BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models

王子恒
SID:12310401

SUSTech
CS308 Computer Vision
期中论文汇报

2025 年 4 月

目录

南方科技大学 SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

背景与动机

- 视觉-语言预训练（VLP）近年来迅速发展，在图像字幕生成、VQA 等任务中表现突出
- 当前主流方法多采用端到端训练大型模型，资源消耗极高（如 Flamingo80B 参数达 80B）
- 难以复用已有单模态模型（如 CLIP, LLaMA, T5）
- **核心问题：如何高效桥接视觉和语言模态，降低训练开销？**
- BLIP-2 提出两阶段预训练框架，有效利用冻结的图像编码器和大语言模型（LLM）

南方科技大学 SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

相关工作 I: 先行模型对比

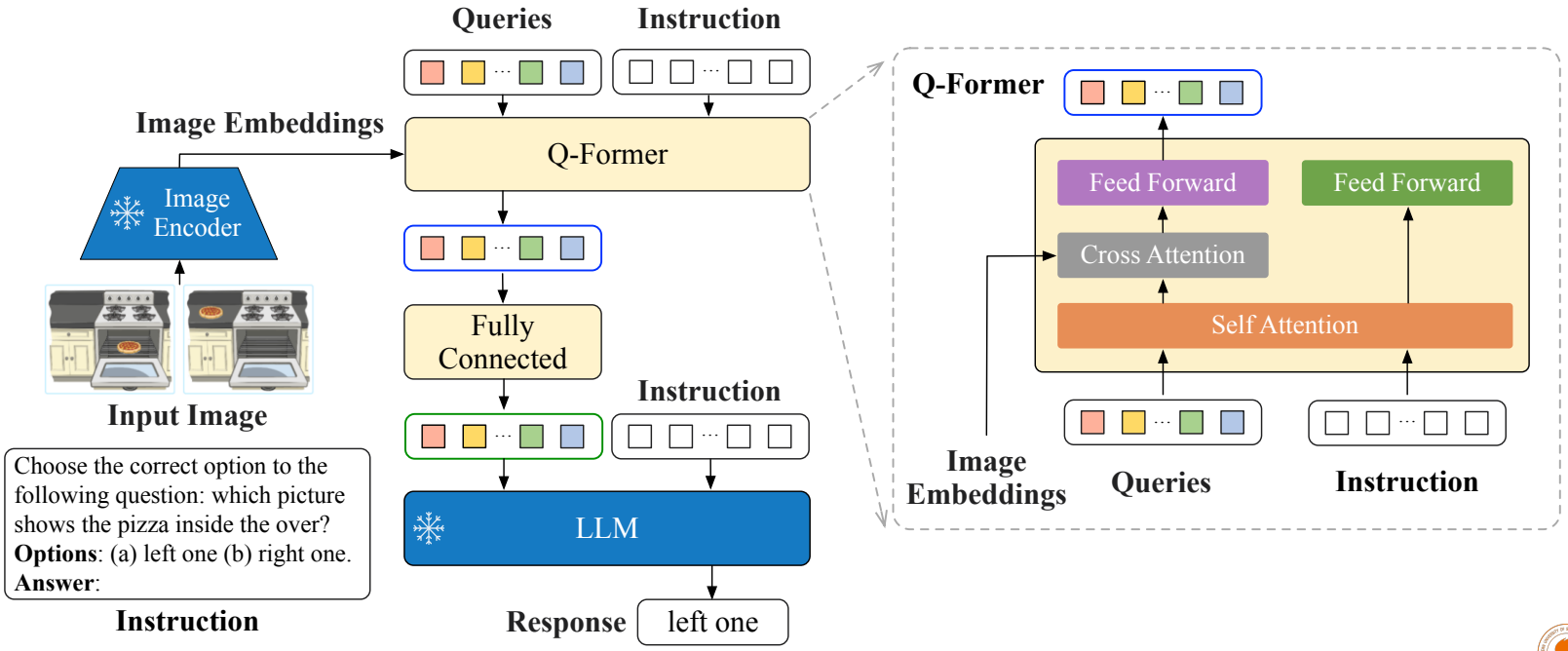
- **CLIP (Radford et al., 2021)**: 双塔结构，通过图文对比学习获取语义对齐特征
- **BLIP (Li et al., 2022)**: 引入多种预训练任务（ITC, ITM, ITG），强化图文建模
- **Frozen, Flamingo (2021-2022)**: 尝试将视觉特征注入冻结的 LLM，实现图文生成
- 存在问题：模态对齐弱、训练成本高、参数量庞大
- BLIP-2 对比：**轻量 Q-Former + 两阶段训练**，兼顾效果与效率

南方科技大学 SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

相关工作 II: 基于 BLIP-2 的扩展研究

InstructBLIP[1]:

- 将任务转换为指令格式训练
- 指令感知 Q-Former 增强泛化性

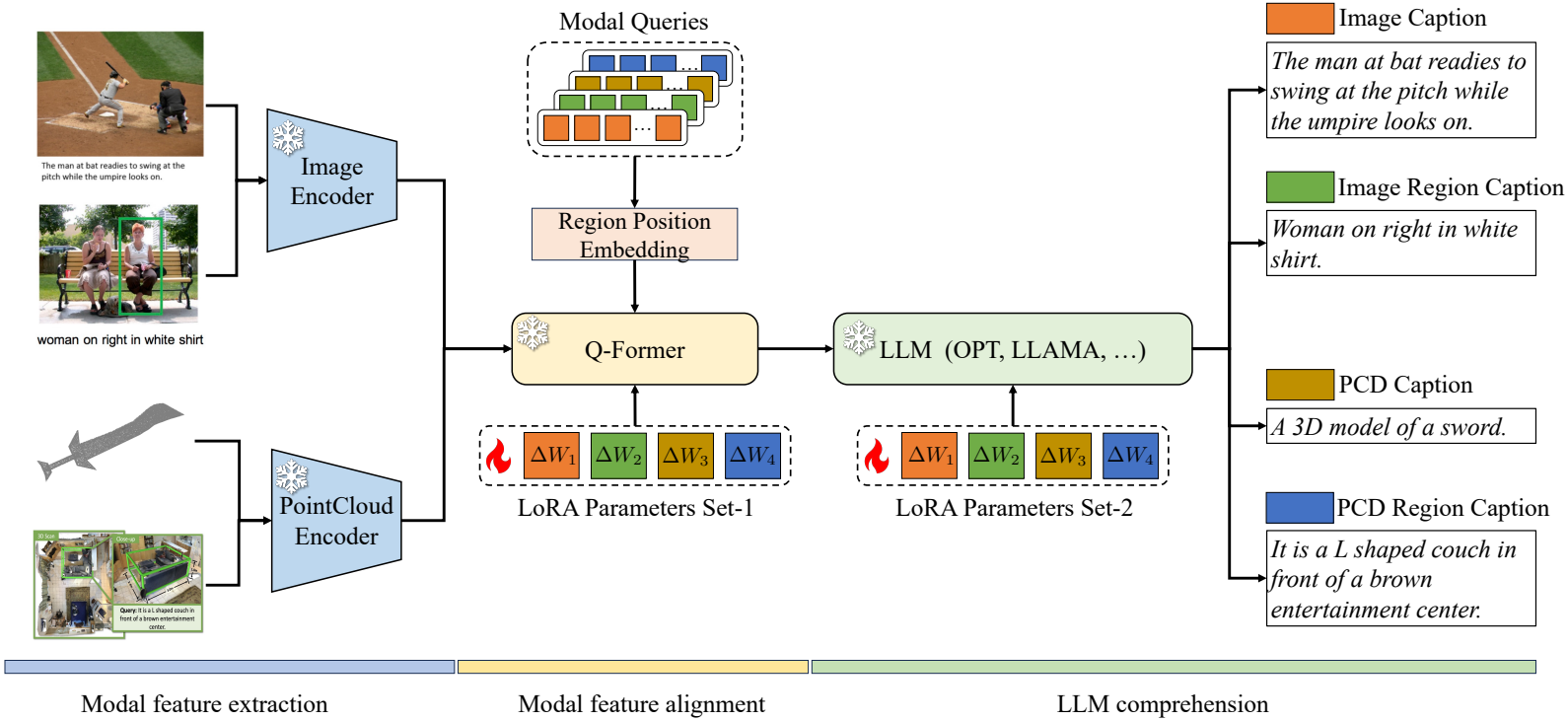


[1] Dai et al., 2023

相关工作 II: 基于 BLIP-2 的扩展研究

RegionBLIP[2]:

- 面向区域目标的多模态扩展框架

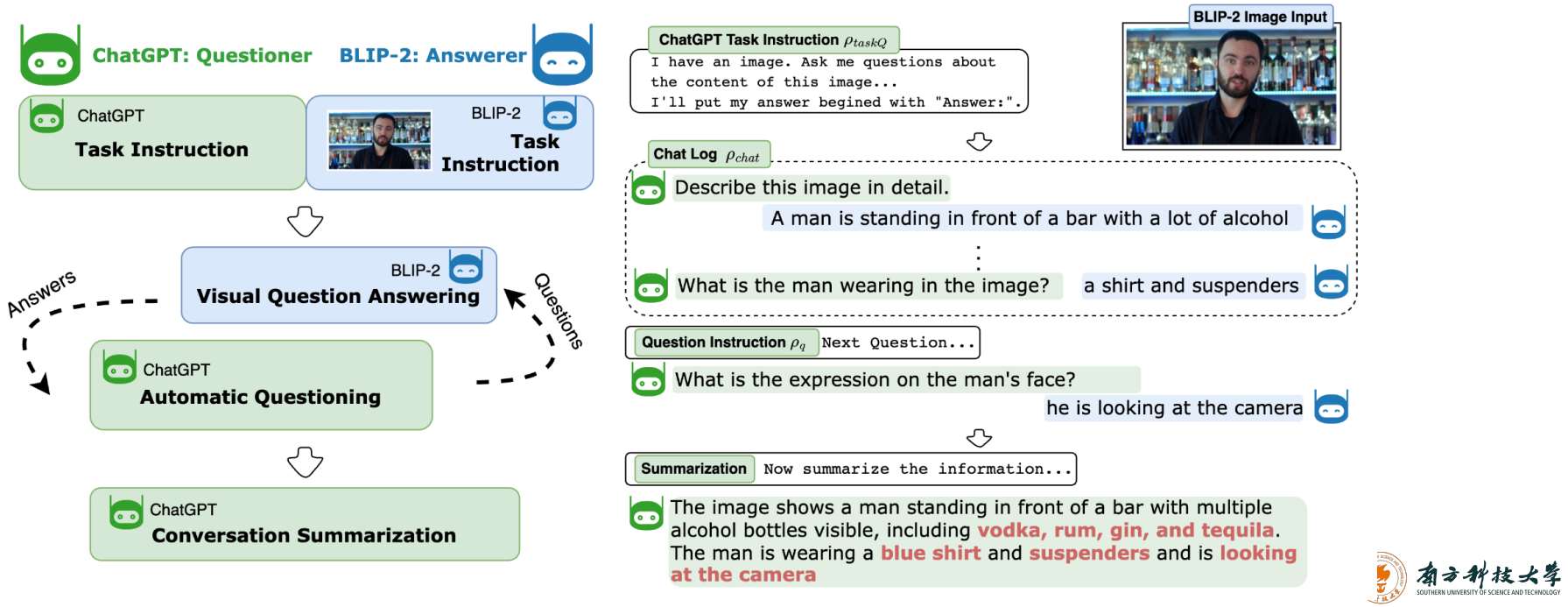


[2]Zhou et al., 2023

相关工作 III: ChatCaptioner 图示

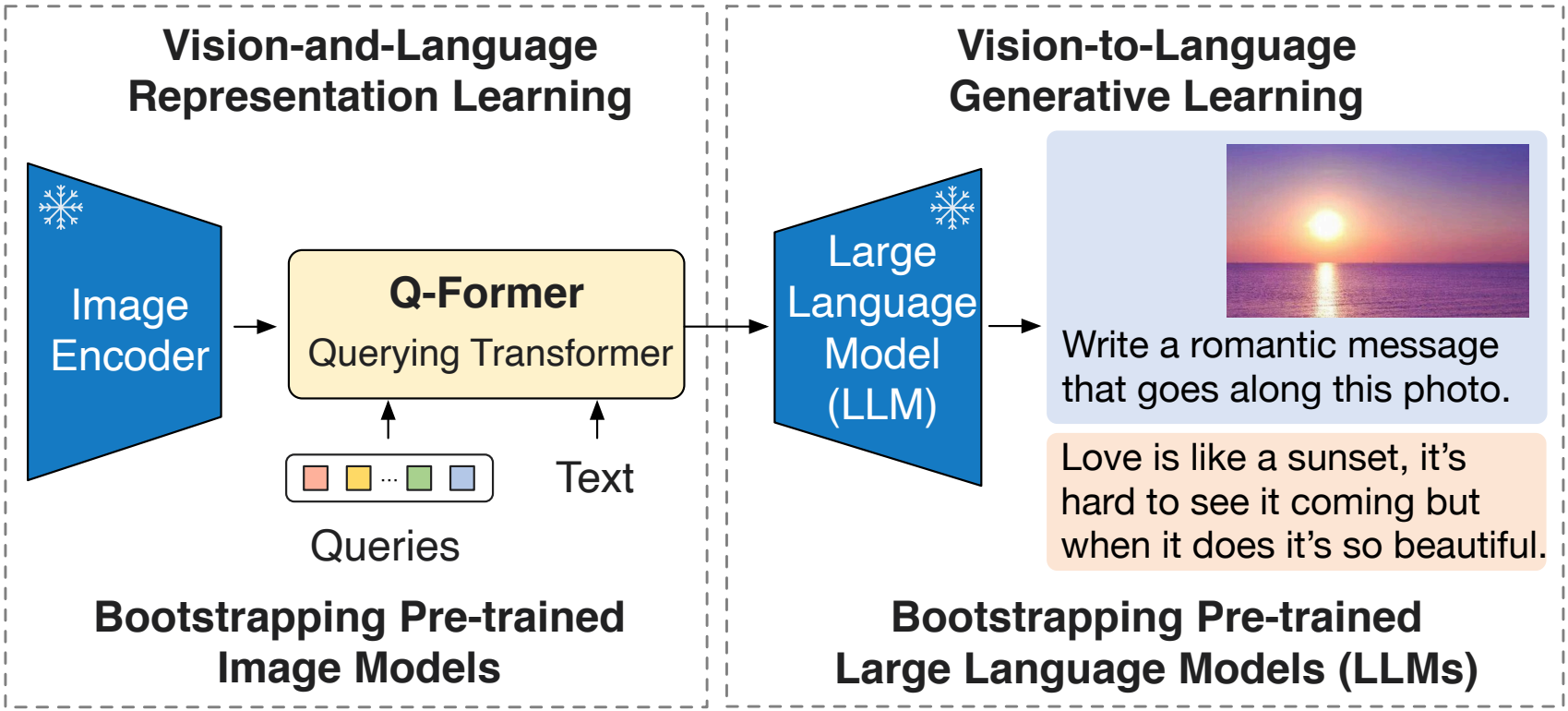
ChatCaptioner[3]:

- 通过 ChatGPT 提问循环，引导 BLIP-2 生成信息丰富的图像描述
- 自动多轮问答提升生成质量



[3] Zhou et al., 2023

BLIP-2 整体架构



BLIP-2 通过 Q-Former 连接冻结的视觉编码器和语言模型

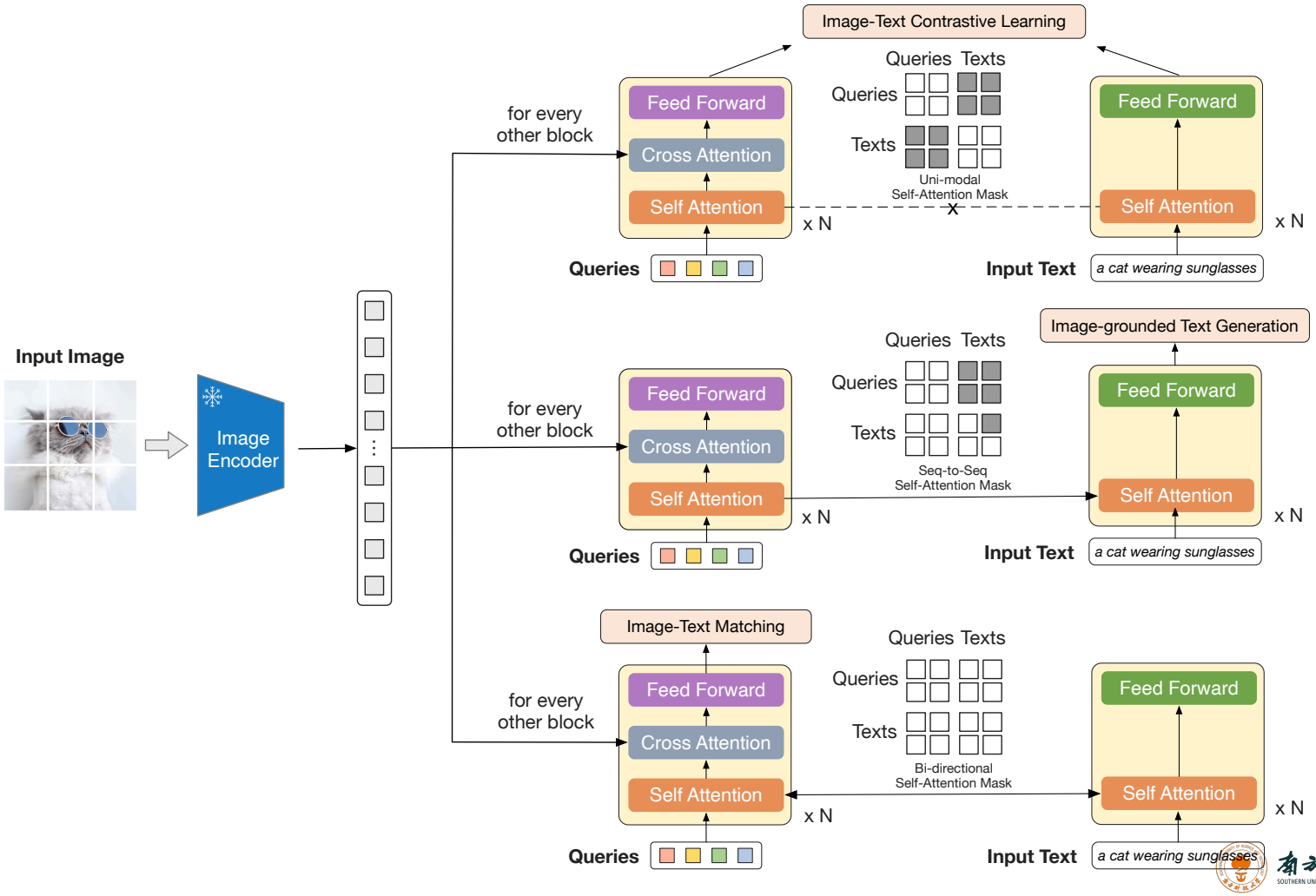
南方科技大学 SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

Q-Former: 核心创新

- **功能**：作为视觉编码器和语言模型之间的桥梁
- **结构**：
  - 包含两组输入：图像特征和查询嵌入 (query embeddings)
  - 基于 Transformer 架构，采用交叉注意力机制
  - 查询嵌入可学习，通过注意力从图像特征中提取关键信息
- **优势**：
  - 减少参数数量，降低计算成本
  - 实现高效信息提取和模态对齐
  - 可灵活连接不同规模的语言模型

南方科技大学 SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

Q-Former 详细结构

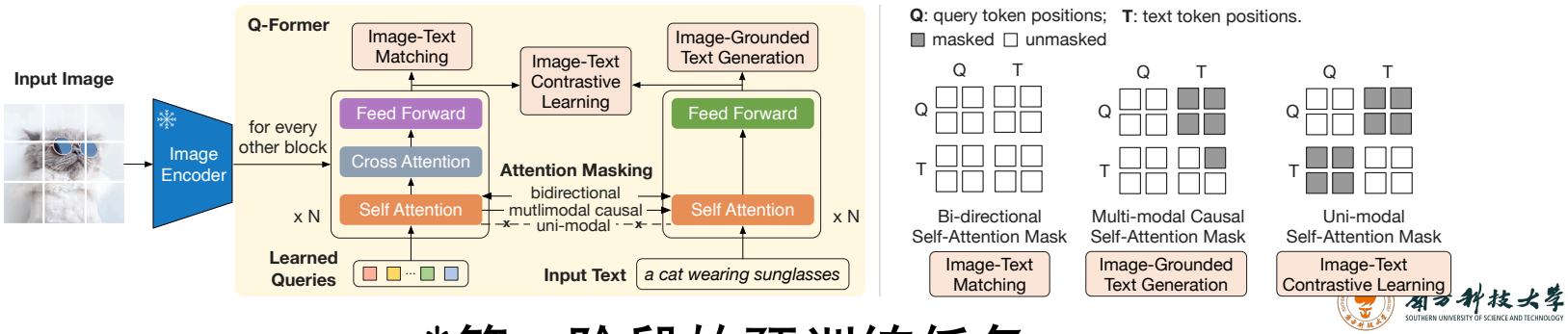


Q-Former 的内部结构与注意力机制

两阶段预训练策略

第一阶段：表征学习

- 冻结 ViT 图像编码器
- 训练 Q-Former 理解视觉信息
- 使用三个目标函数：
  - 图像-文本对比学习
  - 图像-文本匹配
  - 图像条件文本生成

