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）

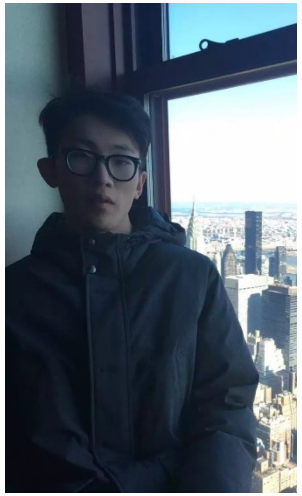
复旦大学
数据智能与社会计算实验室 (Fudan DISC)
自然语言处理组 (Fudan-NLP)

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



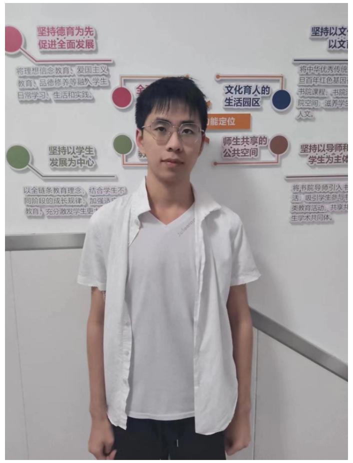
王晔

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

任务	匹配	生成	推理	导航
模块	跨模态语义表示		跨模态语义对齐	
语言	字	短语	句子	段落
视觉	像素	区域	图片	相册

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

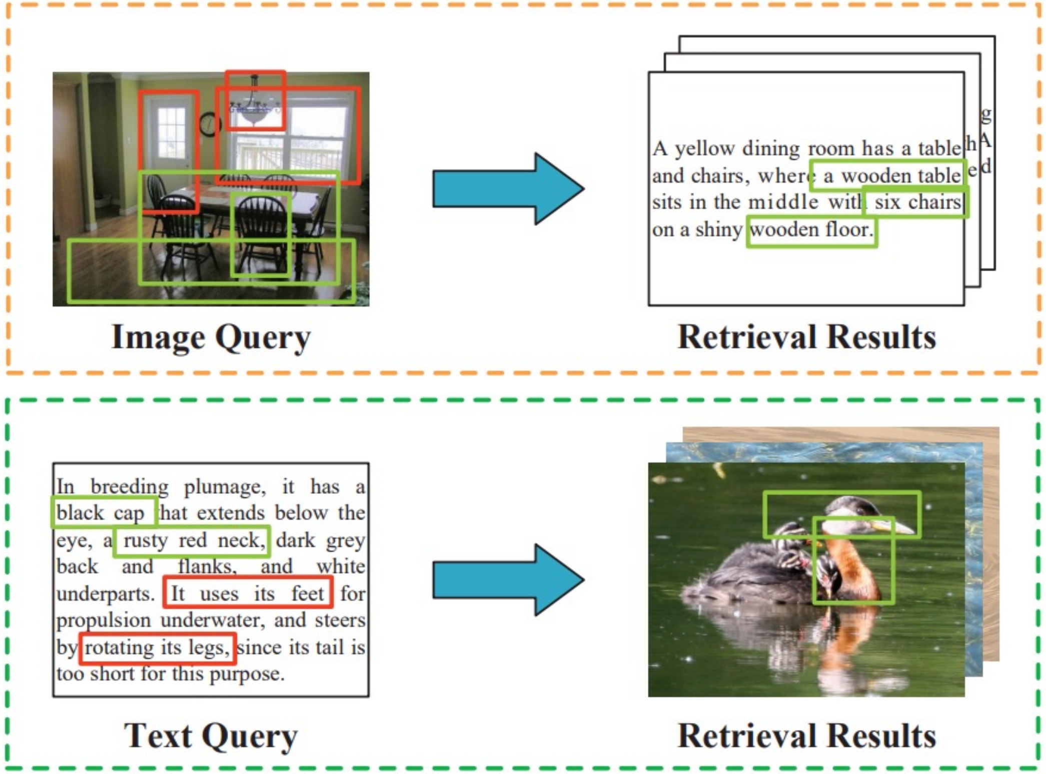
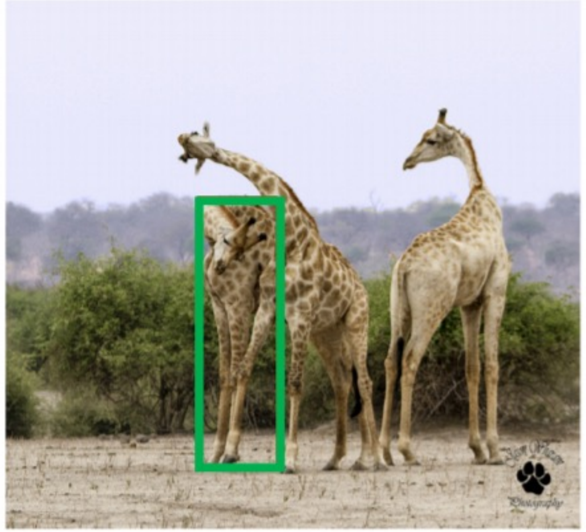


	Image-train	Image-dev	Image-test	caption
MSCOCO	113,287	5,000	5,000	5 for each image
Flickr30K	29,000	1,000	1,000	

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
- <u>评测指标</u>: BLUE, ROUGE, MEOTER, SPICE

The dog was ready to go. | He had a great time on the hike. | And was very happy to be in the field. | His mom was so proud of him. | It was a beautiful day for him.

A person riding a motorcycle on a dirt road.

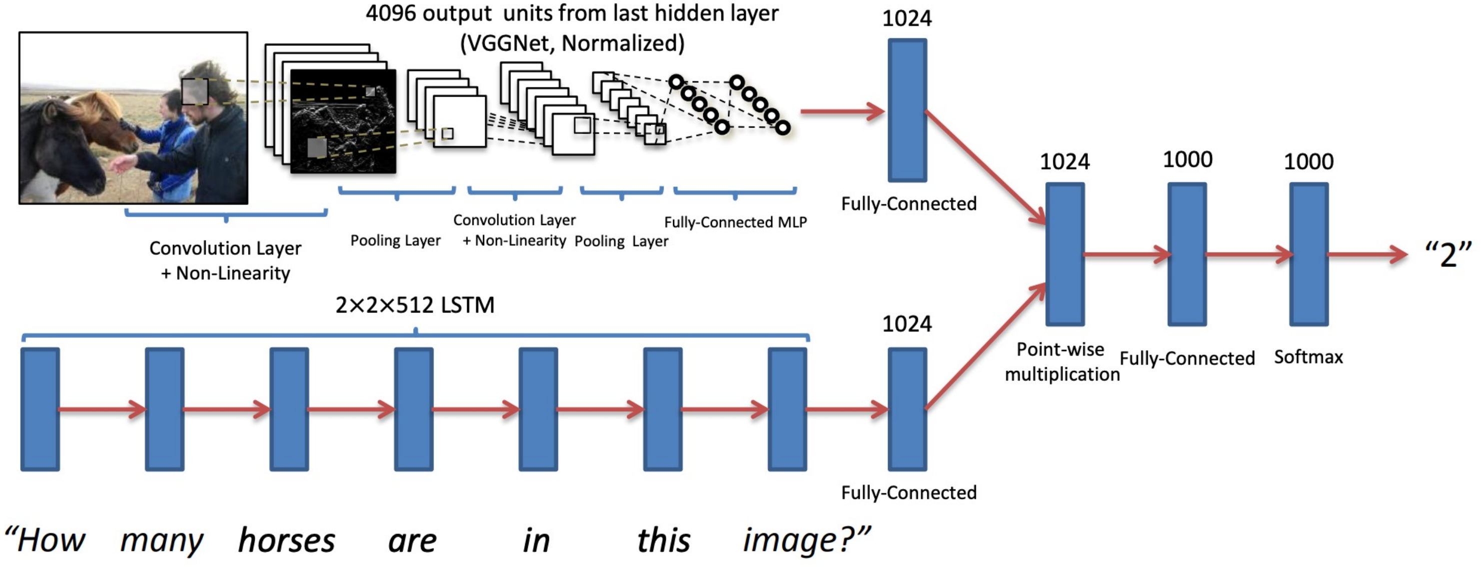
Two dogs play in the grass.

A group of young people playing a game of frisbee.

Two hockey players are fighting over the puck.

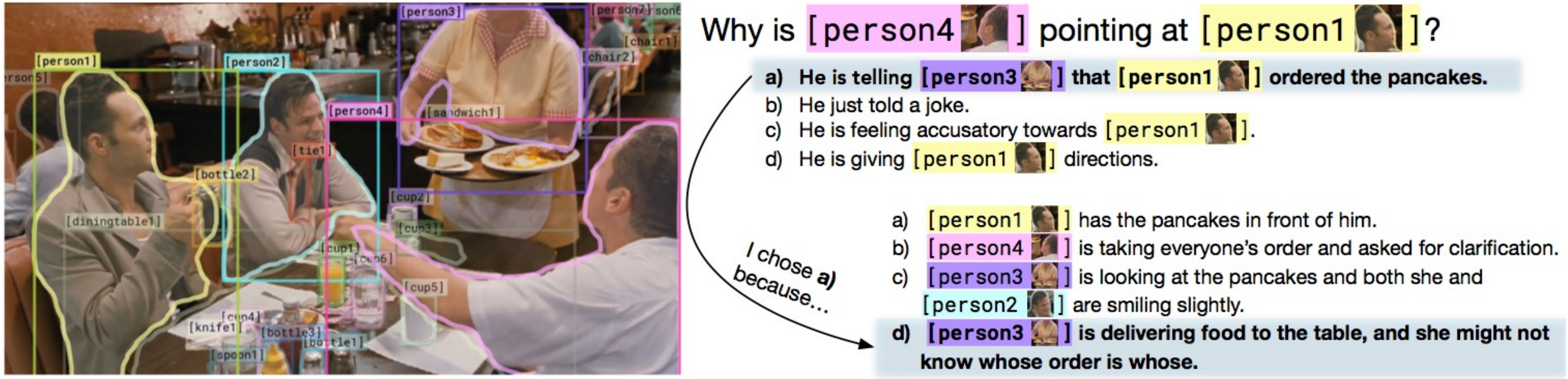
数据集	图片个数	描述个数
MSCOCO	300,000 +	5 per image
Flickr30K	30,000	158,000 in total
Flickr8K	8,000	5 per image
Visual Genome	108,000 +	1,445,322 in total
Instagram	~10,000	5 per image
FlikrStyle10K	10,000	Romantic, humorous, factual

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.

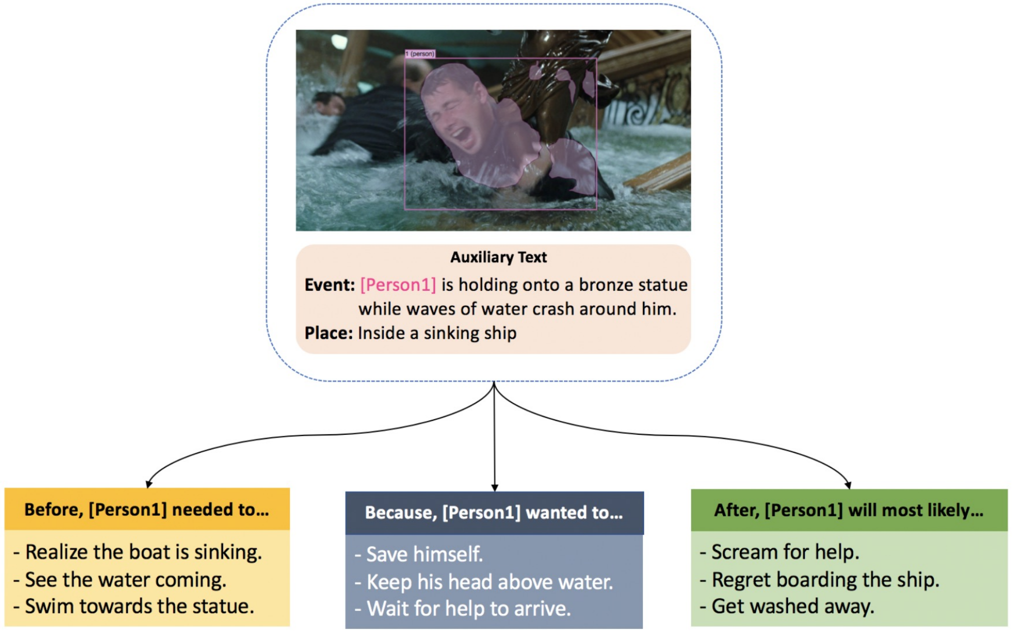
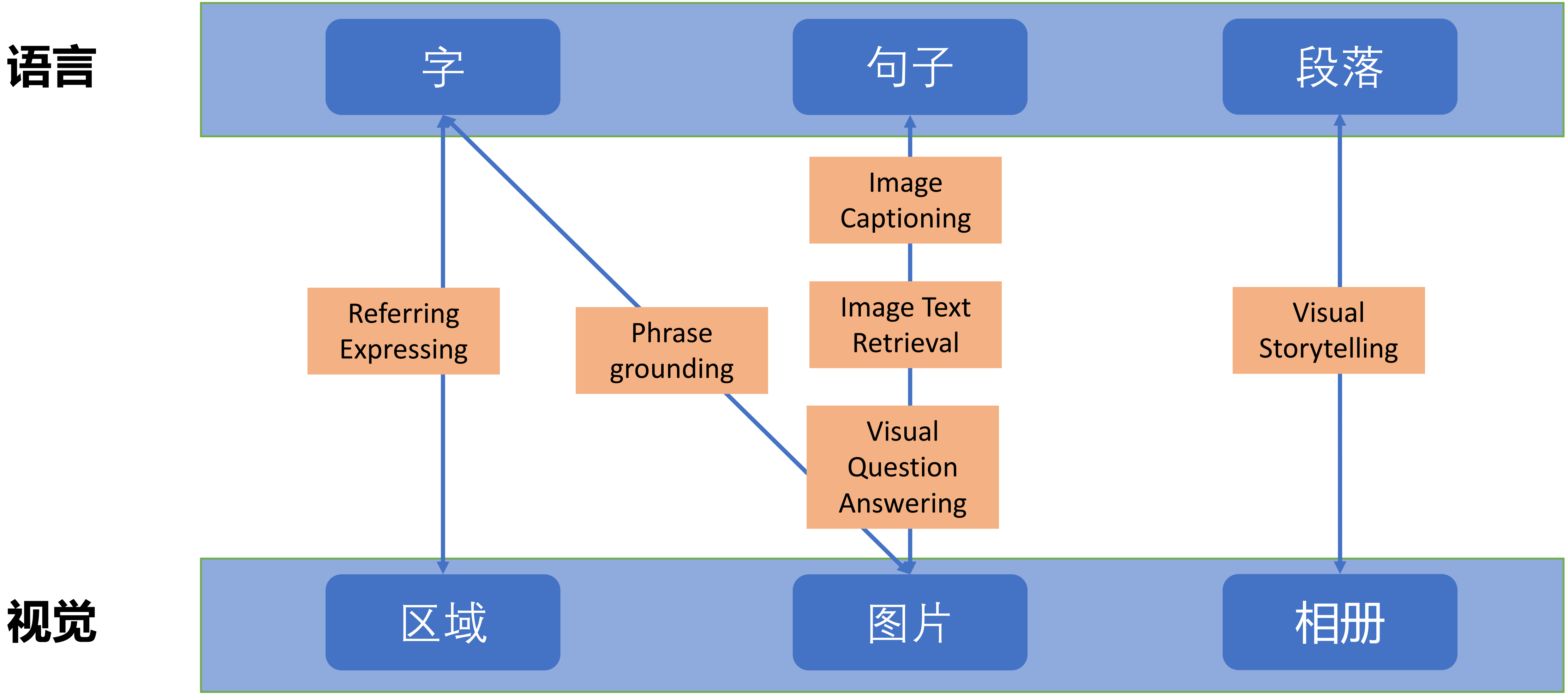


Fig. 2: **Task Overview:** Our proposed task is to generate commonsense inferences of **events before**, **events after** and **intents at present**, given an image, a description of an **event at present** in the image and a plausible scene / location of the image.

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

- <u>基本设定:</u> 以图片-文本对作为输入，联合学习语言和图片的语义表示。
- <u>输入表示:</u> 单字， 图片区域， 整体占位符（CLS）
- <u>跨模态交互学习:</u>
  - 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

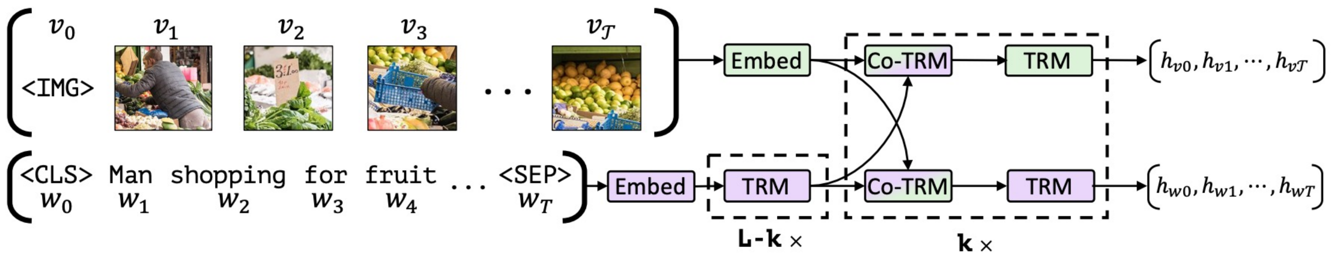


Figure 1: Our ViLBERT model consists of two parallel streams for visual (green) and linguistic (purple) processing that interact through novel co-attentional transformer layers. This structure allows for variable depths for each modality and enables sparse interaction through co-attention. Dashed boxes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

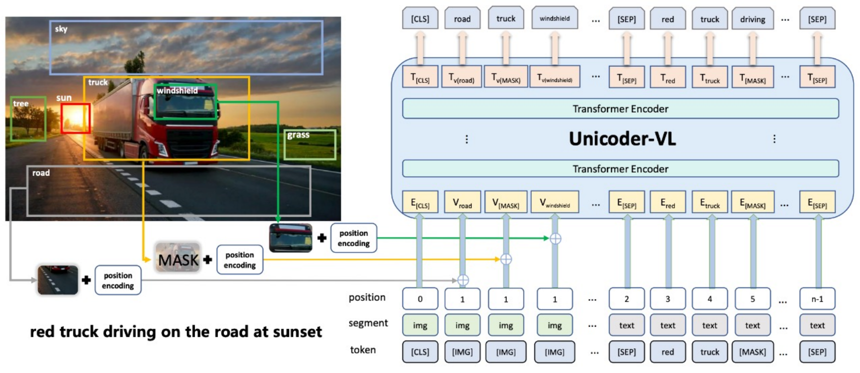


Figure 1: Illustration of Unicoder-VL in the context of an object and text masked token prediction, or *cloze*, task. Unicoder-VL contains multiple Transformer encoders which are used to learn visual and linguistic representation jointly.

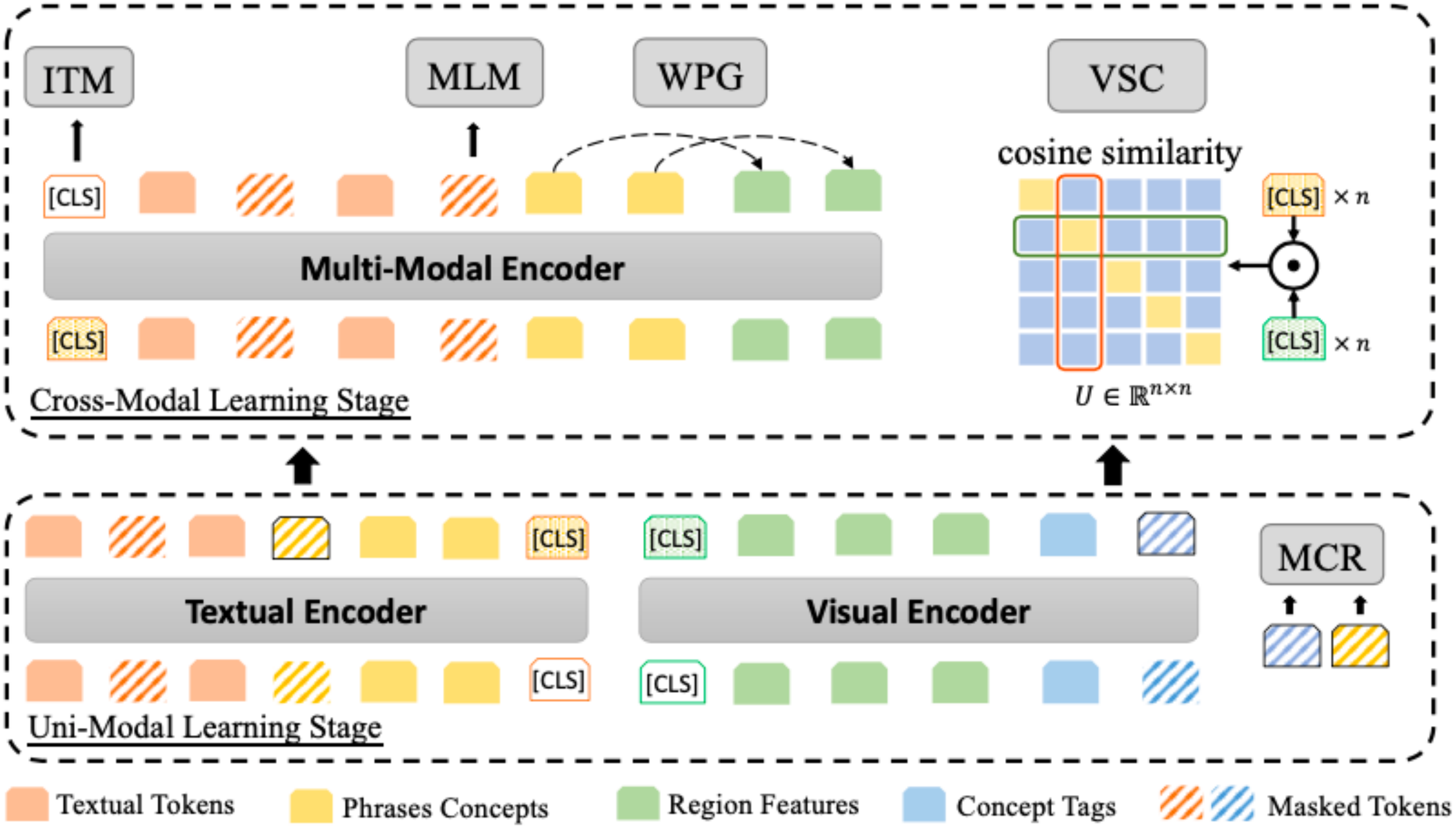
ViLBERT, NIPS 2019

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）

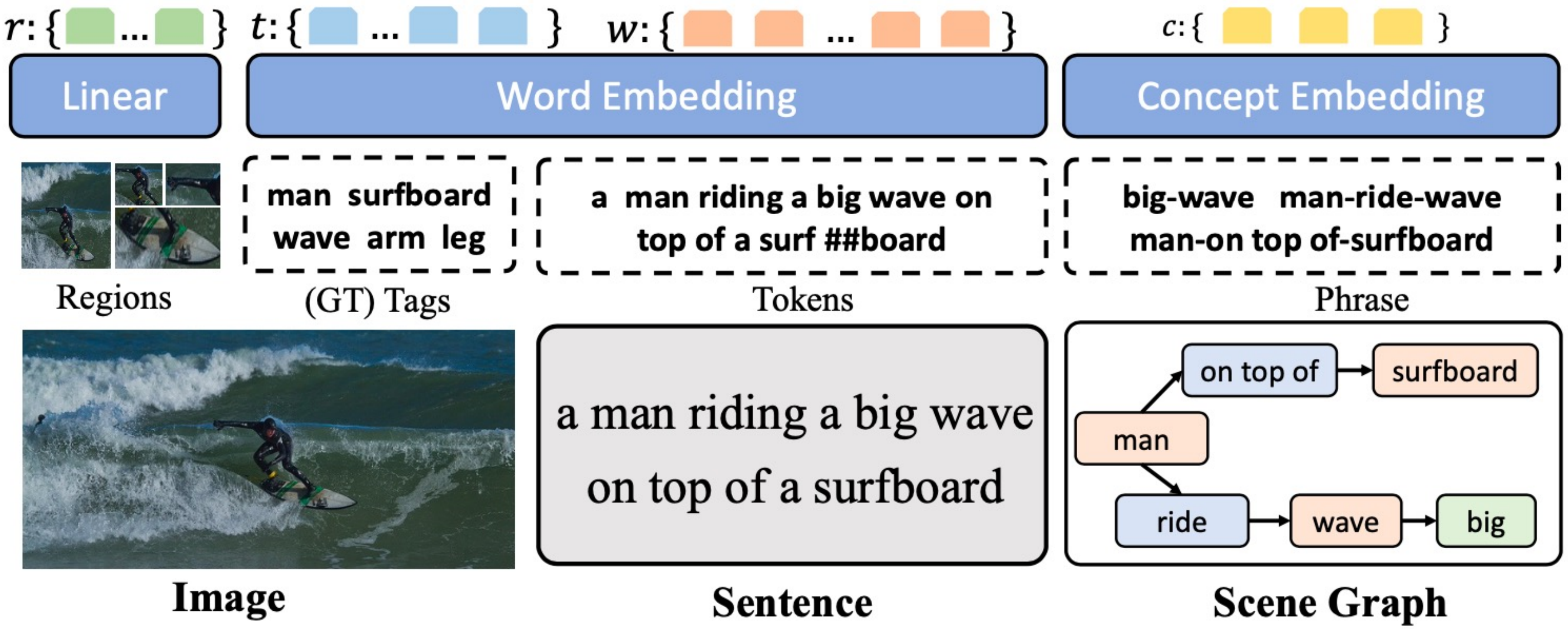
MVPTR：多层次语义关联的模型预训练



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

多层次语义输入

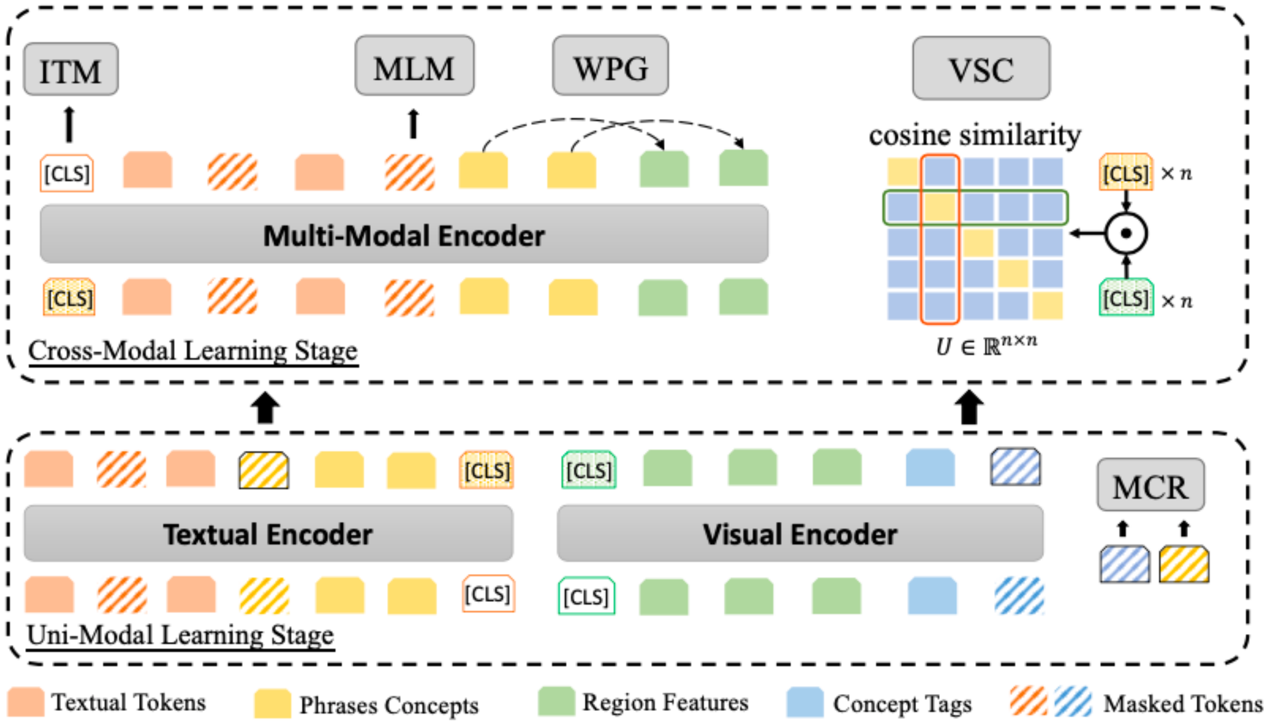
- 文本序列： 单字 + 短语概念 (场景图中的属性和关系)
- 视觉序列： 区域特征 + 实体标签



