Vision-Language Learning

Hongguang Zhu

Image-text Tasks Overview

Close-set classification

Popular Image-text Tasks:
VQA, GQA, VisDial, VCR, NLVR2, image-text matching

Open-ended text sequence

Popular Image-text Tasks:
Image captioning, paragraph captioning, storytelling, open-ended VQA

Unified Image-Text Modeling

Box/mask localization

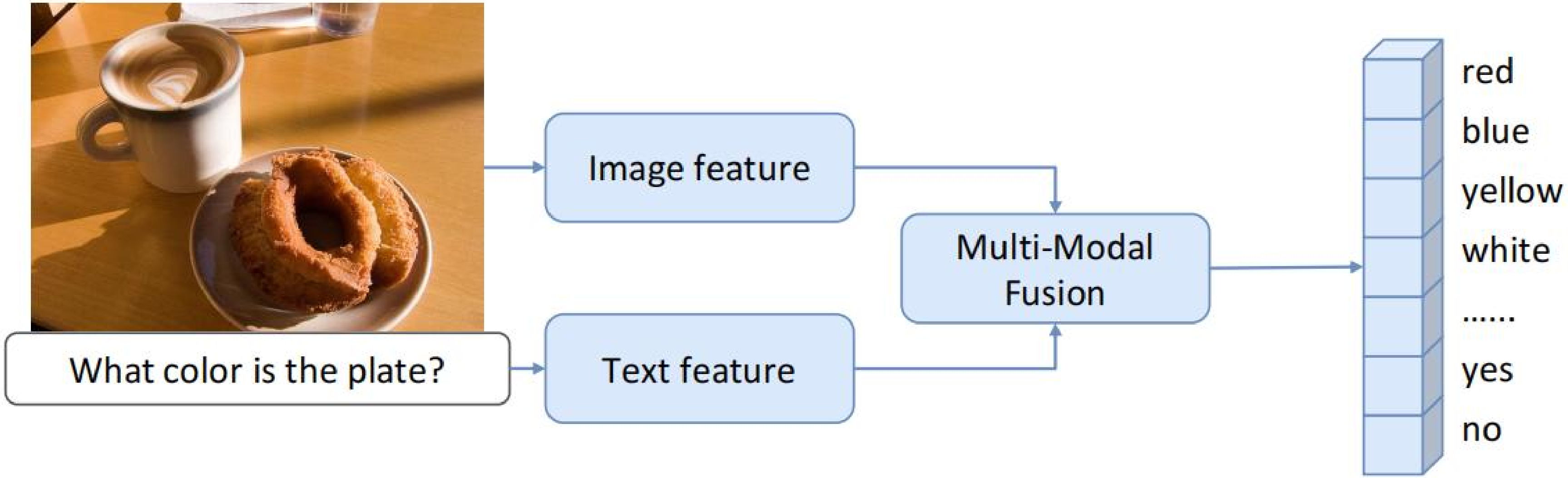
Popular Image-text Tasks:
Referring expression comprehension/ segmentation, phrase grounding, grounded captioning

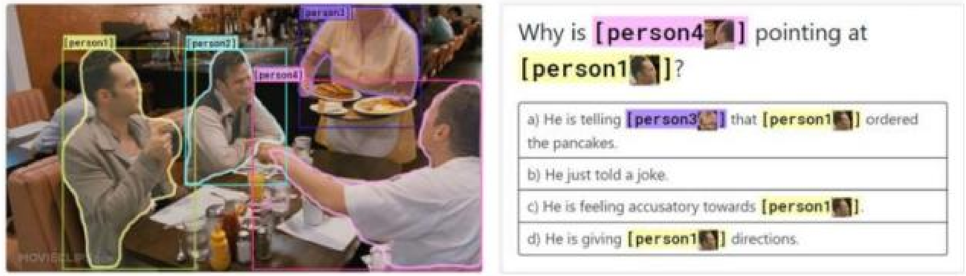
Pixel prediction

A donut on a white plate next to a cup of latte.

Popular Image-text Tasks:
Text-to-image synthesis, text-based image editing

Close-set Classification





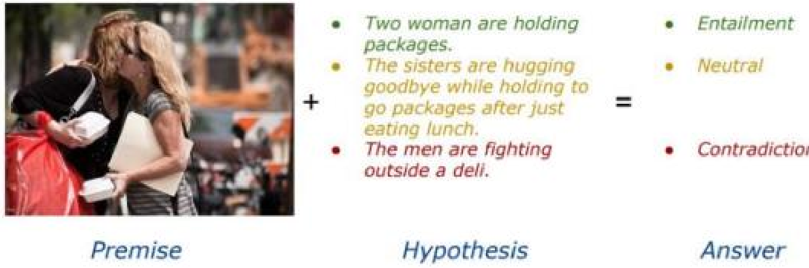


Image-text matching, NLVR2

True
False

VCR

a)
b)
c)
d)

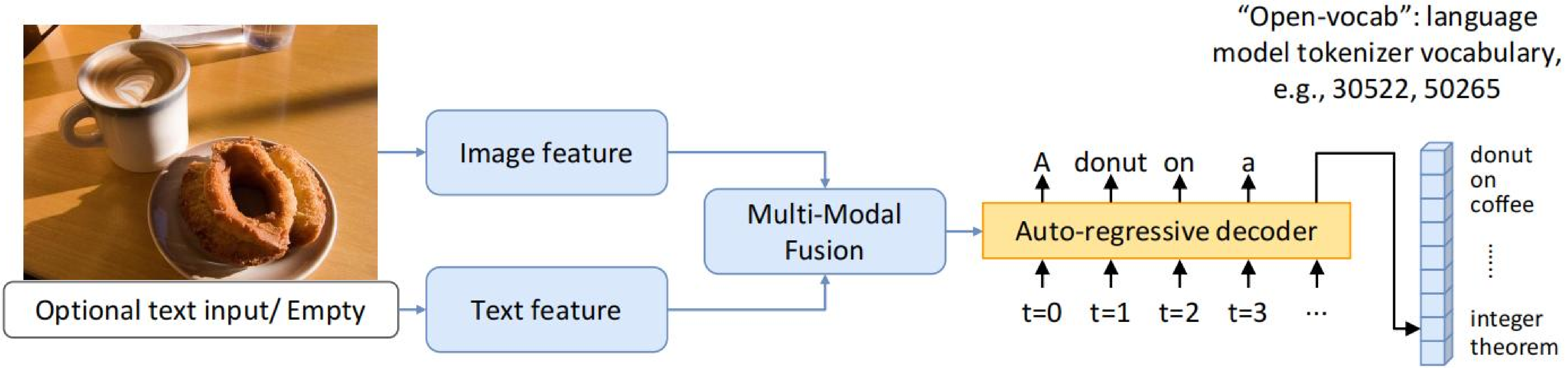
Visual Entailment

Entailment
Neutral
Contradiction

visual commonsense reasoning

Natural Language for Visual Reasoning

Open-ended Text Sequence



A donut on a white plate next to a cup of latte.

Image captioning

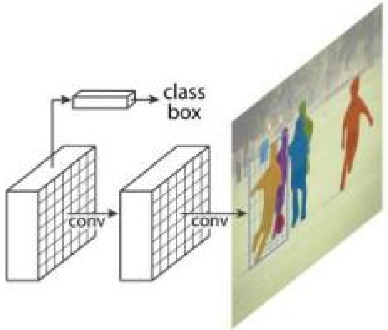
This image is of a family celebrating Christmas. They are all gathered around a dinner table, with a turkey and other food on it. The family is smiling and seems to be enjoying themselves. There is a Christmas tree in the background and some Christmas lights on the walls.

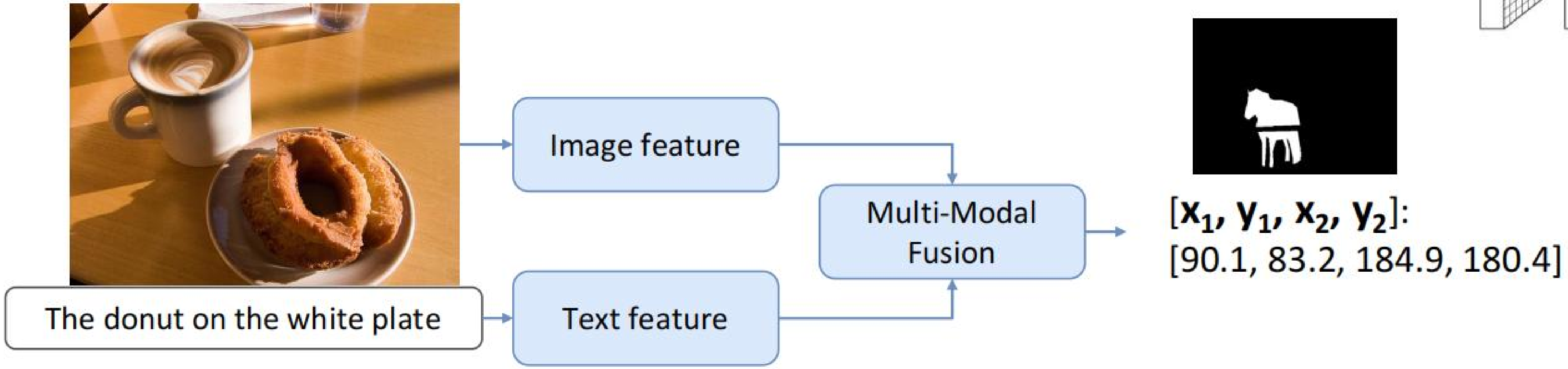
Paragraph Captioning

What color is the plate? The plate is white.

Open-ended VQA

Box/mask Localization





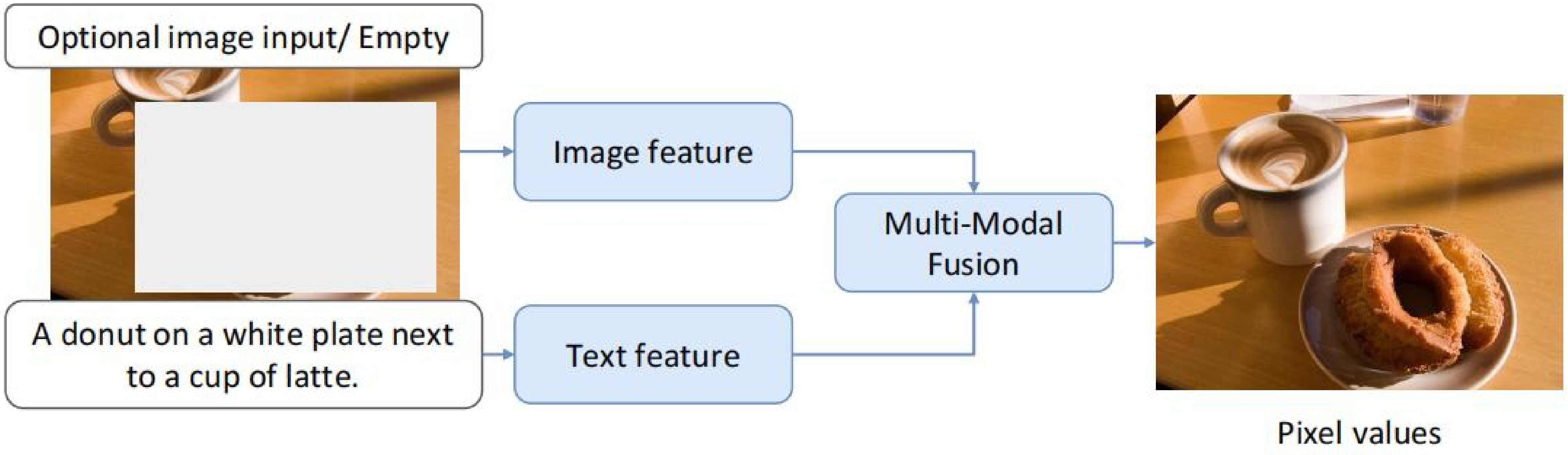
A man with pierced ears is wearing glasses and an ora
A man with glasses is wearing a beer can crotched ha
A man with gauges and glasses is wearing a Blitz hat
A man in an orange hat starring at something
A man wears an orange hat and glasses.

Visual grounding
(REC, phrase grounding)

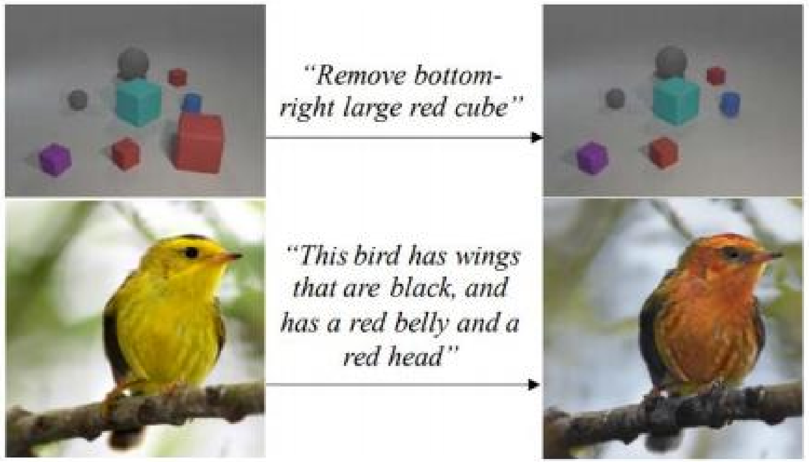
glass bottles blonde hair pedestrian crosswalk

Language-based segmentation
(RES)

Pixel Prediction



Text-to-image synthesis



Text-based image editing

5.下游任务

分类任务:

Visual Question Answering (VQA). VQA 就是对于一个图片回答图片内容相关的问题。将图片和问题输入到模型中，输出是答案的分布，取概率最大的答案为预测答案。

Visual Reasoning and Compositional Question Answering (GQA). 是VQA的升级版，旨在推进自然场景视觉推理的研究。其数据集中的图像、问题和答案具有匹配的语义表示。

Video-Language Inference (VLI). 给定一个以对齐字幕为前提的视频片段，再加上基于视频内容的自然语言假设，模型需要推断该假设是否与给定视频片段相矛盾。

Natural Language for Visual Reasoning (NLVR). 同时输入两张 Image 和一个描述，输出是描述与 Image 的对应关系是否一致，label 只有两种（true/false）。

Visual Entailment (VE). 在 Visual Entailment 中，Image 是前提，Text 是假设，模型的目标是预测 Text 是不是 “Entailment Image” ，一共有三中 label，分别是 Entailment、Neutral 和 Contradiction。

Visual Commonsense Reasoning (VCR). Visual Commonsense Reasoning 中，任务是以选择题形式存在的，对于一个问题有四个备选答案，模型必须从四个答案中选择出一个答案，然后再从四个备选理由中选出选择这个答案的理由。

Grounding Referring Expressions (GRE). 给定文本，选择文本所关联的图片区域。即输入是一个句子，模型要在图片中圈出对应的 region。对于这个任务，我们可以对每一个 region 都输出一个 score，score 最高的 region 作为预测 region。