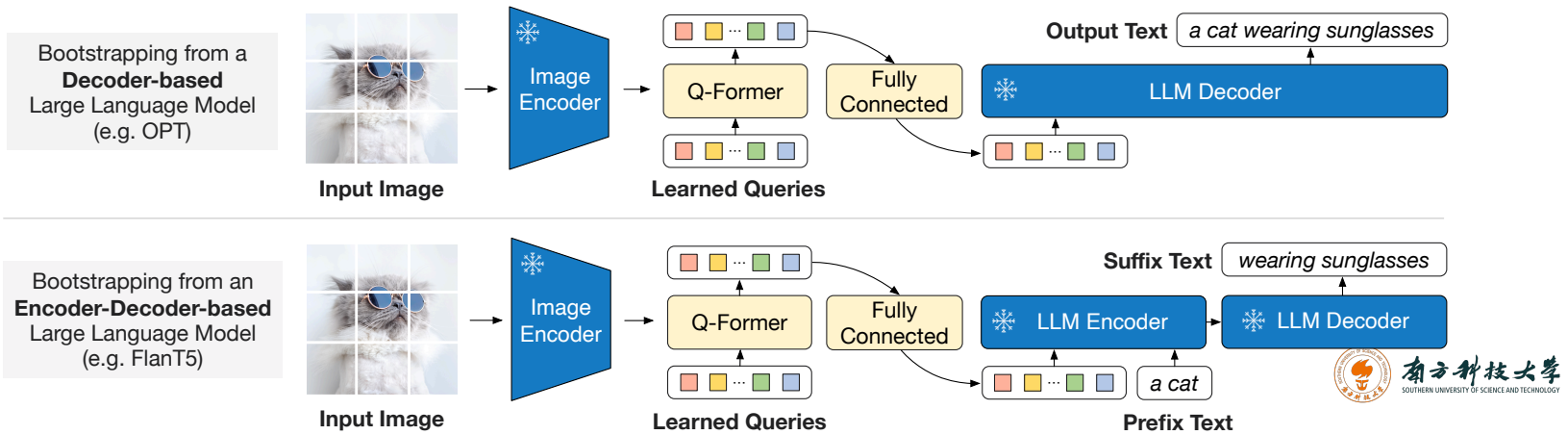


*第一阶段的预训练任务

两阶段预训练策略

第二阶段：指令调优

- 冻结第一阶段训练的 Q-Former
- 冻结预训练 LLM
- 只训练一个线性投影层
- 将视觉特征转换为 LLM 可理解的表示
- 使用指令数据集微调



*第二阶段的预训练任务

实验设置概述

- **预训练数据**：1.4 亿图像-文本对
- **视觉编码器**：ViT-L/14 和 ViT-G (从 CLIP 预训练)
- **语言模型**：
  - 指令调优型：Flan-T5 (XL 和 XXL)
  - 无监督型：OPT (2.7B 和 6.7B)
- **评估任务**：视觉问答、图像字幕生成、图像-文本检索、零样本生成能力

南方科技大学 SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

零样本视觉语言任务性能

BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models

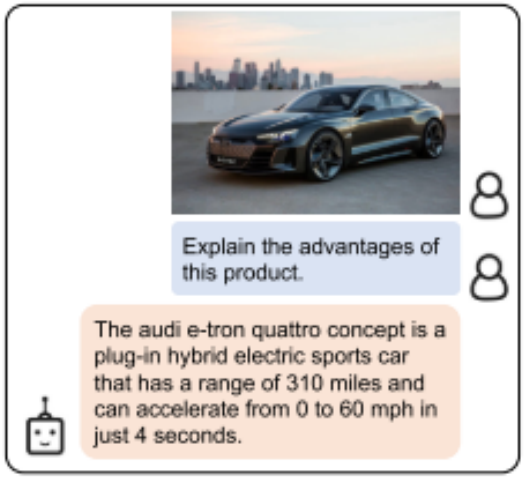
Models	#Trainable Params	Open-sourced?	Visual Question Answering VQAv2 (test-dev) VQA acc.	Image Captioning NoCaps (val) CIDEr	Image Captioning NoCaps (val) SPICE	Image-Text Retrieval Flickr (test) TR@1	Image-Text Retrieval Flickr (test) IR@1
BLIP (Li et al., 2022)	583M	✓	-	113.2	14.8	96.7	86.7
SimVLM (Wang et al., 2021b)	1.4B	✗	-	112.2	-	-	-
BEIT-3 (Wang et al., 2022b)	1.9B	✗	-	-	-	94.9	81.5
Flamingo (Alayrac et al., 2022)	10.2B	✗	56.3	-	-	-	-
BLIP-2	188M	✓	**65.0**	**121.6**	**15.8**	**97.6**	**89.7**

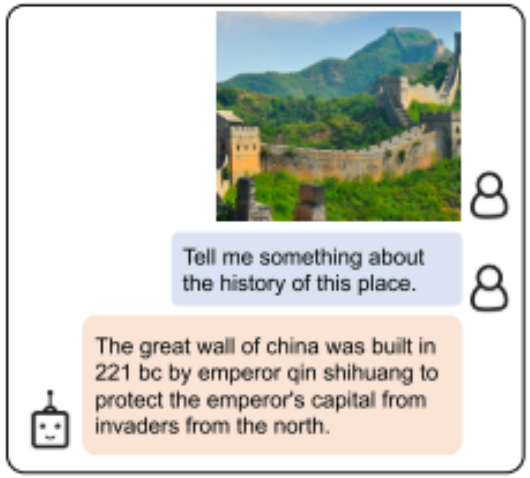
Table 1. Overview of BLIP-2 results on various **zero-shot** vision-language tasks. Compared with previous state-of-the-art models. BLIP-2 achieves the highest zero-shot performance while requiring the least number of trainable parameters during vision-language pre-training.

