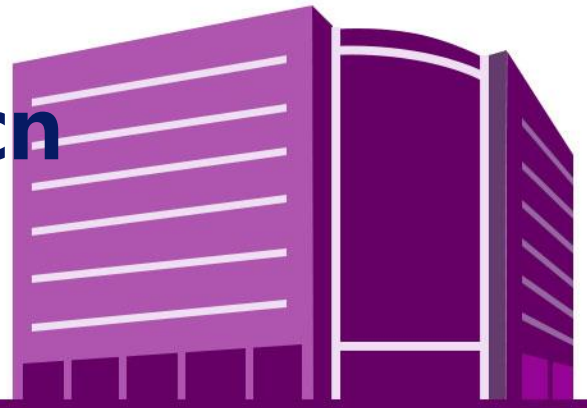
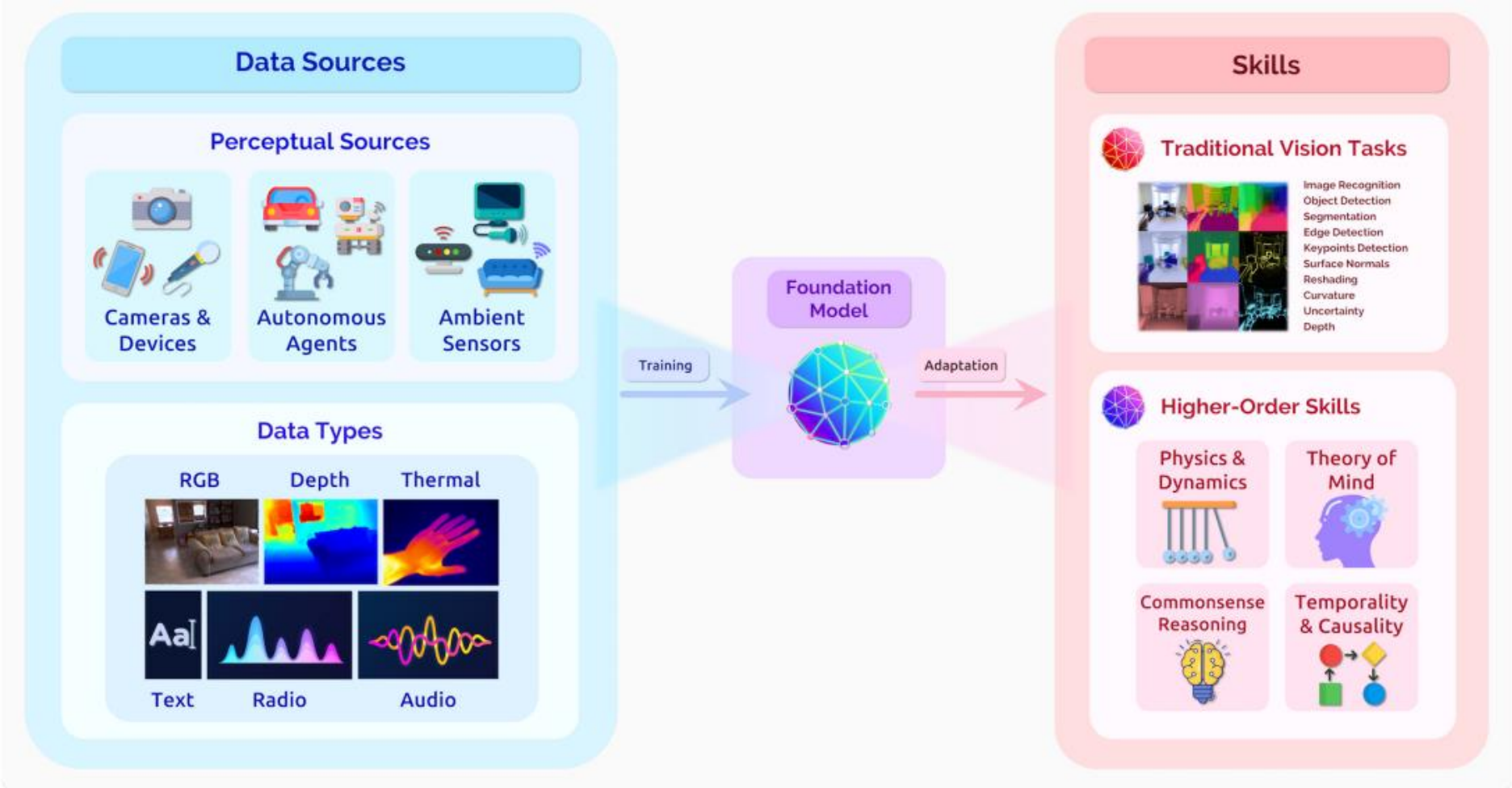
从通用感知模型到通用智能体模型

