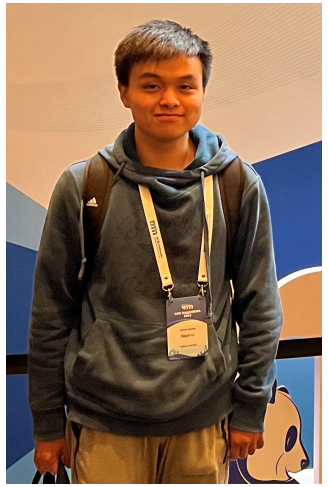
从多模态联合预训练到多模态大语言模型：架构、训练、评测、趋势概览

魏忠钰（Wei, Zhongyu）

复旦大学
数据智能与社会计算实验室 (Fudan DISC)
自然语言处理组 (Fudan-NLP)

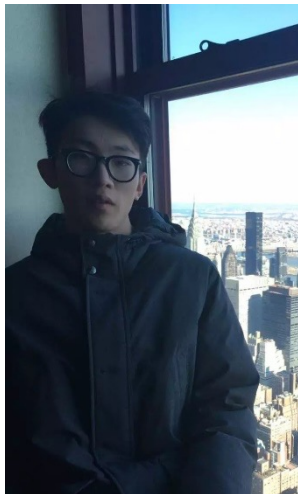
2023年12月03日
中国中文信息学会前沿技术讲习班

合作者



李泽君

张霁雯



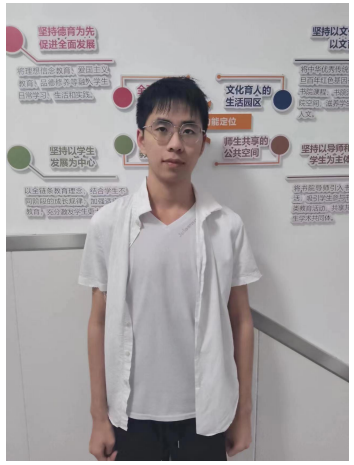
王晔

罗瑞璞

杜梦飞

吴斌浩

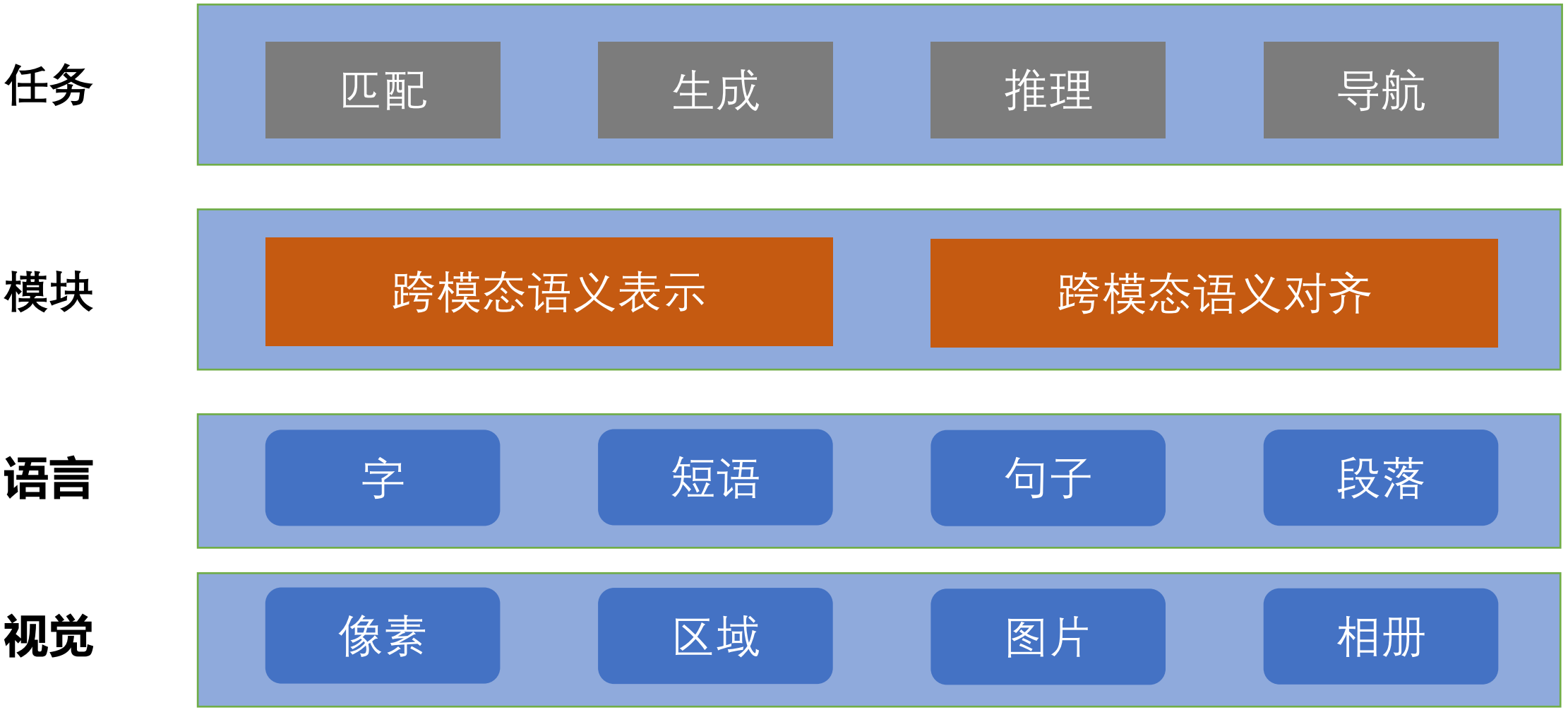
周呈星



陈汉夫

目录

跨视觉语言模态的研究场景



图像文本的语义匹配

- 给定一张图片，从句子集合中检索语义相关的句子。
- 给定一个句子，从图片集合中检索语义相关的图片。
- 评测指标: R@1(Recall@1), R@5, R@10