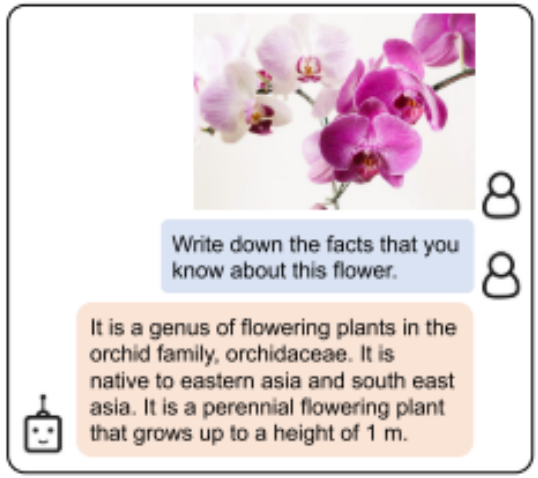
- BLIP-2 达到或超越 SOTA 性能，同时训练参数量显著减少
- Flamingo 训练 80B 参数，而 BLIP-2 只需训练 188M 参数
- 通过复用预训练模型能力，大幅提高参数效率

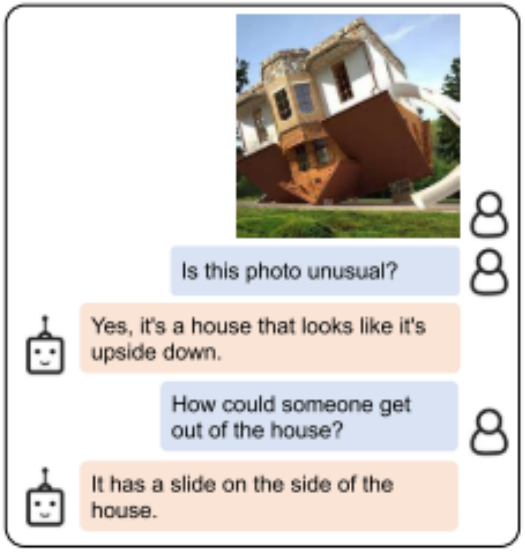
南方科技大学 SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

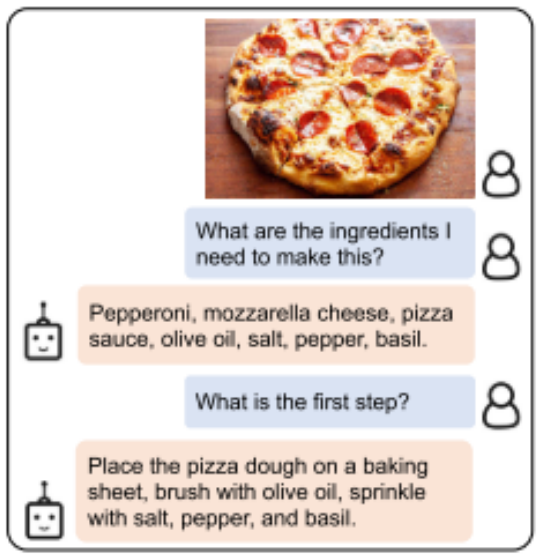
零样本图像到文本生成能力











零样本 VQA 性能分析

Models	#Trainable Params	#Total Params	VQAv2 val	VQAv2 test-dev	OK-VQA test	GQA test-dev
VL-T5$_{\text{no-vqa}}$	224M	269M	13.5	-	5.8	6.3
FewVLM (Jin et al., 2022)	740M	785M	47.7	-	16.5	29.3
Frozen (Tsimpoukelli et al., 2021)	40M	7.1B	29.6	-	5.9	-
VLKD (Dai et al., 2022)	406M	832M	42.6	44.5	13.3	-
Flamingo3B (Alayrac et al., 2022)	1.4B	3.2B	-	49.2	41.2	-
Flamingo9B (Alayrac et al., 2022)	1.8B	9.3B	-	51.8	44.7	-
Flamingo80B (Alayrac et al., 2022)	10.2B	80B	-	56.3	**50.6**	-
BLIP-2 ViT-L OPT$_{2.7B}$	104M	3.1B	50.1	49.7	30.2	33.9
BLIP-2 ViT-G OPT$_{2.7B}$	107M	3.8B	53.5	52.3	31.7	34.6
BLIP-2 ViT-G OPT$_{6.7B}$	108M	7.8B	54.3	52.6	36.4	36.4
BLIP-2 ViT-L FlanT5$_{XL}$	103M	3.4B	62.6	62.3	39.4	44.4
BLIP-2 ViT-G FlanT5$_{XL}$	107M	4.1B	63.1	63.0	40.7	44.2
BLIP-2 ViT-G FlanT5$_{XXL}$	108M	12.1B	**65.2**	**65.0**	45.9	**44.7**

Table 2. Comparison with state-of-the-art methods on zero-shot visual question answering.

- 关键发现：
  - 更强的图像编码器或更大的 LLM 都能提升 BLIP-2 性能
  - 指令调优的 FlanT5 显著优于无监督的 OPT
  - ViT-G + FlanT5-XXL 组合取得最佳性能：VQAv2 达 65.0%，GQA 达 44.7%
- 证明了 BLIP-2 架构的灵活性和可扩展性