代季峰
电子工程系
daijifeng@tsinghua.edu.cn
2025年5月30日



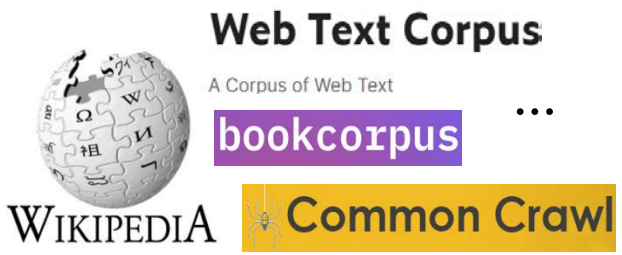
基石模型（Foundation Model）



A foundation model is any model that is trained on broad data (generally using self-supervision at scale) that can be adapted (e.g., fine-tuned) to a wide range of downstream tasks.

R. Bommasani et al., "**On the opportunities and risks of foundation models**", arxiv (2021).

LLM强大的通用能力



Pretraining

Task: Sort the given list ascendingly.
Input: [10, 92, 2, 5, -4, 92, 5, 101]
Output: [-4, 2, 5, 5, 10, 92, 92, 101]
Input: [9.99, 10, -5, -1000, 5e6, 999]
Output: ...

Instruction Tuning

LLM

Few-shot Prompting
Instruction Prompting

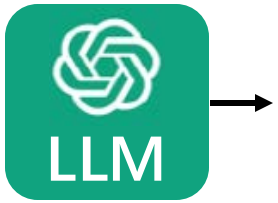
Question answering
Reading comprehension
Sentence classification
Translation
Common sense reasoning
Arithmetic reasoning
Summarization
Sentence completion
Code generation
...

Downstream Tasks

LLM的通用性：

- 统一的训练方式：Pretraining + Instruction Tuning
- 统一的任务形式：Few-shot Prompting / Instruction Prompting
- 多样的下游任务：支持问答、理解、推理、生成等NLP领域内几乎所有下游任务

LLM强大的通用能力



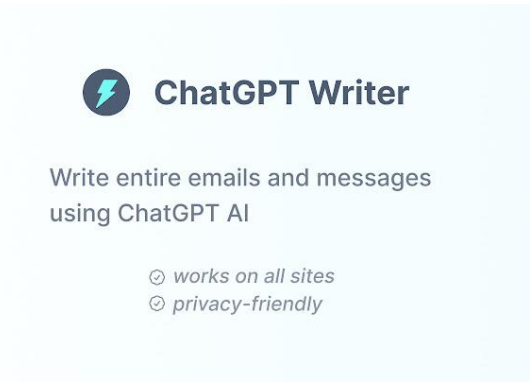

```php
<?php

// Replace "example.com" with the hostname you want to scan
$host = "example.com";

// Loop through common ports (1-1024) and try to connect to each one
for ($port = 1; $port <= 1024; $port++) {
    // Create a new socket
    $sock = socket_create(AF_INET, SOCK_STREAM, SOL_TCP);
    // Set the socket to be non-blocking
    socket_set_nonblock($sock);
    // Try to connect to the port on the host
    $connection = socket_connect($sock, $host, $port);
    // If the connection is successful, the port is open
    if ($connection === true) {
        echo "Port $port is open\n";
    }
    // Close the socket
    socket_close($sock);
}
```

写代码

生日宴会策划

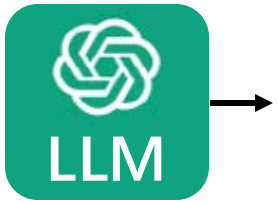


写作辅助

LLM的通用性带来的生产力变化：

- 训练成本较已有模型极高
  - 上万块GPU训练数月时间
  - 几十到几百位精英研究员
- 应用边际成本较已有模型极低
  - 无需任务特定微调
  - 仅有推理计算成本

如何拓展LLM通用能力到更多应用上?