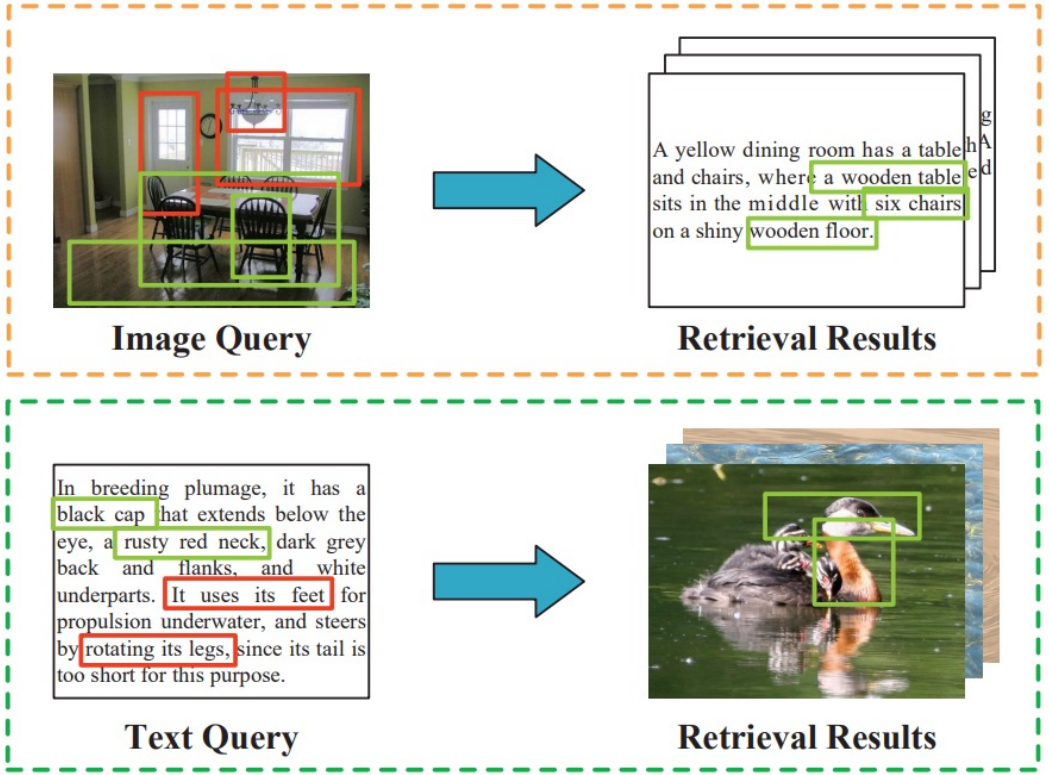
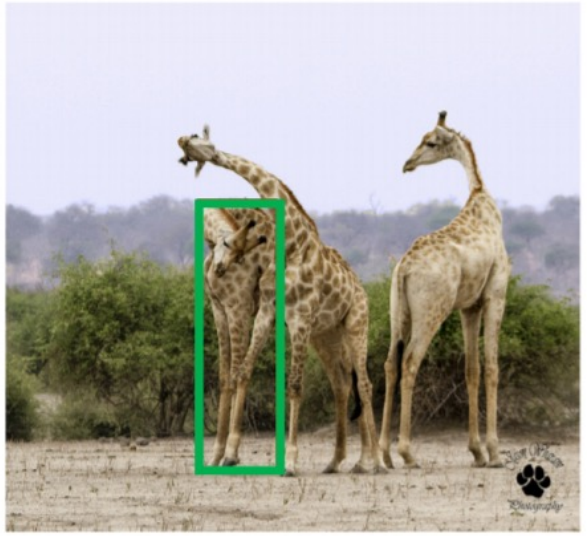


	Image-train	Image-dev	Image-test	caption
MSCOCO	113,287	5,000	5,000	5 for each image
Flickr30K	29,000	1,000	1,000	

视觉指代理解（Visual Referring Expression）

- 给定一个语言表达，确定图片中指代的目标物体。
- 重叠比例Intersection over Union (IoU)：真实和预测的物体框。
- 如果 IoU 超过 0.5, 被认为真，否则为假。



RefCOCO:
1. giraffe on left
2. first giraffe on left

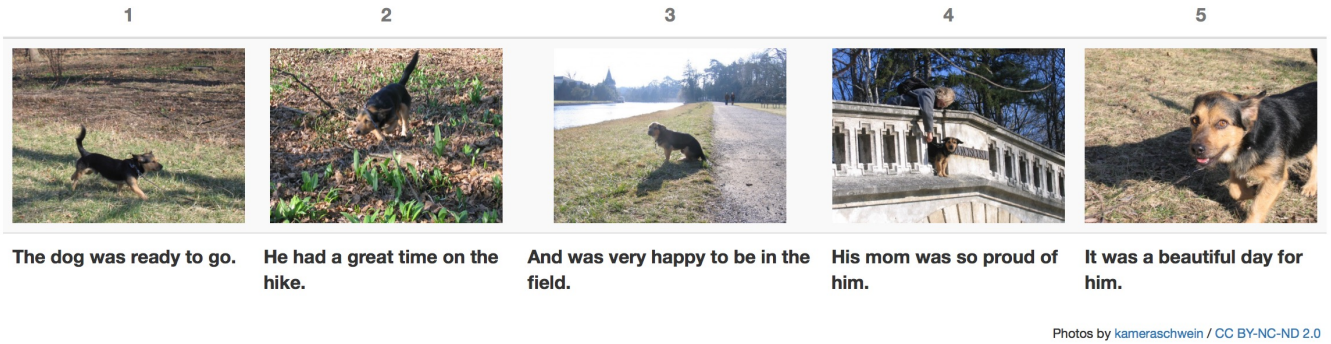
RefCOCO+:
1. giraffe with lowered head
2. giraffe head down

RefCOCOg:
1. an adult giraffe scratching its back with its horn
2. giraffe hugging another giraffe

	图片数	目标物体数	文本表达	平均长度
RefCOCO	50,000	19,994	142,209	3.61
RefCOCO+	49,856	19,992	141,4564	3.53
RefCOCOg	26,711	54,822	85,474	8.43

基于视觉的文本生成

- 图片描述生成
- 相册故事生成
- 图片对话生成
- 评测指标: BLUE, ROUGE, MEOTER, SPICE




A person riding a motorcycle on a dirt road.

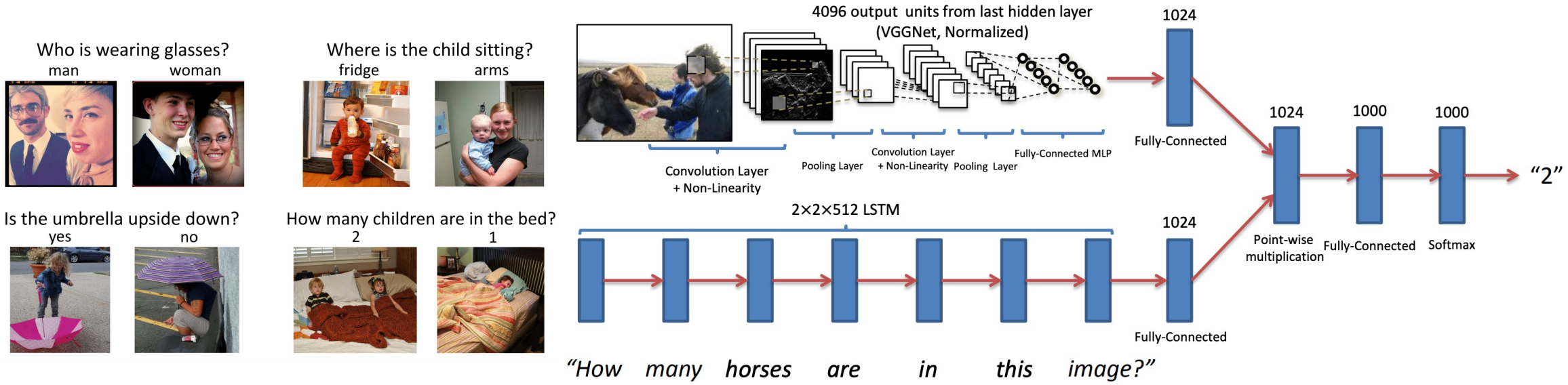

Two dogs play in the grass.


A group of young people playing a game of frisbee.


Two hockey players are fighting over the puck.

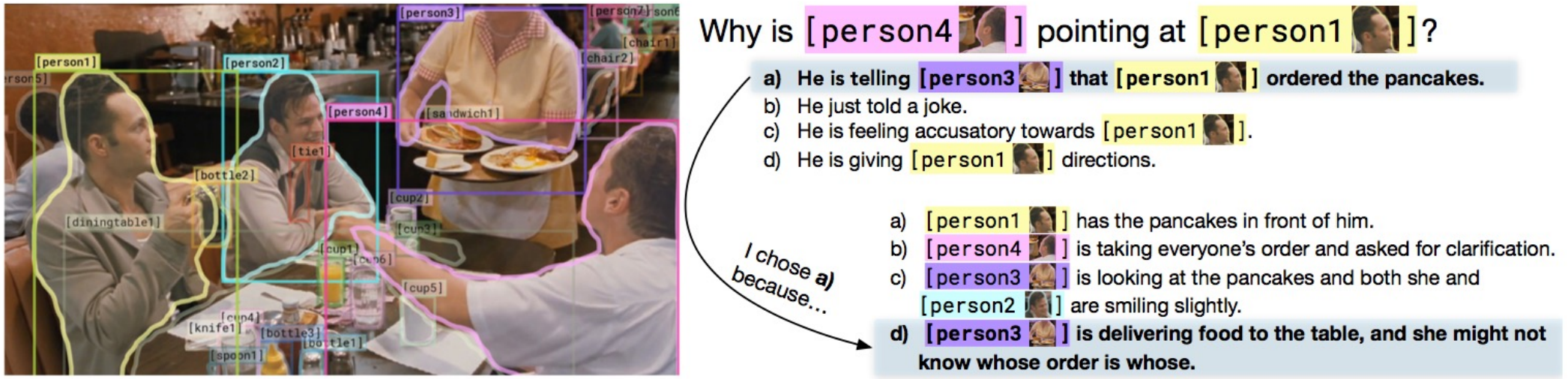
数据集	图片个数	描述个数
MSCOCO	300,000 +	5 per image
Flickr30K	30,000	158,000 in total
Flickr8K	8,000	5 per image
Visual Genome	108,000 +	1,445,322 in total
Instagram	~10,000	5 per image
FlikrStyle10K	10,000	Romantic, humorous, factual