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

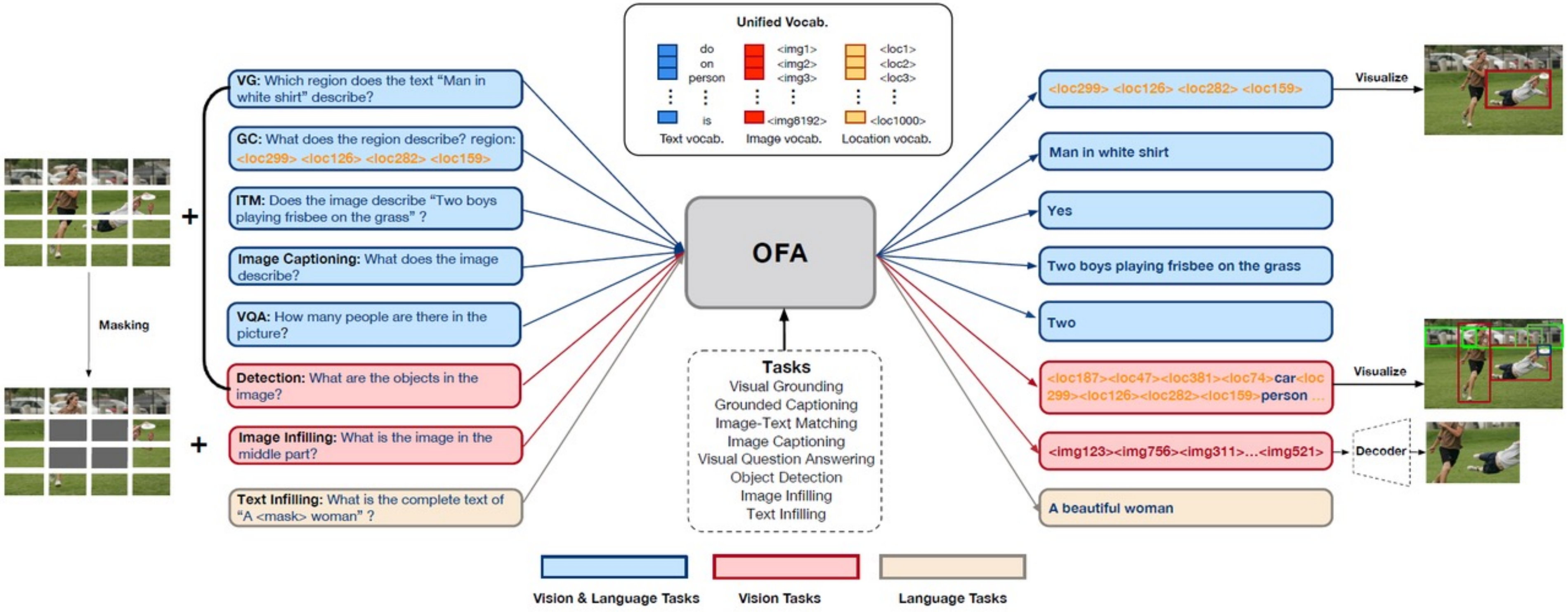


- 单模态阶段：
  - 高层次概念遮罩 (MCR)
- 跨模态阶段：
  - 全局的粗匹配：visual-semantic contrastive learning (VSC)
  - 细粒度匹配：weakly-supervised phrase grounding (WPG)
  - 细粒度图文匹配 (ITM) 和跨模态推理 (MLM)

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)

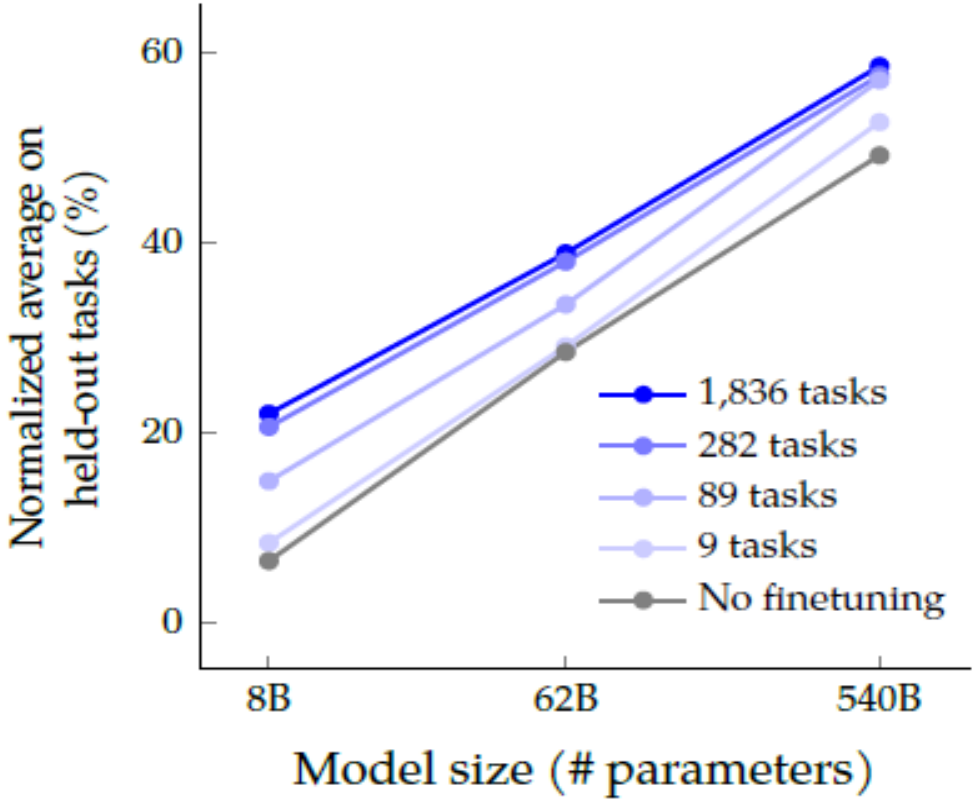
目录

- ChatGPT之前的视觉语言预训练
- **大视觉语言模型的架构和训练**
- 大视觉语言模型的评测
- 大视觉语言模型的能力扩充
- 大语言模型支撑的具身智能（视觉导航）

2022年底大语言模型横空出世

- 大语言模型背后的因素
  - **生成式预训练**
  - **指令微调**
  - 基于人类反馈的强化学习

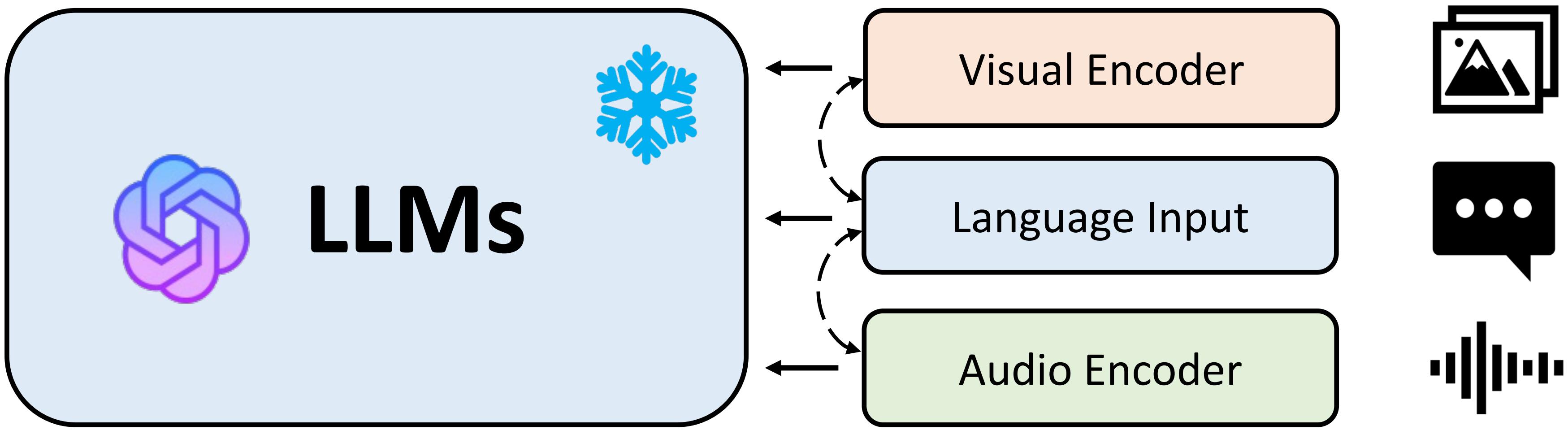
- 从任务特定的微调到指令微调
  - 语言是一个自然的交互方式
  - 使语言模型与人类偏好一致
  - 强大的泛化能力



- 对于视觉-语言等跨模态设定有什么启发呢？

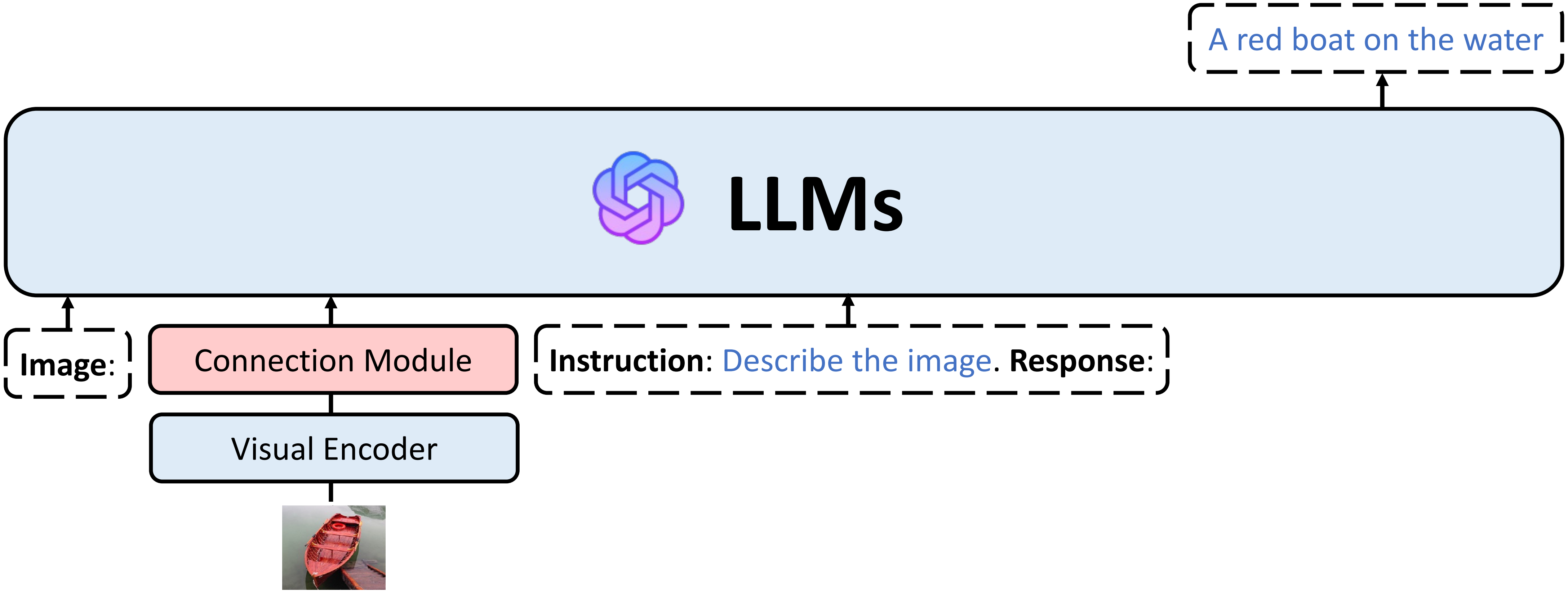
大语言模型如何帮助多模态模型构建?

- 大语言模型可以充当大脑，处理各种模态的信息
- 将其它模态信息对齐到大语言模型的语义空间



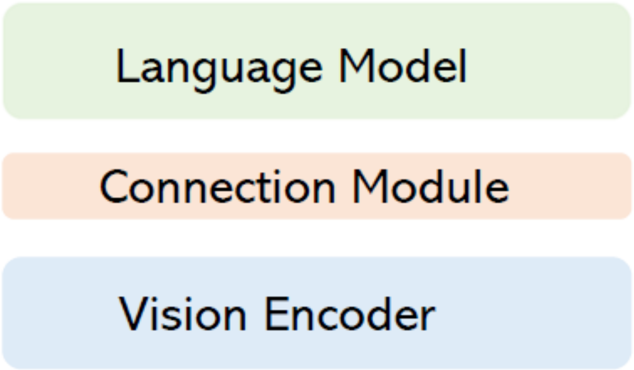
大规模视觉语言模型（Large Vision Language Model)

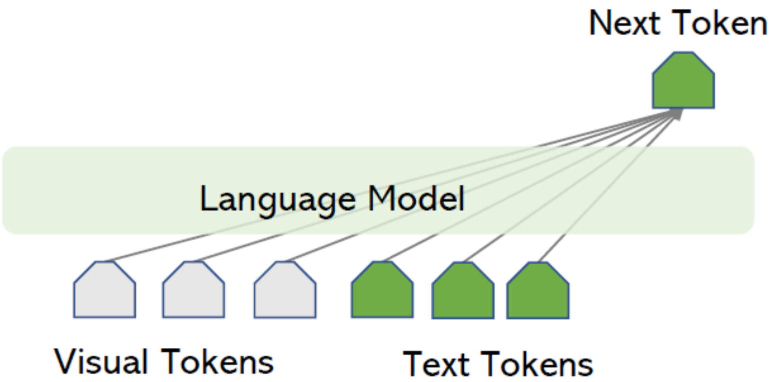
- 大视觉语言模型的通用解决方案 (开源)
  - 使用大语言模型（LLMs）作为骨干 + 视觉编码器
  - 通过多模态数据进行生成式预训练 + 指令微调



大视觉语言模型的训练　步骤一：预训练

- 让视觉表征对齐到大语言模型的语义空间
  - 视觉表征：　使用预训练的视觉编码器 (e.g., CLIP)
  - 大语言模型：使用现有的大语言模型
  - 连接方式：　线性、适配器、感知器、Q-former
- 使用图片-文本对进行语义对齐，如，图片描述生成任务
- 通过自回归语言模型进行训练，最大化生成目标的似然概率





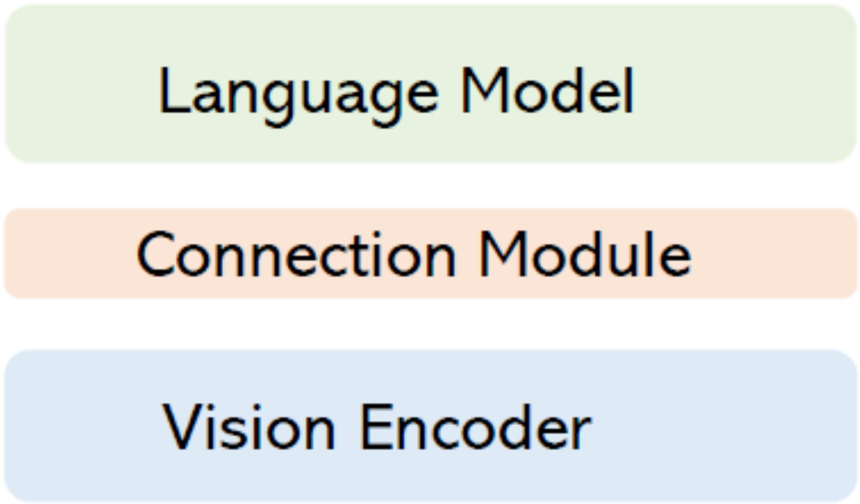
大视觉语言模型的训练 步骤二：指令微调

- 指令数据集构建
  - 基于现有有标注数据集合
  - 由ChatGPT / GPT-4辅助生成指令样本
- 数据形式
  - 仅文本的指令数据集
  - 图文对的指令数据集
- 指令微调
  - Loss：在回复的部分应用文本生成损失

Response type 1: conversation
Question: What type of vehicle is featured in the image?
Answer: The image features a black sport utility vehicle (SUV).
Question: Where is the vehicle parked?
Answer: The vehicle is parked in an underground parking area, likely in a public garage.
Question: What are the people in the image doing?
Answer: The people are trying to fit all of their luggage into the SUV, likely preparing for a trip.

大视觉语言模型的主流架构

- 视觉编码器
  - ViT-L/14， ViT-G/14， ImageBind
- 大语言模型
  - FlanT5， LLaMA， Vicuna， LLaMA-2 Chat
- 连接模块
  - 线性层: LLaVA, PandaGPT, Shikra
  - 适配器: LLaMA-Adapter V2, ImageBind-LLM
  - Q-Former: BLIP-2, InstructBLIP, MiniGPT-4, Cheetor, BLIVA
  - 感知器: Lynx, Multimodal-GPT, mPLUG-Owl



LLaVA：基于线性层的连接模块

- **视觉编码器：** ViT-L/14
- **大语言模型：** Vicuna, LLaMA, L
- **连接模块：** Projection **W**

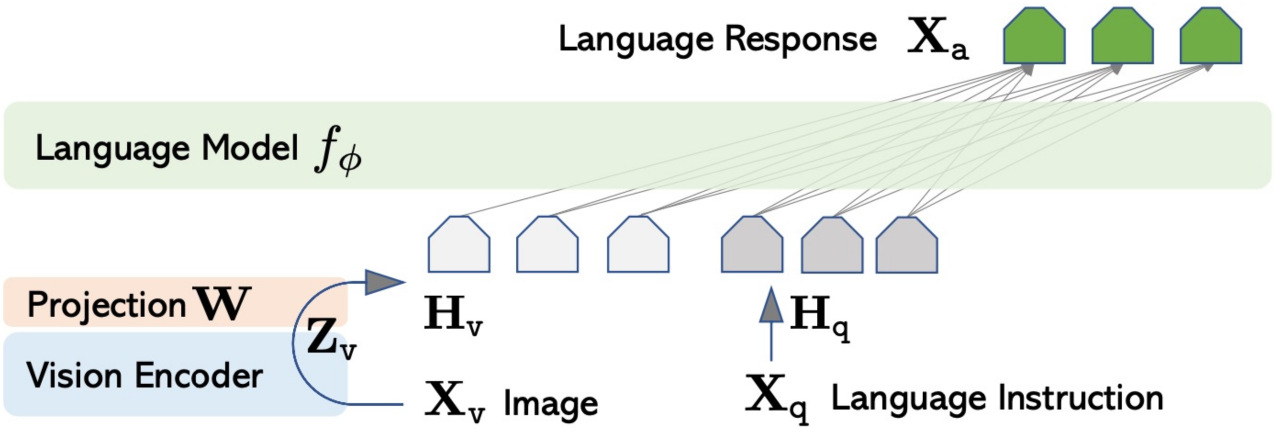
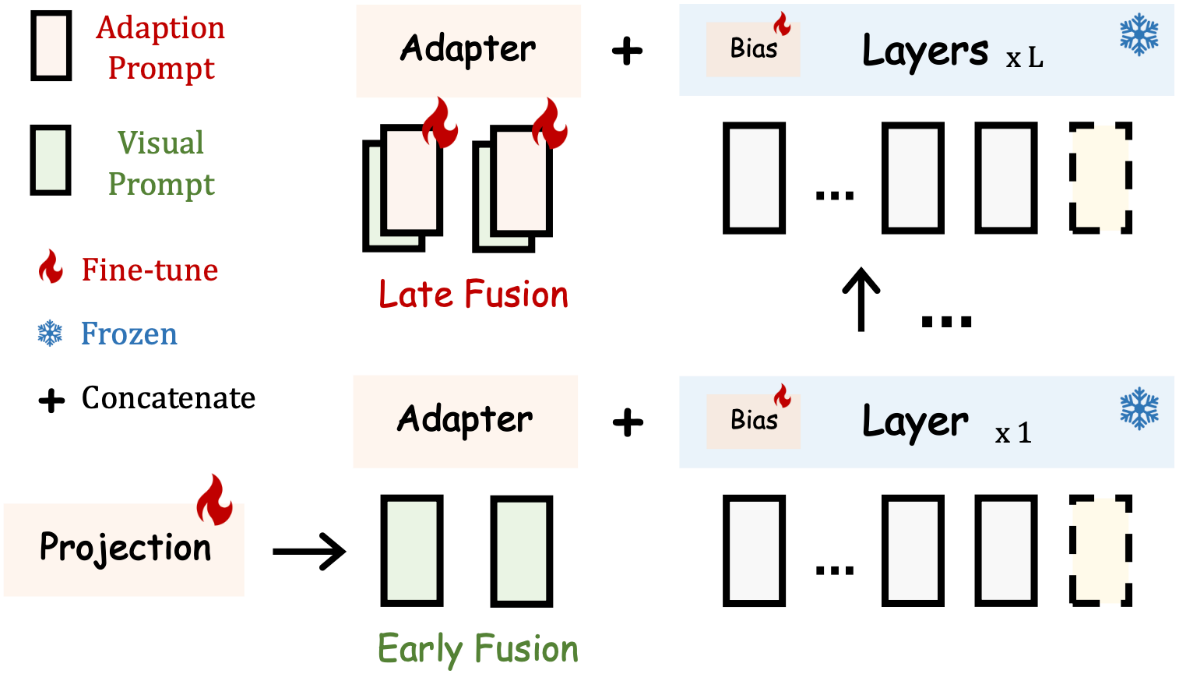


Figure 1: LLaVA network architecture.

- 预训练
  - 冻结 **视觉编码器** 和 **大语言模型** 的权重，并且最大化生成目标的似然概率（595K 文本编辑对 CC3M）
- 指令微调 （158K Multi-Intruct 微调数据）
  - 冻结 **视觉编码器**权重
  - 更新 **线性层**和**大语言模型**

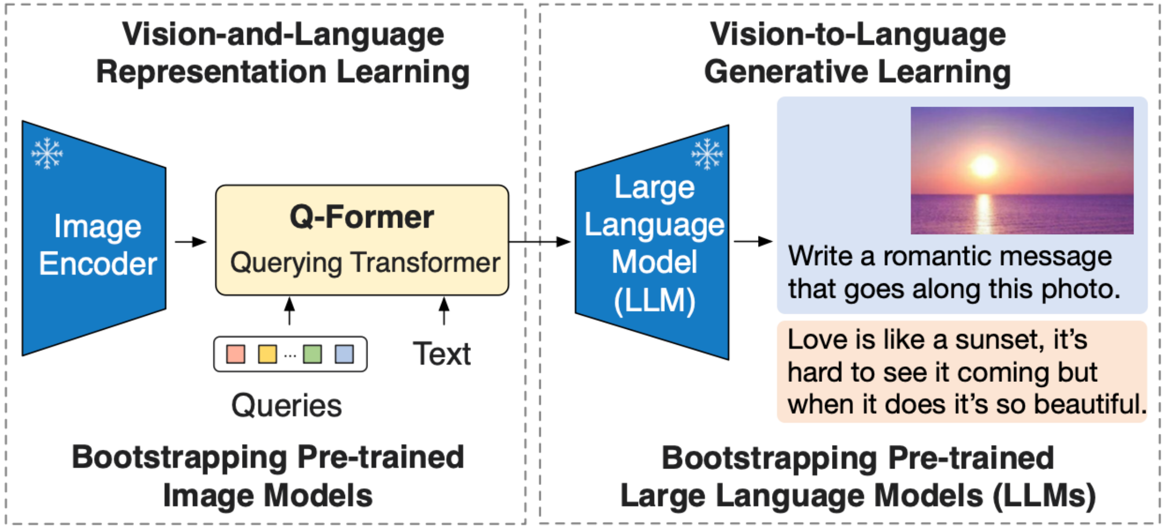
LLaMA Adapter V2: 基于适配器的连接模块

- **视觉编码器：** ViT-L/14
- **大语言模型：** LLaMA
- **连接模块：** Linear, Adapter, Gate
- 在预训练阶段更新连接模块的参数，没有指令微调



BLIP-2：基于Q-Former的连接模块

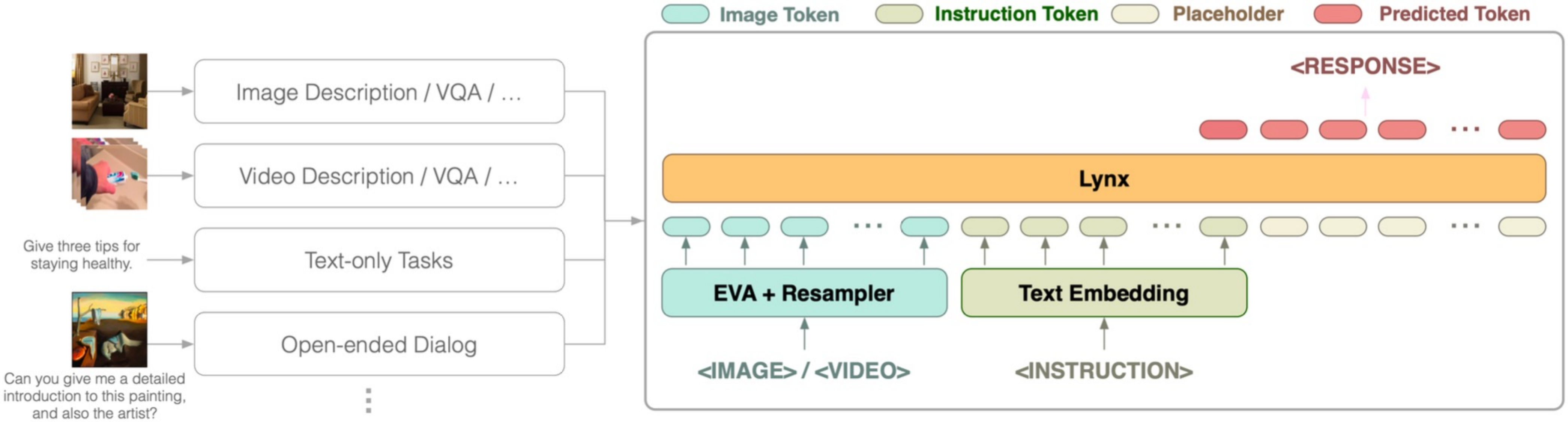
- **视觉编码器：** ViT-L/14
- **大语言模型：** Flan-T5, Vicuna
- **连接模块：** Q-Former



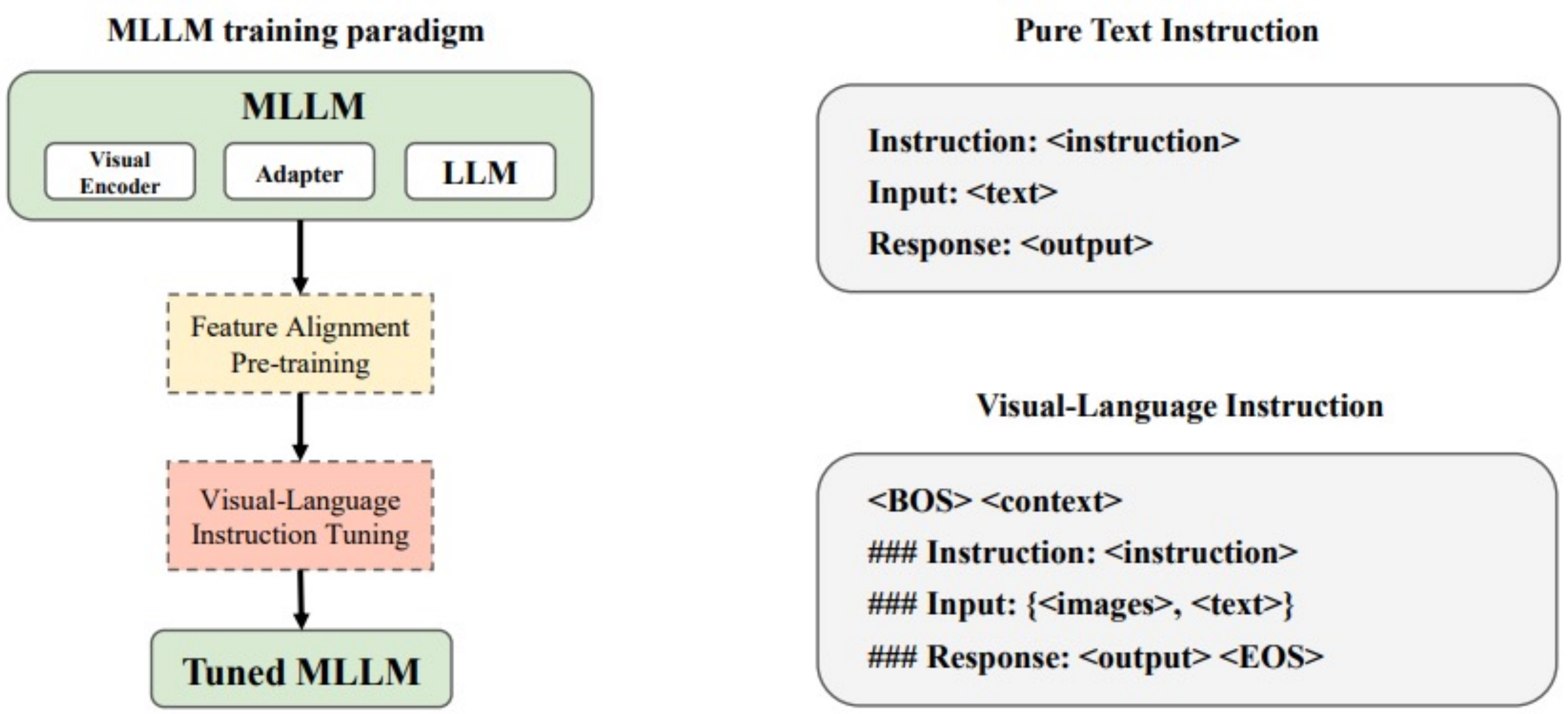
- 生成式训练阶段
  - 将 **Q-Former** (单流的多模态编码器) 和一个 冻结的大语言模型 连接起来
- **视觉编码器** 和 **大语言模型** 冻结， 只有一个轻量级的**Q-Former**被**训练**， 用于弥合模态间差距

Lynx：基于感知器的连接模块

- **视觉编码器：** ViT-L/14
- **大语言模型：** LLaMA, LLM Adapter (add new layers)
- **连接模块：** Perceiver
- 预训练： 训练大语言模型适配器和感知器
- 指令微调： **包含仅文本、图片文本对和视频文本对**



指令微调方法概览



基于现有的标注: **MULTIINSTRUCT, InstructBLIP**,**MiniGPT-4,** KOSMOS2.M3IT, etc

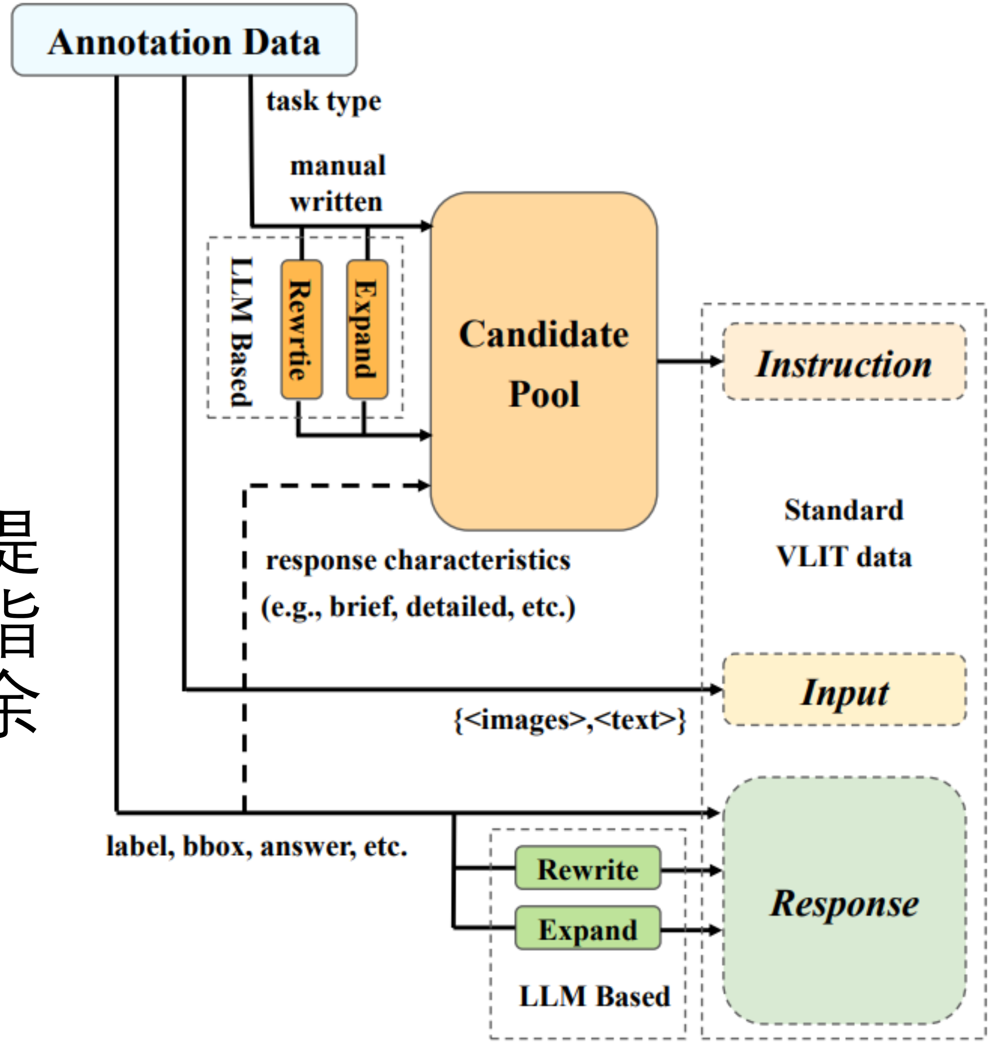
自生成指令微调: **LLaVA,Syphus**,**LVIS-INSTRUCT4V,**LMEye,LAMM,MosIT,etc

模态数据混合: Mplug-Owl, PandaGPT, LAVIN, BLIVA,etc

Vision-Language Instruction Tuning: A Review and Analysis, arXiv:2311.08172, 2023

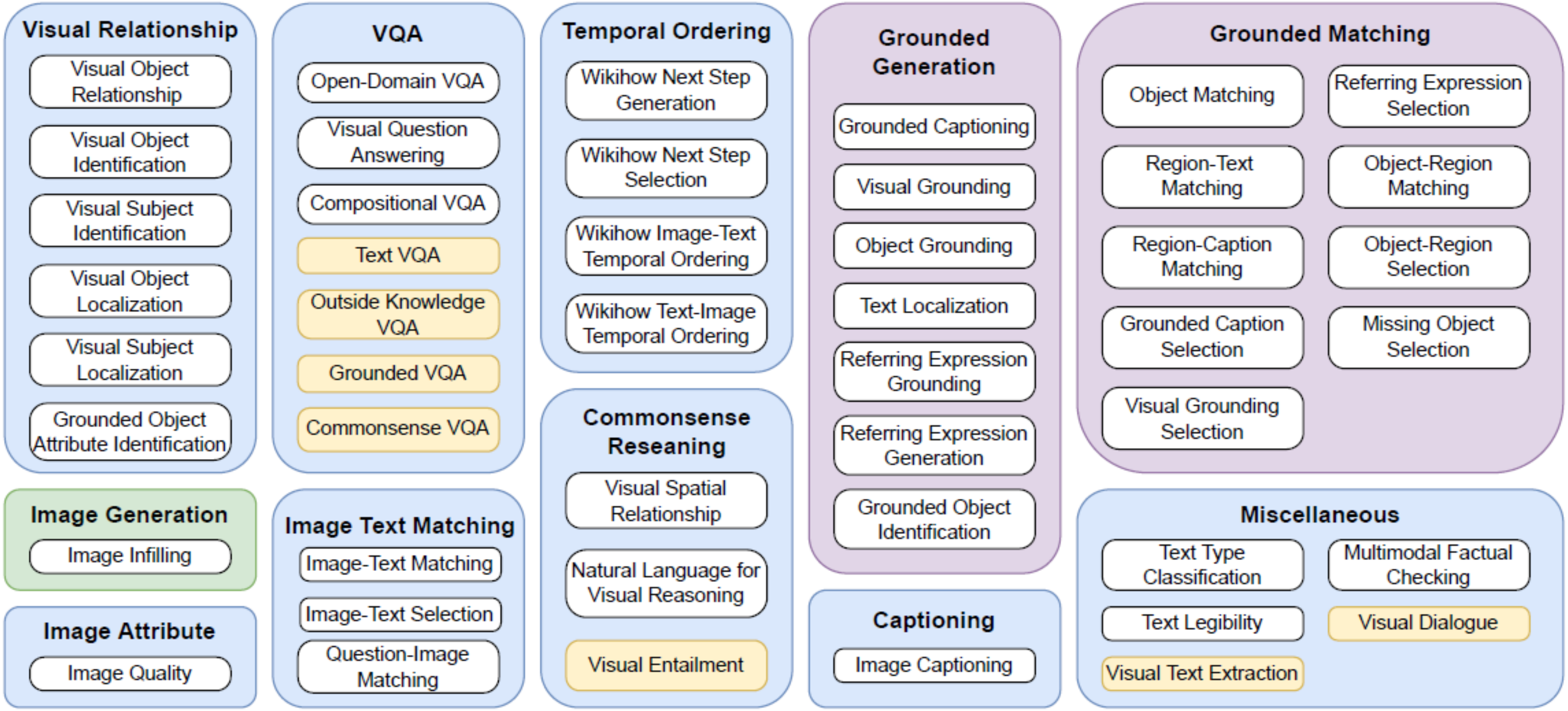
基于现有标注的指令构建

- **输入**：从原始的标注数据中获取
- **回复**：从原始数据标注中提取
- **指令**：从原始标注中通过人工筛选和提取完成构造，或根据人工构造的几种指令模板种子，由语言模型辅助生成剩余数据的指令



Annotation Adaption

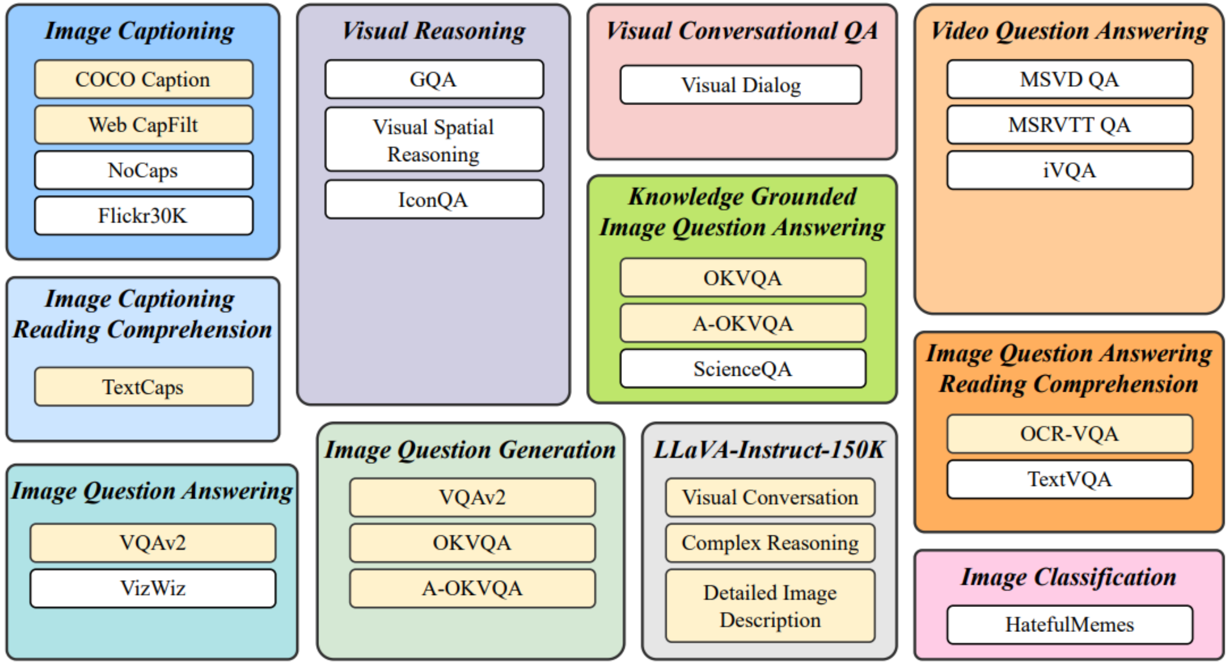
Multi-Instruct：多模态指令微调数据集



MULTIINSTRUCT: Improving Multi-Modal Zero-Shot Learning via Instruction Tuning, arxiv 2212.10773

InstructBlip: 基于现有标注数据的典型示例

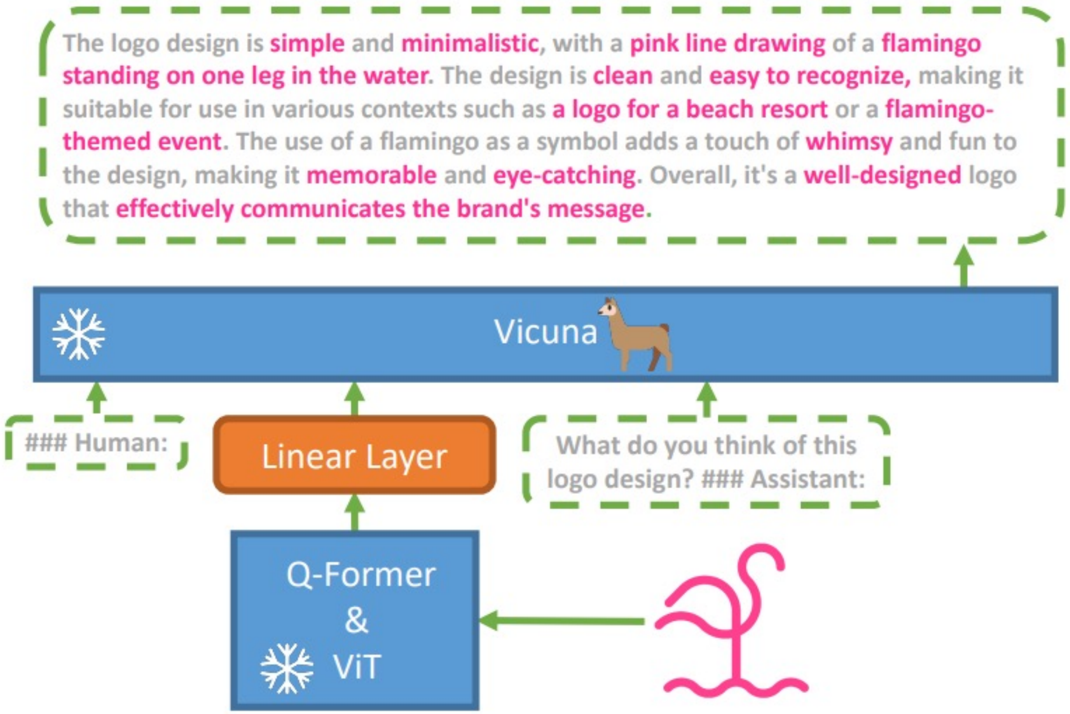
- 数据构造：收集CV领域11中不同任务的26个公开数据集
- 对于每个任务，**人为设计**了10-15个自然语言指令模板
- 该数据集用于**InstructBlip2**模型的指令微调训练



InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning, arXiv:2305.06500

MiniGPT-4：基于现有标注数据的典型示例

- 使用第一阶段预训练得到的模型来生成图像的初步描述
- 调用ChatGPT 优化描述
- 手动验证图像-描述的一致性

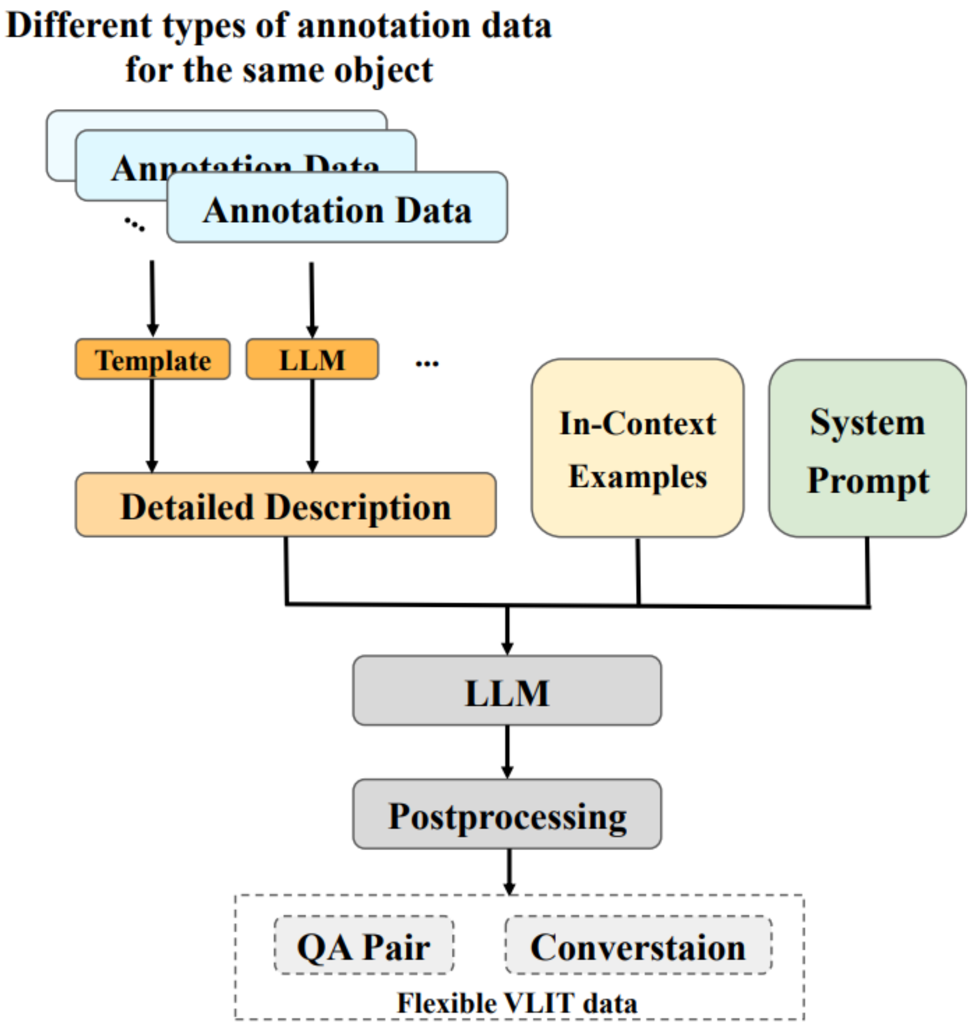


- format：
- ###Human: Describe this image in detail. Give as many details as possible. Say everything you see. ###Assistant:

MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models, arXiv:2304.10592

自生成指令微调（Self-Instruct）

- 使用小样本调用大语言模型来生成指令
- 首先利用语言模型，将原始数据的标注信息转换为对图像的详细描述文本
- 将图像描述、示例模板、输出格式的提示，作为语言模型的输入，最终从输出信息中获取到问答对或多轮对话等其他格式的数据



Self-Instruct

LLaVA：自生成指令微调数据

- 将COCO图像数据集的标注信息转换为两种描述图像的文本格式，一种描述图像的内容，一种描述图像中各个物体的方位信息

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], person: [0.63, 0.222, 0.686, 0.516], person: [0.444, 0.233, 0.487, 0.34], backpack: [0.384, 0.696, 0.485, 0.914], backpack: [0.755, 0.413, 0.846, 0.692], suitcase: [0.758, 0.413, 0.845, 0.69], suitcase: [0.1, 0.497, 0.173, 0.579], bicycle: [0.282, 0.363, 0.327, 0.442], car: [0.786, 0.25, 0.848, 0.322], car: [0.783, 0.27, 0.827, 0.335], car: [0.86, 0.254, 0.891, 0.3], car: [0.261, 0.101, 0.787, 0.626]

LLaVA：自生成指令微调数据

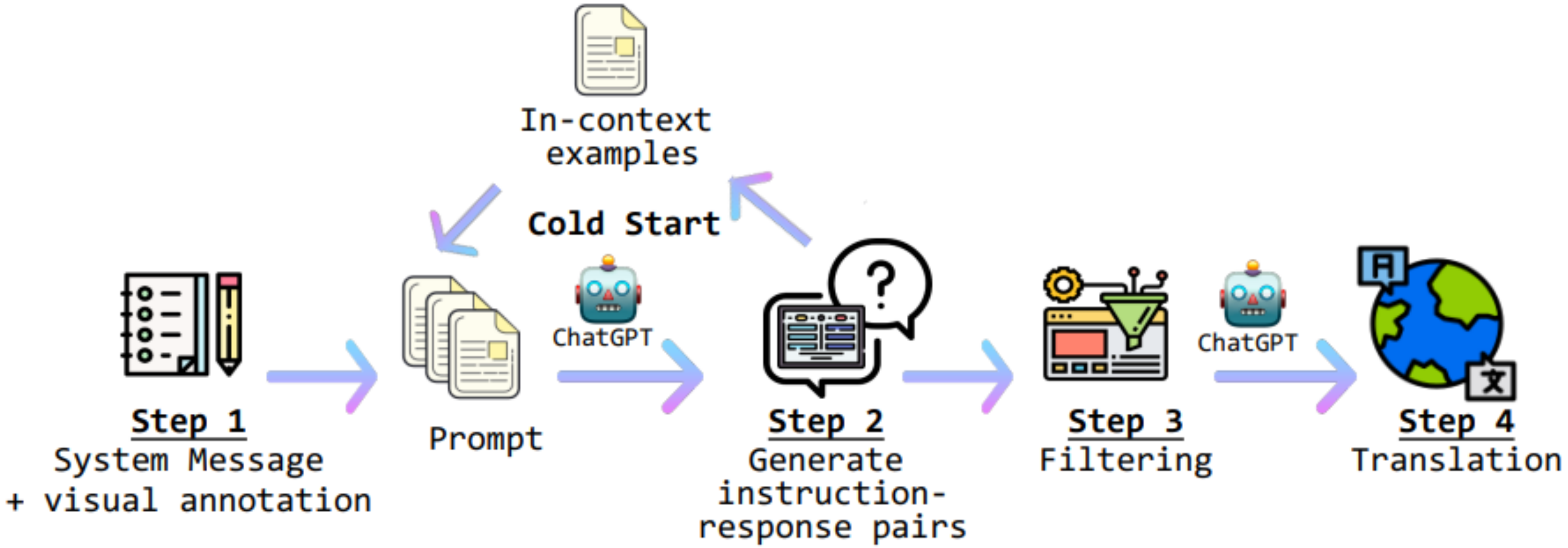
- 挑选少量图像，人为制作三种类型的指令样本，对话，详细描述，复杂推理
- GPT根据提示中约束，参照上下文示例对每个图像数据的描述文本信息都输出三种类型的指令样本

Response type 1: conversation
Question: What type of vehicle is featured in the image?
Answer: The image features a black sport utility vehicle (SUV).
Question: Where is the vehicle parked?
Answer: The vehicle is parked in an underground parking area, likely in a public garage.
Question: What are the people in the image doing?
Answer: The people are trying to fit all of their luggage into the SUV, likely preparing for a trip.

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.
In the parking area, various luggage items are scattered around the vehicle. There are two backpacks, one located near the left rear wheel and the other closer to the right side of the vehicle. Additionally, there are two suitcases, one on the right side of the car and another further away near the center of the parking area. A bicycle can also be seen on the left side of the vehicle.
Other cars are parked around the main SUV, with one car positioned behind it and slightly to the left, another behind and slightly to the right, and the third car further behind on the right side.

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 to accommodate. They might have to strategize and arrange the luggage efficiently to ensure that everything fits properly into the vehicle. Additionally, they need to consider the comfort of the passengers and visibility while driving, so the placement of the luggage must not obstruct the driver's view or make the passengers uncomfortable during the trip.

Sythus: 冷启动的自生成方法



- 实施冷启动(cold start)的方法，将生成的指令响应对作为上下文示例重新输入到GPT，以此反复优化上下文示例，再将示例作为模板进行后续步骤

LVIS-Instruct4V：利用GPT4V生成数据

图像数据
+
标注信息

prompt

GPT-4V

task1

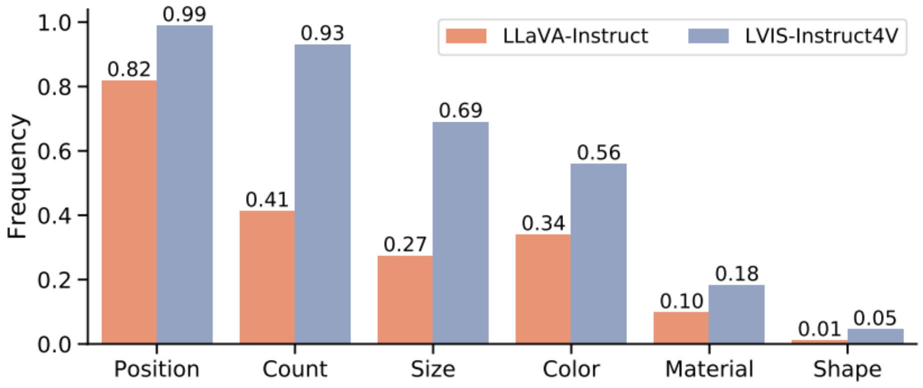
通过自我推理生成对话的QA数据

task2

生成图像描述信息

- 将LVIS数据集的图像数据和标注信息输入GPT4V，约束输出格式，获取两种类型的指令。

- 使用生成的数据集**LVIS-Instruct4V** 训练LLaVA，比原始数据训练得到模型有更丰富的结果输出。



LVIS-Instruct4V与LLaVA-instruct数据集效果对比

To See is to Believe: Prompting GPT-4V for Better Visual Instruction Tuning, arXiv:2311.07574

Valley⛰️： 大视频语言模型

VALLEY: VIDEO ASSISTANT WITH LARGE LANGUAGE MODEL ENHANCED ABILITY

Ruipu Luo[1,2*], Ziwang Zhao[1,3*], Min Yang[1*], Junwei Dong[1,4], Minghui Qiu[1], Pengcheng Lu[1], Tao Wang[1], Zhongyu Wei[2]
[1]ByteDance Inc [2]Fudan University [3]Beijing University of Posts and Telecommunications [4]Chongqing University

arXiv Code Demo

ABSTRACT

Recently, several multi-modal models have been developed for joint image and language understanding, which have demonstrated impressive chat abilities by utilizing advanced large language models (LLMs). The process of developing such models is straightforward yet effective. It involves pre-training an adaptation module to align the semantics of the vision encoder and language model, followed by fine-tuning on instruction-following data. However, despite the success of this pipeline in image and language understanding, its effectiveness in joint video and language understanding has not been widely explored. In this paper, we aim to develop a novel multi-modal foundation model capable of perceiving video, image, and language within a general framework. To achieve this goal, we introduce Valley: Video Assistant with Large Language model Enhanced abilitY. Specifically, our proposed Valley model is designed with a simple projection module that bridges video, image, and language modalities, and is further unified with a multi-lingual LLM. We also collect multi-source vision-text pairs and adopt a spatio-temporal pooling strategy to obtain a unified vision encoding of video and image input for pre-training. Furthermore, we generate multi-task instruction-following video data, including multi-shot captions, long video descriptions, action recognition, causal relationship inference, etc. To obtain the instruction-following data, we design diverse rounds of task-oriented conversations between humans and videos, facilitated by ChatGPT. Qualitative examples demonstrate that our proposed model has the potential to function as a highly effective multilingual video assistant that can make complex video understanding scenarios easy. Code, data, and models will be available at https://github.com/RupertLuo/Valley.

https://valley-vl.github.io/

⛰️Valley: Video Assistant with Large Language model Enhanced abilitY

Understanding Complex Videos Relying on Large Language and Vision Models

[Project Page] [Paper]~~[demo]~~

The online demo is no longer available, because we released the code for offline demo deployment

Video Assistant with Large Language model Enhanced abilitY
Ruipu Luo*, Ziwang Zhao*, Min Yang* (*Equal Contribution)

Generated by stablecog via "A cute llama with valley"

https://github.com/RupertLuo/Valley

Valley: Video Assistant with Large Language model Enhanced Ability, 2023

Valley⛰️: 基于大模型指令构造的缺陷

[0s - 9.81s] A young child wearing a helmet is climbing on monkey bars on a playground.
[0s - 9.36s] A person riding a bike goes around the young girl and she climbs until the last bar

Can you describe the scene that unfolds in the video?

In the video, a young child, wearing a helmet, climbs on monkey bars on a playground with a small boy underneath her. Meanwhile, a person riding a bike goes around the young girl until she climbs the last bar

Dense Caption

Instruction Data Generated by ChatGPT

Florida usa-1962: view of clothes hanging from clothesline on a sunny day

Illustrate the video through a descriptive explanation, painting a vivid picture of its content.

First, we see a backyard with a red and white tent covering an area, and a little girl standing under an umbrella. Next, ... We also see a woman sitting at a table and a stack of books on the table. Then ... A shirtless young man and a red awning over the building can also be seen in the background. After that... Finally... We can also see a woman sitting on a chair, ... , and a large brown bear in the background.

Short Caption

Instruction Data Generated by ChatGPT

- 基于ChatGPT的描述往往比较短，或者丰富性有限
- 自动化的方法会产生与图片不一致的描述信息（幻觉）

Valley⛰️： 视频文本微调数据集 Valley-Instruct-703k

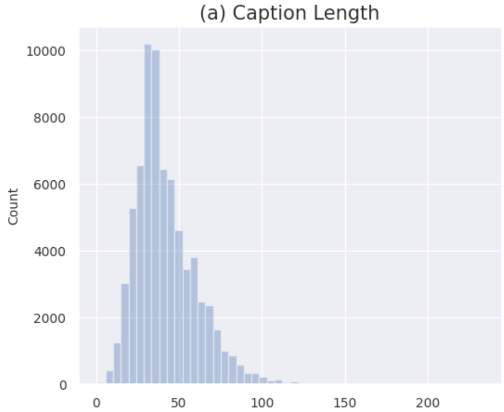
This guy did an amazing martial arts stunt. He jumped on the mat and did a backflip. At the same time, a volunteer was standing at the end of a mat holding a stick with a board at the top at the height of four meters. He amazingly kicked the board and landed safely on the mat.

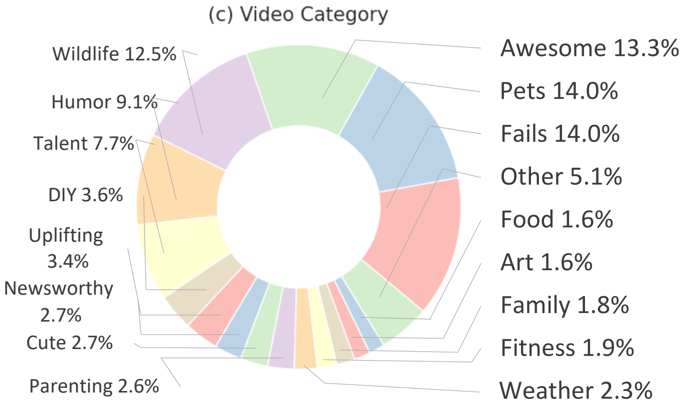
Illustrate the video through a descriptive explanation, painting a vivid picture of its content.

In this impressive video, a skilled individual showcases their martial arts prowess with a jaw-dropping stunt. The video begins with the person standing on a mat, poised for action. With a burst of energy, they launch themselves into the air, executing a flawless backflip mid-flight. Simultaneously, a volunteer positioned at the end of the mat holds a stick with a board balanced on top, an astonishing four meters above the ground. With ……

Detailed Caption

Instruction Data Generated by ChatGPT





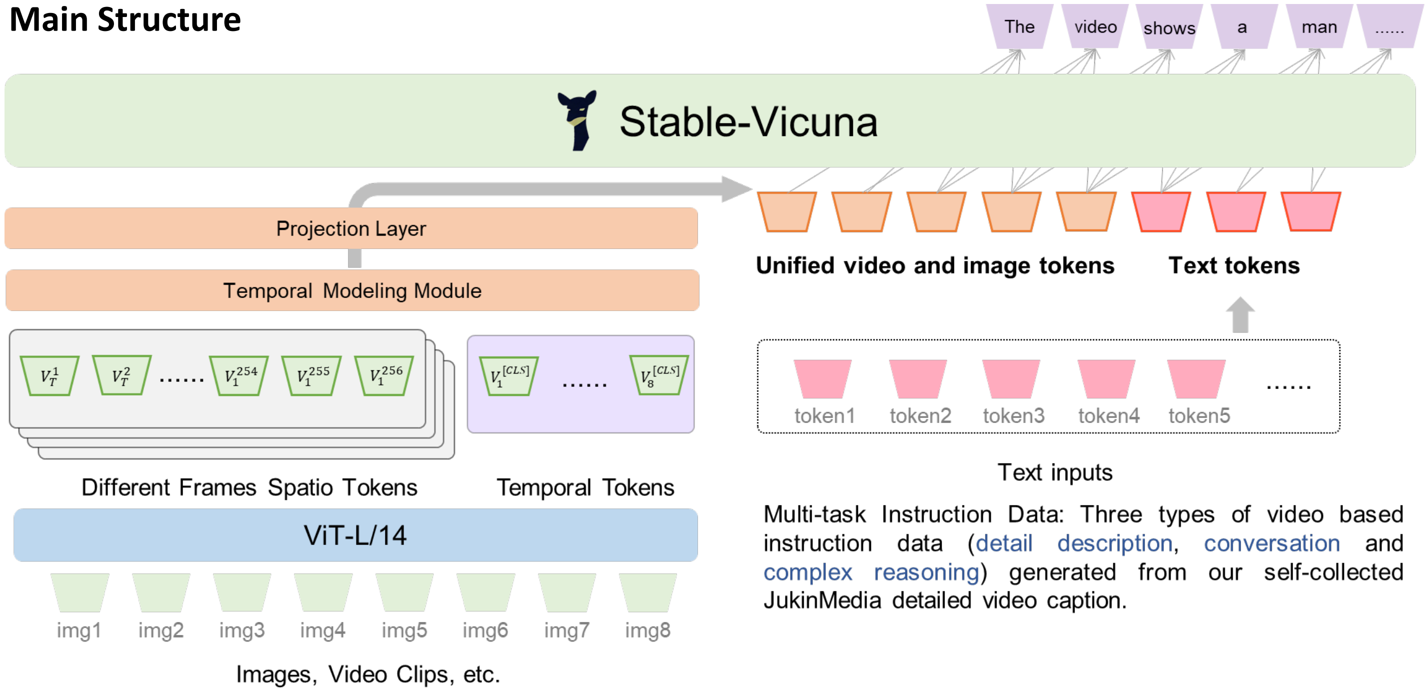
The data has been open-sourced in
https://huggingface.co/datasets/luoruipu1/Valley-Instruct-73k

- 基于 Jukinmedia 视频网站（视频描述很长，而且丰富）

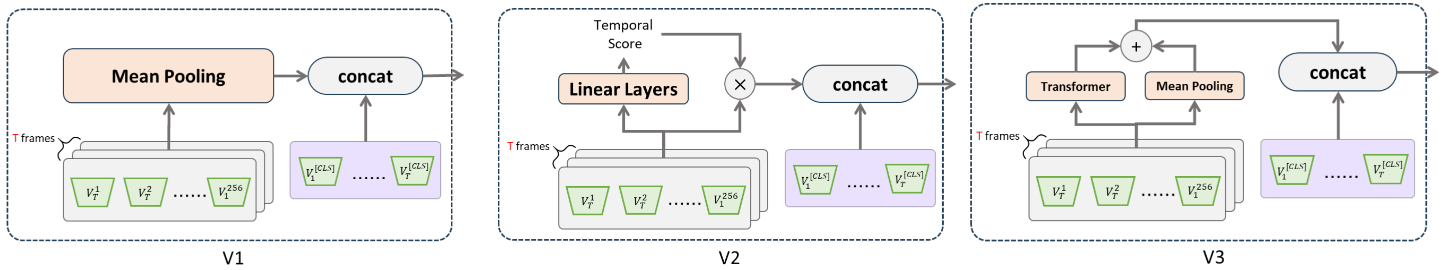
Valley: Video Assistant with Large Language model Enhanced Ability, 2023

Valley⛰️: 视频语言模型的架构介绍

Main Structure



Temporal Modeling Module



- 语言模型： stable-vicuna LLM
- 视觉编码： VIT
- 使用了三种方法进行视频编码

Valley⛰️: 样例演示

视频中的场景在哪?

The video shows a snowy forest with trees in the background. The two people are riding their snowmobiles through this scenic landscape.

视频中有几个人?

The video shows a snowy forest with trees in the background. The two people are riding their snowmobiles through this scenic landscape.

能为这个视频写一个故事吗?

Sure! Here's a story for the video: The video shows a winter wonderland, where two friends decided to go for a thrilling ride on their snowmobiles. They both chose their favorite color - red - for their jackets, making them stand out against the snowy backdrop. As they started their journey, they passed by trees and bushes, making the scenery even more picturesque. They rode deeper into the forest, and soon the trees became taller and closer together, creating a feeling of adventure. The friends were having a great time, enjoying the crisp winter air and the exhilarating speed of their snowmobiles. Finally, they emerged from the forest and rode off into the distance, leaving behind a trail of snow and excitement. This video captures the essence of winter fun and the joy of exploring the great outdoors on a snowmobile.

Valley ⛰️ 开源啦

Code: *https://github.com/RupertLuo/Valley*

Model Zoo:

- Valley-13b: *https://huggingface.co/luoruipu1/valley-13b-v1-delta*
- Valley-7b: *https://huggingface.co/luoruipu1/Valley2-7b*
- Chinese-Valley-13b: *https://huggingface.co/Zhaoziwang/chinese_valley13b_v1*
- *Chinese-Valley-7b: https://huggingface.co/Zhaoziwang/chinese_valley13b_v1*

Data:

- Instruction Tuning Data: *https://huggingface.co/datasets/luoruipu1/Valley-Instruct-73k*
- Pretrain-data: *https://huggingface.co/datasets/luoruipu1/Valley-webvid2M-Pretrain-703K*

Project Page: *https://valley-vl.github.io/*

小结

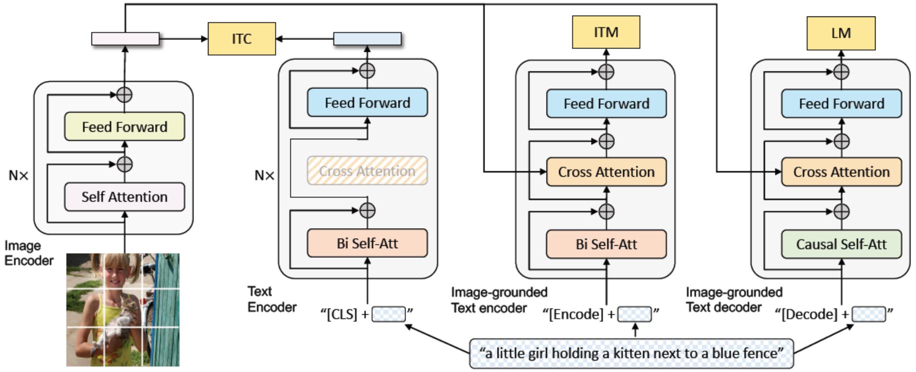
- 多模态大模型 或许是 **<u>“视觉扩充的大语言模型”</u>**

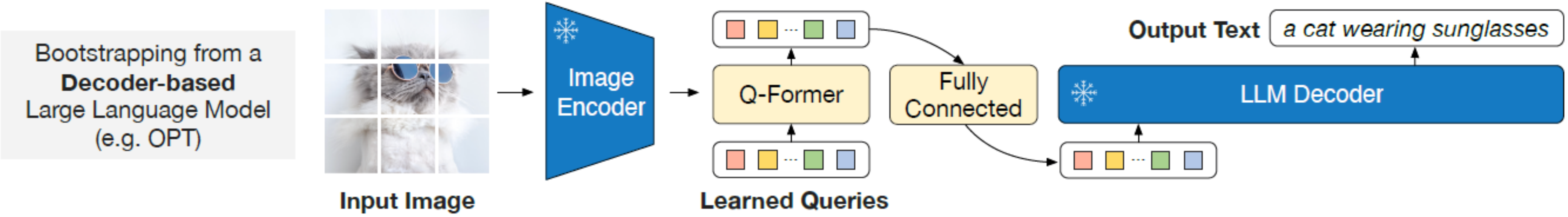
- 由于语言模型的研究得到了突破，多模态模型的训练中，模态地位开始不平等

- 预训练（将视觉对齐到语言）→ 指令微调 （适配到任务场景）

从BLIP 到 BLIP 2 看多模态大模型构建思路的演变

- BLIP-2中的Q-former与BLIP中的跨模态编码器结构相同，但是部件的定位已经从<u>跨模态交互</u>转变为<u>模态连接</u>
- 这种架构陈述的转变，是由语言模型从原本的与视觉编码器等参数部件转变为大语言模型带来的。
- 也指示了多模态大模型构建思路从模态平等到以语言为中心

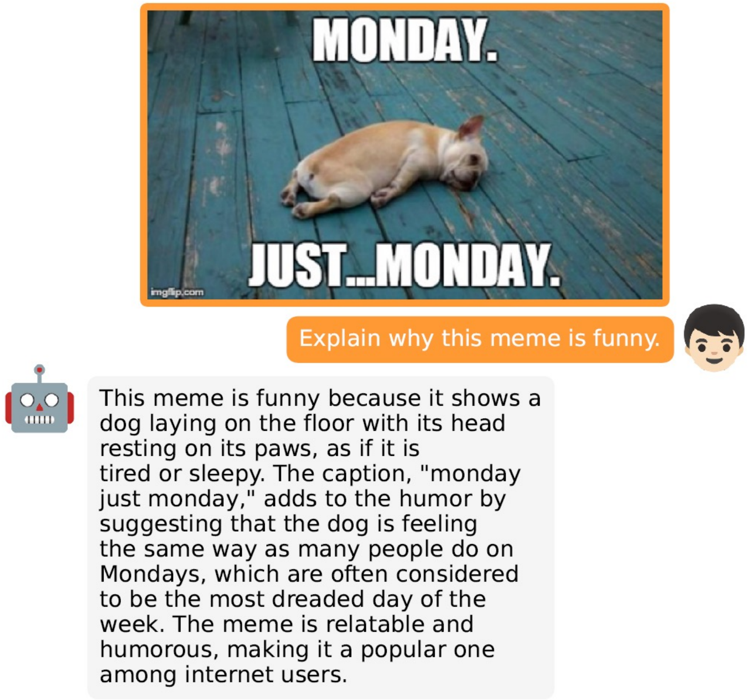




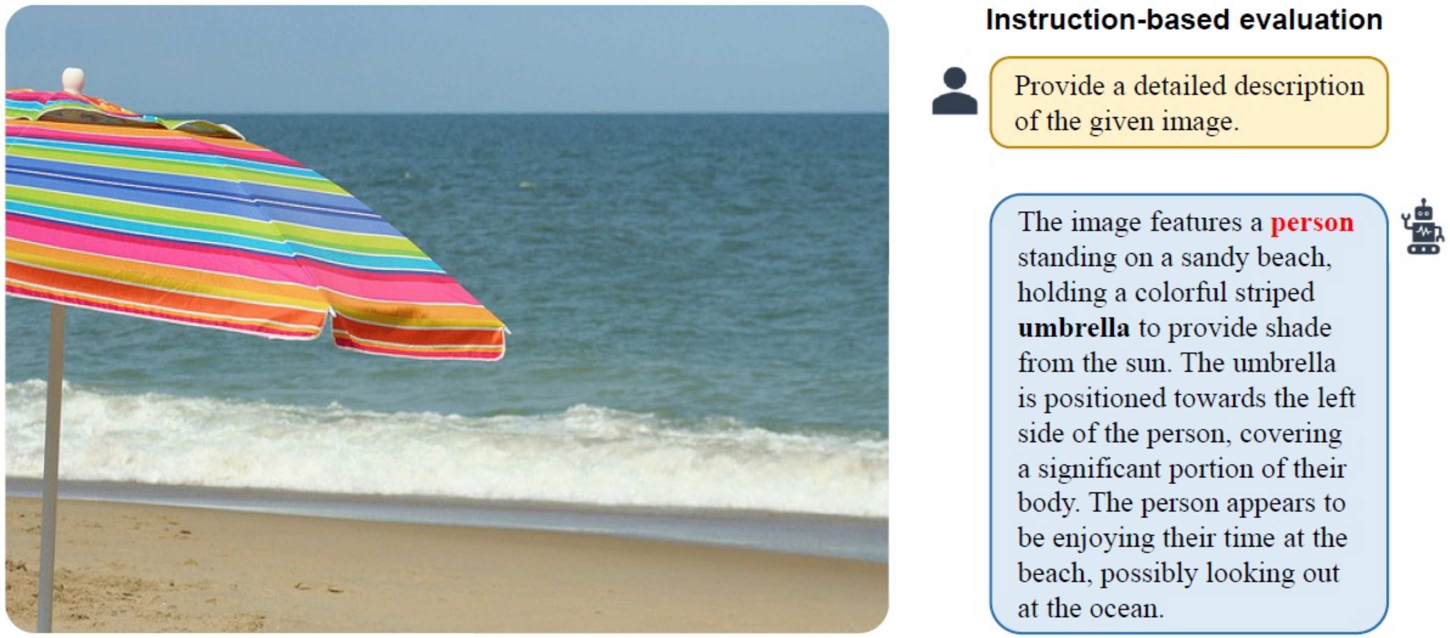
目录

大视觉语言模型（LVLM）的一体两面

- **展现强大的能力：** 光学字符识别（OCR），常识推理（Commonsense Reasoning），表情符理解（meme understanding）
- **难以解决根本的问题：** 物体幻觉（Object hallucination）



MiniGPT-4, 2023



Evaluating Object Hallucination in Large Vision-Language Models, 2023

大视觉语言模型间缺乏定量的分析和比较

- 现存的基准主要是**面向任务的：** 特定的输入输出格式
- 大视觉语言模型是**灵活的**，倾向于输出**详细回答:** 自由文本输入输出
- 如何自动地判断模型输出与任务特定的标签是否等价?

Benchmark: VQA v2
Q: Where is skateboarder looking?
GT: down

Benchmark: Visual Entailment
Claim: The player is well-skilled.
GT: entailment

Benchmark: Object Counting
Q: How many persons are there?
GT: 17

LVLMs

Formulation: Open-Ended QA
Prediction: He is looking at down the board.
Judgement: False [EM] / True [Human]

Formulation: Classification
Prediction: The image support the claim.
Judgement: False [EM] / True [Human]

Formulation: Number-Related QA
Prediction: There are more than 17 persons.
Judgement: True [Contain] / False [Human]

[*] indicates the evaluation method. Red and Green represent Wrong and Correct judgement. EM is short for “exactly matched”.

需要一个全面、可靠且易于使用的评价基准

- 评价目标
  - 任务或能力级别的评测

- 核心元素
  - 数据和标签
  - 问题形式化
  - 评价指标

MME：一个系统化的多模态大模型评测基准

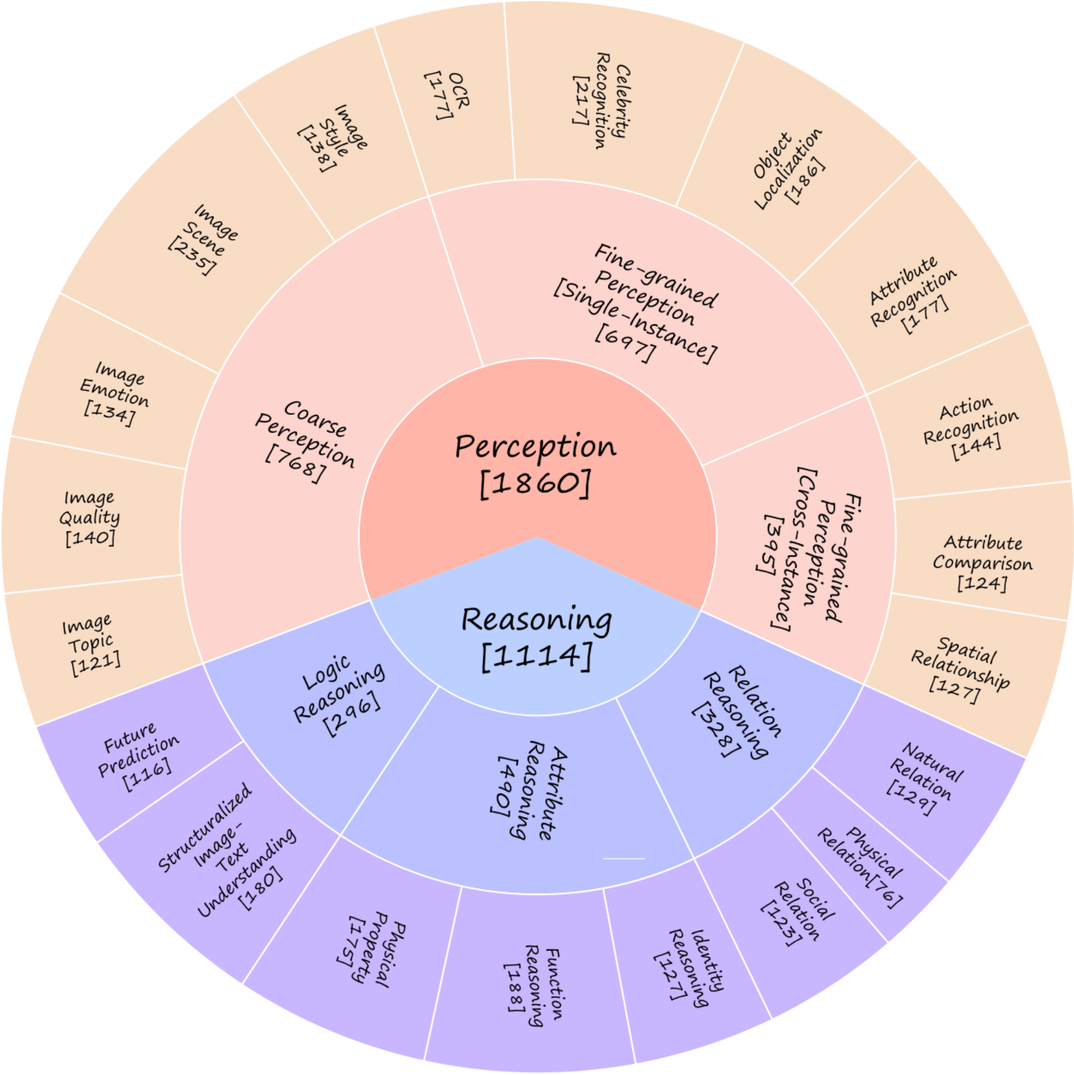
- 感知和认知，一共包括14个子任务
- 二元形式化：让模型回答yes [Y]或no [N]
- 值得注意的是，所有的指令都是人工设计的

MME的评价策略

- **让模型回答“yes”或“no”**
  - 指令包括两部分，分别是一个简明的问题和一个描述“Please answer yes or no.”。
  - 稳定性测试：对于每张测试图片，人工地设计两条指令，两条指令的问题不同，回答分别是“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。

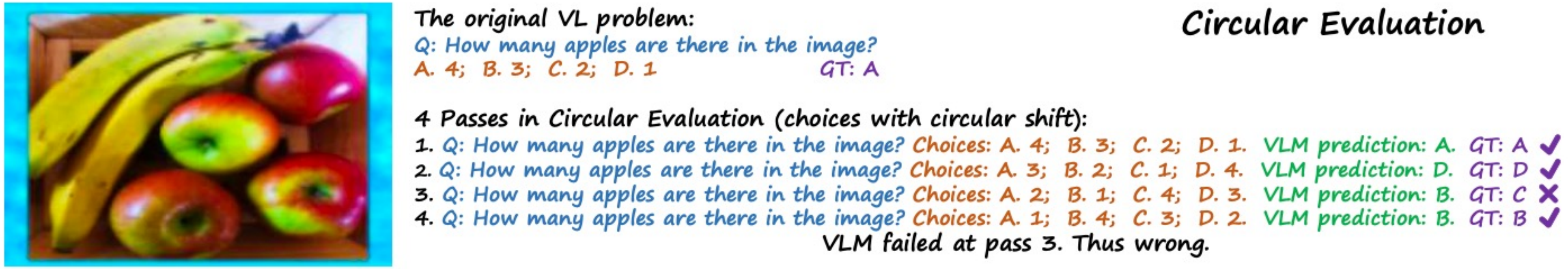
MMBench：一个综合全面的评测基准

- 三个水平的能力维度（L-1到L-3），其中包括20种不同的子能力。
- **L-1：感知**和**推理**
- L-2 感知：1.粗粒度感知，2.细粒度单实例感知，3.细粒度跨实例感知
- L-2认知：1.属性推理，2.关系推理，3.逻辑推理
- **L-3能力**是进一步从L-2能力中划分出来的。



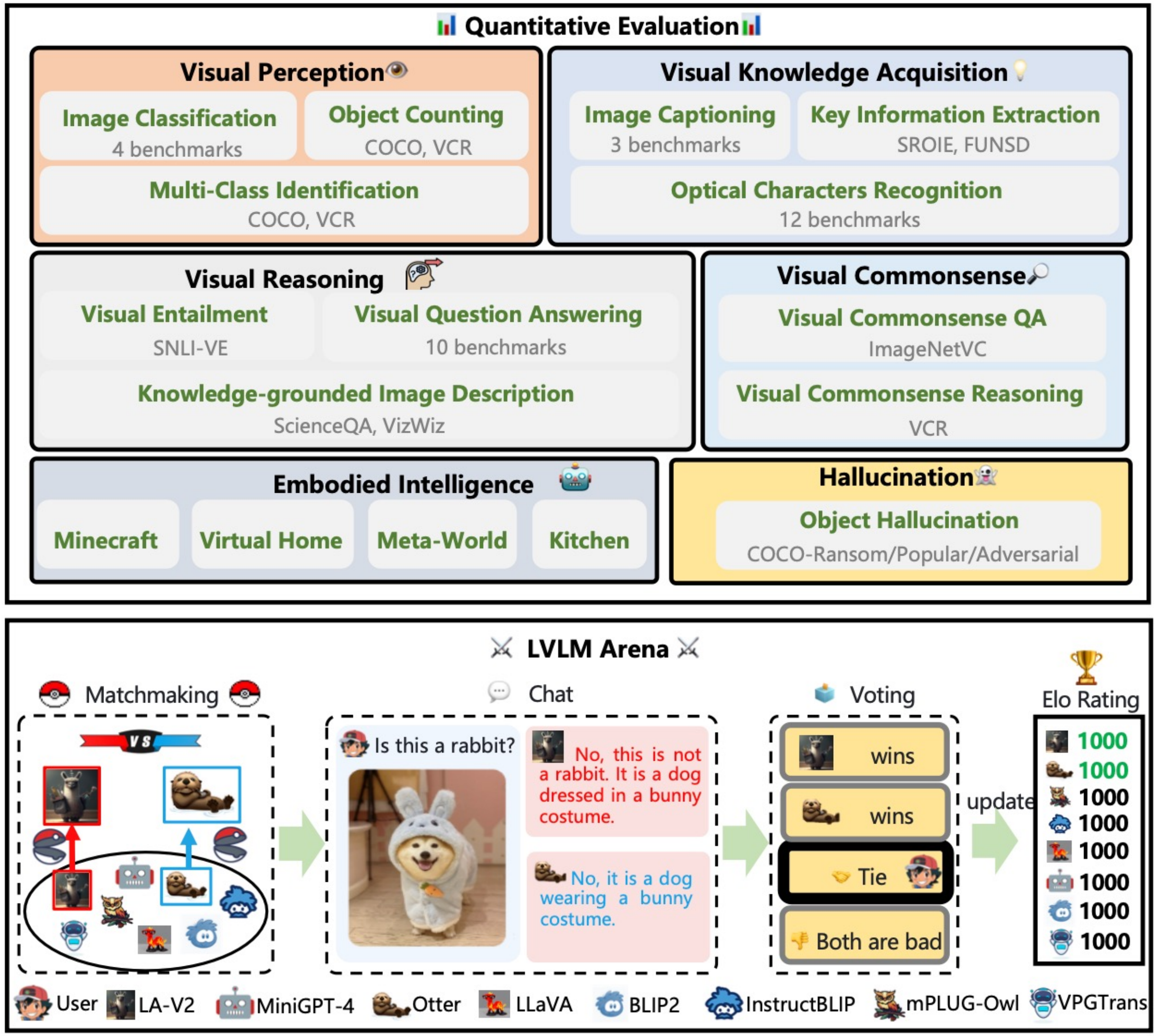
MMBench的评价策略

- **循环评价策略**
  - 循环评价将问题提供给VLM多次（使用不同的提示，调换答案的位置），并检查VLM是否在所有尝试中都成功解决了问题。
- **基于ChatGPT的答案抽取**
  - 为了解决VLM自由形式输出的问题，ChatGPT被利用来帮助抽取选择。



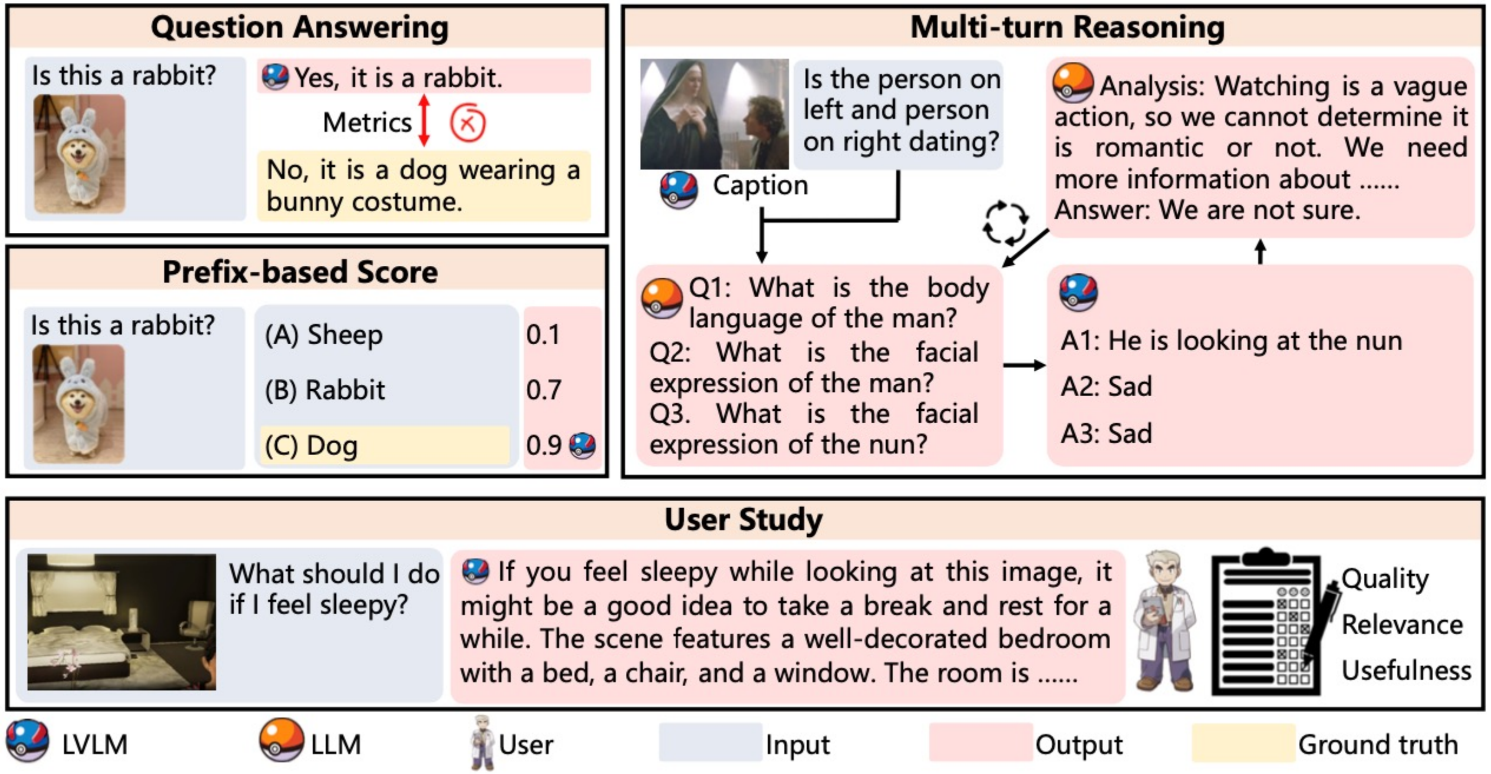
LVLM-eHub：模型评测擂台赛

- 模型在定量评测中的**六个关键能力**。
- 针对不同任务/数据集量身定制的评测方法。
- 在一个在线平台LVLM Arena上，用户可以参与在线评价，通过与两个匿名模型聊天并选择他们偏好的模型。



LVLM-eHub的在线评价

- 从模型集合中抽取两个模型。
- 用户与保持匿名的模型交谈。随后，**用户投票选出更好的模型**。
- 包括三个主要组成部分：
  - **配对**
  - **聊天**
  - **投票**



ReForm-Eval：“新瓶装旧酒”的基准构建方法

- 将面向任务的数据样本重新制定为与LVLM兼容的统一格式
  - 特殊的文本生成问题：对于光学字符识别（ORC）和图像描述任务
  - 多选题：对于剩余的其它任务
- 使用统一和兼容的形式实现通用和高效的评估

Unified Benchmark: ReForm-Eval

Q1: Answer the question “Where is skateboarder looking?” with the options. **Options**: (A) Down; (B) Up; (C) Right.

Q2: Does the image indicate that the player is well-skilled? Select the correct option. **Options**: (A) No; (B) Yes; (C) Maybe.

Q3: How many persons are there? Make your choice from the provided options. **Options**: (A) 17; (B) 7; (C) 15; (D) 20.

→ **LVLMs** →

Unified Formulation: Multiple-Choice

Prediction: The answer is (A) Down.
Judgement: True [Option Matching]

Prediction: The selected answer is (B) Yes.
Judgement: True [Option Matching]

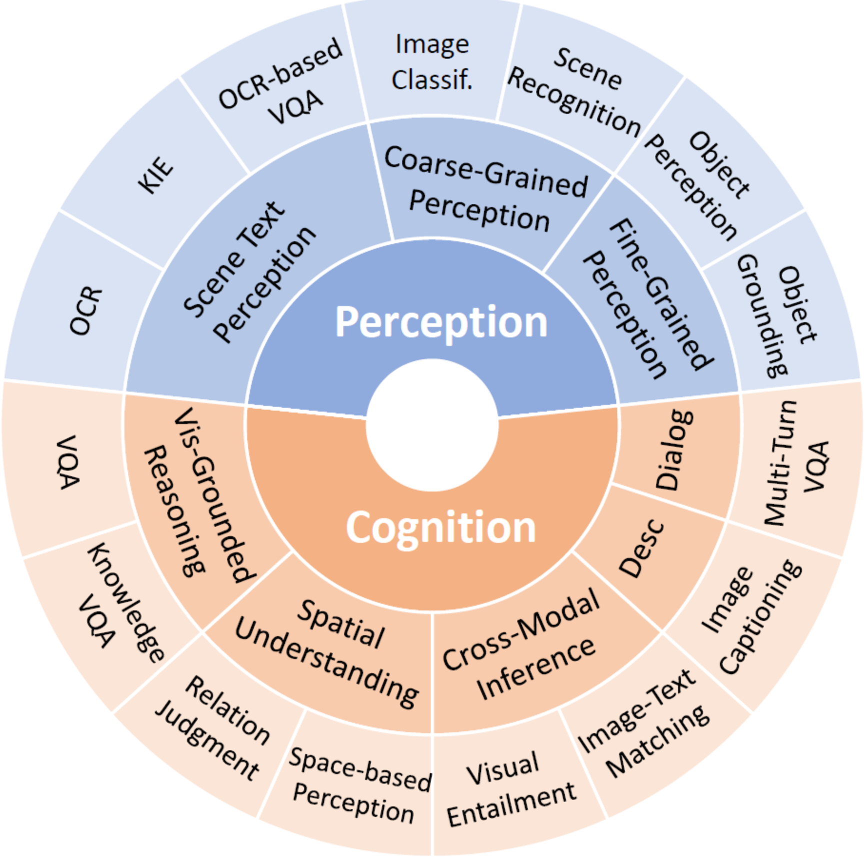
Prediction: The correct answer is (B) 7.
Judgement: False [Option Matching]

[*] indicates the evaluation method. Red and Green represent Wrong and Correct judgement. EM is short for “exactly matched”.

ReForm-Eval: Evaluating Large Vision Language Models via Unified Re-Formulation of Task-Oriented Benchmarks，2023

ReForm-Eval构建

- 61个现有基准数据集，来自2个主要类别、8个子类别和15个任务
- 专门的文本生成：
  - 视觉描述任务
  - OCR相关任务
- 多选题：
  - 标签 → 正选项
  - 难负选项：
    - 分类：类别之间的语义关系
    - 开放式QA: ChatGPT生成
    - 其它：任务特定的策略

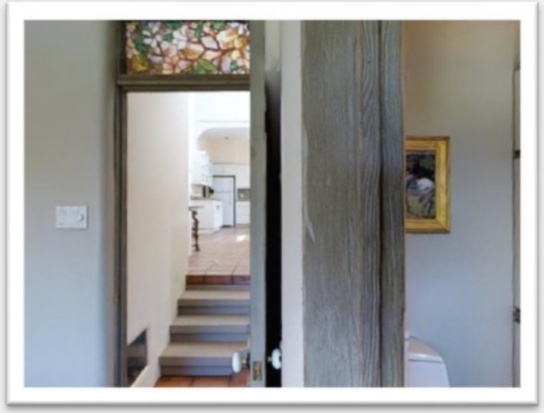


ReForm-Eval: Evaluating Large Vision Language Models via Unified Re-Formulation of Task-Oriented Benchmarks，2023

ReForm-Eval构建：空间理解

- 从Matterport3D中构建MP3D-spatial，用于在真实世界的VLN评估

Question: Describe the spatial connection between vase and mantel within the image.
Options:
(A) The vase is inside the mantel;
(B) The vase is right of the mantel;
(C) The vase is next to the mantel;
(D) The vase is on the top of the mantel.
Answer: (D)



Question: In the image, point out the object that has the greatest distance from you.
Options:
(A) picture;
(B) refrigerator;
(C) stairs;
(D) unknown.
Answer: (B)

ReForm-Eval: Evaluating Large Vision Language Models via Unified Re-Formulation of Task-Oriented Benchmarks，2023

ReForm-Eval评价：统一的形式

- 文本生成问题的评估取决于场景：
  - 视觉描述：
    - 为了简洁的输出，限制了最大的输出长度
    - 指标：CIDEr（参照BLIP-2）
  - OCR相关：
    - 指标：token-level精度，出现在输出中的目标token的比例
- 多选题
  - 输出中的选项匹配：检测输出中的“(A)”等选项标记
  - 指标：准确性
  - 挑战：现存的LVLM可能并不会遵循多选指令
    - E.g., 生成 “Blue” 而不是“(A) Blue” 或 “(A)”

解耦LVLM的指令遵循能力

- 黑盒方法：上下文学习（In-Context Learning）
  - 指导LVLM通过ICL以所需格式生成：

> $X_{\text{system-message}}$
> Human: Can you see the image? Options: (A) Yes; (B) No; (C) Not Sure; (D) Maybe.
> Assistant: The answer is (A) Yes.
> Human: X_{question} Options: X_{options}
> Assistant: The answer is

  - 其中上下文样本仅有文本内容并且不提供图片的信息。
- 白盒方法：似然率的（likelihood）评价
  - 计算每个选项的likelihood，并选择可能性最高的选项：

$$\hat{c} = \arg\max_{c^i \in C} P_\theta(c^i | v, q) = \arg\max_{c^i \in C} \sum_{t=1}^{t_c} P_\theta(c^i_t | v, q, c^i_{<t})$$

  -
  - 其中$C = \{c^i\}_{i=1}^N$ 是选项，v 是图像，q 是问题，P_θ 通过LVLM建模。

大视觉语言模型的输出稳定性评价

- 大模型对提示敏感：
  - 每个样本使用不同但等效的提示进行多次测试
  - 不同的指令模板、打乱选项、随机选项标记
  - 最终性能是多次测试的平均值
- 不稳定性测量：
  - 预测分布的熵（仅适用于多选题）
  - $e = -\sum_{i=1}^{N} p_i \log(p_i)$ where $p_i = \frac{1}{M}\sum_{j=1}^{M} \mathbb{I}(\hat{c}_j = c_i)$
  - 其中 M 是多次测试的数量， $\hat{c}_j$是第j次测试的结果。

ReForm-Eval: Evaluating Large Vision Language Models via Unified Re-Formulation of Task-Oriented Benchmarks，2023

Reform-Eval的总体实验

- 总体性能：13种方法的16个模型，具有不同的基座

Model	Vis Encoder	LLM	Connection Module	#oP	#oTP	#oVT
BLIP-2	ViT-G/14	FlanT5-XL	Q-Former	3.94B	106.7M	32
InstructBLIP$_F$	ViT-G/14	FlanT5-XL	Q-Former	4.02B	187.2M	32
InstructBLIP$_V$	ViT-G/14	Vicuna-7B	Q-Former	7.92B	188.8M	32
LLaVA$_V$	ViT-L/14	Vicuna-7B	Linear	7.05B	6.74B	256
LLaVA$_{L_2}$	ViT-L/14	LLaMA2-7B	Linear	7.05B	6.74B	256
MiniGPT4	ViT-G/14	Vicuna-7B	Q-Former+Linear	7.83B	3.1M	32
mPLUG-Owl	ViT-L/14	LLaMA-7B	Perceiver	7.12B	384.6M	65
PandaGPT	ImageBind	Vicuna-7B+LoRA	Linear	7.98B	37.8M	1
IB-LLM	ImageBind	LLaMA-7B+LoRA+BT	BindNet+Gate	8.61B	649.7M	1
LA-V2	ViT-L/14	LLaMA-7B+BT	Linear+Adapter+Gate	7.14B	63.1M	10
mmGPT	ViT-L/14	LLaMA-7B+LoRA	Perceiver+Gate	8.37B	23.5M	64
Shikra	ViT-L/14	Vicuna-7B	Linear	6.74B	6.44B	256
Lynx	ViT-G/14	Vicuna-7B+Adapter	Perceiver	8.41B	688.4M	64
Cheetor$_V$	ViT-G/14	Vicuna-7B	Query+Linear+Q-Former	7.84B	6.3M	32
Cheetor$_{L_2}$	ViT-G/14	LLaMA2-Chat	Query+Linear+Q-Former	7.84B	6.3M	32
BLIVA	ViT-G/14	Vicuna-7B	Q-Former+Linear	7.92B	194.6M	32

PS: Underlined represents a trainable component. “BT” represents bias-tuning . “BindNet” represents bind network.

Table 7: Model architecture of different LVLMs. “#oP”, “#oTP”, and “#oVT” are number of total parameters, number of trainable parameters, and number of visual tokens, respectively.

ReForm-Eval的优点

- 数据丰富的全面评价:
  - 评估维度包含感知到推理
  - 重新制定了61个基准数据集，丰富的数据收集（~3000样本每维度（MMBench/MME大小的10倍））
  - 无需人工标注
- 高效的评价:
  - 基于统一形式的通用评价方法
  - 无需任务特定的评价方法（ 在LVLM-ehub中的）
  - 无需ChatGPT或人工的帮助（在LAMM和MMBench中的）

ReForm-Eval: Evaluating Large Vision Language Models via Unified Re-Formulation of Task-Oriented Benchmarks，2023

ReForm-Eval的优点

- 可靠的评价：
  - 黑盒和白盒两种辅助LVLM多项选择题的评估方法
- 不稳定性感知的评价：
  - 使用不同但等效的提示对同一样本进行多次测试
  - 对多选题的直接的不稳定性测量

ReForm-Eval开源了！

- https://github.com/FudanDISC/ReForm-Eval/

ReForm-Eval

Version v1.0 | Licence Apache 2.0 | DISC Repositories | Stars 5 | Visitors 49 / 227

Paper PDF | Paper Arxiv | Hugging Face Dataset | Google Drive Dataset

ReForm-Eval: EVALUATING LARGE VISION LANGUAGE MODELS VIA UNIFIED RE-FORMULATION OF TASK-ORIENTED BENCHMARKS

Zejun Li[1†], Ye Wang[1†], Mengfei Du[1†], Qingwen Liu[1†], Binhao Wu[1†], Jiwen Zhang[1†], Chengxing Zhou[2], Zhihao Fan[3], Jie Fu[4], Jingjing Chen[1], Xuanjing Huang[1], Zhongyu Wei[1*].

[1]Fudan University [2]Northeastern University [3]Alibaba Group [4]Hong Kong University of Science and Technology

[†]Equal Contribution [*]Corresponding Author

ReForm-Eval Paper | ReForm-Eval-Data | Google Drive

小结

- 大模型展示了强大的综合能力，对于它的评价也变得复杂

- （1）任务/能力的多样性
- （2）评价方法的高效性
- （3）输出结果的稳定性
- （4）测试样本的不可见性

目录

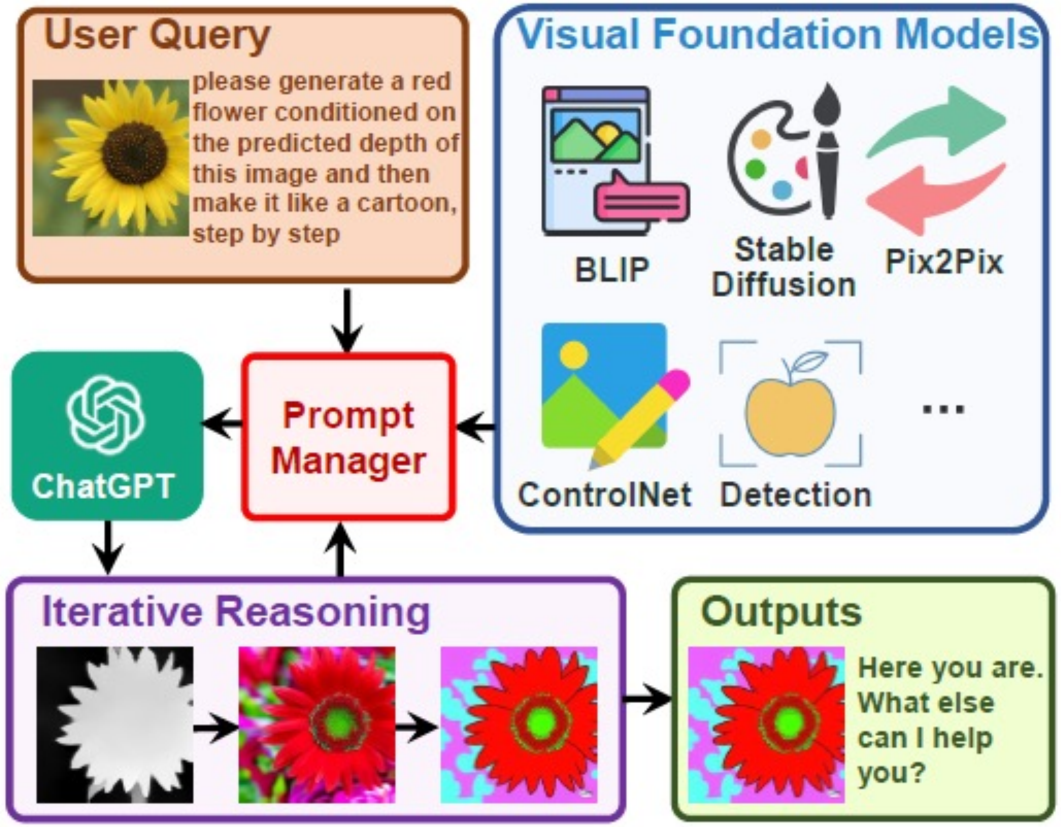
- ChatGPT之前的视觉语言预训练
- 大视觉语言模型的架构和训练
- 大视觉语言模型的评测
- **大视觉语言模型的能力扩充**
- 大语言模型支撑的具身智能（视觉导航）

LVLM的能力扩充: 输出空间的扩展

- 从 LLM 到 LVLM:
  - 完成了输入空间的扩展
  - 通过图文对进行输入空间的对齐
  - 自然地通过LLM基座以文本方式进行输出
- 多模态大模型可以输出离散token以外的输出吗?
  - 连续型输出: 坐标, 标记框 ...
  - 其他模态: 图片, 音频, 3D 点云...

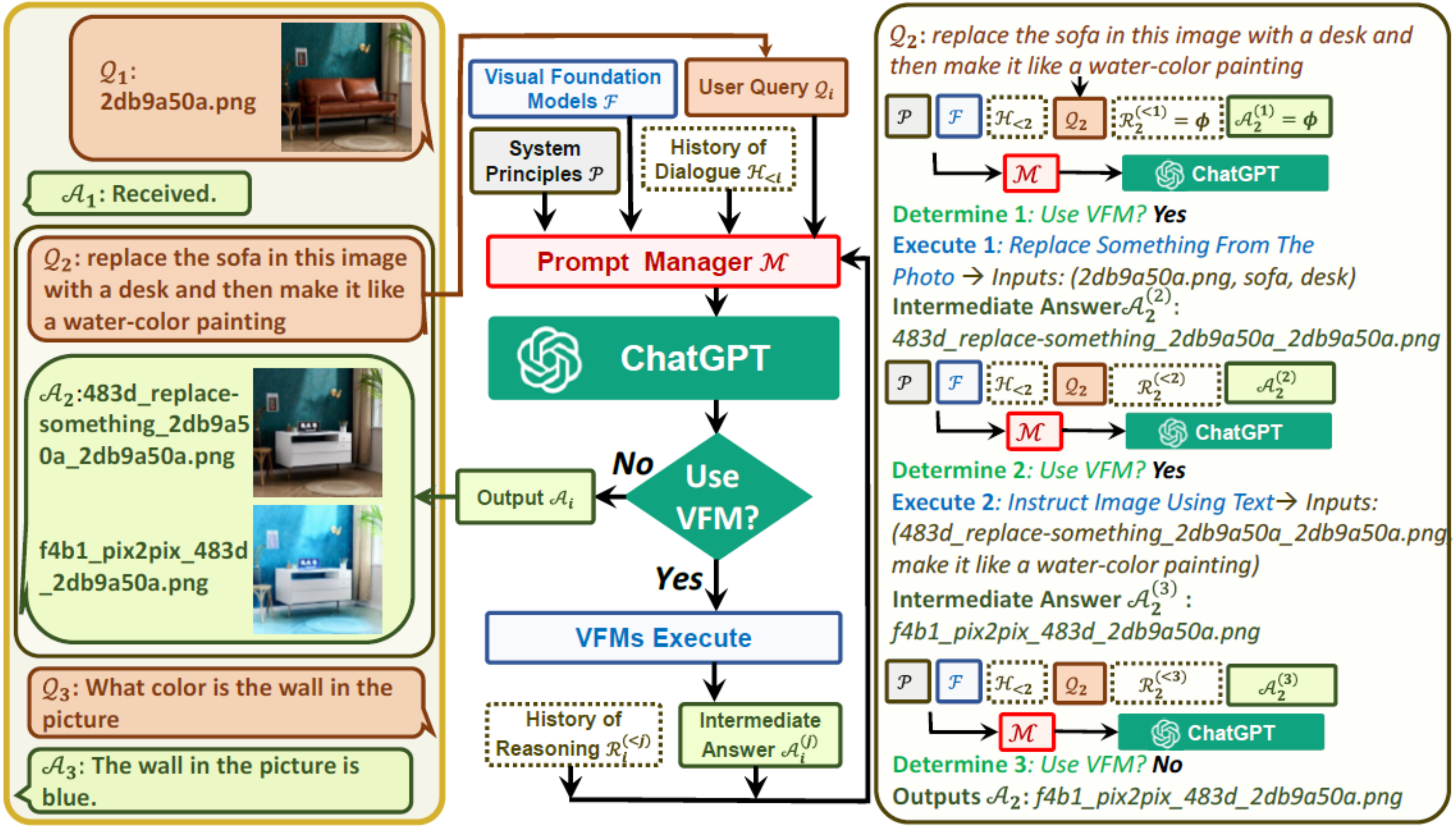
Visual ChatGPT：以Zero-Shot方式使用工具

- **通过文本指令来使用工具!**
- **基座：** ChatGPT
  - 泛用而灵活的系统
  - 局限于文本输入 / 输出
- **工具：** 视觉基础模型 (VFM)
  - 具有特定方面的视觉能力
- **输出空间：** 基于工具得到扩展
  - 图片, 物体标记框...
- **拓展方法:**
  - 通过“**prompts** manager”
  - **Zero-shot** 拓展方式，基于ChatGPT



Visual chatgpt: Talking, drawing and editing with visual foundation models，2023

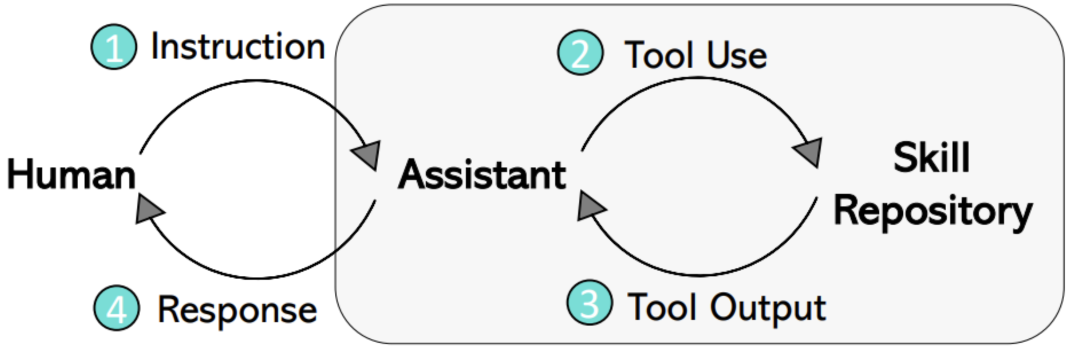
Visual ChatGPT :以Zero-Shot方式使用工具



Visual chatgpt: Talking, drawing and editing with visual foundation models，2023

LLaVA-Plus：训练模型学习工具的使用

- **基座：** LLaVA (或其他任意 LVLMs)
- **输出空间:**
  - 图片: 基于 Stable Diffusion
  - 图分割: 基于 SAM
  - 标记框: 基于物体检测器
  - …



- **拓展方法:**
  - 4-轮对话的形式
  - **通过构建的数据训练模型学习**遵循使用工具的指令

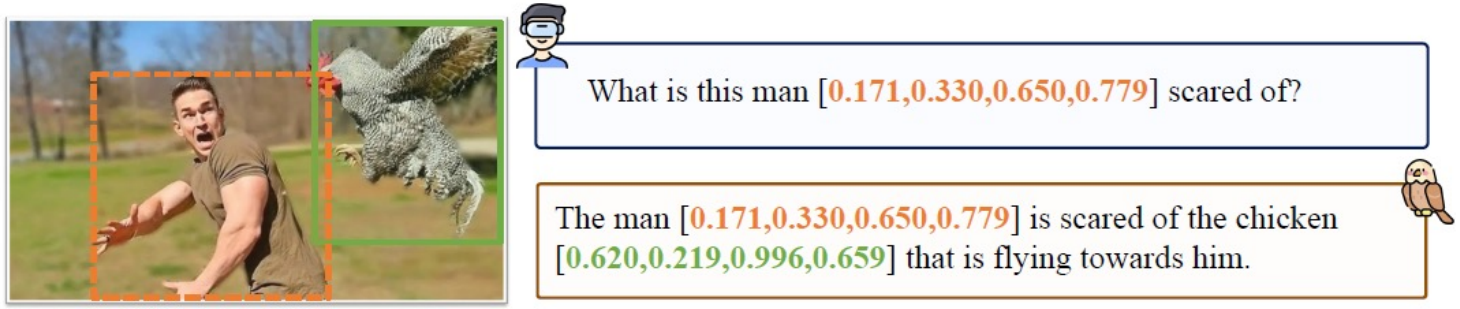
$$\texttt{Human}:\mathbf{I}_{\mathrm{q}}\ <\backslash n>\ \mathbf{X}_{\mathrm{q}}\texttt{<STOP>}\ \texttt{Assistant}:\mathbf{X}_{\texttt{skill\_use}}\texttt{<STOP>}$$

$$\texttt{Human}:\mathbf{X}_{\texttt{skill\_result}}\texttt{<STOP>}\ \texttt{Assistant}:\mathbf{X}_{\texttt{anwser}}\texttt{<STOP>}$$

Llava-plus: Learning to use tools for creating multimodal agents，2023

Shikra：以文本表示连续的数值

- **输出空间：** 连续的坐标
- **拓展方法：**
  - 以 **自然语言的形式** 来表示连续的数值



Shikra: Unleashing Multimodal LLM's Referential Dialogue Magic，2023

  - 指令遵循训练数据的构建:
    - 重构已有的数据: RefCOCO, Visual-7W, visual genome, Flickr30k entities
    - 生成的QA数据: 基于 Flickr30k entities 数据通过GPT-4生成
  - 训练阶段-1: 使用重构的数据
  - 训练阶段-2: LLaVA + 生成的QA数据

Kosmos2：以扩展词表的形式进行Grounding

- **输出空间:** 标记框
- **拓展的方法:** 拓展词表
  - 位置 tokens: P*P tokens 来表示 P*P 个图片里的分块
  - 特殊 tokens: 以markdown里 **超链接** 形式进行表示
    - <p>文本描述</p><box>标记框</box>
    - <grounding> 作为一个**开关**来指示模型是否需要进行grounding

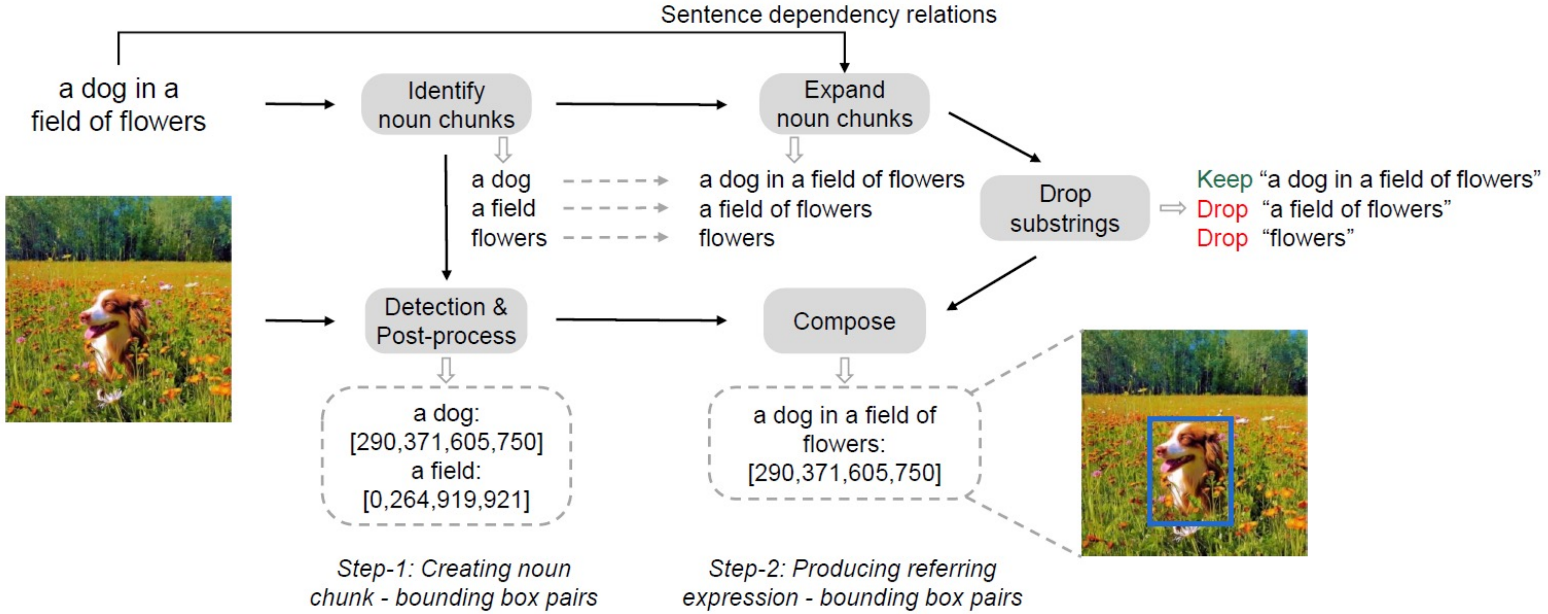
<s> <image> Image Embedding </image> <grounding> <p> It </p><box><loc$_{44}$><loc$_{863}$></box> seats next to <p> a campfire </p><box><loc$_{4}$><loc$_{1007}$></box> </s>

- 预训练: 图文对 + 文本数据 + GRIT
- 指令微调: LLaVA + unnatural instructions + GRIT

Kosmos-2: Grounding Multimodal Large Language Models to the World，2023

Kosmos2: GRIT （Grounded Image-Text） 构建

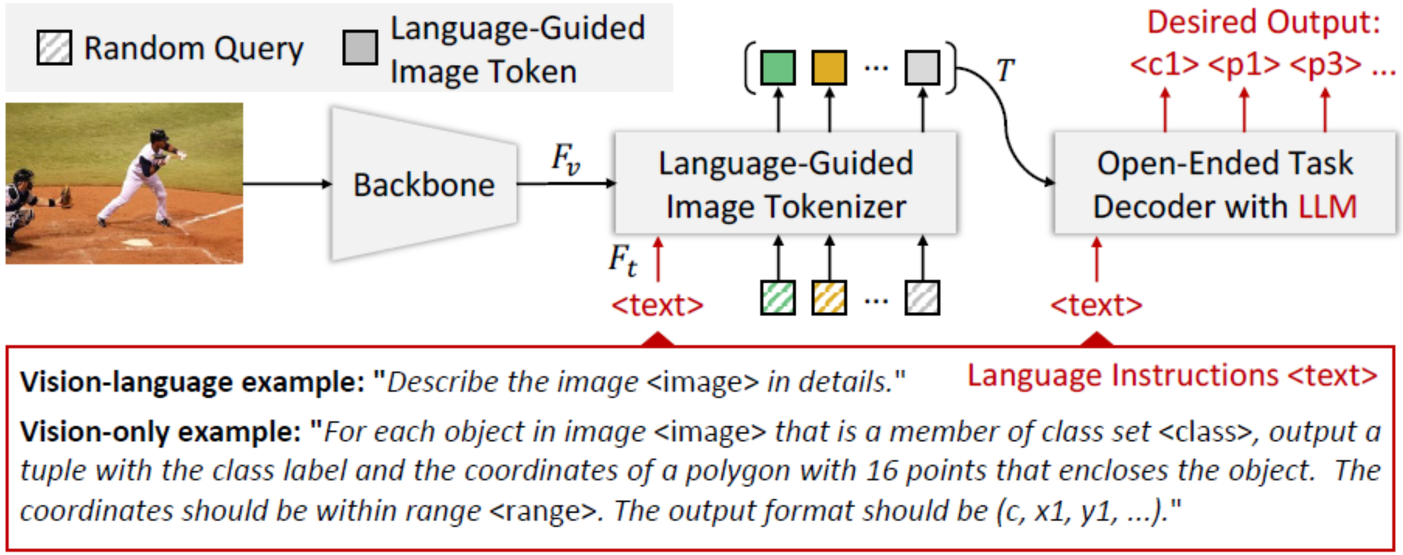
- 步骤-1 构造名词短语-标记框对: SpaCy + GLIP
- 步骤-2 构造referring-expression-标记框对
- 通过不断聚合语义树的节点将名词短语拓展到referring expression
- 舍弃被其他描述（referring expression）包含的项



Kosmos-2: Grounding Multimodal Large Language Models to the World，2023

VisionLLM: 更丰富的词表扩充

- **输出空间:** 分类类别 + 坐标
- **拓展方法：** 在输入指令里扩展: 任务描述 + 输出格式的定义

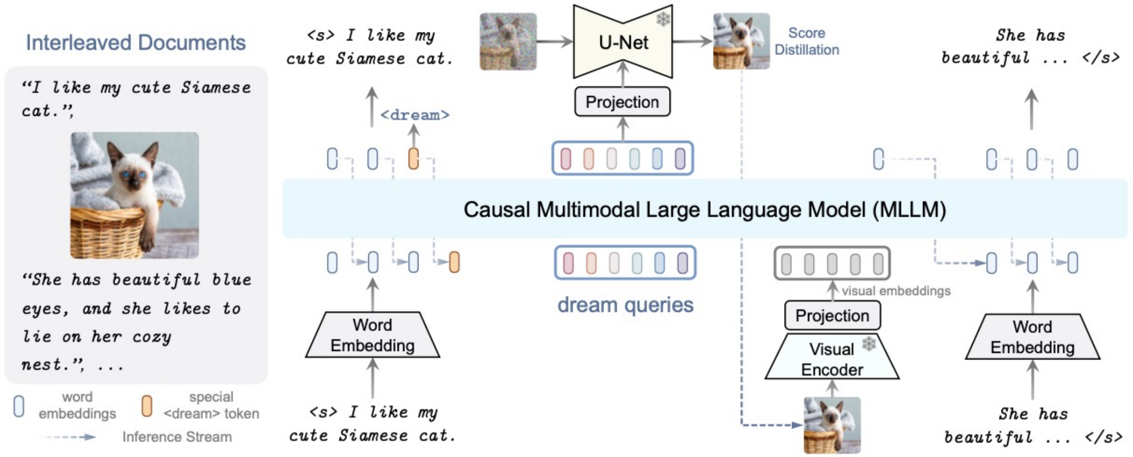


- 词表内进行扩充:
  - 512 位置 tokens: 表示坐标 + 类别 tokens: 作为类别的index
  - 输出tokens: 用来表述输出的形式，并在已知输出形式的情况下进行高效的解码

Visionllm: Large language model is also an open-ended decoder for vision-centric tasks，2023

DreamLLM：引入图文交错的输出形式

- **基座：**CLIP + Vicuna + Stable Diffusion
- **输出空间：**与文本交错的图片
- **拓展方法:**
  - <dream> token 占位符指示产生图片的位置
  - 引入可学习的dream queries

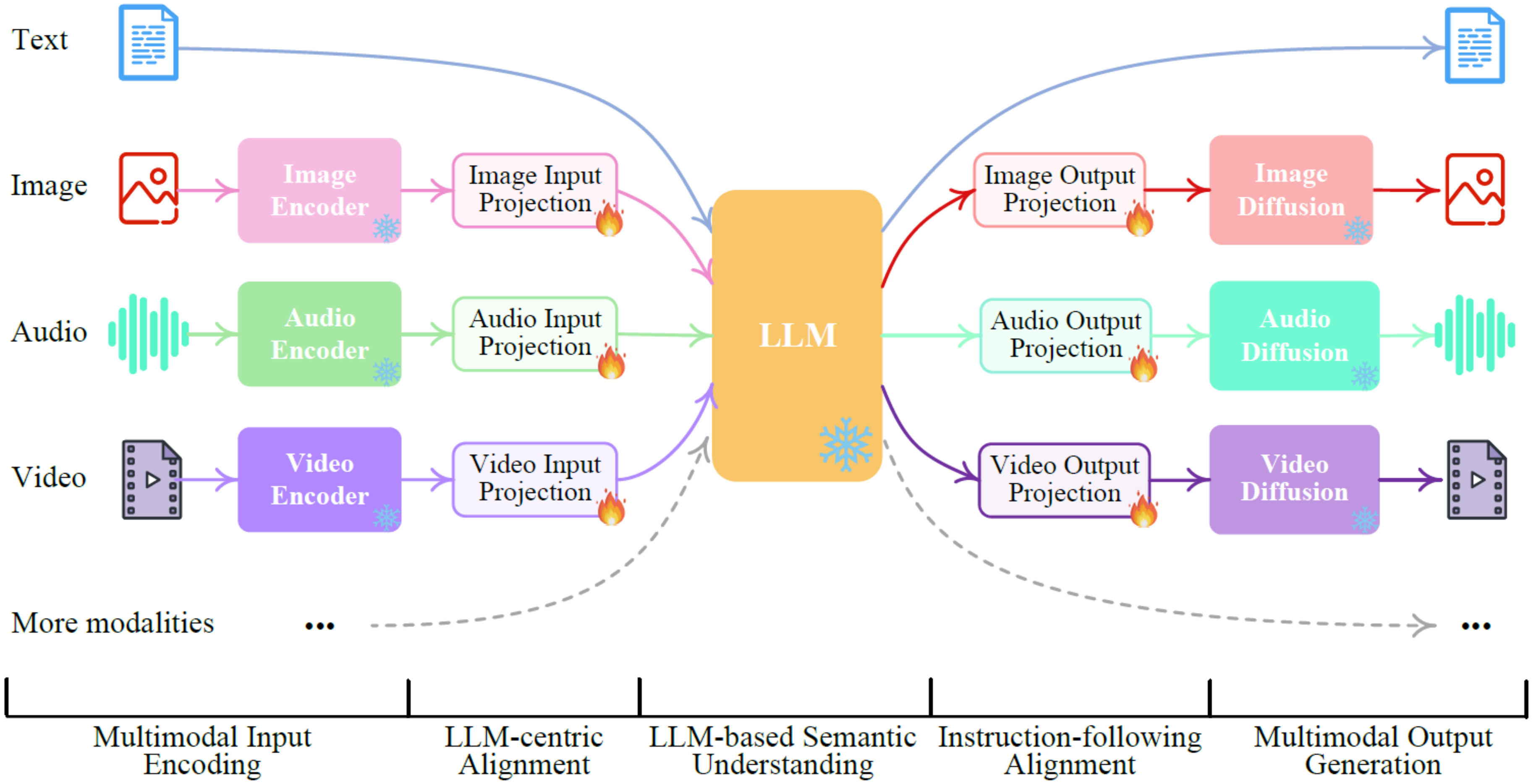


- **数据构建:**
  - MMC4中的多模态文档
  - 利用GPT-4从文档构建指令相关的QA对

Dreamllm: Synergistic multimodal comprehension and creation，2023

Next-GPT: 任意模态的生成

- **输出空间：**任意模态、模态交错的信息



Next-gpt: Any-to-any multimodal llm, 2023

Next-GPT：任意模态的生成

- **输出空间：**任意模态、模态交错的信息
- **拓展方法：**在生成端引入模态信息占位符
  - **指示特定位置生成特定模态信息，**E.g. <IMG0><IMG1><IMG2><IMG3> 指示图片生成，占位符对应的表示作为**对应模态解码器的输入**
- 指令微调数据集:
  - **文本 + X – 文本**: LLaVA, miniGPT-4, VideoChat
  - **文本 – 文本 + X**: 基于 X-描述 数据构造
  - **MosIT**: 构造的 5K 对话
    - 基于GPT-4的Self-instruct方法: 构造多轮、多模态、模态交互的对话
    - 搜集最匹配的对应模态数据: Youtube, StableDiffusion, Midjourney
    - 人工筛选，保证质量

Next-gpt: Any-to-any multimodal llm, 2023

LVLM的能力扩充：特定任务上能力的增强

- 在特定任务上 LVLMs 和对应的 SOTAs 仍有差距 (数据来自 Qwen-VL):

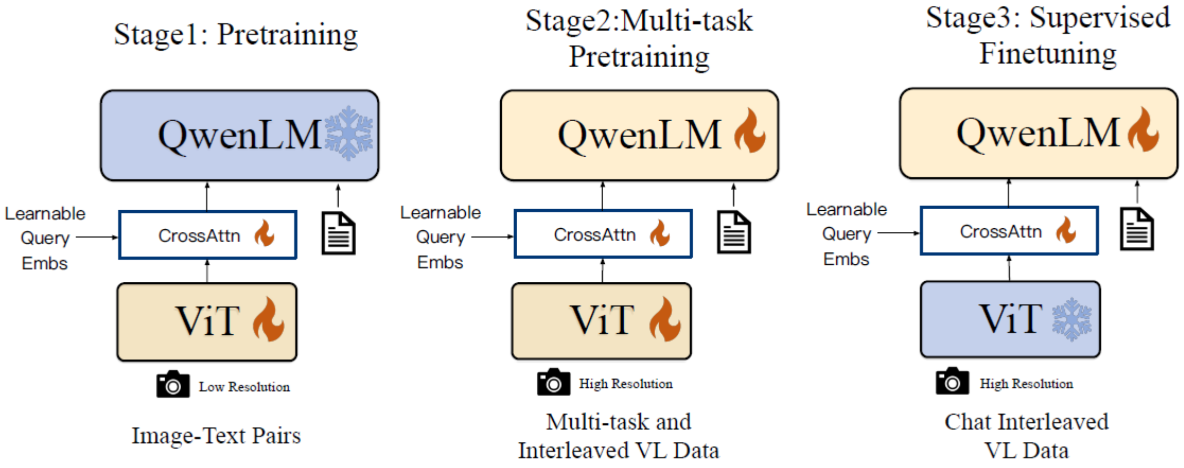
Model	Nocaps	Flickr30K	VQA v2	OKVQA	GQA	SciQA-Img	Vizwiz
BLIP-2	103.9	71.6	65.0	45.9	32.3	61.0	19.6
Specialist SOTAs	127.0 (PALI)	84.5 (InstructBLIP)	86.1 (PALI-X)	66.1 (PALI-X)	72.1 (CFR)	92.5 (LLaVA)	70.9 (PALI-X)

  - **Zero-shot** LVLMs v.s **Fine-tuned** SOTAs
  - LVLM没有学习过特定任务的输出输入结构信息

- LVLM能在特定任务上缩小和SOTA的差距吗?
  - 主要关注的任务: VQA, Object Grounding, Image Captioning

Qwen-VL：多任务学习

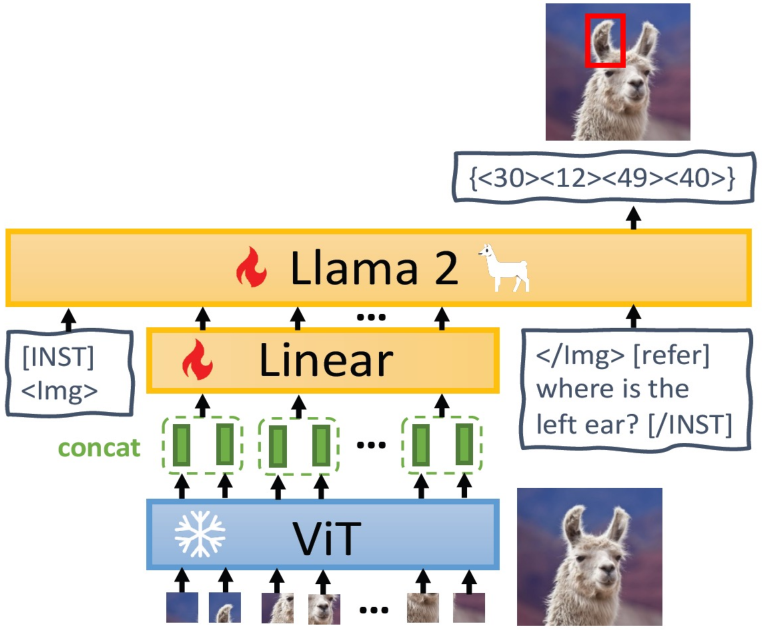
- **视觉编码器:** OpenCLIP ViT-bigG (448 px)
- **LLM：** Qwen-7B; **连接模块:** 单层的cross-attention模块
- **3-阶段的训练框架：**
  - **预训练:** 大规模, 弱关联的图文对
  - **多任务学习:** 高质量数据 (VQA, Caption, Grounding, OCR)
  - **指令微调:** 基于指令遵循数据 (多模态 + 文本)



Qwen-vl: A frontier large vision-language model with versatile abilities，2023

MiniGPT4-v2：额外引入任务指示符

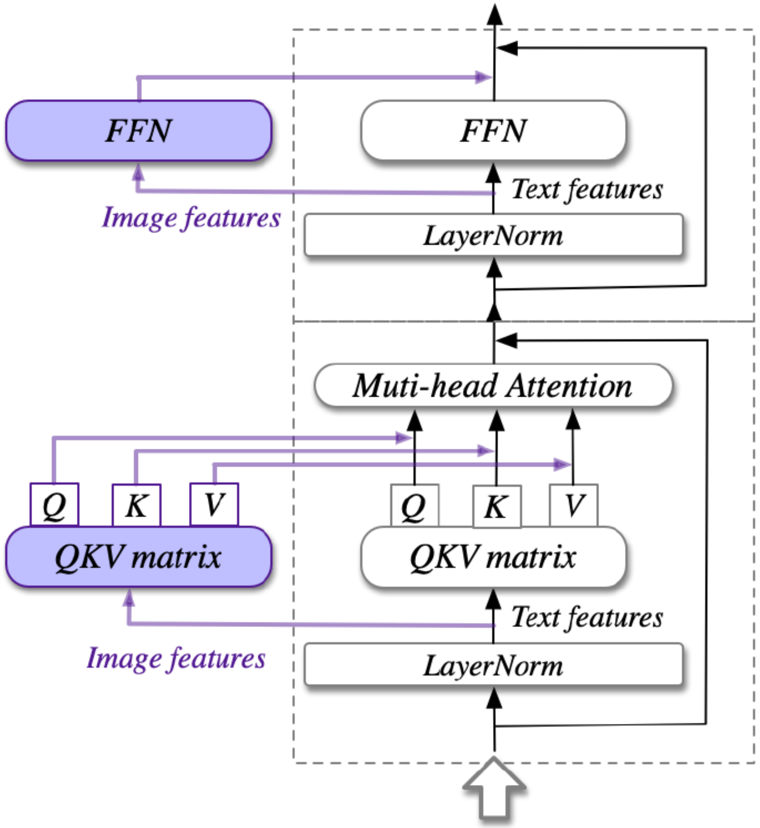
- **视觉编码器:** EVA-ViT (448 px)
- **LLM：** LLaMA-2
- **连接模块:** 拼接邻接的4个tokens进行Linear
- **3-阶段训练:**
  - **引入任务指示符:** [vqa], [grounding], [refer]..
  - **预训练:** 主要学习弱关联的图文对
  - **多任务学习:** 仅细粒度的任务**数据** (VQA + Caption + Grounding)
  - **指令微调:** 指令遵循数据（多模态 + 文本）



Minigpt-v2: large language model as a unified interface for vision-language multi-task learning，2023

CogVLM:引入视觉专家模块

- **视觉编码器:** EVA2-CLIP-E (490 px)
- **LLM：** Vicuna-7B-v1.5 + 视觉专家模块;
- **连接模块:** MLP层
- **预训练：**
  - LAION + 基于 Kosmos2 构造的 grounding 数据
- **SFT对齐训练：**
  - LLaVA, LLaVAR, LRV-Instruction, 非公开数据
- 下游任务上的 **Fine-tuning:**
  - Captioning, VQA, visual grounding



CogVLM: Visual Expert for Pretrained Language Models，2023

小结

- 真正的多模态模型必然是全模态支撑的
- 以大语言模型作为大脑是目前的主流架构
- 编码端可以进行语义对齐
- 语义空间引入其他模态的词汇，扩充输出可能性
- 解码端引入其他工具，完成输出

- **训练数据的生成：多模态混合的数据样本还是远远小于文本模态**

目录

- ChatGPT之前的视觉语言预训练
- 大视觉语言模型的架构和训练
- 大视觉语言模型的评测
- 大视觉语言模型的能力扩充
- **大语言模型支撑的具身智能（视觉导航）**

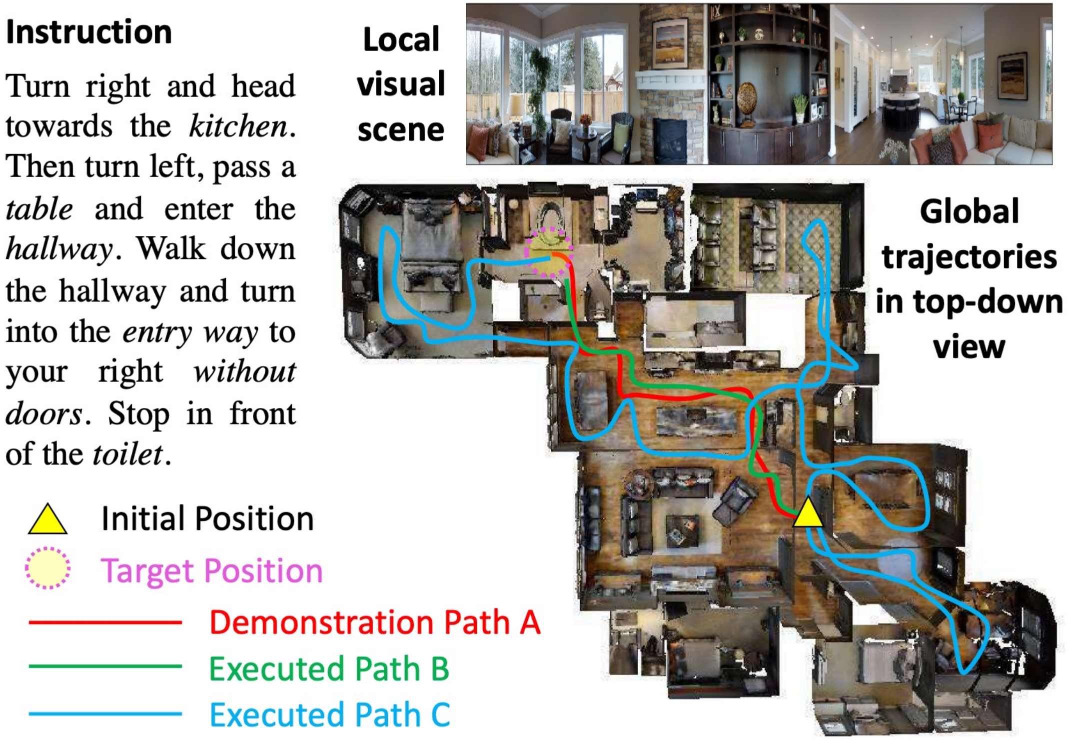
具身智能

目前，通过具身智能来解决任务的研究要求 AI 具备以下的能力

- 观测 (usually in an egocentric view)
- 交流 (via texts or audios)
- 推理 (understand surroundings and plan)
- 行动 (through motor controls or high-level actions).

➡ 视觉语言导航
(Vision-and-Language Navigation)

构建真实机器人的一个好的原型



一个典型的VLN样例

Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments，2018

Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation，2019.

基于大语言模型的机器视觉导航

- LLM展示了强大的文本理解和生成能力。
- LLM具有良好的泛化能力，展示出卓越的零样本性能。

Title: Star's Tux Promise Draws Megyn Kelly's Sarcasm
Subtitle: Joaquin Phoenix pledged to not change for each awards event
Article: A year ago, Joaquin Phoenix made headlines when he appeared on the red carpet at the Golden Globes wearing a tuxedo with a paper bag over his head that read, "I am a shape-shifter. I can't change the world. I can only change myself." It was a promise to not change to fit into the Hollywood mold: "I think that's a really special thing, to not change yourself. I think it's a really special thing to say, 'This is what's inside of me, I'm proud of it, and I'm not going to be ashamed because of the way that someone else thinks I should be.'" Now, it's the Oscars, and Phoenix is at it again. But this time, his publicist is saying he'll be wearing a tux no matter what.
Megyn Kelly was not impressed, and she let him have it on The Tonight Show. "You know, I feel like, I feel like you could have worn the tux," she says. "But you're saying you're a shape-shifter. I don't know if you can change your tux, but you can change your mind. You can change your mind. You can change your mind." Phoenix says he did, but it didn't stick. "I was like, 'Okay, I'm going to wear a tuxedo to this thing.' And then I thought, 'I don't want to wear a tuxedo to this thing.'" Kelly goes on to encourage him to change his mind again, but Phoenix says it's too late: "I'm committed to wearing this."

Figure 3.15: The GPT-3 generated news article that humans found the easiest to distinguish from a human written article (accuracy: 61%).

Sample question using prompt template

The following are multiple choice questions (with answers) about medical knowledge.
Question: A 40-year-old woman has had hypercalcemia for 1 year and recently passed a renal calculus. Serum parathyroid hormone and calcium concentrations are increased, and serum phosphate concentration is decreased. Parathyroid hormone most likely causes an increase in the serum calcium concentration by which of the following mechanisms?
(A) Decreased degradation of 25-hydroxycholecalciferol
(B) Direct action on intestine to increase calcium absorption
(C) Direct action on intestine to increase magnesium absorption
(D) Increased synthesis of 25-hydroxycholecalciferol
(E) Inhibition of calcitonin production
(F) Stimulation of 1,25-dihydroxycholecalciferol production
Answer:(F

Zero-shot medical analysis (GPT-4)

➡ 使用大模型来构建更好的VLN智能体！

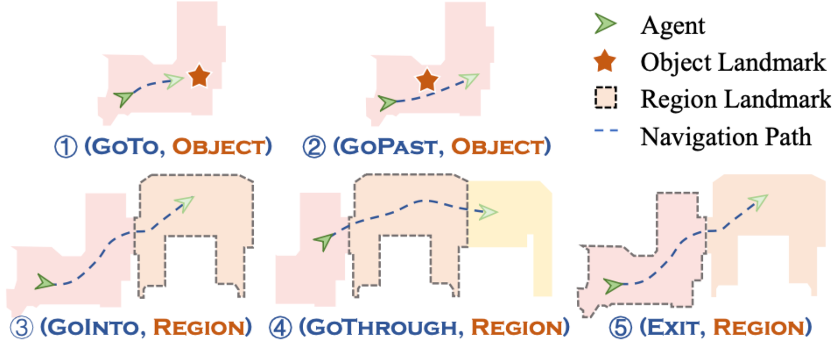
Language models are few-shot learners，2020

A comprehensive capability analysis of gpt-3 and gpt-3.5 series models，2023

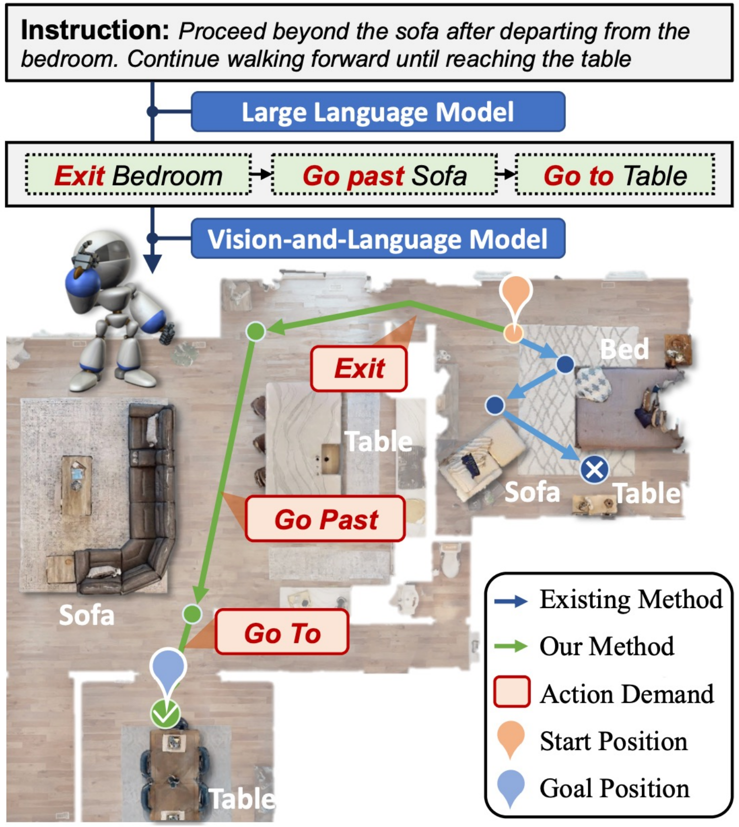
A^2 Nav: 大语言模型作为指令解析器

利用LLM优越的文本理解和推理能力将指令分解成若干个子任务。

- 每个子任务包含
  - 一个地标
  - 一个与该地标相关的特定动作



- 动作感知的导航策略:
  - 零样本物体导航器 (Zero-Shot Object Navigator, ZSON)
  - 每个ZSON负责一个特定类型的子任务

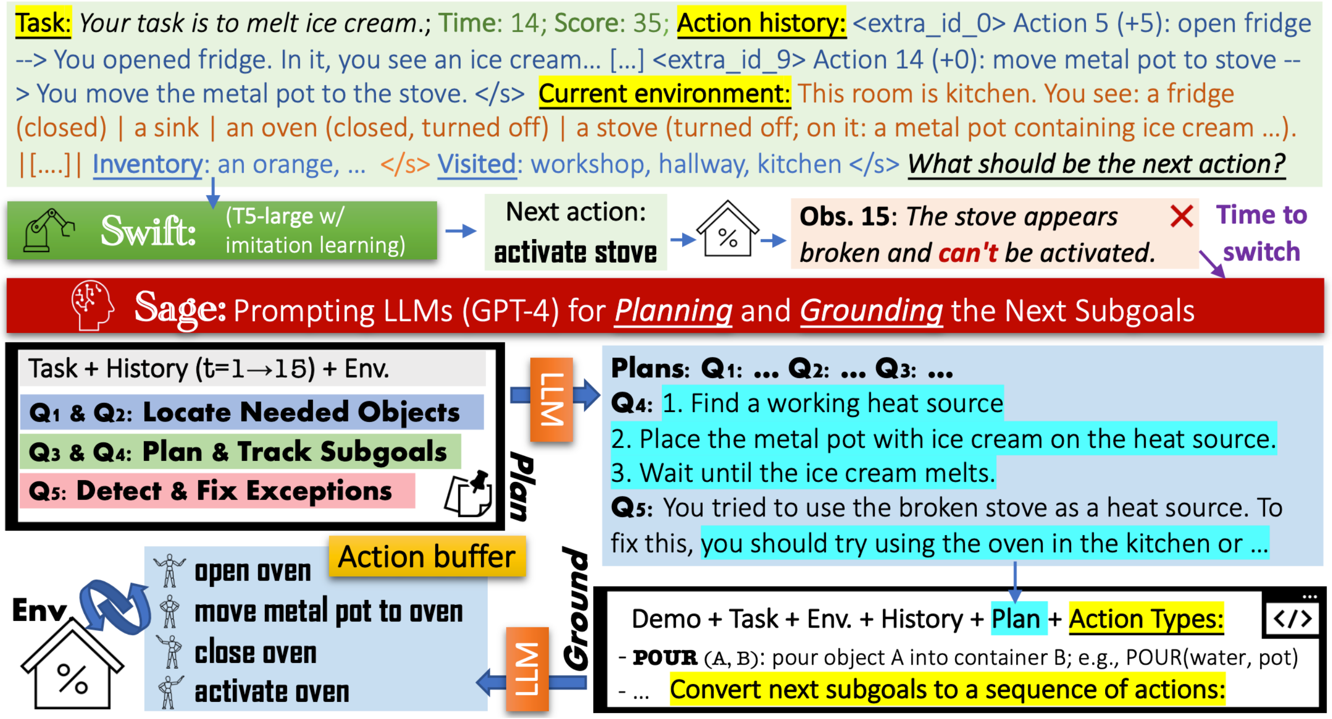


A^2 Nav: Action-Aware Zero-Shot Robot Navigation by Exploiting Vision-and-Language Ability of Foundation Models，2023

SWIFTSAGE：引入大语言模型进行决策反思

让LLM对失败进行反思，然后进行纠错（给出下一步的目标）

- SWIFT 模块 (T-5 Large):
  - 通过模仿学习进行简单决策。
- SAGE 模块 (GPT-4):
  - 纠正SWIFT模块的错误决策。



- SAGE的两阶段策略:
- **规划**: 通过回答问题对历史和任务执行情况进行总结
- **匹配**: 促使LLM专注于下一步目标，并将其转换成一系列行动。

1) There are five consecutive time steps with zero reward ($\sum_{i=t-5}^{t-1} R_i = 0$).
2) The SWIFT's prediction for the next action (A_t') is invalid in the current environme
3) A_t' can result in a critical decision, such as giving the final answer for the experime
4) The observation of A_t' suggests that an exception is encountered.

何时从SWIFT模块切换到SAGE模块？

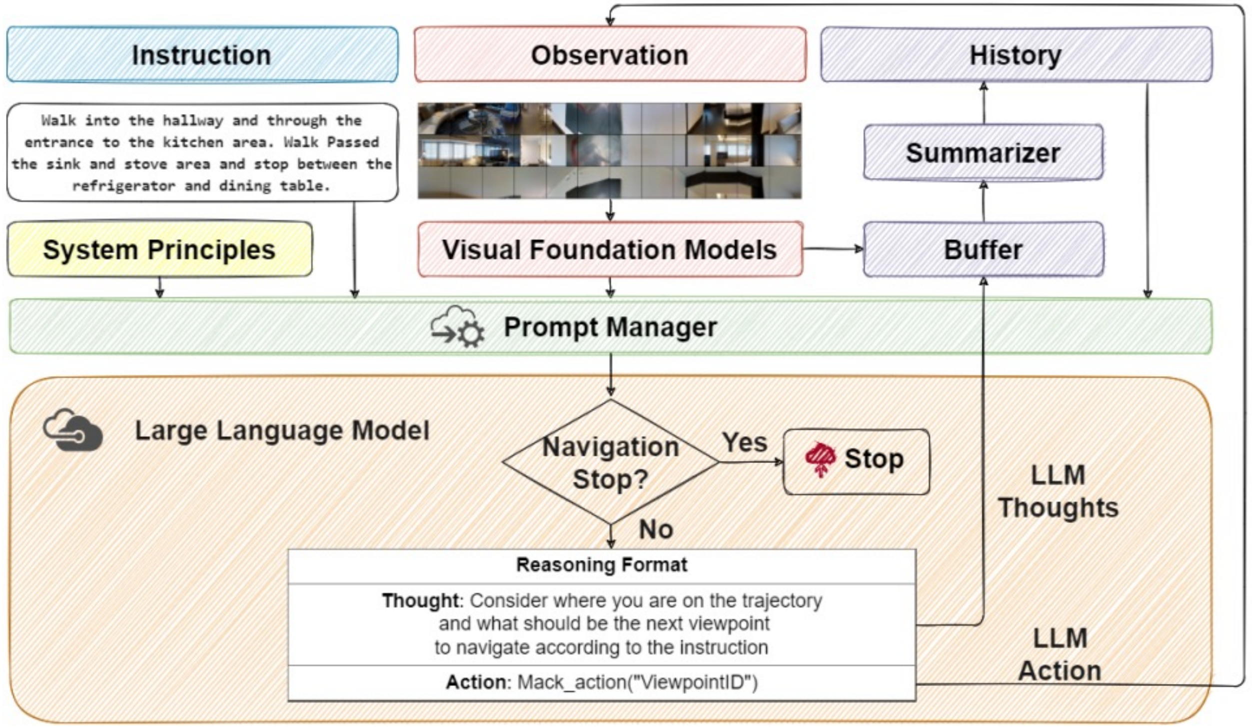
SwiftSage: A Generative Agent with Fast and Slow Thinking for Complex Interactive Tasks，2023

NavGPT：大语言模型作为动作决策模块

- 一个零样本视觉语言导航框架，使用大语言模型作为动作决策模块。
- 以文本的形式表示当前的视觉观测和过去的历史轨迹
- 在决策时，使用 chain-of-thought，融合思维（推理）和行动（决策）

模型架构

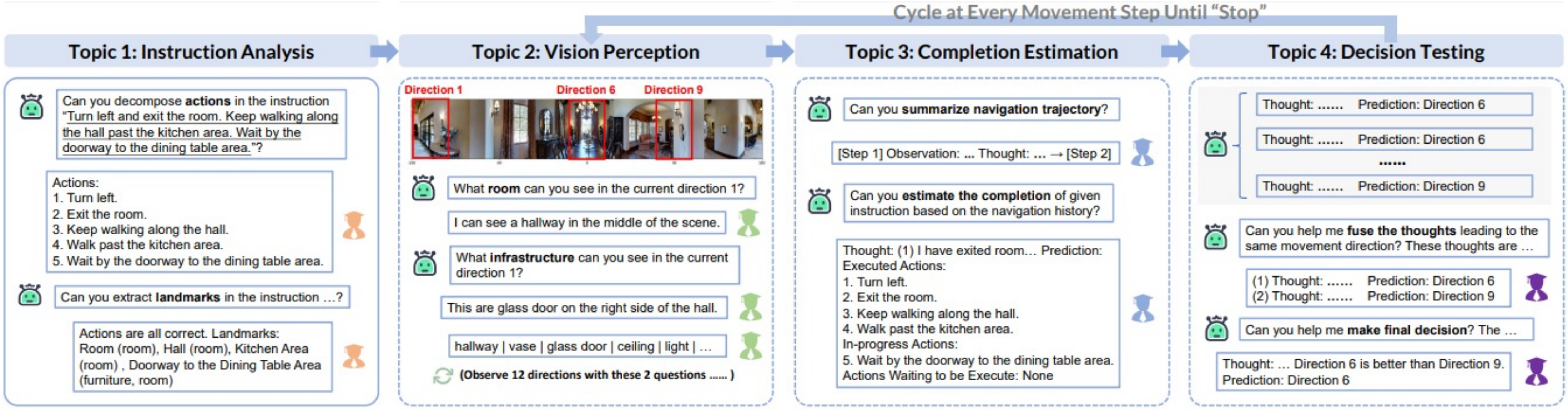
- **Visual Foundation Models:**
  - 视觉描述: BLIP-2
  - 物体检测: Faster-RCNN
  - 方位归纳: GPT-3.5
- **History Summarizer:** GPT-3.5
- **LLM:** GPT-4



NavGPT: Explicit Reasoning in Vision-and-Language Navigation with Large Language Models，2023

DiscussNav：大语言模型作为领域专家

- 将导航过程转换为基于多位专家讨论的决策过程（文本形式）。



GPT-4	InstructBLIP + RAM	ChatGPT & GPT4	ChatGPT & GPT4
extract actions and identify landmark as well as their types	observe the scene-level vision and objects	summarize the trajectory; estimate what actions have been executed	fuse given thought from multi-experts; select the final action

Discuss Before Moving: Visual Language Navigation via Multi-expert Discussions，2023

DiscussNav：大语言模型作为领域专家

RESULTS ON R2R VALIDATION UNSEEN SPLIT.

Training Schema	Method	TL	NE↓	OSR↑	SR↑	SPL↑
Train Only	Seq2Seq [1]	8.39	7.81	28	21	-
	Speaker Follower [5]	-	6.62	45	35	-
	EnvDrop [6]	10.70	5.22	-	52	48
Pretrain + Finetune	PREVALENT [7]	10.19	4.71	-	58	53
	VLN ⟳ BERT [8]	12.01	3.93	69	63	57
	HAMT [9]	11.46	2.29	73	66	61
	DuET [10]	13.94	3.31	81	72	60
Zero-shot	DuET (Init. LXMERT [51])	22.03	9.74	7	1	0
	NavGPT [11]	11.45	6.46	42	34	29
	DiscussNav (Ours)	9.69	5.32	61	43	40

REAL ROBOT EXPERIMENTS IN INDOOR SCENE.

Training Schema	Method	SR↑
Pretrain + Finetune	DuET [10]	0
Zero-shot	NavGPT [11]	10
	DiscussNav (Ours)	**25**

- 在zero-shot 设定下，DiscussNav相比其他方法的性能更好
- 但未能超过专有的小模型

- Sim-2-Real Transfer实验：结果表明 NavGPT 单轮 CoT 的能力受限，
- 而 DiscussNav 在理解指令、观察细粒度地标和做出决策方面表现更好

Discuss Before Moving: Visual Language Navigation via Multi-expert Discussions，2023

使用大语言模型导致信息丢失&误差累积

- 将图像表示为文本将会丢失许多细节。

Image Captioning
(GPT-4V)

No chairs

No direction

Useless information

A modern living room with cream walls, a beige sofa, and a black fireplace. A decorative mirror, indoor plants, and a dark wooden staircase with black railings enhance the space. Natural light streams in through ample windows.

Go to the left of the bed and out of the bedroom. Then go down the hall and make a right at the top of the stairs, go past the stairs and go a couple steps into the bedroom and wait there.

Instruction Parser
(GPT-4)

1.Identify Bed
2.Move to Left of Bed
3.Exit Bedroom
4.Enter Hallway
5.Proceed Down Hallway
6.Turn Right at Top of Stairs
7.Go Past Stairs
8.Enter Second Bedroom
9.Wait

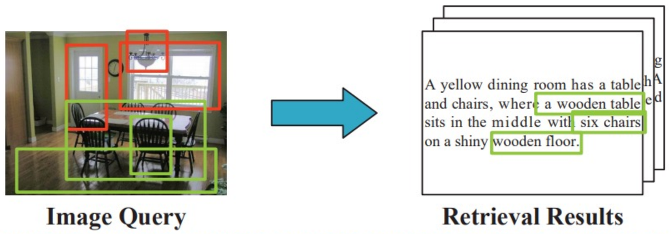
hallucination

- 如果任何模块出错，错误将会累积并传播到策略决策模块中。

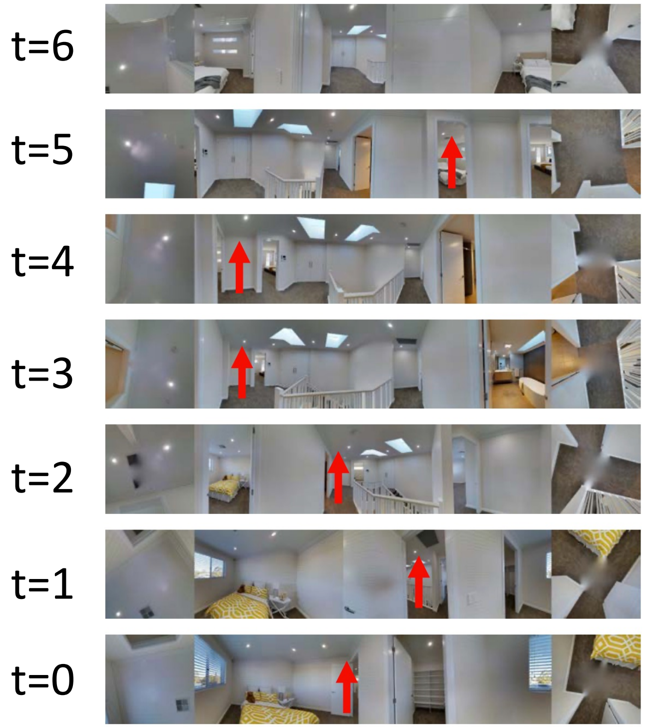
如何在具身智能任务中使用多模态大模型?

- 使用多模态模型 (LVLMs) 来理解视觉和语言已经是常用的做法。
- 然而，LVLMs并非为具身人工智能任务而设计。

➢ Image Text Retrieval



➢ Vision-based Text Generation

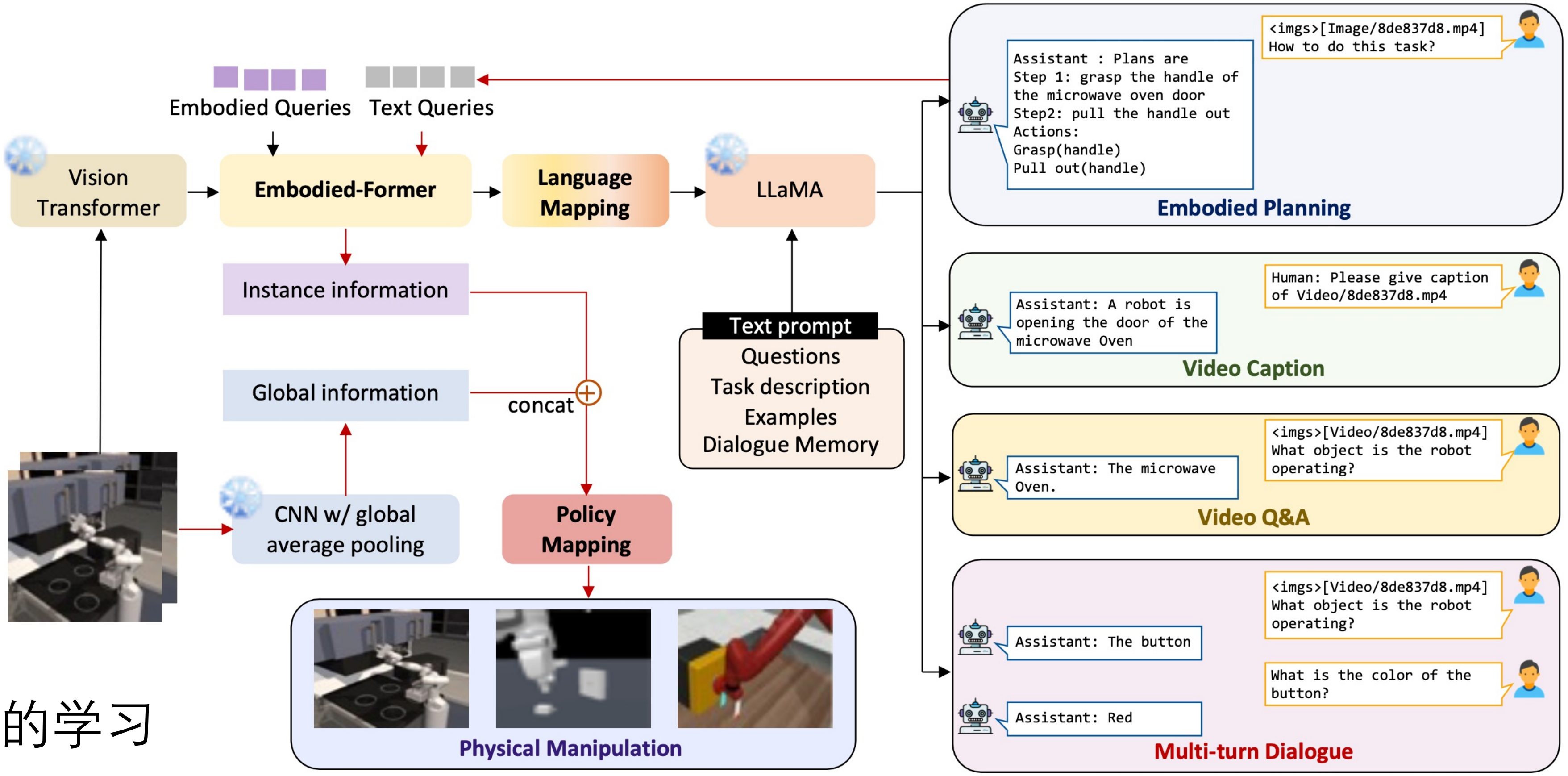


单图的、单步的推理、理解、生成任务

多图的、多步的推理、时间顺序、空间方位关系

EmbodiedGPT: 采用大规模预训练

对LVLMs进行预训练，以构建一个端到端的具身智能基础模型。



三阶段预训练:

1. 基础认知能力的学习
2. 复杂理解和推理能力的学习
3. 执行具身任务能力的学习

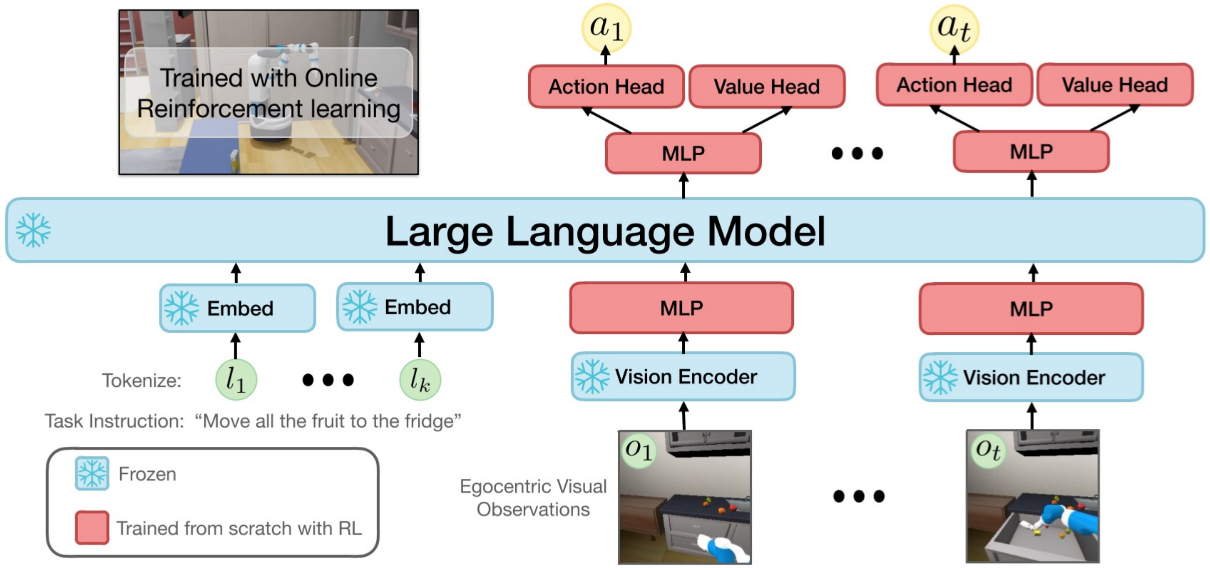
Embodiedgpt: Vision-language pre-training via embodied chain of thought，2023

LLaRP：通过强化学习训练 LVLMs

- 通过在线强化学习将LVLMs训练为适配多种具身任务的视觉-语言策略

<u>模型架构</u>

- **Visual Encoder:** VC-1
- **LLM:** LLaMA-7B V1
- **Connection:** Linear
- **Action Decoder:** MLP with ReLU



强化学习采用DD-PPO算法。

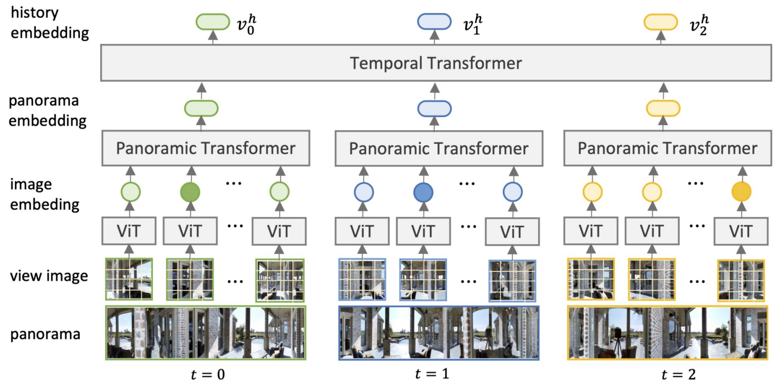
模型在一个新的基准数据集 Language Rearrangement 上进行训练，该数据集包含 150k 个训练任务和 1k 个测试任务，例如复杂的操控、导航和探索。
由于数据集部署在Habitat 2.0模拟器中，智能体动态的与环境进行交互，从而进行在线强化学习。

Large Language Models as Generalizable Policies for Embodied Tasks，2023

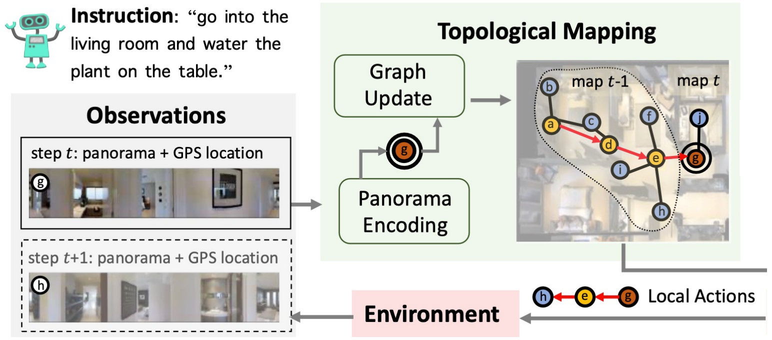
挑战: 对长期依赖性的建模

导航是一个顺序决策问题，因此历史信息至关重要。

Transformer-based Agent



History as sequential observation features



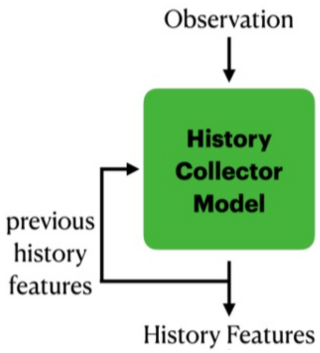
History as Topological graph

LLM-based Agent

Instruction: Walk into the house, and go through the kitchen. Leave the kitchen through the door on the right of the large black bookcase. Go down the stairs. Stop on the third stair from the top.
Init observation: …with hardwood floors and a dining table…
Thought: First, I need to go through the kitchen. The kitchen is in the front left direction, so I should move to the viewpoint '68aafa779b9c41eca16156cfddcedd2b'.
Action: action_maker("68aafa779b9c41eca16156cfddcedd2b")
Observation: …A dining room and kitchen with hardwood floors.
Thought: I am now in the kitchen. Next, I need to find the door on the right of the large black bookcase. I should move to the viewpoint 'e685b29863794f00b09c23e1536412d8' to get a better view of the kitchen and find the door.
Action: action_maker("e685b29863794f00b09c23e1536412d8")
Observation: …A kitchen with white cabinets, varying flooring and counter top materials, and white ceiling lights.
Thought: I can see the kitchen clearly now, but I still can't find the large black bookcase and the door on its right. I should move to the viewpoint '5b01ffae96874e60a9decd1fbac7e4ee' to explore the kitchen further.
Action: action_maker("5b01ffae96874e60a9decd1fbac7e4ee")
Observation: …A kitchen with a marble top island, counter top, and fireplace.
Thought: I found the large black bookcase in the rear left direction, and the door on its right. I should move to the viewpoint 'e62661ec94d04adda7fee9a81d429d2b' to get closer to the door.
Action: action_maker("e62661ec94d04adda7fee9a81d429d2b")
Observation: …A hallway with hardwood floors and a door leads to a room with a bookcase and a ceiling fan, while another hallway features a bookcase and a bookshelf.
Thought: I am now closer to the door on the right of the large black bookcase. I should move to the viewpoint 'ea374bef649549d8a12bb06d67ed206e' to go through the door and reach the stairs.
Action: action_maker("ea374bef649549d8a12bb06d67ed206e")
Observation: …A hallway with hardwood floors, a staircase, a painting.
Thought: I am now in the hallway with the staircase. I need to go down the stairs and stop on the third stair from the top. I should move to the viewpoint 'a8311575bf3d485a81af8350aa25bddb' to start going down the stairs.
Action: action_maker("a8311575bf3d485a81af8350aa25bddb")
Observation: …A view of a staircase with a black railing, a hallway with a white railing and art, and a bathroom with a white ceiling and a light fixture.
Thought: I am now on the stairs. I should move to the viewpoint '998dfb266ae84564b136c862fae09dd5' to go down the stairs and stop on the third …

History as Long Context

LVLM-based Agent



History as summarized features

Any better methods to modelling the long-term dependency of this task ?

History aware multimodal transformer for vision-and-language navigation，2021

挑战：目前缺乏对模型空间理解能力的探索

Model	Space-based Perception		Spatial Relation Judgment				Avg.	
	CLEVR		VSR		MP3D-Spatial			
	Acc	Instability	Acc	Instability	Acc	Instability	Acc	Instability
Generation Evaluation								
BLIP-2$_F$	42.67	0.28	46.95	0.21	39.87	0.32	43.16	0.27
InstructBLIP$_F$	44.84	0.39	52.37	0.25	**41.01**	0.37	**46.07**	0.34
InstructBLIP$_V$	**46.32**	0.51	52.37	0.49	34.59	0.50	44.43	0.50
LLaVA$_V$	19.01	1.24	40.00	0.88	27.19	1.13	28.73	1.08
LLaVA$_{L_2}$	36.52	0.61	**52.54**	0.21	34.67	0.64	41.24	0.49
MiniGPT4	33.74	0.84	36.44	0.81	33.62	0.84	34.60	0.83
mPLUG-Owl	27.48	1.01	28.81	0.97	24.23	1.04	26.84	1.01
PandaGPT	29.65	0.90	35.76	0.86	34.50	0.80	33.30	0.85
IB-LLM	31.45	0.96	40.00	0.94	35.22	0.83	35.56	0.91
LA-V2	21.39	1.05	23.05	1.04	27.06	1.01	23.83	1.03
mmGPT	22.26	1.13	28.98	1.01	29.30	0.98	26.85	1.04
Shikra	23.82	0.77	46.27	0.60	29.77	0.84	33.29	0.74
Lynx	40.58	0.68	45.76	0.66	34.38	0.78	40.24	0.71
Cheetor$_V$	24.72	1.03	35.76	0.77	31.21	0.88	30.56	0.89
Cheetor$_{L_2}$	29.10	0.77	40.85	0.69	33.53	0.73	34.49	0.73
BLIVA	30.64	0.85	35.25	0.61	34.12	0.59	33.34	0.68

ChatGPT-4：
在图片的左侧，有一个高大的不锈钢冰箱。冰箱上方是橱柜。还有一个内置烤箱，嵌入橱柜中。台面似乎是浅色的，与浅色调的地砖相得益彰。此外，还有一个红色圆柱形物体，可能是一个香料研磨器或厨房用具架。在它旁边，台面上放着一份可能是报纸或杂志的印刷品。

- LVLM对空间信息的理解较差（<50%）。
- 很少有模型报告它们在与空间建模相关的数据集上的表现。
- 缺乏使模型掌握空间理解能力的相关探索。

ReForm-Eval: Evaluating Large Vision Language Models via Unified Re-Formulation of Task-Oriented Benchmarks，2023

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