



北京航空航天大學  
BEIHANG UNIVERSITY

人工智能研究院  
Institute of Artificial Intelligence

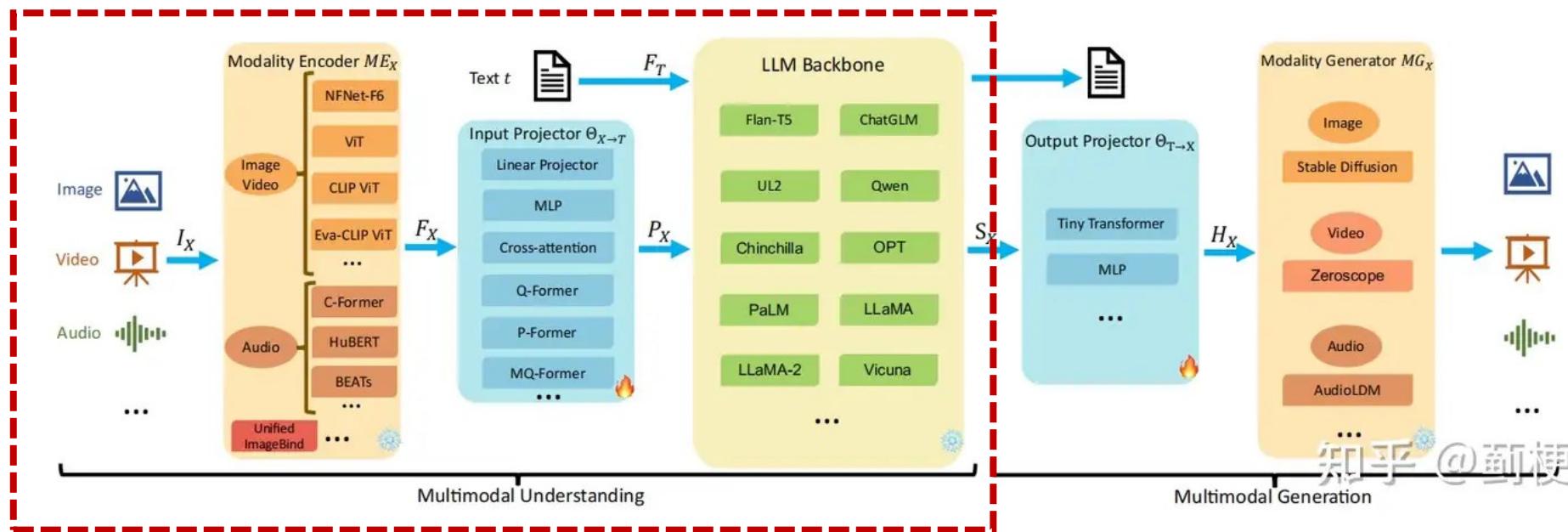
# 多模态大语言模型：基础原理与前沿应用

韦星星

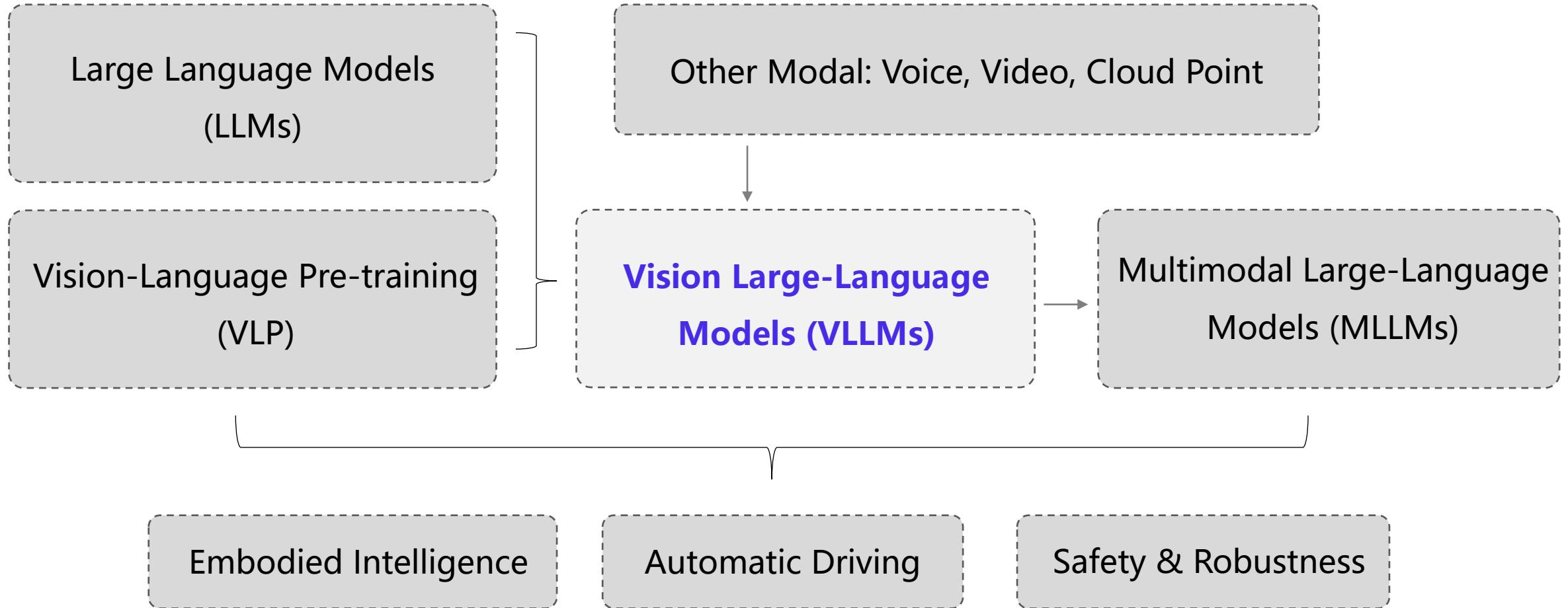
2024年5月

# Why & What is MLLM

- 自然语言是人类对生活环境中不同模态信息（视觉、音频、动作）的一种高度抽象，能够将复杂的概念和信息以人类可理解的形式呈现
- 多模态大语言模型（MLLMs）本质上是一种结合了多种感知模式（如图像、声音等）的大规模语言模型
- 从输入输出形式而言，多模态大语言模型以文本、图像等不同模态信息作为输入，并使用文本模态作为输出



# A RoadMap



# Contents

## 1 预备知识

1.1 大语言模型

## 2 视觉-语言预训练 (VLP)

2.1 Vision-Language (VL) 发展历史

2.2 多模态任务 / 下游任务

2.3 训练目标及模型架构总结

2.4 早期代表性工作

2.5 编码器架构：CLIP及相关改进工作

2.6 编码-解码器架构：BLIP

## 3 视觉-大语言模型 (VLLMs)

3.1 BLIP-2

3.2 MiniGPT4V

3.3 视觉指令微调：LLaVa系列工作

## 4 多模态大语言模型 (VLLMs) 及应用

4.1 3D模态：LION

## 5 多模态大语言模型的安全研究

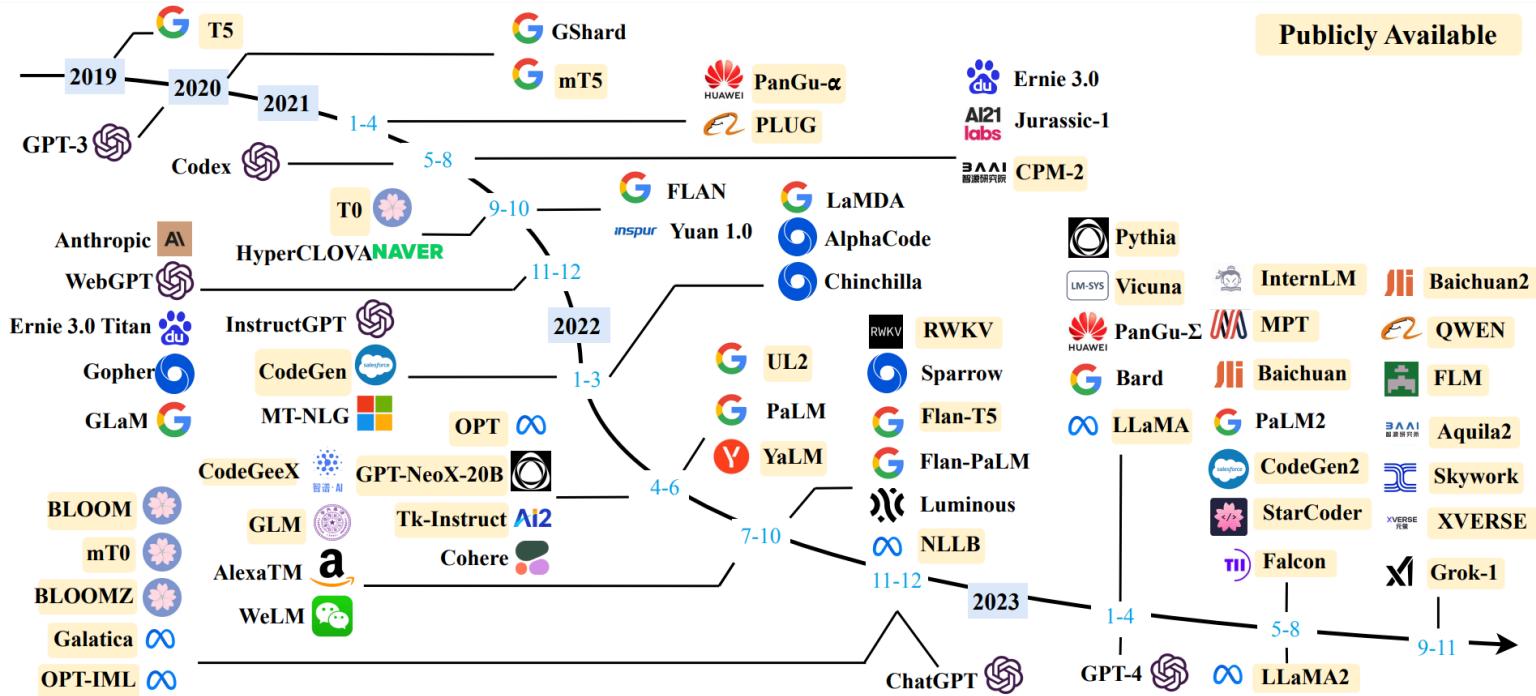


北京航空航天大學  
BEIHANG UNIVERSITY

人工智能研究院  
Institute of Artificial Intelligence

# Intro to Large Language Models ( LLMs )

# 以GPT为代表的基础大模型技术在近年来发展迅猛

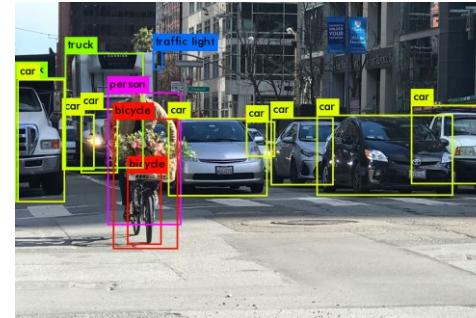


- ✓ 2018.06: GPT1.0 [117M Param]
- ✓ Generative Pre-training + Finetuning
- ✓ 2019.02: GPT2.0 [1.5B Param]
- ✓ Unsupervised Multitask Learners
- ✓ 2020.05: GPT3.0 [175B Param]
- ✓ Few-Shot Learner
- ✓ ... , ...
- ✓ 2022.11: ChatGPT(GPT3.5) [200B Param]
- ✓ Finetuning on GPT3.0 & 4096 Tokens
- ✓ 2023.03: GPT4.0 [? Param]
- ✓ 32768 Tokens

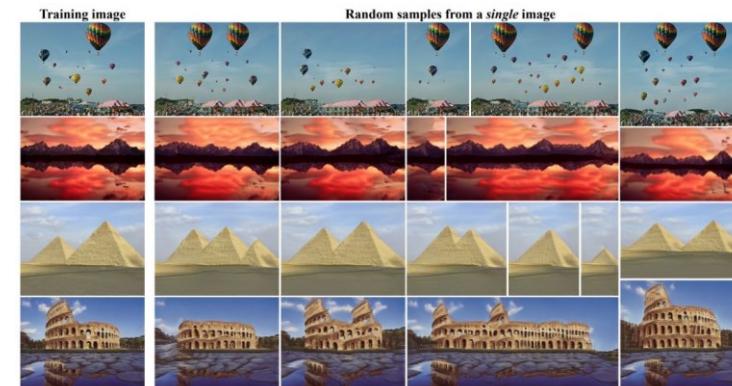
# 以ChatGPT为代表的大语言模型是典型的 生成式人工智能技术 (AIGC)

AI模型可大致分为判别式/分析式AI (Discriminant/Analytical AI) 和生成式AI (Generative AI) 两类

1. 判别式AI：学习数据中的**条件概率分布** $P(Y|X)$ ，  
试图建立从输入变量到输出变量的直接映射，根  
据已有数据进行分析、判断、预测。



2. 生成式AI：学习数据中的**联合概率分布** $P(X,Y)$ ，  
并非简单分析已有数据而是学习归纳已有数据后  
进行演技创造，基于历史进行模仿式、缝合式创  
作，生成全新的内容，也能解决判别问题。



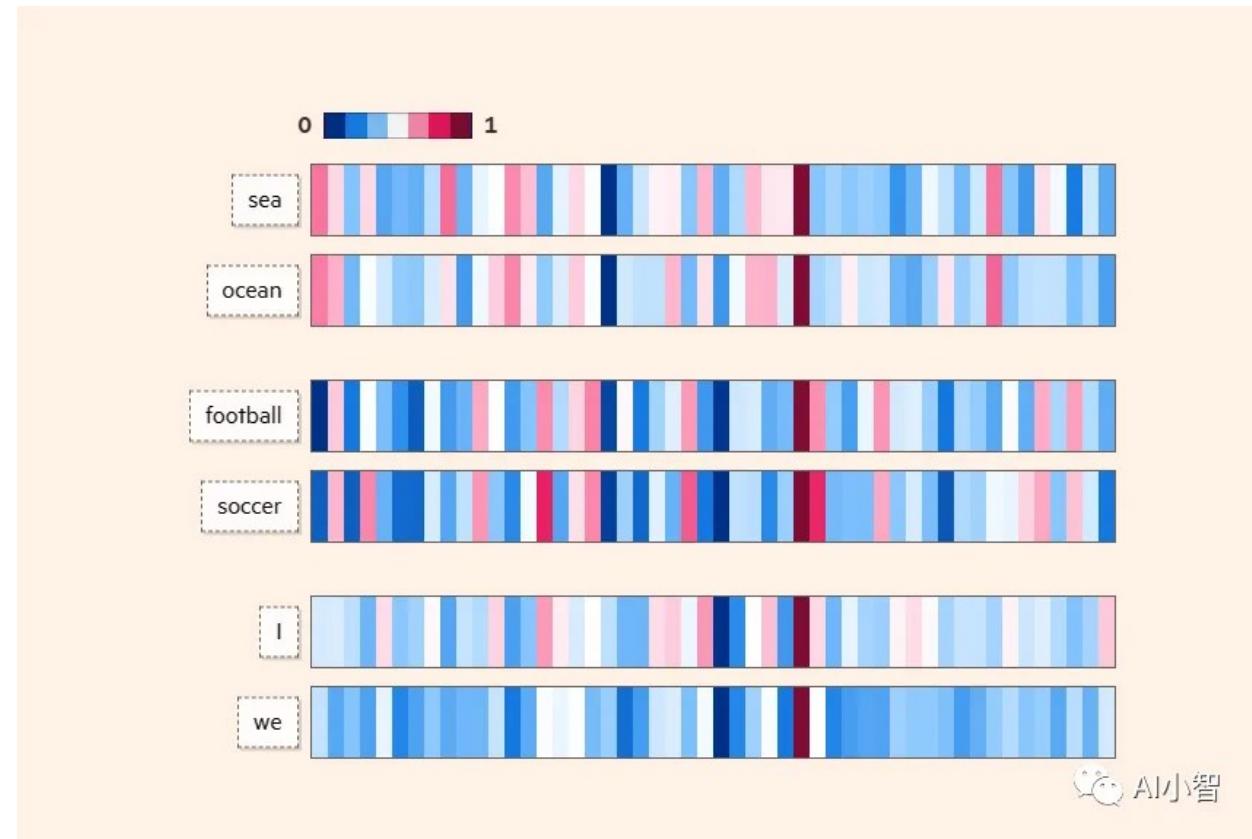
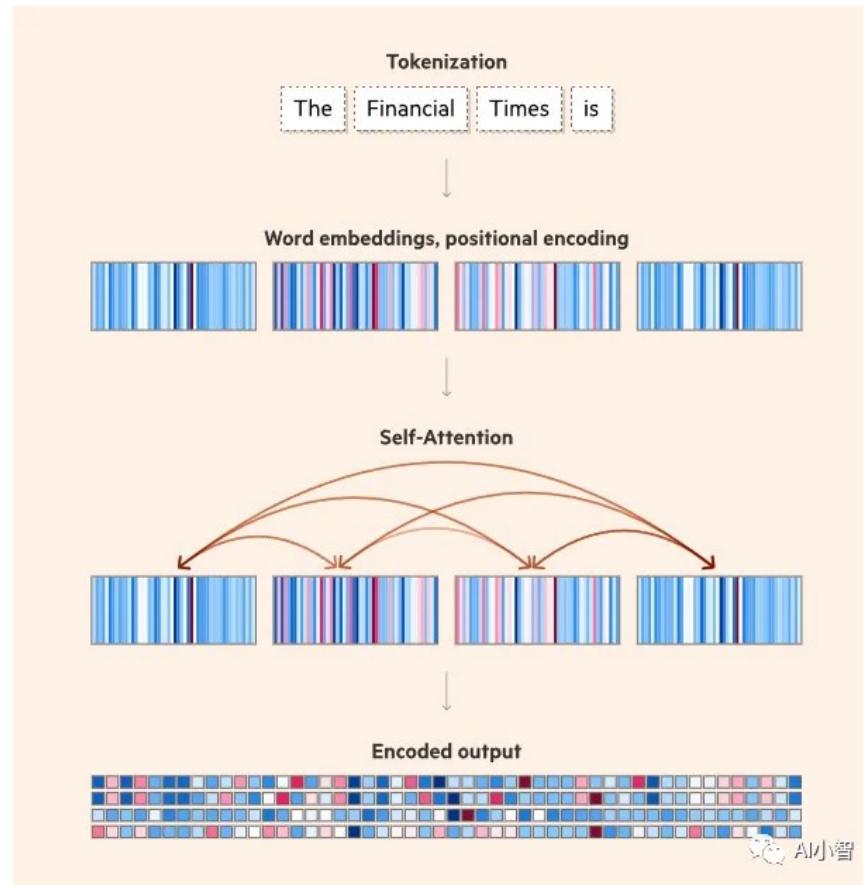
# 以ChatGPT为代表的大语言模型是典型的 生成式人工智能技术 (AIGC)



# GPT的技术途径

## • Tokenization (分词)

1. 在自然语言处理 (NLP) 中, token是指文本中最小的语义单元
2. 分词化指将单词 —> 机器能理解的语言, 即数字形式 (词嵌入), 相似的词具有相近的嵌入



## GPT的技术途径

- **Next-Token-Prediction & Masked-Language-Modeling**

“The cat sat on the”

“The [MASK] sat on the”

请各位做一个完形填空： \_\_\_\_\_ 和阿里、腾讯一起并成为中国互联网 BAT 三巨头。

1. 从大规模的上亿的文本预料中，随机地扣掉一部分字，形成上面例子的完形填空题型，不断地学习空格处到底该填写什么。**所谓语言模型的训练和学习，就是从大量的数据中学习复杂的上下文联系。**
2. 这一目标允许模型学习语言的统计结构，例如常见的词序列和词使用模式。这通常有助于模型生成更自然、更流畅的文本，并且是每个语言模型预训练阶段的重要步骤。

# GPT的技术途径

## • In-Context Learning ( Prompt Learning )

从GPT-3 率先开始 提出了一种 in-context 学习方式，即上下文学习

用户输入到 GPT-3：你觉得 JioNLP 是个好用的工具吗？

GPT-3输出1：我觉得很好啊。

GPT-3输出2：JioNLP是什么东西？

GPT-3输出3：你饿不饿，我给你做碗面吃.....

GPT-3输出4：Do you think jionlp is a good tool?

用户输入到 GPT-3：请把以下中文翻译成英文：你觉得 JioNLP 是个好用的工具吗？

用户输入到 GPT-3：请把以下中文翻译成英文：苹果 => apple; 你觉得 JioNLP 是个好用的工具吗？ =>

The three settings we explore for in-context learning

### Zero-shot

The model predicts the answer given only a natural language description of the task. No gradient updates are performed.



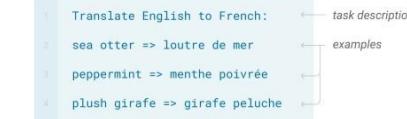
### One-shot

In addition to the task description, the model sees a single example of the task. No gradient updates are performed.



### Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.



Traditional fine-tuning (not used for GPT-3)

### Fine-tuning

The model is trained via repeated gradient updates using a large corpus of example tasks.



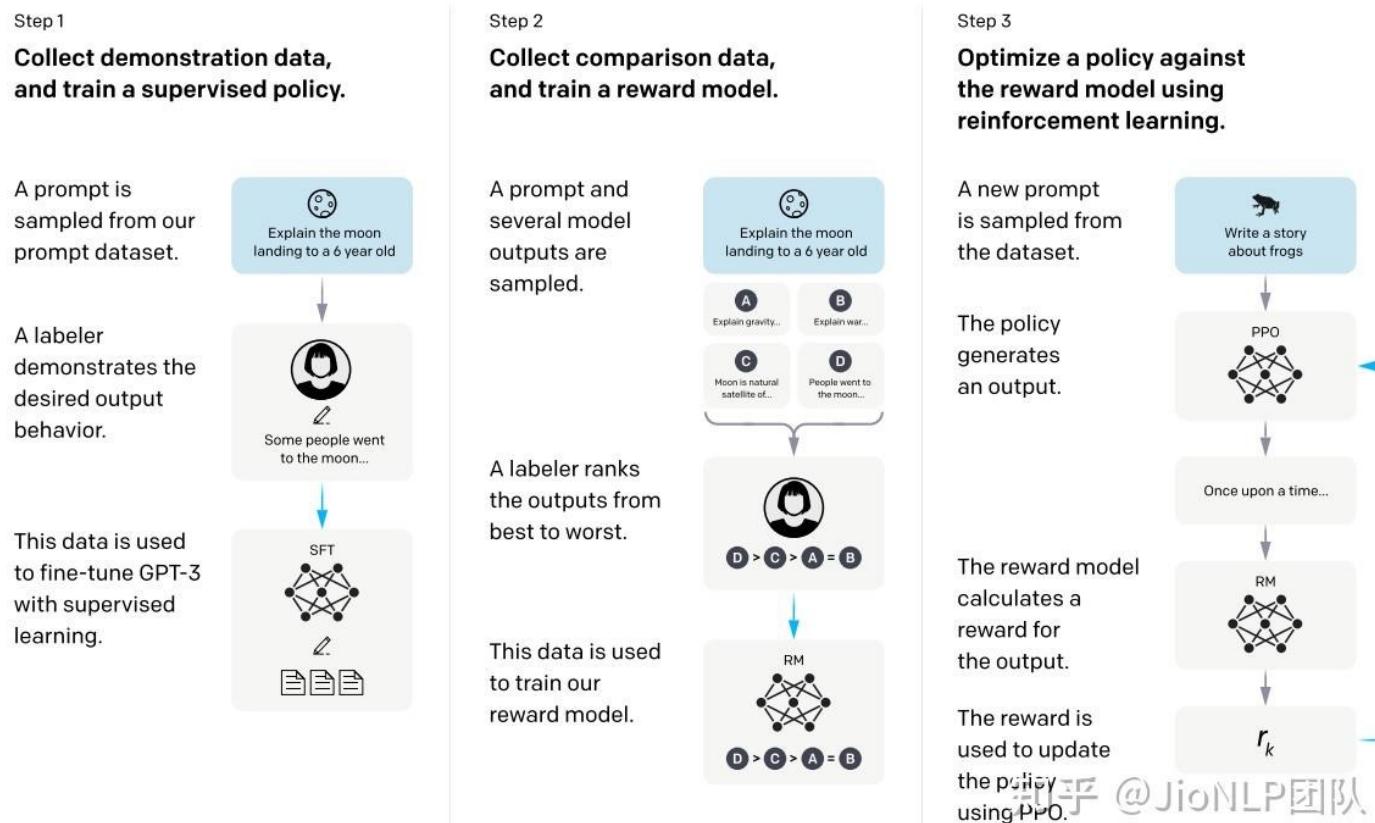
知乎 @JioNLP团队

其中 苹果翻译成 *apple*, 是一个示范样例，用于让模型感知该输出什么。只给提示叫做 zero-shot, 给一个范例叫做 one-shot, 给多个范例叫做 few-shot。

# GPT的技术途径

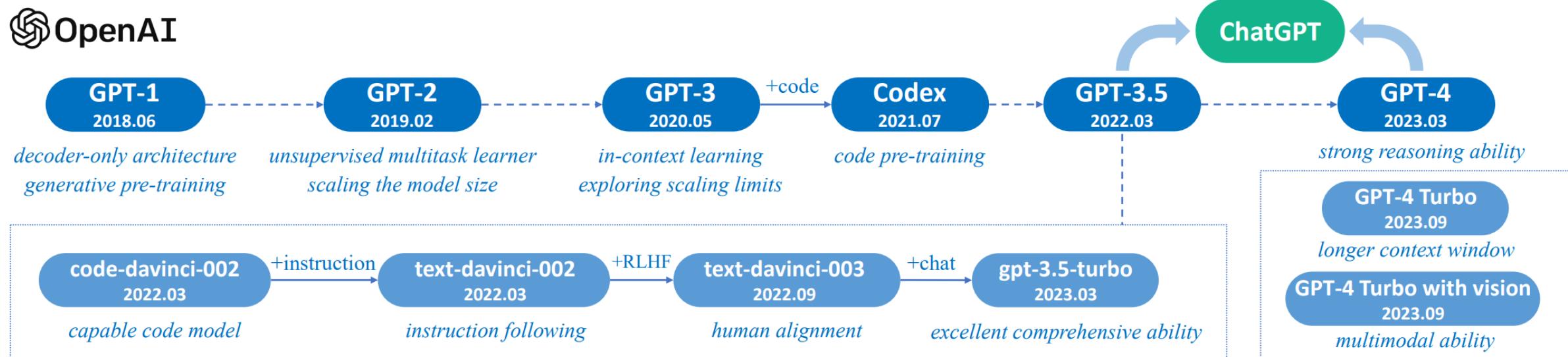
## • Reinforcement Learning from Human Feedback (人类反馈强化学习)

- 为了让回答更加符合人类偏好，采用强化学习+人类反馈奖励函数的方式来对模型进行进一步微调。
- 在ChatGPT里，通过40名外包人员不断地从模型的输出结果中筛选，判断哪些句子是好的，哪些是低质量的，这样就可以训练得到一个 **reward 母体模型**。



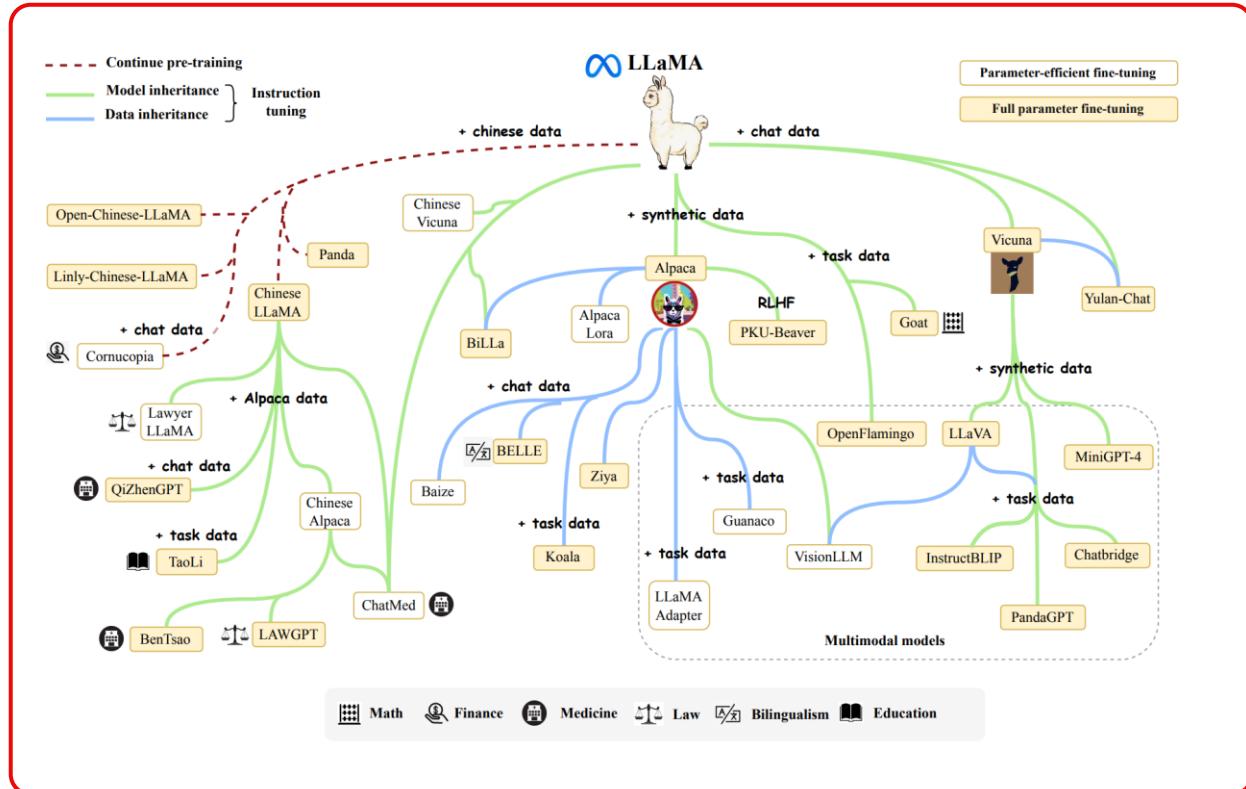
# 基础大模型的发展趋势: 从纯语言模型发展到多模态大模型

OpenAI GPT-4已经具备处理 **图像** 模态的能力



# 基础大模型的发展趋势: 从纯语言模型发展到多模态大模型

开源领域: 基于大语言架构 LLaMa 衍生出了众多的多模态大模型 具备对 图像 视频 音频 等模态的处理能力



## X-InstructBLIP: A Framework for aligning X-Modal instruction-aware representations to LLMs and Emergent Cross-modal Reasoning

Artemis Panagopoulou, Le Xue\*, Ning Yu\*, Junnan Li, Dongxu Li, Shafiq Joty, Ran Xu, Silvio Savarese, Caiming Xiong, and Juan Carlos Niebles

arXiv 1234.56789 | Open in Colab \*equal mentorship contribution.

### Overview

## LLVA: Large Language and Vision Assistant

Visual instruction tuning towards large language and vision models with GPT-4 level capabilities.

[!] [LLVA-NeXT Blog](#) | [Project Page](#) | [Demo](#) | [Data](#) | [Model Zoo](#)

Community Contributions: [llama.cpp](#) | [Colab](#) | [Space](#) | [Replicate](#) | [AutoGen](#) | [BakLLVA](#)

## OpenFlamingo

pypi package 2.0.1

[Paper](#) | Blog posts: [1](#), [2](#) | [Demo](#)



北京航空航天大學  
BEIHANG UNIVERSITY

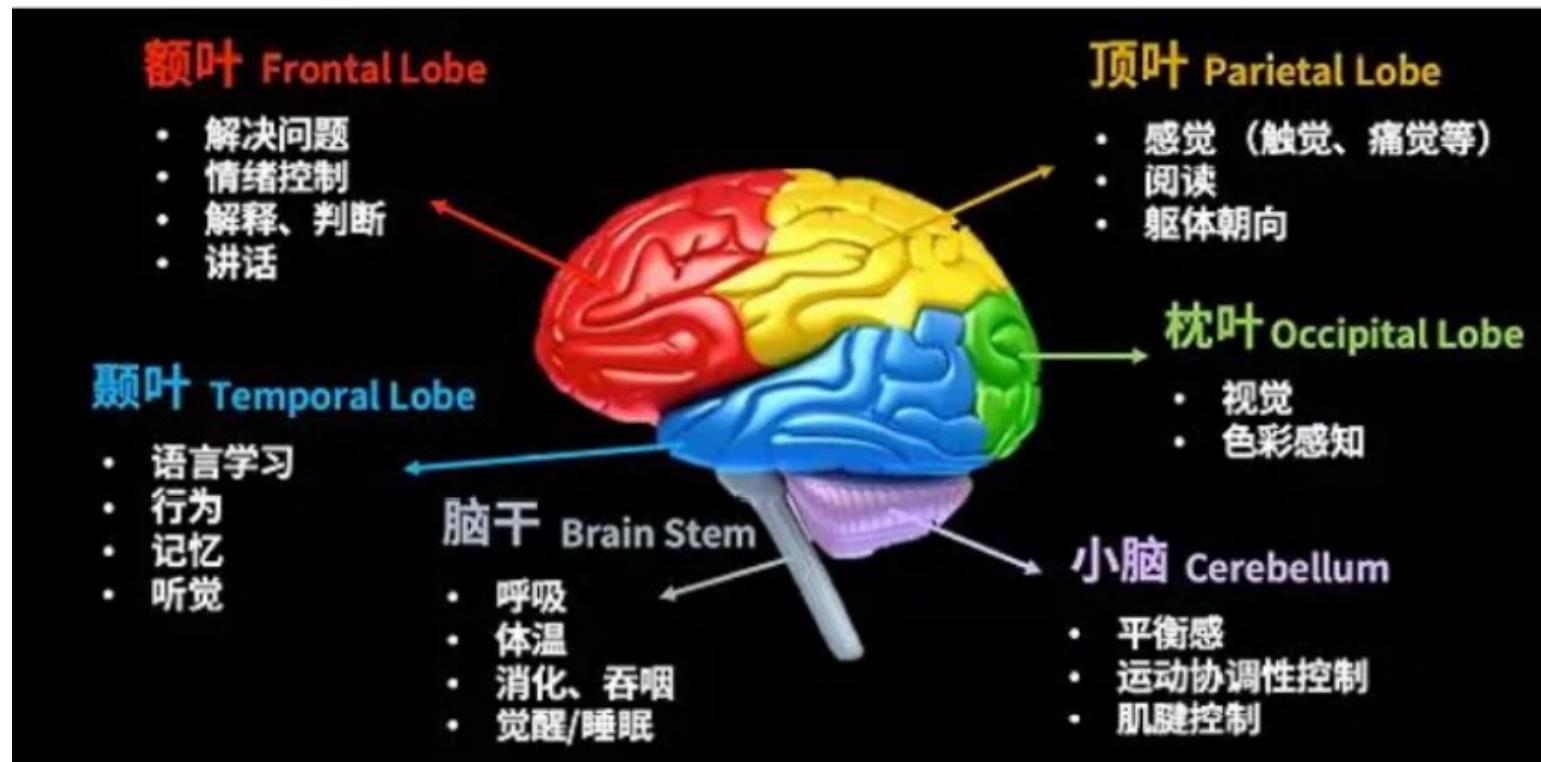
人工智能研究院  
Institute of Artificial Intelligence

# Intro to Vision-Language Pre-training (VLP)

# 为什么需要关注多模态预训练问题？从人脑机制启发

人脑“视觉概念与语言概念一致性”的认知机制是我们采取多模态预训练模型的可靠性依据之一

1. 人类约有70%的信息依靠视觉获得，剩余约20%~30%的信息依靠听觉和触觉
2. 互联网提供了90%的图文大数据，而文本数据只占了10%



# Vision-Language (VL) 发展历史

## VL学习的发展分为三个阶段

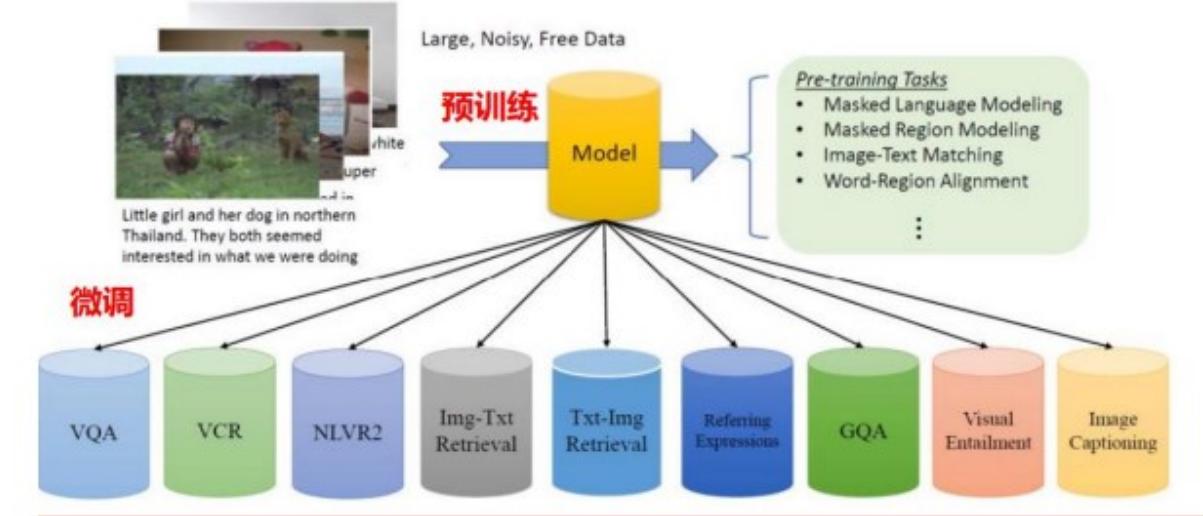
1. 2014-2018年：专门的模型被设计用于不同的任务
2. 2019-2021年 (BERT诞生) : 通过使用高质量标签的VL数据集进行预训练，模型能够学习视觉和语言的联合表征
  - ViLBERT
  - UNITER



16

### VL预训练模型 (VL-PTM)

- 2017年-2018年, NLP-PTMs → PTMs+微调模式→SOTA性能
- 2019年 VL-PTMs 开启研究工作



# Vision-Language (VL) 发展历史

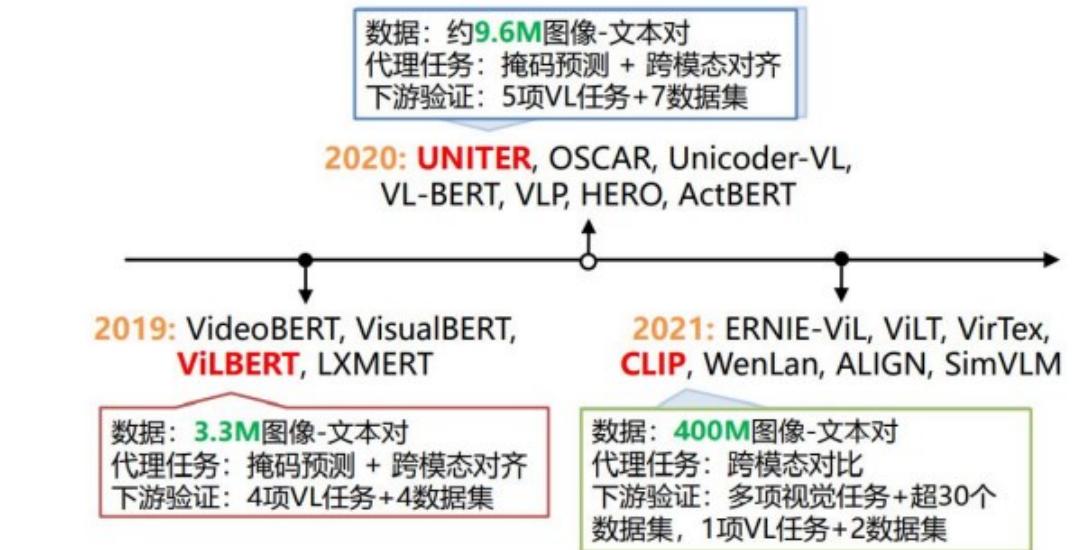
## VL学习的发展分为三个阶段

3. 2021年-今：随着CLIP的出现，期望在更大的弱标签数据集上预训练VL模型，并通过VL预训练获得性能强大的基于零样本或少样本的视觉模型

- CLIP及改进工作（纯编码器架构）
- BLIP系列工作（编码-解码器架构）

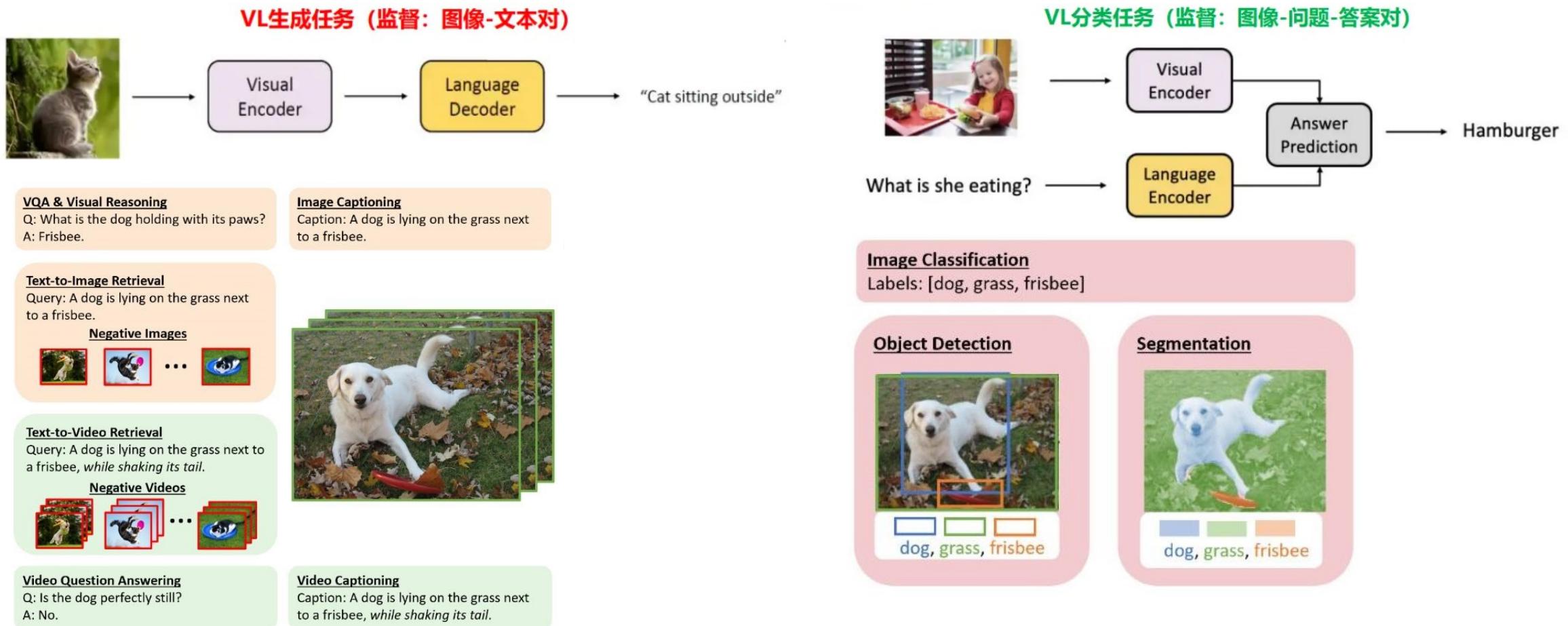


### VL-PTMs的演进 (2019-2021)



# 多模态任务 / 下游任务

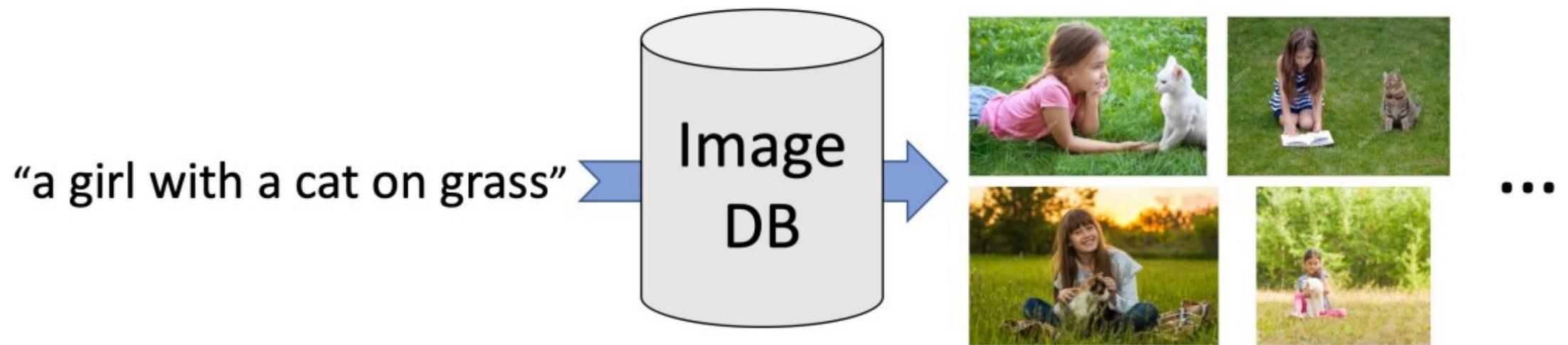
VL任务构成了多模态领域各种任务的基础 包括生成式任务和判别式任务



## 多模态任务 / 下游任务

一些常见的下游任务：

1. Text-Image / Image-Text Retrieval



## 多模态任务 / 下游任务

一些常见的下游任务：

### 2. Captioning & Visual Question Answering



What color are her eyes?  
What is the mustache made of?



How many slices of pizza are there?  
Is this a vegetarian pizza?



Is this person expecting company?  
What is just under the tree?



Does it appear to be rainy?  
Does this person have 20/20 vision?

## 多模态任务 / 下游任务

一些常见的下游任务：

### 3. Visual Entailment

视觉蕴涵是最近提出的多模态推理任务，其目标是预测一段文本与一幅图像的逻辑关系（蕴含/中立/矛盾）



+

- *Two woman are holding packages.*
- *The sisters are hugging goodbye while holding to go packages after just eating lunch.*
- *The men are fighting outside a deli.*

=

- *Entailment*
- *Neutral*
- *Contradiction*

Premise

Hypothesis

Answer

## 多模态任务 / 下游任务

一些常见的下游任务：

### 4. Natural Language for Visual Reasoning



The left image contains twice the number of dogs as the right image, and at least two dogs in total are standing.

true



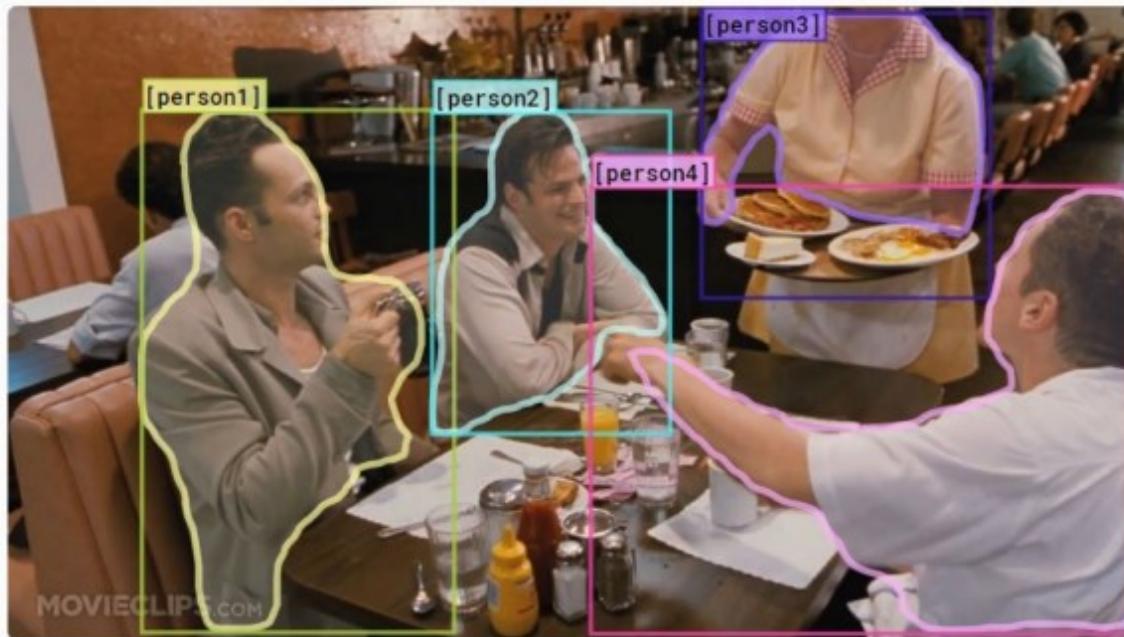
One image shows exactly two brown acorns in back-to-back caps on green foliage.

false

# 多模态任务 / 下游任务

一些常见的下游任务：

## 5. Visual Commonsense Reasoning



Why is [person4] pointing at [person1]?

- a) He is telling [person3] that [person1] ordered the pancakes.
- b) He just told a joke.
- c) He is feeling accusatory towards [person1].
- d) He is giving [person1] directions.

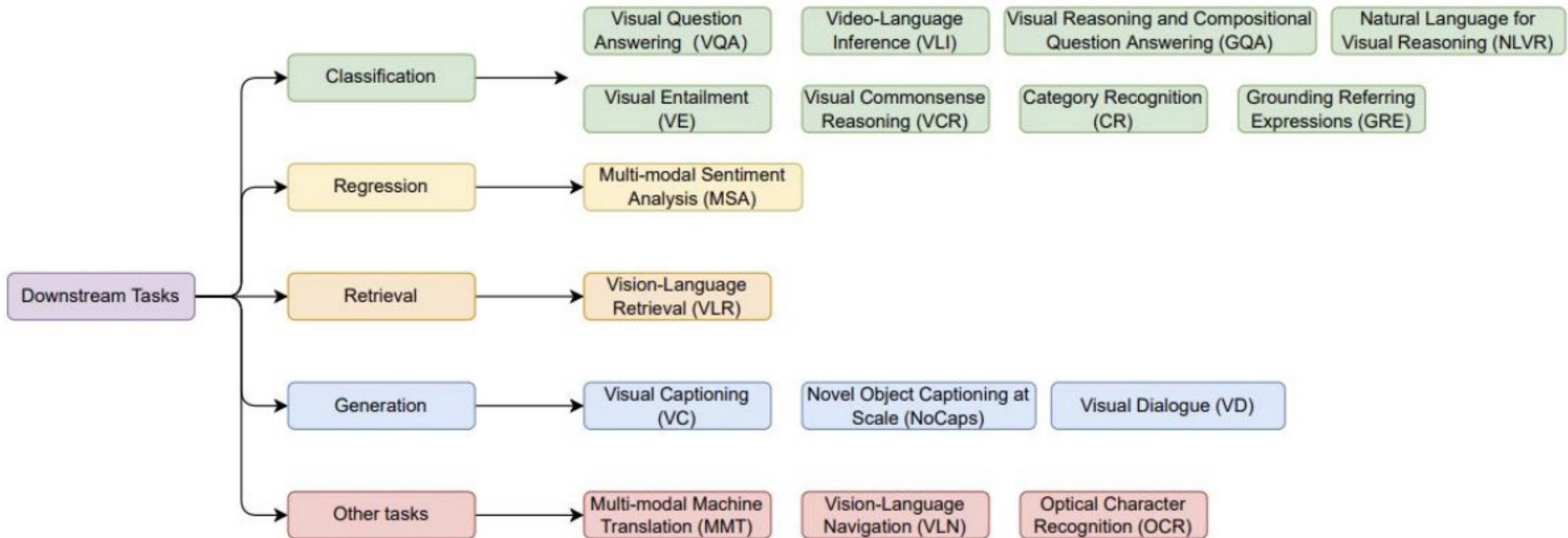
I choose (a) because:

- a) [person1] has the pancakes in front of him.
- b) [person4] is taking everyone's order and asked for clarification.
- c) [person3] is looking at the pancakes and both she and [person2] are smiling slightly.
- d) [person3] is delivering food to the table, and she might not know whose order is whose.

# 多模态任务 / 下游任务

一些常见的下游任务：

6. More...



# VLP的常用模型架构

## 用于 Feature Extraction

### 1. 文本模态的特征提取

- LSTM
- BERT

### 2. 图像模态的特征提取

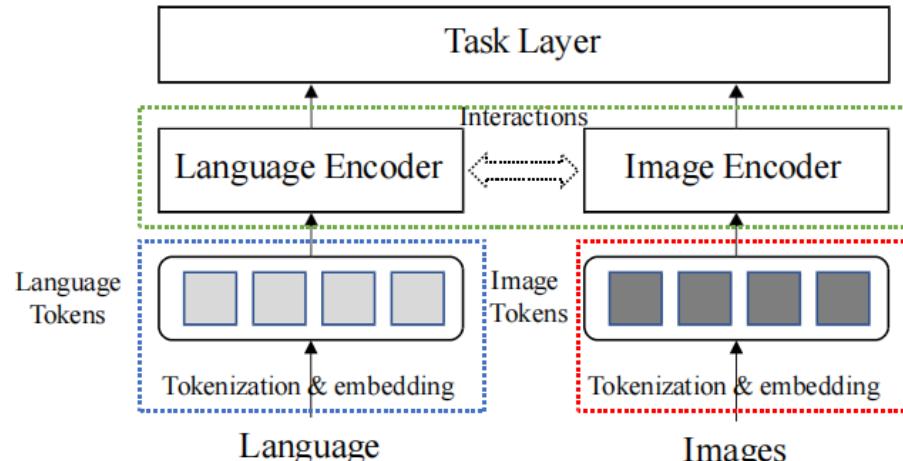
- CNN: Feature Maps
- ROI: Detected Objects
- ViT: 16\*16 patch
- Pixel

### 3. 视频模态的特征提取

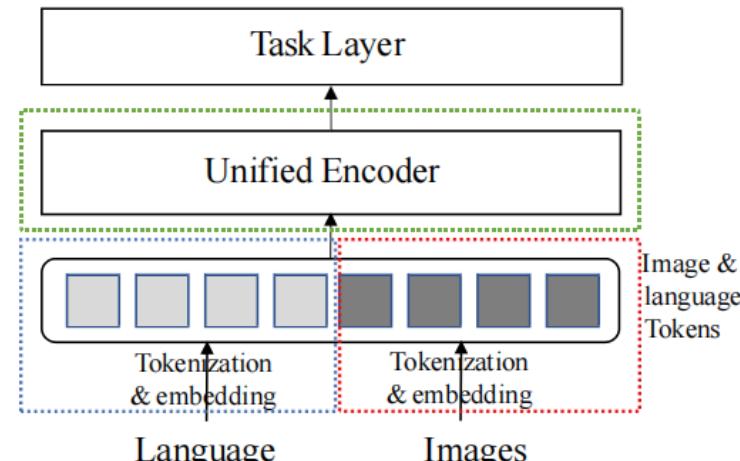
- 先抽取帧图片特征，再把时序图片特征变成定长特征
- temporal maxpooling / mean pooling

# VLP的常用模型架构

## Model Architecture



(a) Dual Stream Models



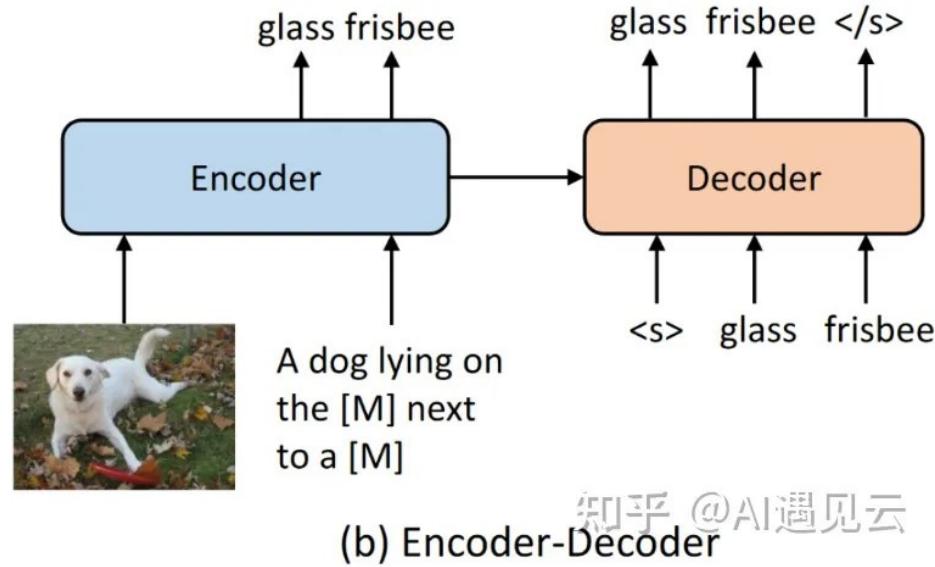
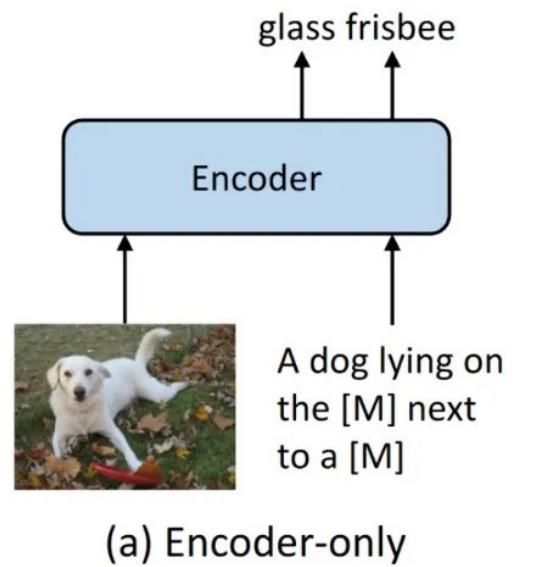
(b) Single Stream Models

### 1. Single-Stream VS. Dual-Stream

- 单流模式指的是将视觉和文本编码特征组合在一起，然后输入单个 Transformer 块中，通过合并注意力来融合多模态输入，通常也被叫做 merged attention。参数效率更高。
- 双流模式是指视觉和文本编码特征没有组合在一起，而是独立输入到两个不同的 Transformer块，不共享参数，而是通过交叉注意力实现跨模态交互，因此也被叫做 co-attention。

# VLP的常用模型架构

## Model Architecture



## 2. Encoder-only VS. Encoder-decoder

- 大多数 VLP 模型仅包含编码器，其中跨模态表示直接输入到基于 MLP 的输出层来生成最终输出。这种设计自然适合 VL 理解任务，例如 VQA 和视觉推理
- 近期衍生除了一些基于 Transformer 的编码器-解码器架构，跨模态表示首先被送入解码器，然后送入输出层。在这些模型中，解码器同时关注编码器表示和先前生成的标记，以自回归地方式产生输出。可以实现各种图像-文本任务的统一和 VLP 模型的零样本/少样本学习，并且也很自然地适合生成类的任务。

## VLP的常用训练目标

### 预训练形式1. Masked Language Modeling (MLM)

$$\mathcal{L}_{\text{MLM}}(\theta) = -E_{(\mathbf{w}, \mathbf{v}) \sim D} \log P_\theta (\mathbf{w}_m \mid \mathbf{w}_{\setminus m}, \mathbf{v})$$



- Image Regions:  $\mathbf{v} = \{v_1, \dots, v_K\}$
- Sentence Tokens:  $\mathbf{w} = \{w_1, \dots, w_T\}$
- Masking Indices:  $\mathbf{m} \in \mathbb{N}^M$

### Masked Language Modeling (MLM)

- 十分straightforward的一种思路，把图像块处理为一种特殊的token
- 迁移自LLMs中的Masked-Language-Modeling目标，结合图像和文本模态进行mask预测
- 最大化 mask内预测为正确token的期望

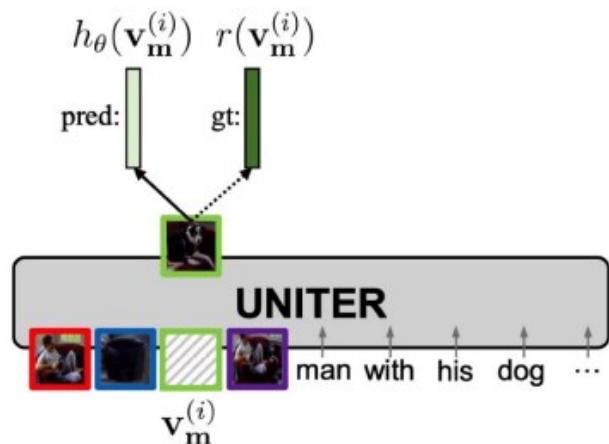
## 预训练形式2. Masked Vision Modeling (MVM)

- 模仿MLM，只不过是对图片或者图片patch的 token 进行随机mask
- 使需要在给定剩余的视觉特征和所有文本特征的情况下重建被 mask 的视觉特征

### 2.1 Masked Region Feature Regression (MRFR)

$$f_{\theta} (\mathbf{v}_m \mid \mathbf{v}_{\setminus m}, \mathbf{w}) = \sum_{i=1}^M \left\| h_{\theta} \left( \mathbf{v}_m^{(i)} \right) - r \left( \mathbf{v}_m^{(i)} \right) \right\|_2^2$$

- Image Regions:  $\mathbf{v} = \{v_1, \dots, v_K\}$
- Sentence Tokens:  $\mathbf{w} = \{w_1, \dots, w_T\}$
- Masking Indices:  $\mathbf{m} \in \mathbb{N}^M$



- 将 mask 特征的模型输出转换为与原始视觉特征相同维度的向量，并在原始视觉特征和向量之间应用 L2 回归

# VLP的常用训练目标

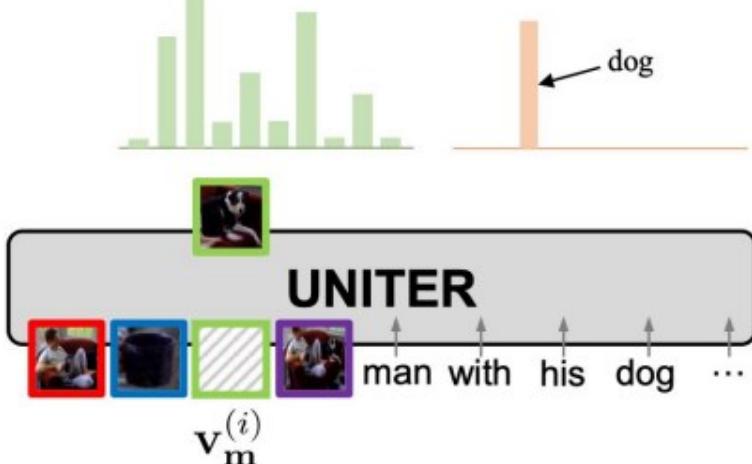
## 预训练形式2. Masked Vision Modeling (MVM)

- 模仿MLM，只不过是对图片或者图片patch的 token 进行随机mask
- 使需要在给定剩余的视觉特征和所有文本特征的情况下重建被 mask 的视觉特征

### 2.2 Masked Region Classification (MRC)

$$f_{\theta} (\mathbf{v}_m \mid \mathbf{v}_{\setminus m}, \mathbf{w}) = \sum_{i=1}^M \text{CE} \left( c \left( \mathbf{v}_m^{(i)} \right), g_{\theta} \left( \mathbf{v}_m^{(i)} \right) \right)$$

$$g_{\theta}(\mathbf{v}_m^{(i)}) \in \mathbb{R}^K \quad c(\mathbf{v}_m^{(i)}) \in \mathbb{R}^K$$



- Image Regions:  $\mathbf{v} = \{v_1, \dots, v_K\}$
- Sentence Tokens:  $\mathbf{w} = \{w_1, \dots, w_T\}$
- Masking Indices:  $\mathbf{m} \in \mathbb{N}^M$

- 学习预测 mask 特征的对象语义类。VLP 模型首先将 mask 特征的输出馈送到 FC 层以预测对象类别的分数，然后进一步通过 softmax 函数将其转换为预测归一化分布

# VLP的常用训练目标

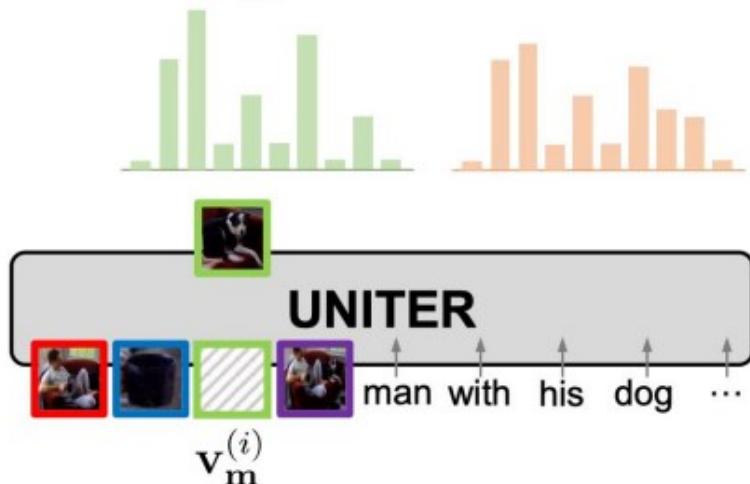
## 预训练形式2. Masked Vision Modeling (MVM)

- 模仿MLM，只不过是对图片或者图片patch的 token 进行随机mask
- 使需要在给定剩余的视觉特征和所有文本特征的情况下重建被 mask 的视觉特征

### 2.3 Masked Region Classification – KL Divergence (MRC-kl)

$$f_{\theta} (\mathbf{v}_m \mid \mathbf{v}_{\setminus m}, \mathbf{w}) = \sum_{i=1}^M D_{KL} \left( \tilde{c} \left( \mathbf{v}_m^{(i)} \right) \parallel g_{\theta} \left( \mathbf{v}_m^{(i)} \right) \right)$$

$$g_{\theta}(\mathbf{v}_m^{(i)}) \in \mathbb{R}^K \quad \tilde{c}(\mathbf{v}_m^{(i)}) \in \mathbb{R}^K$$



- Image Regions:  $\mathbf{v} = \{v_1, \dots, v_K\}$
- Sentence Tokens:  $\mathbf{w} = \{w_1, \dots, w_T\}$
- Masking Indices:  $\mathbf{m} \in \mathbb{N}^M$

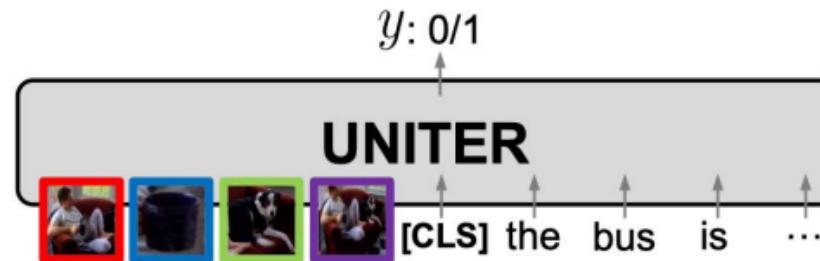
- 与MRC不同，在这里，不使用绝对的类语义标签作为ground truth，而是预测一个 distribution over semantic classes (soft label)，使用KL散度计算损失

## VLP的常用训练目标

### 预训练形式3. Image-Text Matching (ITM)

- 判断当前 `text-image` pair是不是匹配：将图片的`[IMG]` token和文本的`[CLS]` token进行匹配

$$\mathcal{L}_{\text{ITM}}(\theta) = -E_{(\mathbf{w}, \mathbf{v}) \sim D} [y \log s_\theta(\mathbf{w}, \mathbf{v}) + (1 - y) \log (1 - s_\theta(\mathbf{w}, \mathbf{v}))]$$



### 最常见形式：Vision-Language Contrastive Learning

- VLP 模型将两种模式的融合表示提供给 FC 层和 sigmoid 函数以预测 0 到 1 之间的分数，其中 0 表示视觉和语言不匹配，1 表示视觉和语言匹配。
- 在训练期间，VLP 模型在每一步从数据集中采样正对或负对。负对是通过用从其他样本中随机选择的替换成对样本中的视觉或文本来创建的。

## A summary

**VLP的特征提取：**文本模态 / 图像模态 / 视频模态

**VLP的架构区分** (模态特征是否进行融合处理) Single-Stream VS. Dual-Stream

**VLP的架构区分** (是否包含编码器解码器) Encoder-only VS. Encoder-decoder

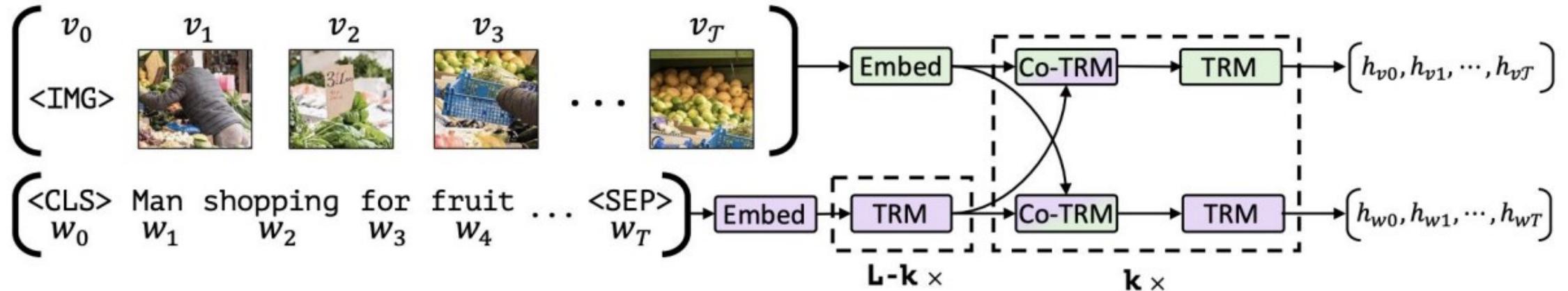
**预训练形式1. Masked Language Modeling (MLM)**

**预训练形式2. Masked Vision Modeling (MVM)**

**预训练形式3. Image-Text Matching (ITM)**

## 代表性的早期工作

### ViLBERT



- dual-stream
- Image Features: detected objects (RoI) + 5-d spatial location vector
  - 5-d vector: region position (normalized top-left and bottom-right coordinates) and the fraction of image area covered (区域坐标+区域面积)

## 代表性的早期工作

### ViLBERT

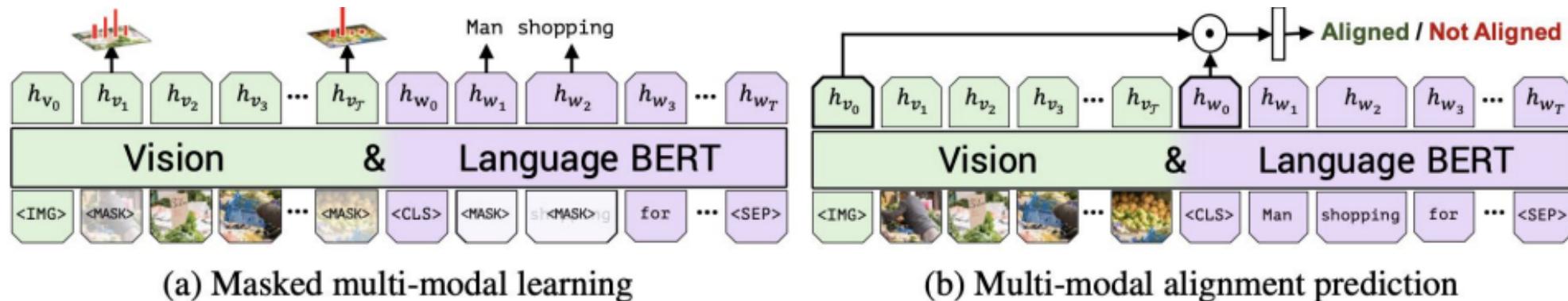
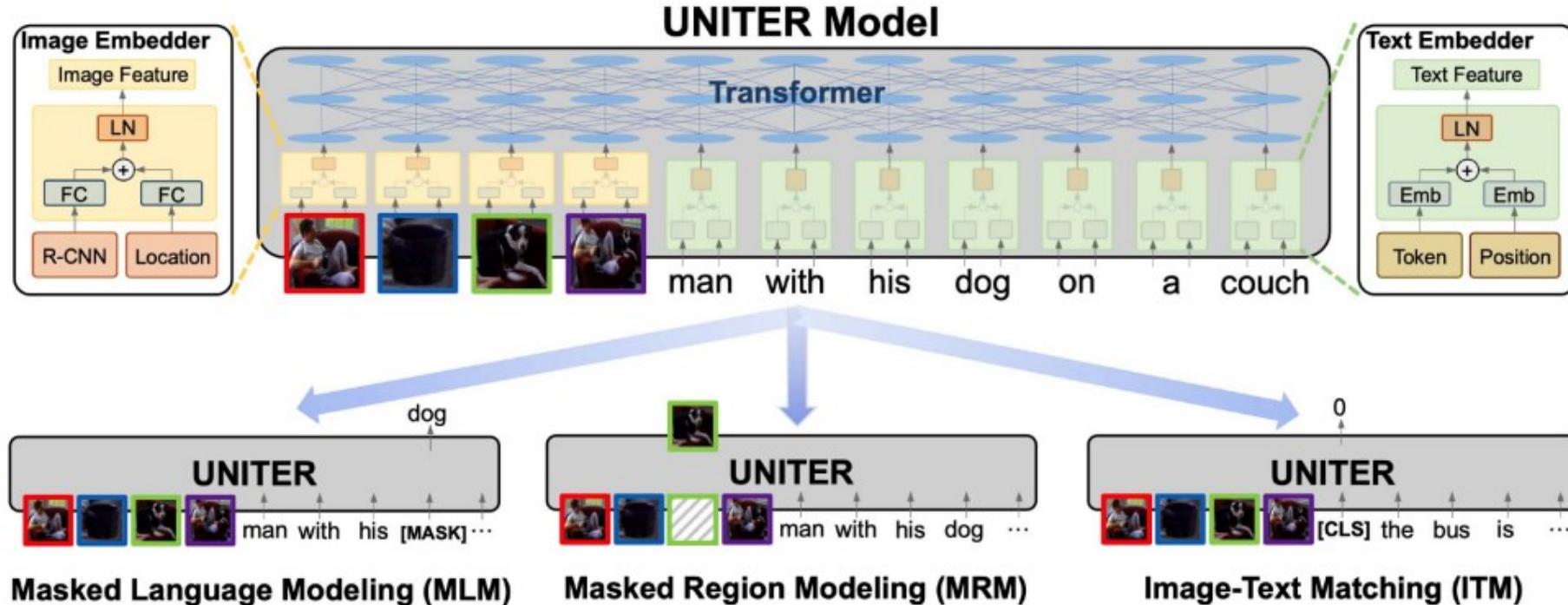


Figure 3: We train ViLBERT on the Conceptual Captions [24] dataset under two training tasks to learn visual grounding. In masked multi-modal learning, the model must reconstruct **image region categories** or words for masked inputs given the observed inputs. In multi-modal alignment prediction, the model must predict whether or not the caption describes the image content.

- MVM: the model predicts a distribution over semantic classes for the corresponding image region; KL divergence
- ITM: Predict Aligned / Not Aligned

# 代表性的早期工作

## UNITER



- single-stream
- Image Input: detected objects (RoI)
- Text Input: tokens
- Pretrained on MLM / MVM / ITM

## 双流编码器架构：CLIP及相关改进工作

**CLIP: Connecting text and images with Contrastive Language–Image Pre-training**

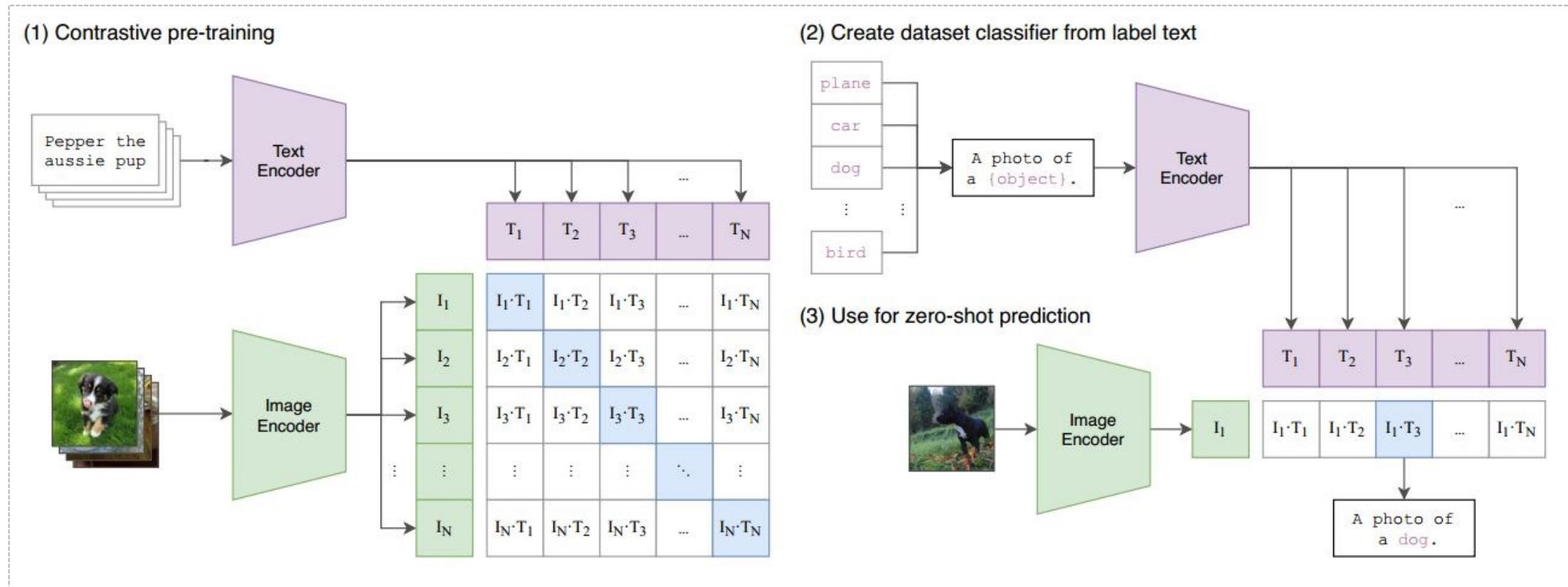
带领视觉语言预训练技术迈向新的时代



- Motivation: 当前的计算机视觉系统被训练成一种分类系统，严重影响了其在未见类别上的泛化性（需要用额外的有标注数据）。因此直接从嘈杂的原始文本中学习是一个更有前景的可替代方法，其能带来大量广泛的有监督数据。

# 双流编码器架构：CLIP及相关改进工作

## CLIP: Connecting text and images with Contrastive Language–Image Pre-training



- dual-stream
- train on large weakly paired text-image examples (WIT数据集：包含400M互联网采集的图文对)

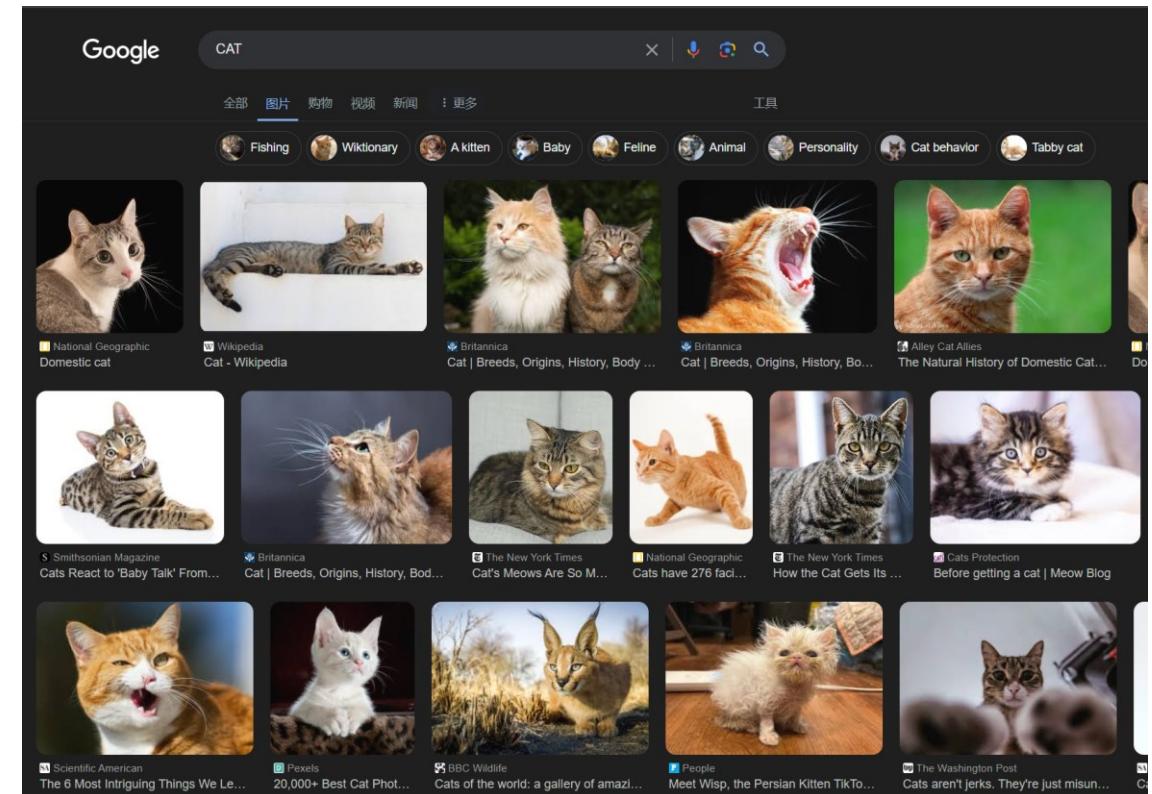
## 双流编码器架构：CLIP及相关改进工作

### CLIP: Connecting text and images with Contrastive Language–Image Pre-training

- Dataset: 基于互联网采集一个足够大的图像-文本对数据集

在此以前，用于VLP的大规模数据集主要有3个，MS-COCO、Visual Genome、YFCC100M。前两者是人工标注的，质量高但是数据量小，大约只有0.1M。YFCC100M大约有100M图片，但是质量堪忧，有些标题只有无意义的名字，过滤后，这个数据集缩小至15M，大约和ImageNet差不多大。

考虑到现有数据集太小，OpenAI自己构建了一个400M的数据集，使用500K个查询进行搜索，每个查询大约有20K个“图像一文本”对。该数据集被称为WIT (webimagetext)，数据量和GPT-2使用的差不多。



## 双流编码器架构：CLIP及相关改进工作

### CLIP: Connecting text and images with Contrastive Language–Image Pre-training

- 学习策略：多模态对比学习 属于ITM的一种

对比学习的目的是学习一个**嵌入函数**： $f: \chi \rightarrow E$ ，以使得相似样本在嵌入空间内的距离度量相近而不相似的远离。早期通常采用 triplet loss 训练：

$$L = \max(d(a, p) - d(a, n) + margin, 0)$$

多模态对比学习，则是针对不同模态的数据对，进行对比学习的操作（例如图像-文本对），对于给定的数据集中的文本和图像，我们分别利用映射函数将其映射到相同嵌入空间。在训练时，给定训练样本对，最大化匹配样本对的嵌入内积，最小化其他不匹配样本对的内积。实现这一途径的目标函数，我们称之为图文对比损失：image-text contrastive (ITC) loss

$$\mathcal{L}_{ITC} = \frac{1}{2}(\mathcal{L}_{I \rightarrow T} + \mathcal{L}_{T \rightarrow I}), \quad \mathcal{L}_{I \rightarrow T} = -\frac{1}{N} \sum_{i=1}^N \log \frac{\exp(d(z_i^I, z_i^T)/\tau)}{\sum_{k=1}^N \exp(d(z_i^I, z_k^T)/\tau)}, \quad \mathcal{L}_{T \rightarrow I} = -\frac{1}{N} \sum_{i=1}^N \log \frac{\exp(d(z_i^T, z_i^I)/\tau)}{\sum_{k=1}^N \exp(d(z_i^T, z_k^I)/\tau)},$$

# 双流编码器架构：CLIP及相关改进工作

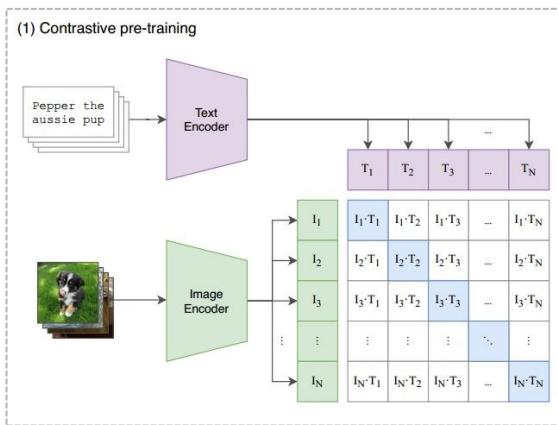
## CLIP: Connecting text and images with Contrastive Language–Image Pre-training

- Pretrained on ITM Task

$$\mathcal{L}_{ITC} = \frac{1}{2}(\mathcal{L}_{I \rightarrow T} + \mathcal{L}_{T \rightarrow I}),$$

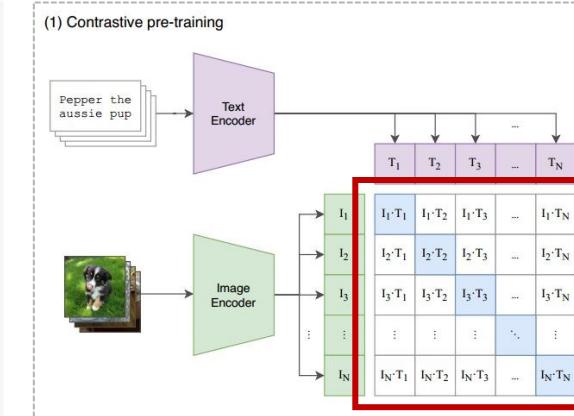
$$\mathcal{L}_{I \rightarrow T} = -\frac{1}{N} \sum_{i=1}^N \log \frac{\exp(d(z_i^I, z_i^T)/\tau)}{\sum_{k=1}^N \exp(d(z_i^I, z_k^T)/\tau)},$$

$$\mathcal{L}_{T \rightarrow I} = -\frac{1}{N} \sum_{i=1}^N \log \frac{\exp(d(z_i^T, z_i^I)/\tau)}{\sum_{k=1}^N \exp(d(z_i^T, z_k^I)/\tau)},$$



对图像和文本，分别应用两个Encoder，Image encoder可以是resnet或vit，而Text encoder采用transformer结构得到文本的embedding  $[T_1, T_2, \dots, T_N]$  以及图像embedding  $[I_1, I_2, \dots, I_N]$ ，将其分别进行内积计算，得到一个**内积矩阵**

```
# image_encoder - ResNet or Vision Transformer
# text_encoder - CBOW or Text Transformer
# I[n, h, w, c] - minibatch of aligned images
# T[n, 1] - minibatch of aligned texts
# W_i[d_i, d_e] - learned proj of image to embed
# W_t[d_t, d_e] - learned proj of text to embed
# t - learned temperature parameter
# extract feature representations of each modality
I_f = image_encoder(I) #[n, d_i]
T_f = text_encoder(T) #[n, d_t]
# joint multimodal embedding [n, d_e]
I_e = l2_normalize(np.dot(I_f, W_i), axis=1)
T_e = l2_normalize(np.dot(T_f, W_t), axis=1)
# scaled pairwise cosine similarities [n, n]
logits = np.dot(I_e, T_e.T) * np.exp(t)
# symmetric loss function
labels = np.arange(n)
loss_i = cross_entropy_loss(logits, labels, axis=0)
loss_t = cross_entropy_loss(logits, labels, axis=1)
loss = (loss_i + loss_t)/2
```



内积矩阵

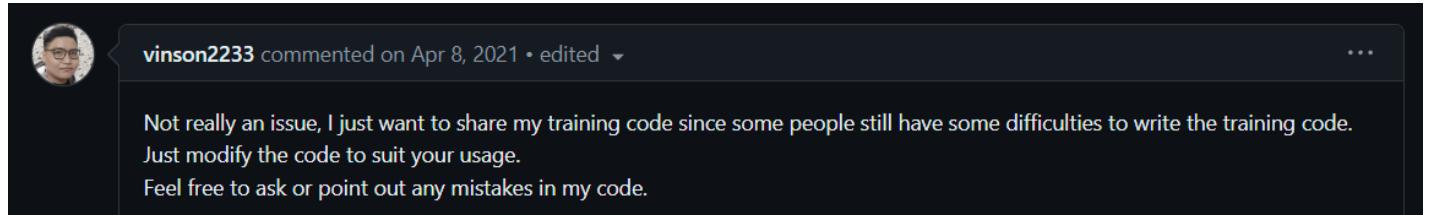
优化目标是最大化矩阵对角线上的元素，而最小化与不相关特征的内积（矩阵的其他元素）

损失函数：使用 **对称的cross-entropy loss**，对内积矩阵的每一行（同一张图片，不同文本标签）和每一列（同一个文本，不同的图片）分别求cross-entropy loss，将其加和作为总的损失

# 双流编码器架构：CLIP及相关改进工作

## CLIP: Connecting text and images with Contrastive Language–Image Pre-training

- 如何实现CLIP训练—以pytorch为例



vinson2233 commented on Apr 8, 2021 • edited

Not really an issue, I just want to share my training code since some people still have some difficulties to write the training code.  
Just modify the code to suit your usage.  
Feel free to ask or point out any mistakes in my code.

```
device = "cuda:0" if torch.cuda.is_available() else "cpu" # If using GPU then use mixed precision training.
model, preprocess = clip.load("ViT-B/32",device=device,jit=False) #Must set jit=False for training

class image_title_dataset(Dataset):
    def __init__(self, list_image_path,list_txt):

        self.image_path = list_image_path
        self.title = clip.tokenize(list_txt) #you can tokenize everything at once in here(slow at the beginning), or tok

    def __len__(self):
        return len(self.title)

    def __getitem__(self, idx):
        image = preprocess(Image.open(self.image_path[idx])) # Image from PIL module
        title = self.title[idx]
        return image,title

# use your own data
list_image_path = ['folder/image1.jpg','folder2/image2.jpg']
list_txt = ['description for image1.jpg' , 'description for image2.jpg']
dataset = image_title_dataset(list_image_path,list_txt)
train_dataloader = DataLoader(dataset,batch_size = BATCH_SIZE) #Define your own dataloader
```

# 双流编码器架构：CLIP及相关改进工作

## CLIP: Connecting text and images with Contrastive Language–Image Pre-training

- 如何实现CLIP训练—以pytorch为例

```
loss_img = nn.CrossEntropyLoss()
loss_txt = nn.CrossEntropyLoss()
optimizer = optim.Adam(model.parameters(), lr=5e-5,betas=(0.9,0.98),eps=1e-6,weight_decay=0.2) #Params used from paper,
# add your own code to track the training progress.
for epoch in range(EPOCH):
    for batch in train_dataloader :
        optimizer.zero_grad()

        images,texts = batch

        images= images.to(device)
        texts = texts.to(device)

        logits_per_image, logits_per_text = model(images, texts)

        ground_truth = torch.arange(len(images),dtype=torch.long,device=device)

        total_loss = (loss_img(logits_per_image,ground_truth) + loss_txt(logits_per_text,ground_truth))/2
        total_loss.backward()
        if device == "cpu":
            optimizer.step()
        else :
            convert_models_to_fp32(model)
            optimizer.step()
            clip.model.convert_weights(model)
```

- 思考：为什么ground-truth是一个  
arange张量？ $[0,1,2,3,\dots,\text{len}(\text{images})]$

# 双流编码器架构：CLIP及相关改进工作

## CLIP: Connecting text and images with Contrastive Language–Image Pre-training

- 如何实现CLIP训练—以pytorch为例

```
loss_img = nn.CrossEntropyLoss()
loss_txt = nn.CrossEntropyLoss()
optimizer = optim.Adam(model.parameters(), lr=5e-5,betas=(0.9,0.98),eps=1e-6,weight_decay=0.2) #Params used from paper,
# add your own code to track the training progress.
for epoch in range(EPOCH):
    for batch in train_dataloader :
        optimizer.zero_grad()

        images,texts = batch

        images= images.to(device)
        texts = texts.to(device)

        logits_per_image, logits_per_text = model(images, texts)

        ground_truth = torch.arange(len(images),dtype=torch.long,device=device)

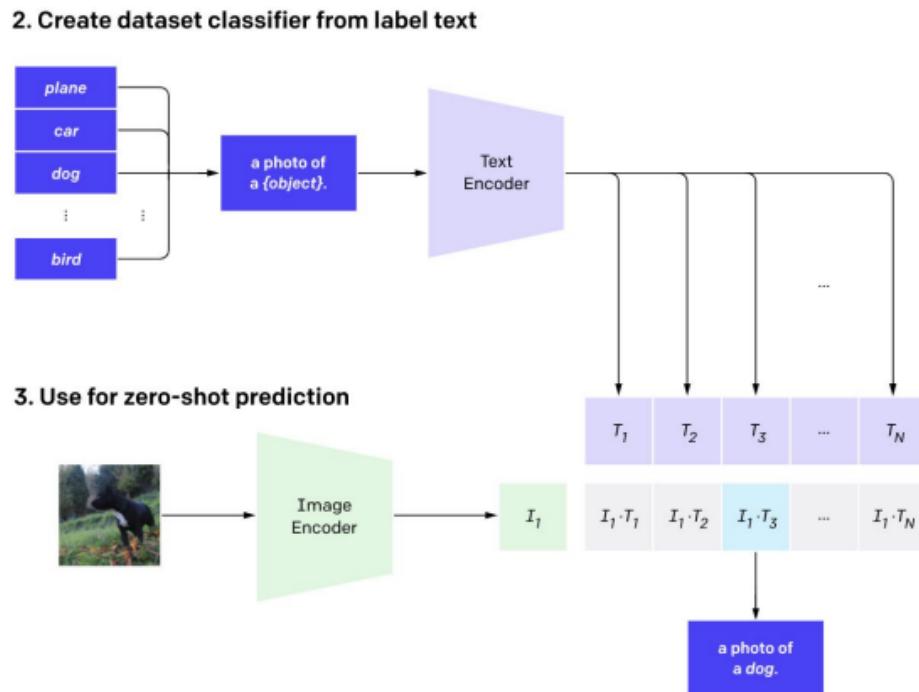
        total_loss = (loss_img(logits_per_image,ground_truth) + loss_txt(logits_per_text,ground_truth))/2
        total_loss.backward()
        if device == "cpu":
            optimizer.step()
        else :
            convert_models_to_fp32(model)
            optimizer.step()
            clip.model.convert_weights(model)
```

- 思考：为什么ground-truth是一个arange张量？ $[0,1,2,3,\dots,\text{len}(\text{images})]$
- 和以往的分类任务不同，假设我有[苹果，苹果，西瓜]的图像，那么交叉熵损失的输入label是 $[0,0,1]$
- 现在，对于每个图像，对应都有不同的文本模态，为了对齐两个模态，我们假设每个图像文本都是同一类的数据。那么，就需要在batch中创建一个arange张量代表虚拟的标签
- 使用对称的交叉熵，对齐文本到虚拟标签，以及图像到虚拟标签，从而使配对的图文对在特征上逼近

# 双流编码器架构：CLIP及相关改进工作

## CLIP: Connecting text and images with Contrastive Language–Image Pre-training

- CLIP出色的下游任务表现

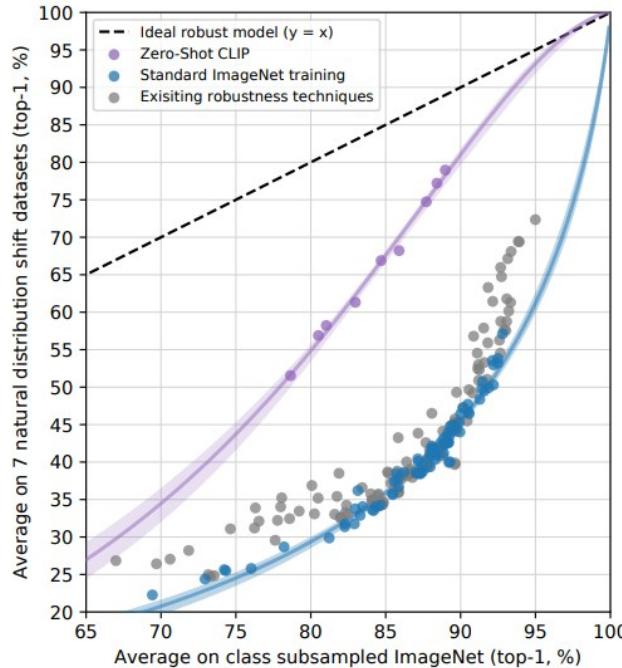


考虑大部分数据集通常是以单词作为标签，如“bird”、“plane”。因此采用**指示上下文的方式对标签进行扩展**：采用a photo of a <LABEL>. 作为推理阶段的文本输入。<LABEL>即为需要预测的zero-shot标签。将不同的标签扩展文本的embedding和待预测图像embedding求内积，得到最终预测的结果。而标签扩展可以是更详细的形式：A photo of a <LABEL>, a type of XXX 可以提升预测性能

# 双流编码器架构：CLIP及相关改进工作

## CLIP: Connecting text and images with Contrastive Language–Image Pre-training

- CLIP出色的下游任务表现
- 在zero-shot分类中能够准确预测各种不同的OOD样本



	Dataset Examples			ImageNet	Zero-Shot	CLIP	$\Delta$ Score
	ResNet101	CLIP	$\Delta$ Score				
ImageNet	76.2	76.2	0%				
ImageNetV2	64.3	70.1	+5.8%				
ImageNet-R	37.7	88.9	+51.2%				
ObjectNet	32.6	72.3	+39.7%				
ImageNet Sketch	25.2	60.2	+35.0%				
ImageNet-A	2.7	77.1	+74.4%				

相比于传统task-specific模型  
(ResNet101等)  
CLIP 在各类OOD基准数据集  
上的表现具有突破性的提升

# 双流编码器架构：CLIP及相关改进工作

## CLIP: Connecting text and images with Contrastive Language–Image Pre-training

- CLIP出色的下游任务表现
- 应用CLIP进行few-shot和Linear probe分类

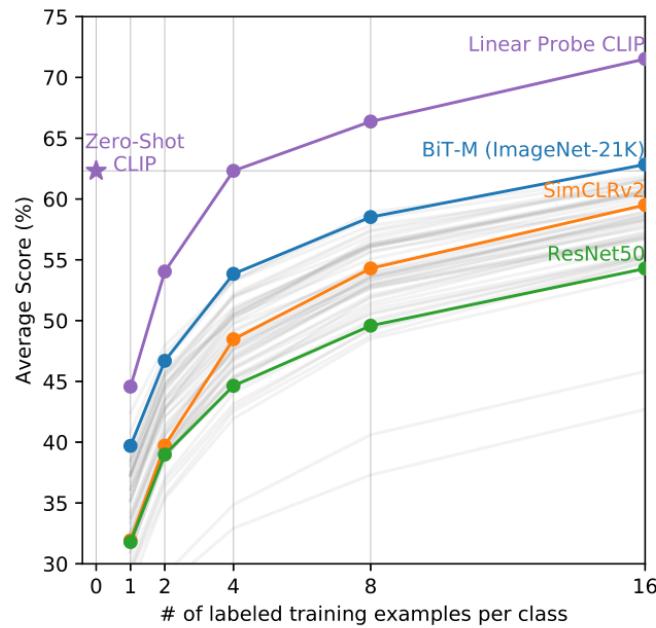
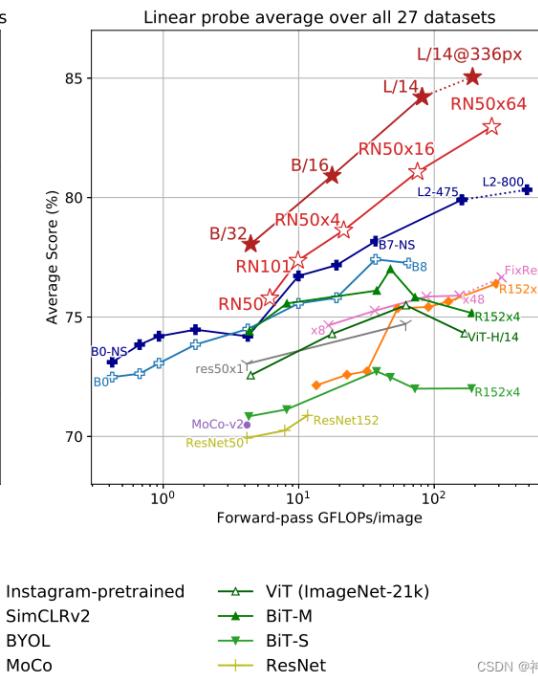
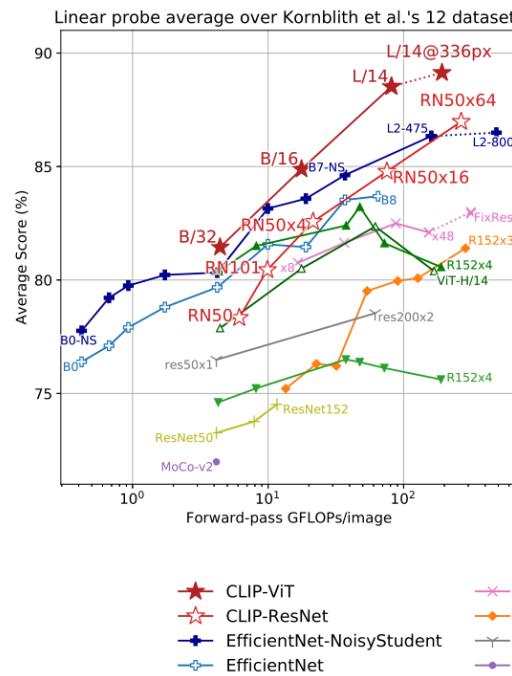


Figure 6. Zero-shot CLIP outperforms few-shot linear probes.

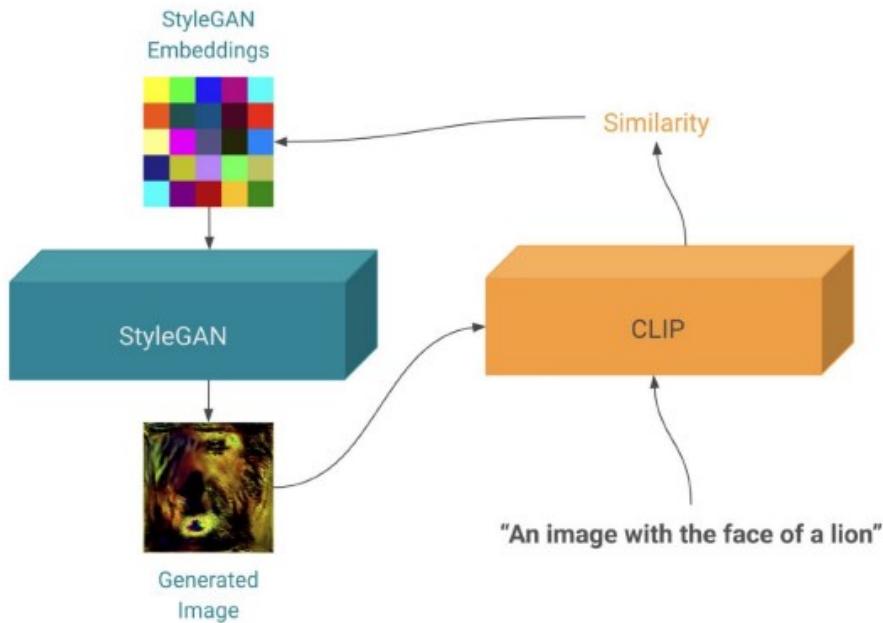


CSDN @神洛华

## 双流编码器架构：CLIP及相关改进工作

### CLIP: Connecting text and images with Contrastive Language–Image Pre-training

- CLIP出色的下游任务表现
- Image Generation: CLIP + StyleGAN



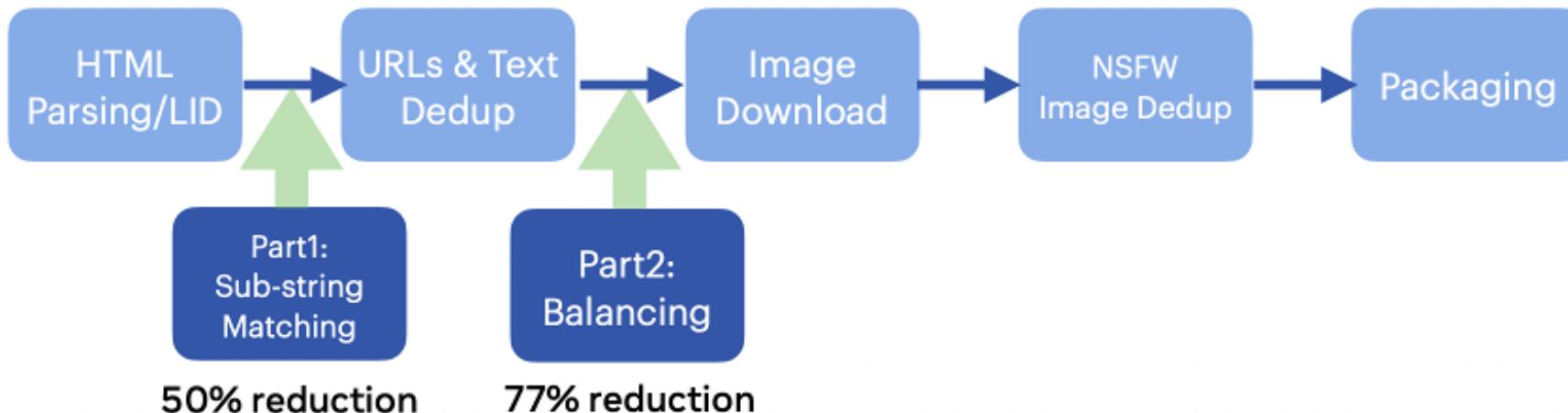
"An image with the face of Elon Musk with blonde hair"



## 双流编码器架构：CLIP及相关改进工作

### MetaCLIP: Demystifying CLIP Data

- 证明CLIP的数据质量对模型表现至关重要
- OpenAI并未提供训练集和数据筛选算法，本文提出了CLIP可能采用的数据算法，简化和产生高质量的训练数据
- 创建元数据：使用WordNet和Wikipedia的高质量视觉概念（visual concept）
- 数据筛选算法：包括元数据字符串匹配和平衡数据分布，使得1.6B的原始数据被下采样成400M的训练数据，降低头部分布的冗余无效信息和头部数据的噪声（比如untitled photo），所有长尾分布的视觉概念全部保留



# 双流编码器架构：CLIP及相关改进工作

## MetaCLIP: Demystifying CLIP Data

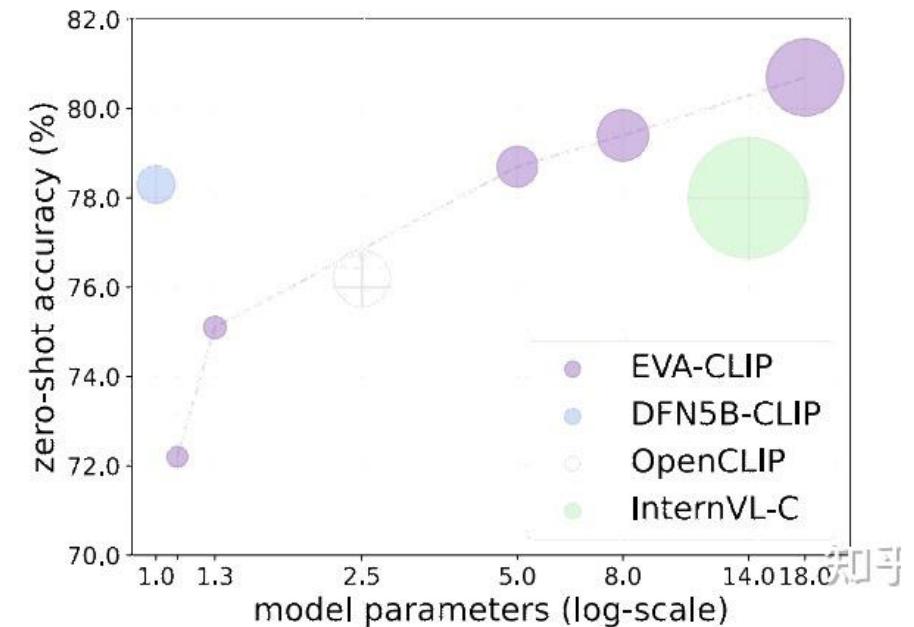
- 证明CLIP的数据质量对模型表现至关重要
- OpenAI并未提供训练集和数据筛选算法，本文提出了CLIP可能采用的数据算法，简化和产生高质量的训练数据
- 创建元数据：使用WordNet和Wikipedia的高质量视觉概念（visual concept）
- 数据筛选算法：包括元数据字符串匹配和平衡数据分布，使得1.6B的原始数据被下采样成400M的训练数据，降低头部分布的冗余无效信息和头部数据的噪声（比如untitled photo），所有长尾分布的视觉概念全部保留

Average	ImageNet	Food-101	CIFAR10	CIFAR100	CUB	SUN397	Cars	Aircraft	DTD	Pets	Caltech-101	Flowers	MNIST	FER-2013	STL-10	EuroSAT	RESISC45	GTSRB	KITTI	Country211	PCAM	UCF101	Kinetics700	CLEVR	HatefulMemes	SST2	
ViT-B/32																											
CLIP, our eval.	56.6	63.4	83.7	89.8	65.1	53.7	62.0	59.7	19.6	44.0	87.2	87.4	66.9	48.2	46.6	97.1	44.9	61.0	32.6	28.7	17.2	62.5	63.9	48.0	23.6	56.4	58.6
OpenCLIP, our eval.	57.6	62.9	80.7	90.7	70.6	61.2	66.4	79.2	16.7	54.5	86.5	90.7	66.1	37.4	48.2	95.6	52.2	58.0	42.0	38.0	14.8	50.1	63.0	42.8	22.5	53.3	52.3
<b>MetaCLIP</b>	<b>58.2</b>	65.5	80.6	91.3	70.2	63.4	63.0	70.7	26.8	52.8	88.7	91.9	68.5	41.5	35.9	95.4	52.6	64.2	35.8	30.7	17.2	55.5	66.1	45.4	30.6	56.4	53.4
ViT-B/16																											
CLIP, our eval.	59.6	68.3	88.8	90.8	68.2	55.6	64.0	64.6	24.0	45.1	88.9	89.1	69.4	51.8	53.0	98.2	54.8	65.5	43.3	21.7	22.8	56.3	68.5	52.3	25.5	58.7	60.5
OpenCLIP, our eval.	60.4	67.0	85.8	91.7	71.4	65.3	69.2	83.6	17.4	51.0	89.2	90.8	66.5	66.3	46.1	97.0	52.2	65.7	43.5	23.7	18.1	51.7	67.0	46.2	33.9	54.5	54.4
<b>MetaCLIP</b>	<b>61.1</b>	70.8	86.8	90.1	66.5	70.8	66.6	74.1	27.9	55.9	90.4	93.8	72.3	47.8	44.6	97.2	55.4	68.8	43.8	33.4	22.6	52.9	68.0	49.5	22.8	54.8	60.6
ViT-L/14																											
CLIP, our eval.	65.7	75.5	93.0	95.6	78.3	63.3	66.8	77.8	31.3	55.3	93.6	93.3	79.3	76.4	56.9	99.4	61.9	70.9	50.6	19.2	31.9	50.1	75.7	60.2	22.3	59.7	68.9
OpenCLIP, our eval.	64.5	72.7	90.0	94.7	78.0	73.9	72.4	89.5	24.7	60.2	91.6	93.6	73.0	76.1	54.3	98.1	63.9	69.6	49.9	16.0	23.0	51.7	71.5	51.6	25.4	55.3	56.0
<b>MetaCLIP</b>	<b>67.1</b>	76.2	90.7	95.5	77.4	75.9	70.5	84.7	40.4	62.0	93.7	94.4	76.4	61.7	46.5	99.3	59.7	71.9	47.5	29.9	30.9	70.1	75.5	57.1	35.1	56.6	65.6

## 双流编码器架构：CLIP及相关改进工作

### EVA-CLIP-18B: Scaling CLIP to 18 Billion Parameters

- 用更大规模的数据和模型架构训练CLIP
- 原始的CLIP最大权重为ViT-L/14-336px，大约500M的参数量，模型越大是否表现越好？
- 模型规模：180亿参数量，迄今为止最强大的开源CLIP模型
- 训练样本量：该模型在观看了60亿个训练样本后取得显著性能。
- 性能表现：在27个广泛认可的图像分类基准测试中平均实现了80.7%的零样本顶级准确率。
- 性能饱和：**EVA-CLIP-18B未显示出性能饱和的迹象，意味着有进一步提高其性能的潜力，尤其在视觉模型的扩展方面。



# 双流编码-解码器架构：BLIP系列

## BLIP: Bootstrapping Language-Image Pretraining

**Bootstrapping** (improve your situation by your own efforts) — 在线的数据标注和清理，把处理好的数据继续用来迭代原模型

salesforce在2022年提出的多模态框架，是**理解和生成的统一**，引入了跨模态的编码器和解码器，实现了跨模态信息流动，在多项视觉和语言任务取得SOTA。在AIGC中通常用来给图像生成prompt，好的prompt对交叉注意力的微调非常关键，例如ControlNet中的Automatic Prompt就是BLIP生成的。

### Image Description Generation

Try the provided image or upload your own:  
Drag and drop file here  
Limit 200MB per file

Image

Generate

Description

the golden gate bridge is illuminated with bright lights in blue twilight sky

### Zero-shot Classification

Image

Categories

category 1	category 2
merlion	elephant
category 3	category 4
giraffe	fountain
category 5	category 6
marina bay	singapore river

Submit

Prediction

Category	Probability (%)
merlion	60.78%
elephant	0.00%
giraffe	0.00%
fountain	8.77%
marina bay	28.06%
singapore river	2.38%

### Text Localization

Person playing with dog on the beach.

Submit

Image

GradCam

person      playing      with      dog

### Visual Question Answering

Image

Question

Input your question!

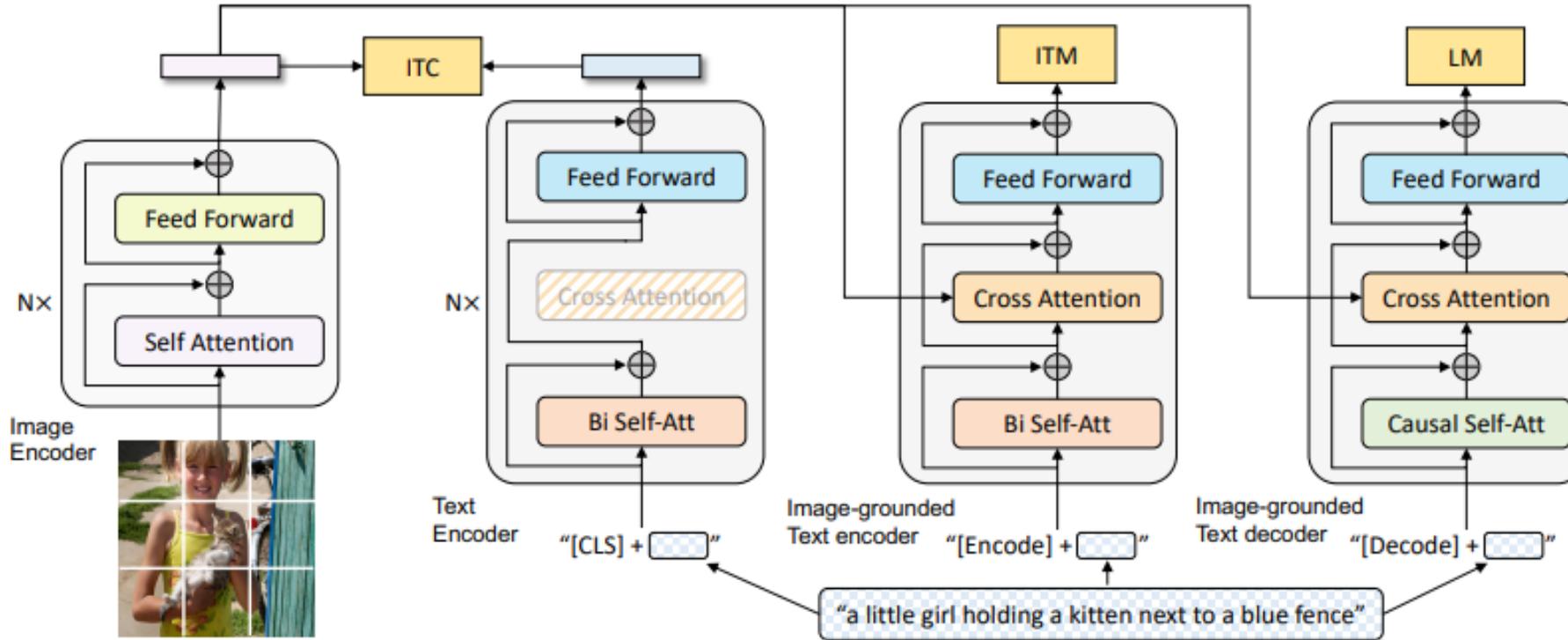
What is the dog's breed?

Answer

golden retriever

## 双流编码-解码器架构：BLIP系列

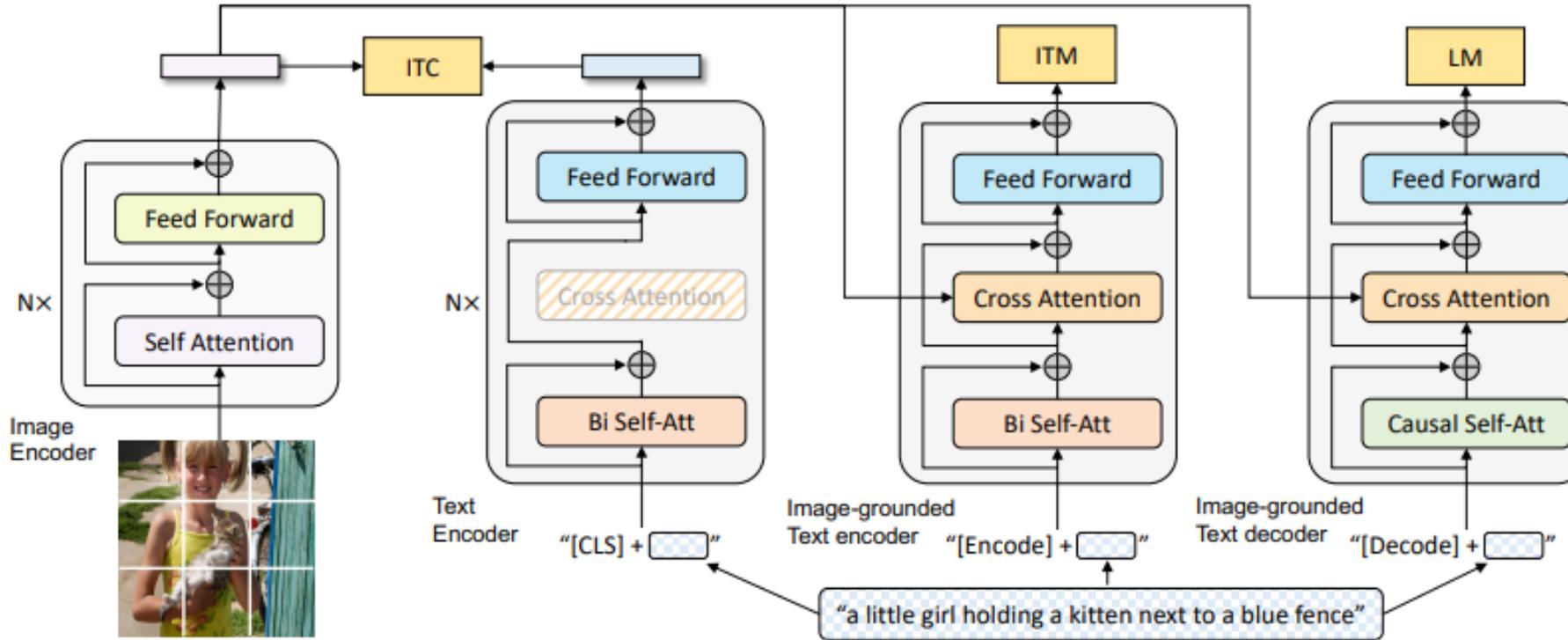
### BLIP: Bootstrapping Language-Image Pretraining



- BLIP引入了编码器-解码器的多模态混合结构MED ( Multimodal mixture of Encoder-Decoder)
- 能够有效地进行多任务预学习和迁移学习。MED包括两个单模态编码器 (Image Encoder, Text Encoder) , 一个以图像为基础的编码器 (image-grounded text encoder) 和一个以图像为基础的解码器 (image-grounded text decoder)

## 双流编码-解码器架构：BLIP系列

### BLIP: Bootstrapping Language-Image Pretraining

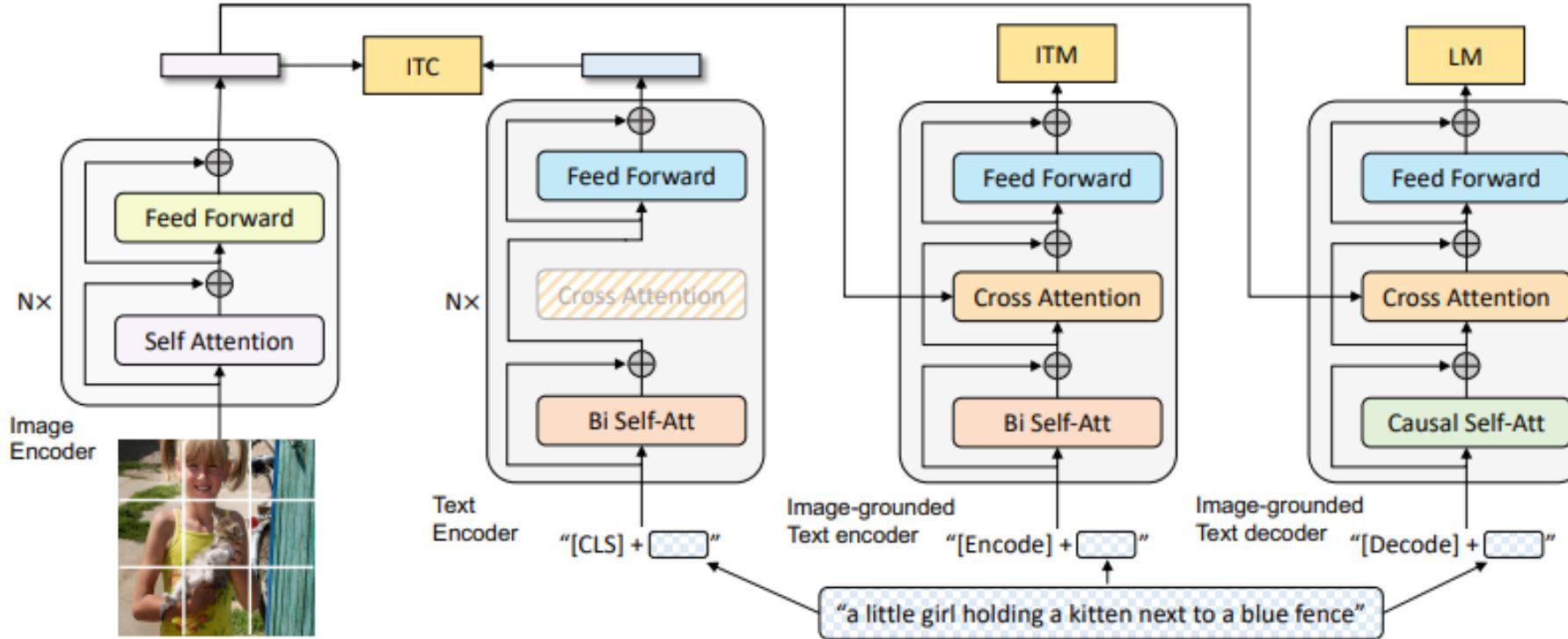


- **训练目标：通过三个损失函数联合进行预训练：**

1. 图像-文本对比损失 ITC (Image-Text Contrastive Loss)：针对图像编码器和文本编码器，通过正负图文对的对比学习，来对齐图像和文本的潜在特征空间。

# 双流编码-解码器架构：BLIP系列

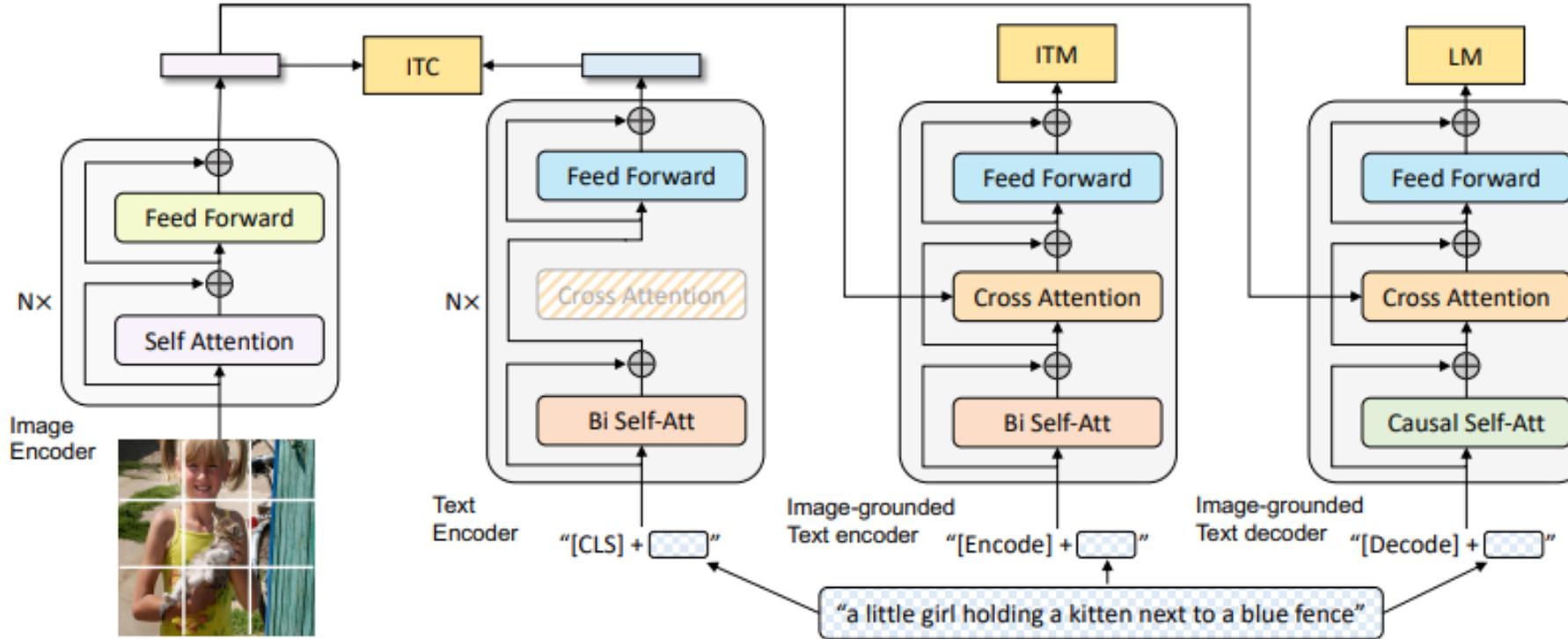
## BLIP: Bootstrapping Language-Image Pretraining



- **训练目标：通过三个损失函数联合进行预训练：**
  2. 图像-文本匹配损失 ITM (Image-Text Matching Loss) : 针对以图像为基础的文本编码器，通过对图文匹配性进行二分类，建模图文多模态信息的相关性。

## 双流编码-解码器架构：BLIP系列

### BLIP: Bootstrapping Language-Image Pretraining

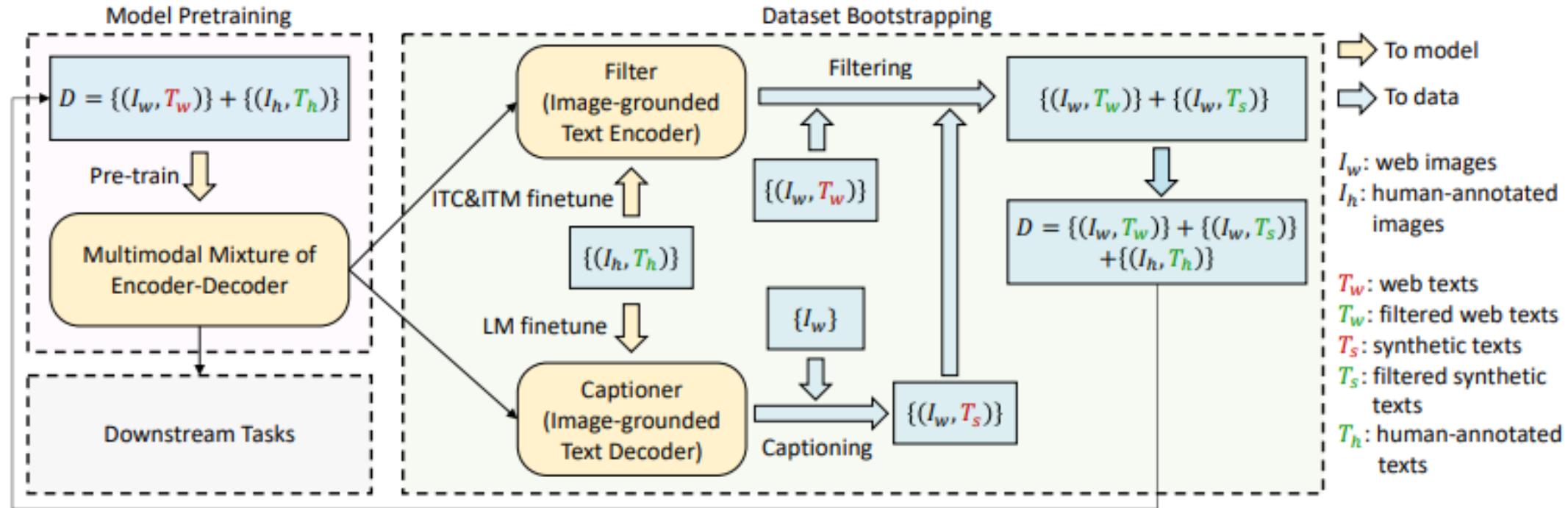


- **训练目标：通过三个损失函数联合进行预训练：**

3. 语言建模损失 LM (Language Modeling Loss) : 针对以图像为基础的文本解码器，通过交叉熵损失进行优化，训练模型以自回归的方式生成目标caption。

# 双流编码-解码器架构：BLIP系列

## BLIP: Bootstrapping Language-Image Pretraining

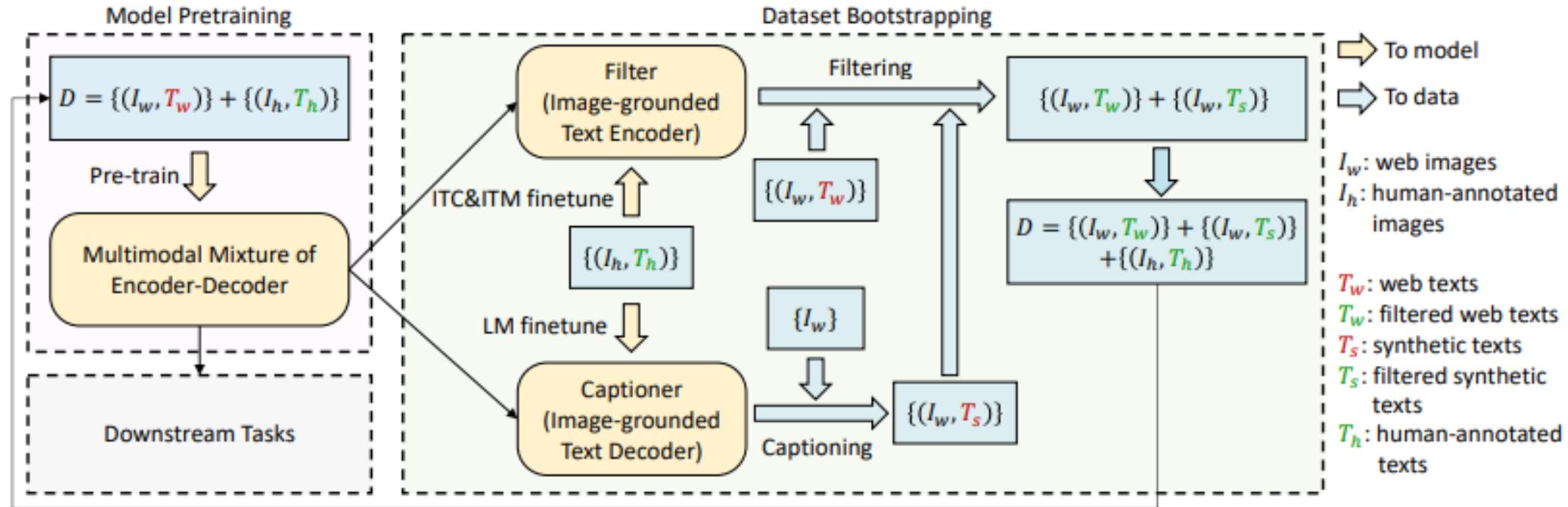


- 训练方法：Bootstrapping的方式获取更好的训练样本

网络上获得的图文对通常包含许多不准确甚至错误的信息，为了有效利用这种形态的数据，BLIP提出caption生成和过滤模块CapFilt (Captioning and Filtering)，首先从噪声图文对中学习，然后生成和过滤产生新的数据集，再去迭代优化原模型。

# 双流编码-解码器架构：BLIP系列

## BLIP: Bootstrapping Language-Image Pretraining



- 训练方法：Bootstrapping的方式在训练的同时获取更好的训练样本

CapFilt包含两个模块：一个是captioner，给网络图像生成caption，另一个是Filter，过滤原始网络文本和合成文本中的噪声caption。这一过程十分类似于GPT中引入的Reward model，通过给予反馈不断提升模型的性能



北京航空航天大學  
BEIHANG UNIVERSITY

人工智能研究院  
Institute of Artificial Intelligence

# Intro to Vision Large-Language Models ( VLLMs )

# VLLMs 和 VLP 的区别与关联

## 区别：

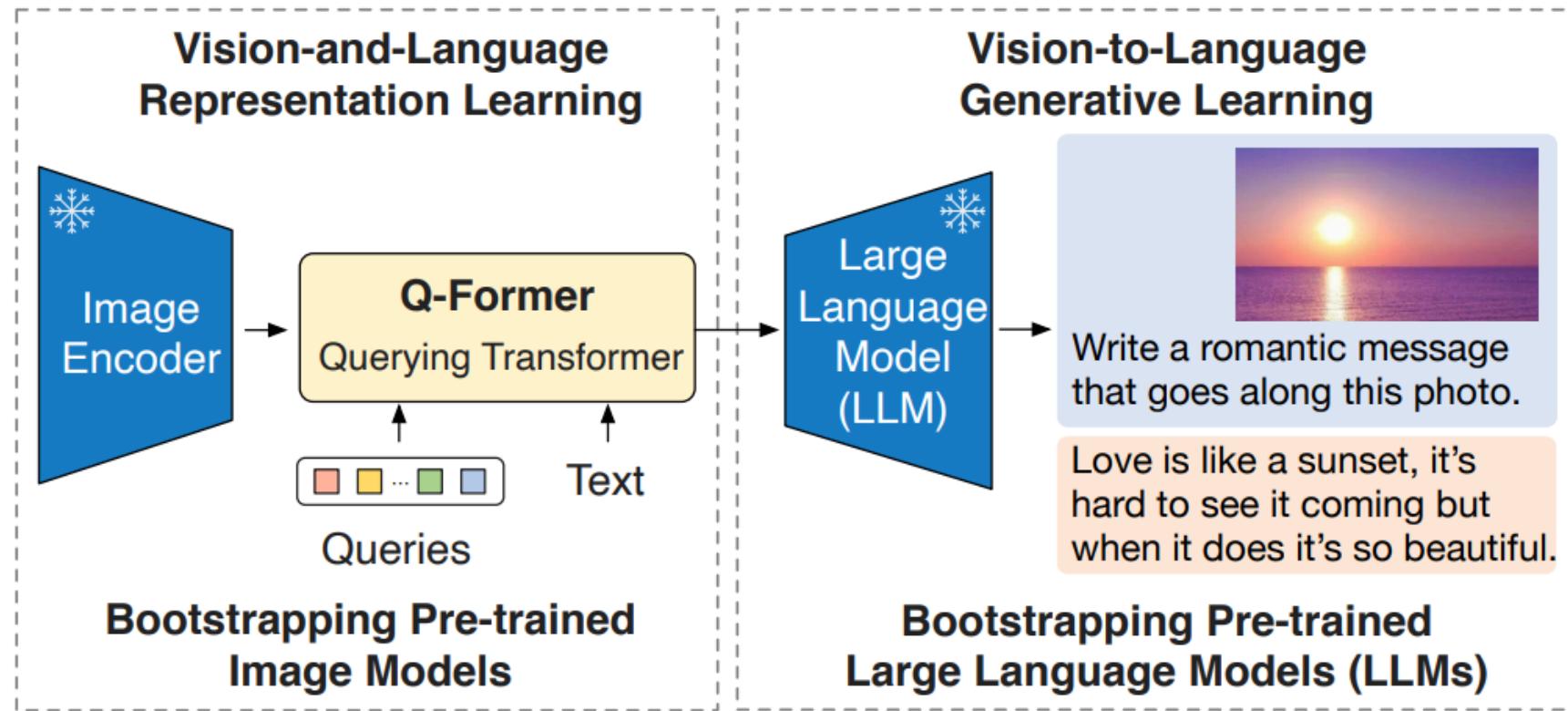
- VLP：专注于如何通过结合视觉和语言信息来预训练模型。以生成可以广泛应用于多种下游任务的通用特征和表示。
- VLLMs：更广泛地指代任何结合视觉和语言能力的大型模型，不限于预训练策略。VLLMs可以是专门为特定任务设计或优化的，也可以是基于VLP技术进一步发展的模型。

## 关联：

- VLLM 可能包括VLP作为其组成之一，更具体而言，VLP是实现VLLMs的一种技术途径和重要组成模块（常作为基础的特征提取 / 解码器模块）
- 随着研究的深入和技术的发展，VLP可以被视为是VLLMs发展的一个分支或初级阶段，VLLMs则代表了在视觉语言模型领域内更广泛和深入应用的发展方向。

## 使用Q-former连接图像编码器和LLMs: BLIP-2

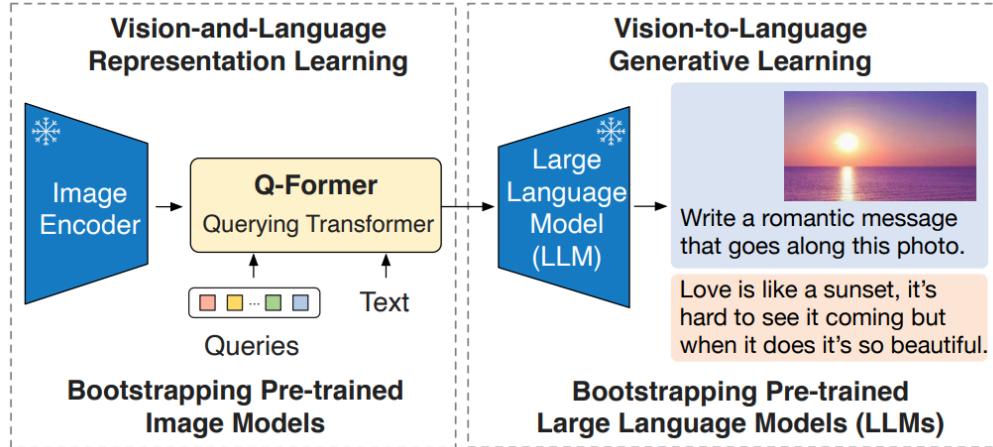
### BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and LLMs



Salesforce在2023年提出 BLIP-2，通过利用预训练的视觉模型和语言模型来提升多模态效果和降低训练成本，预训练的视觉模型能够提供高质量的视觉表征，预训练的语言模型则提供了强大的语言生成能力。

# 使用Q-former连接图像编码器和LLMs: BLIP-2

## BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and LLMs



```
# pretrained image encoder
self.vision_model = Blip2VisionModel(config.vision_config)
# input query
self.query_tokens = nn.Parameter(torch.zeros(1, config.num_query_tokens, config.qformer_config.hidden_size))
self.qformer = Blip2QFormerModel(config.qformer_config)

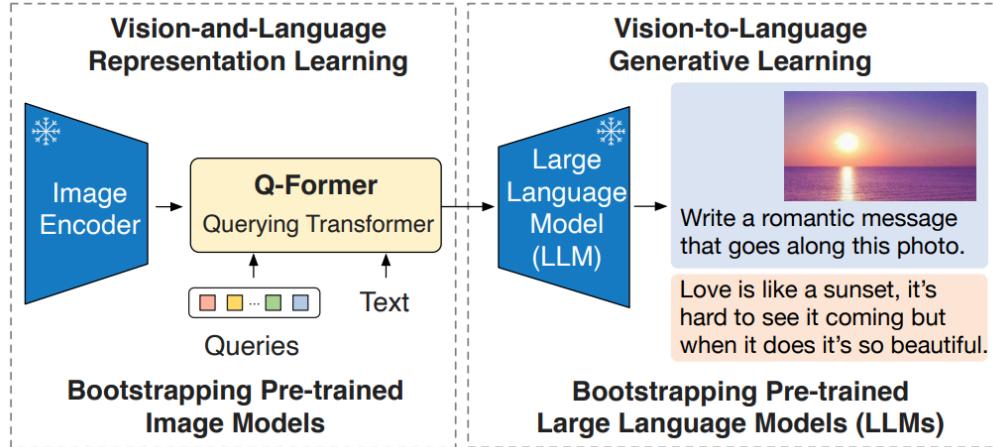
self.language_projection = nn.Linear(config.qformer_config.hidden_size, config.text_config.hidden_size)
if config.use_decoder_only_language_model:
    language_model = AutoModelForCausalLM.from_config(config.text_config)
else:
    language_model = AutoModelForSeq2SeqLM.from_config(config.text_config)
# pretrained large language model
self.language_model = language_model
```

**模型架构:** BLIP-2 由预训练的Image Encoder, 预训练的Large Language Model, 和一个可学习的 Q-Former 组成

- Image Encoder: 负责从输入图片中提取视觉特征, 本文试验了两种网络结构, CLIP 训练的 ViT-L/14 和 EVA-CLIP 训练的 ViT-g/14
- Large Language Model: 负责文本生成, 本文试验了 decoder-based LLM and encoder-decoder-based LLM
- **Q-Former:** 负责弥合视觉和语言两种模态的差距, 由 Image Transformer 和 Text Transformer 两个子模块构成, 它们 **共享相同自注意力层**。

# 使用Q-former连接图像编码器和LLMs: BLIP-2

## BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and LLMs



```
# pretrained image encoder
self.vision_model = Blip2VisionModel(config.vision_config)
# input query
self.query_tokens = nn.Parameter(torch.zeros(1, config.num_query_tokens, config.qformer_config.hidden_size))
self.qformer = Blip2QFormerModel(config.qformer_config)

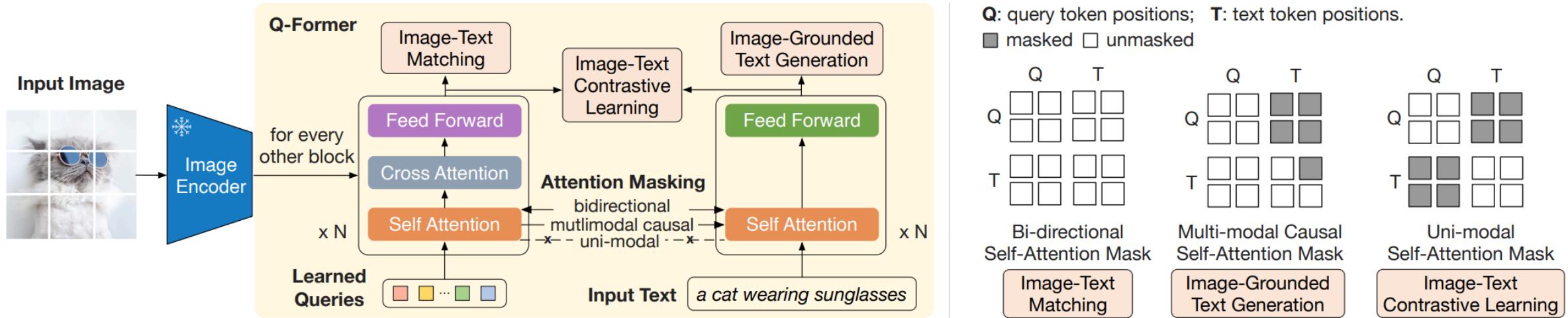
self.language_projection = nn.Linear(config.qformer_config.hidden_size, config.text_config.hidden_size)
if config.use_decoder_only_language_model:
    language_model = AutoModelForCausalLM.from_config(config.text_config)
else:
    language_model = AutoModelForSeq2SeqLM.from_config(config.text_config)
# pretrained large language model
self.language_model = language_model
```

模型架构: BLIP-2 由预训练的Image Encoder, 预训练的Large Language Model, 和一个可学习的 Q-Former 组成

- **Q-Former:** 负责弥合视觉和语言两种模态的差距, 由**Image Transformer**和**Text Transformer**两个子模块构成
- **Image Transformer**通过与**图像编码器**进行交互提取视觉特征, 它的输入是**可学习的 Query**, 这些Query通过**自注意力层**相互交互, 并通过**交叉注意力层与冻结的图像特征交互**, 还可以通过**共享的自注意力层与文本进行交互**。
- **Text Transformer**作为文本编码器和解码器, 它的自注意力层与**Image Transformer**共享, 根据预训练任务, 应用不同的自注意力掩码来控制Query和文本的交互方式。

# 使用Q-former连接图像编码器和LLMs: BLIP-2

## BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and LLMs

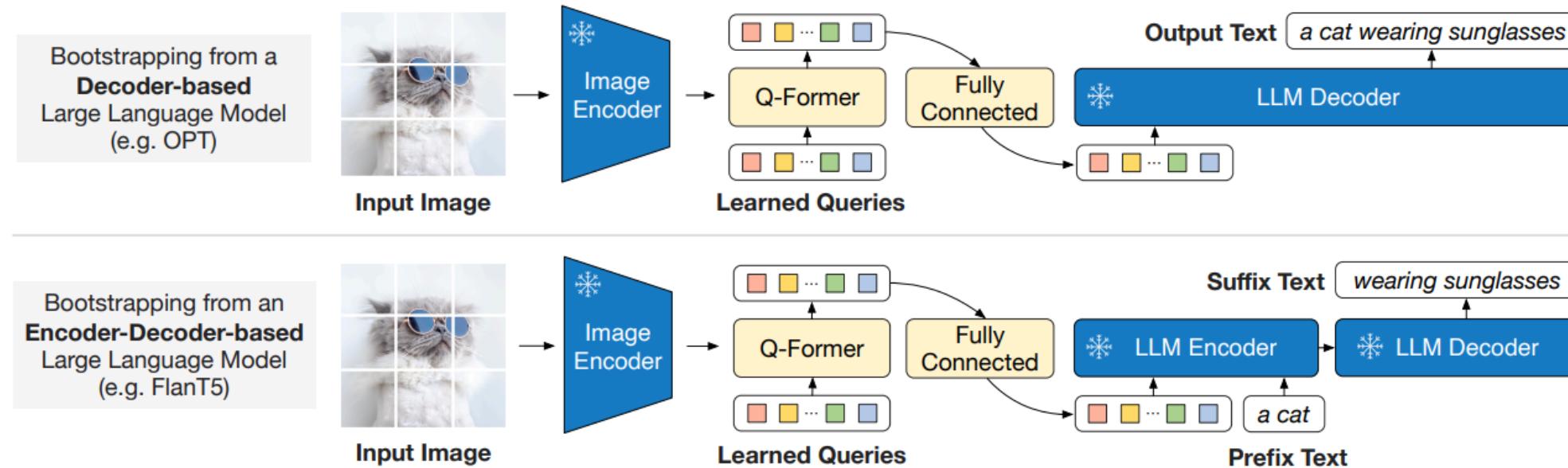


### 训练范式：两阶段训练Q-former—表示学习阶段和生成学习阶段

- 将 Q-Former 连接到冻结的 Image Encoder，训练集为图像-文本对，通过联合优化三个预训练目标，在Query和Text之间分别采用不同的注意力掩码策略，从而控制Image Transformer和Text Transformer的交互方式。
- 目标1： **ITC(Image-Text Contrastive Learning)** 对齐图像嵌入和文本嵌入
- 目标2： **ITG(Image-grounded Text Generation)** 给定输入图像作为条件的情况下，训练 Q-Former 生成文本。
- 目标3： **ITM( Image-Text Matching)** 预测图像-文本对是正匹配还是负匹配，学习图像和文本表示之间的细粒度对齐

# 使用Q-former连接图像编码器和LLMs: BLIP-2

## BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and LLMs

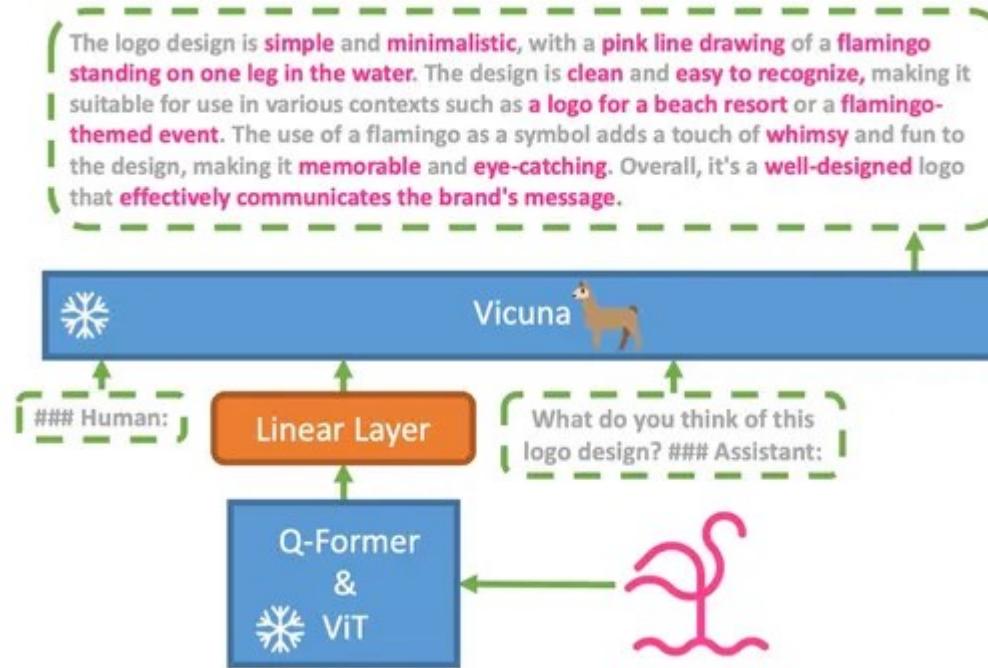


### 训练范式: 两阶段训练Q-former—表示学习阶段和生成学习阶段

- 将 Q-Former 连接到冻结的 LLM，以利用 LLM 的语言生成能力。使用全连接层将输出的 Query 嵌入线性投影到与 LLM 的文本嵌入相同的维度，然后将投影的 Query 嵌入添加到输入文本嵌入前面
- 由于 Q-Former 已经过预训练，可以提取包含语言信息的视觉表示，因此它可以有效地充当信息瓶颈，将最有用的信息提供给 LLM，同时删除不相关的视觉信息，减轻了 LLM 学习视觉语言对齐的负担。

## 使用投影层链接视觉模型和LLMs: MiniGPT4系列

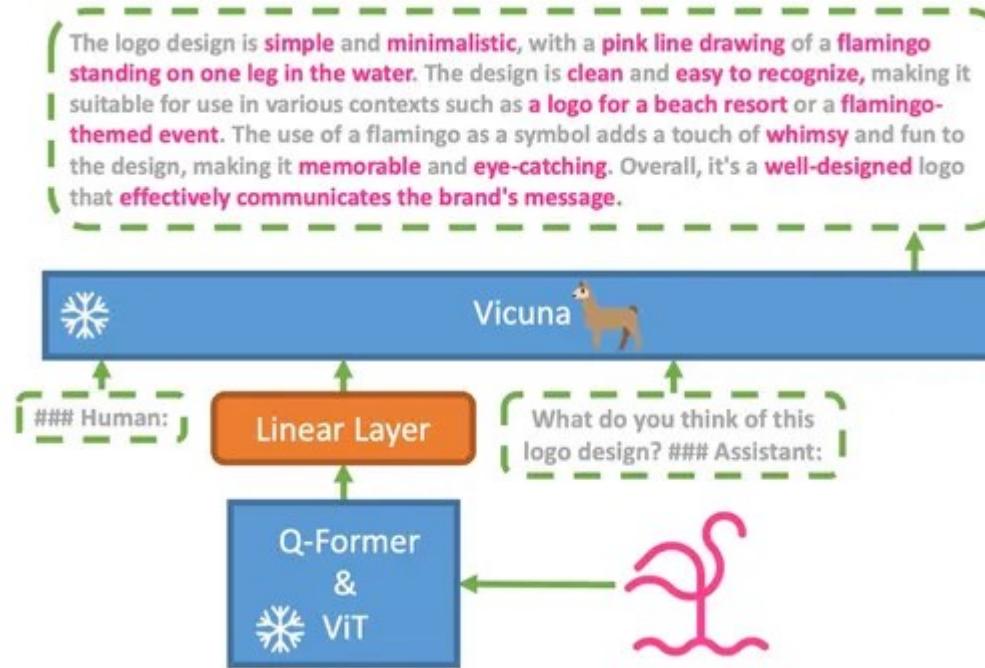
### MiniGPT-4: Enhancing Vision-language Understanding with Advanced Large Language Models



- 和BLIP-2的思想类似，MiniGPT4认为增强多模态模型的生成能力一定要利用**大语言模型的知识**。因此也是采用**“对齐”**的思路来训练一个投影层将预训练好的LLM和视觉模型链接在一起。这也是当前VLLMs的主要实现途径
- 简单概括：通过投影层来将图片的特征转化为LLM能够理解的token，剩下的就和LLM做的一样了（进行视觉语言的自回归训练），这也是猜测GPT-4V所可能采用的范式

# 使用投影层链接视觉模型和LLMs: MiniGPT4系列

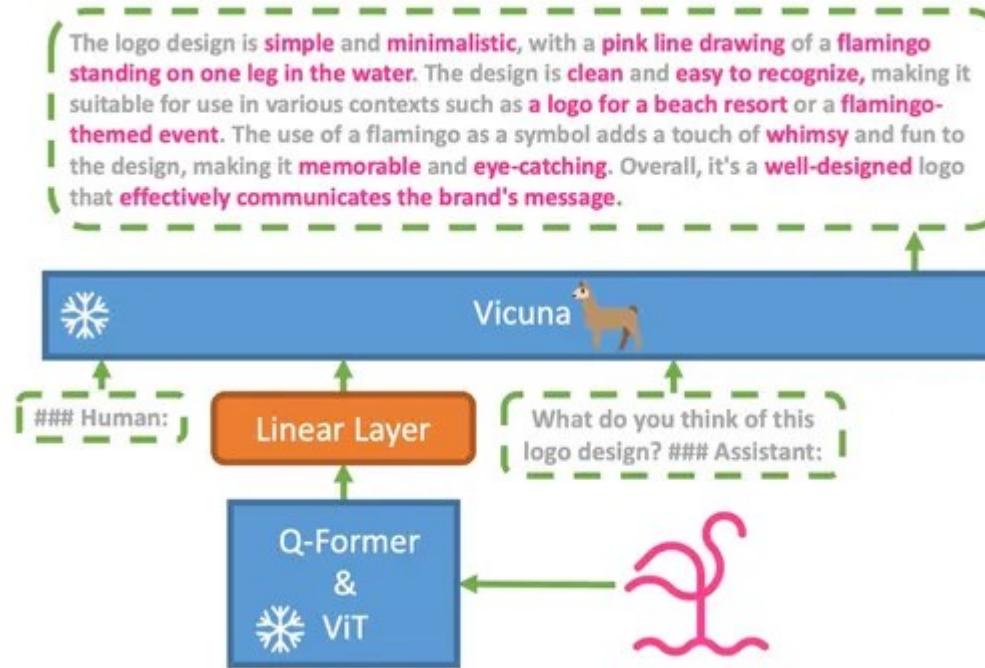
## MiniGPT-4: Enhancing Vision-language Understanding with Advanced Large Language Models



- image encode采用的是预训练的ViT和QFormer
- LLM采用的是Vicunna
- 对齐层采用的是简单的线性层

# 使用投影层链接视觉模型和LLMs: MiniGPT4系列

## MiniGPT-4: Enhancing Vision-language Understanding with Advanced Large Language Models



- **训练范式: 两阶段训练 (多模态大多用多阶段训练的思路)**
- 第一阶段: 使用大规模图文对数据, 优化投影层的参数, 将image encoder提取的图片特征转变成LLM能够理解的形式
- 第二阶段: 指令集数据集微调, 优化模型生成文本的质量

# 使用投影层链接视觉模型和LLMs: MiniGPT4系列

## MiniGPT-4: Enhancing Vision-language Understanding with Advanced Large Language Models

- 如何制作指令集数据：取决于希望模型解决什么类型的任务
- 1. Conceptual Caption数据集随机挑选5000图片。用stage1的模型来生成详细的图片描述。所用的prompt如下：

```
###Human: <Img><ImageFeature></Img>Describe this image in detail. Give as many details as possible. Say everything you see. ###Assistant:
```

- 为了避免生成caption答非所问和语法错误的情况，作者对生成的每一个图文对进行了人工check，以此保证数据集的质量。清洗后获得了总计3500个图文对。

虽然作者主要构建了image caption的数据。我们训练自己的多模态模型的时候，可以根据需要构建自己的instruction数据集。

【例如：open set检测】

prompt: ###Human: <Img><ImageFeature></Img> . 请返回图中所有任务的坐标，返回  
格式以json输出，模版如：{person1: [xmin, ymin, xmax, ymax]},其中坐标做了归一化处理  
###Assistant:

label: {person1: [0.14, 0.24, 0.45, 0.65], person2: [0.11, 0.]...}

【例如：表格信息抽取】

prompt: ###Human: <Img><ImageFeature></Img> . 请抽取表格中的所有3元组

###Assistant:

label: [(分类1, 分类2, value), (分类1, 分类2, value), ...] 总结来说就是把任务的统一建模成sequence-to-sequence的架构。这样就能联合起来一起训练。相当于同一种统一的架构做多任务学习。

# 视觉指令微调：LLaVa系列

## Visual Instruction Tuning

### LLaVA

Haotian Liu<sup>1\*</sup>, Chunyuan Li<sup>2\*</sup>, Qingyang Wu<sup>3</sup>, Yong Jae Lee<sup>1</sup>

<sup>1</sup>University of Wisconsin–Madison    <sup>2</sup>Microsoft Research    <sup>3</sup>Columbia University  
<https://llava-vl.github.io>

#### Improved Baselines with Visual Instruction Tuning

### LLaVA-1.5

Haotian Liu<sup>1</sup>    Chunyuan Li<sup>2</sup>    Yuheng Li<sup>1</sup>    Yong Jae Lee<sup>1</sup>  
<sup>1</sup>University of Wisconsin–Madison    <sup>2</sup>Microsoft Research, Redmond  
<https://llava-vl.github.io>

## LLAVA-PLUS: LEARNING TO USE TOOLS FOR CREATING MULTIMODAL AGENTS

### LLaVA-PLUS

Shilong Liu<sup>♣\*</sup>, Hao Cheng<sup>♦</sup>, Haotian Liu<sup>◇\*</sup>, Hao Zhang<sup>♡\*</sup>, Feng Li<sup>♡\*</sup>,  
Tianhe Ren<sup>▶</sup>, Xueyan Zou<sup>◇\*</sup>, Jianwei Yang<sup>♦</sup>, Hang Su<sup>♣</sup>, Jun Zhu<sup>♣</sup>,  
Lei Zhang<sup>▶</sup>, Jianfeng Gao<sup>♦</sup>, Chunyuan Li<sup>♣</sup><sup>■</sup>

<sup>♣</sup>Dept. of Comp. Sci. & Tech., Institute for AI, BNRIst, Tsinghua University

<sup>♦</sup>Microsoft Research, Redmond

<sup>◇</sup>University of Wisconsin-Madison    <sup>♡</sup>HKUST    <sup>▶</sup>IDEA Research

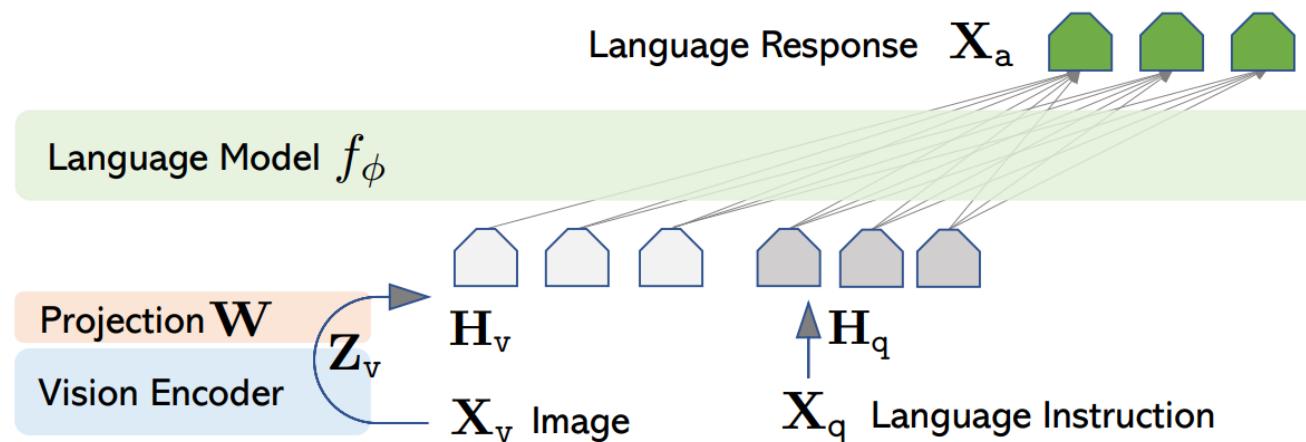
\*Work performed during an internship at Microsoft    <sup>■</sup> Project Lead

- LLaVA(Large Language and Vision Assistant), 即大型语言和视觉助手，是一个端到端训练的大型多模态模型，将视觉编码器和大语言模型连接起来实现通用的视觉和语言理解。当前开源社区中应用较广的VLLMs

# 视觉指令微调：LLaVa系列

## LLaVa: Visual Instruction Tuning

- LLaVa的主要贡献
- **多模态指令跟随数据集：**使用 ChatGPT/GPT-4 将图像文本对转换为适当的指令遵循数据格式，生成三种类型的指令跟随数据：对话、详细描述、复杂推理。总共收集了 158K 个语言图像指令跟随数据样本，其中对话样本 58K，详细描述样本 23K，复杂推理样本 77K。
- 将 CLIP 的视觉编码器与LLaMA语言解码器连接起来，并在生成的视觉语言指令跟随数据集上对它们进行端到端的微调。
- 开源：向公众发布生成的多模态指令数据、用于数据生成和模型训练的代码库、模型权重文件和可视化聊天demo。



## A summary

- 包括LLaVa在内，目前主流的VLLMs大多属于遵循指令的大型多模态模型（Instruction-following LMM），常见架构包括用于编码视觉特征的**预训练视觉主干网络**、用于理解用户指令并产生响应的**预训练大语言模型 (LLM)**，以及用于将视觉编码器输出与语言模型对齐的**视觉语言跨模态连接器**。训练此类模型通常遵循两阶段的方法：
  - 视觉语言对齐预训练阶段：**利用图像文本对将视觉特征与语言模型的词嵌入空间对齐。
  - 视觉指令调整阶段：**根据视觉指令调整模型，使模型能够遵循用户的不同指令要求。

Discovering unusual content



Describe the image.

The image shows a cactus plant standing in the middle of a frozen lake. The cactus is surrounded by large ice crystals that have formed on the surface of the water. In the background, there is a range of mountains with snow-capped peaks. The sky is clear and blue, with a few fluffy clouds floating in the distance. The overall mood of the image is peaceful and serene, with the cactus standing tall and proud in the center of the frozen lake.



Can this happen in real world?

No,





北京航空航天大學  
BEIHANG UNIVERSITY

人工智能研究院  
Institute of Artificial Intelligence

# Intro to Other Multimodal Large-Language Models



# AN EMBODIED GENERALIST AGENT IN 3D WORLD

**Jiangyong Huang<sup>1,2\*</sup>, Silong Yong<sup>1,3\*</sup>, Xiaojian Ma<sup>1\*</sup>, Xiongkun Linghu<sup>1\*</sup>, Puhaao Li<sup>1,4</sup>,  
Yan Wang<sup>1</sup>, Qing Li<sup>1</sup>, Song-Chun Zhu<sup>1,2,4</sup>, Baoxiong Jia<sup>1</sup>, Siyuan Huang<sup>1</sup>**

<sup>1</sup>Beijing Institute for General Artificial Intelligence (BIGAI)

<sup>2</sup>Peking University <sup>3</sup>Carnegie Mellon University <sup>4</sup>Tsinghua University

<https://embodied-generalist.github.io>

(ICLR2024 Under Review)

首个3D 具身Agents



## Motivation



**Describe object.**

There is a rectangular brown table. It is in the center of the room.

LEO



**Where is the bathtub located in the bathroom?**

To right of toilet

LEO



**Is the door closed or open to my left side?**

Closed

LEO

- **任务：构建具备在3D物理世界中感知、推理、规划、行动的Embodied Agents**
- **相关工作：**
- 目前的Embodied Agents框架（基于LLM和VLM进行任务规划、环境感知，并对齐下层的行动控制）取得了巨大进步，但他们的能力都局限于2D领域内，限制了其在3D真实场景下的应用，因此本文解决的核心问题是：**如何让Embodied Agents全面的理解3D环境并具备与真实3D世界交互的能力？**
- **挑战：合适的数据集 / 统一的模型框架 / 设计合理的优化目标**



## Method detail

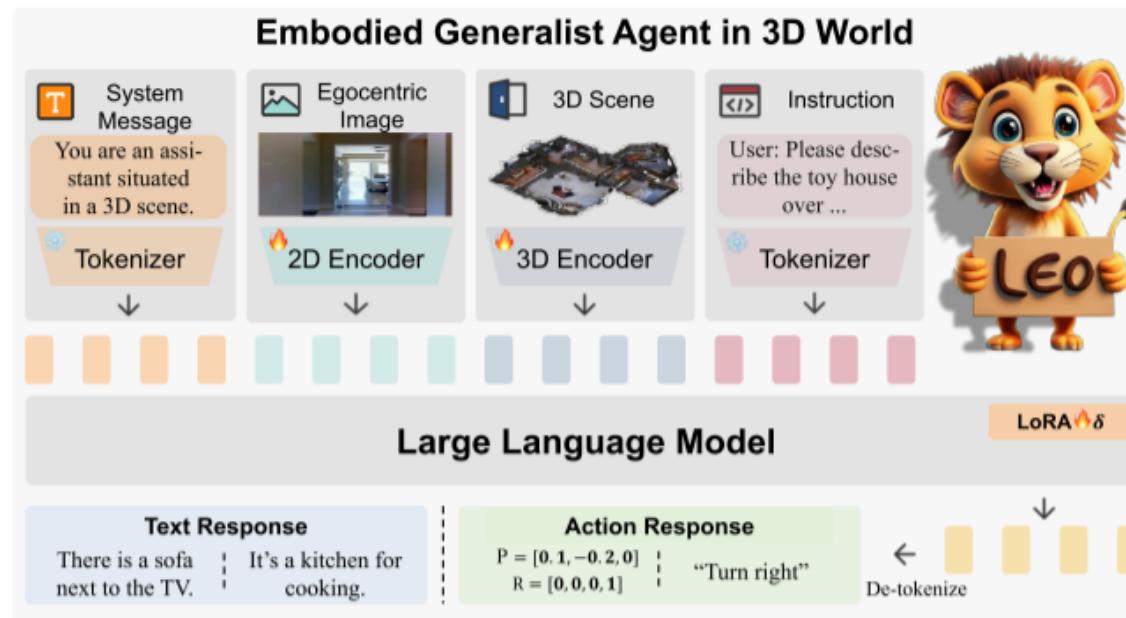
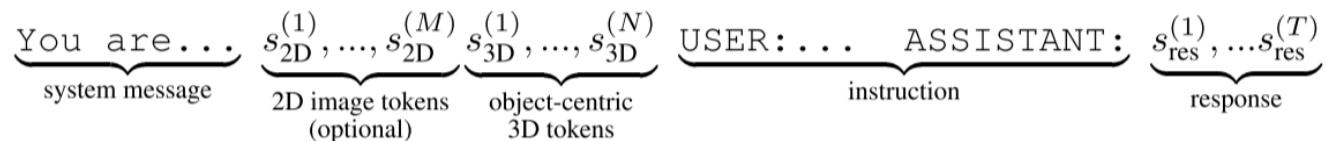
- 挑战：合适的数据集 / 统一的模型框架 / 设计合理的优化目标
- Overview: 将物体中心图像、3D点云场景、文本作为输入，将各种3D任务指定为自回归序列进行学习





## Method detail

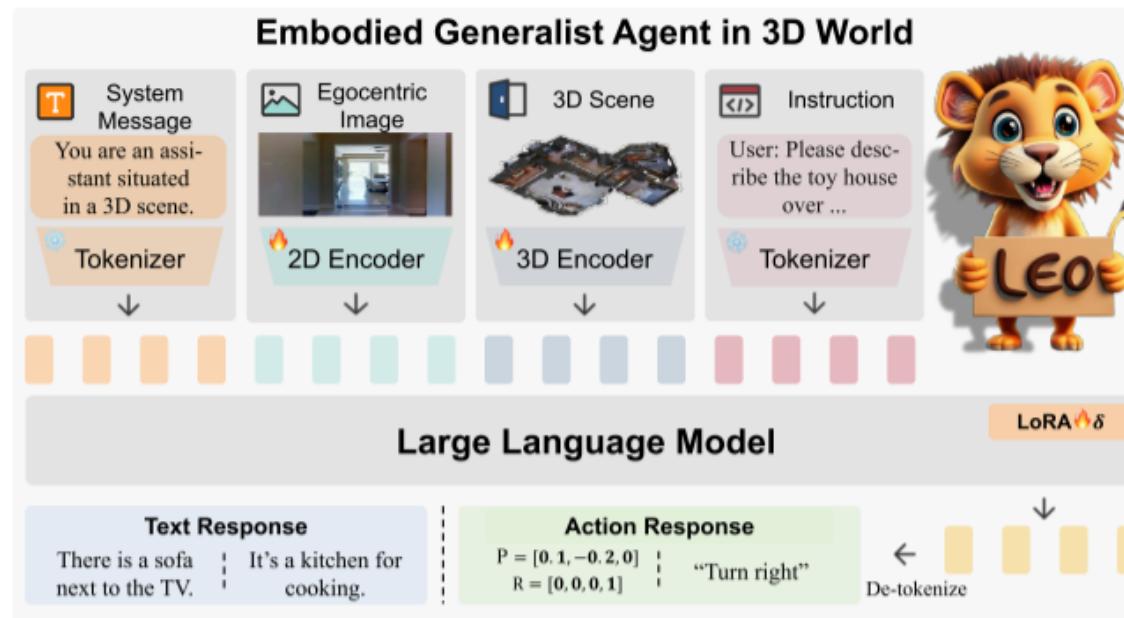
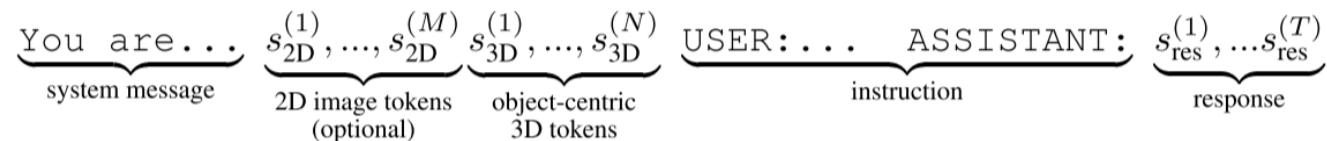
- **Model:** LEO的设计原则有两个方面，1. 能够处理自我为中心的2D信息、全局3D信息、以及文本指令，并具有文本输出和具身动作指令的统一架构 2. 应该利用预训练LLM作为先验
- 因此，将不同模态的数据转化为tokens序列，使用LLM进行处理，并采用类似GPT的自回归建模





## Method detail

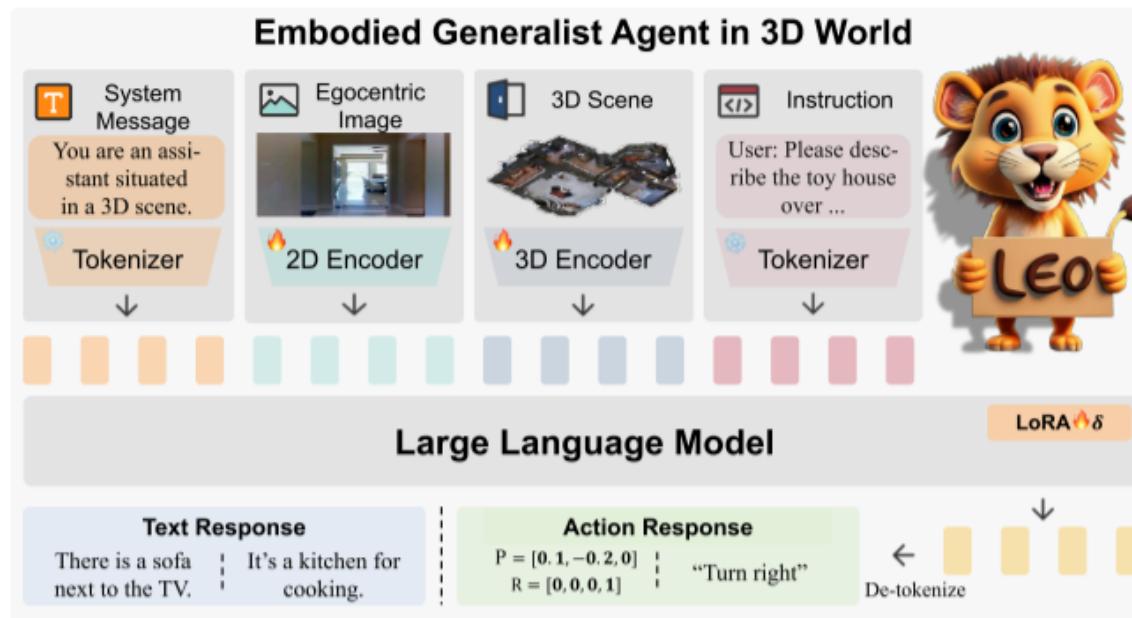
- **Tokenization (编码) :** 使用2D VLM和3D VLM编码多模态的数据，使用SentencePiece tokenizer进行文本编码。对于embodied action 指令，将连续动作离散化并映射到SentencePiece tokenizer中最少被使用的tokens，所有的数据被tokenization后，连接为式子中的形式





## Method detail

- **Training:** 使用Vicuna-7B作为预训练的LLM处理token序列，使用LoRA的方式微调（仅使用少量额外的可学习参数，并引入冻结的LLM）
- **LoRA:** 低秩适应(low-rank adaptation), (Aghajanyan et. al.) 预训练大模型通常具有极小的内在维度，即存在一个极低维度的参数，微调它和在全参数空间中微调能起到相同的效果。这也是大模型具备few-shot能力的原因之一
- **训练目标:** 前缀语言建模，即给定前缀，预测后缀（回答或动作指令的输出）的token

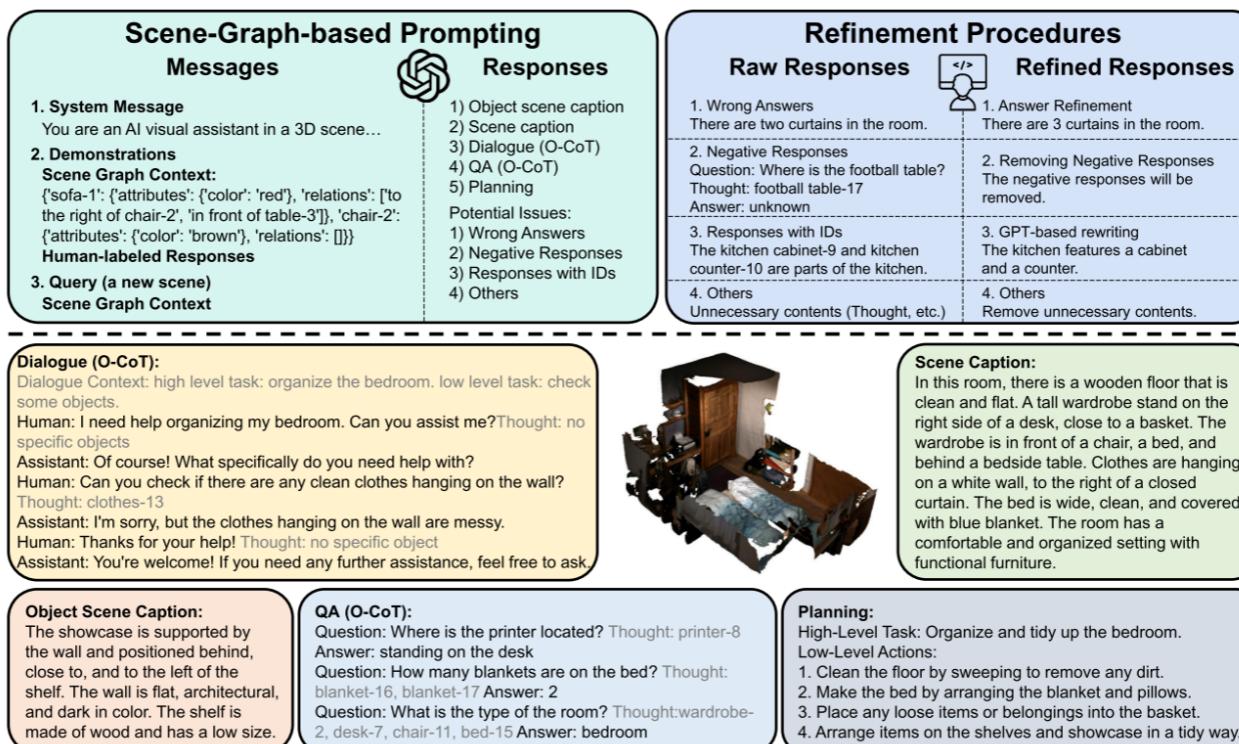


$$\mathcal{L}(\theta, \mathcal{B}) = - \sum_{b=1}^{|\mathcal{B}|} \sum_{t=1}^T \log p_\theta(s_{\text{res}}^{(b,t)} | s_{\text{res}}^{(b,<t)}, s_{\text{prefix}}^{(b,1)}, \dots, s_{\text{prefix}}^{(b,L)}),$$



## Method detail

- Datasets:** 使用LLM辅助生成数据，训练数据分为两组（LEO-align和LEO-instruct），分别用于LEO的两阶段训练，即3D视觉-语言对齐和3D世界任务的Instruction Tuning
- 辅助数据生成的过程，首先生成带有3D场景的信息，输入包括系统信息，场景上下文关系，输出包括caption/对话/QA/规划等形式，接着进行人工纠错和修改提升数据质量





## Experiment

### 在3D理解和具身推理、机器人操作、物体跟踪任务取得SoTA

Table 2: Quantitative comparison with state-of-the-art models on 3D VL understanding and embodied reasoning tasks. “C” stands for “CIDEr”, “B-4” for “BLEU-4”, “M” for “METEOR”, “R” for “ROUGE”, “Sim” for sentence similarity, and “EM@1” for top-1 exact match. The n-gram metrics for Scan2Cap are governed by IoU@0.5. <sup>†</sup> indicates answering questions via prompting GPT-3 with the generated scene caption.

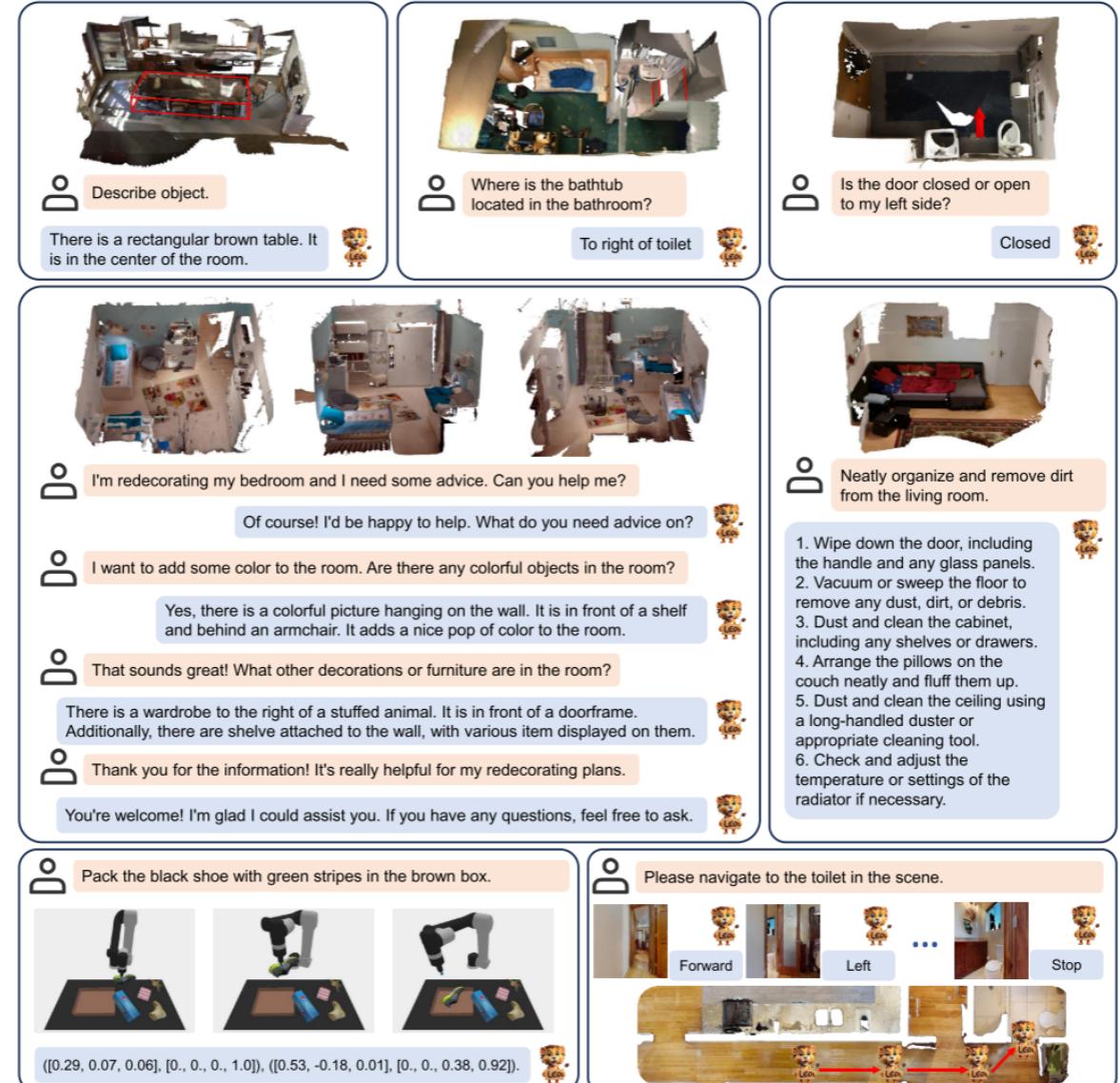
	Scan2Cap (val)					ScanQA (val)					SQA3D
	C	B-4	M	R	Sim	C	B-4	M	R	EM@1	EM@1
<i>Task-specific models</i>											
Scan2Cap (GPT-3) (Chen et al., 2021)	35.2	22.4	21.4	43.5	-	-	-	-	-	-	41.0 <sup>†</sup>
3DJCG (Cai et al., 2022)	47.7	31.5	24.3	51.8	-	-	-	-	-	-	-
Vote2Cap-DETR (Chen et al., 2023)	61.8	34.5	26.2	54.4	-	-	-	-	-	-	-
ScanRefer+MCAN (Chen et al., 2020)	-	-	-	-	-	55.4	7.9	11.5	30.0	18.6	-
ClipBERT (Lei et al., 2021)	-	-	-	-	-	-	-	-	-	-	43.3
ScanQA (Azuma et al., 2022)	-	-	-	-	-	64.9	10.1	13.1	33.3	21.1	47.2
<i>Task-specific fine-tuned</i>											
3D-VisTA (Zhu et al., 2023c)	66.9	34.0	27.1	54.3	53.8	69.6	10.4	13.9	35.7	22.4	48.5
3D-LLM (FlanT5) (Hong et al., 2023)	-	-	-	-	-	69.4	<b>12.0</b>	14.5	35.7	20.5	-
LEO	<b>68.4</b>	<b>36.9</b>	<b>27.7</b>	<b>57.8</b>	<b>54.7</b>	<b>80.0</b>	11.5	<b>16.2</b>	<b>39.3</b>	<b>36.6</b>	<b>53.7</b>

Table 3: Results on CLIPort robot manipulation. We compare with results from Shridhar et al. (2021). *seen* indicates in-domain tasks. *unseen* marks OOD tasks with novel colors or objects.

	separating-piles		packing-google-objects-seq		put-blocks-in-bowls	
	seen	unseen	seen	unseen	seen	unseen
Transporter	48.4	52.3	46.3	37.3	64.7	18.7
CLIP-only	90.2	71.0	95.8	57.8	97.7	44.5
RN50-BERT	46.5	44.9	94.0	56.1	91.8	23.8
CLIPort (single)	98.0	<b>75.2</b>	<b>96.2</b>	71.9	<b>100</b>	25.0
CLIPort (multi)	89.0	62.8	84.4	70.3	<b>100</b>	<b>45.8</b>
LEO	<b>98.8</b>	<b>75.2</b>	76.6	<b>79.8</b>	86.2	35.2

Table 4: Results on object navigation. We compare LEO with similar imitation learning agents in Habitat-web (H.w.) (Ramrakhy et al., 2022) and CortexBench (VC-1) (Majumdar et al., 2023). S: success rate; L: SPL.  
<sup>†</sup> LEO is not trained on HM3D scenes.

	MP3D-val		HM3D-val	
	S(↑)	L(↑)	S(↑)	L(↑)
H.w. (shortest)	4.4	2.2	-	-
H.w. (70k demo)	<b>35.4</b>	10.2	-	-
VC-1 (ViT-B)	-	-	<b>57.1</b>	<b>31.4</b>
LEO	23.1	<b>15.2</b>	23.1 <sup>†</sup>	19.1 <sup>†</sup>





北京航空航天大學  
BEIHANG UNIVERSITY

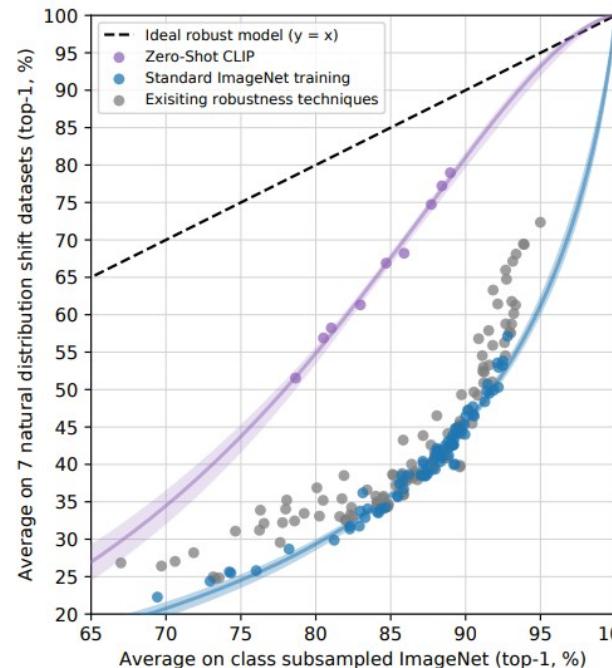
人工智能研究院  
Institute of Artificial Intelligence

# Intro to VLLMs Safety



## 多模态大模型的显著优势：应对不同数据分布的强大泛化性

基于图文对比学习预训练的CLIP 在zero-shot分类中能够准确预测各种不同的OOD样本



相比于传统task-specific模型  
(ResNet101等)  
CLIP 在各类OOD基准数据集  
上的表现具有突破性的提升



## 多模态大模型的显著优势: 应对不同数据分布的强大泛化性

### 基于指令微调的视觉语言大模型LLaVa 能够准确理解图像中的unusual和复杂对象

Visual input example, Extreme Ironing:



User  
LLaVA

Source: <https://www.barnorama.com/wp-content/uploads/2016/12/03-Confusing-Pictures.jpg>

What is unusual about this image?

The unusual aspect of this image is a man ironing clothes on the back of a minivan or van. This is not a typical place to perform this activity, as one would usually iron clothes in a more stationary and safe location, such as a home, using a regular ironing board. The scene depicted in the image is peculiar as it involves a makeshift ironing setup on a vehicle, which can be both unsafe and unconventional. Additionally, it is not clear how the man is able to maintain balance and stability while ironing clothes in such an unstable environment.

[Start a new conversation, and clear the history]

What's happening in the scene?

The image depicts a man standing on top of a yellow SUV in a busy city street. He is holding a portable ladder, seemingly ironing clothes while standing on the vehicle. Around the scene, there are other cars, a traffic light, and a fire hydrant. Additionally, there are two other people in the vicinity, one close to the man on the car and another further away. The man on the ladder appears to be attracting attention due to his unconventional choice of ironing his clothes on top of a moving car. The city street around him is bustling with activity, adding to the unique nature of the scene.

User  
GPT-4 [36]

What is unusual about this image?

The unusual thing about this image is that a man is ironing clothes on an ironing board attached to the roof of a moving taxi.

User  
BLIP-2

What is unusual about this image?

a man is sitting on the back of a yellow cab

User  
OpenFlamingo

What is unusual about this image?

The man is drying his clothes on the hood of his car.

Challenging examples from LLaVA-Bench (In-the-Wild):



ICHIRAN Ramen [[source](#)]



Filled fridge [[source](#)]

Annotation

A close-up photo of a meal at ICHIRAN. The chashu ramen bowl with a spoon is placed in the center. The ramen is seasoned with chili sauce, chopped scallions, and served with two pieces of chashu. Chopsticks are placed to the right of the bowl, still in their paper wrap, not yet opened. The ramen is also served with nori on the left. On top, from left to right, the following sides are served: a bowl of orange spice (possibly garlic sauce), a plate of smoke-flavored stewed pork with chopped scallions, and a cup of matcha green tea.

An open refrigerator filled with a variety of food items. In the left part of the compartment, towards the front, there is a plastic box of strawberries with a small bag of baby carrots on top. Towards the back, there is a stack of sauce containers. In the middle part of the compartment, towards the front, there is a green plastic box, and there is an unidentified plastic bag placed on it. Towards the back, there is a carton of milk. In the right part of the compartment, towards the front, there is a box of blueberries with three yogurts stacked on top. The large bottle of yogurt is Fage non-fat yogurt, and one of the smaller cups is Fage blueberry yogurt. The brand and flavor of the other smaller cup are unknown. Towards the back, there is a container with an unknown content.

Question 1 What's the name of the restaurant?

What is the brand of the blueberry-flavored yogurt?

Question 2 Describe this photo in detail.

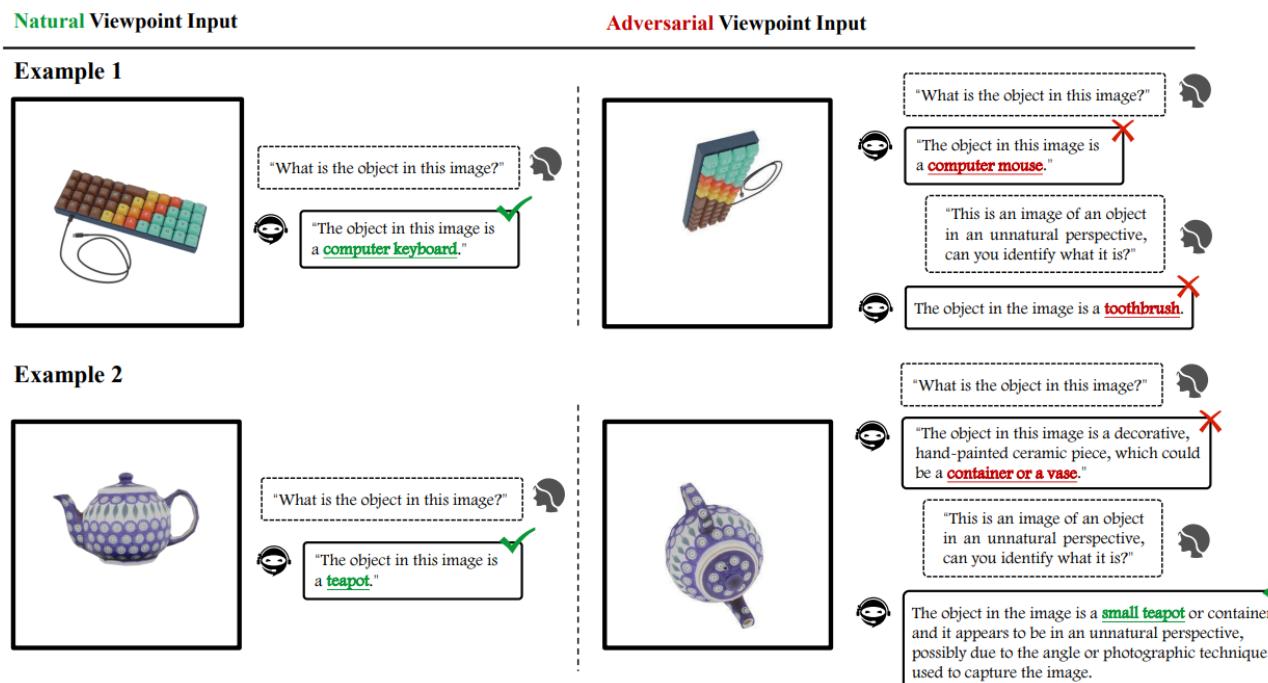
Is there strawberry-flavored yogurt in the fridge?



然而目前的评估大多关注2D-OOD变换 忽略了3D视角变换

我们观察到：在应对视角OOD样本时 视觉语言大模型的性能发生显著下降

- Vision question answering



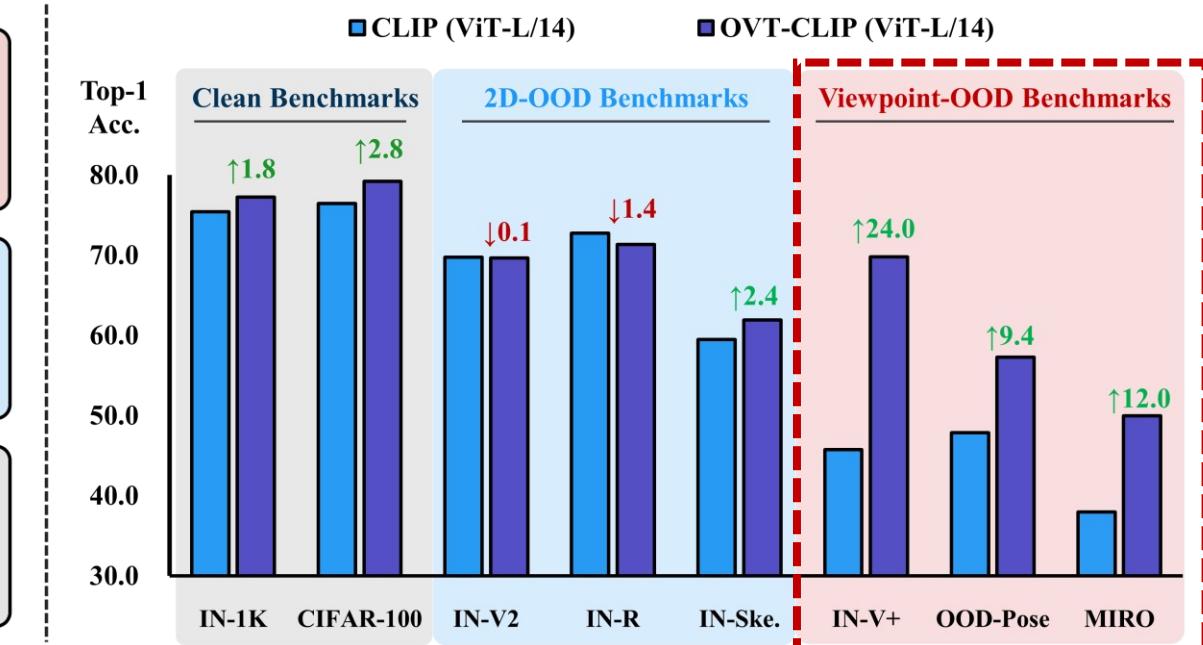
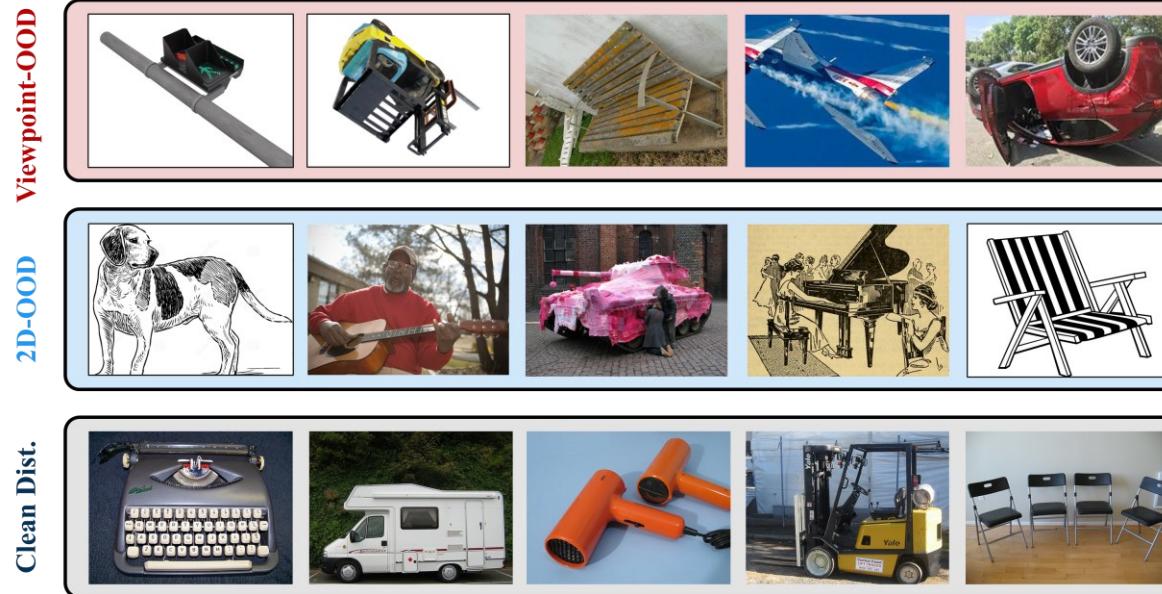
- Zero-shot classification

Models	Vision Encoder	# Toatal Params	ImageNet	IM3D	ImageNet-V+
CLIP [49]	ALFBF [36]	ViT-B/16	210M	66.06% (↓ 13.18%)	52.88% (↓ 39.84%)
	ResNet-50	102M	65.12%	53.36% (↓ 11.76%)	2.51% (↓ 62.61%)
	ViT-B/16	150M	76.94%	66.60% (↓ 10.34%)	37.99% (↓ 38.95%)
	ViT-B/32	151M	72.74%	58.59% (↓ 14.15%)	29.24% (↓ 43.50%)
	ViT-L/14	428M	81.96%	76.16% (↓ 5.80%)	48.49% (↓ 33.47%)
BLIP [35]	ViT-B/16	224M	70.02%	70.73% (↑ 0.71%)	40.08% (↓ 29.94%)
	ViT-L/14	449M	73.86%	76.38% (↑ 2.52%)	50.05% (↓ 23.81%)
BLIP-2 [34]	ViT-G/14	1.2B	77.40%	83.76% (↑ 6.36%)	57.92% (↓ 19.48%)



和常见的2D-OOD变换相比 3D视角变换对于视觉语言模型更具挑战性

针对CLIP 我们在常用的2D-OOD基准和视角OOD基准进行性能的比较：





**应对3D视角变换的鲁棒性 对于AI系统在安全敏感领域的应用 具有重要意义**

如自动驾驶领域 系统需要能够准确的识别来自异常视角的视觉输入 否则将造成安全事故

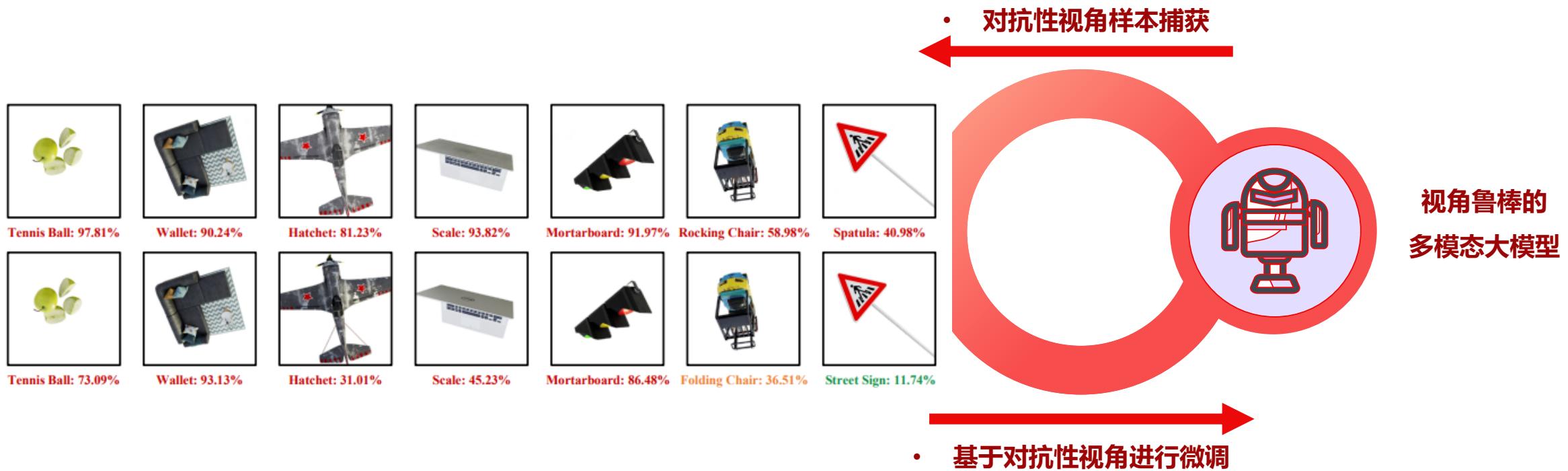


**鉴于当前多模态大模型在自动驾驶等安全敏感领域开始逐渐应用 增强其视角鲁棒性具有重要意义**



## 如何提高多模态大模型对于视角变换的鲁棒性？

1. 从数据增强和对抗性训练的范式考虑：通过某种手段 挖掘大模型的脆弱视角
2. 基于这些脆弱视角进行微调 从而提高模型对不同视角的一致性表征能力



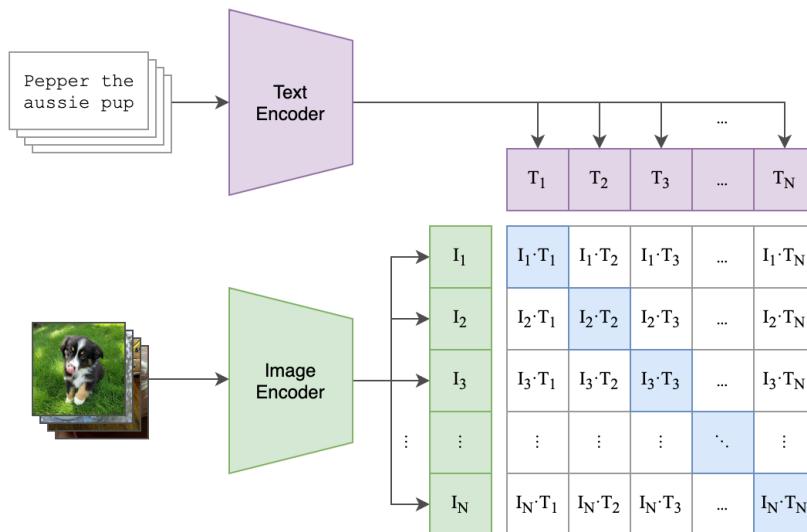


## 当前主流多模态基础模型的训练范式

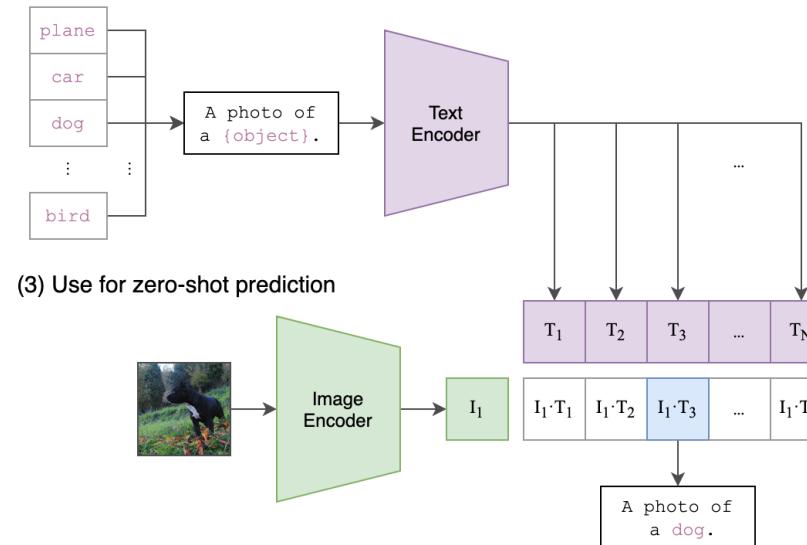
尽管多模态学习的范式众多，但以 **CLIP & ALIGN & BLIP** 等模型为代表的图文对比学习范式占据主流

特点： ➤ 对比学习形式的目标 ➤ 图像/文本 双流编码器架构 ➤ 海量的图文对 作为训练数据集

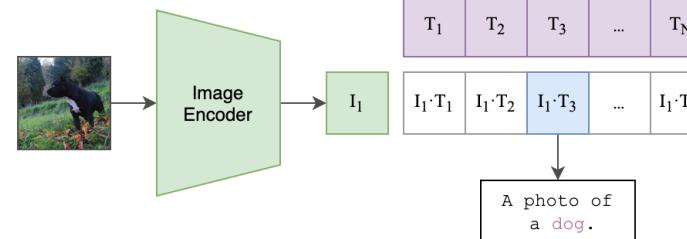
(1) Contrastive pre-training



(2) Create dataset classifier from label text



(3) Use for zero-shot prediction



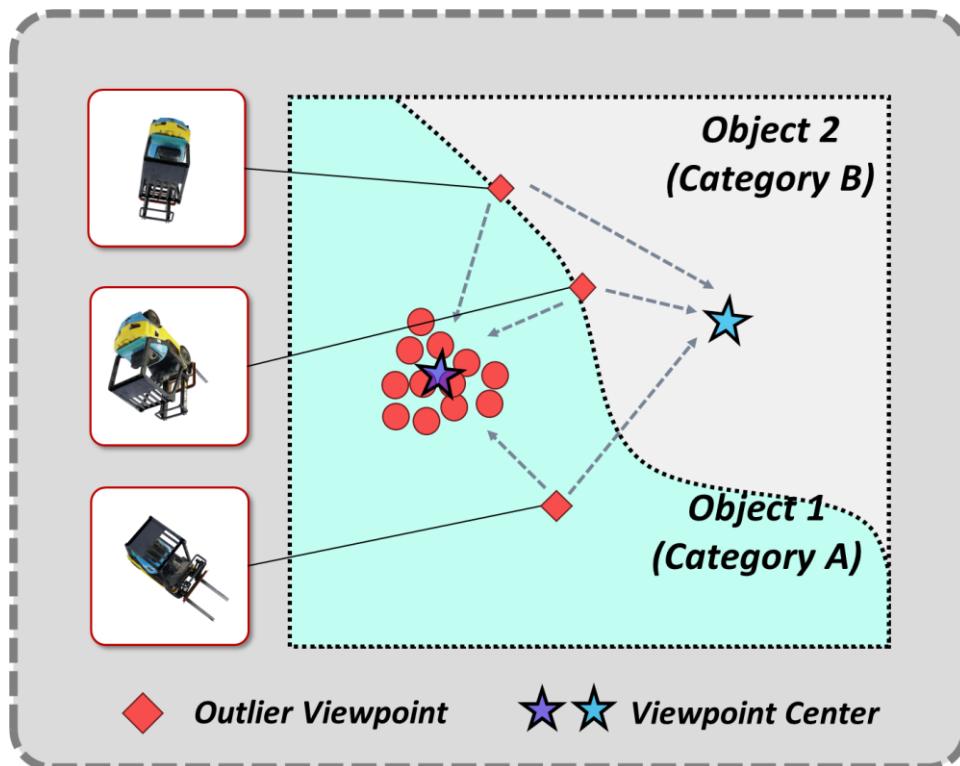
- 因此，有必要考虑对这类模型展开视角鲁棒性的增强



## 工作二：增强多模态模型对脆弱视角的一致性表征

我们的思路是在所挖掘的脆弱视角的基础上

强化多模态模型对于这些视角的表征能力使其能够和正常的视角在表征空间中足够接近



类似于图-文对齐目标

我们针对多模态模型提出了跨视角对齐目标：

在高维嵌入空间中 **最小化** 脆弱视角和中心视角的表征距离

$$\min_{\mathbf{W}_v, \mathbf{W}_t} [\mathcal{L}_{ITC} + \lambda \cdot \underbrace{\sum_i \sum_{j \neq j'} d(z_{ij}^I, z_{ij'}^I)}_{\mathcal{L}_{VC}},$$

$$\min_{\mathbf{W}_v, \mathbf{W}_t} \left[ \mathcal{L}_{ITC} + \lambda \cdot \max_{\mathcal{O} = \{O_i\}_{i=1}^N, |O_i|=K} \sum_{i=1}^N \sum_{j \in \mathcal{O}} l(z_{ij}^I, z_{C_i}^I) \right],$$

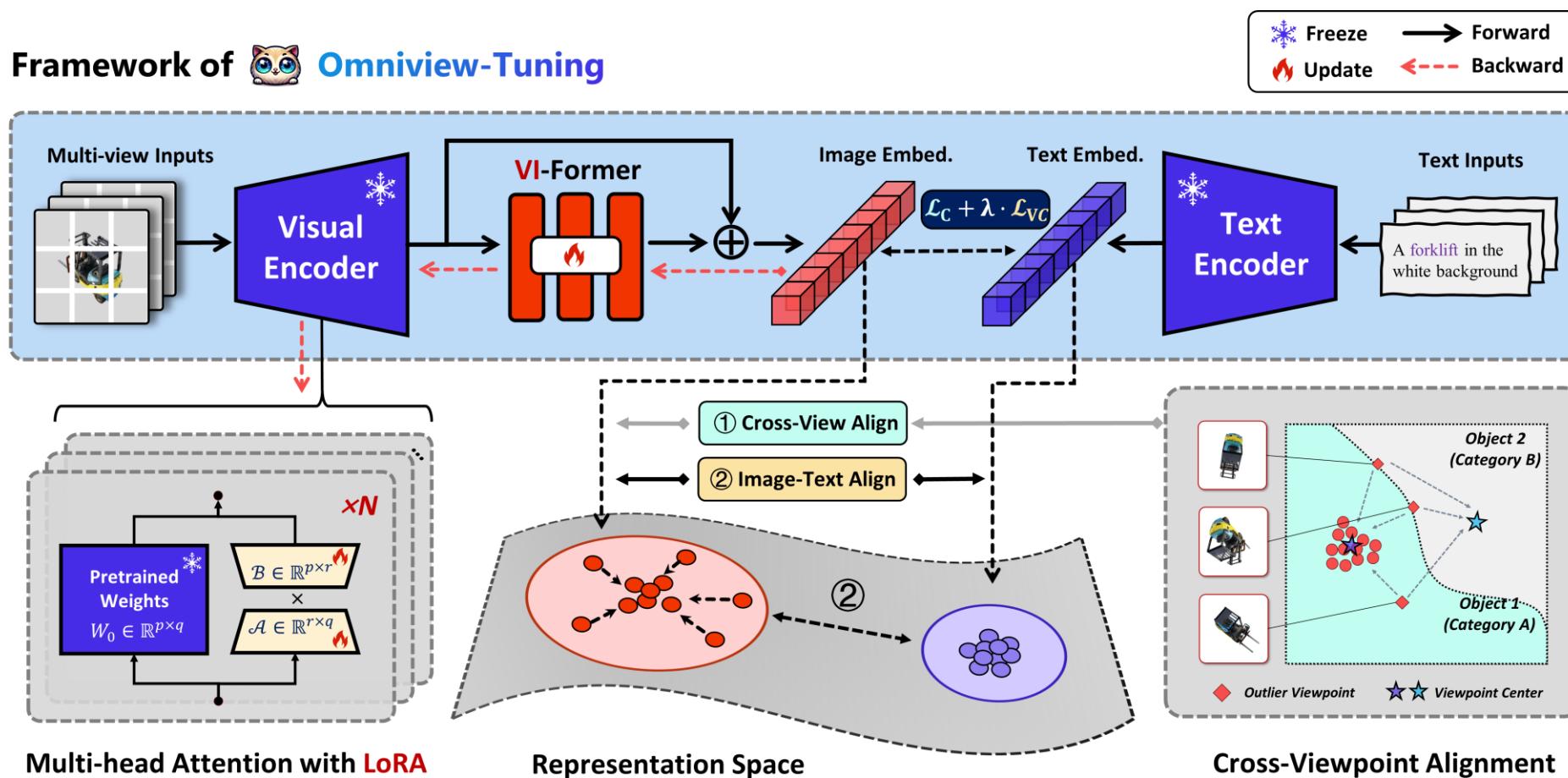
$$\text{where } l(z_{ij}^I, z_{C_i}^I) = \max [d(z_{ij}^I, z_{C_i}^I) + m, 0],$$



## 工作二：增强多模态模型对脆弱视角的一致性表征

我们提出了Omniview-Tuning (OVT) 框架：通过高效参数微调增强多模态模型的视角鲁棒性

Framework of Omnidview-Tuning



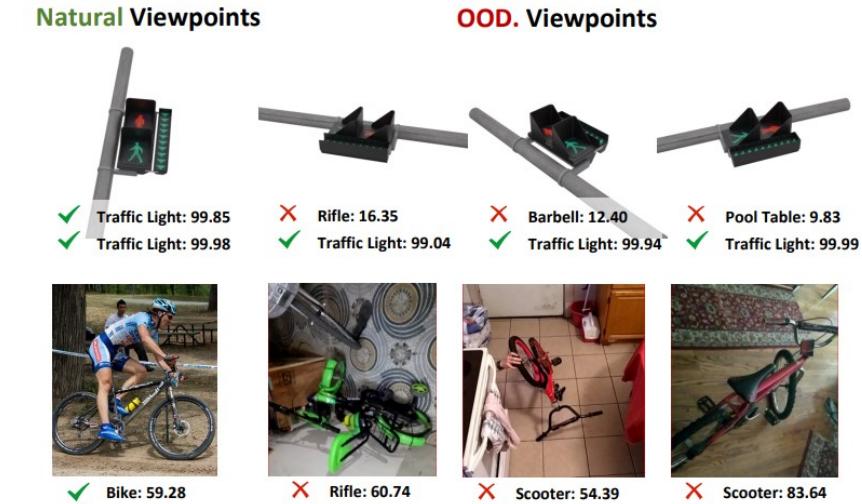


## 工作二：增强多模态模型对脆弱视角的一致性表征

我们的方法能够应用于各种主流的图像-文本多模态架构（如CLIP, BLIP）

在横跨不同数据分布的基准（包含2D-OOD, 视角-OOD, 以及干净数据）上展开zero-shot分类能力的评估

Model	Clean				Common-OOD				Viewpoint-OOD				Total Avg. Acc.				
	ImageNet-100 [47]	ImageNet-1K [13]	Cifar-100 [28]	Avg. Acc.	ImageNet-V2 [45]	ImageNet-Ske. [56]	ImageNet-OOD. [22]	ImageNet-Ren. [20]	OOD-CV [60]	Avg. Acc.	ImageNet-View. [47]	OOD-CV-Pose [60]	MIRO [7]				
A. Comparisons with ViT-B/32 baselines																	
OpenAI CLIP	77.5/93.9	63.3/88.8	64.3/88.1	68.4/90.2	55.8/83.4	42.2/70.3	33.4/62.2	50.7/75.4	50.2/82.6	46.5/74.8	44.5/65.4	27.5/52.4	47.2/84.5	26.5/59.4	36.4/65.4	48.6/75.5	
Open CLIP	<b>81.1</b> /95.3	66.5/89.9	<b>75.8</b> / <b>94.0</b>	<b>74.5</b> / <b>93.0</b>	<b>58.1</b> /83.9	<b>53.6</b> / <b>79.3</b>	34.8/64.4	<b>61.0</b> / <b>81.9</b>	<b>53.5</b> / <b>81.9</b>	<b>52.2</b> / <b>78.3</b>	54.4/72.1	37.1/63.2	46.9/81.6	33.0/69.2	42.8/71.5	54.6/79.7	
OV-T-OpenCLIP	80.9/95.6	67.8/90.8	65.0/89.3	71.2/91.9	58.0/84.2	45.8/73.4	42.8/75.0	50.3/71.4	51.7/79.5	49.7/76.7	<b>61.9</b> / <b>81.2</b>	<b>59.5</b> / <b>85.6</b>	<b>52.8</b> / <b>82.5</b>	<b>35.4</b> / <b>80.1</b>	<b>52.4</b> / <b>82.4</b>	<b>56.0</b> / <b>82.4</b>	
MetaCLIP	80.7/95.6	67.6/90.5	<b>77.7</b> / <b>95.2</b>	<b>75.3</b> / <b>93.8</b>	59.5/85.4	<b>55.9</b> / <b>81.4</b>	32.4/62.5	<b>63.2</b> / <b>83.8</b>	<b>52.0</b> / <b>84.2</b>	<b>52.6</b> / <b>79.5</b>	61.4/76.7	41.0/67.8	48.9/ <b>87.9</b>	34.8/73.2	46.5/76.4	56.3/82.0	
OV-T-MetaCLIP	<b>80.7</b> / <b>95.6</b>	<b>69.7</b> / <b>92.0</b>	71.8/93.0	74.0/93.5	<b>60.6</b> / <b>85.8</b>	47.8/75.8	<b>43.5</b> / <b>73.8</b>	49.0/70.8	50.1/80.1	50.2/77.2	<b>64.0</b> / <b>79.2</b>	<b>54.8</b> / <b>80.4</b>	<b>55.1</b> / <b>84.8</b>	<b>35.6</b> / <b>77.0</b>	<b>52.4</b> / <b>80.3</b>	<b>56.9</b> / <b>82.3</b>	
B. Comparisons with ViT-B/16 baselines																	
OpenAI CLIP	82.1/95.7	68.3/91.9	67.2/89.4	72.5/92.3	61.8/87.4	48.2/76.3	27.7/55.7	59.1/83.0	52.2/ <b>84.6</b>	49.8/77.4	51.6/68.9	36.9/63.8	53.4/86.8	30.1/66.1	43.0/71.4	53.2/79.1	
Open CLIP	83.2/96.2	70.1/91.8	<b>77.0</b> / <b>94.8</b>	<b>76.8</b> / <b>94.3</b>	62.2/87.0	<b>56.0</b> / <b>82.0</b>	30.7/59.8	<b>64.9</b> / <b>85.6</b>	54.3/82.7	<b>53.6</b> / <b>79.4</b>	58.1/74.4	44.2/70.9	48.5/84.0	34.6/74.6	46.4/76.0	57.0/82.0	
OV-T-OpenCLIP	<b>83.9</b> / <b>97.0</b>	<b>71.9</b> / <b>93.1</b>	69.0/90.7	74.9/93.6	<b>64.0</b> / <b>88.6</b>	50.5/77.9	<b>36.8</b> / <b>68.9</b>	57.0/77.2	<b>56.3</b> / <b>84.5</b>	52.9/ <b>79.4</b>	<b>65.4</b> / <b>80.7</b>	<b>61.7</b> / <b>85.8</b>	<b>56.9</b> / <b>87.4</b>	<b>42.4</b> / <b>84.9</b>	<b>56.6</b> / <b>84.7</b>	<b>59.6</b> / <b>84.7</b>	
EVA-CLIP	<b>85.3</b> /96.5	<b>74.6</b> / <b>94.2</b>	87.5/95.8	<b>98.0</b> / <b>82.5</b>	<b>96.3</b>	<b>67.0</b> / <b>89.8</b>	57.6/82.3	21.3/47.3	69.6/87.5	<b>53.1</b> / <b>83.1</b>	53.7/78.0	61.8/76.6	44.3/69.4	53.9/87.4	32.9/73.2	48.2/76.6	59.1/82.1
MetaCLIP	84.3/97.2	72.1/93.4	<b>78.9</b> / <b>95.4</b>	78.4/95.3	65.0/89.3	<b>60.1</b> / <b>84.8</b>	26.2/56.4	<b>70.2</b> / <b>89.3</b>	52.3/85.4	<b>54.8</b> / <b>81.0</b>	64.2/79.4	49.6/76.1	48.9/ <b>90.9</b>	38.5/78.7	50.3/81.2	59.2/84.7	
OV-T-MetaCLIP	83.4/ <b>97.4</b>	73.8/94.1	73.9/93.6	77.0/95.0	65.9/89.4	53.6/81.0	<b>36.2</b> / <b>66.8</b>	59.0/79.6	51.6/ <b>83.8</b>	53.2/80.1	<b>69.7</b> / <b>84.0</b>	<b>64.8</b> / <b>87.3</b>	<b>55.2</b> / <b>87.8</b>	<b>39.2</b> / <b>82.9</b>	<b>57.2</b> / <b>85.5</b>	<b>60.5</b> / <b>85.6</b>	
C. Comparisons with ViT-L/14 baselines																	
OpenAI CLIP	86.5/97.4	75.4/94.6	76.5/93.3	79.5/95.1	<b>69.8</b> /90.9	59.5/84.3	18.6/43.8	72.8/91.4	52.9/88.8	54.7/79.8	60.3/75.6	45.8/71.5	47.9/88.2	38.0/74.1	48.0/77.3	58.6/82.8	
Open CLIP	86.8/97.8	75.2/94.3	<b>83.7</b> / <b>96.7</b>	<b>81.9</b> / <b>96.2</b>	67.7/90.2	<b>63.2</b> / <b>86.4</b>	24.0/50.5	<b>74.5</b> / <b>91.2</b>	54.5/85.0	56.8/80.6	65.7/78.1	53.2/76.7	52.4/90.5	42.3/83.0	53.4/82.1	61.9/85.0	
OV-T-OpenCLIP	<b>89.0</b> / <b>97.8</b>	<b>77.3</b> / <b>95.3</b>	79.2/95.3	81.8/96.1	69.6/ <b>91.5</b>	61.9/86.0	<b>27.5</b> / <b>55.4</b>	71.3/88.7	<b>56.4</b> / <b>87.0</b>	<b>57.3</b> / <b>81.7</b>	<b>72.2</b> / <b>86.6</b>	<b>69.8</b> / <b>89.7</b>	<b>57.3</b> / <b>94.1</b>	<b>50.0</b> / <b>89.3</b>	<b>62.3</b> / <b>89.9</b>	<b>65.1</b> / <b>88.1</b>	
EVA-CLIP	88.5/97.9	<b>79.6</b> / <b>96.0</b>	<b>90.6</b>	<b>98.6</b> / <b>86.3</b>	<b>97.5</b>	<b>72.8</b> / <b>92.7</b>	68.0/89.1	16.3/40.0	<b>82.8</b> / <b>95.7</b>	54.7/87.4	58.9/81.0	71.5/82.3	61.1/81.7	54.4/ <b>94.5</b>	39.6/86.1	56.6/86.1	65.0/86.8
MetaCLIP	88.3/ <b>98.3</b>	79.1/95.9	84.1/96.9	83.8/97.0	72.5/92.6	<b>68.9</b> / <b>89.8</b>	17.0/40.6	81.8/95.1	<b>56.6</b> / <b>87.5</b>	<b>59.3</b> / <b>81.1</b>	77.3/89.3	66.4/87.0	<b>58.9</b> / <b>93.3</b>	<b>48.1</b> / <b>89.6</b>	62.7/89.8	<b>66.6</b> / <b>88.0</b>	
OV-T-MetaCLIP	<b>88.8</b> / <b>97.5</b>	77.7/95.9	84.0/96.9	83.5/96.8	70.8/92.2	64.4/87.9	<b>20.8</b> / <b>47.0</b>	77.0/92.7	56.3/ <b>89.3</b>	57.8/ <b>81.8</b>	<b>79.3</b> / <b>90.6</b>	<b>75.4</b> / <b>93.0</b>	57.0/94.4	46.4/ <b>93.8</b>	<b>64.5</b> / <b>92.9</b>	66.5/ <b>89.3</b>	
D. Comparisons with BLIP VIT-B/16 baselines																	
BLIP	76.6/93.3	52.9/80.2	<b>67.0</b> / <b>88.3</b>	65.5/87.3	47.3/74.7	<b>51.0</b> / <b>76.6</b>	25.6/53.4	<b>64.3</b> / <b>83.8</b>	53.9/ <b>87.6</b>	48.4/75.2	55.2/68.2	36.8/63.3	50.8/ <b>89.9</b>	27.0/66.1	42.4/71.9	50.7/77.1	
OV-T-BLIP	<b>82.2</b> / <b>97.0</b>	61.7/ <b>88.8</b>	66.6/ <b>88.9</b>	70.2/ <b>91.5</b>	53.7/ <b>82.9</b>	46.5/74.2	<b>33.8</b> / <b>62.7</b>	57.4/77.9	<b>56.4</b> / <b>87.3</b>	49.6/ <b>77.0</b>	<b>62.6</b> / <b>79.0</b>	<b>54.8</b> / <b>79.9</b>	<b>55.2</b> / <b>89.5</b>	<b>31.5</b> / <b>73.2</b>	<b>51.0</b> / <b>80.4</b>	<b>55.2</b> / <b>81.8</b>	



结果表明 经过OV-T训练  
模型在视角OOD样本上的表现显著提升  
并维持了在其他数据分布上的泛化性



## 工作二：增强多模态模型对脆弱视角的一致性表征

Model	Visual Encoder	Real-world Domain					Synthetic Domain					
		OOD-CV (iid) [60]			OOD-CV (Pose) [60]		IM3D [47]			ImageNet-V+ [47]		
		$\beta@1.0$	$\beta@0.5$	$\beta@Adp.$	$\beta@1.0$	$\beta@0.5$	$\beta@Adp.$	$\beta@1.0$	$\beta@0.5$	$\beta@Adp.$	$\beta@1.0$	$\beta@0.5$
LLaVa-7b	OpenAI CLIP(ViT-L/14)	44.1	61.1	67.5	46.4	<b>53.6</b>	58.7	46.7	53.3	58.8	20.4	25.5
	<i>TeCoA<sup>4</sup></i> [38](ViT-L/14)	41.9	58.9	65.5	36.1	41.6	49.2	26.3	30.1	42.6	8.7	11.6
	<i>FARE<sup>4</sup></i> [49](ViT-L/14)	42.1	58.9	65.2	40.2	45.9	50.8	35.2	39.2	49.2	12.7	15.8
	OVT-CLIP(ViT-L/14)	43.5	59.5	65.9	<b>46.5</b>	<b>53.6</b>	<b>59.1</b>	49.4	54.0	61.8	<b>26.4</b>	<b>31.9</b>
LLaVa-13b	OpenAI CLIP(ViT-L/14)	45.4	68.0	70.6	<b>48.6</b>	<b>58.6</b>	60.8	48.7	56.7	60.8	21.2	28.4
	<i>TeCoA<sup>4</sup></i> [38](ViT-L/14)	42.4	67.0	72.2	37.4	48.9	51.3	25.0	28.6	41.5	8.4	10.9
	<i>FARE<sup>4</sup></i> [49](ViT-L/14)	43.9	66.7	71.1	41.9	52.1	54.8	36.1	41.4	48.6	12.1	15.9
	OVT-CLIP(ViT-L/14)	45.7	67.3	70.8	48.2	<b>58.6</b>	<b>61.9</b>	50.4	58.9	63.2	<b>26.4</b>	<b>36.2</b>



同时，在应用于视觉语言大模型  
(LLaVa, Openflamingo) 时  
我们的方法也能够有效提升其在  
Image Caption任务上的视角鲁棒性