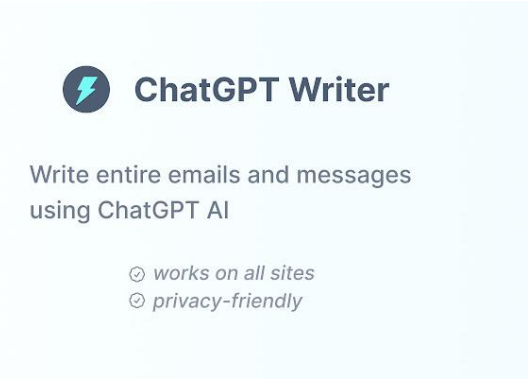

```php
<?php

// Replace "example.com" with the hostname you want to scan
$host = "example.com";

// Loop through common ports (1-1024) and try to connect to each one
for ($port = 1; $port <= 1024; $port++) {
    // Create a new socket
    $sock = socket_create(AF_INET, SOCK_STREAM, SOL_TCP);
    // Set the socket to be non-blocking
    socket_set_nonblock($sock);
    // Try to connect to the port on the host
    $connection = socket_connect($sock, $host, $port);
    // If the connection is successful, the port is open
    if ($connection === true) {
        echo "Port $port is open\n";
    }
    // Close the socket
    socket_close($sock);
}
```