回归任务:

Multi-modal Sentiment Analysis (MSA). 通过利用多模态信号（例如视觉、语言等）来检测视频中的情绪。它是作为一个连续的强度变量来预测话语的情感取向。

检索任务:

Vision-Language Retrieval (VLR). 在 Image-Text Retrieval 任务中，就是给定一个模态的指定样本，在另一个模态的 DataBase 中找到对应的样本。这个任务 Image-Text Matching 任务非常相似，所以在 fine-tune 的过程中就是选择 positive pair 和 negative pair 的方式来训练模型。

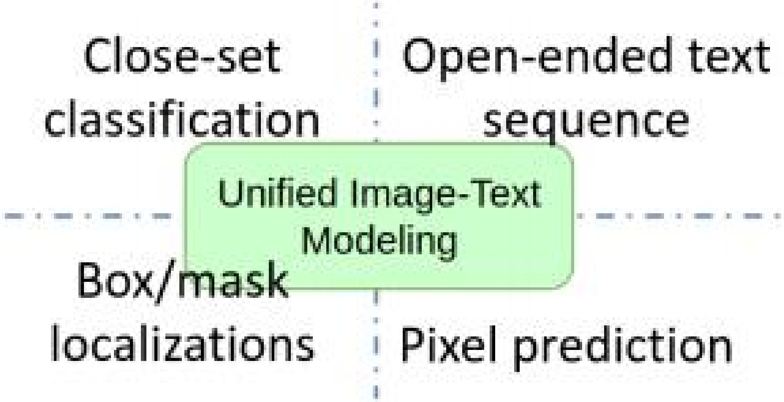
生成任务:

Visual Captioning (VC). 为给定的视觉（图像或视频）输入生成语义和语法上合适的文本描述。

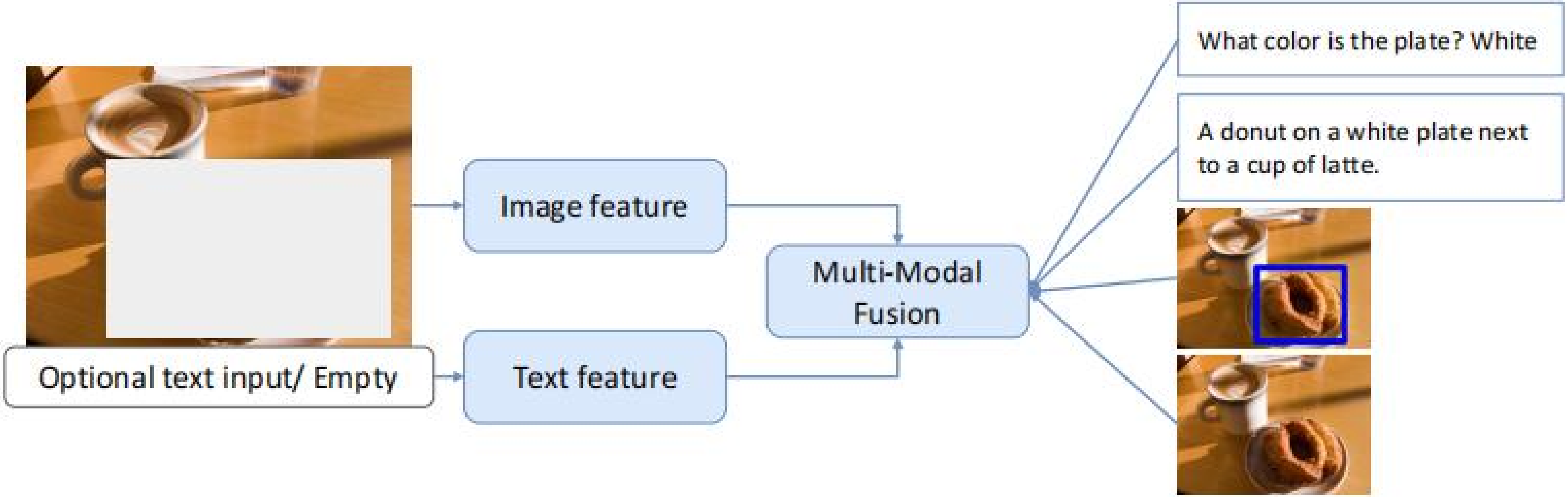
Novel Object Captioning at Scale (NoCaps). 扩展了VC任务，以测试模型从开放图像数据集中描述新对象的能力，这些对象在训练语料库中是没有的。

Visual Dialogue (VD). VD的任务形式是一个图像（或视频）、一段历史对话和一个语言问题，并让模型生成问题的答案。

Why Unified Image-Text Modeling



- Better performance
- New capabilities
- Task-agnostic unified systems



The most recent art

Contrastive Vision-Language Learning

A lot of research works come along the line of vision-language learning for vision

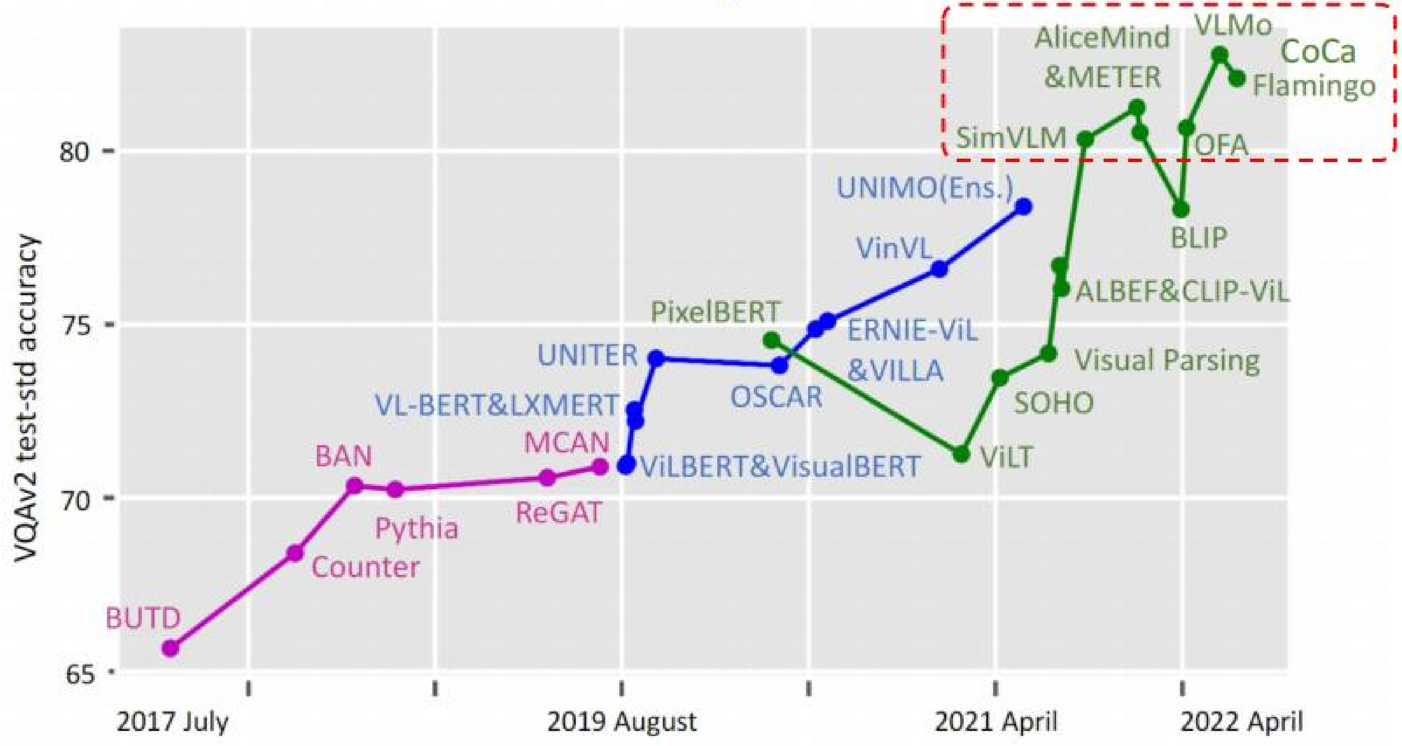
https://github.com/phellonchen/awesome-Vision-and-Language-Pre-training

How about big multimodal models?

- Models that have either billion-level parameters or use billion-level pre-training data are considered as *"big"* in this context
- First, note that foundation models are not necessarily needed to be big
- CLIP-like dual encoders and text-to-image big models (DALLE-2, Imagen) are not considered here (will be covered in the afternoon session)
- Take VQA as an example
  - OD-based models
  - E2E models

Large model sizes and pre-training data have been the driving force for SOTA performance.

We will also briefly talk about what's beyond SOTA chasing in later slides.



A summary of big multimodal models

Model	Model Size				#Pre-training image-text data	Pre-training tasks
	Img Enc	Txt Enc	Fusion	Total		
CLIP ViT-L/14	302M	123M	0	425M	400M	ITC
ALIGN	480M	340M	0	820M	1.8B	ITC
Florence	637M	256M	0	893M	900M	ITC
SimVLM-huge	300M	39M	600M	939M	1.8B	PrefixLM
METER-huge	637M	125M	220M	982M	20M*	MLM+ITM
LEMON	147M	39M	636M	822M	200M	MLM
Flamingo	200M	70B	10B	80.2B	2.1B+27M video-text	LM
GIT	637M	40M	70M	747M	800M	LM
VLMo++	--	--	--	565M	1B	MLM+ITM+ITC
CoCa	1B	477M	623M	2.1B	4.8B (before filtering)	ITC+LM

Note: Some of the numbers here are based on our best estimate
**: excluding the data used to pre-train the Florence image encoder*

Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision
Learning Transferable Visual Models From Natural Language Supervision
Florence: A New Foundation Model for Computer Vision
SimVLM: Simple Visual Language Model Pretraining with Weak Supervision
An Empirical Study of Training End-to-End Vision-and-Language Transformers
Scaling Up Vision-Language Pre-training for Image Captioning
Flamingo: a Visual Language Model for Few-Shot Learning
GIT: A Generative Image-to-text Transformer for Vision and Language
VLMo: Unified Vision-Language Pre-Training with Mixture-of-Modality-Experts
CoCa: Contrastive Captioners are Image-Text Foundation Models

Application：BAIDU

为什么需要大模型

标注数据更少

通过学习少量行业数据，大模型就能够应对特定业务场景的需求

模型效果更优

大模型在各场景上的效果均优于普通模型

创造能力更强

大模型能够进行内容生成（AIGC），助力内容规模化生产

灵活定制场景

通过举例子的方式，定制大模型海量的应用场景

中国人寿

保险合同管理业务中，针对合同中条款文本进行关键字段的提取，是其核心的业务需求；基于文心ERNIE大模型实现了保险合同条款的智能解析，自动提取近40个维度的关键字段，业务处理效率大幅提升。

保险合同条款智能解析

度小满金融

金融风控一直是金融交易过程中非常重要的一个环境，面对用户的信用风险与道德风险，基于文心ERNIE大模型，对业务信息进行语义层面深度风控建模，有效地提升了优质客群人数，降低了贷款风险。

金融风控

国美

电商服务考核业务中，针对用户评论信息，对评论涉及环节的服务人员进行评分考核。基于文心ERNIE大模型实现自动提取用户评论中科学精准的评价信息，完成服务人员考核，提升负面问题处理率。

用户评论观点分析

https://m6.aliyun.com/?spm=a2c6h.12873639.article-detail.5.5de01e6bbb97gt#/
https://wenxin.baidu.com/

Topic List

- **Before CVPR (2021.11)**
  - ALBEF （NIPS21） Align before Fuse: Vision and Language Representation Learning with Momentum Distillation
    Salesforce Research

- **CVPR: (exploration)**
  - Vision-Language Pre-Training with Triple Contrastive Learning
    UTA, Amazon
  - Multi-modal Alignment using Representation Codebook
    UTA, Amazon
  - An Empirical Study of Training End-to-End Vision-and-Language Transformers
    UCLA,Microsoft

- **After CVPR (2022.4-6)-> AGI, Foundational model**
  - OpenAI: Flamingo
  - Google: LIMoE
  - Sensetime: Uni-Perceiver-MoE
  - Allen Institute for AI： UNIFIED-IO

CVPR Trend：

- key word： CLIP， Pretrain， Video, Captioning -> more
- emerging: Sign Language (4)， vision-language Navigation (10+)
- declining: Cooking Recipes(1)

Part I: Before CVPR

Align before Fuse: Vision and Language Representation Learning with Momentum Distillation

Salesforce Research

ALBEF:

- Three key limitations
  - V/T embeddings reside in their own spaces
    - multimodal encoders learn to model their interactions -> challenging
  - object detector is both annotation-expensive and compute-expensive
    - requires bounding box annotations during pre-training
    - high resolution images during inference
  - IT datasets are collected from the web and are inherently noisy
    - pre-training objectives (MLM) may overfit to the noisy text
    - degrade the model's generalization performance.

- intermediate image-text contrastive (ITC) loss
  - align the image features and the text features
    - easier for the multimodal encoder to perform cross-modal learning
  - unimodal encoders to better understand the semantic meaning of images and texts
  - learns a common low-dimensional space to embed images and texts
    - to find more informative samples through our contrastive hard negative mining.

Method	# Pre-train Images	Flickr30K (1K test set) TR			IR		
		R@1	R@5	R@10	R@1	R@5	R@10
UNITER [2]	4M	83.6	95.7	97.7	68.7	89.2	93.9
CLIP [6]	400M	88.0	98.7	99.4	68.7	90.6	95.2
ALIGN [7]	1.2B	88.6	98.7	99.7	75.7	93.8	96.8
ALBEF	4M	90.5	98.8	99.7	76.8	93.7	96.7
ALBEF	14M	**94.1**	**99.5**	**99.7**	**82.8**	**96.3**	**98.1**

Table 3: Zero-shot image-text retrieval results on Flickr30K.

Method	VQA test-dev	VQA test-std	$NLVR^2$ dev	$NLVR^2$ test-P	SNLI-VE val	SNLI-VE test
VisualBERT [13]	70.80	71.00	67.40	67.00	-	-
VL-BERT [10]	71.16	-	-	-	-	-
LXMERT [1]	72.42	72.54	74.90	74.50	-	-
12-in-1 [12]	73.15	-	-	78.87	-	76.95
UNITER [2]	72.70	72.91	77.18	77.85	78.59	78.28
VL-BART/T5 [54]	-	71.3	-	73.6	-	-
ViLT [21]	70.94	-	75.24	76.21	-	-
OSCAR [3]	73.16	73.44	78.07	78.36	-	-
VILLA [8]	73.59	73.67	78.39	79.30	79.47	79.03
ALBEF (4M)	74.54	74.70	80.24	80.50	80.14	80.30
ALBEF (14M)	**75.84**	**76.04**	**82.55**	**83.14**	**80.80**	**80.91**

Table 4: Comparison with state-of-the-art methods on downstream vision-language tasks.

Method	# Pre-train Images	Flickr30K (1K test set) TR			IR			MSCOCO (5K test set) TR			IR		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
UNITER	4M	87.3	98.0	99.2	75.6	94.1	96.8	65.7	88.6	93.8	52.9	79.9	88.0
VILLA	4M	87.9	97.5	98.8	76.3	94.2	96.8	-	-	-	-	-	-
OSCAR	4M	-	-	-	-	-	-	70.0	91.1	95.5	54.0	80.8	88.5
ALIGN	1.2B	95.3	99.8	100.0	84.9	97.4	98.6	77.0	93.5	96.9	59.9	83.3	89.8
ALBEF	4M	94.3	99.4	99.8	82.8	96.7	98.4	73.1	91.4	96.0	56.8	81.5	89.2
ALBEF	14M	**95.9**	**99.8**	**100.0**	**85.6**	**97.5**	**98.9**	**77.6**	**94.3**	**97.2**	**60.7**	**84.3**	**90.5**

Table 2: Fine-tuned image-text retrieval results on Flickr30K and COCO datasets.

1. ITC

default:GT one-hot

$$\mathcal{L}_{v2t} = -\frac{1}{b}\sum_{i}^{b}\log\frac{\exp\left(s\left(i_i^t, t_i^t\right)\right)}{\sum_{j=1}^{B}\exp\left(s\left(i_i^t, t_j^t\right)\right)}$$

softmax-normalized similarity

$$p_m^{\mathrm{i2t}}(I) = \frac{\exp(s(I,T_m)/\tau)}{\sum_{m=1}^{M}\exp(s(I,T_m)/\tau)}, \quad p_m^{\mathrm{t2i}}(T) = \frac{\exp(s(T,I_m)/\tau)}{\sum_{m=1}^{M}\exp(s(T,I_m)/\tau)}$$

cross-entropy = KD loss

$$\mathcal{L}_{\mathrm{itc}} = \frac{1}{2}\mathbb{E}_{(I,T)\sim D}\left[\mathrm{H}(\boldsymbol{y}^{\mathrm{i2t}}(I), \boldsymbol{p}^{\mathrm{i2t}}(I)) + \mathrm{H}(\boldsymbol{y}^{\mathrm{t2i}}(T), \boldsymbol{p}^{\mathrm{t2i}}(T))\right]$$

MoD -> pseudo target label

(1-α) ones(V,T) + α(V,T_m)

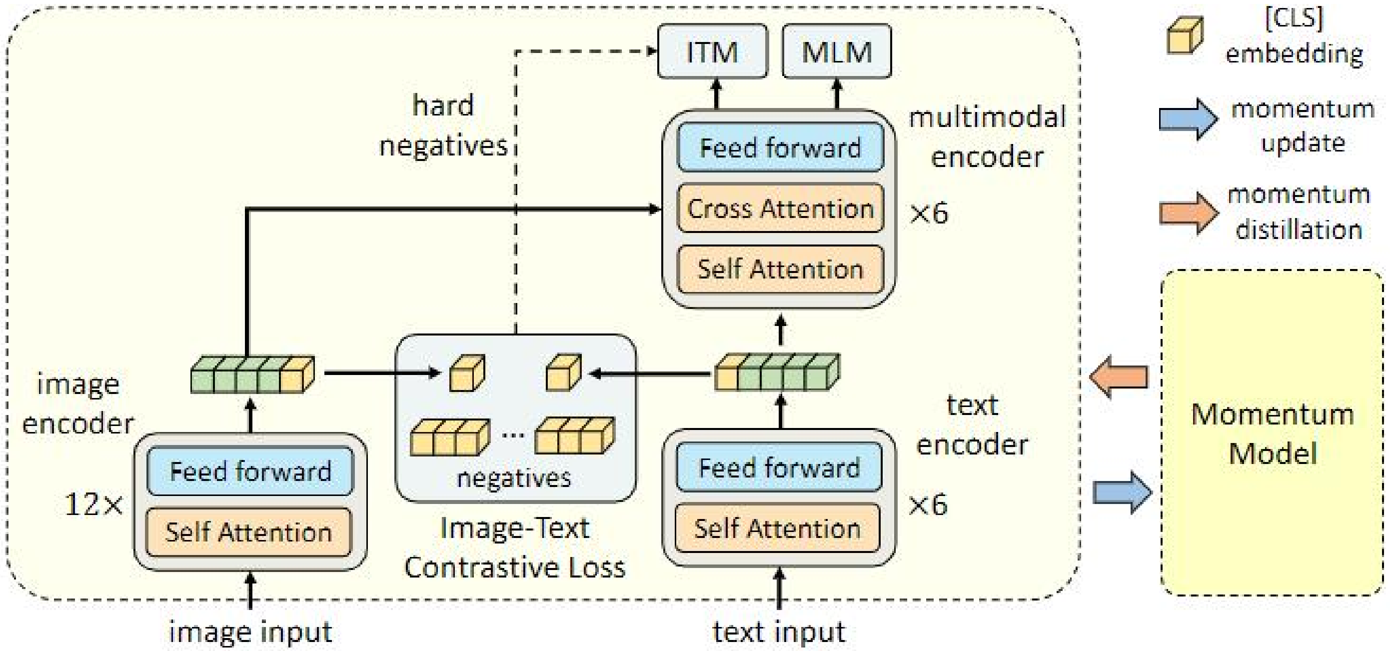


Figure 1: **Illustration of ALBEF.** It consists of an image encoder, a text encoder, and a multimodal encoder. We propose an image-text contrastive loss to align the unimodal representations of an image-text pair before fusion. An image-text matching loss (using in-batch hard negatives mined through contrastive similarity) and a masked-language-modeling loss are applied to learn multimodal interactions between image and text. In order to improve learning with noisy data, we generate pseudo-targets using the momentum model (a moving-average version of the base model) as additional supervision during training.

2. Momentum Distillation

$$\mathcal{L}_{itc}^{mod} = (1-\alpha)\mathcal{L}_{itc} + \frac{\alpha}{2}\mathbb{E}_{(I,T)\sim D}\left[\mathrm{KL}(\mathbf{q}^{i2t}(I) \parallel \mathbf{p}^{i2t}(I)) + \mathrm{KL}(\mathbf{q}^{t2i}(T) \parallel \mathbf{p}^{t2i}(T))\right]$$

$$\mathcal{L}_{mlm}^{mod} = (1-\alpha)\mathcal{L}_{mlm} + \alpha\mathbb{E}_{(I,\hat{T})\sim D}\mathrm{KL}(\mathbf{q}^{msk}(I,\hat{T}) \parallel \mathbf{p}^{msk}(I,\hat{T}))$$

3. MI

ITC $\mathcal{L}_{\mathrm{itc}} = \frac{1}{2}\mathbb{E}_{(I,T)\sim D}\left[\mathrm{H}(\boldsymbol{y}^{\mathrm{i2t}}(I), \boldsymbol{p}^{\mathrm{i2t}}(I)) + \mathrm{H}(\boldsymbol{y}^{\mathrm{t2i}}(T), \boldsymbol{p}^{\mathrm{t2i}}(T))\right]$

MLM $\mathcal{L}_{\mathrm{mlm}} = \mathbb{E}_{(I,\hat{T})\sim D}\mathrm{H}(\boldsymbol{p}^{\mathrm{msk}}(I,\hat{T}), \boldsymbol{y}^{\mathrm{msk}})$

ITM $\mathcal{L}_{\mathrm{itm}} = \mathbb{E}_{(I,T)\sim D}\mathrm{H}(\boldsymbol{p}^{\mathrm{itm}}(I,T), \boldsymbol{y}^{\mathrm{itm}})$

Ablation

MoD
queue:65536

#Pre-train Images	Training tasks	TR (flickr test)	IR (flickr test)	SNLI-VE (test)	$NLVR^2$ (test-P)	VQA (test-dev)
4M	MLM + ITM	93.96	88.55	77.06	77.51	71.40
	ITC + MLM + ITM	96.55	91.69	79.15	79.88	73.29
	ITC + MLM + ITM_{hard}	97.01	92.16	79.77	80.35	73.81
	ITC_{MoD} + MLM + ITM_{hard}	97.33	92.43	79.99	80.34	74.06
	Full (ITC_{MoD} + MLM_{MoD} + ITM_{hard})	97.47	92.58	80.12	80.44	74.42
	ALBEF (Full + $MoD_{Downstream}$)	97.83	92.65	80.30	80.50	74.54
14M	ALBEF	98.70	94.07	80.91	83.14	75.84

Table 1: Evaluation of the proposed methods on four downstream V+L tasks. For text-retrieval (TR) and image-retrieval (IR), we report the average of R@1, R@5 and R@10. ITC: image-text contrastive learning. MLM: masked language modeling. ITM_{hard}: image-text matching with contrastive hard negative mining. MoD: momentum distillation. $MoD_{Downstream}$: momentum distillation on downstream tasks.

finetune：ITM-one negative

ALIGN:98.37

ITC+ITM (k) >ITC

Flickr30K	w/ hard negs s_{itc}	w/ hard negs $k=16$	w/ hard negs $k=128$	w/ hard negs $k=256$	w/o hard negs $k=128$
TR	97.30	98.60	98.57	98.57	98.22 (−0.35)
IR	90.95	93.64	93.99	93.95	93.68 (−0.31)

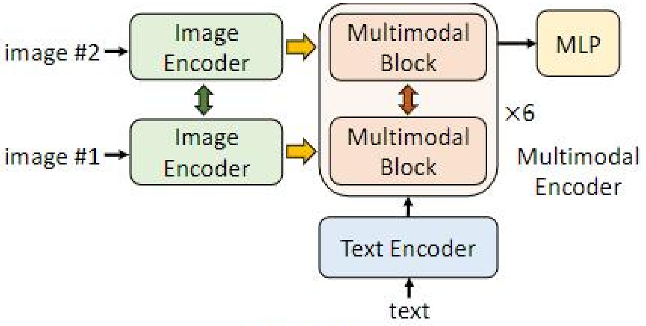
Table 6: Ablation study on fine-tuned image-text retrieval. The average recall on the test set is reported. We use s_{itc} to filter top-k candidates and calculate their s_{itm} score for ranking.

more 128 forward time

text assignment pre-training

$NLVR^2$	w/ TA share all	w/ TA share CA	w/ TA no share	w/o TA share all	w/o TA share CA	w/o TA no share
dev	82.13	82.55	81.93	80.52	80.28	77.84
test-P	82.36	83.14	82.85	81.29	80.45	77.58

Table 7: Ablation study on $NLVR^2$.



(b) $NLVR^2$

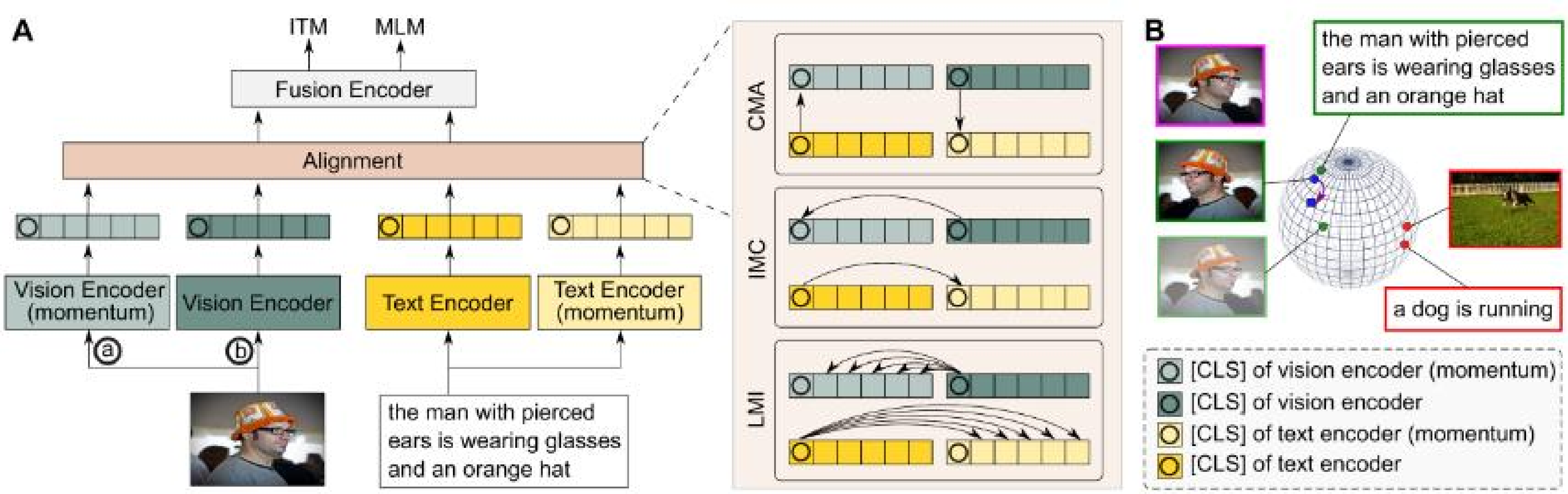
three-way classification, pretrain 1 epoch

Part II: CVPR

Vision-Language Pre-Training with Triple Contrastive Learning

UTA, Amazon

- cross-modal alignment (CMA):maximizing the mutual information (MI)
  - fail to ensure that similar inputs from the same modality stay close by
  - global information -> localized and structural information (max local MI of local regions and global summary)
- triple contrastive learning (TCL)
  - leveraging both cross-modal and intra-modal self-supervision.



I:two views $I_1, I_2 \quad T_+ = T$

two encoder: raw, EMA

$$\mathcal{L}_{nce}(I_1, T_+, \tilde{T}) = -\mathbb{E}_{p(I,T)}\left[log\frac{e^{(\text{sim}(I_1,T_+)/\tau)}}{\sum_{k=1}^{K} e^{(\text{sim}(I_1,\tilde{T}_k)/\tau)}}\right]$$

$$\mathcal{L}_{cma} = \frac{1}{2}[\mathcal{L}_{nce}(I_1, T_+, \tilde{T}) + \mathcal{L}_{nce}(T, I_2, \tilde{I})]$$

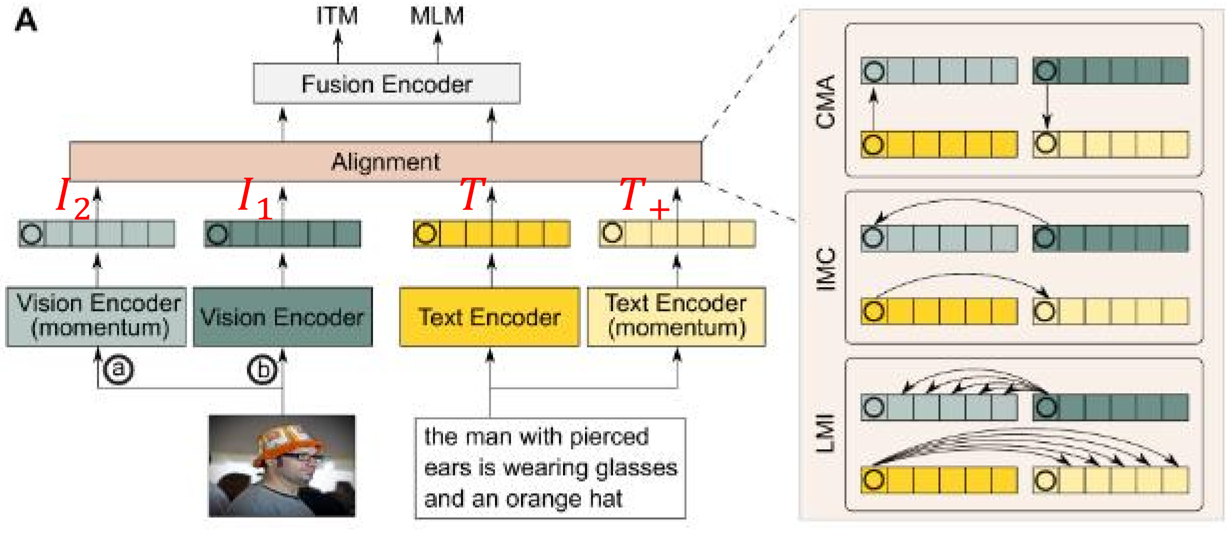
$$\mathcal{L}_{imc} = \frac{1}{2}[\mathcal{L}_{nce}(T, T_+, \tilde{T}) + \mathcal{L}_{nce}(I_1, I_2, \tilde{I})]$$

$$\mathcal{L}_{lmi} = \frac{1}{2}\left[\frac{1}{M}\sum_{i=1}^{M}\mathcal{L}_{nce}(I_1, I_2^i, \tilde{I}_l) + \frac{1}{N}\sum_{j=1}^{N}\mathcal{L}_{nce}(T, T_+^j, \tilde{T}_l)\right]$$

$$\mathcal{L}_{itm} = \mathbb{E}_{p(I,T)} H(\phi(I,T), y^{(I,T)})$$

$$\mathcal{L}_{mlm} = \mathbb{E}_{p(I,T^{msk})} H(\Phi(I, T^{msk}), y^{T^{msk}})$$

$$\mathcal{L} = \mathcal{L}_{cma} + \mathcal{L}_{imc} + \mathcal{L}_{lmi} + \mathcal{L}_{itm} + \mathcal{L}_{mlm}$$



	COCO	VG	SBU	CC	CC12M
# images	113K	100K	859K	2.92M	10.97M
# text	567K	769K	859K	2.92M	10.97M

All of our experiments are performed on 8 NVIDIA A100 GPUs with PyTorch framework [36]. Our vision encoder is implemented by ViT-B/16 with 12 layers and 85.8M parameters. Both the text encoder and the fusion encoder are implemented by a 6-layer transformer. They are initialized by the first 6 layers and the last 6 layers of $\text{BERT}_{\text{base}}$ (123.7M parameters), respectively. We set $K = 65,536$ and $m = 0.995$. For the pre-training stage, the model is trained

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 [38]	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 [8]	4M	64.1	87.7	93.3	48.8	76.7	85.8	80.7	95.7	98.0	66.2	88.4	92.9
ViLT [24]	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 [39]	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 [26]	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 [23]	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

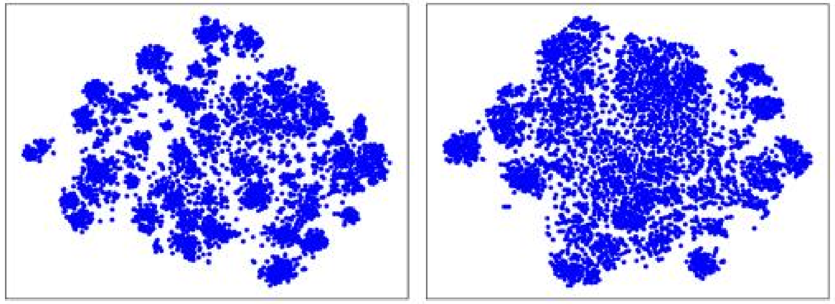
Table 2. Performance comparison of zero-shot image-text retrieval on Flickr30K and COCO datasets. For completeness, we also provide the results of ALIGN [26] which uses 1.8B image-text pairs (1.2B unique images) for pre-training. For text-retrieval (TR) and image-retrieval (IR), we report the average of R@1, R@5 and R@10.

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 [38]	6M	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]	4M	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 [16]	4M	✗	✗	✗	✗	✗	✗	87.9	97.5	98.8	76.3	94.2	96.8
OSCAR [28]	4M	70.0	91.1	95.5	54.0	80.8	88.5	✗	✗	✗	✗	✗	✗
ViLT [24]	4M	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]	4M	✗	✗	✗	✗	✗	✗	89.7	98.4	99.1	74.7	93.47	96.1
SOHO [21]	200K	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 [26]	4M	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	4M	**75.6**	**92.8**	**96.7**	**59.0**	**83.2**	**89.9**	**94.9**	**99.5**	**99.8**	**84.0**	**96.7**	**98.5**
ALIGN [23]	1.2B	77.0	93.5	96.9	59.9	83.3	89.8	95.3	99.8	100.0	84.9	97.4	98.6

Method	#Images	VQA		$NLVR^2$		SNLI-VE	
		test-dev	test-std	dev	test-P	val	test
OSCAR [28]	4M	73.16	73.44	78.07	78.36	✗	✗
UNITER [8]	4M	72.70	72.91	77.18	77.85	78.59	78.28
ViLT [24]	4M	71.26	✗	75.7	76.13	✗	✗
UNIMO [27]	4M	73.29	74.02	✗	✗	80.0	79.1
VILLA [16]	4M	73.59	73.67	78.39	79.30	79.47	79.03
ALBEF [26]	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
VinVL [49]	6M	75.95	76.12	82.05	83.08	✗	✗

Table 3. Performance comparison of fine-tuned image-text retrieval on Flickr30K and COCO datasets. For completeness, we also provide the results of ALIGN [26] which uses 1.8B image-text pairs (1.2B unique images) for pre-training.

Module	Zero-Shot MSCOCO TR	Zero-Shot MSCOCO IR	Zero-Shot Flickr30K TR	Zero-Shot Flickr30K IR	Fine-Tune MSCOCO TR	Fine-Tune MSCOCO IR	Fine-Tune Flickr30K TR	Fine-Tune Flickr30K IR
+IMC (w/o aug) (4M)	71.1	52.2	92.0	78.6	75.0	58.6	94.5	82.9
+IMC (w/o aug) (14M)	72.7	54.1	94.6	83.6	77.9	60.9	96.2	86.0



ALBEF TCL

Figure 1. t-SNE visualization of learned features on the COCO dataset.

better intra-modal representation: uniformly distributed text

m	MSCOCO (5K) Text Retrieval R@1	R@5	R@10	MSCOCO (5K) Image Retrieval R@1	R@5	R@10	Flickr30K (1K) Text Retrieval R@1	R@5	R@10	Flickr30K (1K) Image Retrieval R@1	R@5	R@10
0.995	60.6	85.9	92.2	46.0	74.1	83.1	67.2	89.3	94.4	52.7	79.0	85.7
0.9	59.7	85.1	92.0	45.5	74.1	83.5	68.0	89.6	94.9	53.3	79.8	86.3
0.5	61.6	85.6	92.2	46.5	74.9	84.0	69.7	89.1	94.3	54.7	79.9	86.9
0.0	61.3	85.8	92.7	46.4	75.2	84.4	70.0	88.6	93.0	53.3	78.5	85.6

momentum coefficient: m

Module	Zero-Shot MSCOCO TR	Zero-Shot MSCOCO IR	Zero-Shot Flickr30K TR	Zero-Shot Flickr30K IR	Fine-Tune MSCOCO TR	Fine-Tune MSCOCO IR	Fine-Tune Flickr30K TR	Fine-Tune Flickr30K 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

Table 5. Ablation study of each component on image-text retrieval tasks. The R@1 is reported. For CMA+ITM+MLM, we use the results in ALBEF [26].

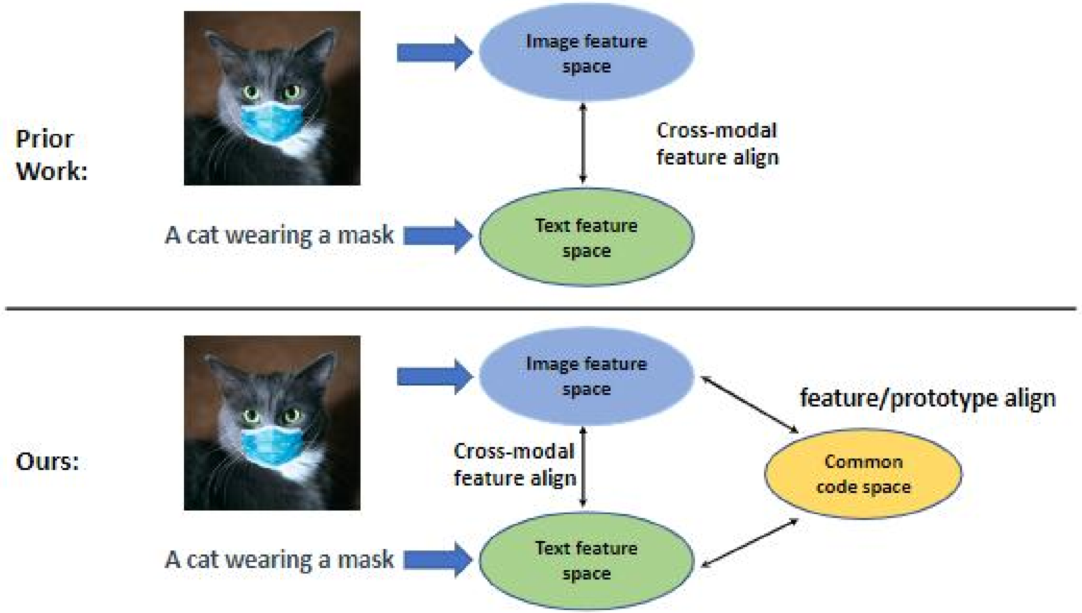
- pooling： local features16*16->GAP->16 patchs
- use last layer

Pooling	Intermediate	Zero-Shot MSCOCO TR	Zero-Shot MSCOCO IR	Zero-Shot Flickr30K TR	Zero-Shot Flickr30K IR	Fine-Tune MSCOCO TR	Fine-Tune MSCOCO IR	Fine-Tune Flickr30K TR	Fine-Tune Flickr30K IR
		71.5	52.9	92.4	79.1	75.7	58.6	94.6	83.3
	✓	71.4	52.9	91.5	77.9	75.7	58.6	94.4	82.3
✓	✓	71.8	53.2	93.2	79.2	75.6	58.7	94.8	82.8
✓		71.4	53.5	93.0	79.6	75.6	59.0	94.9	84.0

Table 6. Ablation study of image patch pooling and intermediate local feature on image-text retrieval. R@1 is reported.

Multi-modal Alignment using Representation Codebook
UTA, Amazon

- Image and text typically reside in different regions of the feature space
  - directly aligning them at instance level is challenging
  - features are still evolving during training.
- align cluster representation (codeword/ learnable prototypes)
  - a dictionary of cluster centers (codebook)
  - optimal transport problem
- teacher-student distillation paradigm



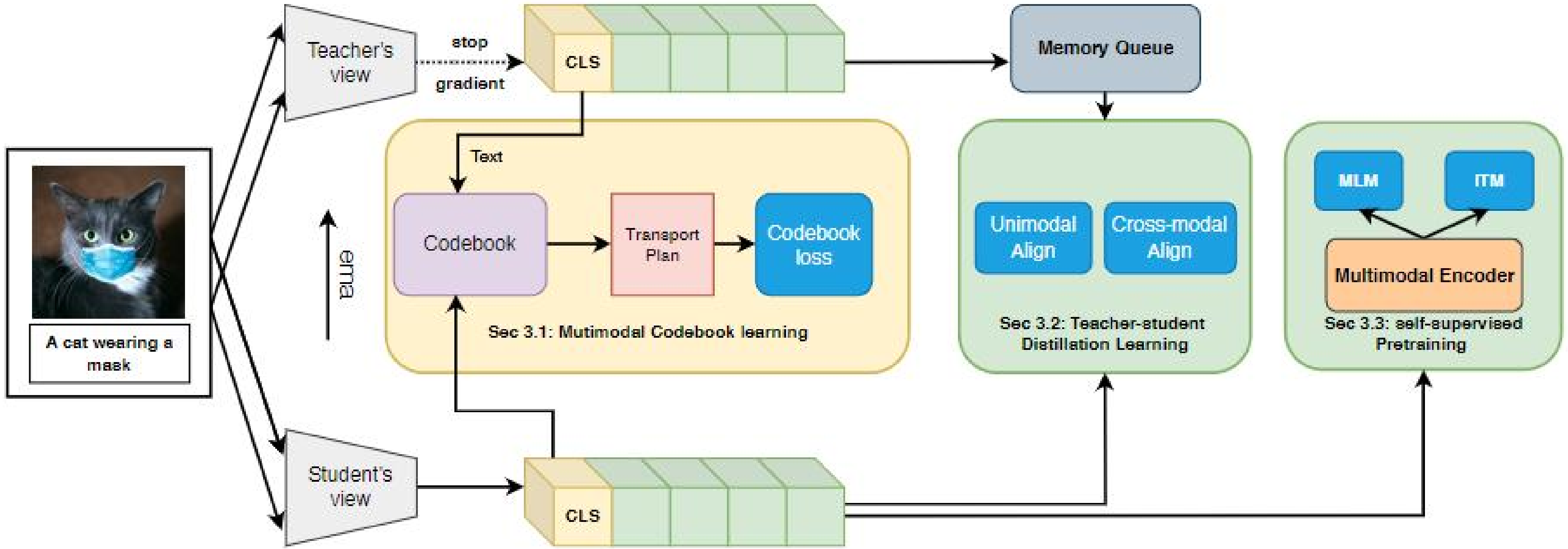


Figure 2. Overview of our framework. For simplicity, we only display a pair of teacher-student encoders (e.g., teacher for the image and student for the text) and similarly for the memory queue. The teacher is updated with an exponential moving average of the student (from the same modality). The codebook helps bridge the gap between the different modalities. The entire framework is end-to-end optimized.

momentum update

$$p_{t2i}(T) = \exp\frac{z_t z_v^{m\top}}{\gamma} \Big/ \sum_{z_v^{m'} \in \mathbf{Q}_v} \exp\frac{z_t z_v^{m'\top}}{\gamma}$$

$$p_{i2t}(I) = \exp\frac{z_v z_t^{m\top}}{\gamma} \Big/ \sum_{z_t^{m'} \in \mathbf{Q}_t} \exp\frac{z_v z_t^{m'\top}}{\gamma}$$

$$p_{i2i}(I) = \exp\frac{z_v z_v^{m\top}}{\gamma} \Big/ \sum_{z_v^{m'} \in \mathbf{Q}_v} \exp\frac{z_v z_v^{m'\top}}{\gamma}$$

$$p_{t2t}(T) = \exp\frac{z_t z_t^{m\top}}{\gamma} \Big/ \sum_{z_t^{m'} \in \mathbf{Q}_t} \exp\frac{z_t z_t^{m'\top}}{\gamma}$$

$$\begin{aligned}\mathcal{L}_{ica} = \mathbb{E}_{I,T\sim p_{\text{data}}} [&H(p_{t2t}, y_{t2t}) + H(p_{i2i}, y_{i2i}) \\ &+ H(p_{t2i}, y_{t2i}) + H(p_{i2t}, y_{i2t})]\end{aligned}$$

$$f_t = \alpha f_t + (1-\alpha) f_s, g_t = \alpha g_t + (1-\alpha) g_s$$

$$\mathcal{L}_{\text{final}} = \mathcal{L}_{\text{mlm}} + \mathcal{L}_{\text{itm}} + \mathcal{L}_{\text{ica}} + \mathcal{L}_{\text{code}}$$

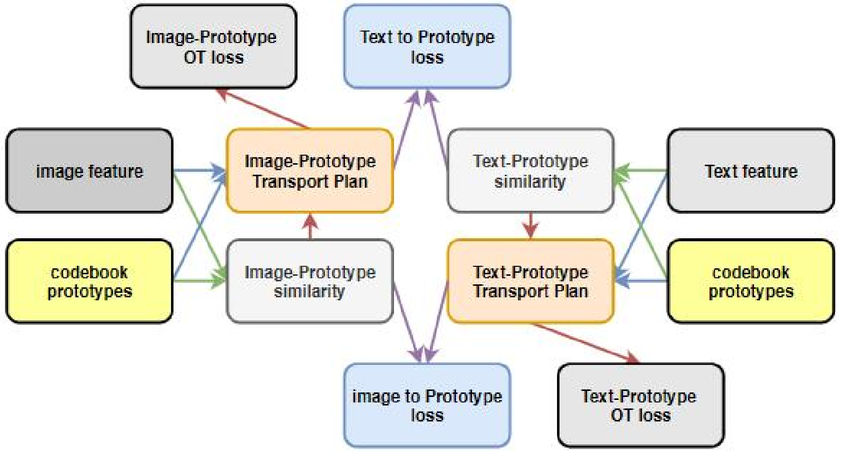


Figure 3. This is the diagram illustrating how to calculate four codebook losses. "→": softmax operator. "→": IPOT algorithm. "→": OT loss. "→": cross entropy.

codebook learning: optimal transport

$$\{c_1, c_2, \ldots, c_K\} \in \mathcal{R}^{d_c \times K},$$

$$\begin{aligned}\mathcal{L}_{\text{t2p}}(\mathbf{Z}_t, \mathbf{C}, \mathbf{T}_{i2p}) &= H(\mathbf{P}_{t2p}, \mathbf{T}_{i2p}), \\ \mathcal{L}_{\text{i2p}}(\mathbf{Z}_v, \mathbf{C}, \mathbf{T}_{t2p}) &= H(\mathbf{P}_{i2p}, \mathbf{T}_{t2p}), \end{aligned} \tag{2}$$

$$\mathbf{P}_{t2p} = \mathbf{SoftMax}(\mathbf{Z}_t\mathbf{C}/\gamma), \mathbf{P}_{i2p} = \mathbf{SoftMax}(\mathbf{Z}_v\mathbf{C}/\gamma)$$

P: student predict
T: teacher

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

$$\begin{aligned}\mathcal{L}_{\text{code}} = &\mathcal{L}_{\text{ot}}(\mathbf{Z}_v^m, \mathbf{C}) + \mathcal{L}_{\text{ot}}(\mathbf{Z}_t^m, \mathbf{C}) \\ &+ \mathcal{L}_{\text{t2p}}(\mathbf{Z}_t, \mathbf{C}, \mathbf{T}_{t2p}) + \mathcal{L}_{\text{i2p}}(\mathbf{Z}_v, \mathbf{C}, \mathbf{T}_{i2p})\end{aligned}$$

Algorithm 2 IPOT Algorithm.

Input: distance/similarity matrix $\mathbf{Z}$, $\mathbf{C}$, ϵ, probability vectors $\boldsymbol{\mu}$, $\boldsymbol{\nu}$

$\boldsymbol{\sigma} = \frac{1}{n}\mathbf{1_n}$, $\mathbf{T}^{(1)} = \mathbf{1}\mathbf{1}^\top$

$D_{ij} = d(z_i, c_j)$, $\mathbf{A}_{ij} = \mathrm{e}^{-\frac{\mathbf{D}_{ij}}{\epsilon}}$

for $t = 1, 2, 3 \ldots$ **do**

$\quad \mathbf{Q} = \mathbf{A} \odot \mathbf{T}^{(t)}$ // $\odot$ is Hadamard product

$\quad$ **for** $k = 1, 2, 3, \ldots K$ **do**

$\quad\quad \boldsymbol{\delta} = \frac{\boldsymbol{\mu}}{n\mathbf{Q}\boldsymbol{\sigma}}$, $\boldsymbol{\sigma} = \frac{\boldsymbol{\nu}}{n\mathbf{Q}^\top\boldsymbol{\delta}}$

$\quad$ **end for**

$\quad \mathbf{T}^{(t+1)} = \mathrm{diag}(\boldsymbol{\delta})\mathbf{Q}\mathrm{diag}(\boldsymbol{\sigma})$

end for

Return $\mathbf{T}$

Algorithm 1 CODIS pseudocode

```
# gs, gt: student/teacher networks for image
# fs, ft: student/teacher networks for text
# C: codebook d-by-K
# Qv, Qt: image/text queue, d-by-M
# tmp, learnable temperature
for (img, txt) in loader: # a minibatch with N samples
    # teacher/student's image view
    img_t, img_s = gt(img), gs(img) # N-by-d

    # teacher/student's text view
    txt_t, txt_s = ft(txt), fs(txt) # N-by-d

    # calculate codebook loss
    I2P, T2P = img_t@C, txt_t@C, # N-by-K
    Tg, Tf = IPOT(1-I2P), IPOT(1-T2P) # refer to Algo 2
    L_ot = Trace(I2P.t()@Tg).sum() + Trace(T2P.t()@Tf).sum()
    L_code = H(img_s@C, Tg) + H(txt_s@C, Tf) + L_ot

    # calculate alignment loss
    L_cross = H(img_s@Qt, img_t@Qt) + H(txt_s@Qv, txt_t@Qv)
    L_unimo = H(img_s@Qv, img_t@Qv) + H(txt_s@Qt, txt_t@Qt)
    L_align = L_cross + L_unimo

    # enqueue/dequeue
    update_queue(Qv, img_t, Qt, txt_t)

    # pretraining loss
    L_pretrain = L_itm + L_mlm

    loss = L_code + L_align + L_pretrain
    loss.backward() # back-propagate

    # student, teacher updates
    update(gs, fs) # SGD
    ema(gs, gt, fs, ft) # momemtum update


def H(s, t):
    t = t.detach() # stop gradient
    s = softmax(s / tmp, dim=1)
    return - (t * log(s)).sum(dim=1).mean()
```

Table 1. Performance comparison of zero-shot image-text retrieval on MSCOCO and Flickr30K datasets.

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

Table 2. Performance comparison of fine-tuned image-text retrieval on MSCOCO and Flickr30K datasets.

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

Table 3. Comparison with variety of state-of-the-art methods on downstream vision-language tasks: VQA, NVLR^2, SNLI-VE.

Method	VQA		NLVR^2		SNLI-VE	
	test-dev	test-std	dev	test-P	val	test
VisualBERT [26]	70.80	71.00	67.40	67.00	-	-
LXMERT [43]	72.42	72.54	74.90	74.50	-	-
12-in-1 [32]	73.15	-	-	78.87	-	76.95
UNITER [8]	72.70	72.91	77.18	77.85	78.59	78.28
ViLT [22]	70.94	-	75.24	76.21	-	-
OSCAR [28]	73.16	73.44	78.07	78.36	-	-
VILLA [14]	73.59	73.67	78.39	79.30	79.47	79.03
ALBEF 4M [25]	74.54	74.70	80.24	80.50	80.14	80.30
Ours	**74.86**	**74.97**	**80.50**	**80.84**	**80.47**	**80.40**

Table 4. Efficiency of our approach under limited pretraining regime using only MSCOCO.

	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

Table 5. Performance comparison of zero-shot image-text retrieval on Flickr30K and COCO datasets for ablation study.

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**

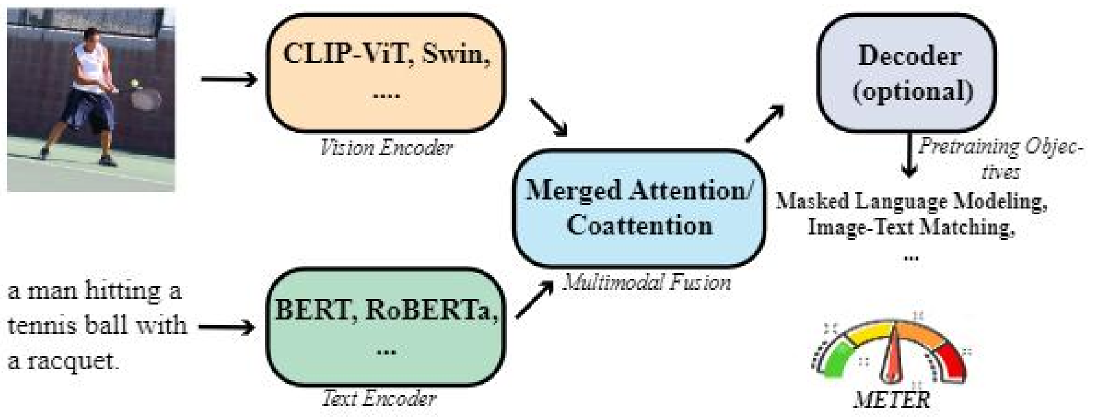
4000

<TCL

An Empirical Study of Training End-to-End Vision-and-Language Transformers

UCLA,Microsoft

- systematically investigate how to design and pre-train transformer-based VL model
    - VIT -> vital role than language transformer
    - the preformance of VIT on ImageNet classification is not a good indicator on VL
    - in multimodal fusion cross attention > self attention
    - encoder-only > encoder-decoder for VQA and zero-shot ITR
    - MIM is not a critical pre-training objective for VLP



Model	Vision Encoder	Text Encoder	Multimodal Fusion	Decoder	Pre-training Objectives
ViLBERT [31]	OD+Xformer	Xformer	Co-attn.	✗	MLM+ITM+MIM
LXMERT [43]					MLM+ITM+MIM+VQA
VisualBERT [23]	OD	Emb.	Merged-attn.		MLM+ITM
VL-BERT [23]					MLM+MIM
UNITER [4]					MLM+ITM+MIM+WRA
OSCAR [4]					MLM+ITM
VinVL [55]					MLM+ITM
VL-T5 [5]				✓	MLM+ITM+VQA+Grounding+Captioning
PixelBERT [16]	CNN	Emb.	Merged-attn.	✗	MLM+ITM
SOHO [15]					MLM+ITM+MIM
CLIP-ViL [40]					MLM+ITM+VQA
SimVLM [50]				✓	PrefixLM
ViLT [19]	Patch Emb.	Emb.	Merged-attn.	✗	MLM+ITM
Visual Parsing [52]	Xformer				MLM+ITM+MIM
ALBEF [22]		Xformer	Co-attn.		MLM+ITM+ITC
METER (Ours)					MLM+ITM
CLIP [34]	CNN/Xformer	Xformer	None	✗	ITC
ALIGN [17]	CNN				

Table 1. Glossary of representative VLP models. OD: object detection. Xformer: transformer. Emb.: embedding. MLM/MIM: masked language/image modeling. ITM: image-text matching. WRA: word-region alginment. ITC: image-text contrastive learning.

tasks: VQA
NLVR
VE
ITR

Dataset	#Images	#Captions
COCO	113K	567K
Visual Genome	108K	5.4M
Conceptual Captions	3.1M	3.1M
SBU Captions	875K	875K

Table 2. Statistics of the pre-training datasets.

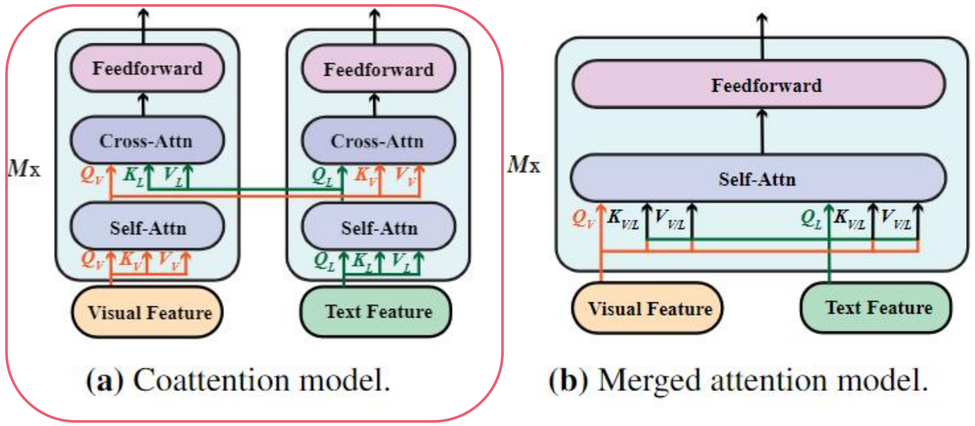


Figure 2. Illustration of two types of multimodal fusion modules: (a) Co-attention, and (b) Merged attention.

no decoder and no parameter shared

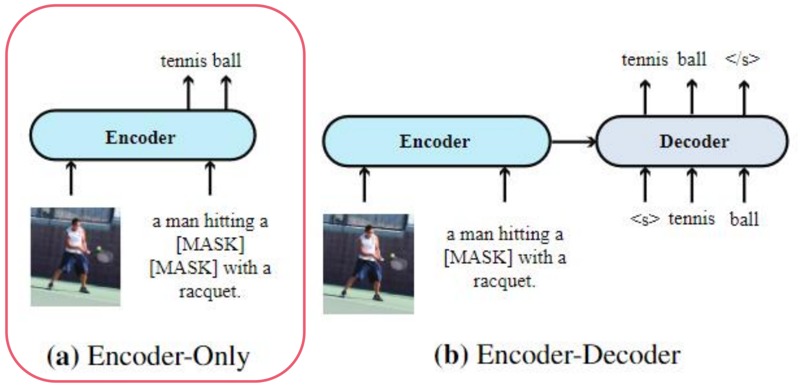


Figure 3. Illustration of the Encoder-Only and Encoder-Decoder model architectures for VLP.

MLM+ITM.

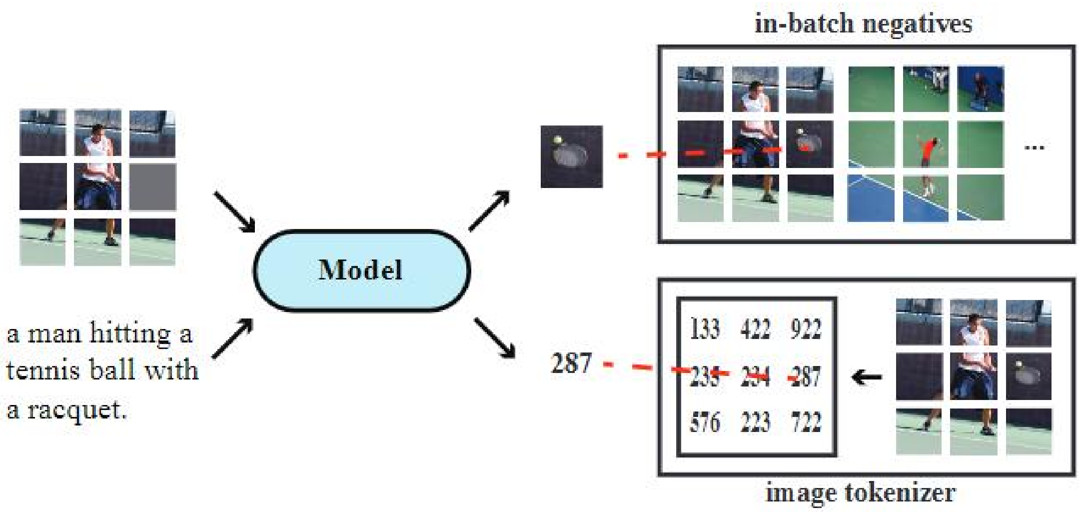


Figure 4. Illustration of masked patch classification.

$$p(\mathbf{v}_i^k|[\mathbf{v}^{k,mask};\mathbf{l}^k]) = \frac{e^{h(\mathbf{v}_i^k)^\top c(\mathbf{v}_i^k)}}{\sum_{j,k'} e^{h(\mathbf{v}_i^k)^\top c(\mathbf{v}_j^{k'})}}.$$

NCE

Masked Image Modeling -> masked patch classification

- MSE feature, predict object label
- ALBEF dont use,
- ViLT: maksed patch regression is not helpul

use VQVAE model in DALLE
predict the discrete tokens

Explorations without VLP

- initialize the bottom layers with specific pre-trained vision and text encoders, and randomly initialize the top layers.
- T: V-> CLIP-ViT-224/32   V: T-> RoBERTa

Text Enc.	VQAv2 Acc.	VE Acc.	IR R@1	TR R@1	SQuAD EM	MNLI Acc.
BERT	69.56	76.27	49.60	66.60	76.3	84.3
RoBERTa	69.69	76.53	49.86	68.90	84.6	87.6
ELECTRA	69.22	76.57	41.80	58.30	86.8	**88.8**
DeBERTa	69.40	**76.74**	51.50	67.70	**87.2**	**88.8**
ALBERT	**69.94**	76.20	52.20	68.70	86.4	87.9
Emb-only	67.13	74.85	49.06	68.20	-	-
CLIP	69.31	75.37	**54.96**	**73.80**	-	-

Table 3. Comparisons of different text encoders without VLP. CLIP-ViT-224/32 is used as the vision encoder. All the text encoders are in base model size, except ALBERT, which is xlarge. Emb-only: only using word embeddings as text encoder. IR/TR: Flickr30k image/text retrieval. EM: exact match. The results of SQuAD and MNLI are copied from their corresponding papers. All the results on VL tasks are from their test-dev/val sets.

RoBERTa -> most robust performace

Vision Encoder	VQAv2	VE	IR	TR	ImageNet
ViT B-384/16	69.09	76.35	40.30	59.80	83.97
DeiT B-384/16	68.92	75.97	33.38	50.90	82.9
Dis. DeiT B-384/16	67.84	76.17	34.84	52.10	85.2
CaiT M-384/32	71.52	76.62	38.96	61.30	86.1
VOLO 4-448/32	71.44	76.42	40.90	61.40	**86.8**
Swin B-384/32	**72.38**	**77.65**	52.30	69.50	86.4
CLIP B-224/32	69.69	76.53	49.86	68.90	-
CLIP B-224/16	71.75	77.54	**57.64**	**76.90**	-
BEiT B-224/16	68.45	75.28	32.24	59.80	85.2

Table 4. Comparisons of different vision encoders without VLP. RoBERTa is used as the default text encoder. IR/TR: Flickr30k image/text retrieval. The results of ImageNet classification are copied from their corresponding papers. All the results on VL tasks are from their test-dev/val sets.

swin without VLP is already comparable to some VLP model in VQA

DEBERTa and BEiT achieve better classification performance but not suitable for VL

Text Enc.	Vision Enc.	VQAv2	Flickr-ZS IR	Flickr-ZS TR
Embed-only	CLIP-32	73.99	60.32	90.38
BERT	CLIP-32	74.98	66.08	78.10
	CLIP-16	76.70	74.52	87.20
RoBERTa	CLIP-32	74.67	65.50	76.60
	CLIP-16	**77.19**	**76.64**	**89.60**
	Swin	76.43	71.68	85.30

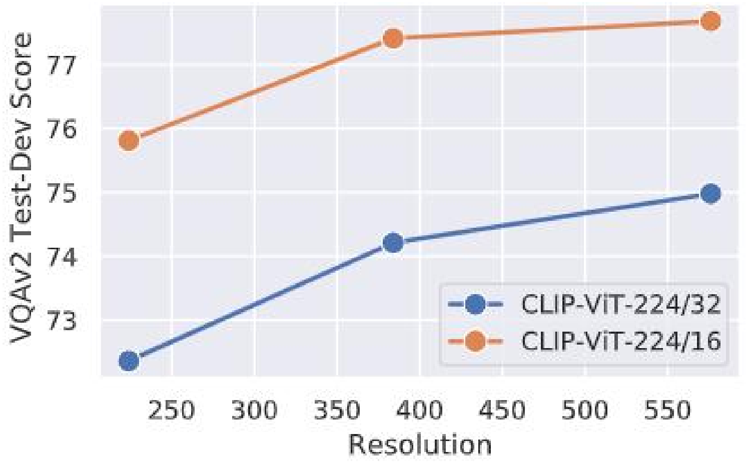
Table 5. Comparisons of different vision and text encoders with VLP. Results on VQAv2 are on test-dev set. ZS: zero-shot.

the difference between BERT and RoBERTa seems to be diminished
>embed-only

Bottom LR	Top LR	VQAv2	Flickr-ZS IR	Flickr-ZS TR
1e-5	1e-5	73.16	48.80	63.70
2e-5	2e-5	73.66	53.14	67.20
3e-5	3e-5	73.77	56.48	70.90
5e-5	5e-5	73.54	52.48	65.90
1e-5	5e-5	**74.98**	**66.08**	**78.10**

Table 6. Using different learning rates for the randomly-initialized and pre-trained parameters is better than using the same learning rate. Results on VQAv2 are on test-dev set. ZS: zero-shot.

use a larger learning rate for the randomly initialized parameters than parameters initialized with pre-trained models



increasing the image resolution during finetuning can improve the model performance by a large margin

Fusion	Decoder	VQAv2	Flickr-ZS IR	Flickr-ZS TR
Merged attention	✗	74.00	57.46	73.10
Co-attention	✗	**74.98**	**66.08**	**78.10**
Co-attention	✓	74.73	48.96	71.60

Mmerged = 12 and Mco = 6

Model	VQAv2	Flickr-ZS IR	Flickr-ZS TR
MLM	74.19	-	-
ITM	72.63	53.74	71.00
MLM+ITM	**74.98**	**66.08**	**78.10**
MLM+ITM + MIM with in-batch negatives	74.01	62.12	76.90
MLM+ITM + MIM with discrete code	74.21	59.80	76.30

Table 10. Masked language modeling (MLM) and image-tech-matching (ITM) can both improve the model performance, but both of our designed masked image modeling (MIM) objectives lead to degraded performance on downstream tasks. Results on VQAv2 are on test-dev set. ZS: zero-shot.

MIM drop

- the conflicts between different objectives.
- image patches can be noisy

Model	VQAv2 test-dev	VQAv2 test-std	NLVR2 dev	NLVR2 test	SNLI-VE dev	SNLI-VE test	Flickr-ZS IR@1	Flickr-ZS IR@5	Flickr-ZS IR@10	Flickr-ZS TR@1	Flickr-ZS TR@5	Flickr-ZS TR@10
Pre-trained with >10M images												
ALBEF (14M) [22]	75.84	76.04	82.55	83.14	80.80	80.91	82.8	96.3	98.1	94.1	99.5	99.7
SimVLM$_{\text{BASE}}$ (1.8B) [50]	77.87	78.14	81.72	81.77	84.20	84.15	-	-	-	-	-	-
SimVLM$_{\text{HUGE}}$ (1.8B) [50]	80.03	80.34	84.53	85.15	86.21	86.32	-	-	-	-	-	-
Pre-trained with <10M images												
UNITER$_{\text{LARGE}}$ [4]	73.82	74.02	79.12	79.98	79.39	79.38	68.74	89.20	93.86	83.60	95.70	97.70
VILLA$_{\text{LARGE}}$ [11]	74.69	74.87	79.76	81.47	80.18	80.02	-	-	-	-	-	-
UNIMO$_{\text{LARGE}}$ [24]	75.06	75.27	-	-	**81.11**	80.63	-	-	-	-	-	-
VinVL$_{\text{LARGE}}$ [55]	76.52	76.60	**82.67**	**83.98**	-	-	-	-	-	-	-	-
PixelBERT [16]	74.45	74.55	76.5	77.2	-	-	-	-	-	-		
CLIP-ViL (ResNet50x4) [40]	76.48	76.70	-	-	80.61	80.20	-	-	-	-	-	-
ViLT [55]	71.26	-	75.70	76.13	-	-	55.0	82.5	89.8	73.2	93.6	96.5
Visual Parsing [52]	74.00	74.17	77.61	78.05	-	-	-	-	-	-	-	-
ALBEF (4M) [22]	74.54	74.70	80.24	80.50	80.14	80.30	76.8	93.7	96.7	90.5	**98.8**	**99.7**
METER-Swin	76.43	76.42	82.23	82.47	80.61	80.45	71.68	91.80	95.30	85.30	97.70	99.20
METER-CLIP-ViT	**77.68**	**77.64**	82.33	83.05	80.86	**81.19**	**79.60**	**94.96**	**97.28**	**90.90**	98.30	99.50

Table 8. Comparisons with previous models on visual question answering, visual reasoning, visual entailment, and Flickr30k zero-shot retrieval tasks. The best scores are in **bold**, and the second best scores are underlined.

Model	Flickr IR@1	Flickr IR@5	Flickr IR@10	Flickr TR@1	Flickr TR@5	Flickr TR@10	COCO IR@1	COCO IR@5	COCO IR@10	COCO TR@1	COCO TR@5	COCO TR@10
Pre-trained with >10M images												
ALBEF (14M) [22]	85.6	97.5	98.9	95.9	99.8	100.0	60.7	84.3	90.5	77.6	94.3	97.2
Pre-trained with <10M images												
UNITER$_{\text{LARGE}}$ [4]	75.56	94.08	96.76	87.30	98.00	99.20	52.93	79.93	87.95	65.68	88.56	93.76
VILLA$_{\text{LARGE}}$ [11]	76.26	94.24	96.84	87.90	97.50	98.80	-	-	-	-	-	-
UNIMO$_{\text{LARGE}}$ [24]	78.04	94.24	97.12	89.40	98.90	99.80	-	-	-	-	-	-
VinVL$_{\text{LARGE}}$ [55]	-	-	-	-	-	-	**58.8**	**83.5**	**90.3**	75.4	92.9	96.2
PixelBERT [16]	71.5	92.1	95.8	87.0	98.9	99.5	50.1	77.6	86.2	63.6	87.5	93.6
ViLT [55]	64.4	88.7	93.8	83.5	96.7	98.6	42.7	72.9	83.1	61.5	86.3	92.7
Visual Parsing [52]	73.5	93.1	96.4	87.0	98.4	99.5	-	-	-	-	-	-
ALBEF (4M) [22]	**82.8**	**96.7**	**98.4**	**94.3**	99.4	99.8	56.8	81.5	89.2	73.1	91.4	96.0
METER-Swin	79.02	95.58	98.04	92.40	99.00	99.50	54.85	81.41	89.31	72.96	92.02	96.26
METER-CLIP-ViT	82.22	96.34	**98.36**	**94.30**	**99.60**	**99.90**	57.08	82.66	90.07	**76.16**	**93.16**	**96.82**

Table 9. Comparisons with previous models on Flickr30k and COCO retrieval tasks. The best scores are in **bold**, and the second best scores are underlined.

