中国科学院大学
University of Chinese Academy of Sciences

自然语言处理

第17讲 多模态大模型

王 石 资康莉 刘 瑜

2026年春季课程

https://ictkc.github.io/teaching/

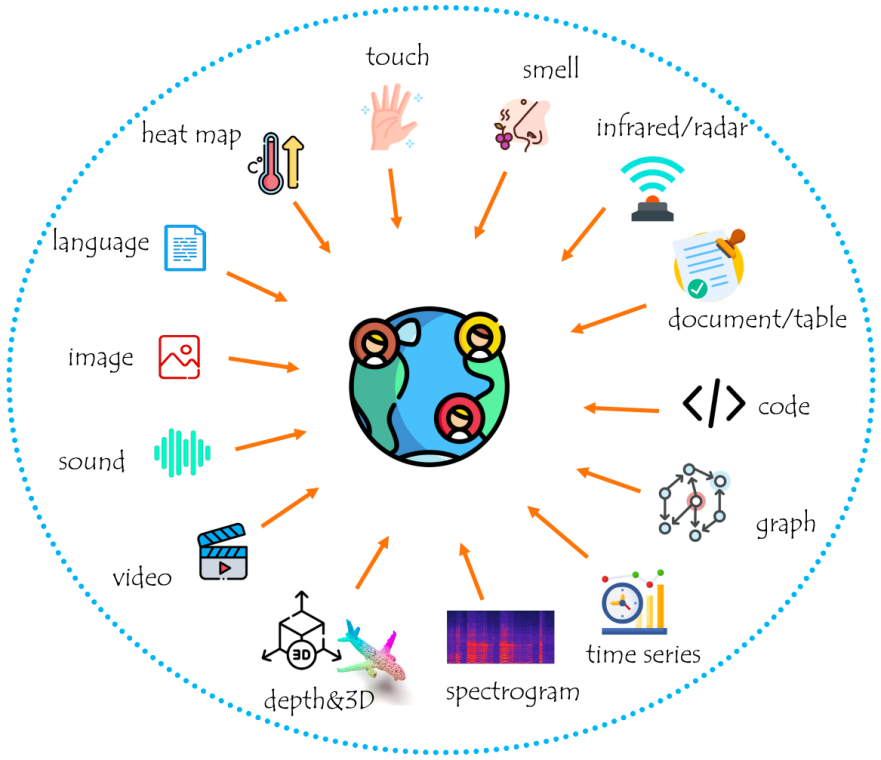
第十七讲
多模态大模型

中国科学院大学
University of Chinese Academy of Sciences

目 录

多模态：大模型发展的必经之路

❑ **多模态预训练模型**通过融合图文音等海量数据，实现类人多模态感知与认知，推动语音、语言和视觉等领域协同发展

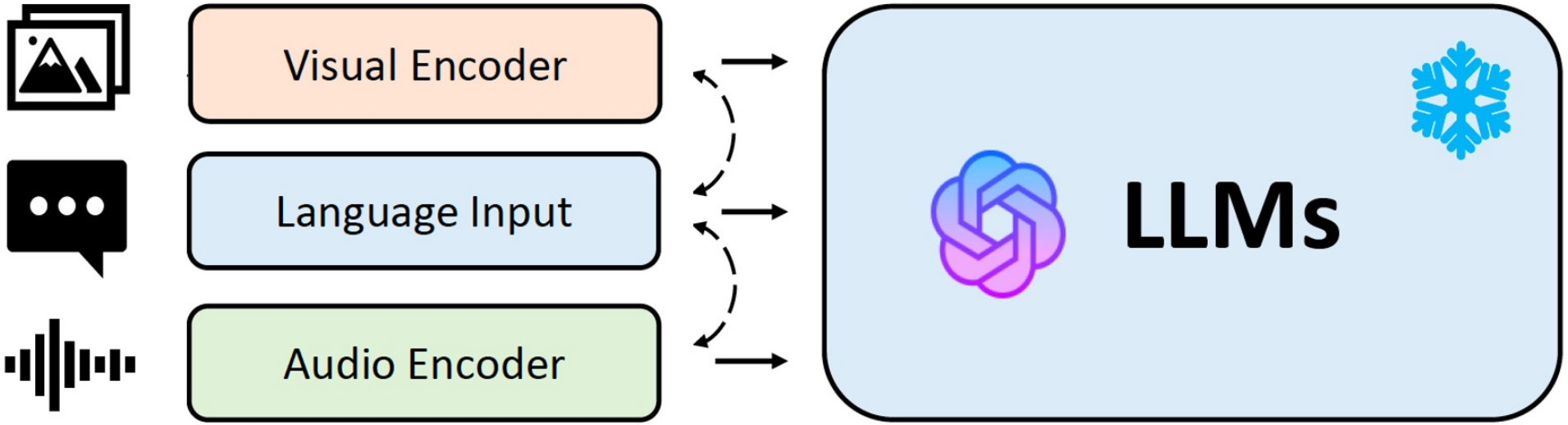


多模态数据无处不在：互联网90%以上是图像与音视频数据，文本不到10%

多模态协同更契合人类感知与表达方式：让机器"看懂、听懂、能说、会读”

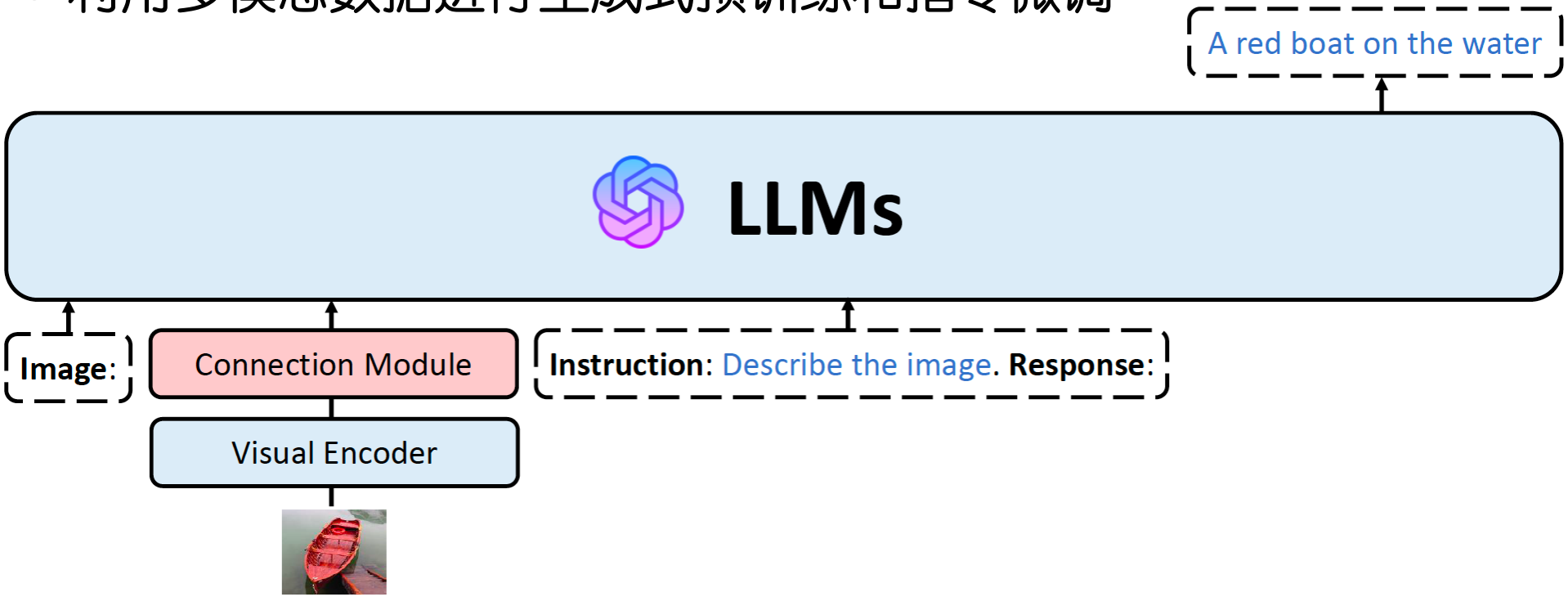
如何将多模态融入大模型?

- 大语言模型作为认知中枢，统一处理多模态信息
- 将不同模态映射到统一语义空间，实现跨模态理解与生成



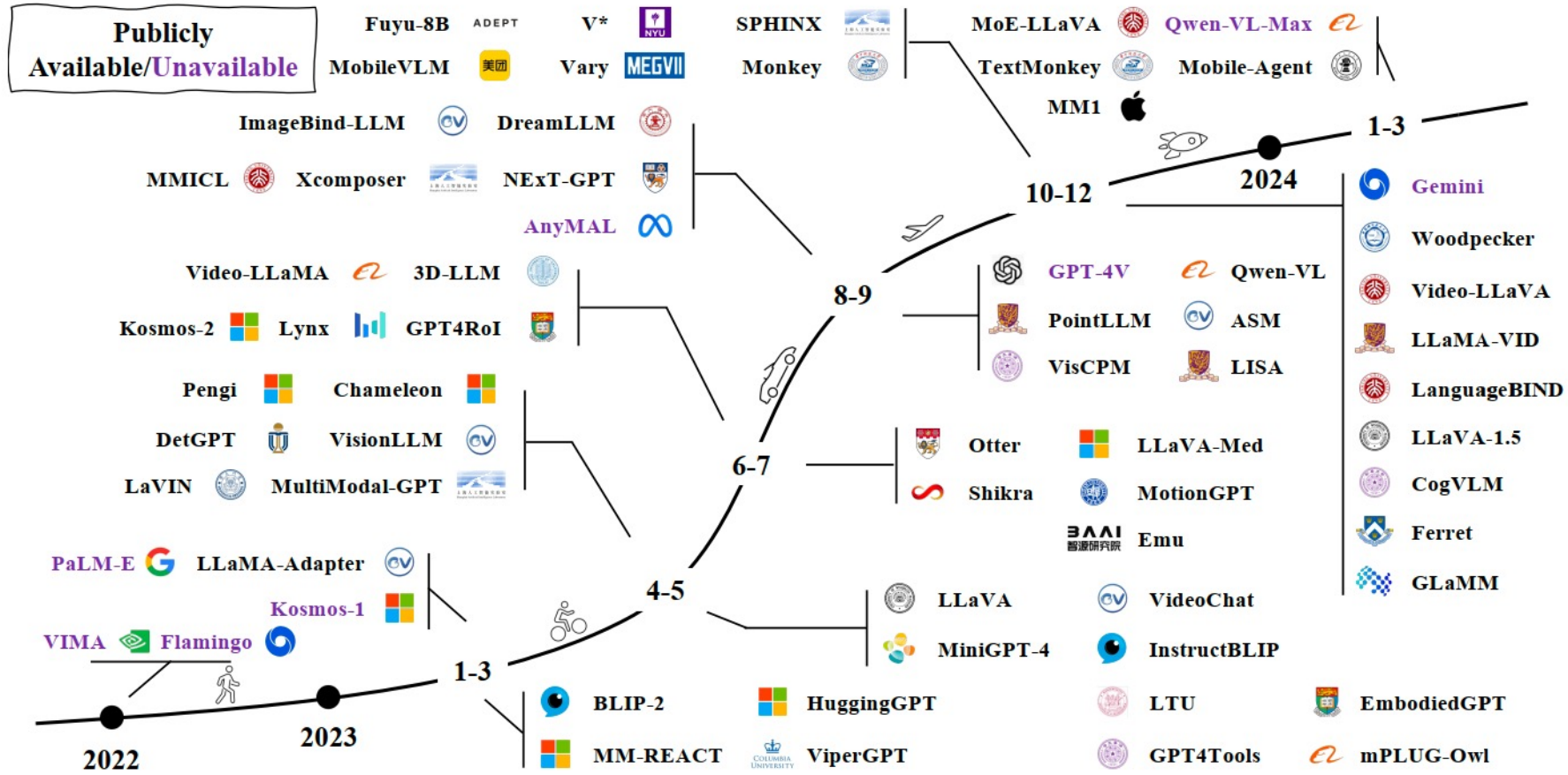
多模态大模型MLLM

- 多模态大模型 （Multimodal Large Language Model, MLLM）
  - 使用大模型LLM作为骨干和模态编码器
  - 利用多模态数据进行生成式预训练和指令微调



MLLM发展历程

- 随着大模型能力的爆发，多模态研究进行快速发展阶段

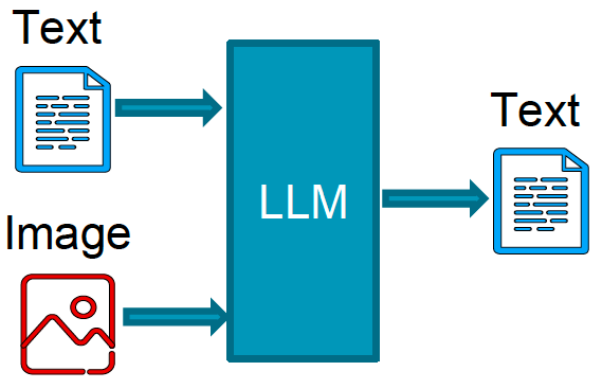


A Survey on Multimodal Large Language Models. 2023

MLLM-多模态理解

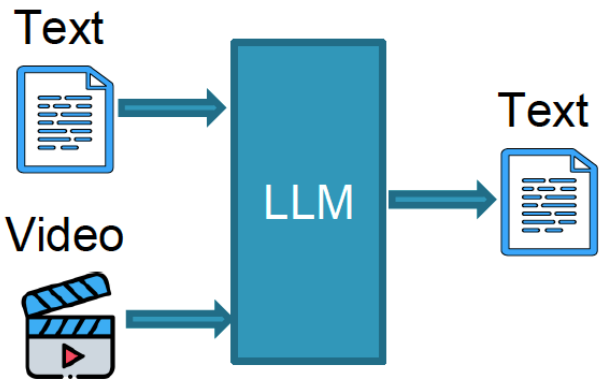
- 利用外部多模态编码器对输入信号进行特征提取和表示映射，并结合LLM实现对多模态内容进行理解与推理

图像理解



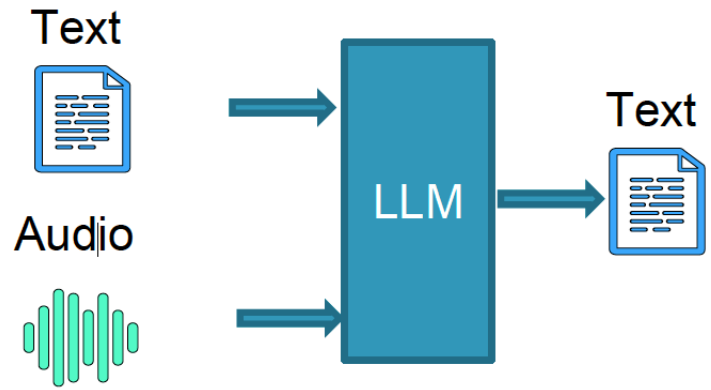
代表模型：BLIP2、LLaVA、Flamingo、Mini-GPT4

视频理解



代表模型：VideoChat、Video-LLaVA、Momentor

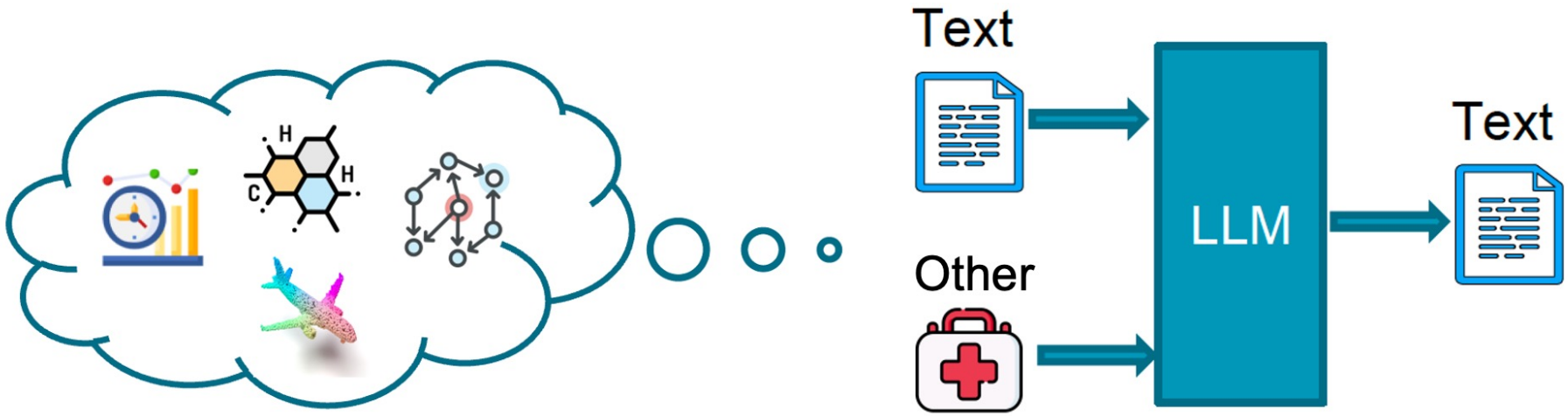
音频理解



代表模型：AudioGPT、VIOLA、SpeechGPT、SALMONN

MLLM-多模态理解

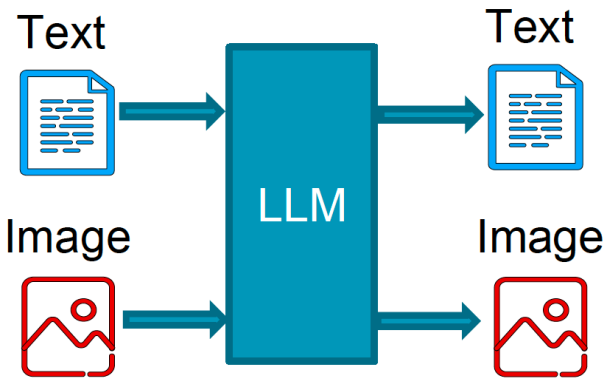
- 此外，多模态还涵盖时序数据、结构化表格、知识图谱、3D空间数据以及生物医药等



MLLM-多模态生成

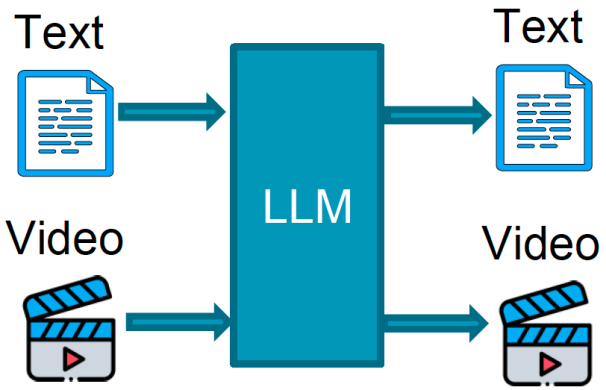
❑ 多模态大模型不仅需要感知能力，还需具备多模态的生成能力

图像生成



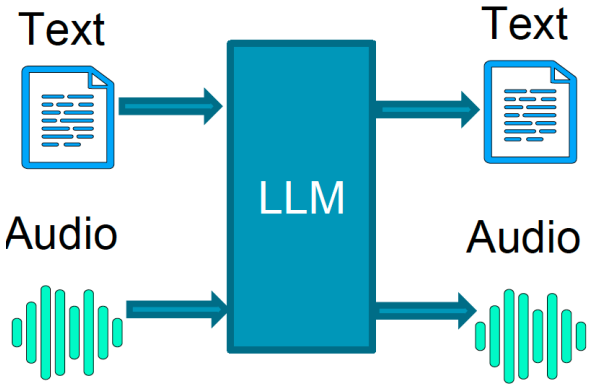
代表模型：GILL、EMU
MiniGPT5、DreamLLM

视频生成



代表模型：GPT4Video、
VideoPoet、Video-LaVIT

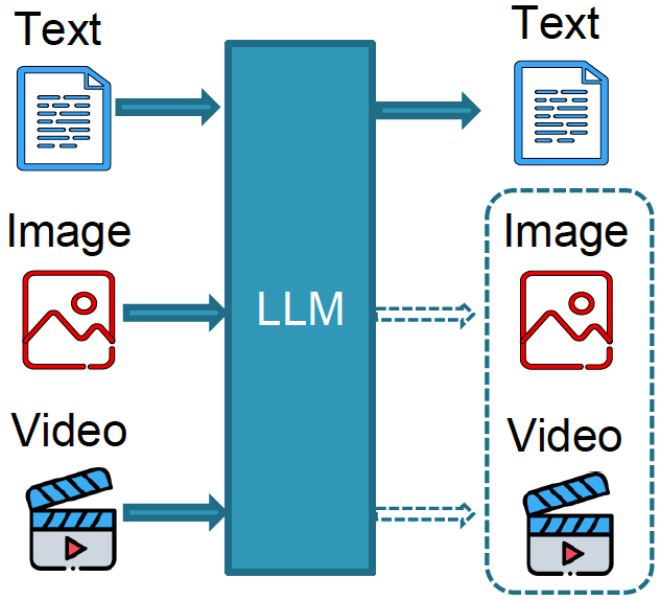
音频生成

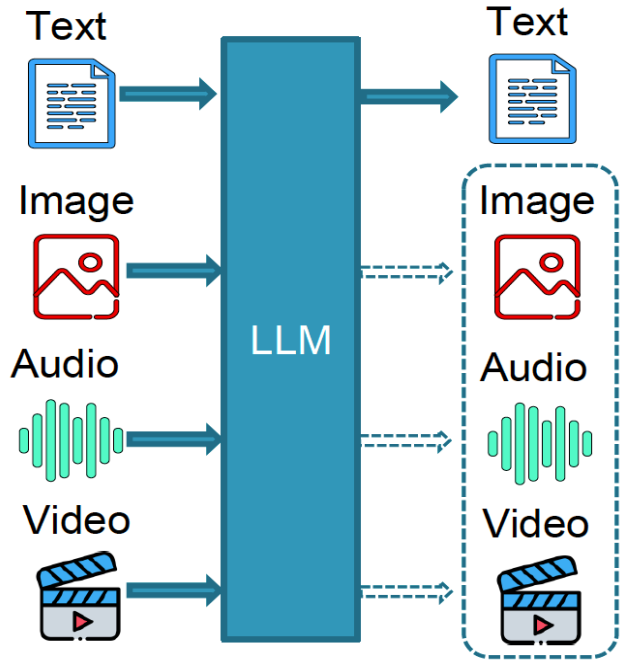


代表模型：AudioGPT、
SpeechGPT、AudioPaLM

MLLM-多模态生成

- 多模态大模型接收文本、图像和视频等多模态输入，通过统一语义理解实现跨模态推理，并生成多种形式的输出



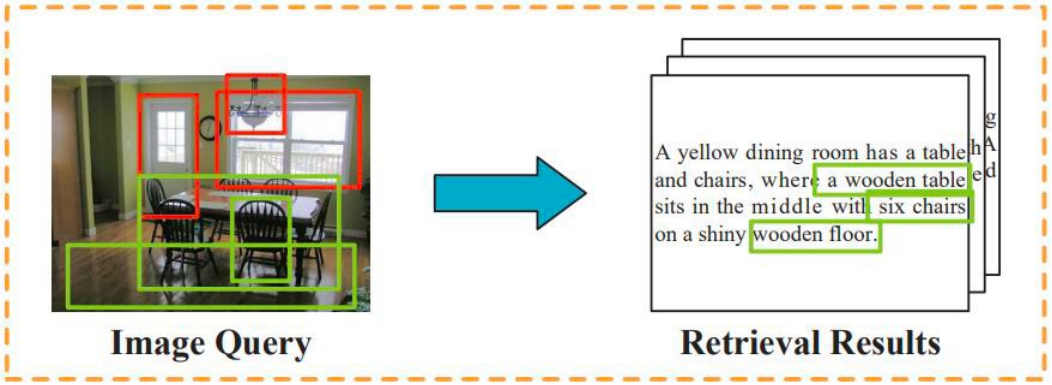


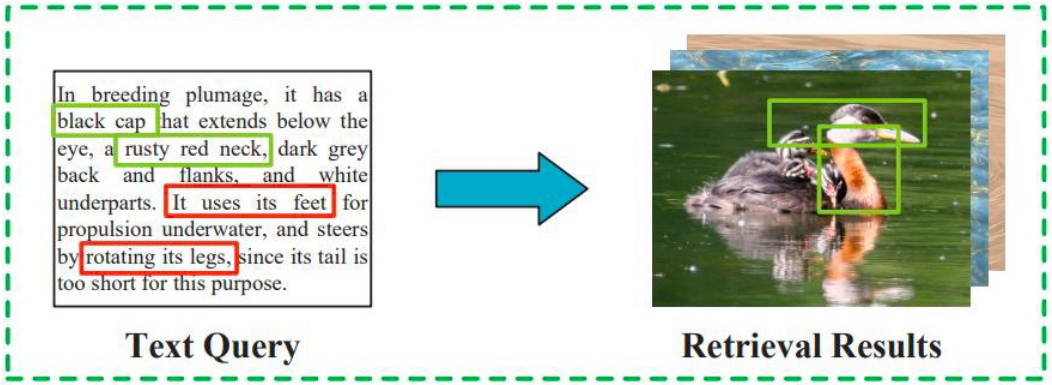
MLLM任务场景

- 面向多模态应用场景，多模态大模型不仅需理解多源信息，还需具备相应模态的生成能力，以支持跨模态理解与生成任务

1. 图像文本的语义匹配

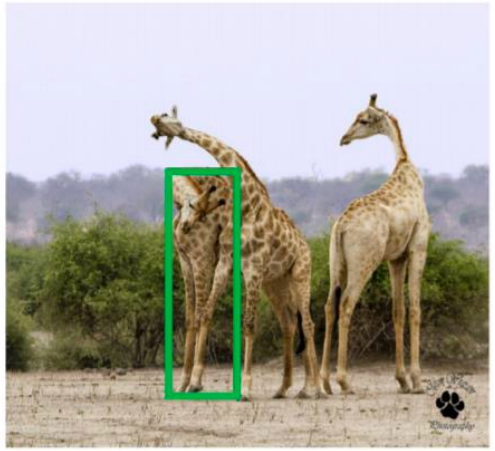
- 给定一张图片，从句子集合中检索语义相关的句子
- 给定一个句子，从图片集合中检索语义相关的图片
- 评测指标: R@1(Recall@1), R@5, R@10





2. 图像指代理解

- 给定一段自然语言描述，定位图像中对应的目标物体
- 重叠比例Intersection over Union(IoU)：真实和预测的物体框
- 当 IoU 大于 0.5 时，认为定位正确；否则定位错误



RefCOCO

1.左边的长颈鹿
2.最左侧那头长颈鹿

RefCOCO+

1.低头的长颈鹿
2.脑袋垂下的长颈鹿

RefCOCOg

1.一头成年长颈鹿在用角挠自己的后背
2.依偎着另一头长颈鹿的长颈鹿

3. 基于图像的文本生成

- 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

A person riding a motorcycle on a dirt road.

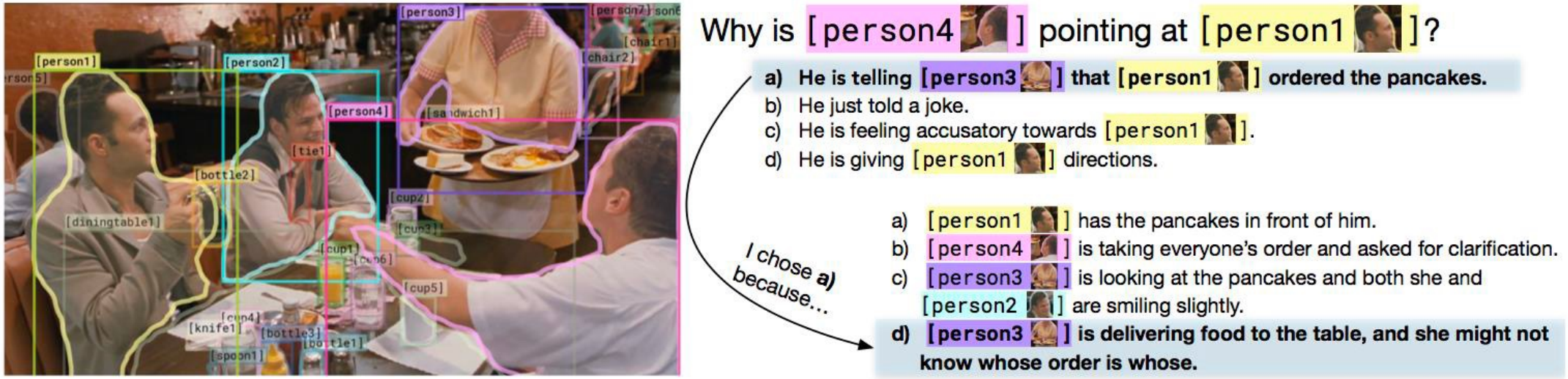
Two dogs play in the grass.

4.视觉语言问答

- 输入一张图像和一段自然语言问题，模型输出贴合图像内容的自然语言答案
- 评测指标：Accuracy , BLEU , CIDEr , ROUGE

5.视觉常识推理

- 给定图片、若干目标物体、一个问题及四个候选答案
  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

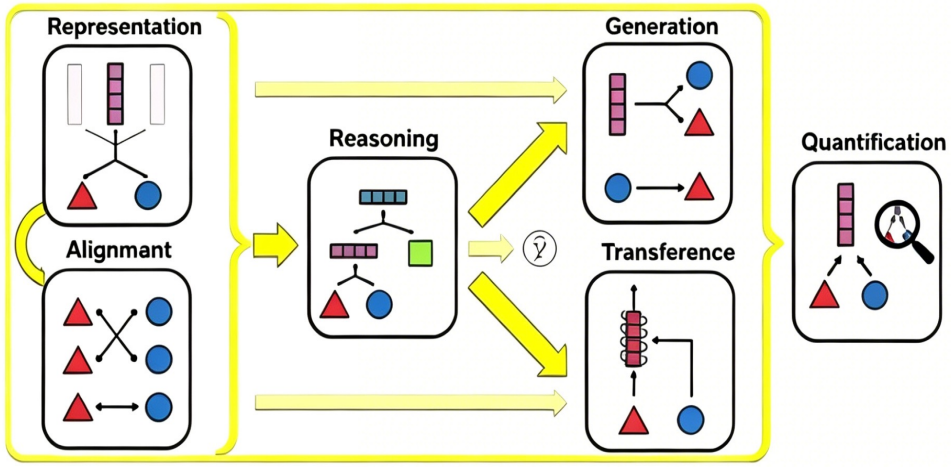


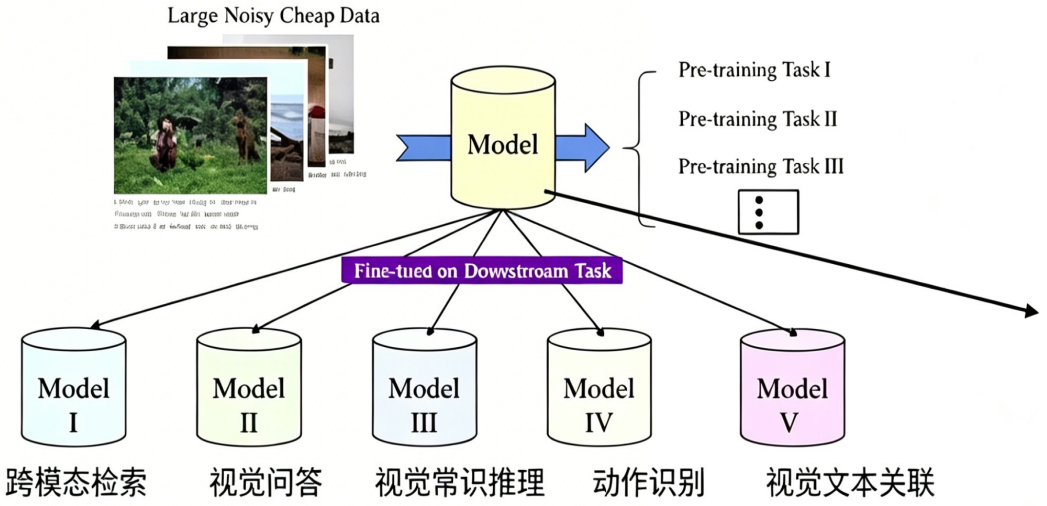
中国科学院大学
University of Chinese Academy of Sciences

目录

多模态预训练

- 通过自监督学习与通用知识迁移，统一框架适配多领域任务
- 聚焦多模态表征、跨模态对齐，提升多模态理解与生成能力





多模态预训练数据集

- 无标注成本的网络数据
  - 图像文本数据：图像及其相关文本(标签、描述、评论等)
  - 视频文本数据：视频及其相关文本(标签、描述、字幕、语音等)
  - 音频文本数据：音频及其相关文本(文稿、标签、歌词、旁白等)

公众号、自媒体

短视频平台

国外APP数据

国内APP数据

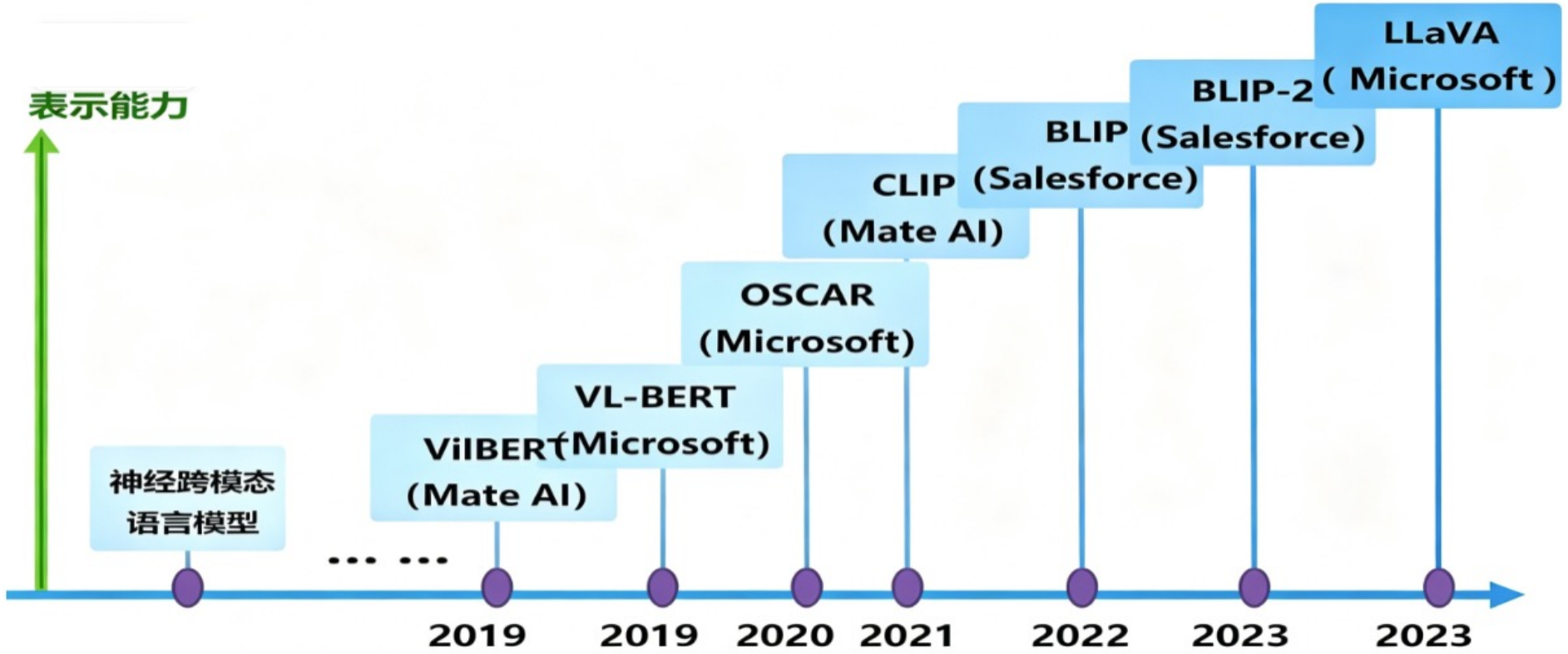
多模态预训练数据集

❑ 万级别强关联人工标注 ➡ 百万/亿级别弱关联无标注

SBU NeurIPS 2011	Microsoft ECCV 2014	Stanfooft EICV 2017	Google ACL 2018	Google ICML 2021	Kalcan Brain 2022
1M 图片标题对	330K 图片/1.5M标题	108K 图片标题对	12M 图片标题对	1.8B 图片标题对	747M 图片标题对
			 Conceptual Captions: a worker helps to clear the debris.	 "motorcycle front wheel"	 The Gate by Pete2453
SBU Caps	COCO Caps	VGD Caps	Conceptual Caps	ALIGN	COYO-700M
HowTo100M		WebVid-2.5M	YT-Temporal-180M		HD-VILA-100M
		 "Billiards, concentrated young woman playing in club"			
136M 视频标题对/134500小时		2.5M 视频标题对/13000小时	1.8M 视频标题对		100M 视频标题对
INRIA ICCV 2019		Oxford ICCV 2021	UW NeurIPS 2021		NISRA CVPR 2022

多模态预训练发展历程

❑ 传统多模态预训练模型 ➡ 基于LLM的多模态预训练语言模型

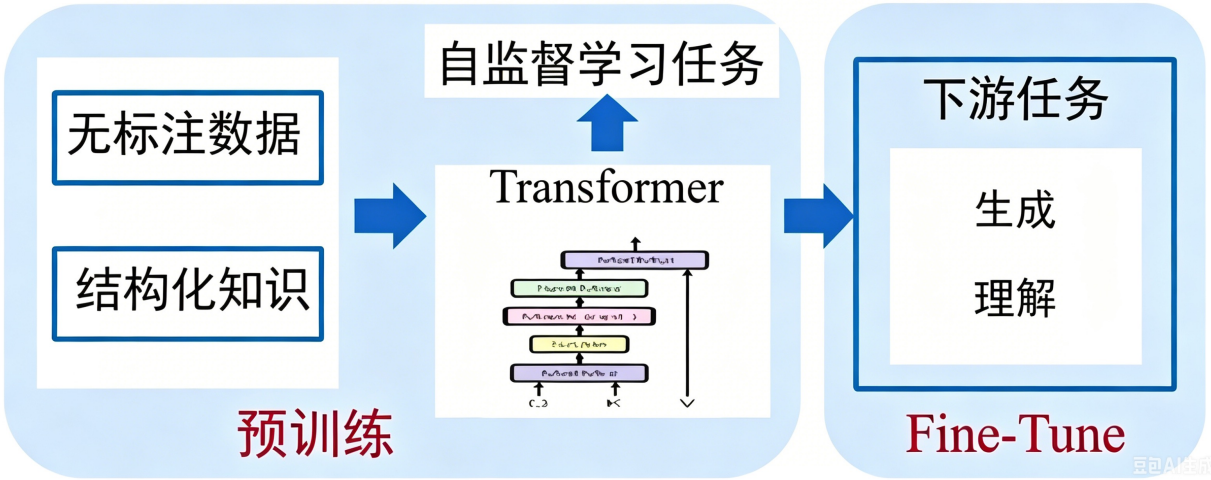


中国科学院大学
University of Chinese Academy of Sciences

目 录

传统多模态预训练

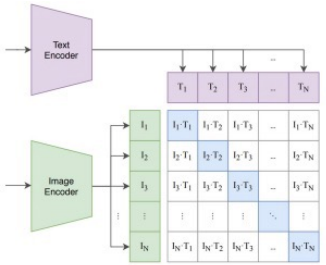
- 预训练：以 Transformer 为基底，构建自监督预训练任务，从海量无标注数据习得通用底层表征
- 微调：适配各类多模态理解、生成下游任务



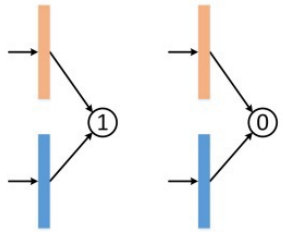
- 训练数据：大规模无标注
- 模型框架：Transformer
- 学习机制：自监督学习
- 下游任务：理解与生成

传统多模态预训练任务

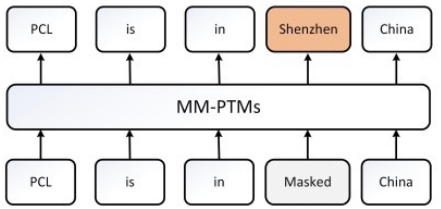
- 自监督学习任务
  - **对比匹配**：对比损失、模态匹配
  - **遮罩重建**：遮罩语言建模、遮罩目标分类、遮罩目标回归、自编码
  - **理解生成**：视觉问答（分类）、图文生成（生成）



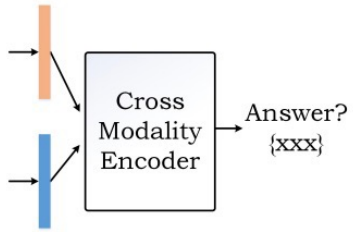
(a). Contrastive Loss



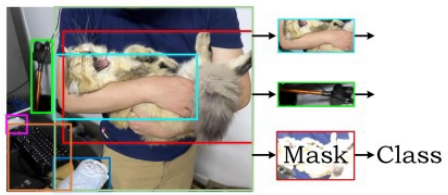
(b). Modality Matching Loss (MML)



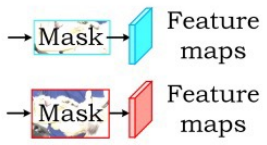
(c). Masked Language Modeling (MLM)



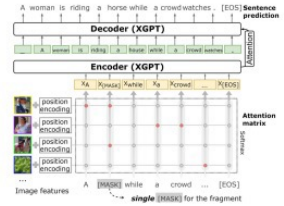
(d). Image Question Answering (QA)



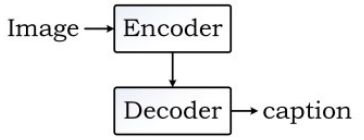
(e). Masked Object Classification (MOC)



(f). Masked Object Regression (MOR)



(g). Image-conditioned Denoising Autoencoding (IDA)

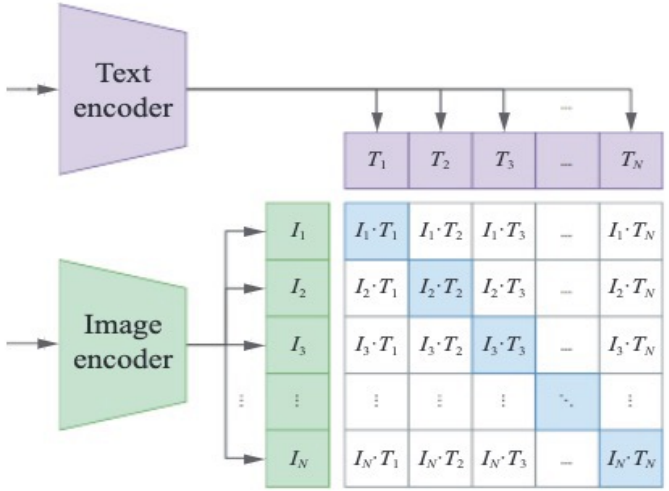


(h). Image-Text Generation (ITG)

1. 对比匹配任务

❑ 对比损失 (Contrastive Loss)

- **训练目标**：利用内积相似度衡量图文语义相关性；提高配对图文相似度，降低非配对图文相似度，实现跨模态对齐
- **代表模型**：CLIP、ALIGN、VinVL



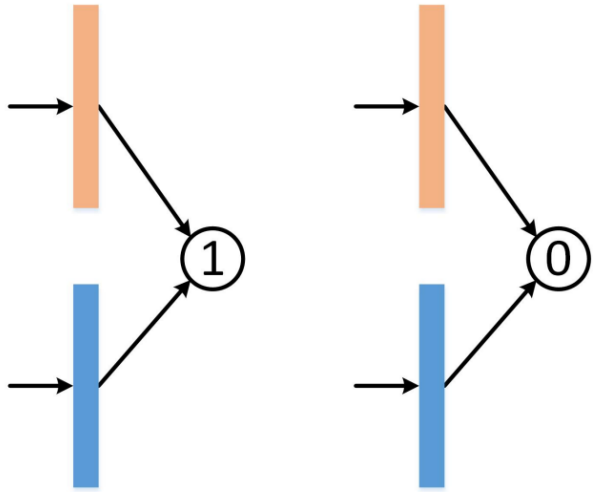
$$\mathcal{L}_{i2t} = -\frac{1}{N}\sum_{i}^{N} log\frac{exp(x_i^T y_i/\sigma)}{\sum_{j=1}^{N} exp(x_i^T y_j/\sigma)}$$

$$\mathcal{L}_{t2i} = -\frac{1}{N}\sum_{i}^{N} log\frac{exp(y_i^T x_i/\sigma)}{\sum_{j=1}^{N} exp(y_i^T x_j/\sigma)}$$

$$\mathcal{L}_{CL} = \mathcal{L}_{i2t} + \mathcal{L}_{t2i}$$

1. 对比匹配任务

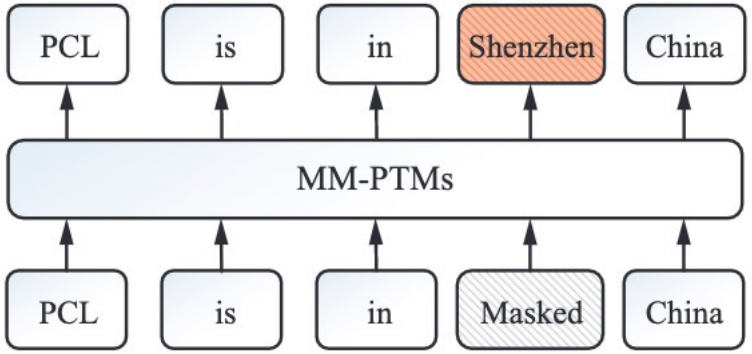
- **模态匹配损失** (Modality Matching Loss)
  - **训练目标**：采用分类损失判断图文是否匹配；与对比损失的差异：更灵活的负样本训练方式，强化困难负样本的识别能力
  - **代表模型**：Unicoder-VL、InterBERT



$$\mathcal{L}_{\mathrm{ITM}} = -\mathbb{E}_{(W,V)\in\mathcal{D}}\left[y\log s_\theta\left(\boldsymbol{h}_{w_{[\mathrm{CLS}]}},\boldsymbol{h}_{v_{[\mathrm{IMG}]}}\right)\right.$$
$$\left.+(1-y)\log\left(1-s_\theta\left(\boldsymbol{h}_{w_{[\mathrm{CLS}]}},\boldsymbol{h}_{v_{[\mathrm{IMG}]}}\right)\right)\right]$$