表征学习的重要性

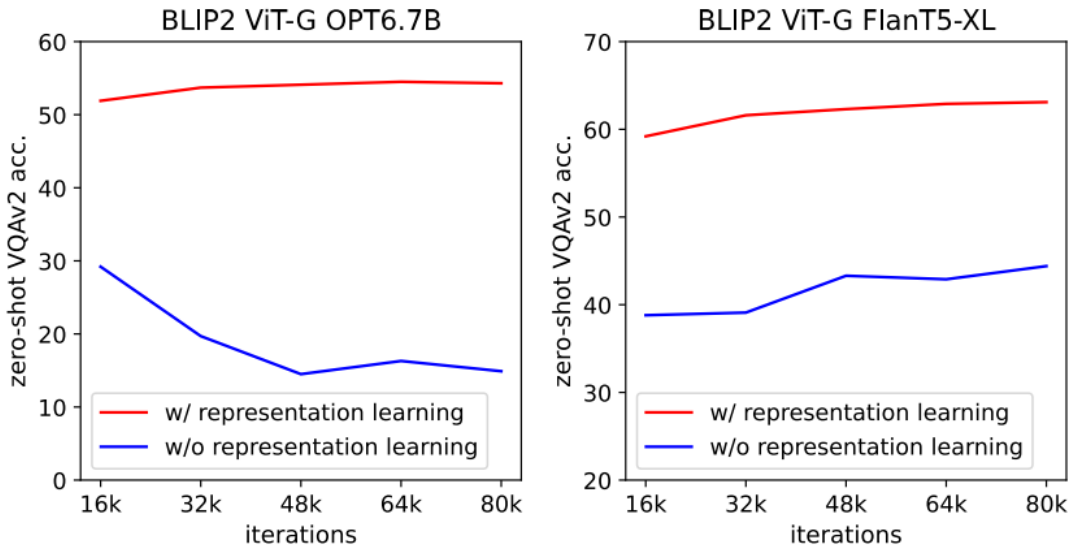


Figure 5. Effect of vision-language representation learning on vision-to-language generative learning. Without representation learning, the Q-Former fails the bridge the modality gap, leading to significantly lower performance on zero-shot VQA. CSDN @'Atlas'

- 第一阶段预训练使 Q-Former 学习提取与文本相关的视觉表征
- 消融实验结果：
  - 没有表征学习阶段，两种 LLM 在零样本 VQA 任务性能大幅下降
  - OPT 模型性能从 50%降至 30%甚至 10%，下降 20 个百分点以上
  - FlanT5 模型性能从 60%降至 40%，下降约 20 个百分点
- 证明两阶段训练策略的有效性和必要性

图像描述生成结果

Models	#Trainable Params	NoCaps Zero-shot (validation set) in-domain C	in-domain S	near-domain C	near-domain S	out-domain C	out-domain S	overall C	overall S	COCO Fine-tuned Karpathy test B@4	COCO Fine-tuned Karpathy test C
OSCAR (Li et al., 2020)	345M	-	-	-	-	-	-	80.9	11.3	37.4	127.8
VinVL (Zhang et al., 2021)	345M	103.1	14.2	96.1	13.8	88.3	12.1	95.5	13.5	38.2	129.3
BLIP (Li et al., 2022)	446M	114.9	15.2	112.1	14.9	115.3	14.4	113.2	14.8	40.4	136.7
OFA (Wang et al., 2022a)	930M	-	-	-	-	-	-	-	-	**43.9**	145.3
Flamingo (Alayrac et al., 2022)	10.6B	-	-	-	-	-	-	-	-	-	138.1
SimVLM (Wang et al., 2021b)	~1.4B	113.7	-	110.9	-	115.2	-	112.2	-	40.6	143.3
BLIP-2 ViT-G $OPT_{2.7B}$	1.1B	123.0	15.8	117.8	15.4	123.4	**15.1**	119.7	15.4	43.7	**145.8**
BLIP-2 ViT-G $OPT_{6.7B}$	1.1B	**123.7**	15.8	119.2	15.3	124.4	14.8	121.0	15.3	43.5	145.2
BLIP-2 ViT-G $FlanT5_{XL}$	1.1B	**123.7**	**16.3**	**120.2**	**15.9**	**124.8**	**15.1**	**121.6**	**15.8**	42.4	144.5

Table 3. Comparison with state-of-the-art image captioning methods on NoCaps and COCO Caption. All methods optimize the cross-entropy loss during finetuning. C: CIDEr, S: SPICE, B@4: BLEU@4.

- BLIP-2 在 COCO 和 NoCaps 数据集上均达到 SOTA 性能：
  - COCO: 145.8 CIDEr 分数
  - NoCaps: 15.8 CIDEr 分数
- 在 NoCaps 上的优异表现展示了强大的跨域生成能力
- 能够准确描述训练中未见过的对象和场景

南方科技大学 SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

视觉问答任务分析

Models	#Trainable Params	VQAv2 val	VQAv2 test-dev
Open-ended generation models			
ALBEF (Li et al., 2021)	314M	75.84	76.04
BLIP (Li et al., 2022)	385M	78.25	78.32
OFA (Wang et al., 2022a)	930M	82.00	82.00
Flamingo80B (Alayrac et al., 2022)	10.6B	82.00	82.10
BLIP-2 ViT-G FlanT5$_{XL}$	1.2B	81.55	81.66
BLIP-2 ViT-G OPT$_{2.7B}$	1.2B	81.59	81.74
BLIP-2 ViT-G OPT$_{6.7B}$	1.2B	**82.19**	**82.30**
Closed-ended classification models			
VinVL	345M	76.52	76.60
SimVLM (Wang et al., 2021b)	∼1.4B	80.03	80.34
CoCa (Yu et al., 2022)	2.1B	82.30	82.30
BEIT-3 (Wang et al., 2022b)	1.9B	**84.19**	**84.03**

Table 4. Comparison with state-of-the-art models fine-tuned for visual question answering.