视觉语言问答 (Visual Question Ansering)



数据集合	图片个数	问题个数	数据集特点
VQA2.0(2015)	204,721(coco)	1,105,904	10 annotated answers : yes/no, number, other
CLEVR(2016)	100,000	864,968	Synthetic; Reason about relationships between objects of different shapes, colors and sizes
Visual Genome(2016)	108,077(coco,flickr)	1,445,322	Region based qa-pair and caption, scene graph, object detection with annotated attribute
GQA(2019)	113,018(coco,flickr, visual genome)	22,669,678	Unbalanced data; scene graph based; full answer; word-object mapping

视觉常识推理（Visual Commonsense Reasoning）



- 任务：给定一张图片、一些目标物体、一个问题、四个答案，（1）让模型选择哪一个描述与图片是一致的，（2）让模型选择输出该答案的解释。

- 数据集 VCR：从110k电影片段中，抽取的290K多选QA.

From recognition to cognition: Visual commonsense reasoning, ICCV. 2019.

带时序的视觉常识推理（Visual COMET）

- 给定一张图片和当前的某一个事件描述以及地点，生成该事件片前的事件，当前事件的原因，后续时间片的事件。

	Train	Dev	Test	Total
# Images/Places	47,595	5,973	5,968	**59,356**
# Events at Present	111,796	13,768	13,813	**139,377**
# Inferences on Events Before	467,025	58,773	58,413	584,211
# Inferences on Events After	469,430	58,665	58,323	586,418
# Inferences on Intents at Present	237,608	28,904	28,568	295,080
# Total Inferences	1,174,063	146,332	145,309	**1,465,704**

Table 1: **Statistics** of our Visual Commonsense Graph repository: there are in total 139,377 distinct Visual Commonsense Graphs over 59,356 images involving 1,465,704 commonsense inferences.

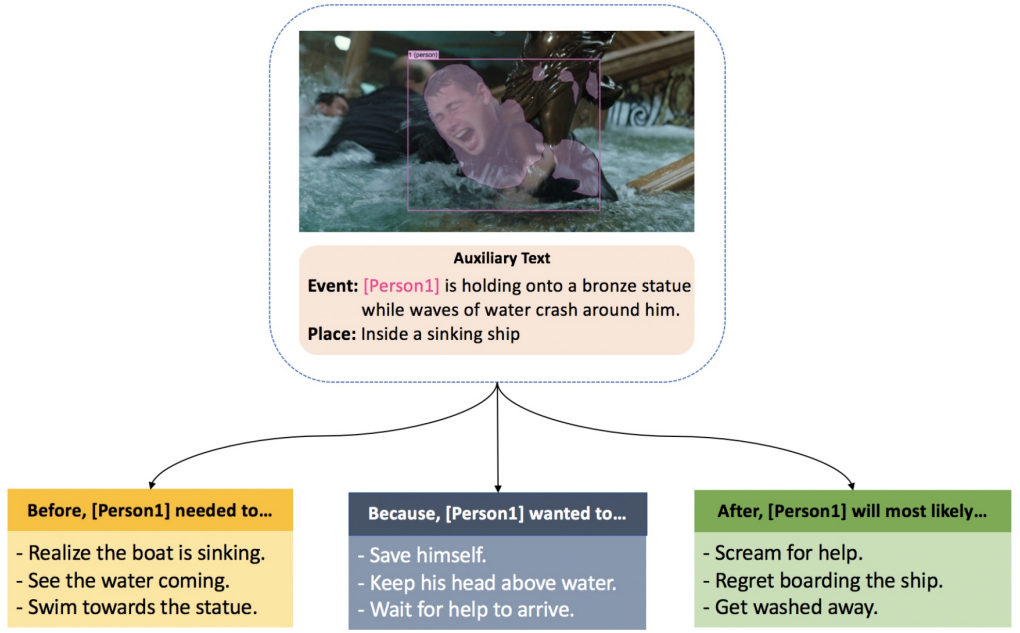
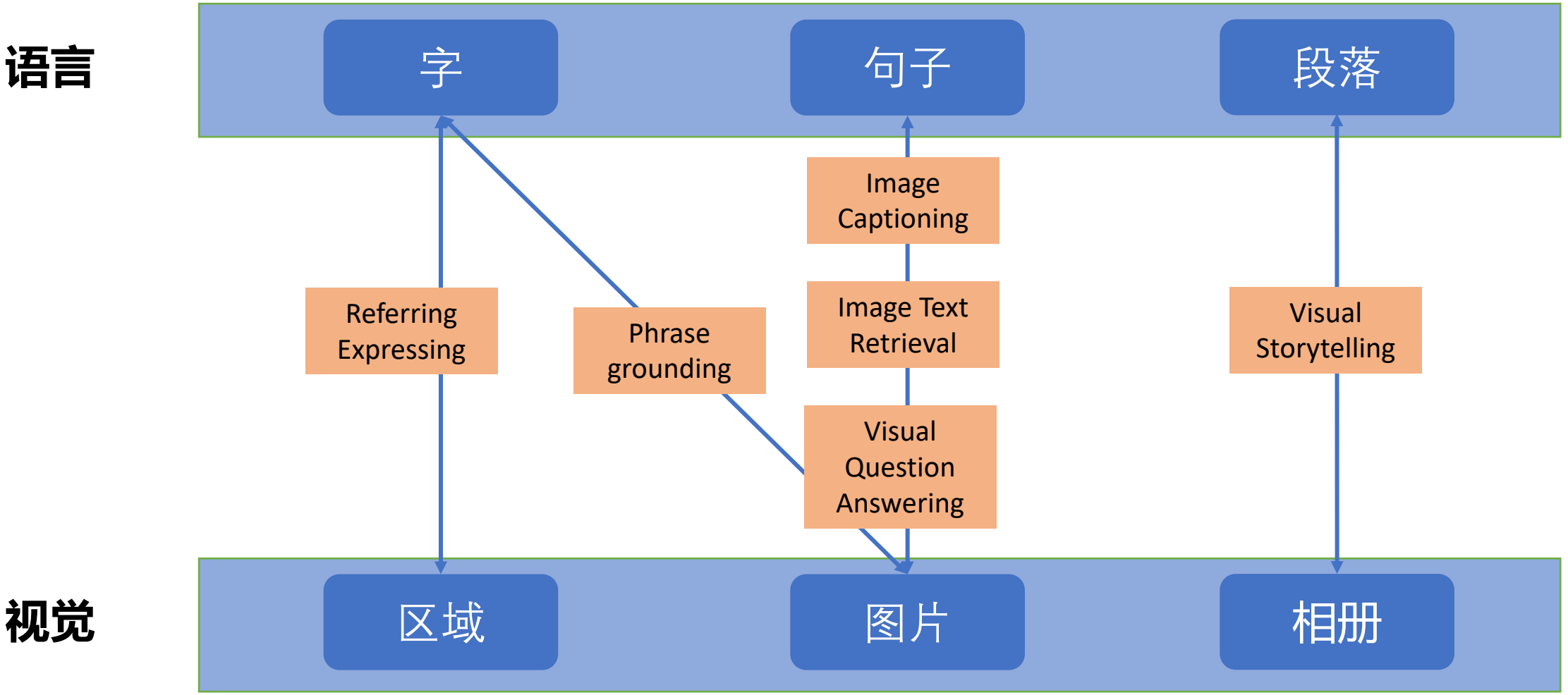


Fig. 2: **Task Overview:** Our proposed task is to generate commonsense inferences of **events before**, **events after** and **intents at present**, given an image, a description of an **event at present** in the image and a plausible scene / location of the image.

VisualCOMET: Reasoning about the Dynamic Context of a Still Image[J]. 2020.

跨模态任务探索不同粒度的语义对齐



跨视觉语言模态的预训练

- 基本设定: 以图片-文本对作为输入，联合学习语言和图片的语义表示。
- 输入表示: 单字， 图片区域，整体占位符（CLS）
- 跨模态交互学习:
  - 双塔模型（ LXMBert, ViLBERT, CLIP） ：浅层语义交互
  - 单塔模型（VLBert, Unicoder-VL） ：深层语义交互

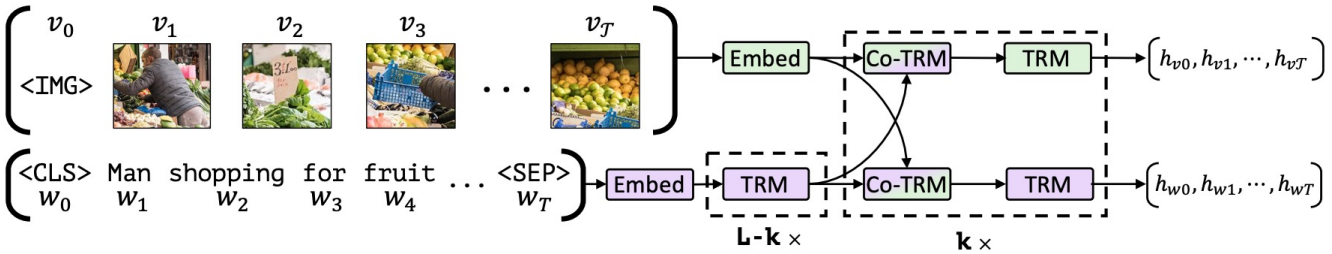


Figure 1: Our ViLBERT model consists of two parallel streams for visual (green) and linguistic (purple) processing that interact through novel co-attentional transformer layers. This structure allows for variable depths for each modality and enables sparse interaction through co-attention. Dashed boxes with multiplier subscripts denote repeated blocks of layers.

ViLBERT, NIPS 2019

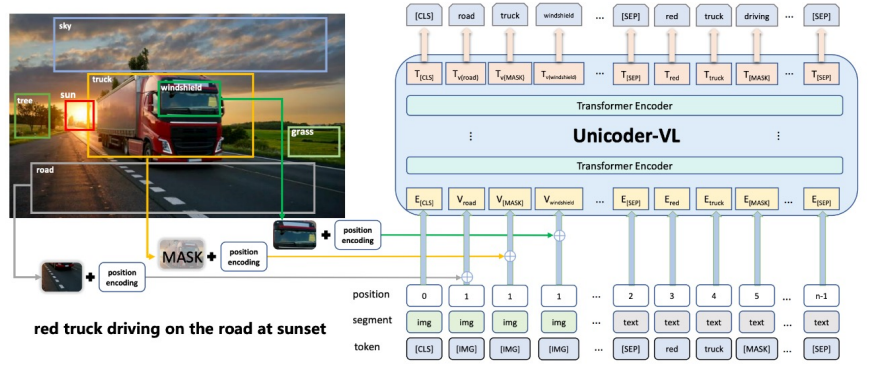


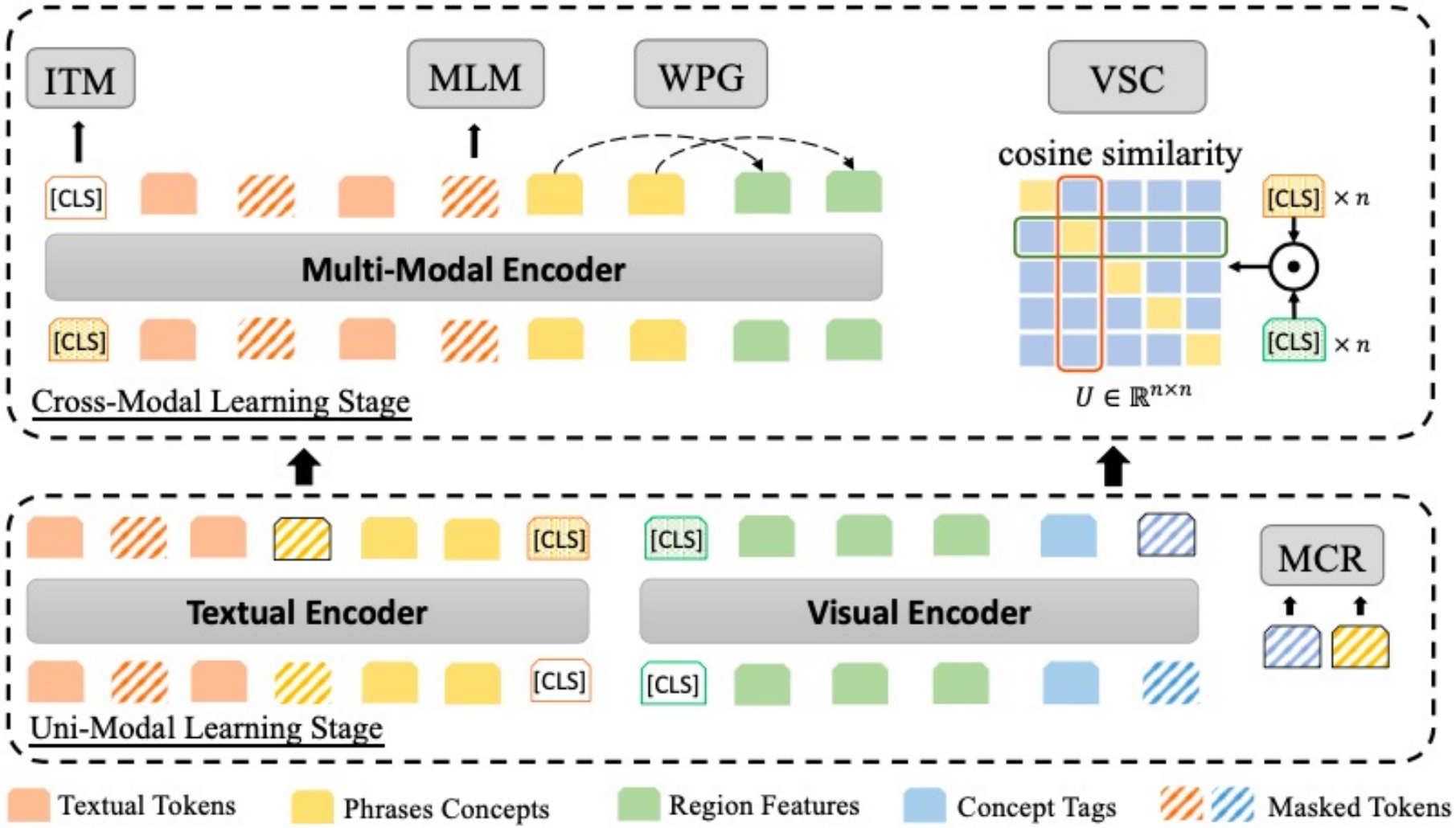
Figure 1: Illustration of Unicoder-VL in the context of an object and text masked token prediction, or *cloze*, task. Unicoder-VL contains multiple Transformer encoders which are used to learn visual and linguistic representation jointly.

VLBert, ICLR 2020

跨视觉语言模态的预训练任务

- **预训练任务**
  - 语言遮罩训练（Masked Language Modeling， MLM）
  - 图片区域遮罩（Mask Region Modeling， MOC）
  - 图 文匹配（Visual-Linguistic Matching， VLM）
- **预训练 + 下游微调**
  - 使用大规模数据集进行训练（COCO, Visual Genome, Conceptual Captions, and SBU Captions）

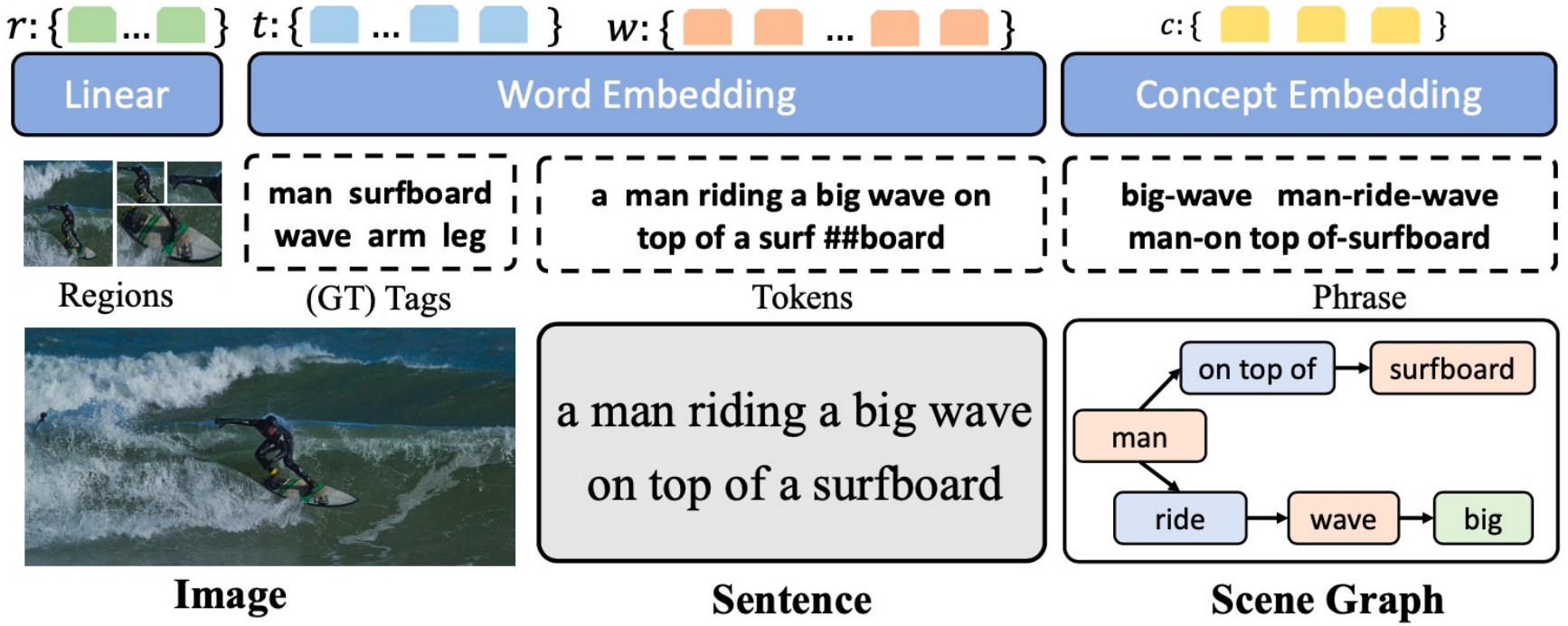
MVPTR： 多层次语义关联的模型预训练



MVPTR: Multi-Level Semantic Alignment for Vision-Language Pre-Training via Multi-Stage Learning, ACM MM 2022

多层次语义输入

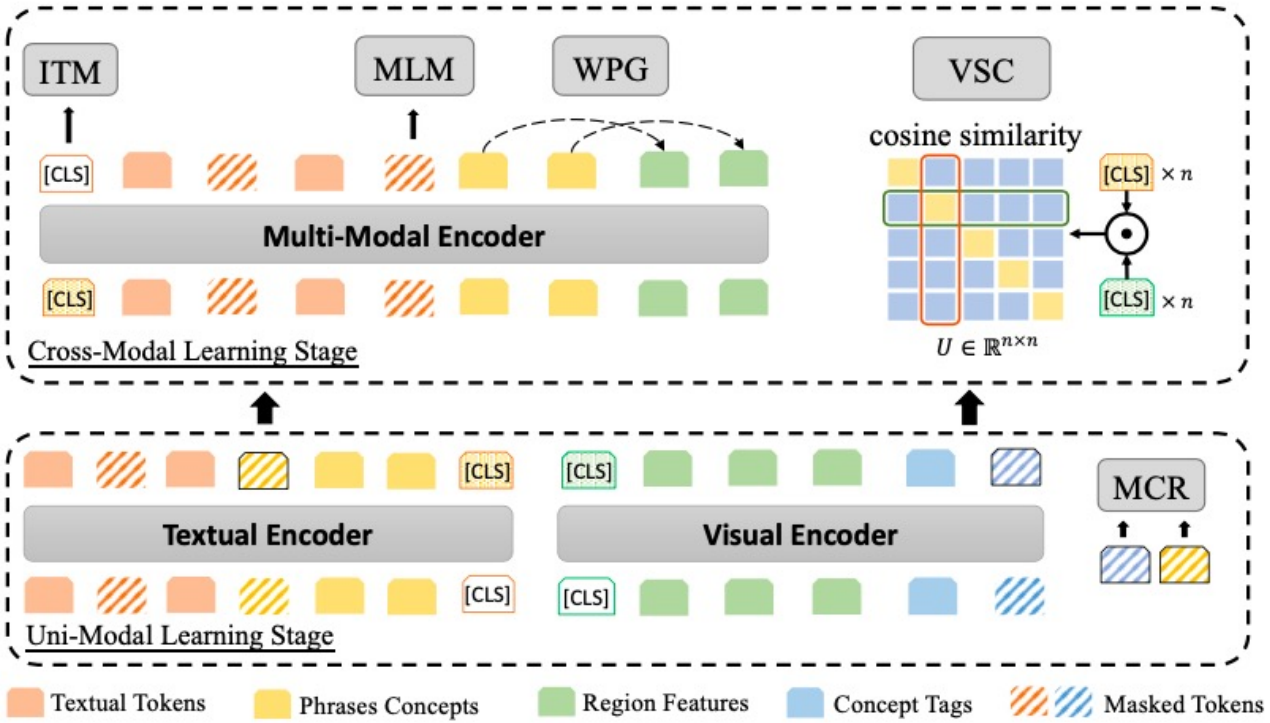
- 文本序列：单字 + 短语概念 (场景图中的属性和关系)
- 视觉序列：区域特征 + 实体标签



MVPTR: Multi-Level Semantic Alignment for Vision-Language Pre-Training via Multi-Stage Learning, ACM MM 2022

MVPTR：两阶段预训练方法

- 双阶段：单模态 + 混合模态
- 双模式：单塔 + 双塔

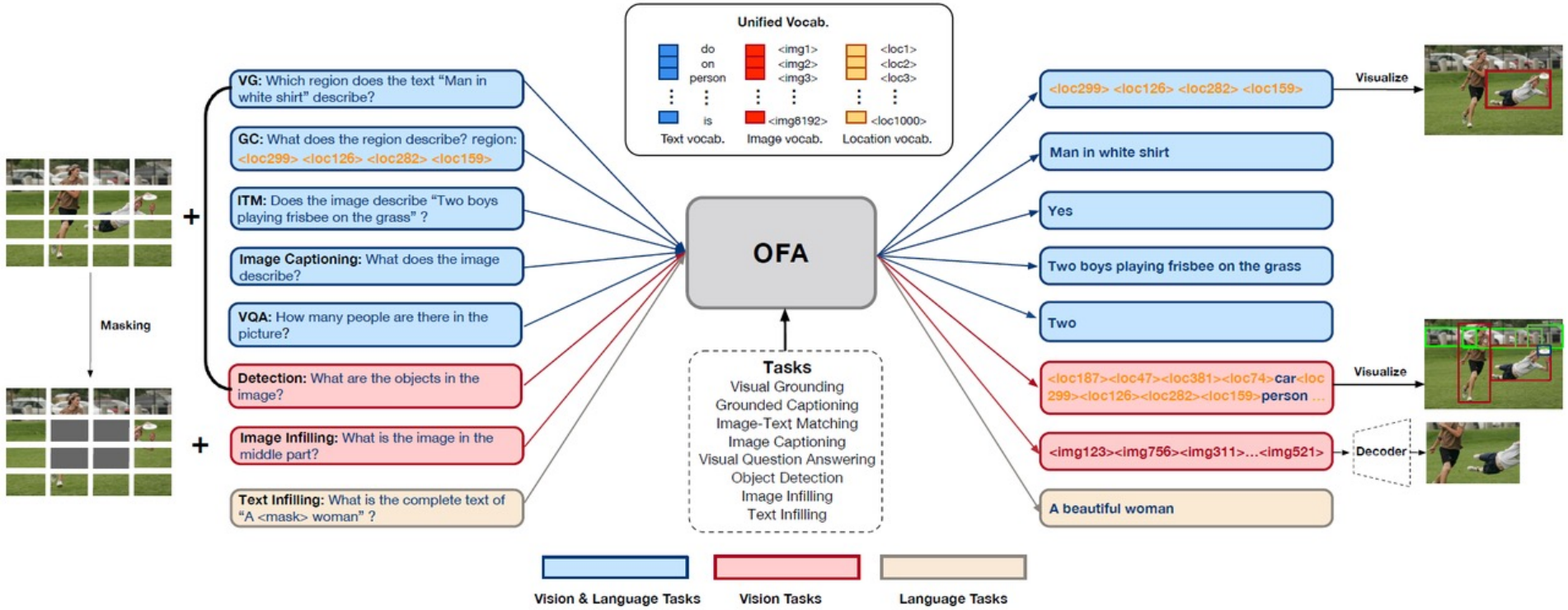


- 单模态阶段：
  - 高层次概念遮罩 (MCR)
- 跨模态阶段：
  - 全局的粗匹配：visual-semantic contrastive learning (VSC)
  - 细粒度匹配：weakly-supervised phrase grounding (WPG)
  - 细粒度图文匹配 (ITM) 和跨模态推理 (MLM)

MVPTR: Multi-Level Semantic Alignment for Vision-Language Pre-Training via Multi-Stage Learning, ACM MM 2022

训练-推理统一的多模态预训练框架

- VL-BART 和 OFA 将所有的任务改造成序列到序列的格式
- 在预训练阶段收集多个任务的样本（多模态、视觉模态、文本）
- 扩充词汇表（视觉、文本、位置）



OFA: Unifying Architectures, Tasks, and Modalities Through a Simple Sequence-to-Sequence Learning Framework，2022

小结

- 在训练阶段利用不同粒度的语义对齐完成多模态语义表示学习
- 在推理阶段使用不同的决策参数进行下游任务推理（初代预训练）
- 使用序列到序列的模式规整多种推理任务 （OFA）
- **假设：视觉模态和文本模态是平等的**

预训练多模态模型：规模并不够大

- 参数规模 (以 2022 数据为准)

模型	BLIP	OFA	CoCa	BeiT-3	GIT	PaLi
参数	0.3B	0.9B	2.1B	1.9B	5.1B	17B

- 训练的数据规模
  - 14M 高质量图文匹配对 (COCO, VG, CC, SBU)
  - 100M~5B 弱匹配对 (LAION, in-house data)