写代码

生日宴会策划



写作辅助

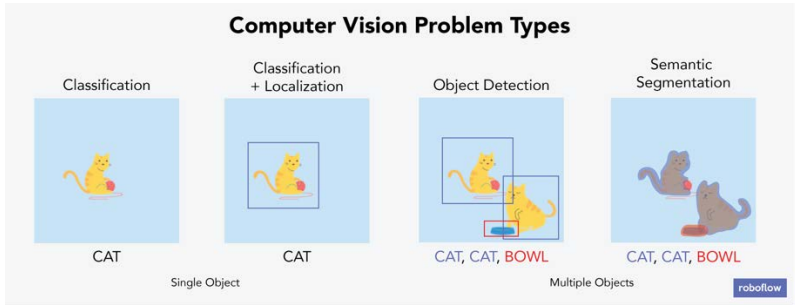
计算机视觉

虚拟环境

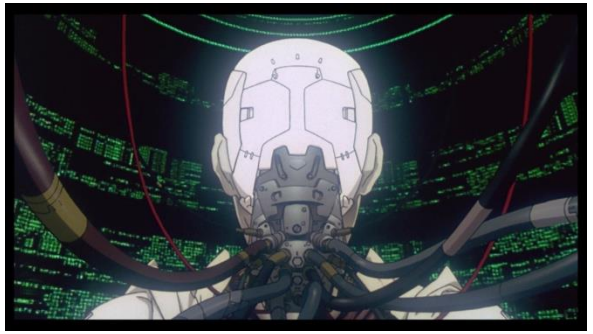
机器人

如何拓展LLM通用能力到更多应用上？

研究目标： 多模态通用模型，为大语言模型装上手脚和眼睛，与现实世界交互



视觉感知能力

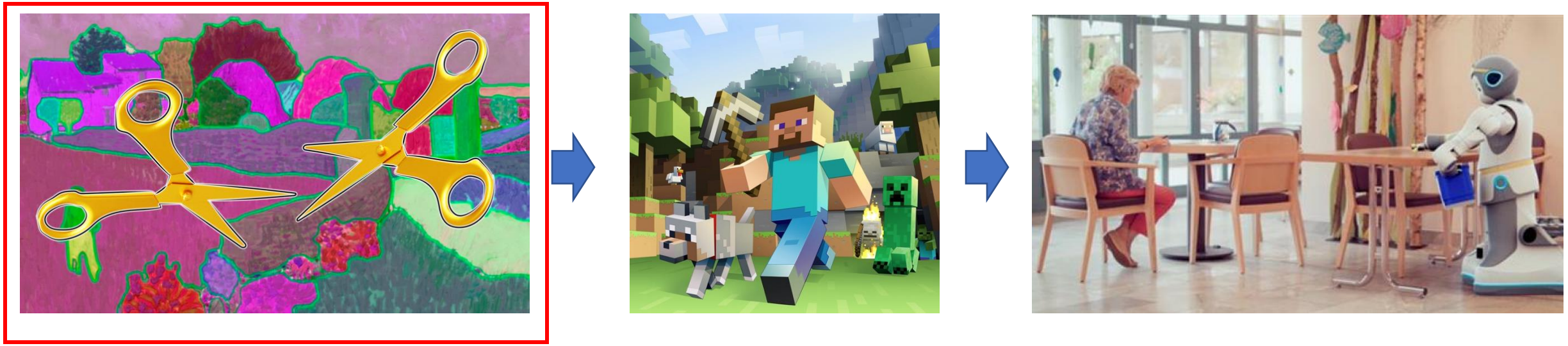


LLM高阶智能中枢

环境交互能力

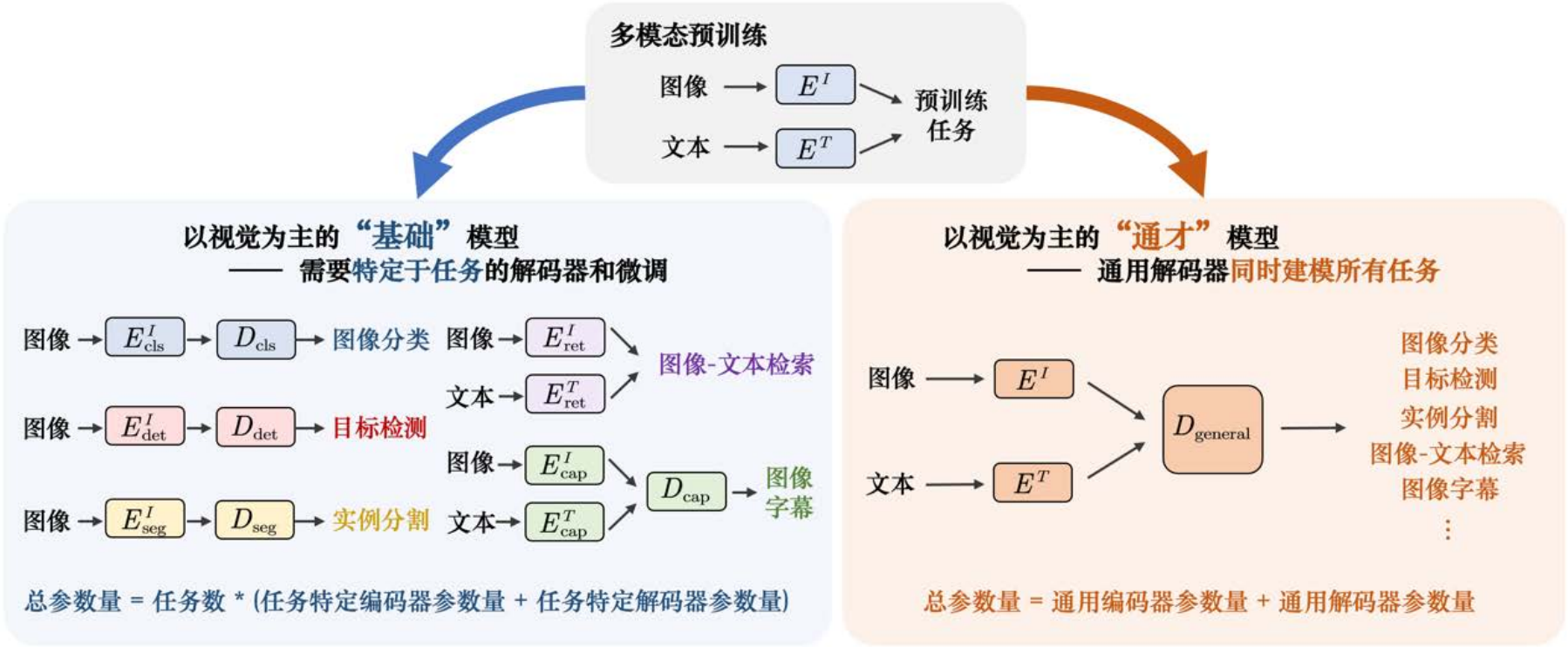
如何拓展LLM通用能力到更多应用上？