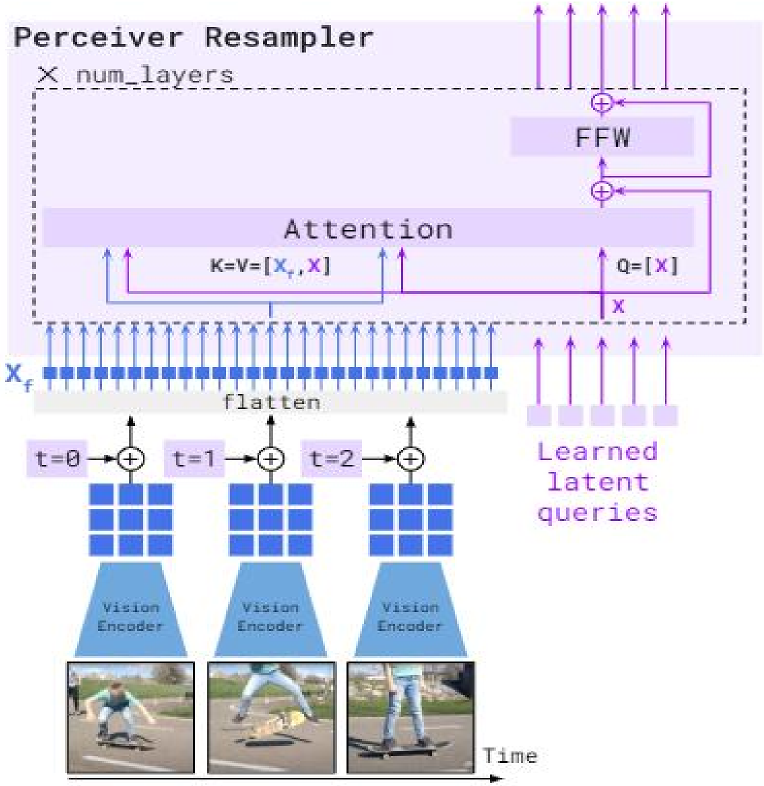
ALBEF has specially-designed objectives for retrieval but it use CLIP_V (300M VLP)->no ITC

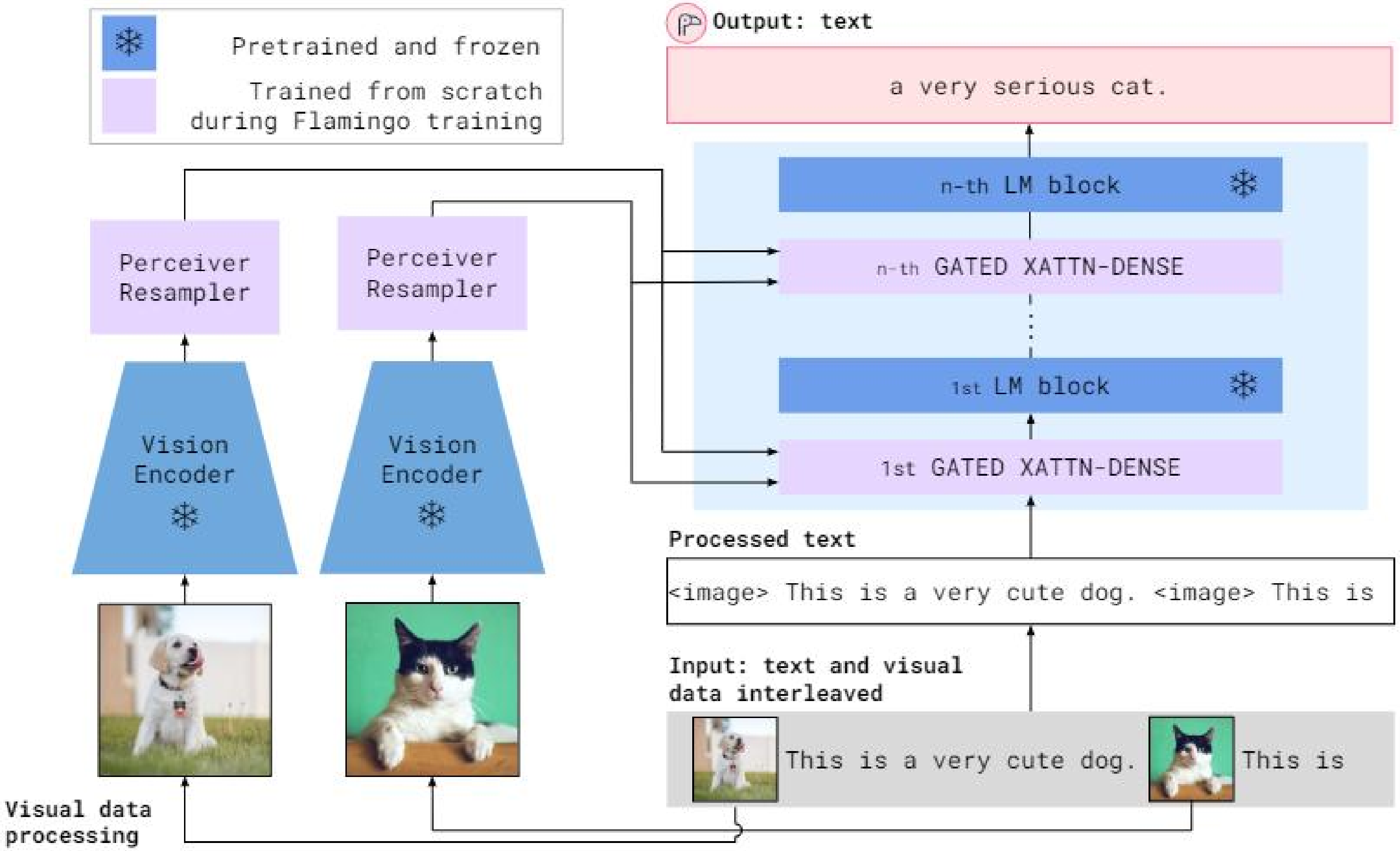
Part III: After CVPR

Flamingo (2022.4) ： Analogical learning

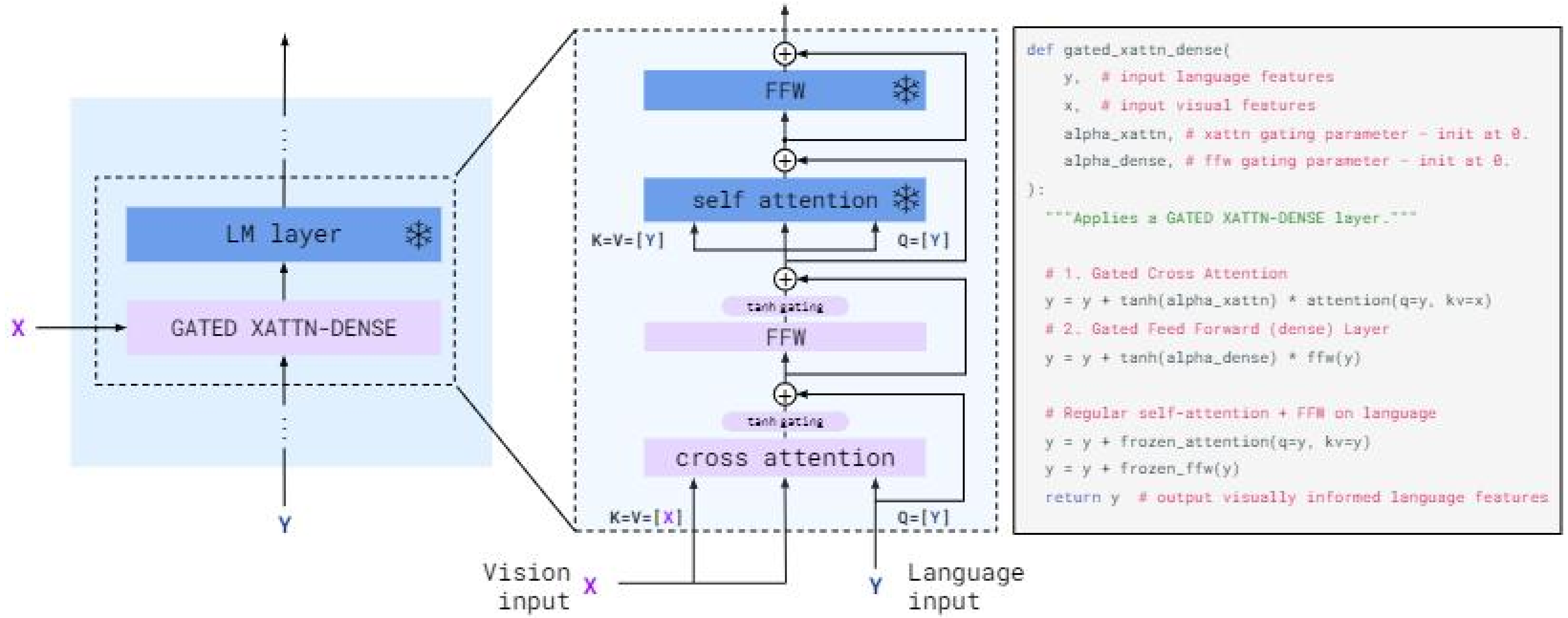
Multimodal generative modeling

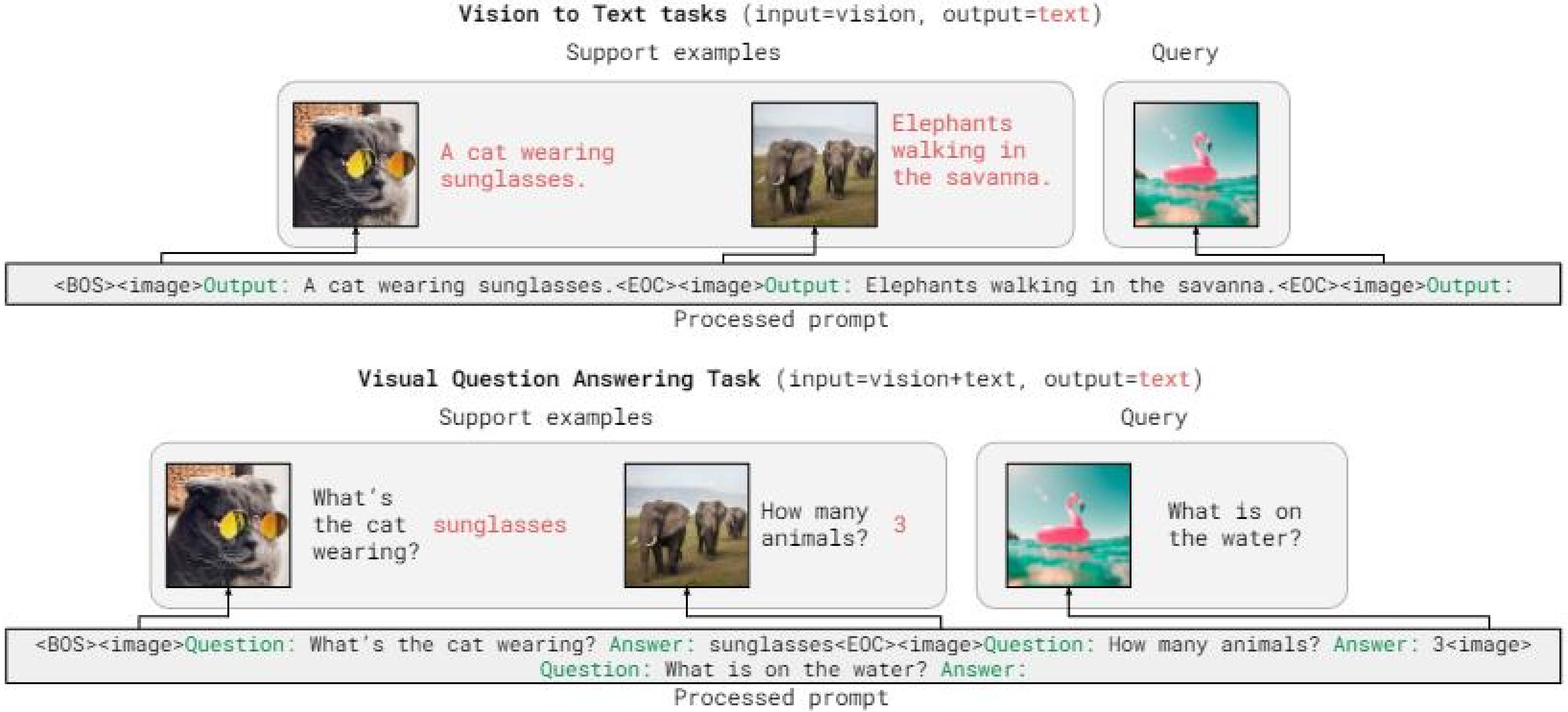
- Unifying strong single-modal models
  - Interleave CA and SA (frozen) -> initialization->specific gating (stability/performance)
- Supporting both images and videos
  - Native addition visual token sequence - > memory limitation.
  - Local 2D priors (inductive bias) - >improve efficiency but not suitable for text.
  - -> Perceiver-based architecture (fixed number : a hundred tokens)
- Mixture of dataset： M3W + ALIGN + …= 3B
- Frozen encoder： V： CLIP， T： BERT decoder



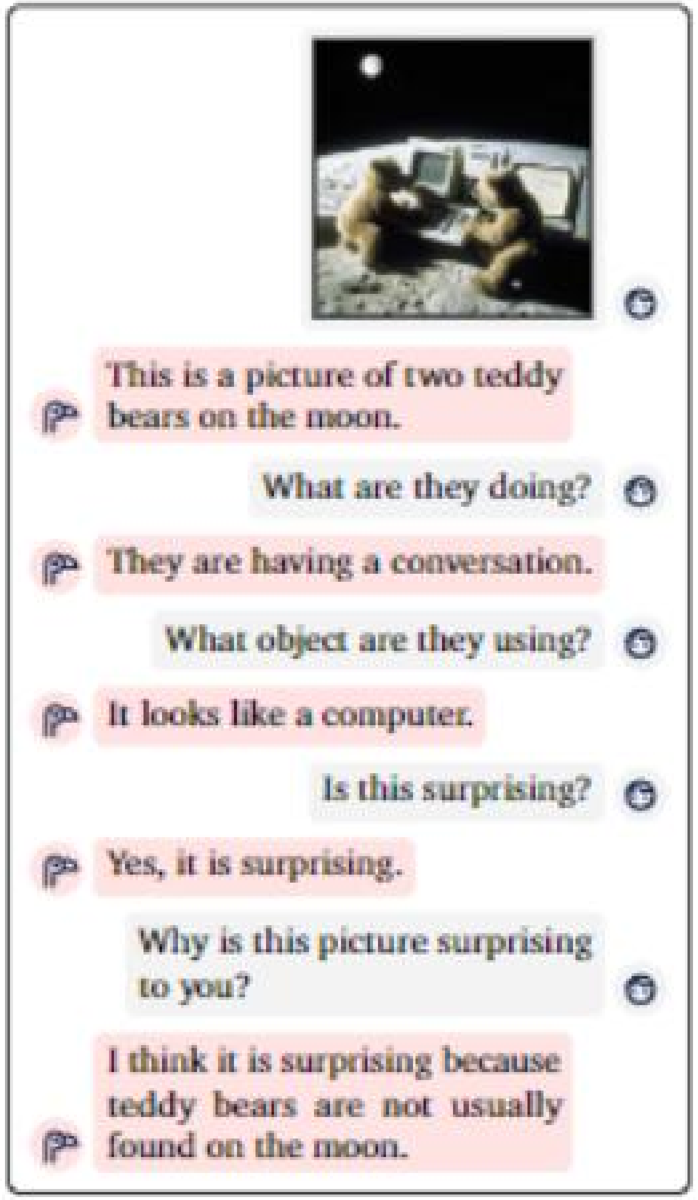


Interleaved prompt





interleaved visual and textual data



MoE (2022.6): multi-task Generalist Models -> Multi-architecture (task-interference)

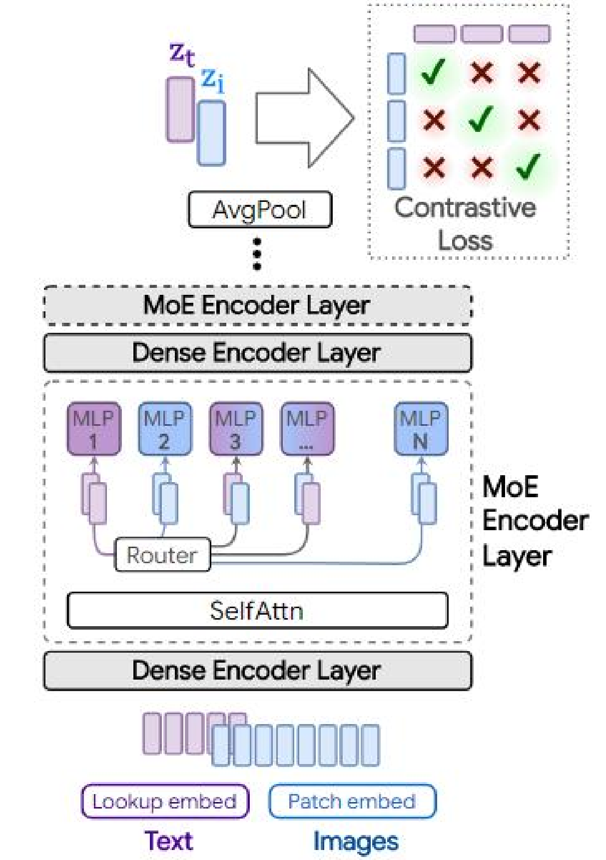


Figure 1: LIMoE, a sparsely activated multimodal model, processes both images and texts, utilising conditional computation to allocate tokens in a modality-agnostic fashion.