2. 遮罩重建任务

- **遮罩语言建模** (Masked Language Modeling)
  - **训练目标**：根据上下文预测被掩码词，提升语义理解能力。同时，基于上文让模型预测接下来的文本，提高模型的生成能力
  - **代表模型**：SIMVLM

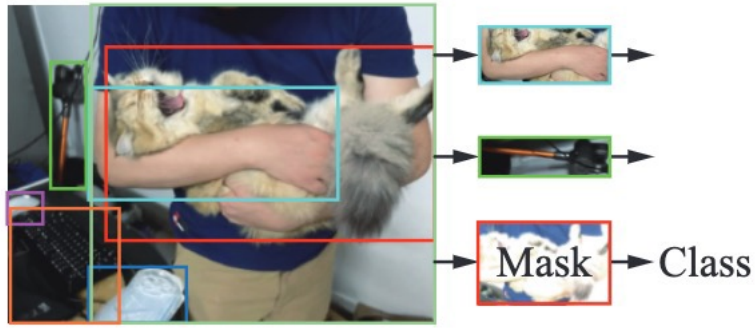


$$\mathcal{L}_{MLM}(\theta) = -\mathrm{E}_{(x,v)} \log P_\theta(x_m | x_{\neg m}, v)$$

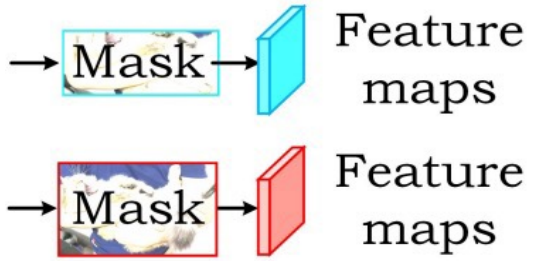
$$\mathcal{L}_{PrefixLM}(\theta) = -\mathrm{E}_{\boldsymbol{x}\sim D} \log P_\theta(\boldsymbol{x}_{\geq T_p} | \boldsymbol{x}_{<T_p})$$

2. 遮罩重建任务

- **遮罩目标分类/分类**（Masked Object Classification/Regression）
  - **训练目标**：通过预测被mask的图像区域的类别，或者把其转变为特征回归任务，提高模型对于图像的理解能力、图文之间的关联能力
  - **代表模型**：Unicoder-VL



遮罩目标分类（MOC）

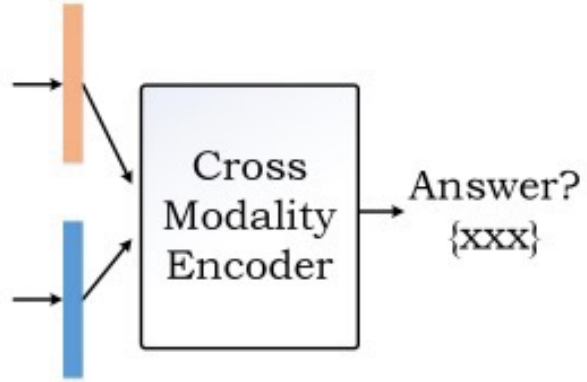


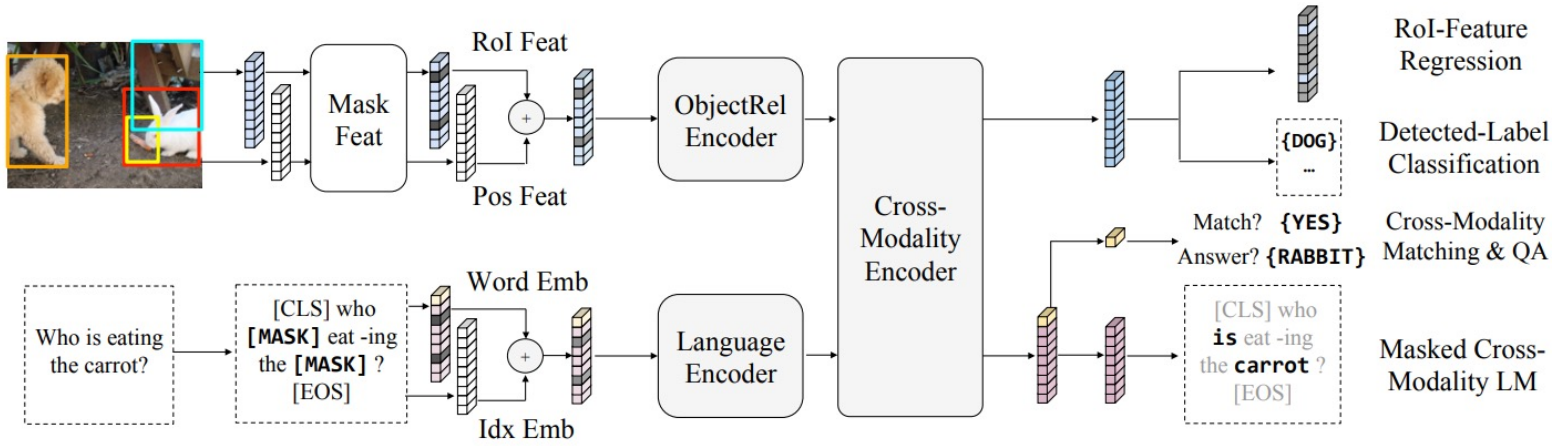
遮罩目标回归（MOR）

3. 理解生成任务

❑ **图像问答损失** (Image Question Answering)

- **训练目标**：利用图像-问题-答案（VQA）数据构建预训练任务；通过答案预测增强模型的跨模态理解与推理能力。
- **代表模型**：LXMERT

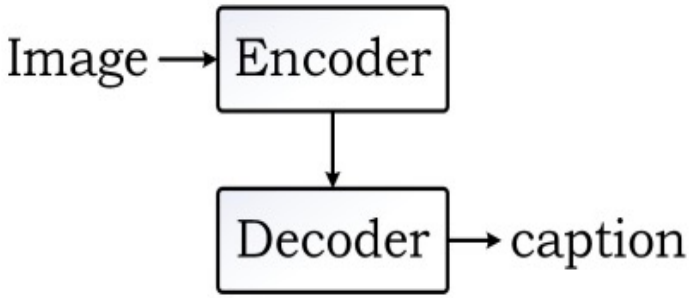




LXMERT模型

3. 理解生成任务

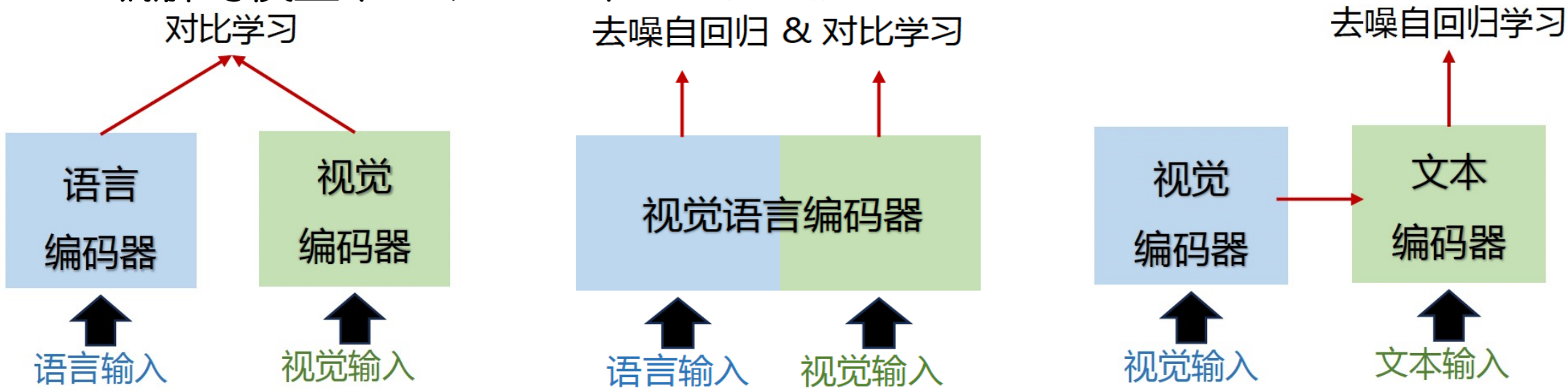
- **图像文本生成** (Image-Text Generation)
  - **训练目标**：根据输入图像生成对应的文本描述，提升模型的跨模态理解与生成能力
  - **代表模型**：E2E-VLP



$$\mathcal{L}_{ITG} = -\sum_{(x,y)\in(\mathcal{X},\mathcal{Y})} log \prod_{t=1}^{n} P(y_t|y_{<t}, x)$$

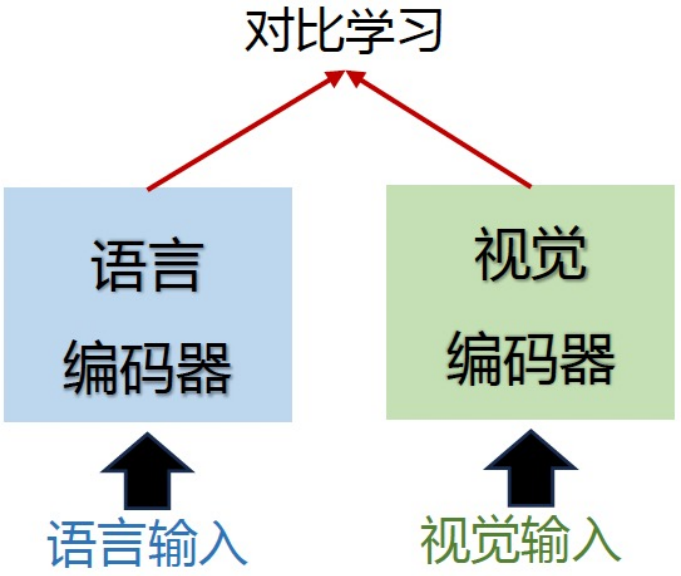
传统多模态预训练模型

- 基本设定：输入图片-文本对，联合学习语言和图片语义表示
- 跨模态交互学习模式:
  - 双流模型 (LXMBERT, ViLBERT,CLIP)：浅层语义交互
  - 单流模型 (VLBERT, Unicoder-VL)：深层语义交互
  - 编解码模型 (BLIP, BLIP-2)：生成式交互



1.双流模型

- 图像和文本单独编码，将视觉表示和语言表示映射到一个统一的共享语义空间



函数关系：

$$H_I = f_{visual(I;\theta_{visual})},$$

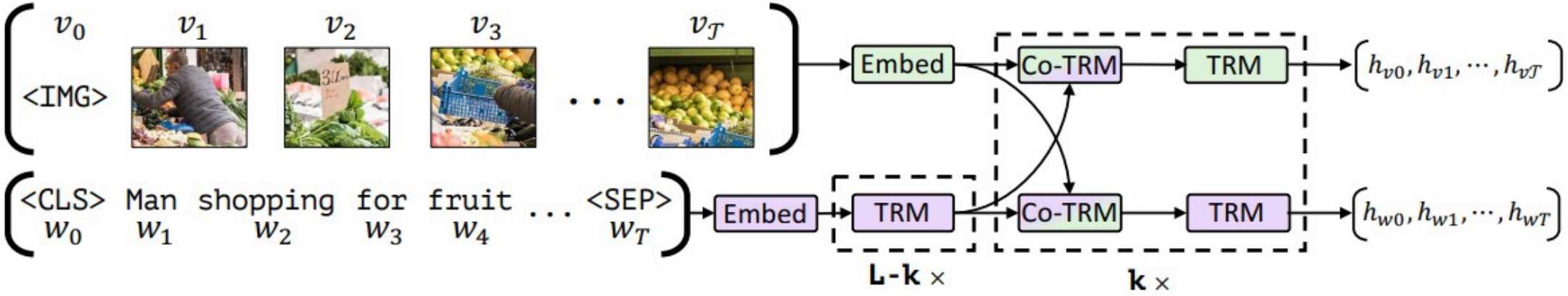
$$H_T = f_{text(T;\theta_{text})},$$

$$H_C = f_{fusion(H_I,H_T;\theta_{fusion})}$$

其中，f_{visual}、f_{text} 是模态特征提取函数，f_{fusion} 是特征融合函数

1.双流模型-ViLBERT

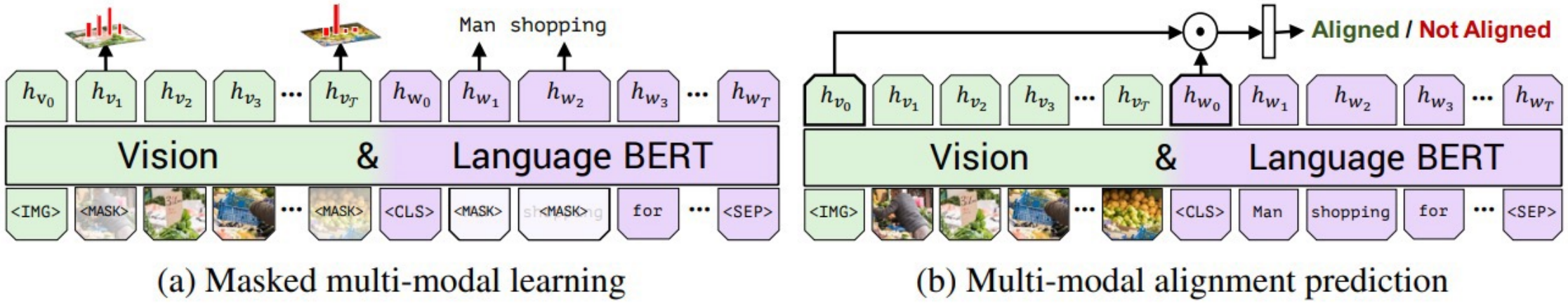
- 将图像与文本分别编码，通过跨模态自注意力实现模态交互
  - **输入编码**：文本中的单词；目标检测得到的目标区域
  - **模态交互**： Q、K、V 分属不同模态，图像可作 Q 检索文本 K/V，文本亦可作 Q 检索图像 K/V



Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. NIPS 2019 .

1.双流模型-ViLBERT

- 多模态预训练任务
  - **文本掩码预测**：随机掩码文本词元，并预测当前位置的词元
  - **图像遮掩区域预测**：随机遮蔽图像区块，预测对应区域类别
  - **图文匹配任务**：取图像 [IMG] 与文本 [CLS] 做点乘融合，判别图文匹配度

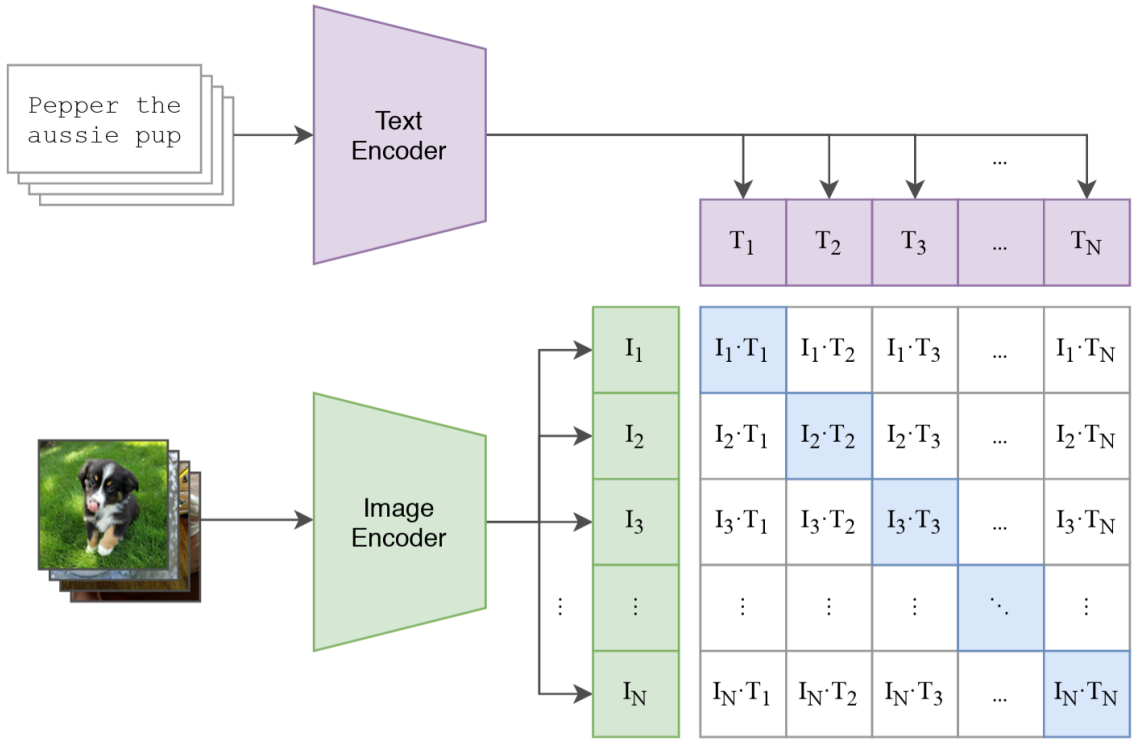


(a) Masked multi-modal learning　　(b) Multi-modal alignment prediction

Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. NIPS 2019 .

1.双流模型-CLIP

- ViLBERT 泛化性弱、检索效率差、算力消耗高
- CLIP基于**大规模图文对比学习**，实现轻量化跨模态对齐

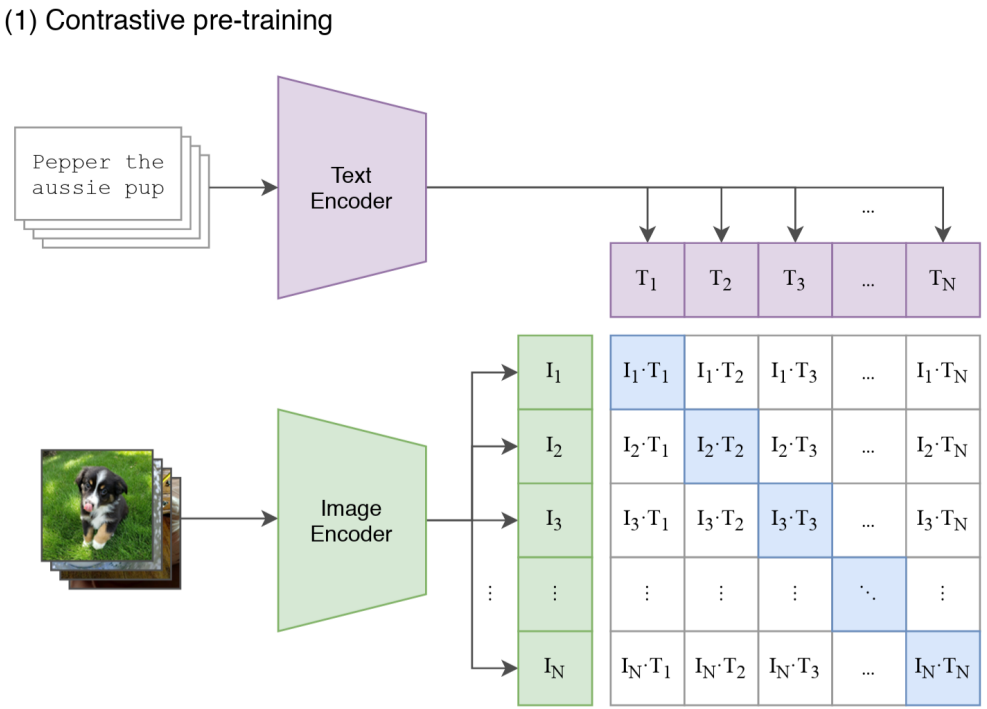


Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. NIPS 2019 .

1. 双流模型-CLIP

❑ 预训练阶段

- 在**同一个Batch**内使用图文配对数据进行对比学习



$$\mathcal{L}_{i2t} = -\frac{1}{N}\sum_{i}^{N} log\frac{exp(x_i^T y_i/\sigma)}{\sum_{j=1}^{N} exp(x_i^T y_j/\sigma)}$$

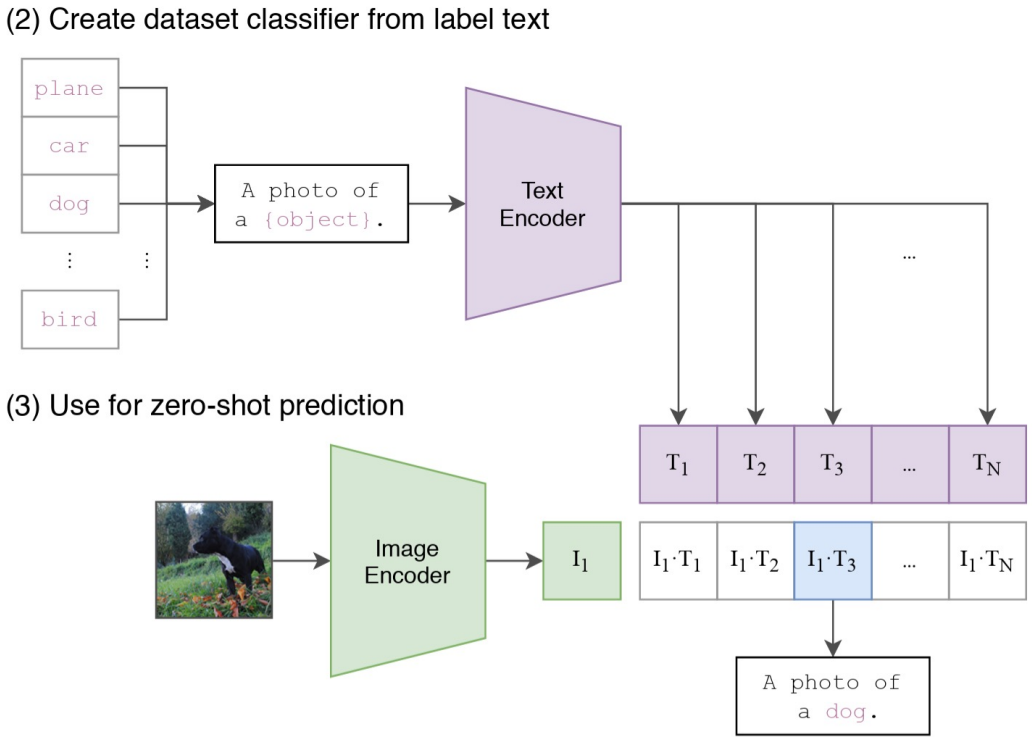
$$\mathcal{L}_{t2i} = -\frac{1}{N}\sum_{i}^{N} log\frac{exp(y_i^T x_i/\sigma)}{\sum_{j=1}^{N} exp(y_i^T x_j/\sigma)}$$

$$\mathcal{L}_{CL} = \mathcal{L}_{i2t} + \mathcal{L}_{t2i}$$

- 使用ViT 提取视觉的<CLS> 词元
- 使用BERT提取文本的<CLS> 词元
- 通过简单的对比学习在超大规模数据上进行预训练

1. 双流模型-CLIP

- 图文表征经对比学习对齐至相同语义空间，可微调适配下游任务
- 此外，CLIP 同时具备零样本图像分类能力

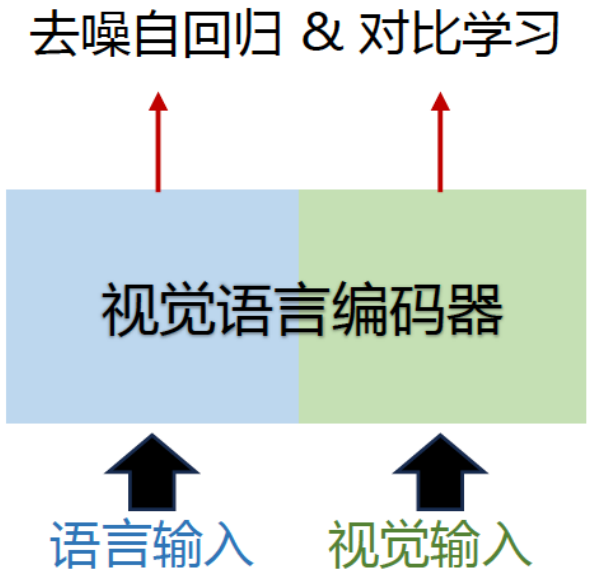


零样本图像分类：

- 对于每个类别标签，构造“A photo of a [类别]”文本提示，
- 经过编码器分别得到图文特征后计算内积，取相似度最高的类别作为预测结果

2. 单流模型

- 如何有效跨越不同模态之间的语义鸿沟，学习视觉与语言的深层交互表征，支撑下游任务推理?



函数关系：

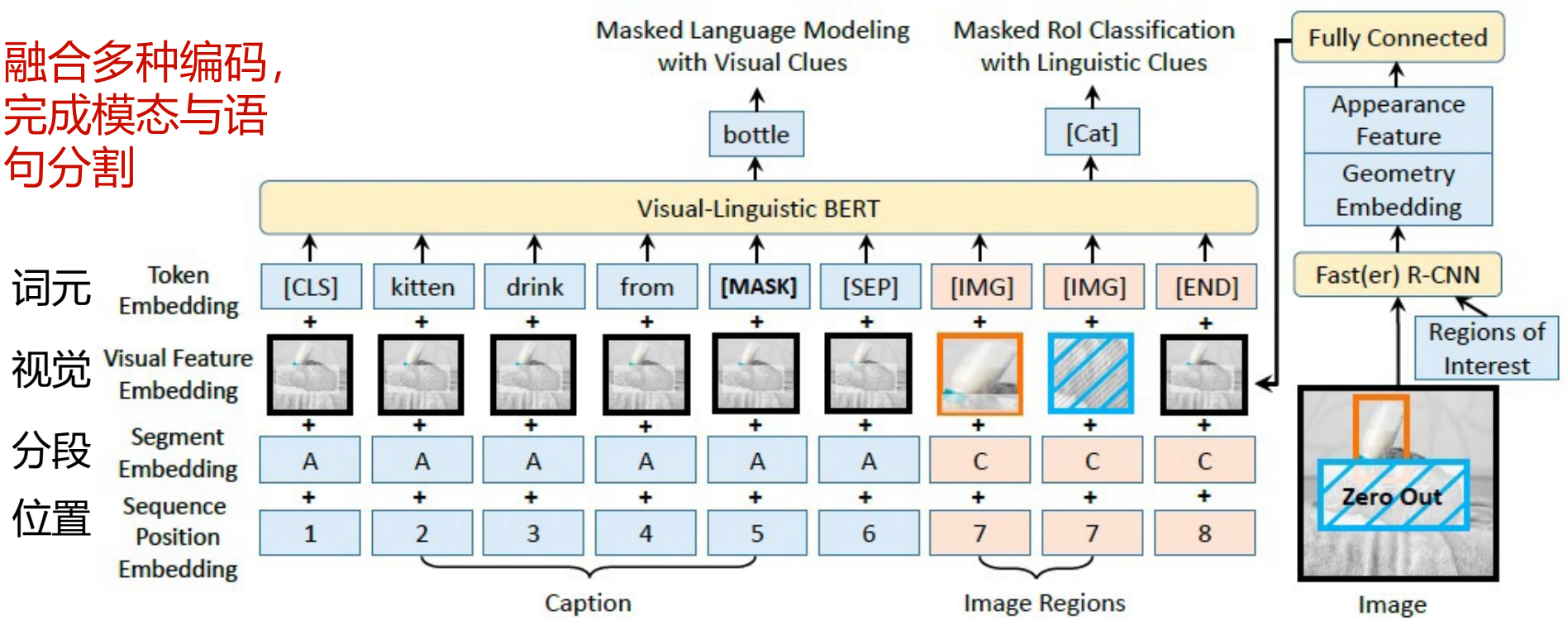
$$H = f_{single}(I, T; \theta_{single})$$

其中，f_{single} 是模型的映射函数

2. 单流模型-VLBERT

- 图像与文本模态进行拼接，输入同一个Transformer，利用注意力机制进行多模态交互

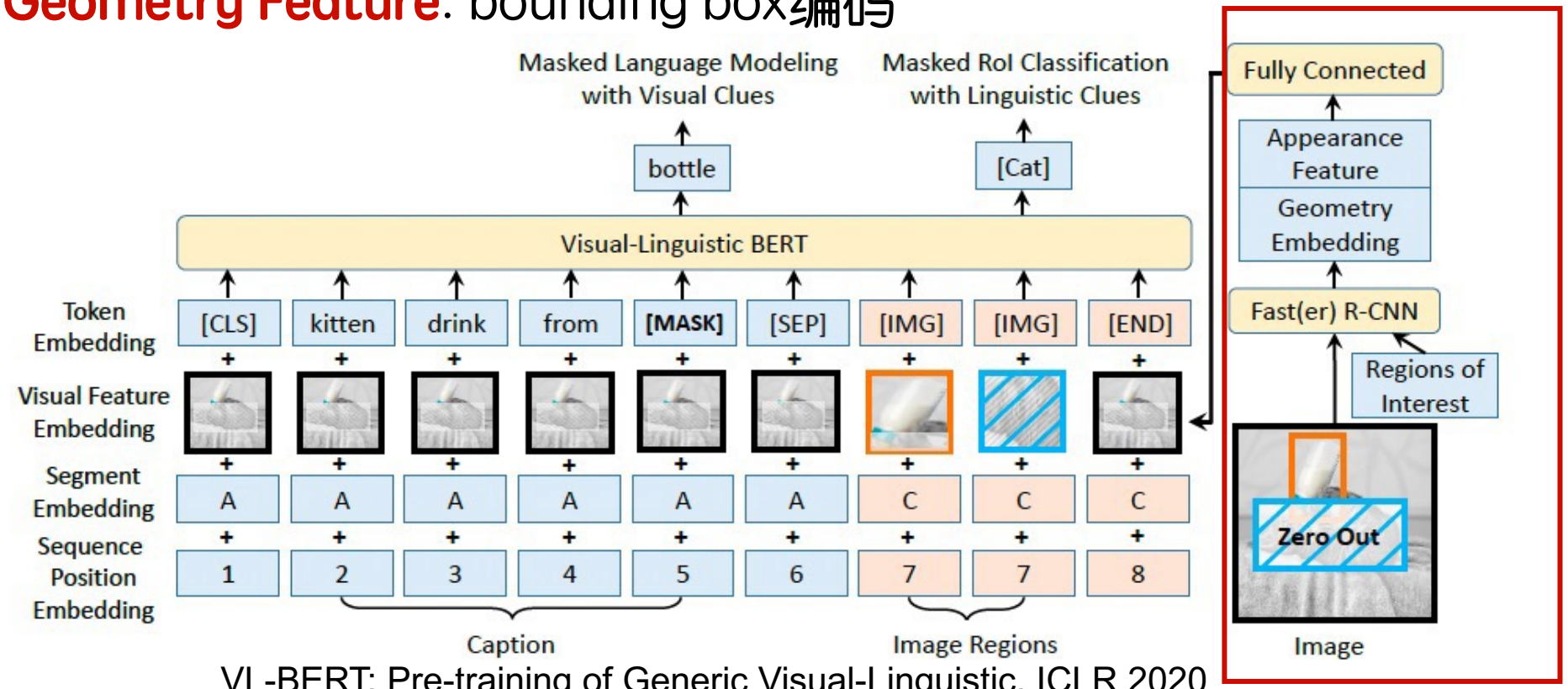
融合多种编码，完成模态与语句分割



VL-BERT: Pre-training of Generic Visual-Linguistic. ICLR 2020

2. 单流模型-VLBERT

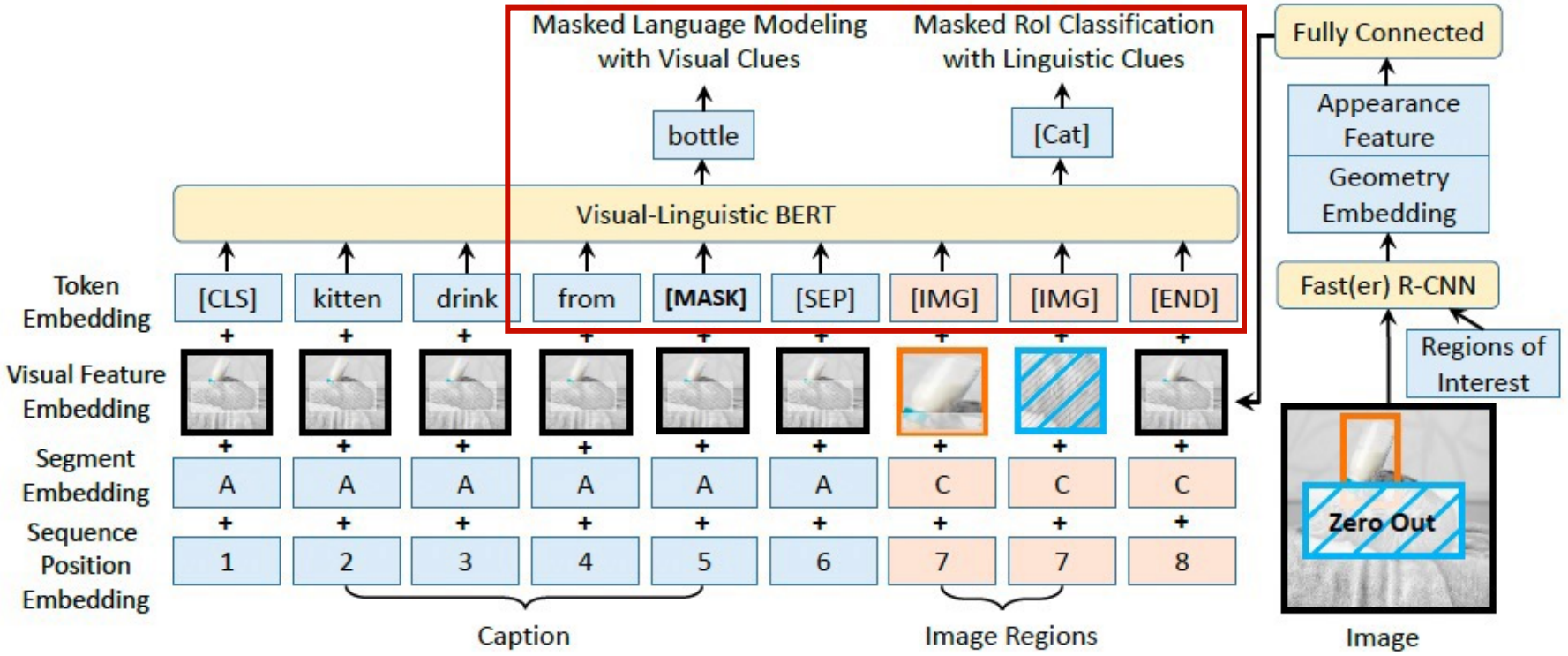
- 图像输入：同等对待图像和文本
  - **Appearance Feature**: Faster R-CNN的目标检测特征
  - **Geometry Feature**: bounding box编码



VL-BERT: Pre-training of Generic Visual-Linguistic. ICLR 2020

2. 单流模型-VLBERT

- 多模态预训练任务
  - **预测缺失的词元**：随机掩码文本词元，并预测当前位置的词元
  - **预测图像目标标签**：随机遮蔽图像区块，预测对应区域类别

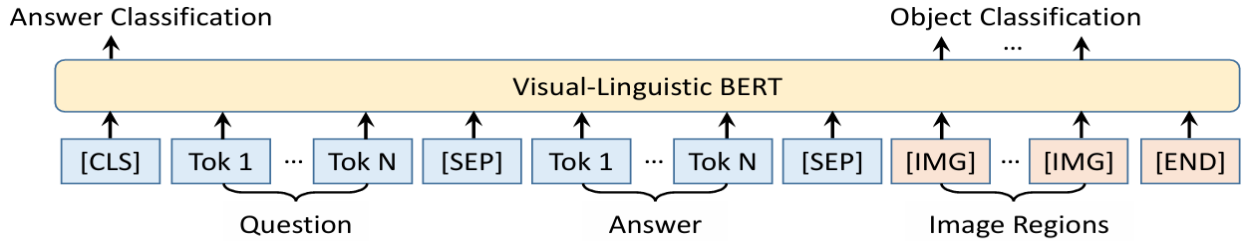


VL-BERT: Pre-training of Generic Visual-Linguistic. ICLR 2020

2. 单流模型-VLBERT

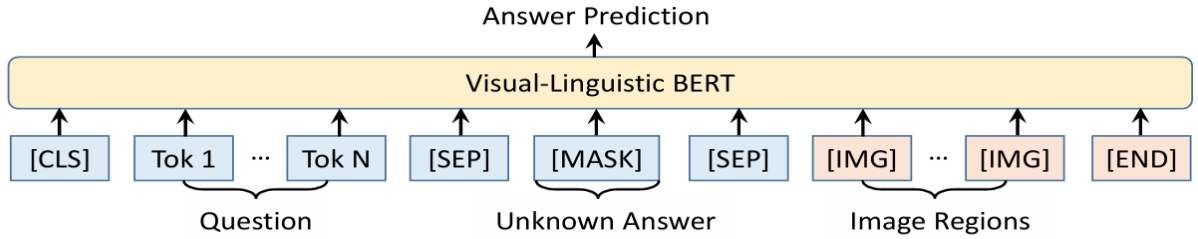
❑ 下游任务微调：利用任务标注数据微调模型参数

- 视觉常识推理（选择题）



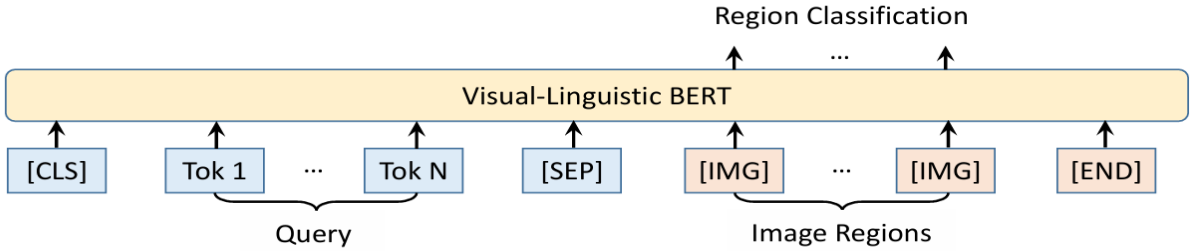
(a) Input and output format for Visual Commonsense Reasoning (VCR) dataset

- 视觉问答（共享答案池）



(b) Input and output format for Visual Question Answering (VQA) dataset

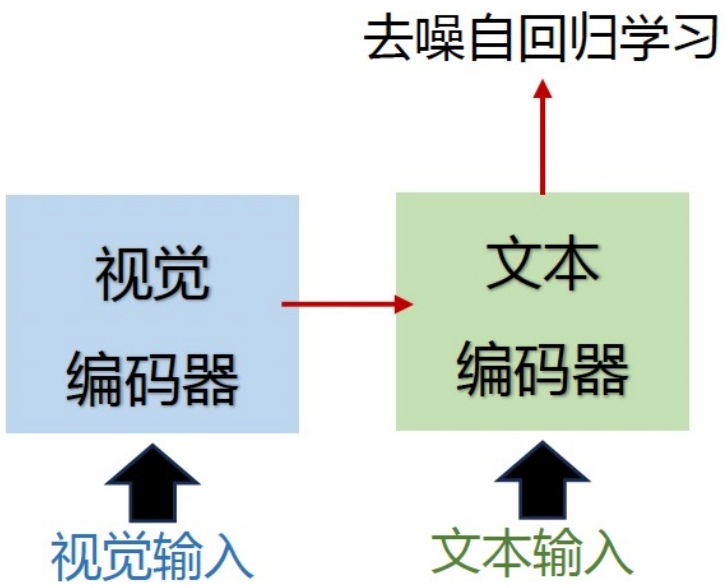
- 指称表达理解（物体定位）



(c) Input and output format for Referring Expression task on RefCOCO+ dataset

3. 编解码模型

❑ 将视觉和语言输入融合成一个统一的表示，并具备文本生成能力



函数关系：

$$H_{enc} = f_{encoder}(I;\theta_{encoder}),$$

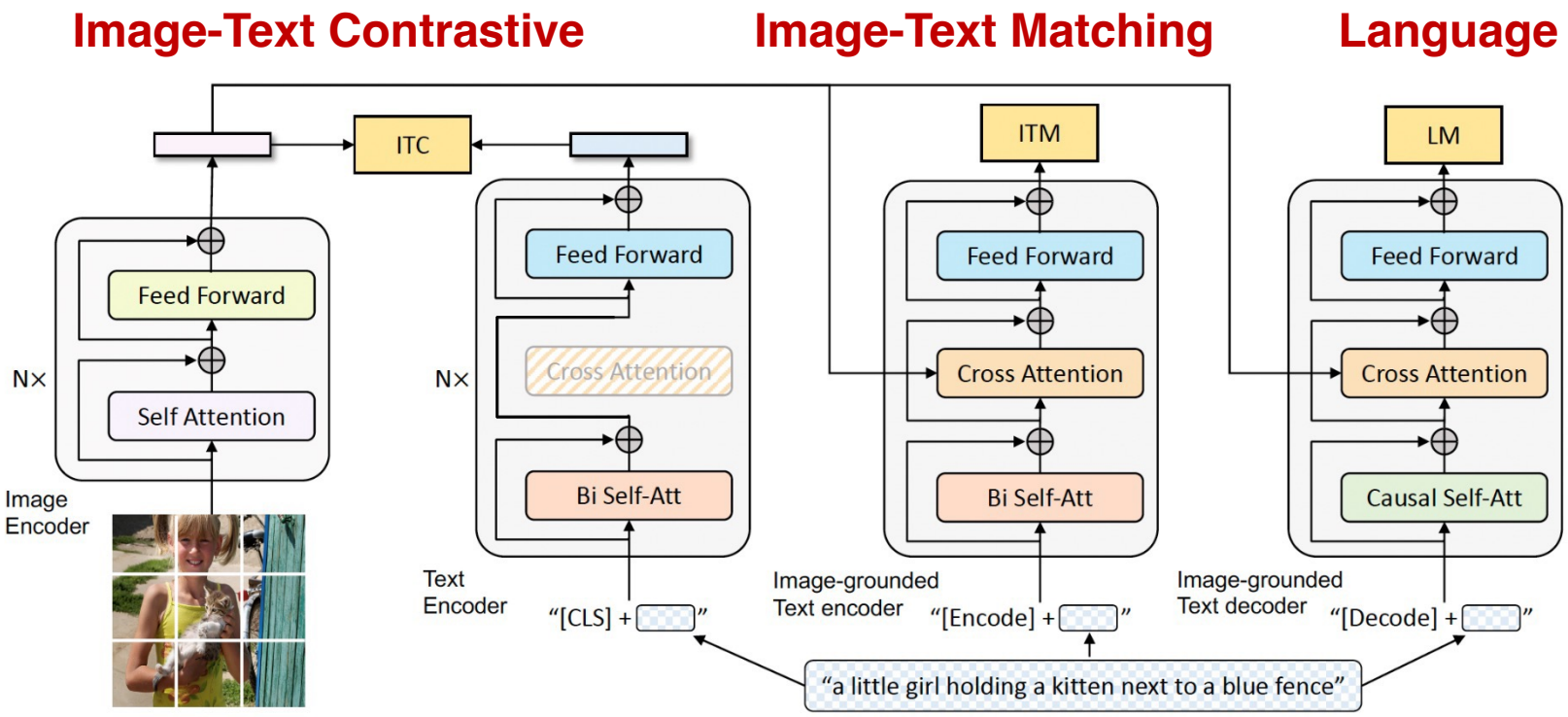
$$T_{dec} = f_{decoder}(H_{enc},T_{prev};\theta_{decoder}).$$

其中，$f_{encoder}$ 是编码器，编码图像特征

$f_{decoder}$ 是编码器，生成新的文本

3. 编解码模型-BLIP

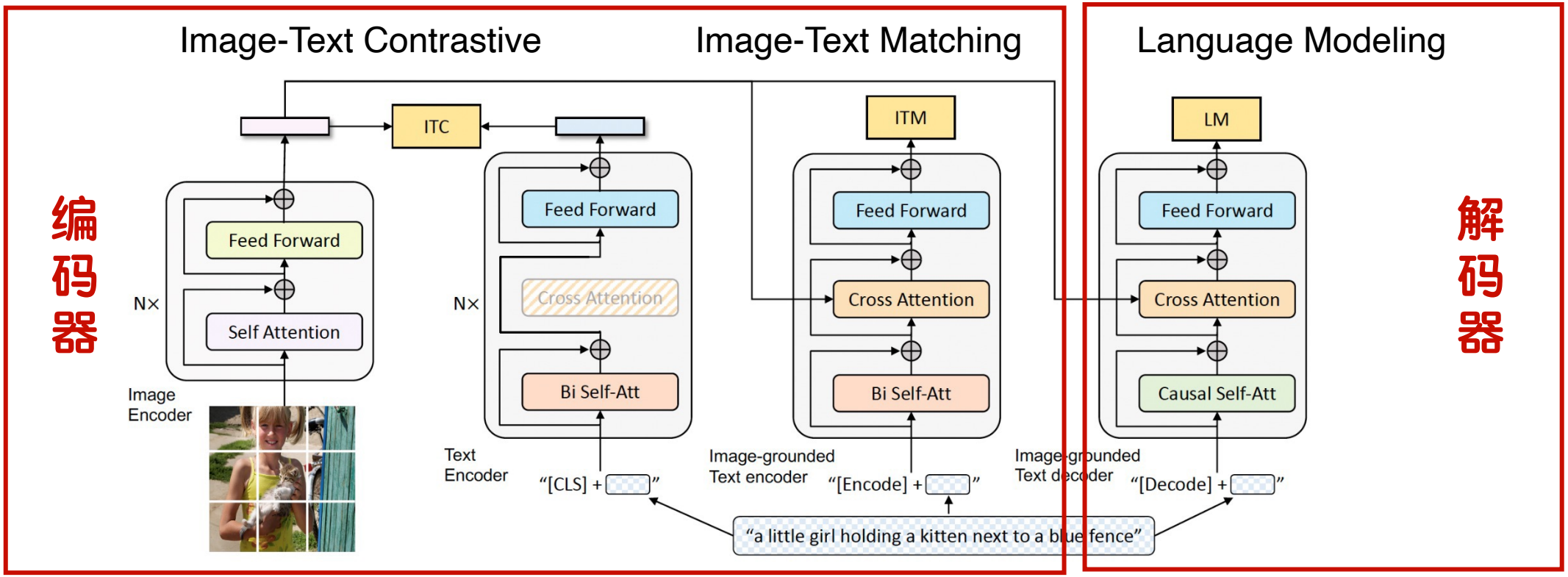
- 上述方法关注于理解和检索，难以实现跨模态生成任务
- 针对不同任务的适配需要，设计多个预训练任务目标



BLIP-Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. ICML 2022.

3. 编解码模型-BLIP

- 编码器：利用ITC和ITM任务，拉近视觉空间和语言空间的双流模型
- 解码器：根据图像信息生成文本描述，采用自回归的交叉熵损失函数训练



BLIP-Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. ICML 2022.

中国科学院大学
University of Chinese Academy of Sciences

目录

传统多模态预训练困境

❑ 截止2022年，代表性模型参数规模

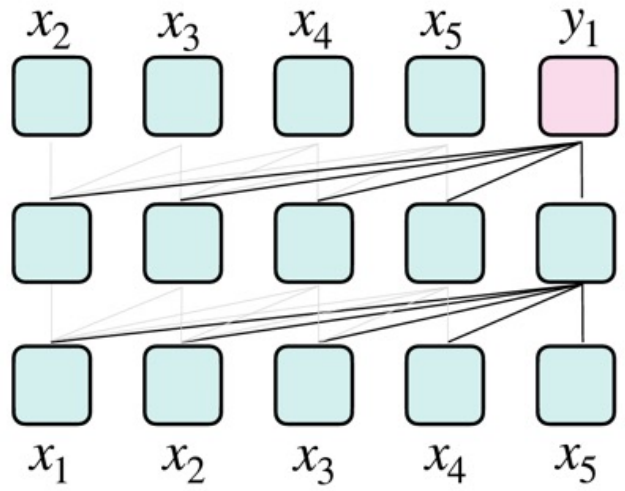
模型	BLIP	OFA	CoCa	BeiT-3	GIT	PaLi
参数	0.3B	0.9B	2.1B	1.9B	5.1B	17B

❑ 数据规模

- 14M高质量图文匹配对(COCO,VG,CC,SBU)
- 100M~5B 弱匹对(LAION, in-house data)

ChatGPT问世

❑ 2022年底ChatGPT横空出世，模型参数规模越大，模型能力越强

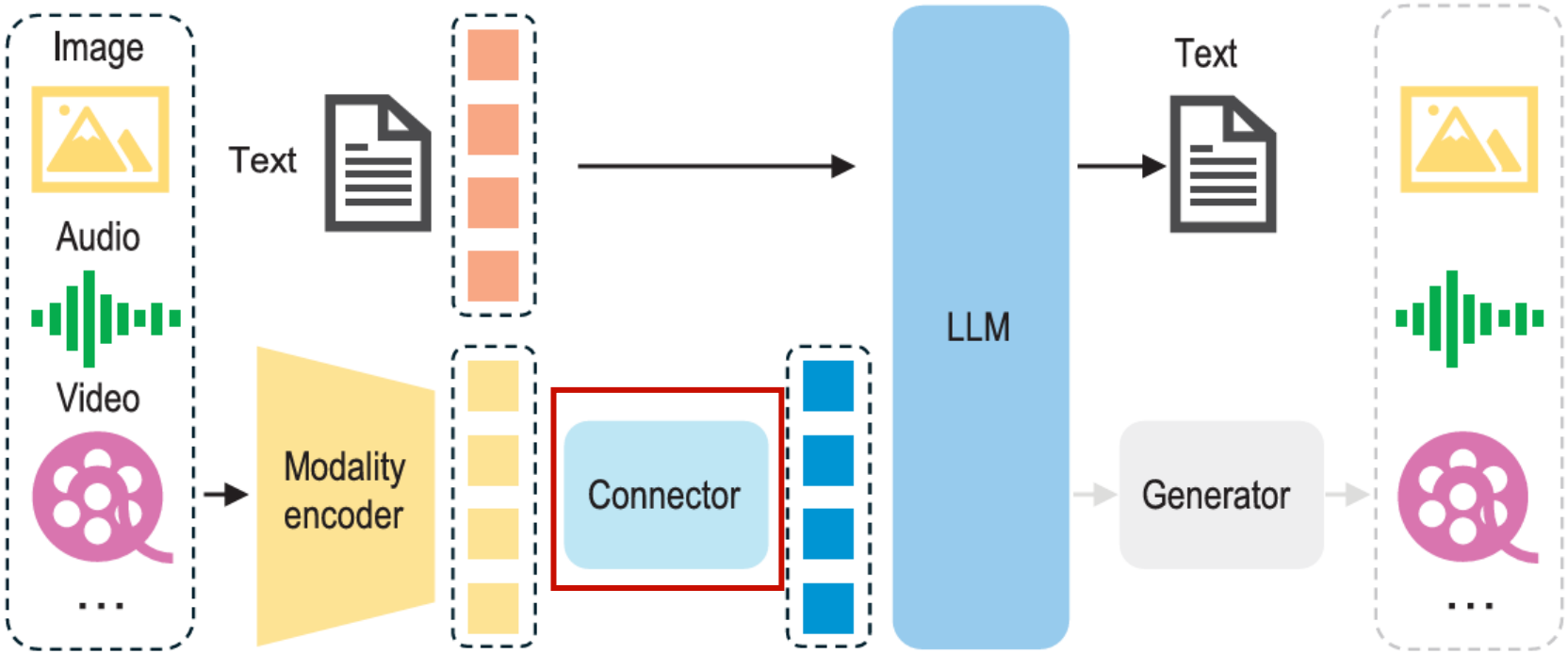


ChatGPT基础模型：GPT

模型	层数	维度	参数
GPT (2018)	12	768	1.17亿
GPT-2 (2019)	48	1600	15亿
GPT-3 (2020)	96	12,288	1750亿

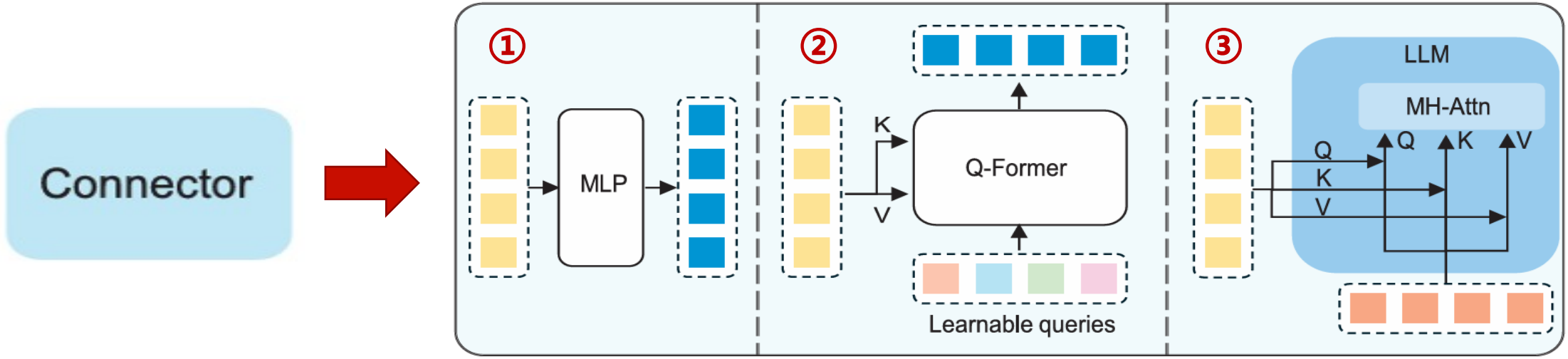
大模型时代的多模态预训练模型

- 将多模态数据作为语言，接入大语言模型，适配和利用大语言模型的推理能力



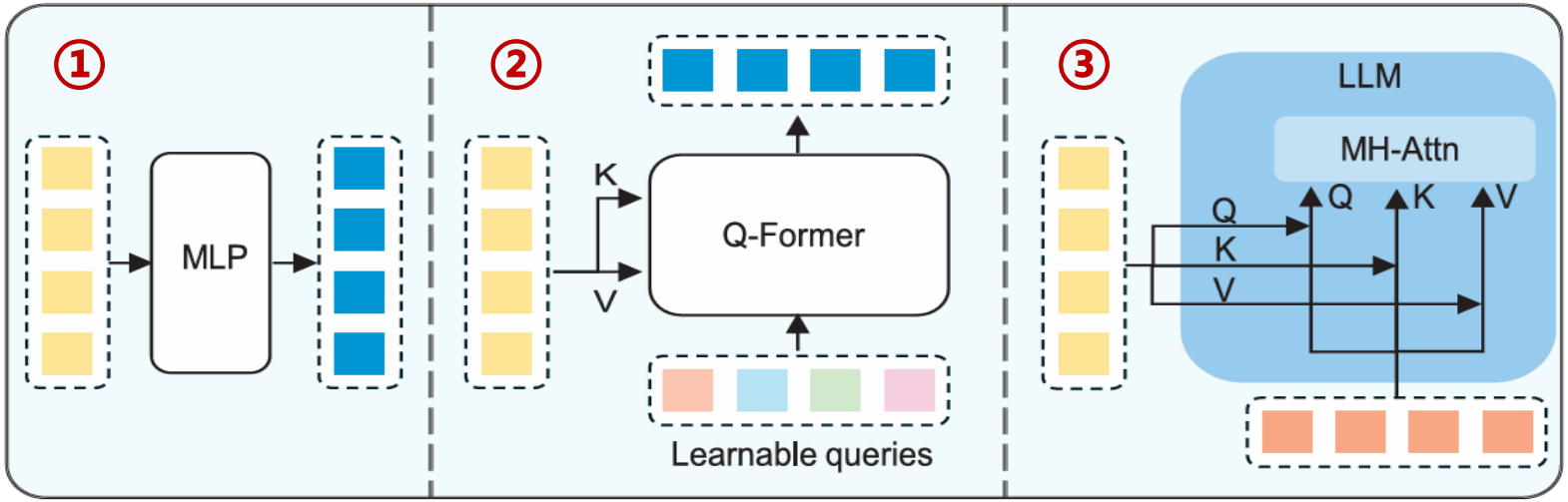
多模态大模型分类

- 根据多模态信息的适配方式，分为以下三类：
  - **基于映射**：直接将视觉特征，映射至语言特征空间
  - **基于查询**：设计可学查询向量，关注视觉特征，获得语言token
  - **基于融合**：设计视觉语言交互，将视觉信息注入LLM中间层



多模态大模型分类

- 对于三类适配方式，各自代表性方法：
  - **基于映射方法**：LLaVA、MiniGPT-4
  - **基于查询方法**：BLIP-2、MiniGPT-5、EMU、Flamingo
  - **基于融合方法**：Flamingo



1. 基于映射的方法-LLaVA

- 如何更好地利用LLM的视觉指令遵循能力?
- 构建视觉指令数据集，微调开源LLaMA使之适应视觉语言输入

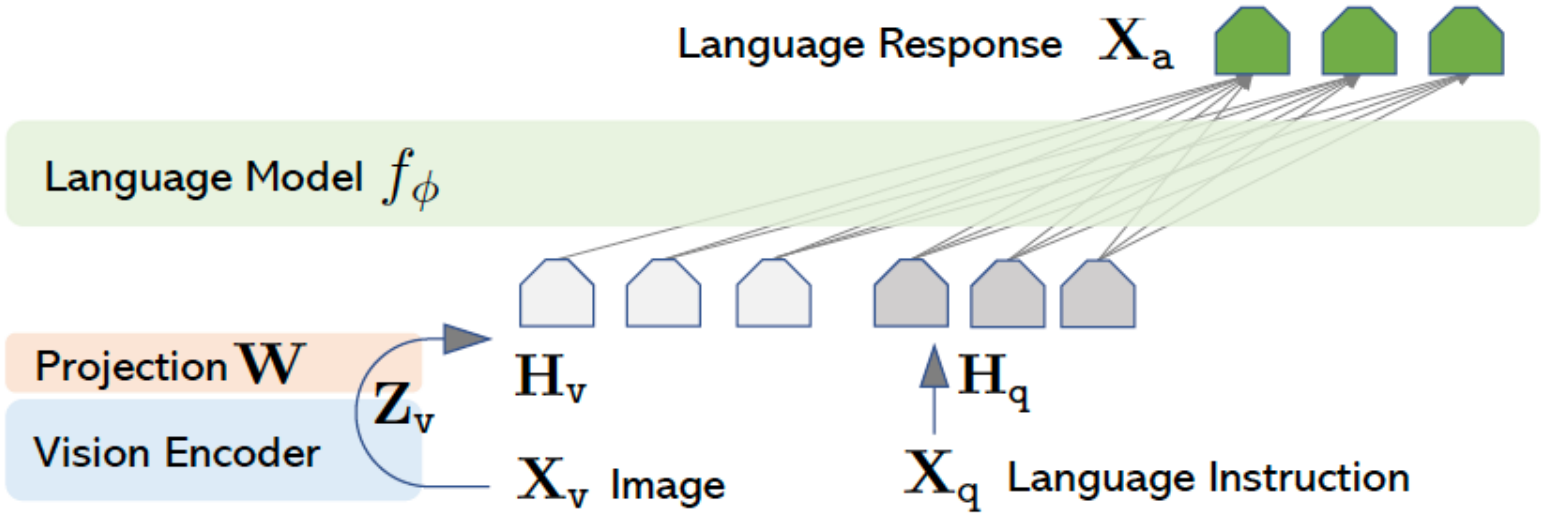


Figure 1: LLaVA network architecture.

Visual instruction tuning. NeurIPS 2023.

1. 基于映射的方法-LLaVA

❑ **模型结构**：在图像编码后接入映射模块，将图像表征序列输入LLM

- **图像编码器**：CLIP ViT-L/14 视觉编码器
- **语言模型**：Vicuna，基于ShareGPT对话数据微调Llama2得到的大模型

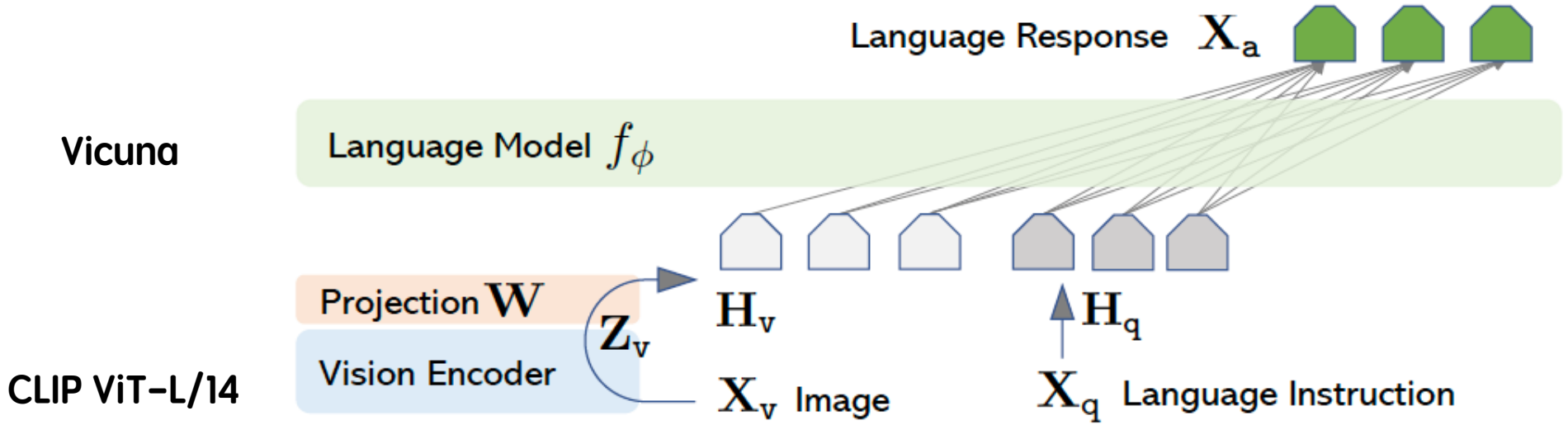


Figure 1: LLaVA network architecture.

Visual instruction tuning. NeurIPS 2023.

1. 基于映射的方法-LLaVA

❑ **数据构造**：构造大规模指令数据，提升大模型指令理解能力

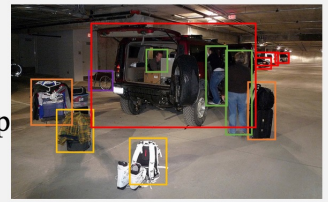
给定信息

- **图像描述**
- **物体位置**

输出信息

- **对话数据**
- **描述数据**
- **复杂推理**

Context type 1: Captions
A group of people standing outside of a black vehicle with various luggage.
Luggage surrounds a vehicle in an underground parking area
People try to fit all of their luggage in an SUV.
The sport utility vehicle is parked in the public garage, being packed for a trip
Some people with luggage near a van that is transporting it.



Context type 2: Boxes
person: [0.681, 0.242, 0.774, 0.694], backpack: [0.384, 0.696, 0.485, 0.914], suitcase: ...<omitted>

Response type 1: conversation
Question: What type of vehicle is featured in the image?
Answer: The image features a black sport utility vehicle (SUV) ...<omitted>
Response type 2: detailed description
The image is an underground parking area with a black sport utility vehicle (SUV) parked. There are three people in the scene, with one person standing closer to the left side of the vehicle, another person in the middle, and the third person on the right side. They are all working together to pack their luggage into the SUV for a trip. ...<omitted>
Response type 3: complex reasoning
Question: What challenges do these people face?
Answer: In the image, a group of people is standing outside a black SUV in a parking area, surrounded by various pieces of luggage, including suitcases and backpacks. They are facing the challenge of fitting all their luggage into the black SUV. There are multiple suitcases and backpacks to be packed, which suggests that the group has a significant amount of belongings ...<omitted>

1. 基于映射的方法-LLaVA

❑ **构造指令数据集**：利用ChatGPT/GPT4获取Visual Instruction数据

系统信息

示例数据

开始提问

```
messages = [ {"role":"system", "content": f"""You are an AI visual assistant, and you are
seeing a single image. What you see are provided with five sentences, describing the same image you
are looking at. Answer all questions as you are seeing the image.

Design a conversation between you and a person asking about this photo. The answers should be in a
tone that a visual AI assistant is seeing the image and answering the question. Ask diverse questions
and give corresponding answers.

Include questions asking about the visual content of the image, including the object types, counting
the objects, object actions, object locations, relative positions between objects, etc. Only include
questions that have definite answers:
(1) one can see the content in the image that the question asks about and can answer confidently;
(2) one can determine confidently from the image that it is not in the image. Do not ask any question
that cannot be answered confidently.

Also include complex questions that are relevant to the content in the image, for example, asking
about background knowledge of the objects in the image, asking to discuss about events happening in
the image, etc. Again, do not ask about uncertain details. Provide detailed answers when answering
complex questions. For example, give detailed examples or reasoning steps to make the content more
convincing and well-organized. You can include multiple paragraphs if necessary."""}
]
for sample in fewshot_samples:
    messages.append({"role":"user", "content":sample['context']})
    messages.append({"role":"assistant", "content":sample['response']} )
messages.append({"role":"user", "content":'\n'.join(query)})
```

1. 基于映射的方法-LLaVA

- **阶段一**：预训练**Projection参数**，用于视觉模型和语言模型的表征空间对齐
- 训练数据： CC3M to 595K image-text pairs

用于训练的指令数据格式：

$$\mathbf{X}_{\text{system-message}} \texttt{<STOP>} \backslash\text{n}$$
$$\text{Human}: \mathbf{X}^1_{\text{instruct}} \texttt{<STOP>} \backslash\text{n}\ \text{Assistant}: \mathbf{X}^1_{\text{a}} \texttt{<STOP>} \backslash\text{n}$$
$$\text{Human}: \mathbf{X}^2_{\text{instruct}} \texttt{<STOP>} \backslash\text{n}\ \text{Assistant}: \mathbf{X}^2_{\text{a}} \texttt{<STOP>} \backslash\text{n} \cdots$$

$$\mathbf{X}^t_{\text{instruct}} = \begin{cases} \text{首轮包含图像和文本：} \\ \text{Random choose } [\mathbf{X}^1_{\text{q}}, \mathbf{X}_{\text{v}}] \text{ or } [\mathbf{X}_{\text{v}}, \mathbf{X}^1_{\text{q}}], \\ \text{后续包含多轮对话问题：} \ \mathbf{X}^t_{\text{q}}, \end{cases}$$

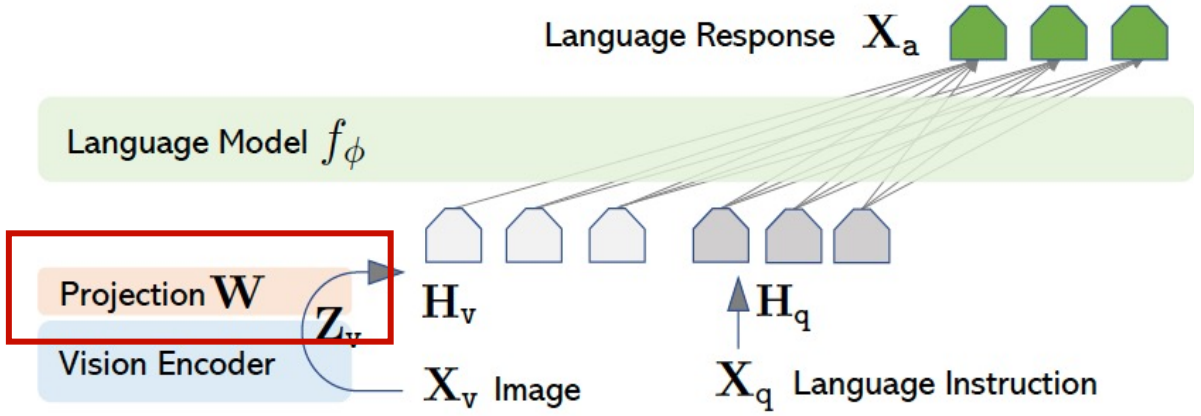


Figure 1: LLaVA network architecture.

$$p(\mathbf{X}_{\text{a}}|\mathbf{X}_{\text{v}}, \mathbf{X}_{\text{instruct}}) = \prod_{i=1}^{L} p_{\boldsymbol{\theta}}(x_i|\mathbf{X}_{\text{v}}, \mathbf{X}_{\text{instruct},<i}, \mathbf{X}_{\text{a},<i}),$$

Visual instruction tuning. NeurIPS 2023.

1. 基于映射的方法-LLaVA

- **阶段二**：固定视觉编码器，微调**Projection**和**LLM参数**，适配指令遵循能力
- 训练数据：
  - 多模态对话数据：前文构建的 158K 图文指令跟随数据集
  - 科学问答数据：大规模多模态科学问答数据集，支持图文问答
  - 统一沿用第一阶段指令数据格式

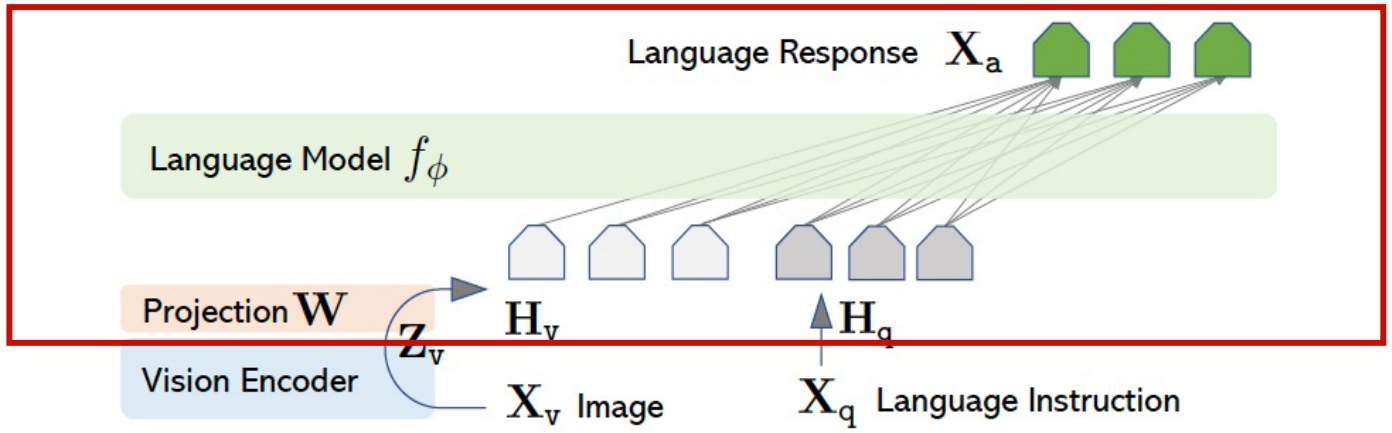
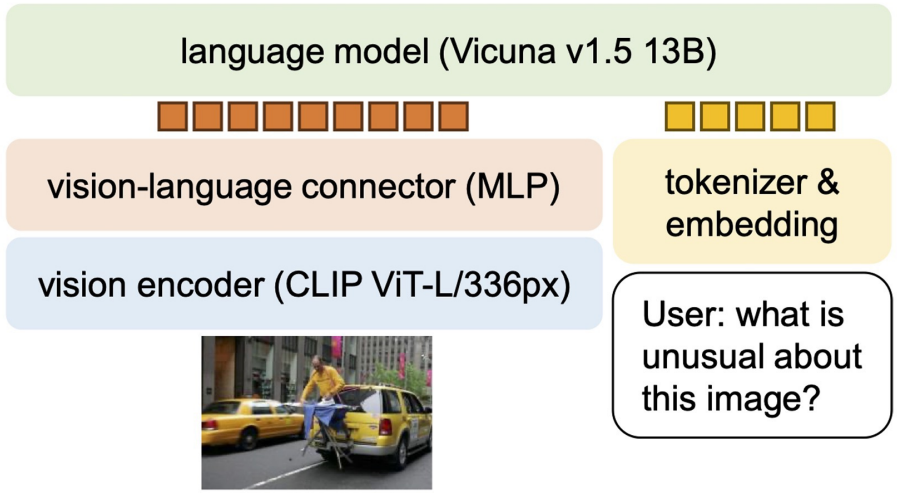


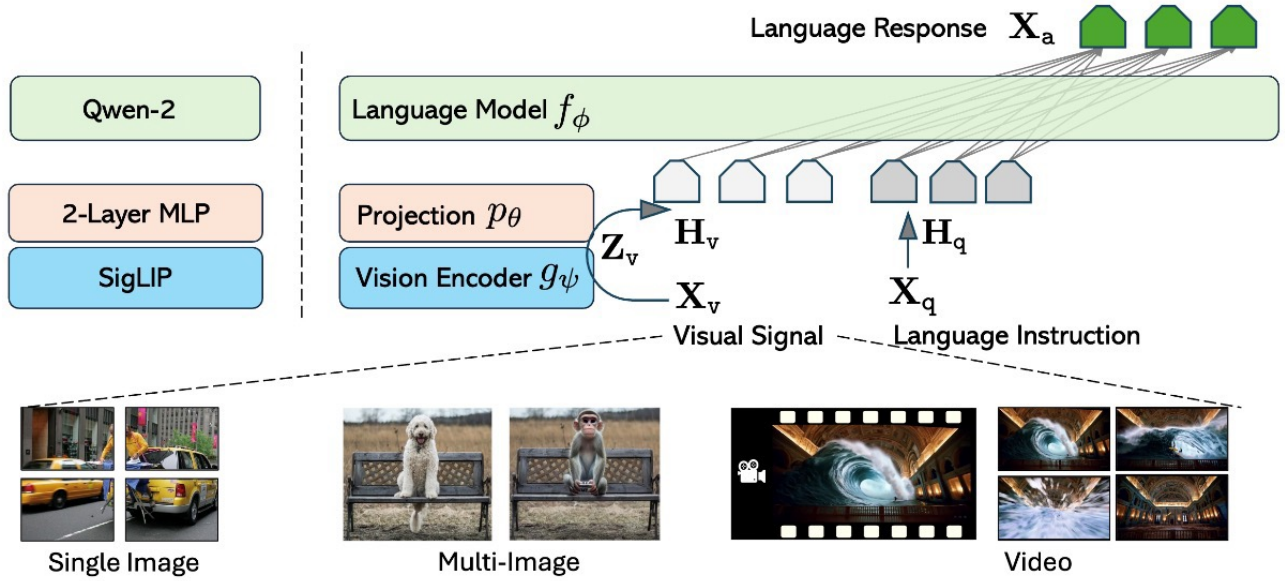
Figure 1: LLaVA network architecture.
Visual instruction tuning. NeurIPS 2023.

1. 基于映射的方法-LLaVA系列

- 在LLaVA的基础上进行升级，对模型结构、数据集数量及丰富度进行改进



LLaVA-1.5



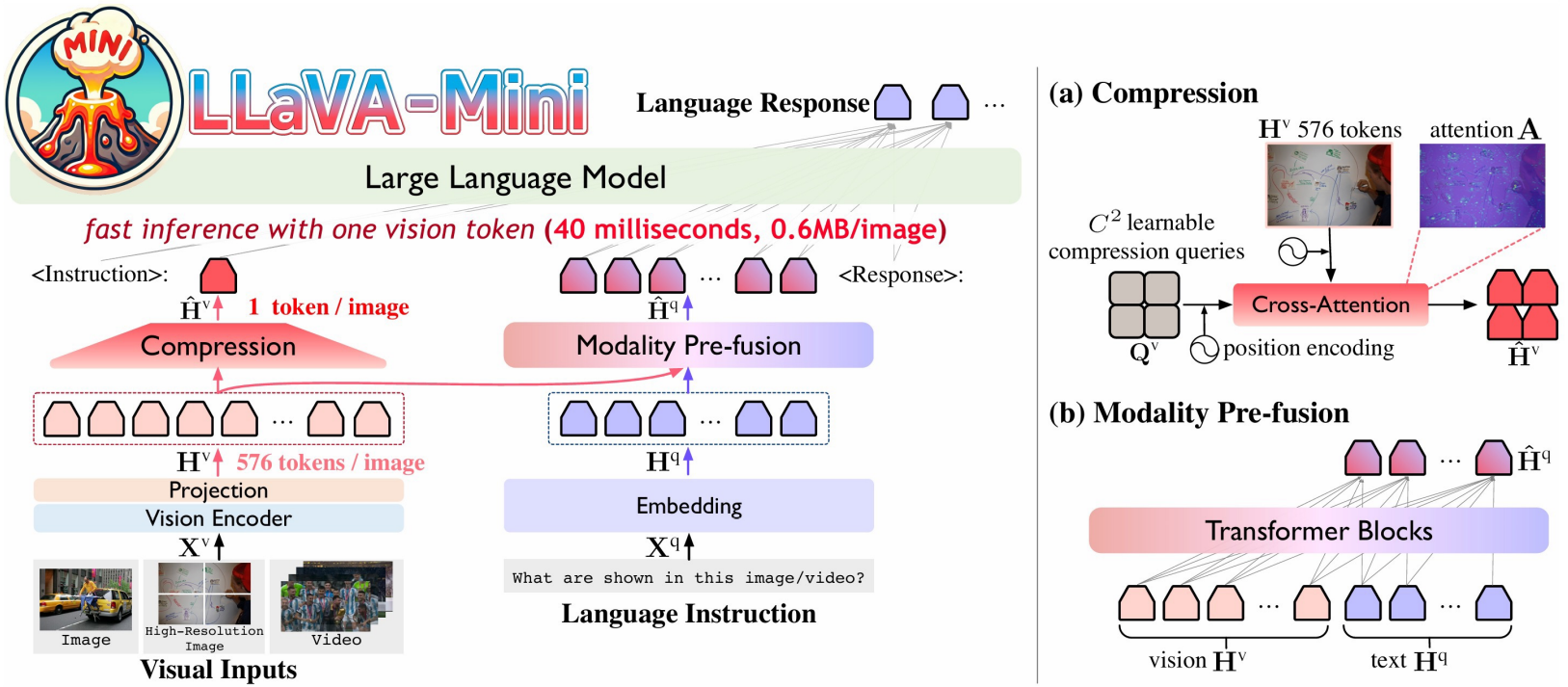
LLaVA-OneVision

Improved Baselines with Visual Instruction Tuning. CVPR 2024.

LLaVA-OneVision: Easy Visual Task Transfer. 2024.

1. 基于映射的方法-LLaVA系列

- 在LLaVA的基础上进行升级，对模型结构、数据集数量及丰富度进行改进

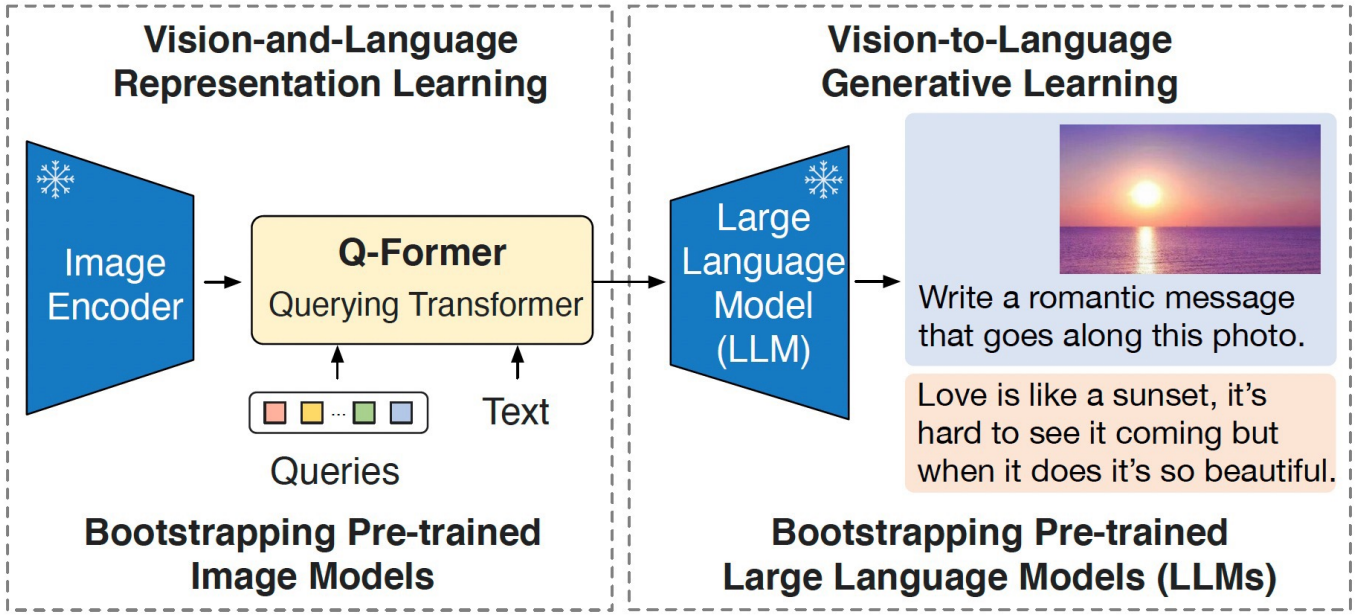


LLaVA-Mini

LLaVA-Mini: Efficient Image and Video Large Multimodal Models with One Vision Token. ICLR 2025.

2. 基于查询的方法-BLIP-2

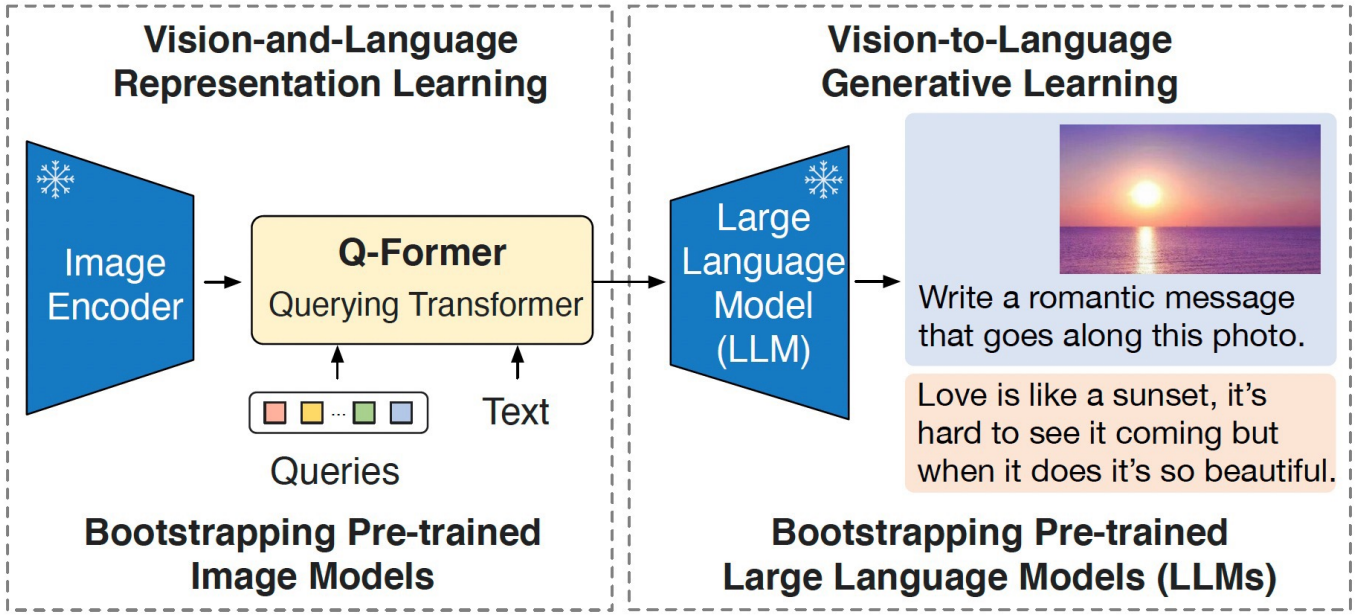
- 不同于表示映射的方法，如何进一步增强跨模态的对齐?
- BLIP-2在图像和LLM之间建立适配器，利用LLM提升生成能力



BLIP-2 Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. ICML 2023

2. 基于查询的方法-BLIP-2

- 两阶段设计：
  - 阶段1: 设计Q-Former，基于文本信息选择视觉信息
  - 阶段2: 利用LLM，将选择的视觉信息对齐到LLM语言空间

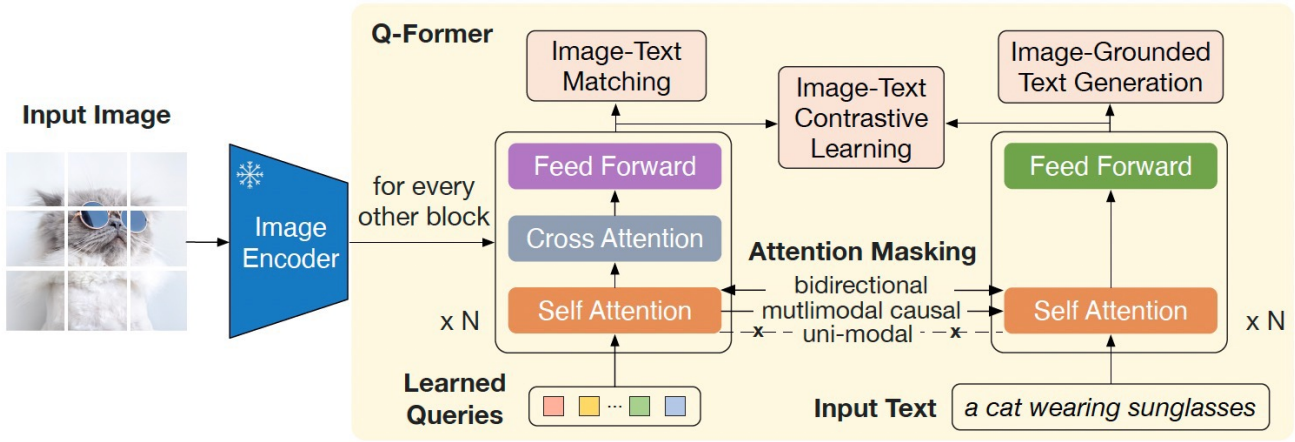


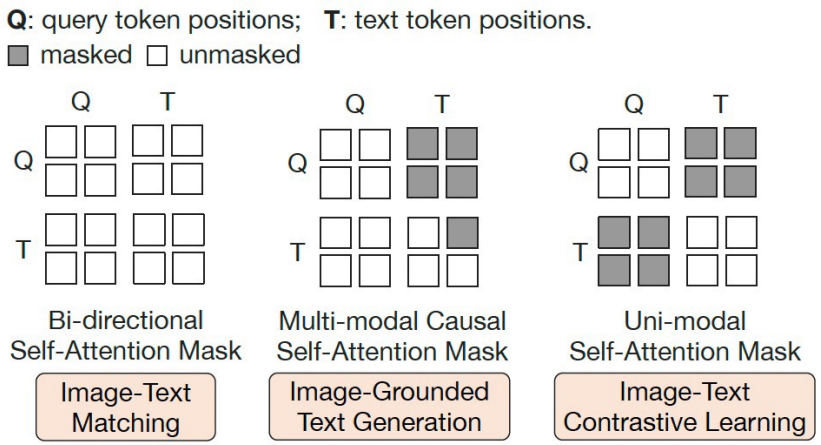
BLIP-2 Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. ICML 2023

2. 基于查询的方法-BLIP-2

❑ 阶段一：预训练Q-Former适配器

- **自注意力机制参数共享**：Query 结合文本上下文筛选视觉信息
- **三种视觉语言损失：**使得交叉注意力机制根据query选择视觉信息，并拉近视觉和文本特征空间

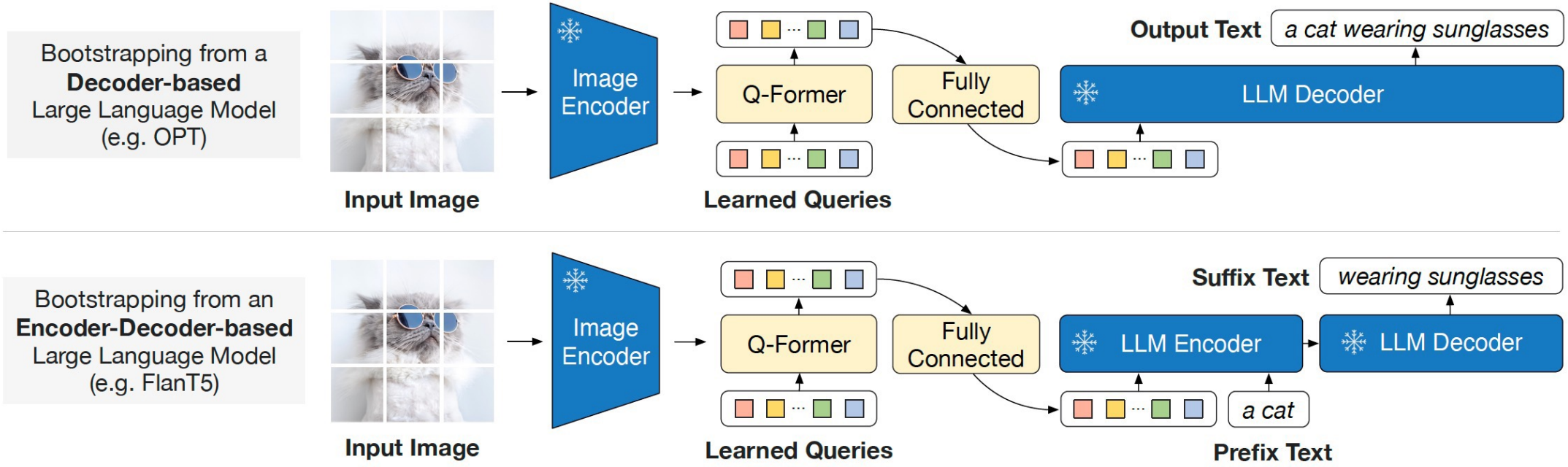




BLIP-2 Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. ICML 2023

2. 基于查询的方法-BLIP-2

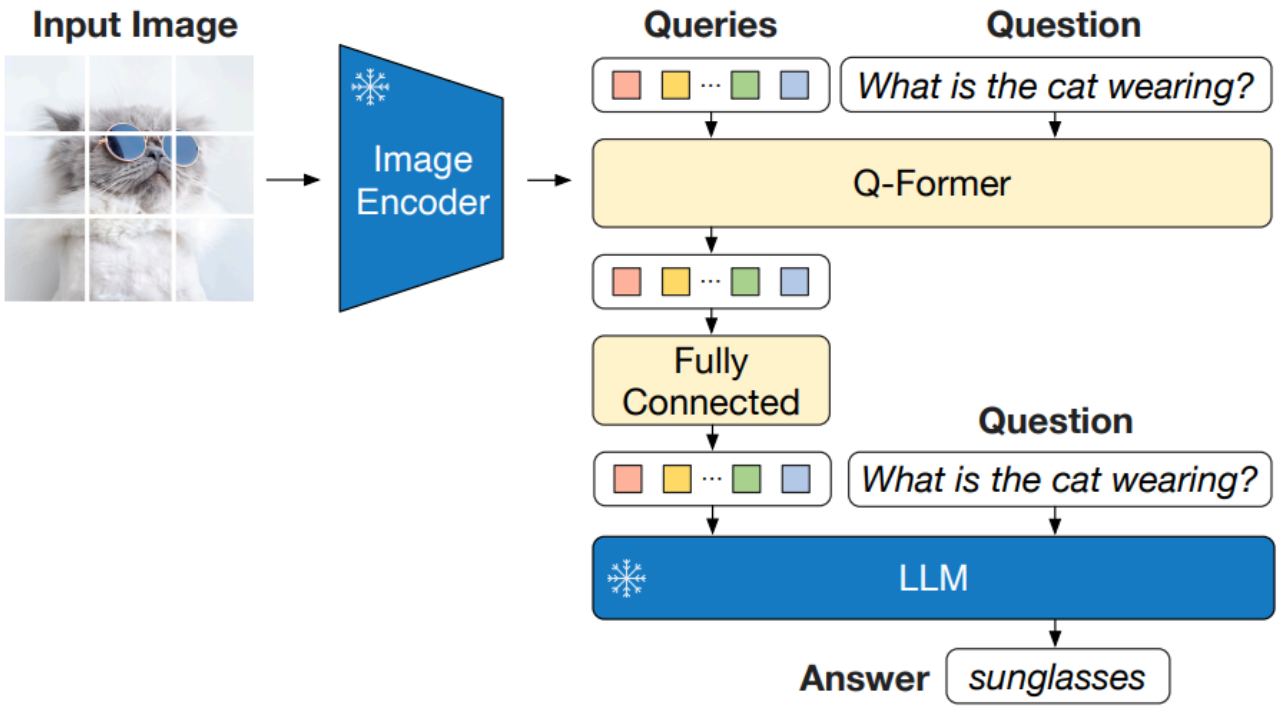
- 阶段二：将图像信息对齐至LLM语言空间
  - 利用全连接层，将Stage 1 中选择的视觉特征映射，将其转换为文本token并输入LLM，以完成各种多模态下游任务



BLIP-2 Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. ICML 2023

2. 基于查询的方法-BLIP-2

- 下游任务：视觉问答
  - 将图像、查询、问题三者作为模型输入，利用大模型进行答案输出



BLIP-2 Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. ICML 2023

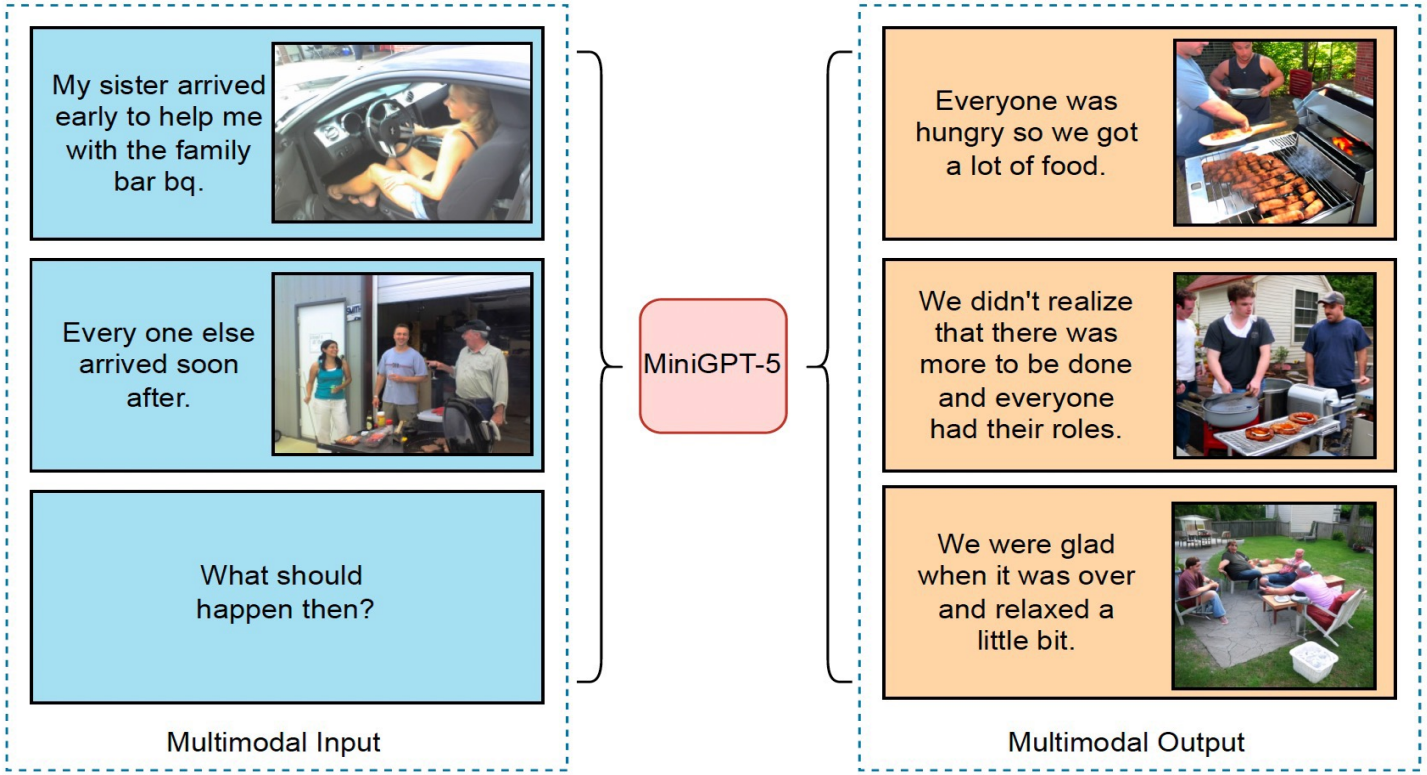
2. 基于查询的方法-BLIP-2

❑ 案例效果

BLIP-2 Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. ICML 2023

2. 基于查询的方法-MiniGPT-5

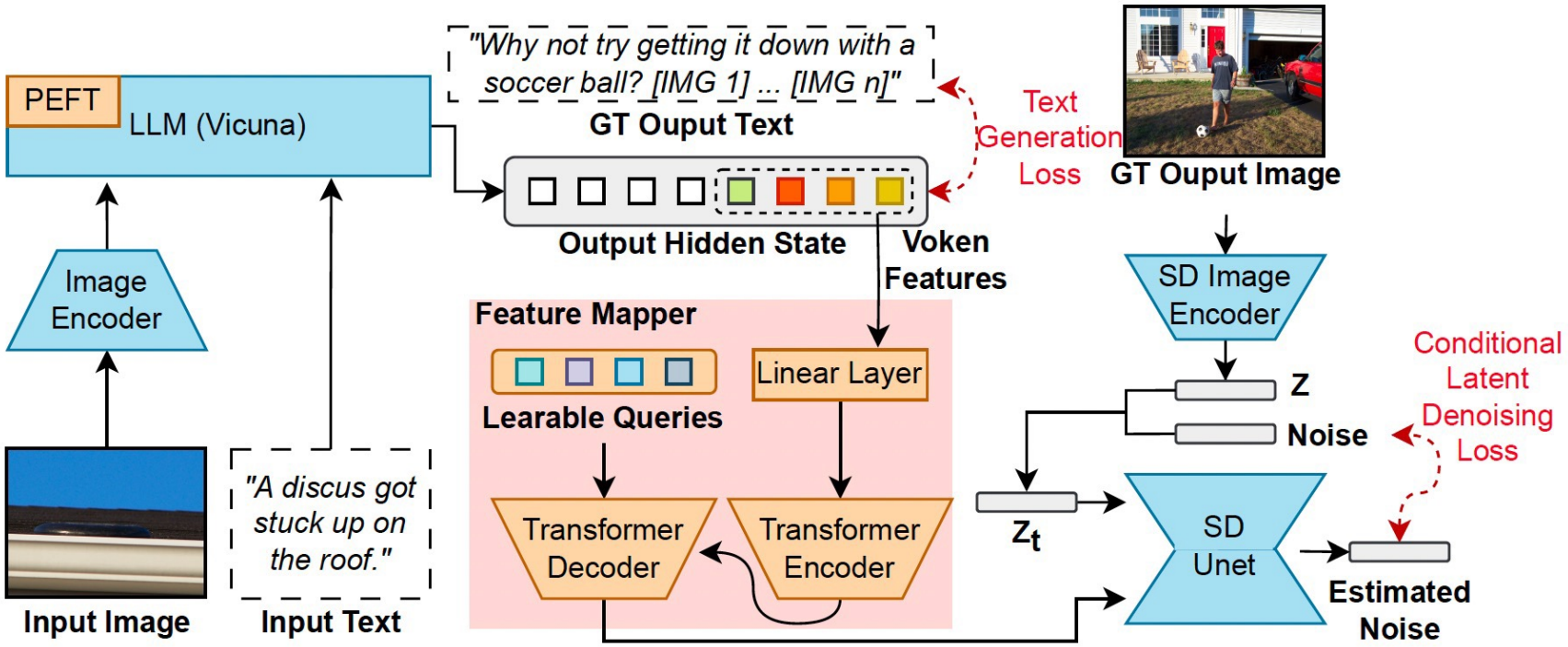
❑ 如何使得LLM-based的视觉预训练模型获得**文到图生成**的能力?



Minigpt-5: Interleaved vision-and-language generation via generative vokens. 2023

2. 基于查询的方法-MiniGPT-5

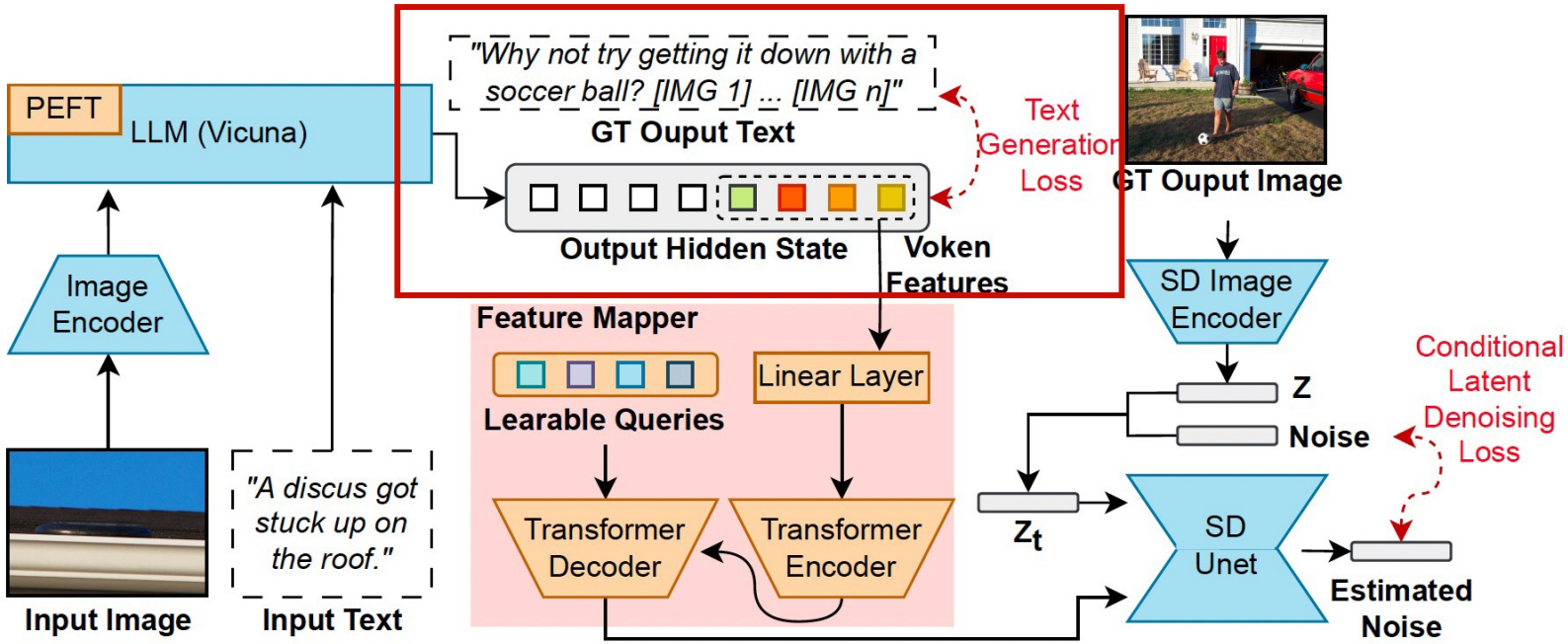
- 基于MiniGPT-4与diffusion model，提出视觉Token（Voken）用于图文对齐，提取多模态表征后利用扩散模型进行图像生成



Minigpt-5: Interleaved vision-and-language generation via generative vokens. 2023

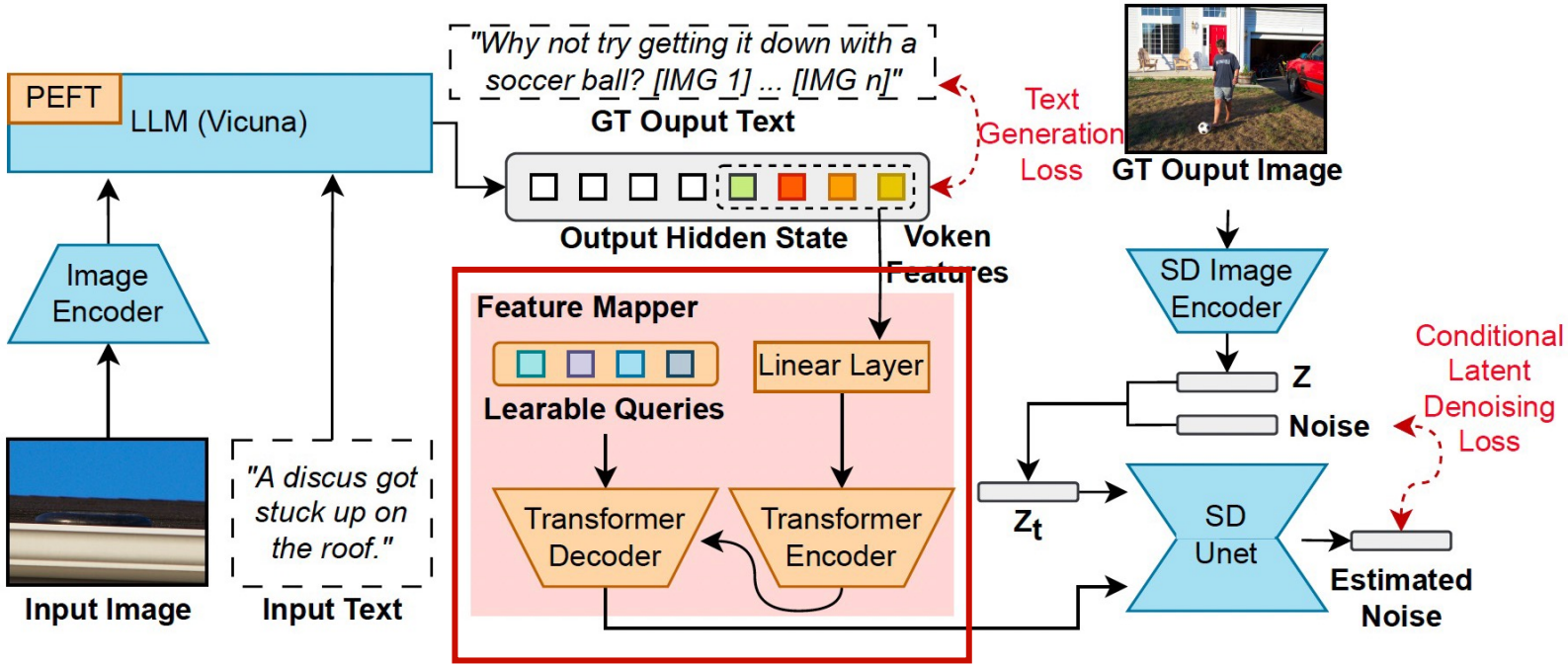
2. 基于查询的方法-MiniGPT-5

- 步骤一：**图文对齐**
  - **Vokens**：拓展大模型词表，输入词表和输出词表各新增8个参数，用 Voken 占位待生成图像位置
  - 将Vokens与文本联合进行自回归预训练，使得模型可预测Vokens



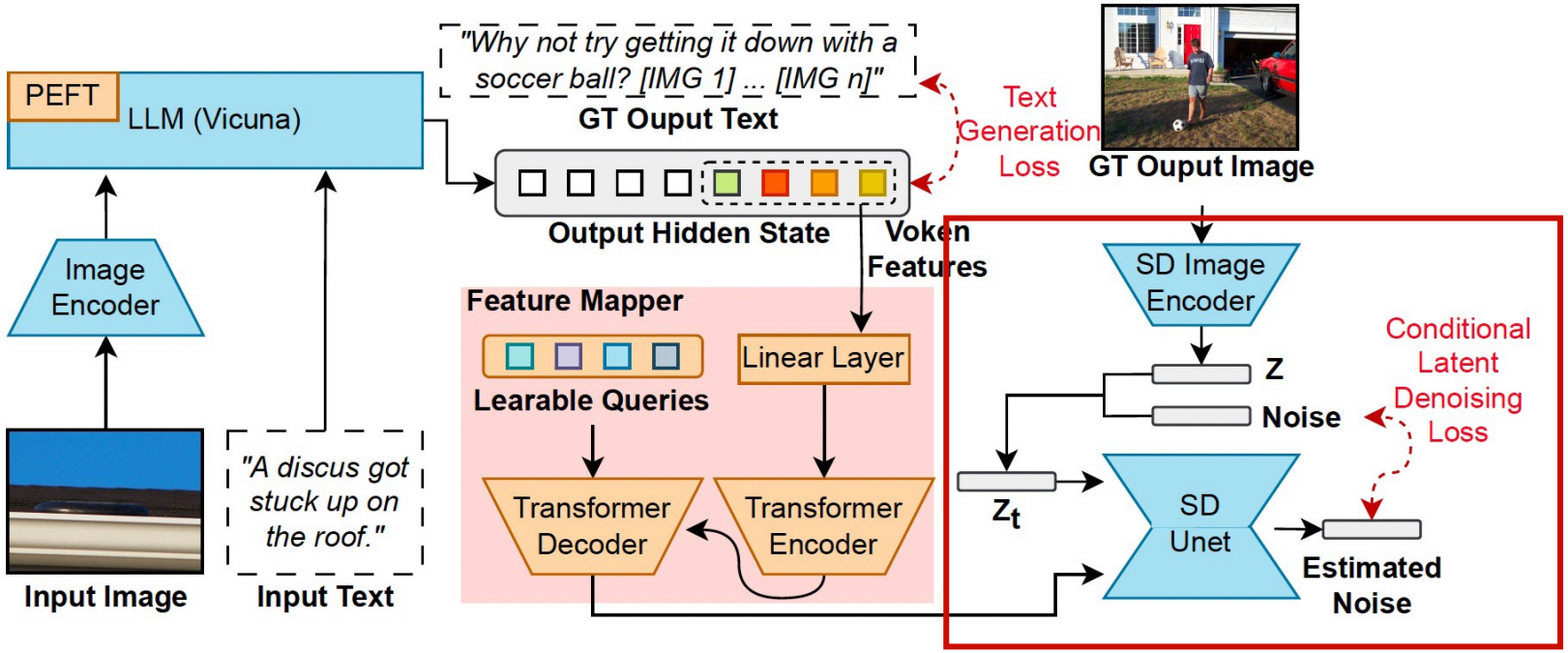
2. 基于查询的方法-MiniGPT-5

- 步骤二：**多模态表征提取**
  - Linear Layer：将输出voken表征映射到视觉生成空间（2层MLP）
  - Enc-Dec：4层transformer结构
  - Queries：保持关键信息的提取能力



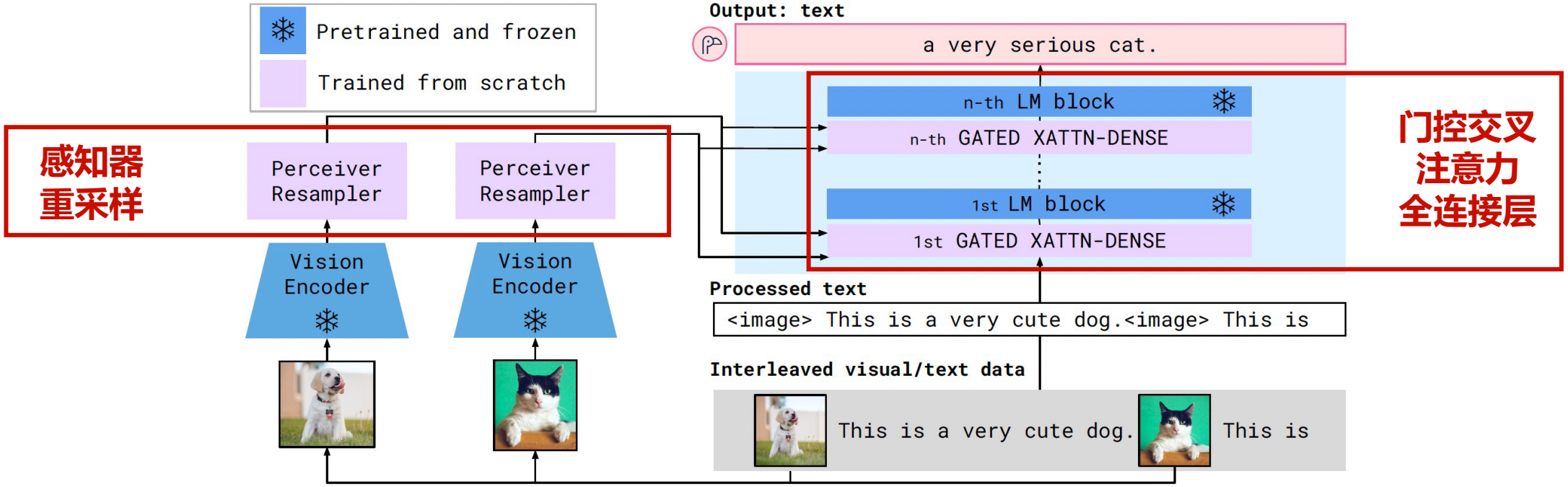
2. 基于查询的方法-MiniGPT-5

- 步骤三：**图像生成模型**
  - 利用Query输出特征，用于Stable Diffusion 的图像生成
  - Stable Diffusion：支持条件控制的分层图像生成模型



3. 基于融合的方法-Flamingo

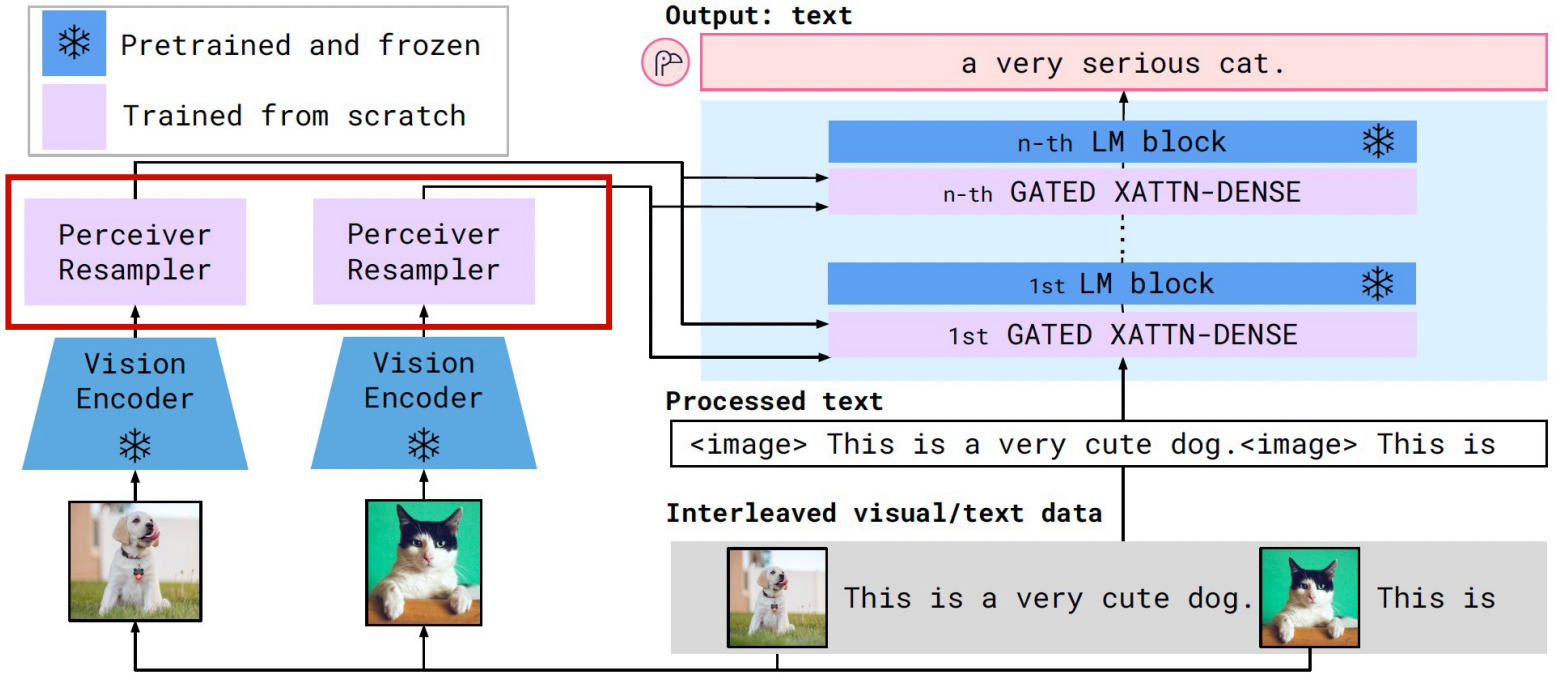
- 固定LLM和视觉编码器，训练门控网络（根据文本上下文选择视觉信息），使得LLM能处理各种多模态Few-shot任务
- 整体框架：① Perceiver Resampler；② Gated XATTN-DENSE



Flamingo: a visual language model for few-shot learning. NeurIPS 2022.

3. 基于融合的方法-Flamingo

- 模块一：**感知器重采样** Perceiver Resampler
  - 将图片以及不同尺寸的图片或视频，进行统一的表征建模，从而确保其具有统一维度的输出（64 tokens）

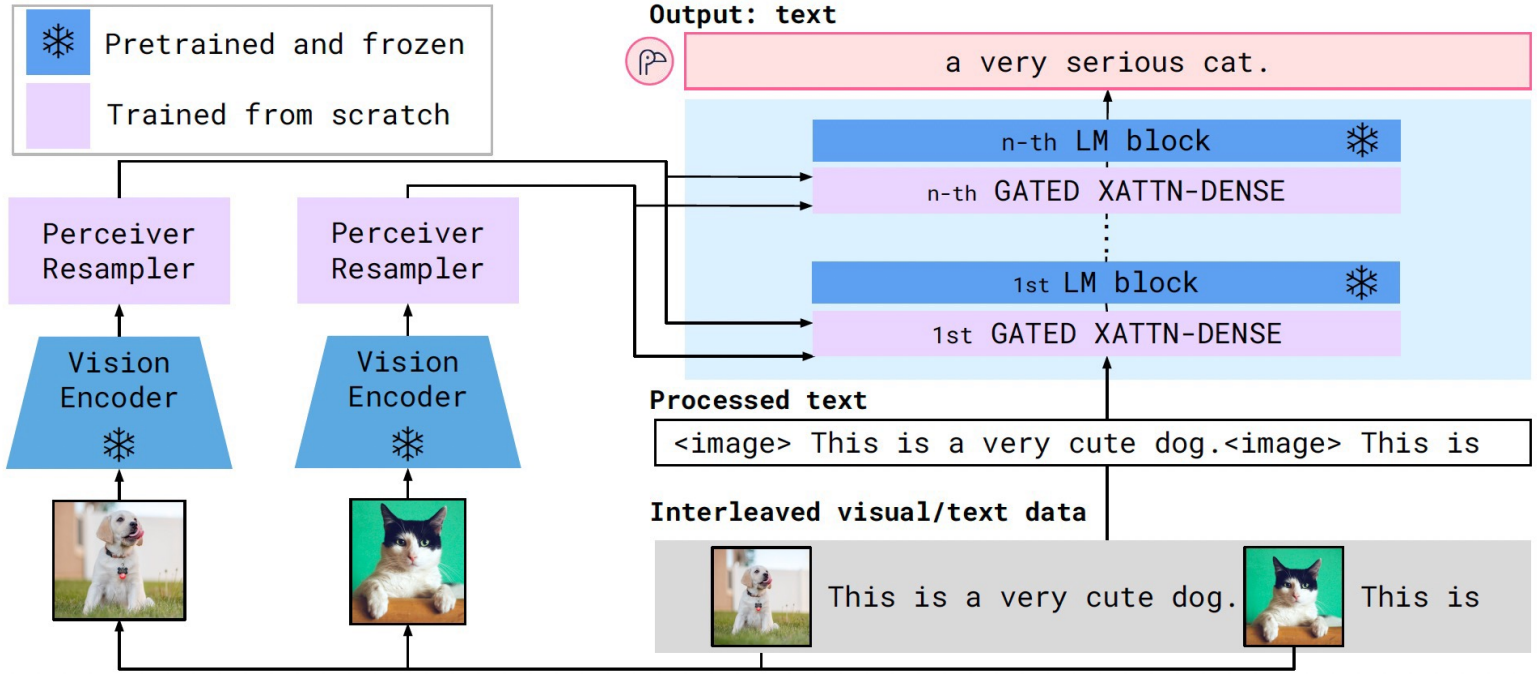


Flamingo: a visual language model for few-shot learning. NeurIPS 2022.

3. 基于融合的方法-Flamingo

❑模块二： **门控交叉注意力全连接层** Gated XATTN-DENSE

- 在不破坏LLM知识的情况下，无缝的将视觉信息嵌入”进来，有效利用语言模型强大的推理能力，帮助实现多模态推理



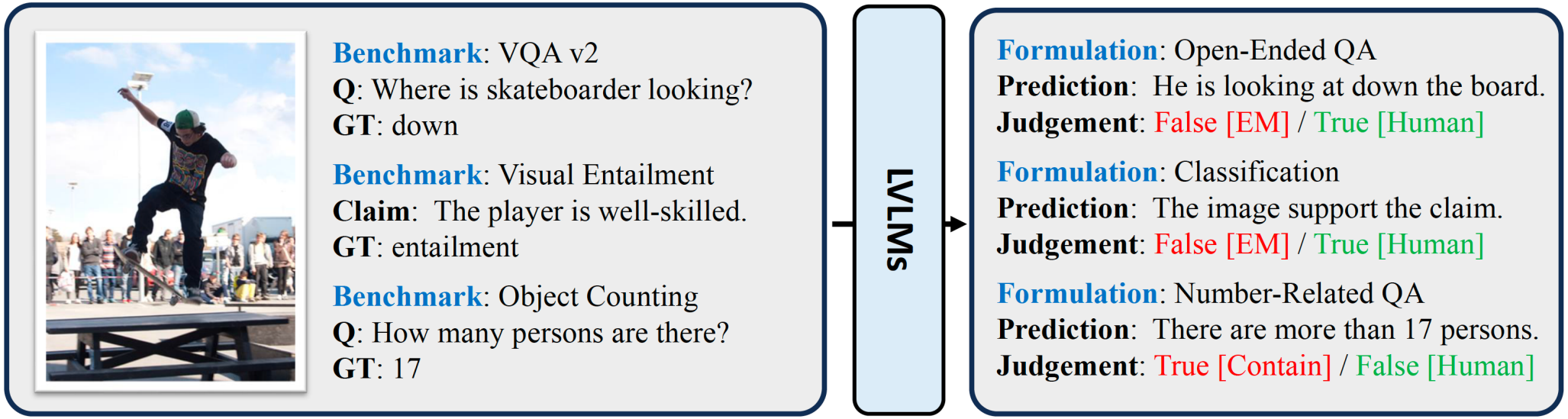
Flamingo: a visual language model for few-shot learning. NeurIPS 2022.

中国科学院大学
University of Chinese Academy of Sciences

目录

多模态大模型评测

- 多模态大模型展现出强大的综合能力，对应的评价体系更为复杂
- 核心难点：自动化判定模型开放式输出与标准标签是否语义等价



评价基准

❑ 多模态大模型评价基准主要包括四个方面：

评价框架的设定

明确需要测试的能力，如知识理解、推理、代码生成等

测试集合的构建

选择具有代表性的测试样例，覆盖不同任务和场景

输出结果的评价

结合自动指标和人工评测，从正确性、相关性等方面进行评价

模型输出的随机性

通过多次测试和结果统计，提高评测结果的可靠性

MME：一个系统性的评测基准

- 感知与认知维度，共涵盖 14 项子任务
- 所有任务指令均为人工设计
- 二元形式化：让模型回答是/否

MME：评价策略

❑ 让模型回答“yes”或“no”

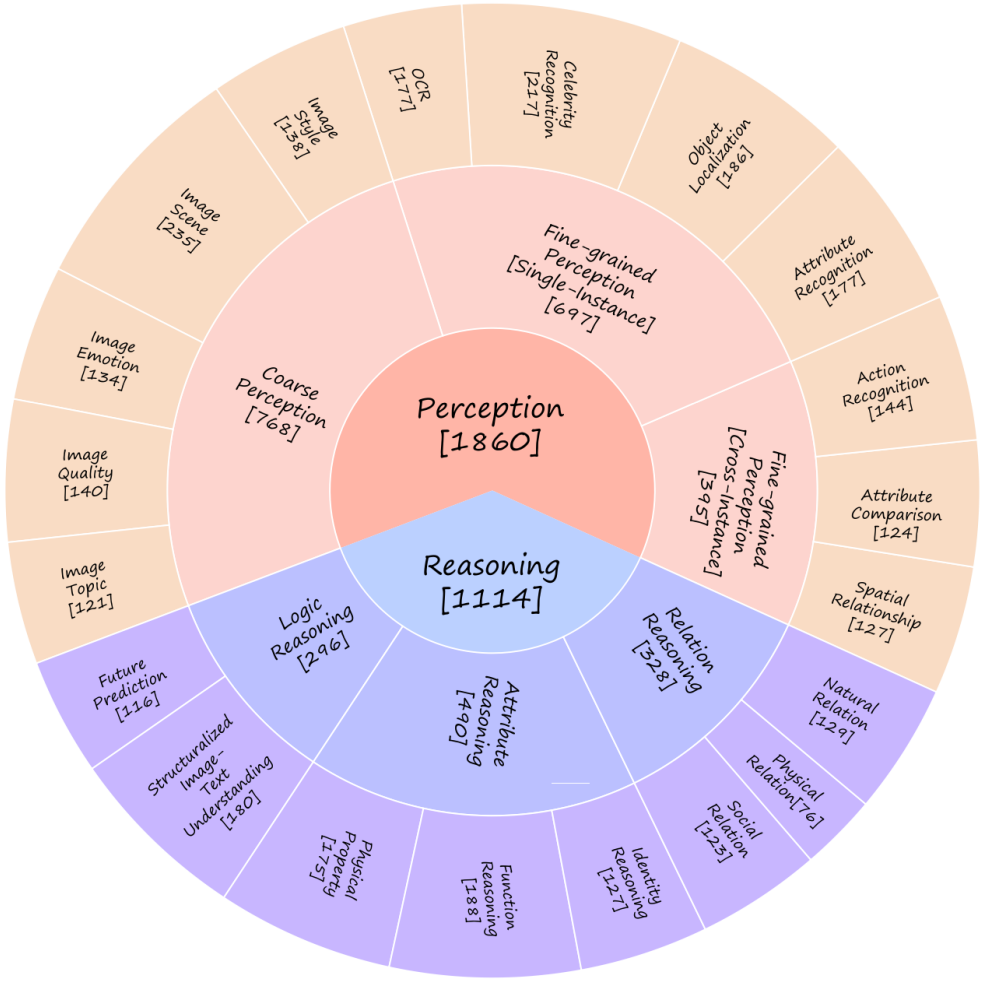
- 指令构成：简明问题+固定提示句“Please answer yes or no.”
- 稳定性测试：每张测试图片，人工设计两条不同指令，对应回答分别为“yes” 和“no”

❑ 评价指标

- “accuracy”是根据每个问题计算的；“accuracy+”是根据每张图片计算的，其中两个问题都需要被正确回答
- **感知分数**是所有感知子任务的分数总和；**认知分数**以相同的方式计算

MMBench：一个综合全面的评测基准

- 三层能力维度（L-1~L-3），总计 20 类子能力：
  - L-1：感知、推理基础维度
  - L-2 细分模块
    - 感知：粗粒度感知、细粒度单实例感知、细粒度跨实例感知
    - 认知：属性推理、关系推理、逻辑推理
  - L-3：基于 L-2 能力进一步拆分所得细分能力



MMBench：评价策略

❑ 循环评价策略

- 对同一问题多次输入 VLM，更换提示、调整答案顺序，校验模型在多轮不同输入条件下能否始终得到正确结果

❑ ChatGPT 辅助答案抽取

- 为了解决VLM自由形式输出的问题，借助 ChatGPT 完成答案抽取

The original VL problem:
Q: How many apples are there in the image?
A. 4; B. 3; C. 2; D. 1 GT: A

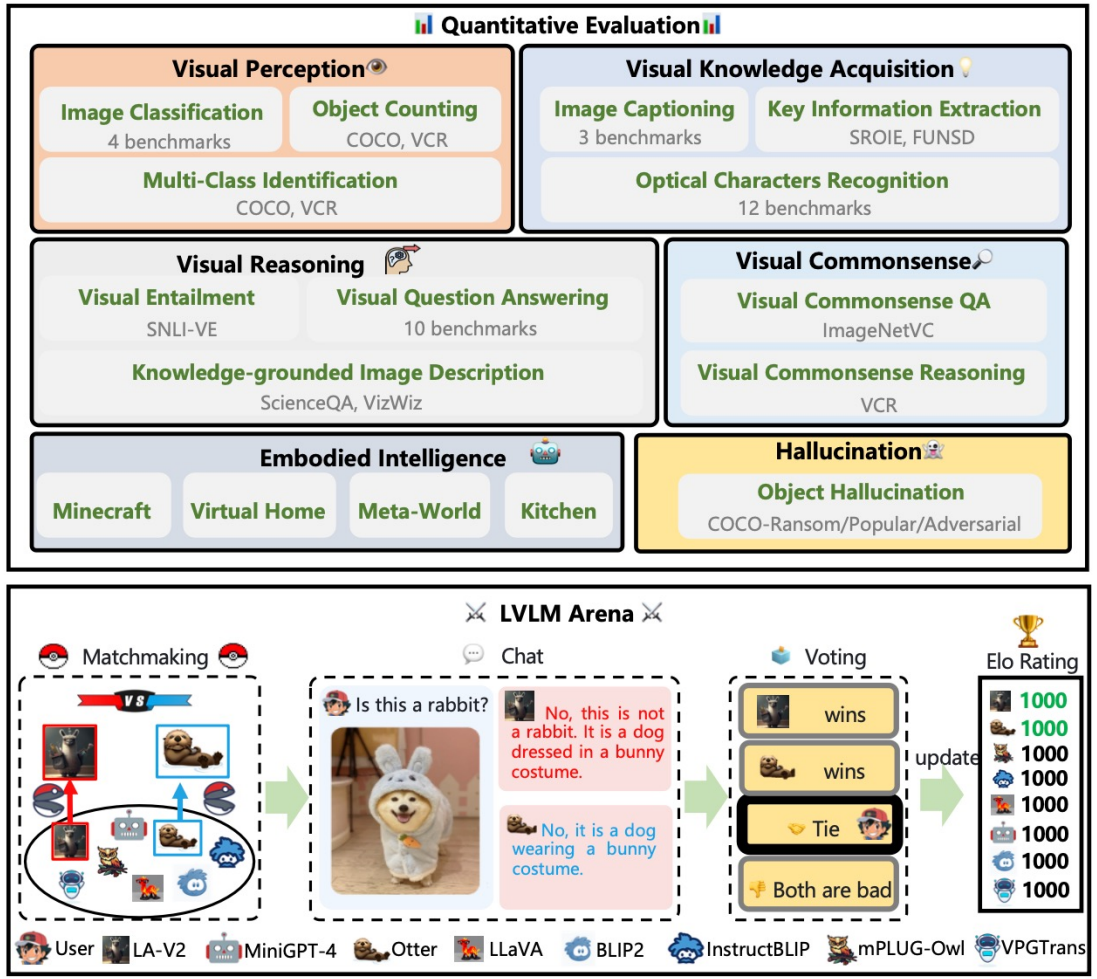
Circular Evaluation

4 Passes in Circular Evaluation (choices with circular shift):
1. Q: How many apples are there in the image? Choices: A. 4; B. 3; C. 2; D. 1. VLM prediction: A. GT: A ✓
2. Q: How many apples are there in the image? Choices: A. 3; B. 2; C. 1; D. 4. VLM prediction: D. GT: D ✓
3. Q: How many apples are there in the image? Choices: A. 2; B. 1; C. 4; D. 3. VLM prediction: B. GT: C ✗
4. Q: How many apples are there in the image? Choices: A. 1; B. 4; C. 3; D. 2. VLM prediction: B. GT: B ✓

VLM failed at pass 3. Thus wrong.

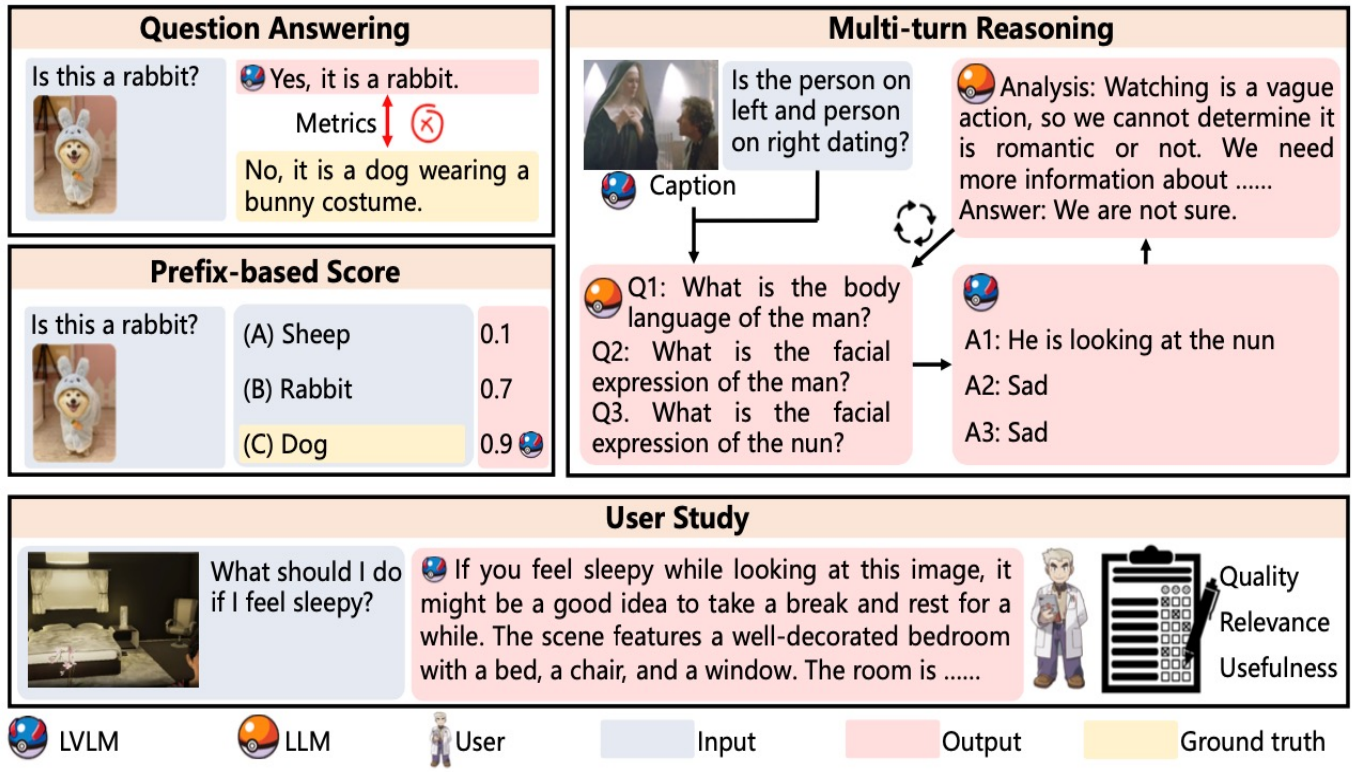
LVLM-eHub：模型评测擂台赛

- 定量评测覆盖六大核心能力维度
- 面向多样任务、数据集定制差异化评测方案
- 上线 LVLM Arena 在线测评平台，支持用户盲测对比：与双匿名模型对话后自主择优



LVLM-eHub：在线评价

- 从模型库中随机抽取一对模型
- 用户与匿名模型开展交互对话，完成体验后投票择优
- 整体流程包含三大核心模块：配对、对话交互、投票打分



目 录

多模态大模型展望

❑ 大规模、高质量的预训练数据

- ●构建大规模不同模态间的对齐数据(弱监督、半监督)
- ●引入知识来筛选大数据

❑ 高效计算的大模型网络结构

- ●改进或替代Transformer的高效模型
- ●超大规模模型分布式并行训练
- ●与下游任务兼容的更优模型
- ●显示知识嵌入与隐式知识学习

多模态大模型展望

❑适合多模态关联建模的自监督学习

- ●单模态、部分模态、全模态混合训练
- ●如何实现多模态信息之间更细粒度的对齐建模
- ●联合无监督强化学习，引入环境反馈

❑预训练模型的下游应用与迁移能力

- ● 模型压缩与推理加速为特定场景应用提供可能
- ● 多模态应用更为丰富，如何拓展更多创新下游应用

多模态大模型展望

❑ 当前大模型研究主要围绕“研究、利用和治理”三个方面展开

研究大模型

研究大模型的原理、能力来源、可解释可控性，探索大模型的能力边界

利用大模型

赋能各个研究任务和研究方向，例如赋能科学研究AI4Scicnce，赋能各行各业AI+

治理大模型

关注大模型安全，确保隐私安全，符合人类根本利益，防止其做出危害人类事情

本节复习

- 多模态预训练模型：VL-BERT、CLIP、BLIP
- 多模态大模型：LLaVA、BLIP2、MiniGPT5
- 多模态大模型评测：MME、MMBench、 LVLM-eHub

致 谢

- 胡玥、曹亚男、方芳：国科大《自然语言处理基础》
- 曹亚男、任昱冰：国科大《深度学习与自然语言处理概述》


THANKS
https://ictkc.github.io/teaching/2026spring-nlp