- **创新输入处理方式**：
  - 问题文本输入到 Q-Former，引导关注图像相关区域
  - 通过 self-attention 机制，query 与问题交互
  - cross-attention 关注图像中与问题相关的区域
- **结果**：在开放式生成模型中达到 SOTA 表现

图像文本检索性能

Model	#Trainable Params	Flickr30K Zero-shot (1K test set)						COCO Fine-tuned (5K test set)					
		Image → Text			Text → Image			Image → Text			Text → Image		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
Dual-encoder models													
CLIP (Radford et al., 2021)	428M	88.0	98.7	99.4	68.7	90.6	95.2	-	-	-	-	-	-
ALIGN (Jia et al., 2021)	820M	88.6	98.7	99.7	75.7	93.8	96.8	77.0	93.5	96.9	59.9	83.3	89.8
FILIP (Yao et al., 2022)	417M	89.8	99.2	99.8	75.0	93.4	96.3	78.9	94.4	97.4	61.2	84.3	90.6
Florence (Yuan et al., 2021)	893M	90.9	99.1	-	76.7	93.6	-	81.8	95.2	-	63.2	85.7	-
BEIT-3(Wang et al., 2022b)	1.9B	94.9	99.9	**100.0**	81.5	95.6	97.8	84.8	96.5	98.3	67.2	**87.7**	**92.8**
Fusion-encoder models													
UNITER (Chen et al., 2020)	303M	83.6	95.7	97.7	68.7	89.2	93.9	65.7	88.6	93.8	52.9	79.9	88.0
OSCAR (Li et al., 2020)	345M	-	-	-	-	-	-	70.0	91.1	95.5	54.0	80.8	88.5
VinVL (Zhang et al., 2021)	345M	-	-	-	-	-	-	75.4	92.9	96.2	58.8	83.5	90.3
Dual encoder + Fusion encoder reranking													
ALBEF (Li et al., 2021)	233M	94.1	99.5	99.7	82.8	96.3	98.1	77.6	94.3	97.2	60.7	84.3	90.5
BLIP (Li et al., 2022)	446M	96.7	**100.0**	**100.0**	86.7	97.3	98.7	82.4	95.4	97.9	65.1	86.3	91.8
BLIP-2 ViT-L	474M	96.9	**100.0**	**100.0**	88.6	97.6	**98.9**	83.5	96.0	98.0	66.3	86.5	91.8
BLIP-2 ViT-G	1.2B	**97.6**	**100.0**	**100.0**	**89.7**	**98.1**	**98.9**	**85.4**	**97.0**	**98.5**	**68.3**	**87.7**	92.6

Table 5. Comparison with state-of-the-art image-text retrieval methods, finetuned on COCO and zero-shot transferred to Flickr30K.

- 检索过程：先根据相似度选 128 个候选，再根据 ITM 分数排序
- BLIP-2 在 COCO 和 Flickr30K 上均显著超越之前方法：
  - COCO 图像 → 文本检索：85.4% (vs. BLIP 的 82.4%)
  - Flickr30K 文本 → 图像检索：89.7% (vs. BLIP 的 86.7%)

实验结果总结

- **关键发现**：
  - BLIP-2 在各类视觉-语言任务上都取得 SOTA 或接近 SOTA 性能
  - 通过冻结预训练模型，极大降低训练参数量和计算成本
  - 视觉编码器和语言模型质量对 BLIP-2 性能有直接影响
  - 两阶段预训练策略对模型性能至关重要
- **优势**：
  - 参数效率高：仅训练 188M 参数即可达到优异性能
  - 灵活性：可与不同规模视觉和语言模型结合
  - 泛化能力强：在跨域和零样本任务中表现突出

南方科技大学 SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

局限性分析

- **技术层面局限**：
  - 对视觉编码器质量的强依赖性
  - Q-Former 可能成为信息瓶颈
  - 难以处理超出 ViT 预训练范围的视觉场景
  - 对时序信息和空间关系建模不足
- **应用层面挑战**：
  - 上下文理解有限，难以把握复杂场景
  - 在专业领域 (如医学、科学图像) 表现不佳
  - 与人类偏好的对齐问题
  - 缺乏多模态的知识推理能力
- **潜在的社会影响**：
  - 可能继承预训练模型中的偏见和刻板印象
  - 对隐私和版权内容的识别与生成问题

南方科技大学 SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

后续研究方向与可能的突破点

架构改进

- 动态适应的 Q-Former，根据任务调整查询数
- 多级别视觉特征融合（局部+全局）
- 探索 Perceiver-IO 等替代架构
- 将 Transformer 替换为更高效结构

训练与应用

- 整合多源知识（如常识、关系等）
- 强化学习用于优化视觉特征选择
- 探索连续学习方法适应新模态
- 用于生成具身 AI 的视觉基础

个人观点：未来视觉-语言模型的关键在于**高效适应性连接**而非模型规模。BLIP-2 的思路比纯粹扩大模型更有前途。

南方科技大学 SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

结论与未来工作

- **主要贡献**：
  - 提出 Q-Former 架构，高效连接视觉和语言模型
  - 实现参数高效的预训练策略，显著降低计算成本
  - 在多项视觉-语言任务上取得 SOTA 结果
- **局限性**：
  - 对特定领域知识的理解仍有不足
  - 在某些复杂推理任务上表现不稳定
- **未来工作**：
  - 扩展到视频、3D 等多模态数据
  - 增强模型的推理和知识整合能力
  - 研究更高效的模态对齐和融合方法

南方科技大学 SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

参考文献

- Li, J. et al. (2023). BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. arXiv preprint arXiv:2301.12597.
- Radford, A. et al. (2021). Learning Transferable Visual Models From Natural Language Supervision. ICML 2021.
- Alayrac, J. et al. (2022). Flamingo: a Visual Language Model for Few-Shot Learning. NeurIPS 2022.

南方科技大学 SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

谢谢观看!

南方科技大学
SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY