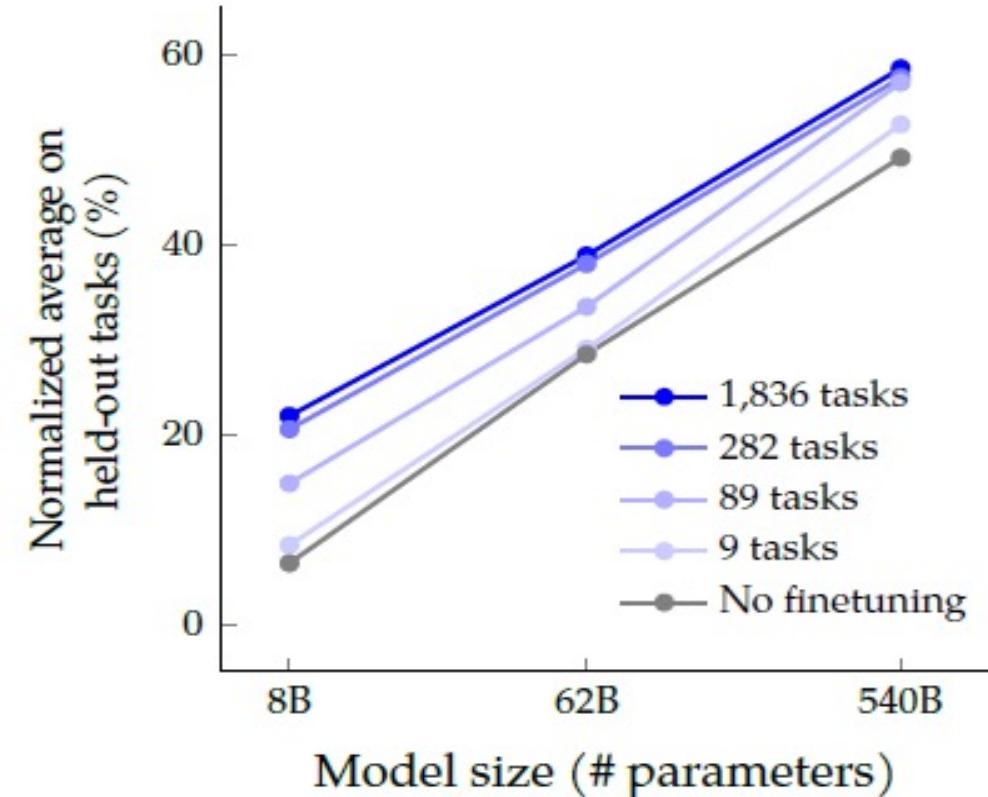


目录

- 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）

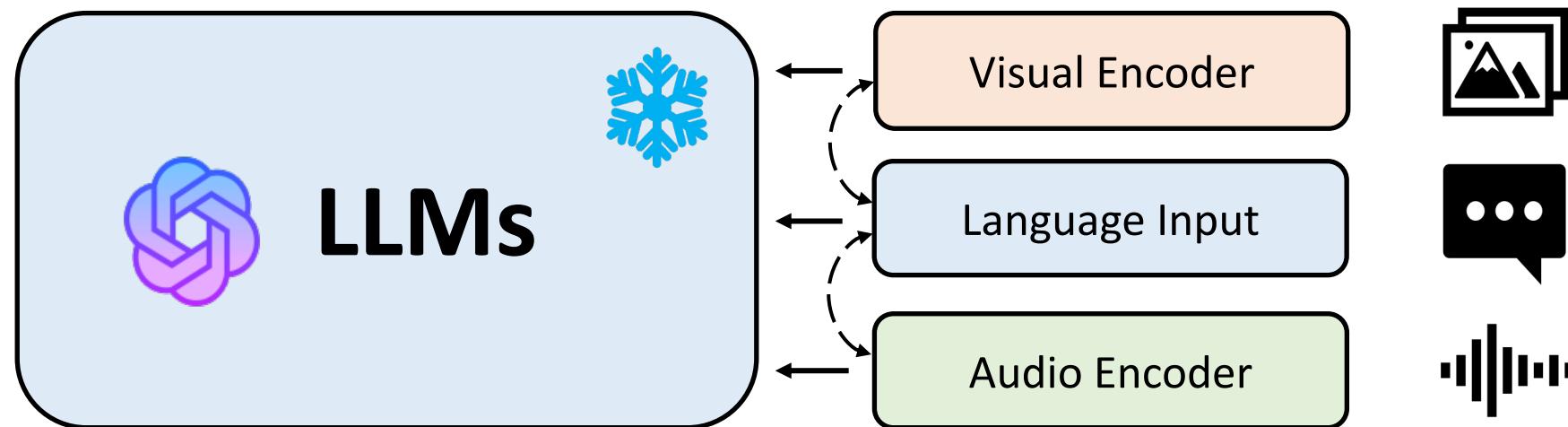
2022年底大语言模型横空出世

- 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？



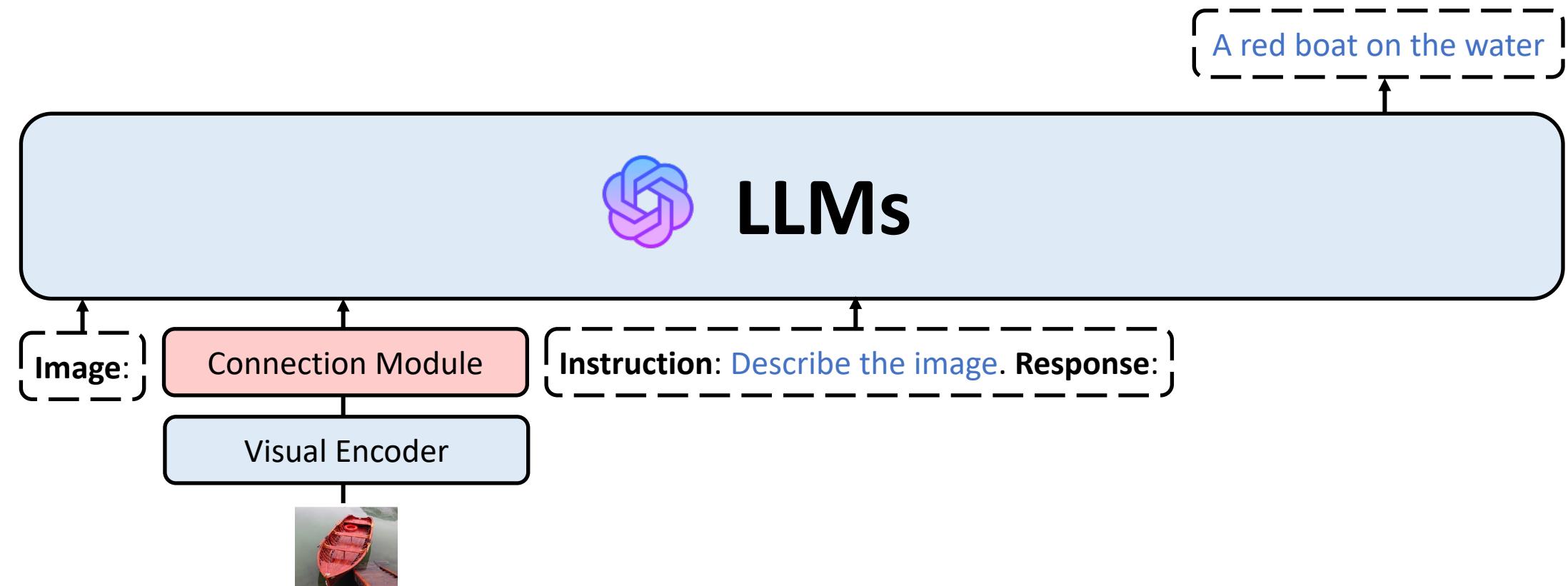
大语言模型如何帮助多模态模型构建？

- 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



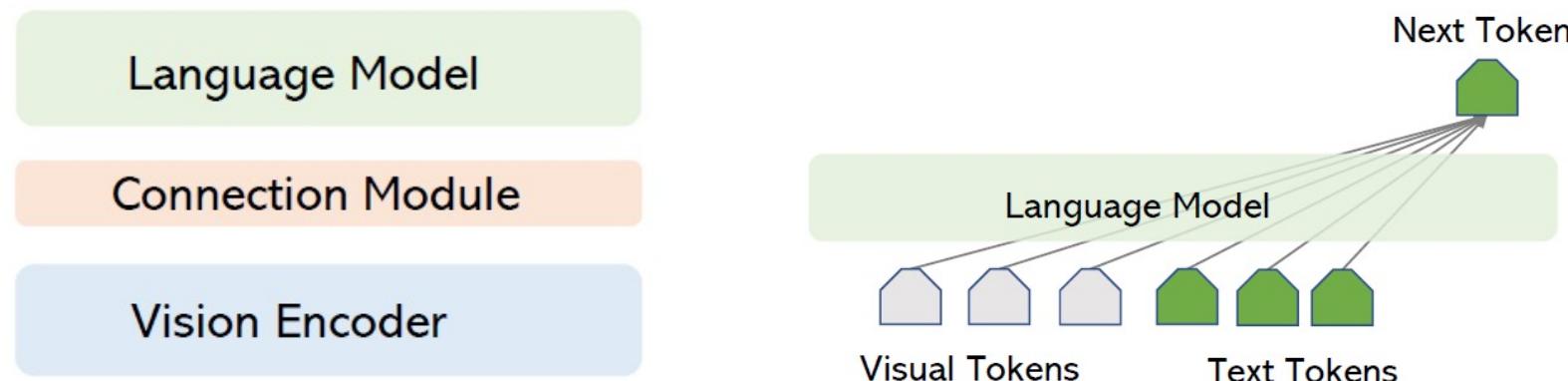
大规模视觉语言模型 (Large Vision Language Model)

- 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

Language Model

Connection Module

Vision Encoder

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 \mathbf{W}

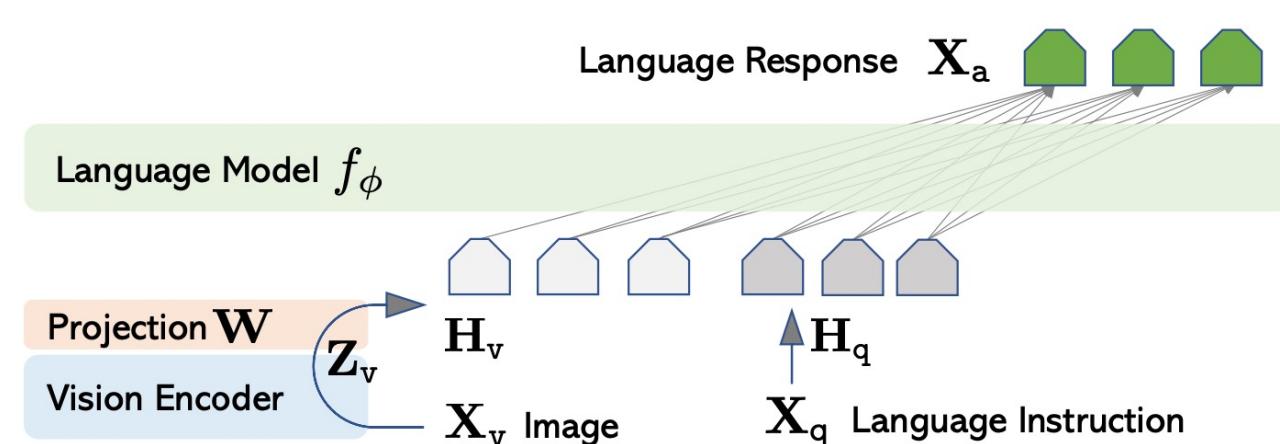
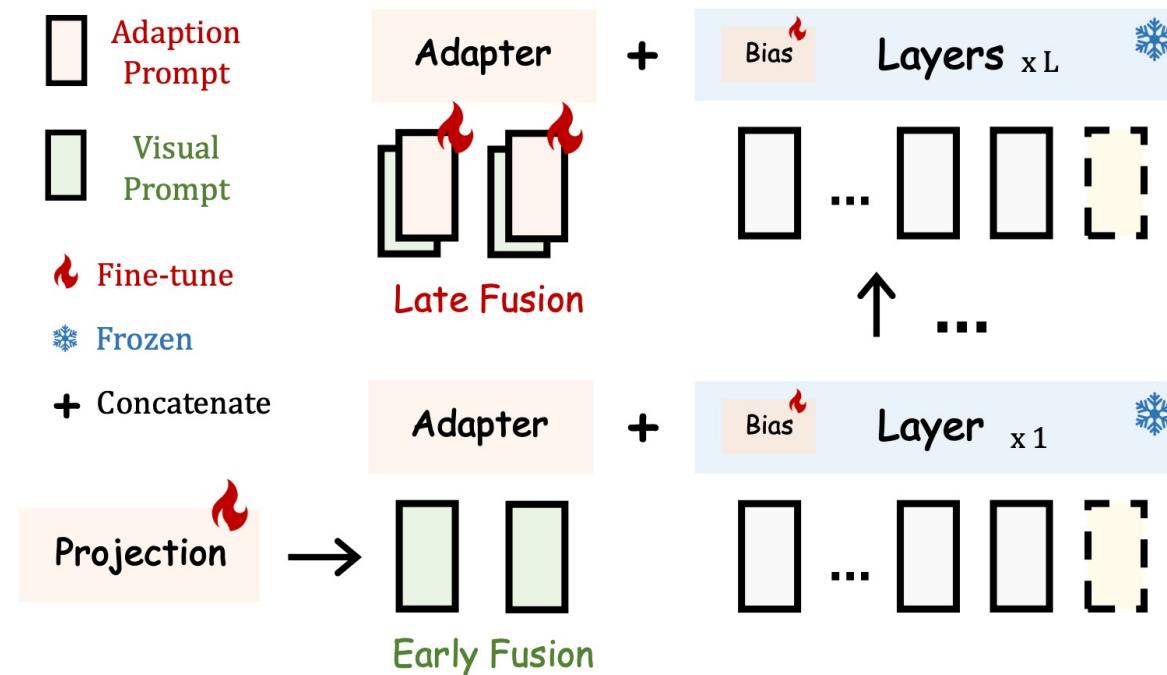


Figure 1: LLaVA network architecture.

- 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-Instruct 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

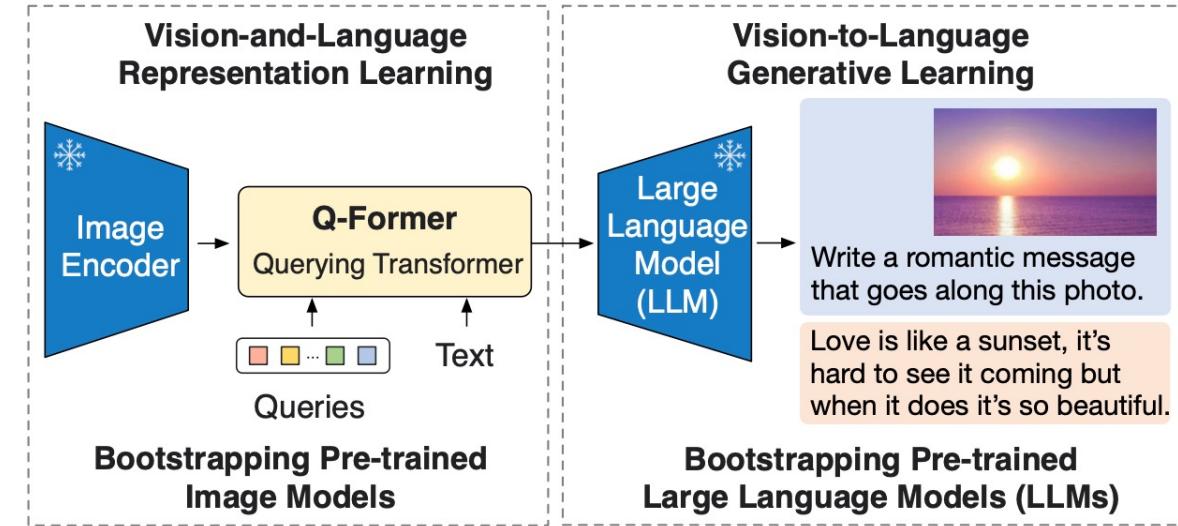
- 视觉编码器 : ViT-L/14
- 大语言模型 : LLaMA
- 连接模块 : Linear, Adapter, Gate
- 在预训练阶段更新连接模块的参数, 没有指令微调



BLIP-2：基于Q-Former的连接模块

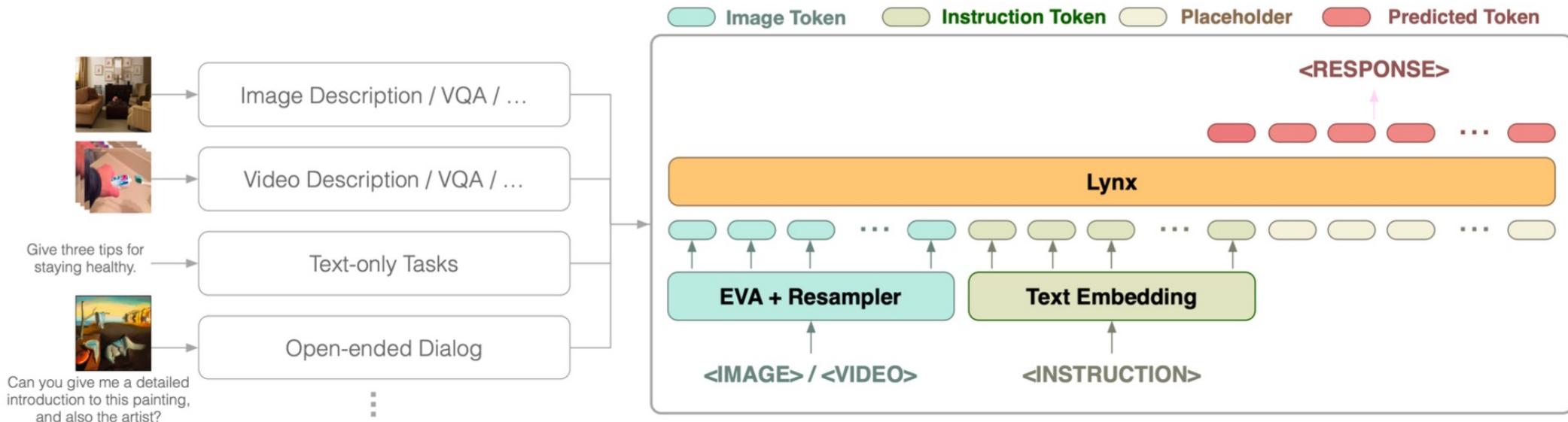
- 视觉编码器：ViT-L/14
- 大语言模型：Flan-T5, Vicuna
- 连接模块：Q-Former

- 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**，用于弥补模态间差距

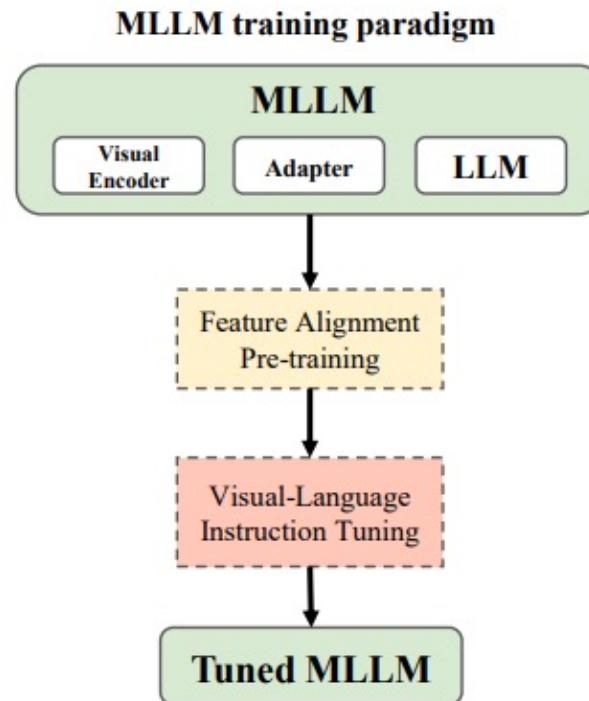


Lynx：基于感知器的连接模块

- 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



Pure Text Instruction

Instruction: <instruction>
Input: <text>
Response: <output>

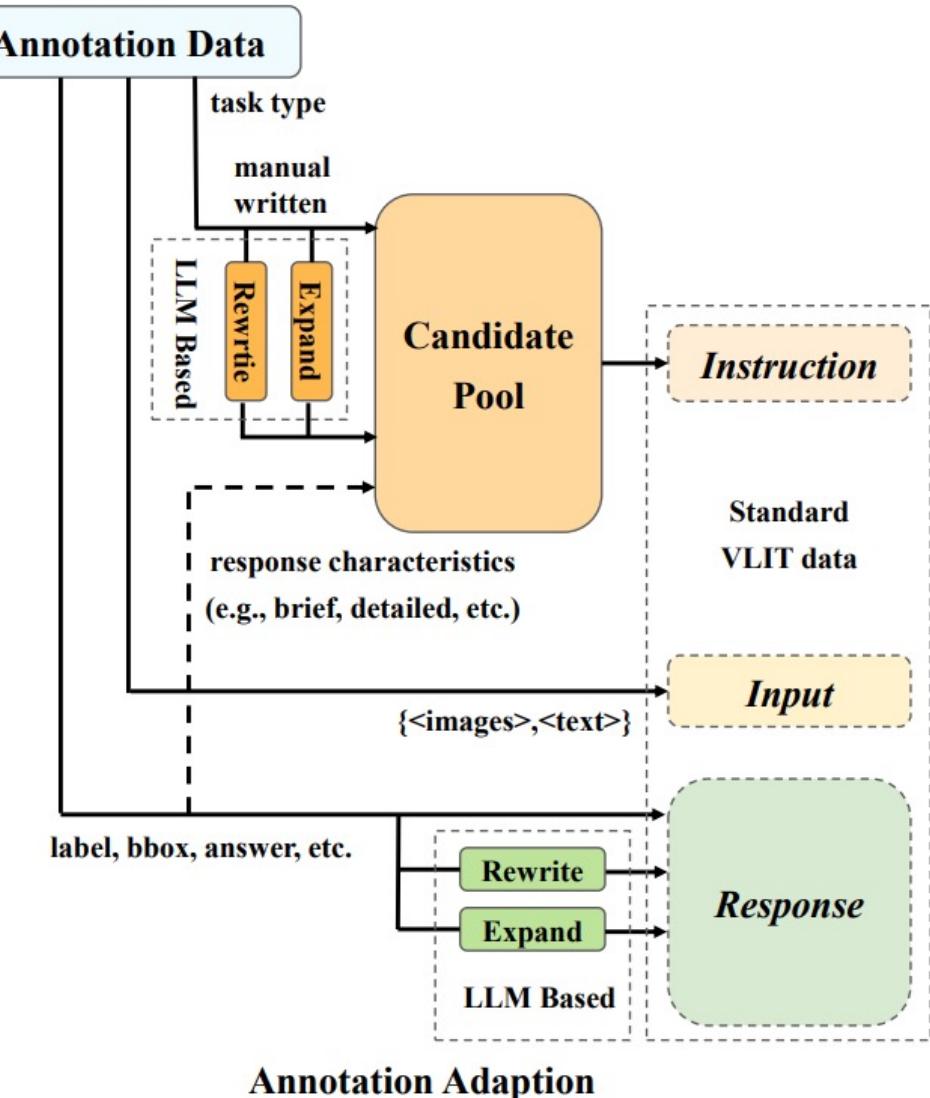
Visual-Language Instruction

<BOS> <context>
Instruction: <instruction>
Input: {<images>, <text>}
Response: <output> <EOS>

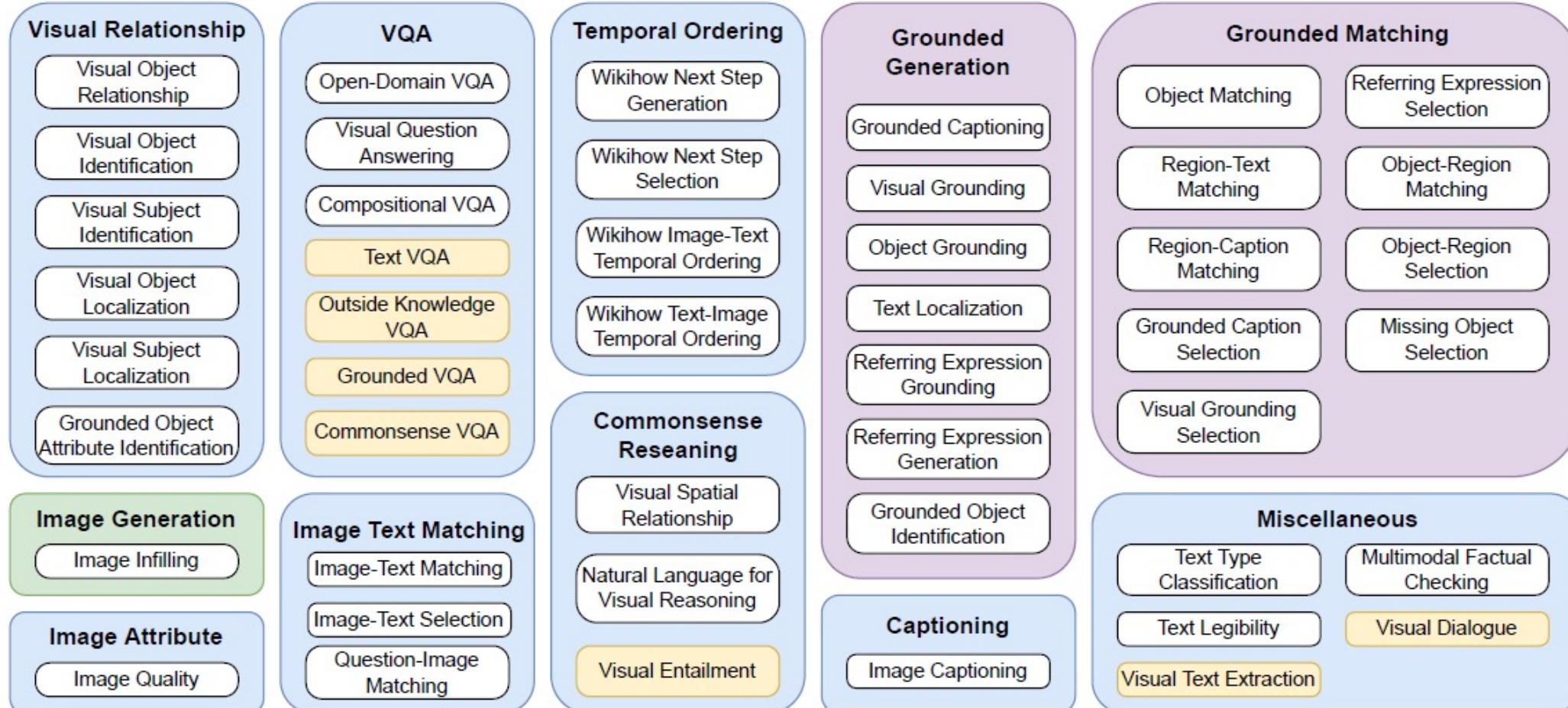
- 基于现有的标注: **MULTIINSTRUCT, InstructBLIP, MiniGPT-4, KOSMOS2.M3IT, etc**
- 自生成指令微调: **LLaVA, Syphus, LVIS-INSTRUCT4V, LMEye, LAMM, MosIT, etc**
- 模态数据混合: **Mplug-Owl, PandaGPT, LAVIN, BLIVA, etc**

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

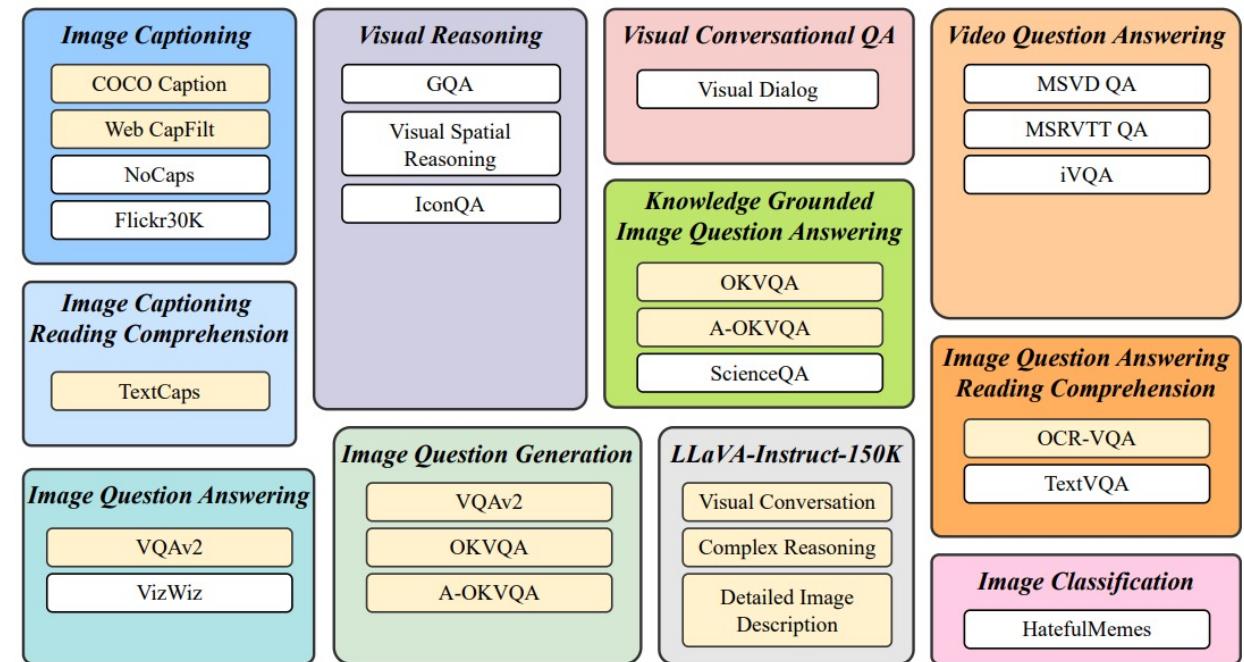


Multi-Instruct : 多模态指令微调数据集



InstructBlip: 基于现有标注数据的典型示例

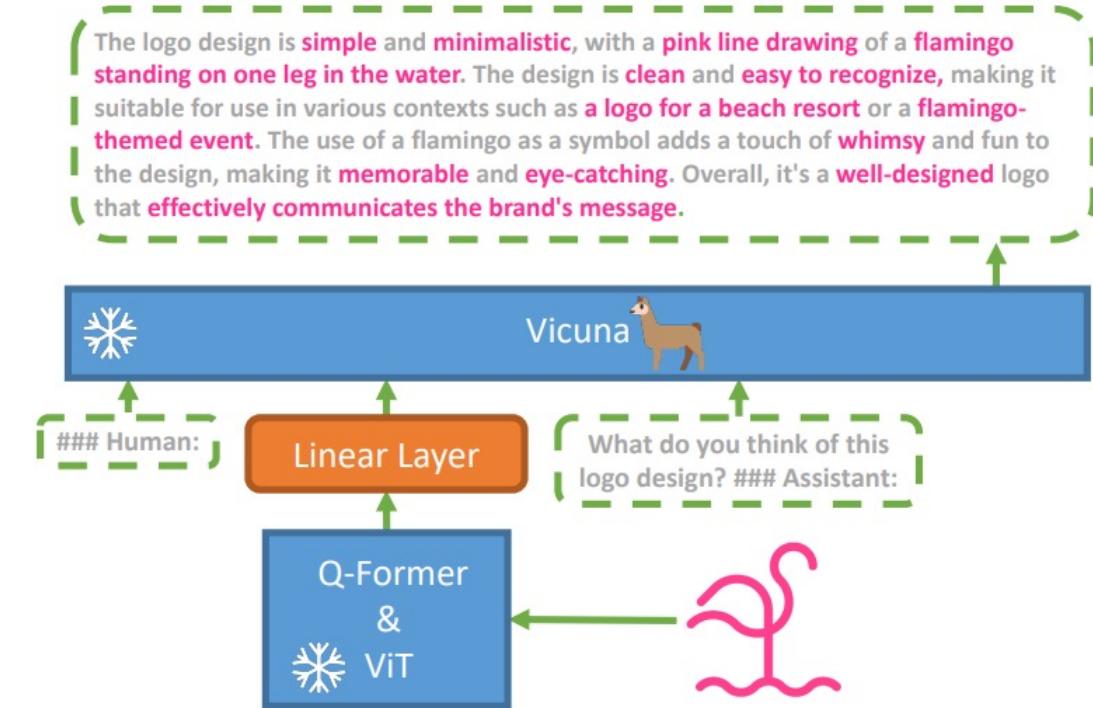
- 数据构造：收集CV领域11中不同任务的26个公开数据集
- 对于每个任务，**人为设计**了10-15个自然语言指令模板
- 该数据集用于**InstructBlip2**模型的指令微调训练



MiniGPT-4：基于现有标注数据的典型示例

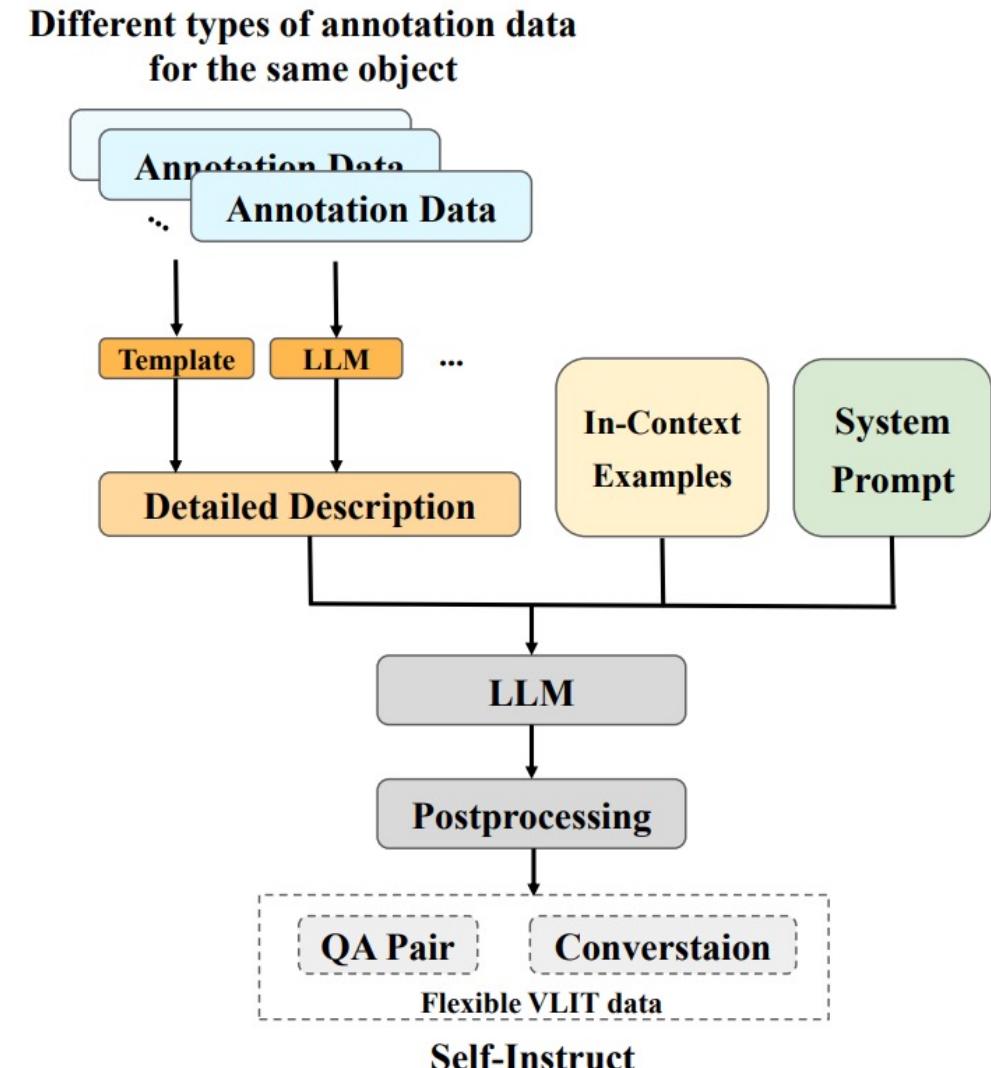
- 使用第一阶段预训练得到的模型来生成图像的初步描述
- 调用ChatGPT 优化描述
- 手动验证图像-描述的一致性

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



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



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，参考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.

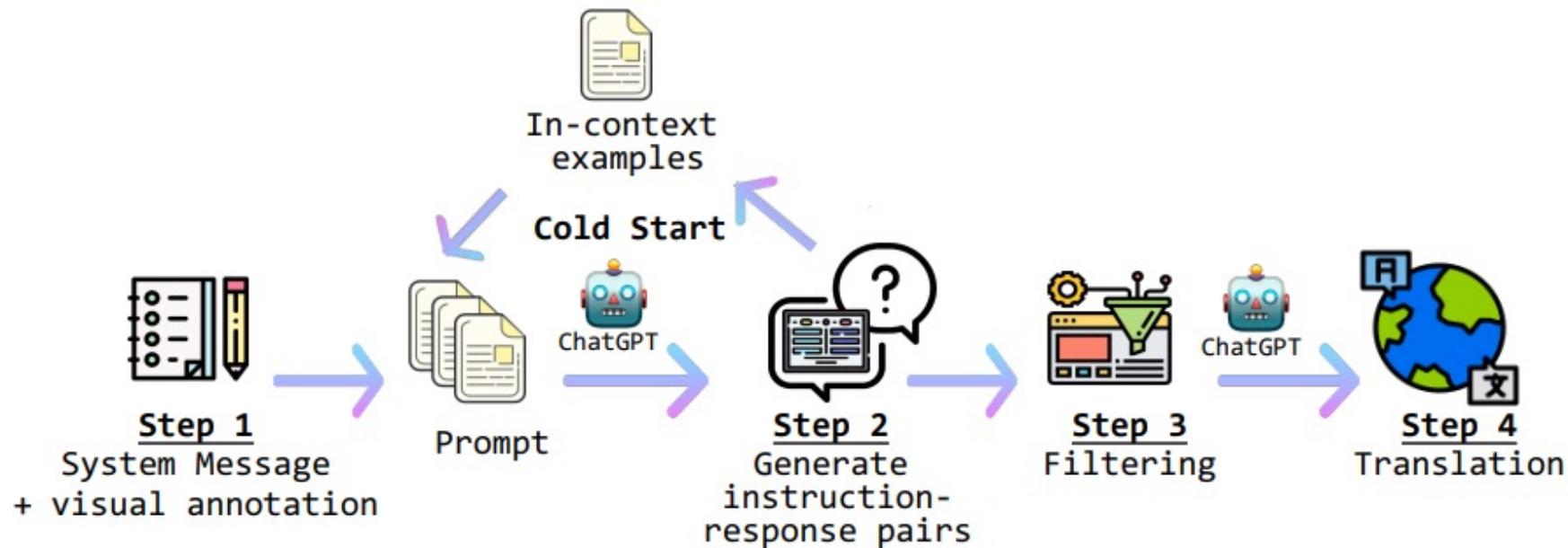
Other cars are parked around the main SUV, with one car positioned behind it and slightly to the left, another behind and slightly to the right, and the third car further behind on the right side.

Response type 3: complex reasoning

Question: What challenges do these people face?

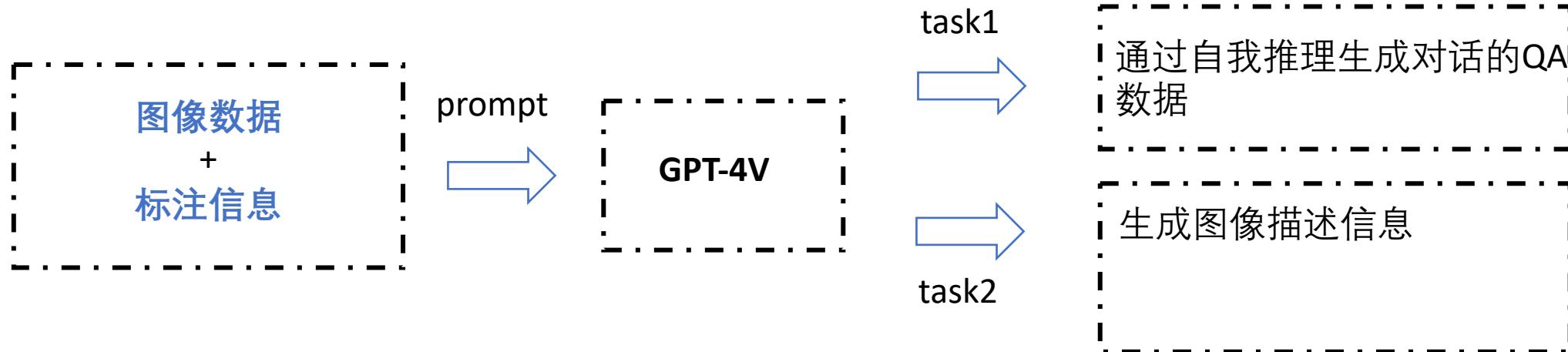
Answer: In the image, a group of people is standing outside a black SUV in a parking area, surrounded by various pieces of luggage, including suitcases and backpacks. They are facing the challenge of fitting all their luggage into the black SUV. There are multiple suitcases and backpacks to be packed, 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

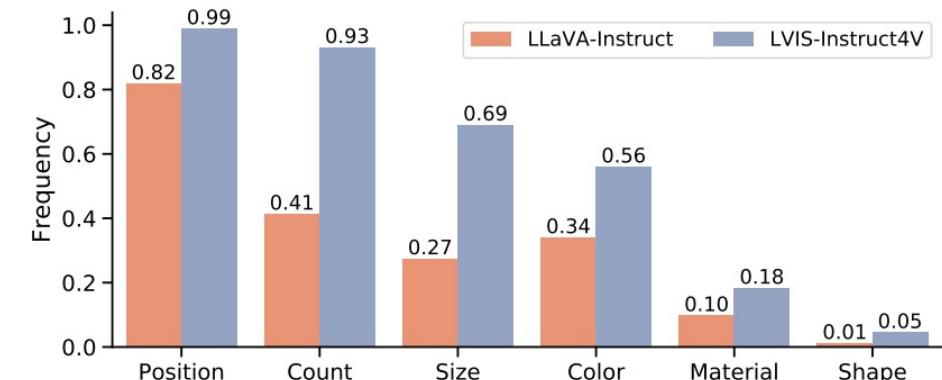


- 实施冷启动(cold start)的方法，将生成的指令响应对作为上下文示例重新输入到GPT，以此反复优化上下文示例，再将示例作为模板进行后续步骤

LVIS-Instruct4V : 利用GPT4V生成数据



- 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地结果输出。



LVIS-Instruct4V与LLaVA-instruct数据集效果对比

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 Qiu¹, Pengcheng Lu¹, Tao Wang¹, Zhongyu Wei²
¹ByteDance Inc ²Fudan University ³Beijing University of Posts and Telecommunications ⁴Chongqing 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 abilityY. 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 abilityY

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 abilityY

Ruipu Luo*, Ziwang Zhao*, Min Yang* (*Equal Contribution)



Generated by [stablecog](#) via "A cute llama with valley"

<https://github.com/RupertLuo/Valley>

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?



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.



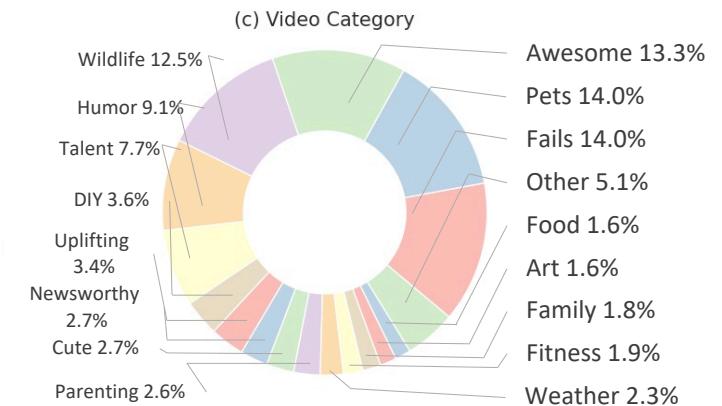
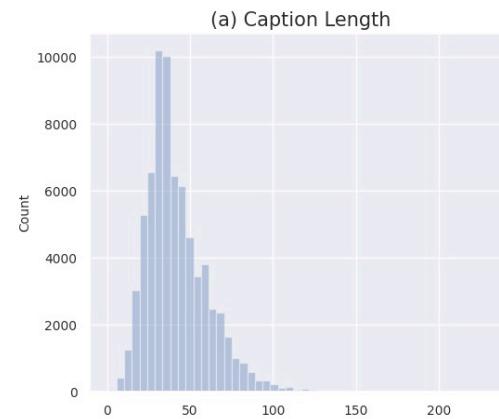
Illustrate the video through a descriptive explanation, painting a vivid picture of its content.



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

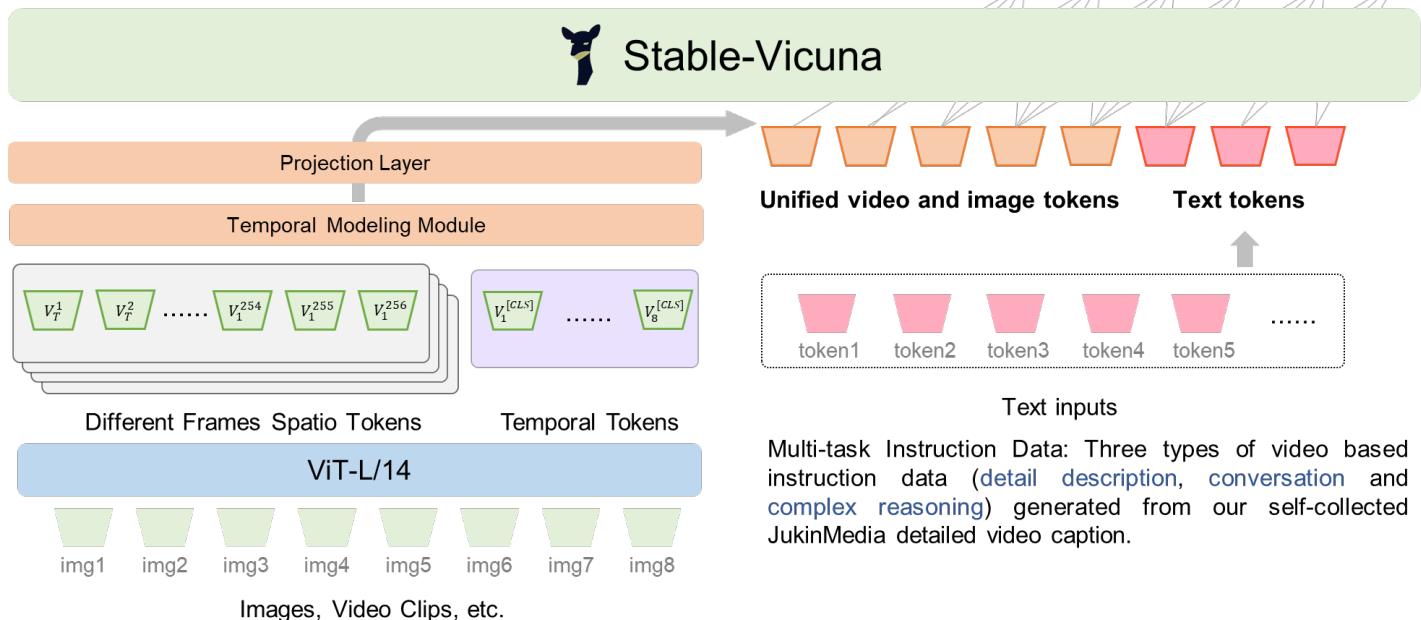


The data has been open-sourced in
<https://huggingface.co/datasets/luoruipu1/Valley-Instruct-73k>

- 基于 Jukinmedia 视频网站 (视频描述很长，而且丰富)

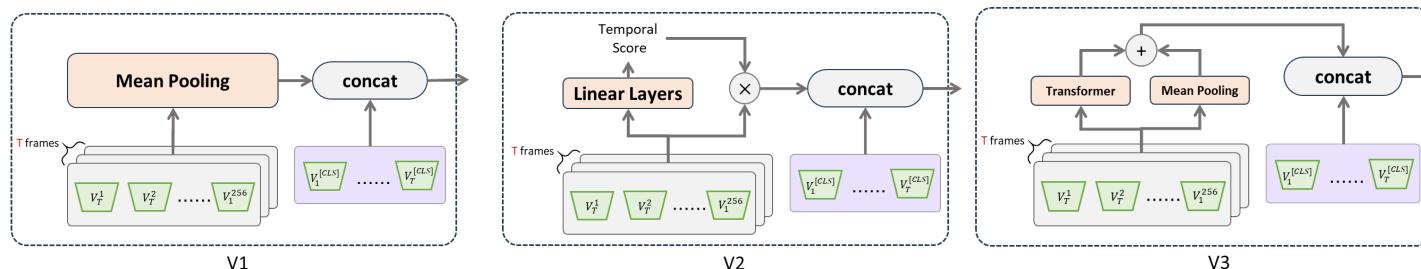
Valley: 视频语言模型的架构介绍

Main Structure

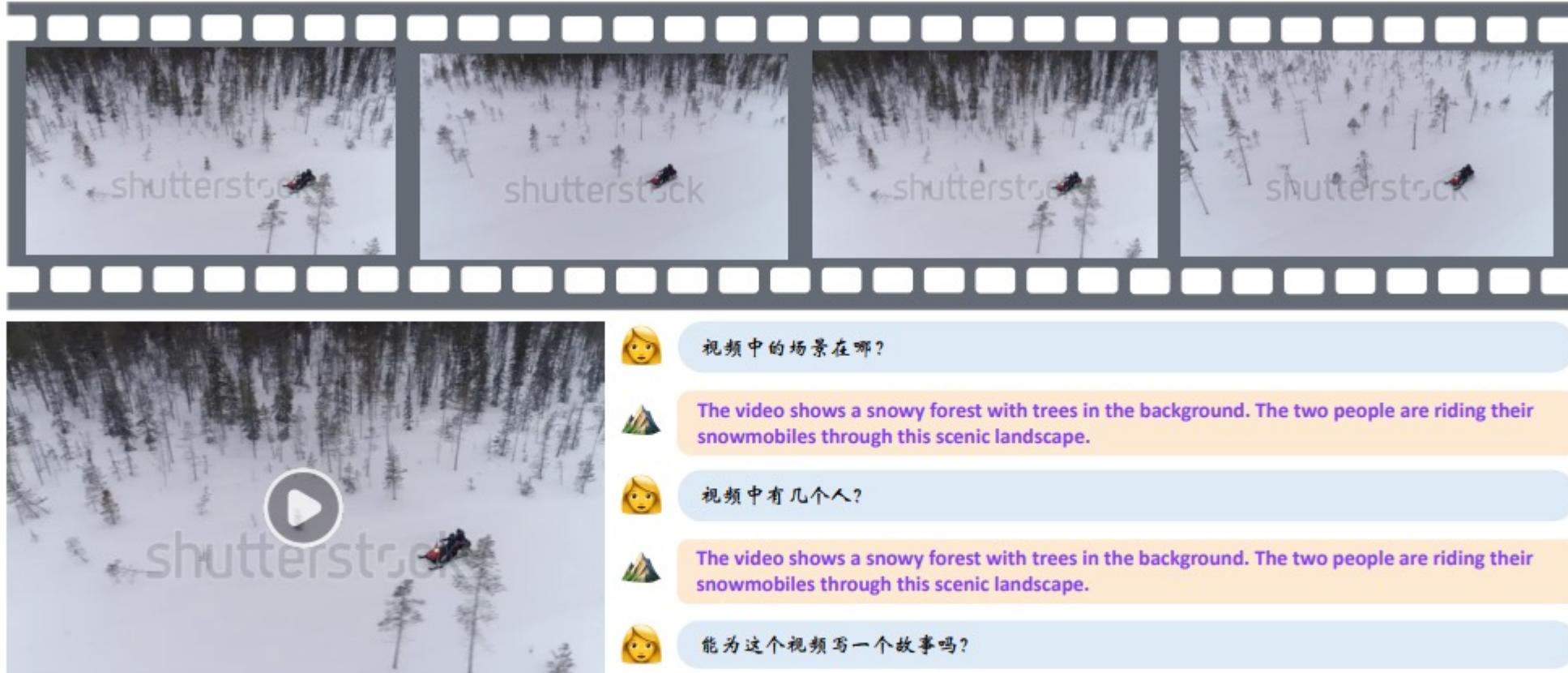


- 语言模型：stable-vicuna LLM
- 视觉编码：ViT
- 使用了三种方法进行视频编码

Temporal Modeling Module



Valley: 样例演示



 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 “视觉扩充的大语言模型”
- 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跨模态交互转变为模态连接
- 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