Google: LiMoE

Only contrastive pretrain

Selected (5) and total (64) experts

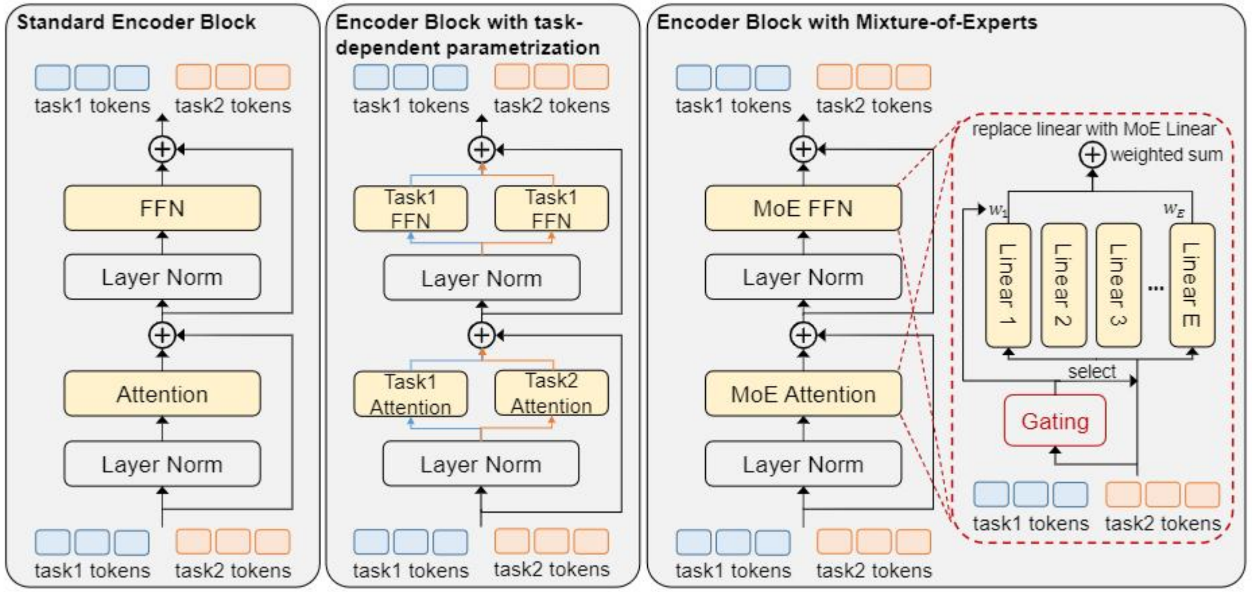
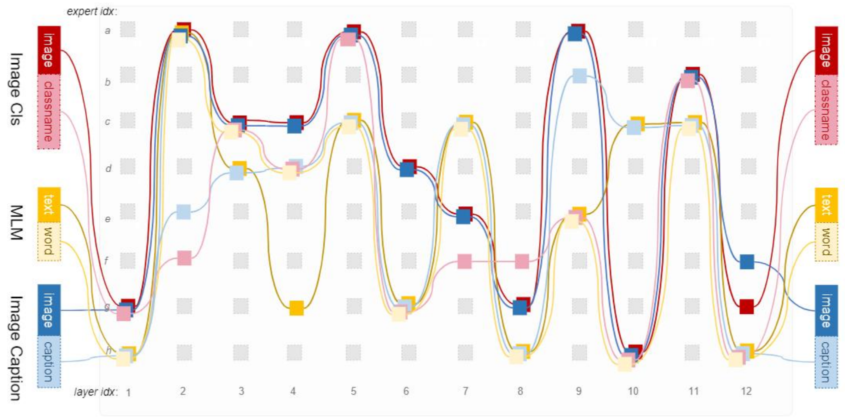


Figure 1: Comparisons of fully-shared standard encoder block, task-specific encoder block with task-dedicated parameters, and encoder block with efficient MoE parameterization.



(a) Gating decisions of the self-attention layers for Uni-Perceiver-MoE-Ti.

Sensetime:Uni-Perceiver-MoE

Index	Descriptions	Yes	No
0	Visual modality exists in the inputs of the current task.	1	0
1	Text modality exists in the inputs of the current task.	1	0
2	Visual modality exists in the targets of the current task.	1	0
3	Text modality exists in the targets of the current task.	1	0
4	The modality of current token is visual.	1	0
5	The modality of current token is text.	1	0
6	The attention mask of the current token is causal.	1	0
7	The current token comes from the inputs, not the targets.	1	0

UNIFIED-IO:A Unified Model For Vision, Language, And Multi-model Tasks

Figure 1: UNIFIED-IO is a sequence-to-sequence model that performs a variety of tasks in computer vision and NLP using a unified architecture without a need for either task or modality specific branches. This broad unification is achieved by homogenizing every task's input and output into a sequence of discrete vocabulary tokens. UNIFIED-IO supports modalities as diverse as images, masks, keypoints, boxes, and text, and tasks as varied as depth estimation, inpainting, semantic segmentation, captioning, and reading comprehension.

https://unified-io.allenai.org/

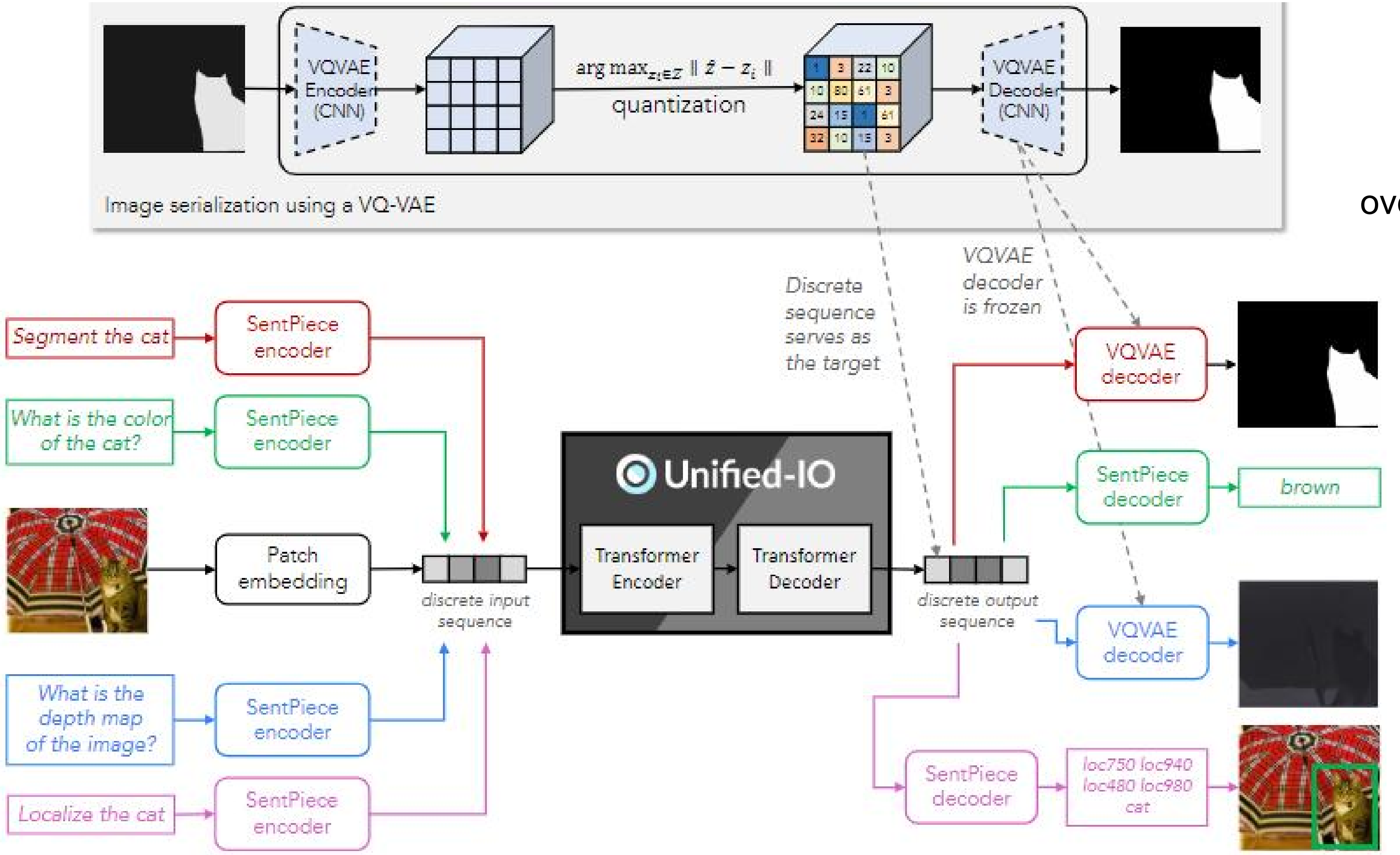


Figure 2: **Unified-IO.** A schematic of the model with four demonstrative tasks: object segmentation, visual question answering, depth estimation and object localization.

over 80 diverse datasets

+pixel (refer expression) inpaint/synthesis
+Depth estimation
+NLP tasks

	NYUv2	ImageNet	Place365	VQAv2	OkVQA	A-OkVQA	VizWizQA	VizWizGround	Swig	SNLI-VE	VisComet	Nocaps	COCO	COCO	MRPC	BoolQ	SciTail
Split	val	val	val	test-dev	test	test	test-dev	test-std	test	val	val	val	val	test	val	val	test
Metric	RMSE	Acc.	Acc.	Acc.	Acc.	Acc.	Acc.	IOU	Acc.	Acc.	CIDEr	CIDEr	CIDEr	CIDEr	F1	Acc	Acc
Unified SOTA	UViM	-	-	-	Flamingo	-	Flamingo	-	-	-	-	-	-	-	T5	PaLM	-
	0.467	-	-	-	57.8	-	49.8	-	-	-	-	-	-	-	92.20	92.2	-
UNIFIED-IO$_{\text{SMALL}}$	0.649	42.8	38.2	57.7	31.0	24.3	42.4	35.5	17.3	76.5	-	45.1	80.1	-	84.9	65.9	87.4
UNIFIED-IO$_{\text{BASE}}$	0.469	63.3	43.2	61.8	37.8	28.5	45.8	50.0	29.7	85.6	-	66.9	104.0	-	87.9	70.8	90.8
UNIFIED-IO$_{\text{LARGE}}$	0.402	71.8	50.5	67.8	42.7	33.4	47.7	54.7	40.4	86.1	-	87.2	117.5	-	87.5	73.1	93.1
UNIFIED-IO$_{\text{XL}}$	0.385	79.1	53.2	77.9	54.0	45.2	57.4	65.0	49.8	91.1	21.2	100.0	126.8	122.3	89.2	79.7	95.7
Single or fine-tuned SOTA	BinsFormer	CoCa	MAE	CoCa	KAT	GPV2	Flamingo	MAC-Caps	JSL	OFA	SVT	CoCa	-	OFA	Turing NLR	ST-MOE	DeBERTa
	0.330	91.00	60.3	82.3	54.4	38.1	65.7	27.3	39.6	91.0	18.3	122.4	-	145.3	93.8	92.4	97.7

Table 3: Comparing the jointly trained UNIFIED-IO to specialized and benchmark fine-tuned state of the art models across Vision, V&L and Language tasks. Benchmarks used for evaluation are: NYUv2 (Nathan Silberman & Fergus, 2012), ImageNet (Deng et al., 2009), Places365 (Zhou et al., 2017), VQA 2.0 (Goyal et al.), A-OKVQA (Schwenk et al., 2022), VizWizVQA (Gurari et al., 2018), VizWizGround (Chen et al., 2022a), Swig (Pratt et al., 2020), SNLI-VE (Xie et al., 2019), VisComet (Park et al., 2020), Nocaps (Agrawal et al., 2019), COCO Captions (Chen et al., 2015), MRPC (Dolan & Brockett, 2005), BoolQ (Clark et al., 2019), and SciTail (Khot et al., 2018).

embedding align
2017 and before

cross attention
2018:SCAN

Transformer + Pretrain
2019.9 vilBERT, UNITER

2020.12 OSCAR,VinVL
stronger region feature

VQA
fusion task->more dataset

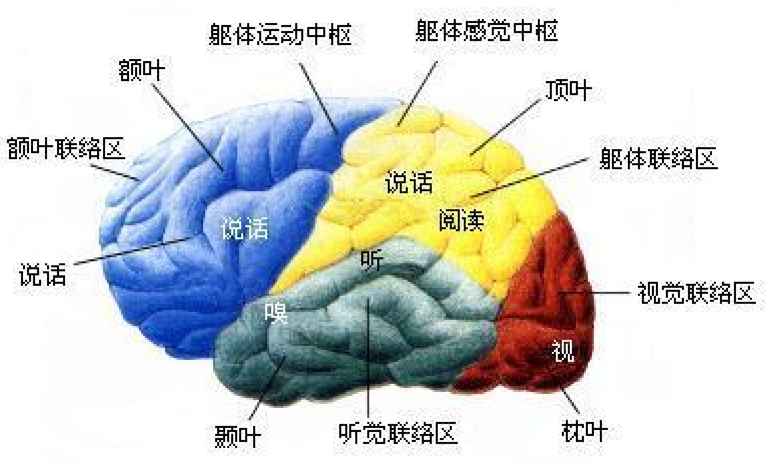
IT retriveal/IT matching
rely modal encoder

2021.2 CLIP
More data 400M

2021.2 ViLT (ICML)
Grid/Patch feature

2021.7 ALBEF
encoder-decoder

2022： AGI?
Foundation model



大脑皮层功能区示意图

multi-task --> Generalist Models

saleforce, facebook -> *encoder-decoder*
OpenAI -> *analogical learning*
Google (pathways)
Alibaba (OFA) -> *MoE*
Sensetime
allenai -> *multi data encoder - transformer - multi data decoder*

https://github.com/phellonchen/awesome-Vision-and-Language-Pre-training