CVPR JUNE 19-24 2022
NEW ORLEANS • LOUISIANA

Multi-modal Alignment using Representation Codebook

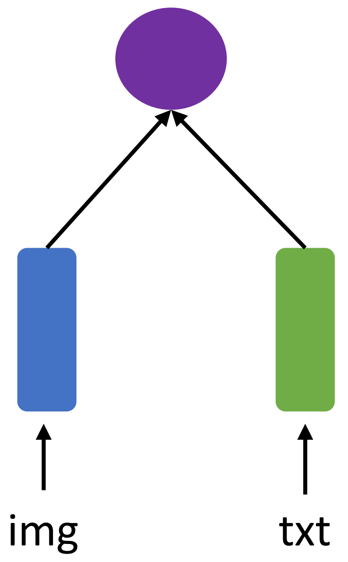




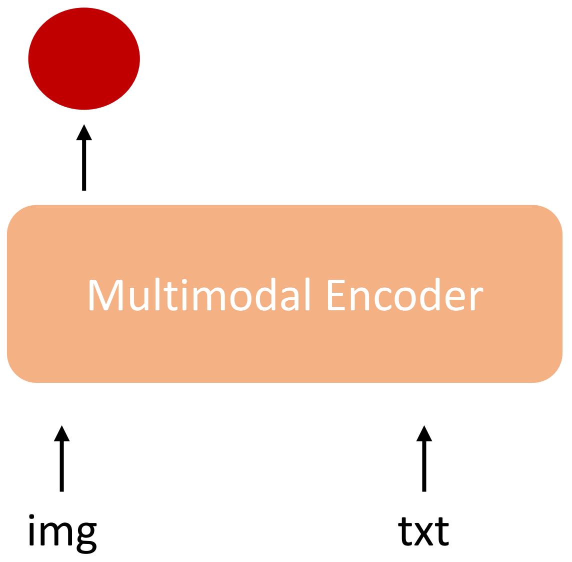


Jiali Duan* Liqun Chen* Son Tran Jinyu Yang Yi Xu Belinda Zeng Trishul Chilimbi

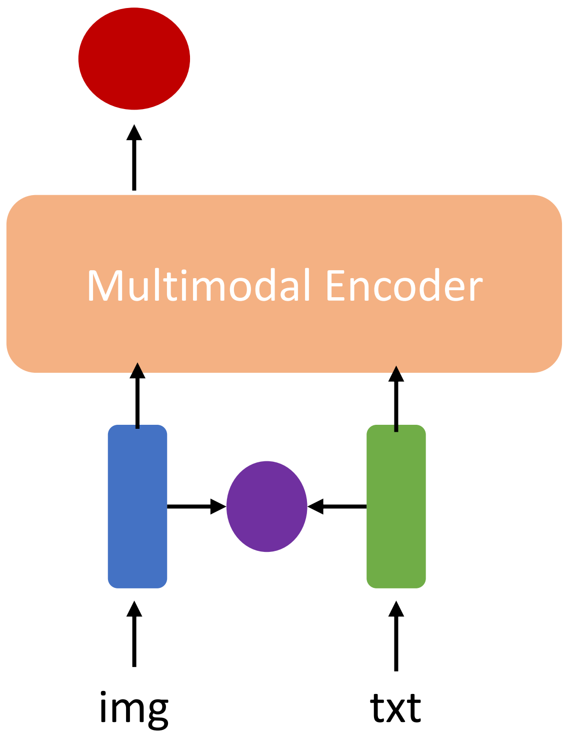
Background: Vision-Language Pretraining



Late fusion: CLIP [1], ALIGN [2]



Early fusion: OSCAR [3], UNITER [4]



Hybrid: ALBEF [5]

