive learning



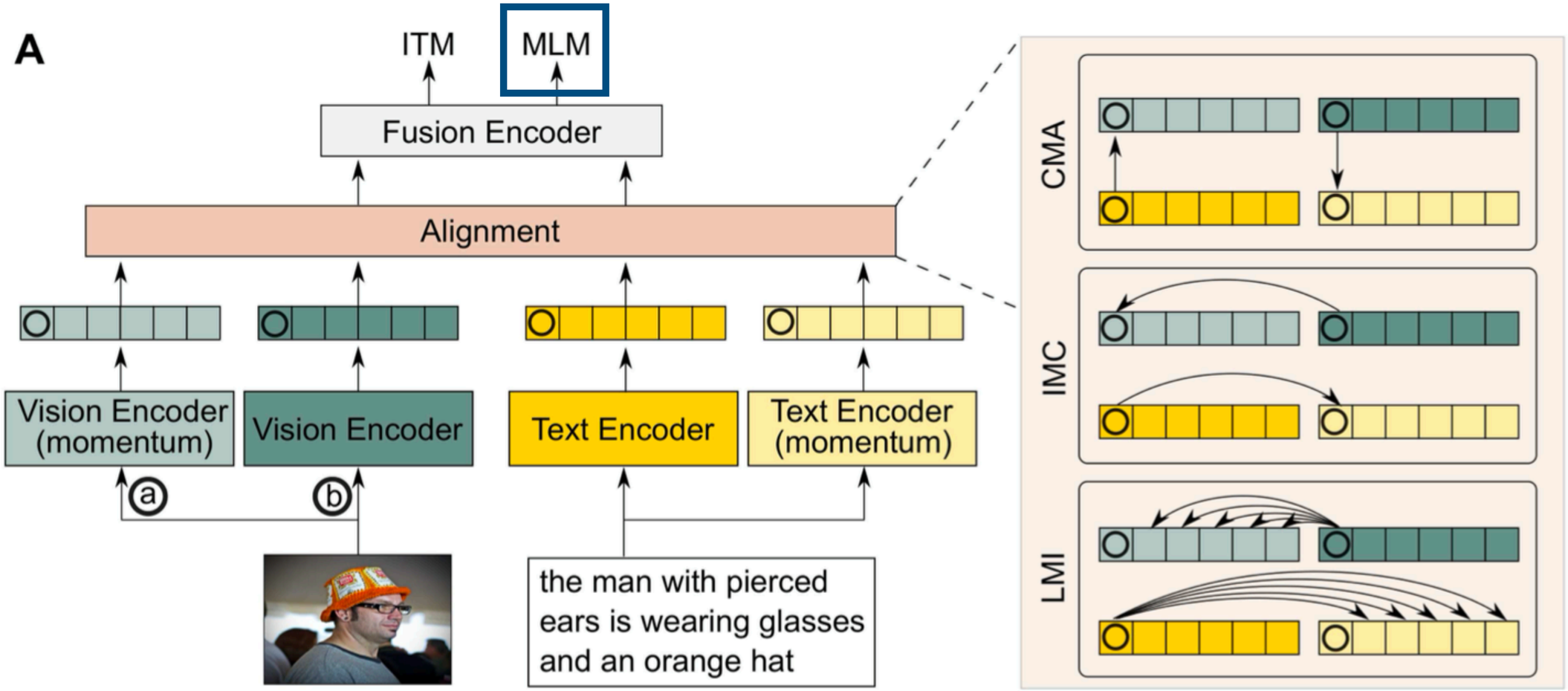
Design proxy tasks

Image-text matching (ITM)



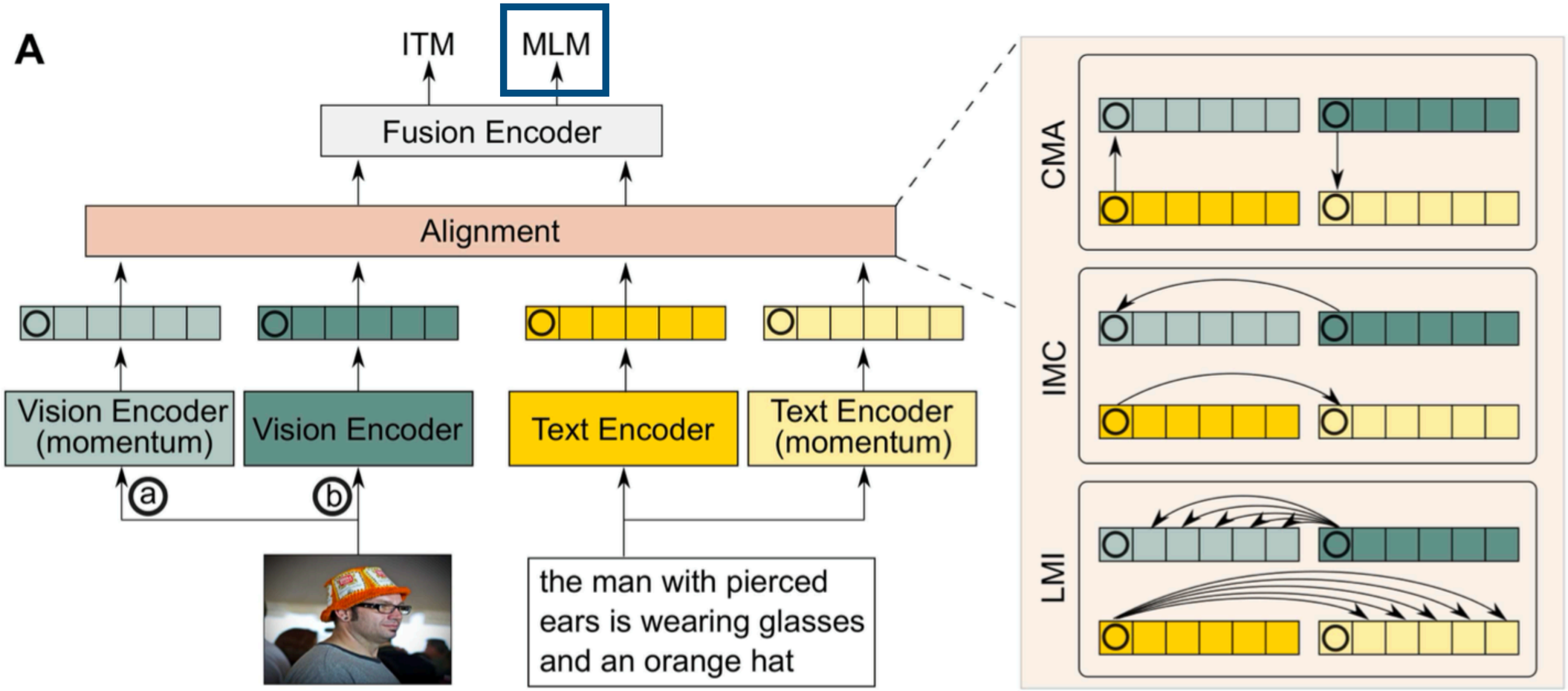
***Classification**: Matched or Not*

Masked language modeling (MLM)



Mask some <u>text</u> tokens and ***reconstruct***

Masked image modeling (MLM)



Mask some <u>image</u> tokens and ***reconstruct***

(Not used)

Experiments

Module	Zero-Shot				Fine-Tune			
	MSCOCO		Flickr30K		MSCOCO		Flickr30K	
	TR	IR	TR	IR	TR	IR	TR	IR
CMA+ITM+MLM	68.7*	50.1*	90.5	76.8	73.1	56.8	94.3	82.8
+IMC (w/o aug)	71.1	52.2	92.0	78.6	75.0	58.6	94.5	82.9
+IMC	71.4	53.3	92.1	78.9	75.6	58.8	95.1	83.1
+IMC+LMI (**Ours**)	71.4	53.5	93.0	79.6	75.6	59.0	94.9	84.0

Ablation studies on image-text retrieval tasks

TR: Text Retrieval
IR: Image retrieval

Baseline:
Cross-modal alignment (CMA)
Image-text matching (ITM)
Masked language modeling (MLM)

Useful:
Intra modal contrastive (IMC)

Marginal:
Local MI maximization (LMI)

Experiments

Method	#Images	MSCOCO (5K)						Flickr30K (1K)					
		Text Retrieval			Image Retrieval			Text Retrieval			Image Retrieval		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
ImageBERT [34]	6M	44.0	71.2	80.4	32.3	59.0	70.2	70.7	90.2	94.0	54.3	79.6	87.5
UNITER [7]	4M	✗	✗	✗	✗	✗	✗	80.7	95.7	98.0	66.2	88.4	92.9
ViLT [20]	4M	56.5	82.6	89.6	40.4	70.0	81.1	73.2	93.6	96.5	55.0	82.5	89.8
CLIP [35]	400M	58.4	81.5	88.1	37.8	62.4	72.2	88.0	98.7	99.4	68.7	90.6	95.2
ALBEF [22]	4M	68.7	89.5	94.7	50.1	76.4	84.5	90.5	98.8	**99.7**	76.8	93.7	96.7
Ours	4M	**71.4**	**90.8**	**95.4**	**53.5**	**79.0**	**87.1**	**93.0**	**99.1**	99.6	**79.6**	**95.1**	**97.4**
ALIGN [19]	1.2B	58.6	83.0	89.7	45.6	69.8	78.6	88.6	98.7	99.7	75.7	93.8	96.8

Ablation studies on image-text retrieval (useful)

Method	#Images	VQA		NLVR^2		SNLI-VE	
		test-dev	test-std	dev	test-P	val	test
OSCAR [24]	4M	73.16	73.44	78.07	78.36	✗	✗
UNITER [7]	4M	72.70	72.91	77.18	77.85	78.59	78.28
ViLT [20]	4M	71.26	✗	75.7	76.13	✗	✗
UNIMO [23]	4M	73.29	74.02	✗	✗	80.0	79.1
VILLA [13]	4M	73.59	73.67	78.39	79.30	79.47	79.03
ALBEF [22]	4M	74.54	74.70	80.24	80.50	80.14	**80.30**
Ours	4M	**74.90**	**74.92**	**80.54**	**81.33**	**80.51**	80.29

Ablation studies on VQA, Visual Reasoning (NLVR2), Visual Entailment (SNLI-VE) (marginal)

Conclusion

Contrastive learning:

Cross-modal alignment (CMA)
Intra modal contrastive (IMC)
Local MI maximization (LMI)

Predictive learning:

Image-text matching (ITM)
Masked language modeling (MLM)
Masked Image modeling (MLM)

Good ensemble of existing techniques