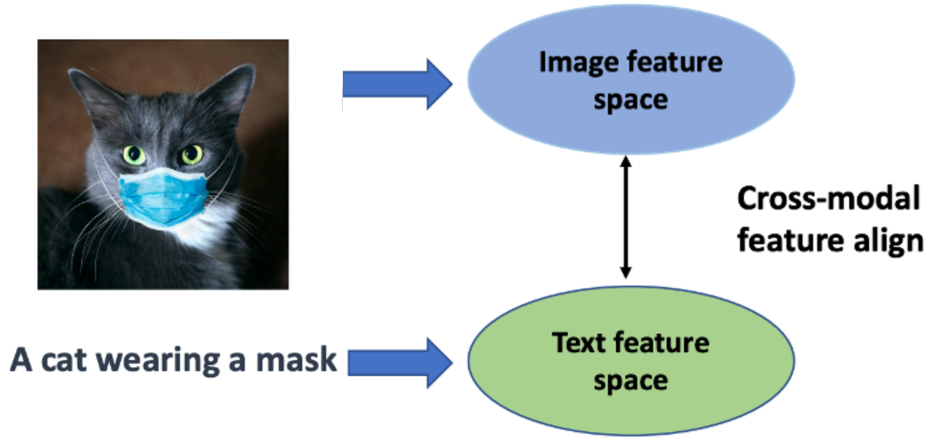
[1] Learning transferable visual models from natural language supervision[C] ICML 2021
[2] Scaling up visual and vision-language representation learning with noisy text supervision[C] ICML 2021
[3] Oscar: Object-semantics aligned pre-training for vision-language tasks[C] ECCV 2020
[4] Uniter: Universal image-text representation learning[C] ECCV 2020
[5] Align before fuse: Vision and language representation learning with momentum distillation[C] Neurips 2021

Background: Self-supervised Learning

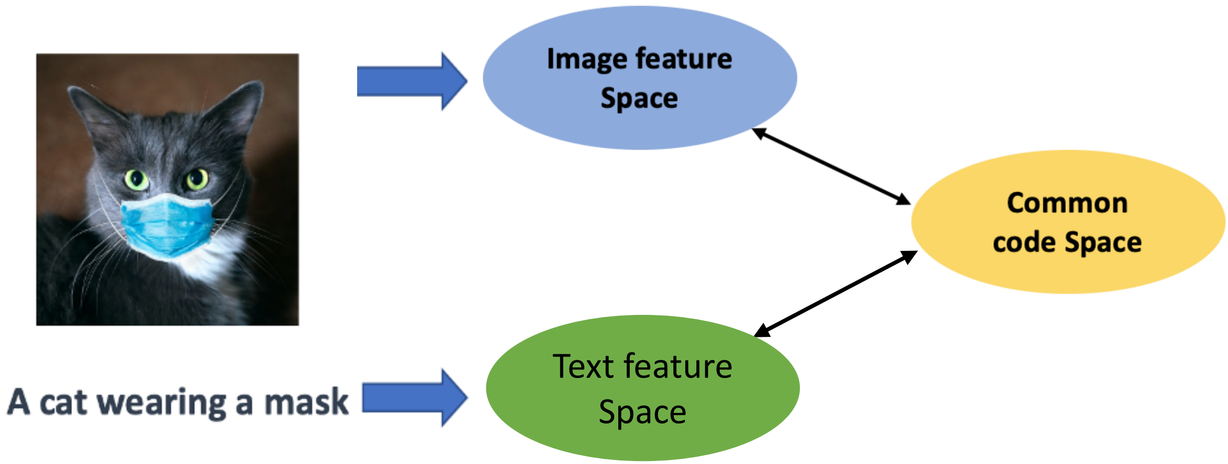
[1] Emerging properties in self-supervised vision transformers[C]. ICCV 2021

Motivation: Multimodal codebook as Semantic Bridge

Prior Work:



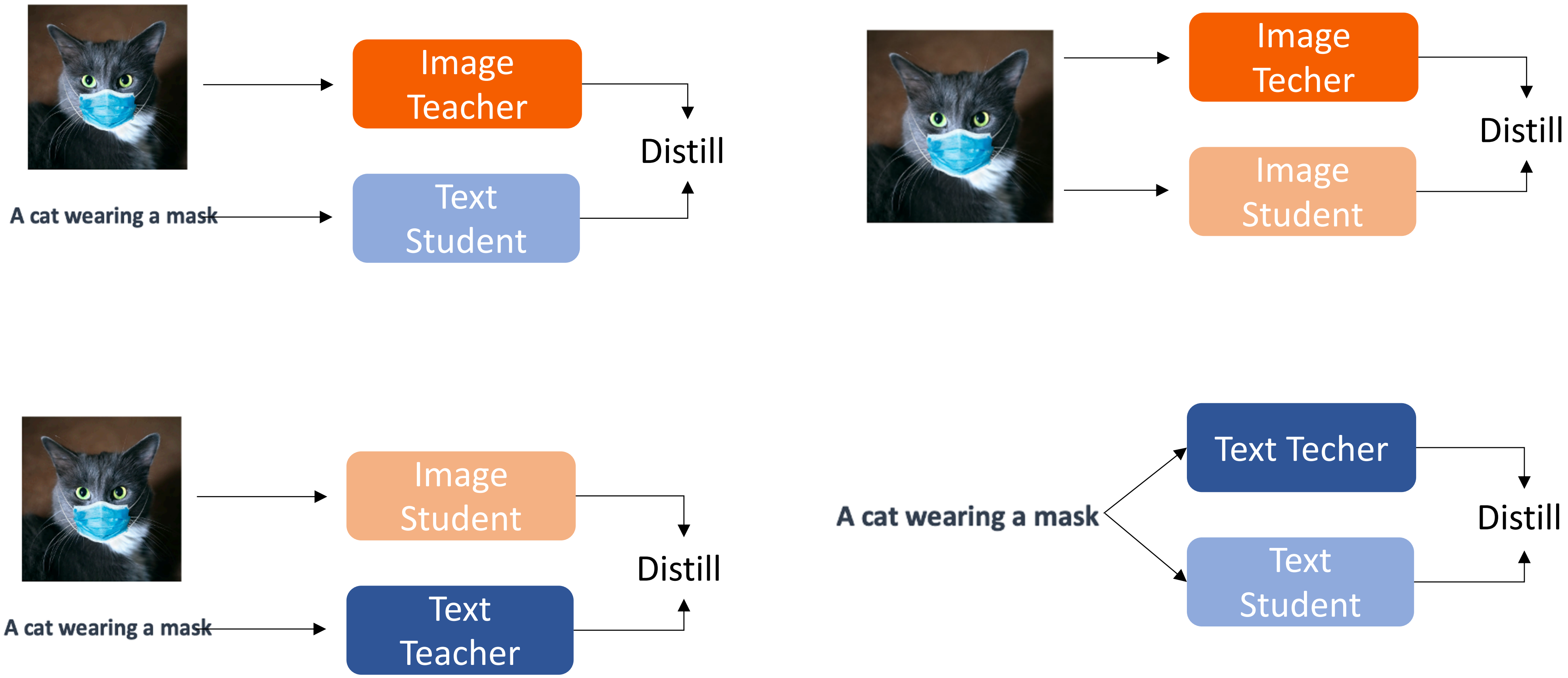
Ours:



[1] Unsupervised learning of visual features by contrasting cluster assignments[J]. Neurips 2020

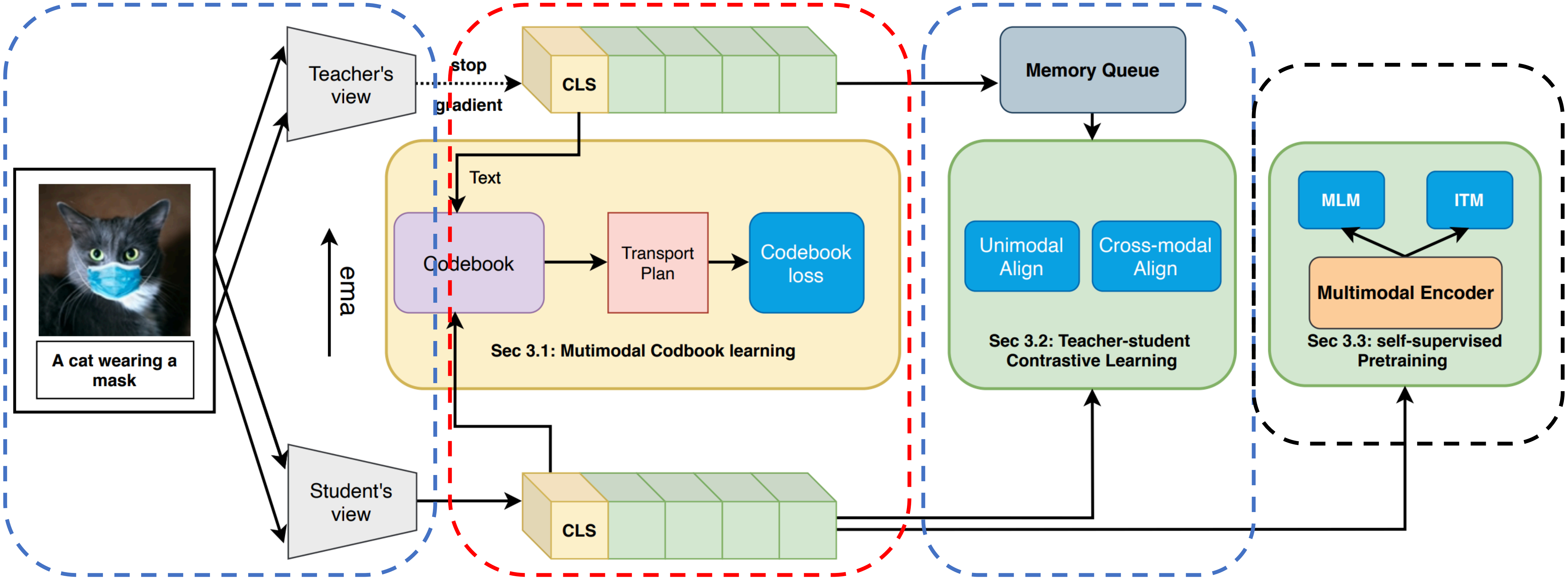
Motivation: Extension of SSL into Multimodal Setting

Image and text as two views of the same entity



[1] Align before fuse: Vision and language representation learning with momentum distillation[C] Neurips 2021

Framework Overview



Part 1: Multimodal Codebook Learning

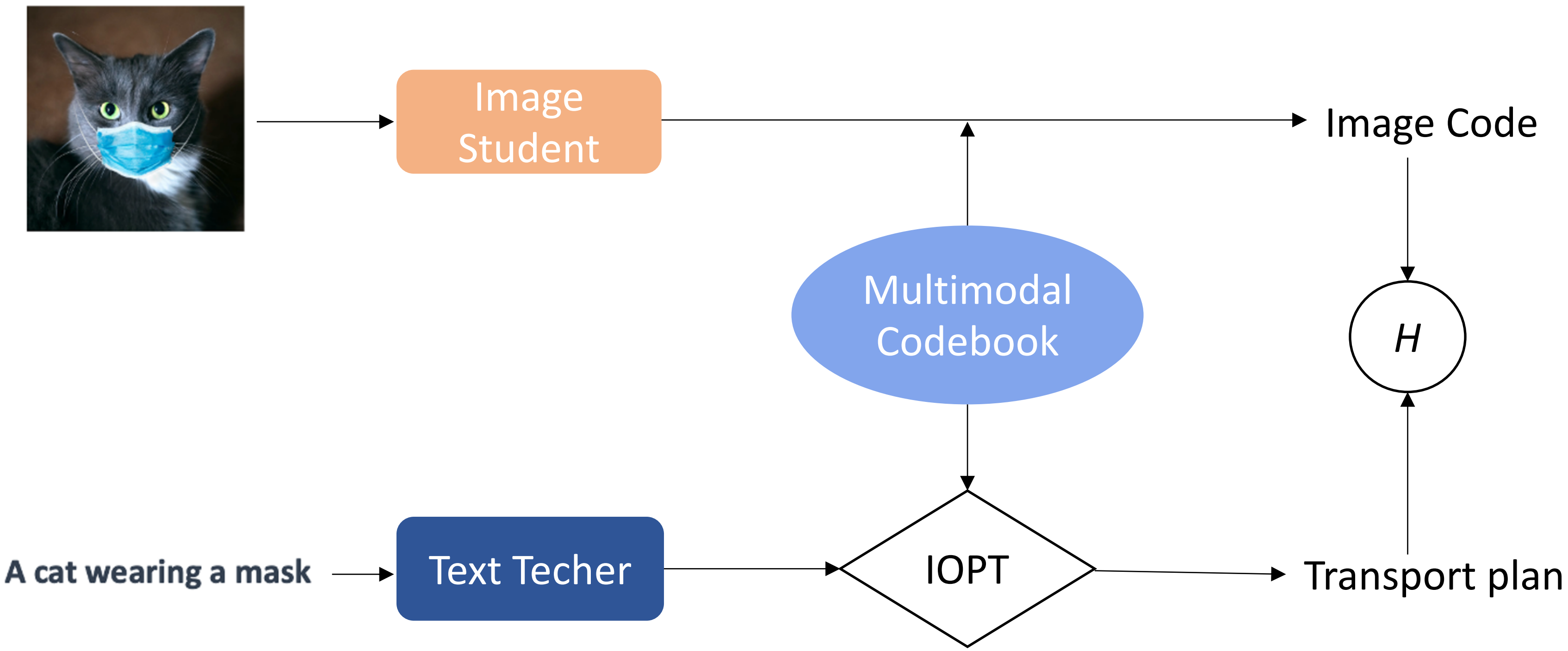
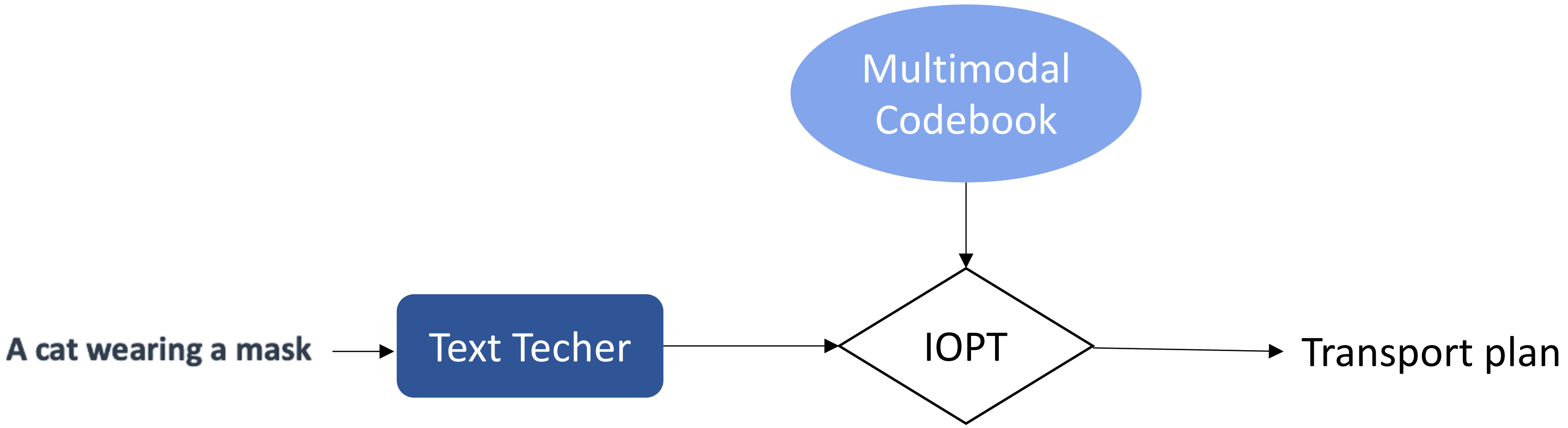


Image instances should distribute to clusters proportionally to the optimal text transport plan

$$L_{code} = L_{i2p}\left(Z_v, C, T_{t2p}\right) + L_{t2p}\left(Z_t, C, T_{i2p}\right)$$

Part 1: Multimodal Codebook Learning



$$\mathcal{L}_{\text{ot}} = \min_{\mathbf{T}\in\Pi(\mathbf{u},\mathbf{v})} \sum_{i=1}^{N}\sum_{j=1}^{K} \mathbf{T}_{ij} \cdot d(\boldsymbol{z}_i^m, \boldsymbol{c}_j) = \min_{\mathbf{T}\in\Pi(\mathbf{u},\mathbf{v})} \langle \mathbf{T}, \mathbf{D} \rangle \,,$$

What's the cost to transport instances to clusters?

$$L_{code} = L_{t2p}\,(Z_t, C, T_{i2p}) \;+\; L_{i2p}\,(Z_v, C, T_{t2p}) \;+\; L_{ot}\,(Z_t, C) \;+\; L_{ot}\,(Z_v, C)$$

Part 2: Teacher-student Contrastive Learning

$$\boldsymbol{p}_{t2i}(T) = \exp \frac{\boldsymbol{z}_t \boldsymbol{z}_v^{m\top}}{\gamma} / \sum_{\boldsymbol{z}_v^{m'} \in \mathbf{Q}_v} \exp \frac{\boldsymbol{z}_t \boldsymbol{z}_v^{m'\top}}{\gamma}$$

How close is text student to image teacher?

$$\boldsymbol{p}_{i2t}(I) = \exp \frac{\boldsymbol{z}_v \boldsymbol{z}_t^{m\top}}{\gamma} / \sum_{\boldsymbol{z}_t^{m'} \in \mathbf{Q}_t} \exp \frac{\boldsymbol{z}_v \boldsymbol{z}_t^{m'\top}}{\gamma}$$

How close is image student to text teacher?

$$\boldsymbol{p}_{i2i}(I) = \exp \frac{\boldsymbol{z}_v \boldsymbol{z}_v^{m\top}}{\gamma} / \sum_{\boldsymbol{z}_v^{m'} \in \mathbf{Q}_v} \exp \frac{\boldsymbol{z}_v \boldsymbol{z}_v^{m'\top}}{\gamma}$$

How close is image student to image teacher?

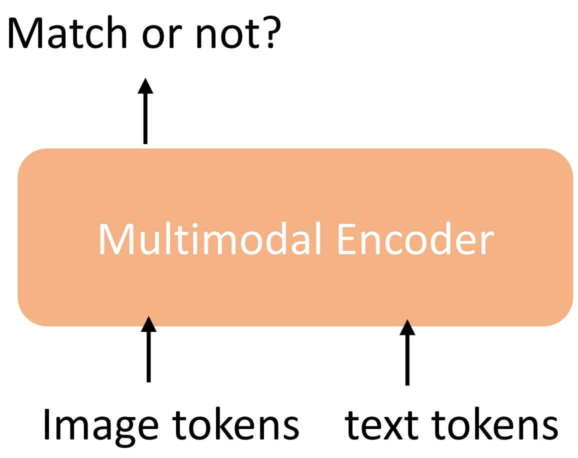
$$\boldsymbol{p}_{t2t}(T) = \exp \frac{\boldsymbol{z}_t \boldsymbol{z}_t^{m\top}}{\gamma} / \sum_{\boldsymbol{z}_t^{m'} \in \mathbf{Q}_t} \exp \frac{\boldsymbol{z}_t \boldsymbol{z}_t^{m'\top}}{\gamma}$$

How close is text student to text teacher?

$$L_{align} = H(P_{t2i}, y_{t2i}) + H(P_{i2t}, y_{i2t}) + H(P_{i2i}, y_{i2i}) + H(P_{t2t}, y_{t2t})$$

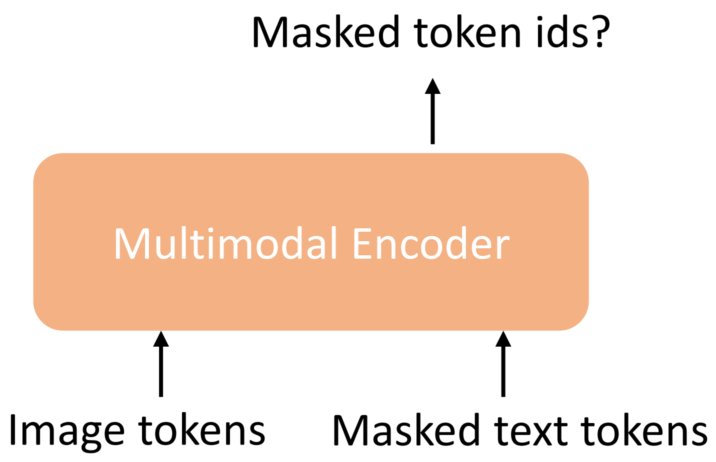
Part 3: Pretraining

Image-Text Matching



$$L_{itm} = H(P_{itm}, y_{itm})$$

Masked Language Modeling



$$L_{mlm} = H(P_{mlm}, y_{mlm})$$

$$L = L_{code} + L_{align} + L_{itm} + L_{mlm}$$

Experiments

Pretraining Data					
	CC3M	SBU	VG	COCO	Total
#images	2.92M	859K	100K	113K	~4.0M
#texts	2.92M	859K	769K	567K	~5.1M

Evaluation Data															
Retrieval				VQA				Visual Reasoning				Visual Entailment			
	Train	Val	Test		Train	Val	Test		Train	Val	Test		Train	Val	Test
COCO	113K	5K	5K	VQA2	83K	41K	81K	NLVR	Ref. [1]	7K	7K	SNLI	29.8K	1K	1K
Flickr	29K	1K	1K												

[1] Suhr A, Zhou S, Zhang A, et al. A corpus for reasoning about natural language grounded in photographs[J]. arXiv 2018

Quantitative Results

zero-shot image/text retrieval performance on MSCOCO and Flickr30K

Method	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 [36]	44.0	71.2	80.4	32.3	59.0	70.2	70.7	90.2	94.0	54.3	79.6	87.5
Unicoder-VL [24]	-	-	-	-	-	-	64.3	85.8	92.3	48.4	76.0	85.2
UNITER [8]	-	-	-	-	-	-	80.7	95.7	98.0	66.2	88.4	92.9
ViLT [22]	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 [37]	58.4	81.5	88.1	37.8	62.4	72.2	88.0	98.7	99.4	68.7	90.6	95.2
ALIGN [21]	58.6	83.0	89.7	45.6	69.8	78.6	88.6	98.7	**99.7**	75.7	93.8	**96.8**
ALBEF 4M [25]	68.6	89.5	94.7	50.1	76.4	84.5	90.5	98.8	**99.7**	76.8	93.7	96.7
Ours	**71.5**	**91.1**	**95.5**	**53.9**	**79.5**	**87.1**	**91.7**	**99.3**	99.8	**79.7**	**94.8**	97.3

finetuned image/text retrieval performance on MSCOCO and Flickr30K

Method	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 [36]	66.4	89.8	94.4	50.5	78.7	87.1	87.0	97.6	99.2	73.1	92.6	96.0
UNITER [8]	65.7	88.6	93.8	52.9	79.9	88.0	87.3	98.0	99.2	75.6	94.1	96.8
VILLA [14]	-	-	-	-	-	-	87.9	97.5	98.8	76.3	94.2	96.8
OSCAR [28]	70.0	91.1	95.5	54.0	80.8	88.5	-	-	-	-	-	-
ViLT [22]	61.5	86.3	92.7	42.7	72.9	83.1	83.5	96.7	98.6	64.4	88.7	93.8
UNIMO [27]	-	-	-	-	-	-	89.7	98.4	99.1	74.6	93.4	96.0
SOHO [20]	66.4	88.2	93.8	50.6	78.0	86.7	86.5	98.1	99.3	72.5	92.7	96.1
ALBEF 4M [25]	73.1	91.4	96.0	56.8	81.5	89.2	94.3	**99.4**	99.8	82.8	96.7	**98.4**
Ours	**75.3**	**92.6**	**96.6**	**58.7**	**82.8**	**89.7**	**95.1**	**99.4**	**99.9**	**83.3**	96.1	97.8

Ablation Studies

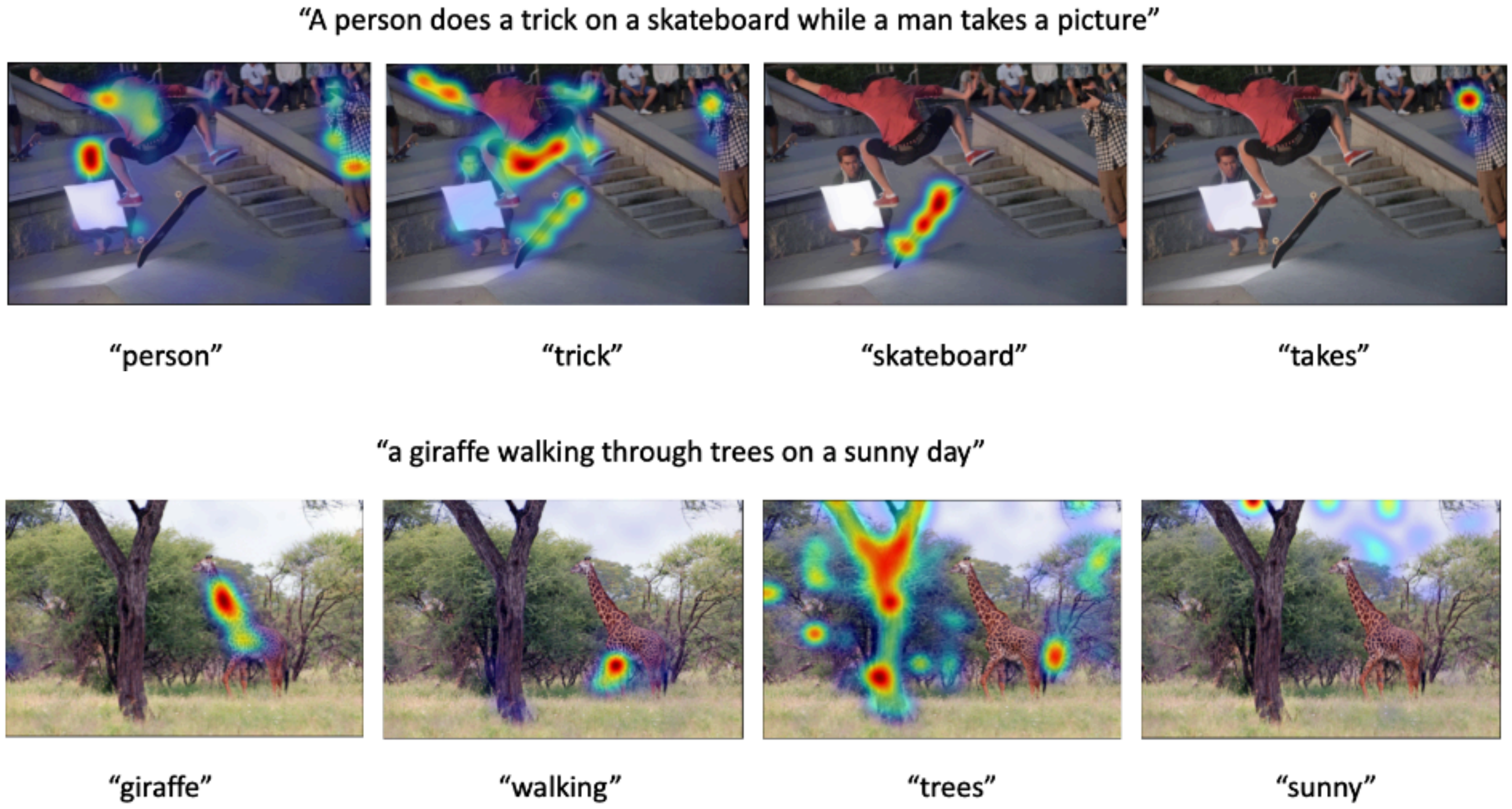
Ablations on different variants of our model for zero-shot image/text retrieval on MSCOCO and Flickr30K

Objective functions	MSCOCO (5K)						Flickr30K (1K)					
	Text Retrieval			Image Retrieval			Text Retrieval			Text 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
a: MLM+ITM+ITC (cross align)	68.60	89.50	94.70	50.10	76.40	84.50	84.90	97.20	99.00	68.18	88.58	93.02
b: MLM+ITM+ITC (intra + cross)	69.86	89.48	94.42	50.52	77.02	85.17	85.80	96.80	98.10	69.70	89.60	93.48
a + codebook (teacher feature)	70.74	89.54	94.88	51.39	77.86	85.60	86.00	97.00	98.20	70.18	90.66	94.44
b + codebook (student feature)	71.12	89.62	94.78	51.40	77.42	85.53	86.30	96.90	98.30	70.34	90.00	93.84
b + codebook (teacher feature)	**71.10**	**90.60**	**95.10**	**52.10**	**78.00**	**85.90**	**86.70**	**97.30**	98.70	**71.40**	**90.82**	**94.62**

	TR@1	TR@5	TR@10	IR@1	IR@5	IR@10
ALBEF	55.70	81.92	88.78	41.08	69.01	78.86
0.5x codebook	58.66	83.9	90.64	43.74	72.10	81.58
2.0x codebook	59.02	84.46	91.06	43.62	71.69	81.12
3K codewords	58.96	84.28	90.98	44.66	72.31	81.68
500 codewords	55.52	81.68	89.28	41.53	68.75	78.43
Ours	59.38	84.04	91.20	44.71	72.63	81.69

Ablations on codebook sizes under limited pretraining regime using only MSCOCO

Qualitative Results



Grad-CAM visualization on the cross-attention maps corresponding to individual worlds

Conclusions

- Propose *multi-modal codebook* to align image and text modality at cluster level
- Connect SSL with vision-language pretraining by generalizing teacher-student distillation to multimodal setting