研究目标： 多模态通用模型，为大语言模型装上手脚和眼睛，与现实世界交互



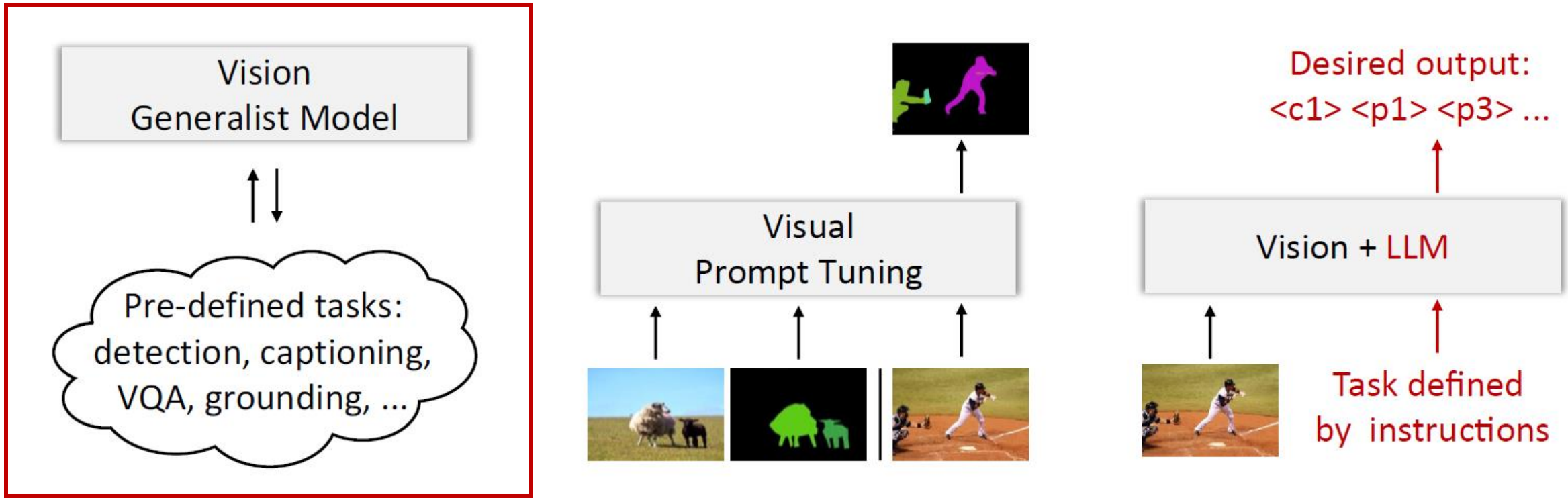
通用视觉模型

- 如何在视觉、视觉-语言任务中构建一个**通用**于各种任务的模型?
- 现有方案：**特定于任务的finetuning**
  - 针对**特定的任务**，采集**特定场景的数据集**并设计和训练**专用的解码器及参数**
  - 针对不同的应用场景，需要单独构建特定场景的数据集进行训练
  - 一个专用模型只擅长处理一项任务，面对千变万化的任务需求时，需独立开发训练成千上万个模型



VisionLLM: LLM范式的开放式视觉模型

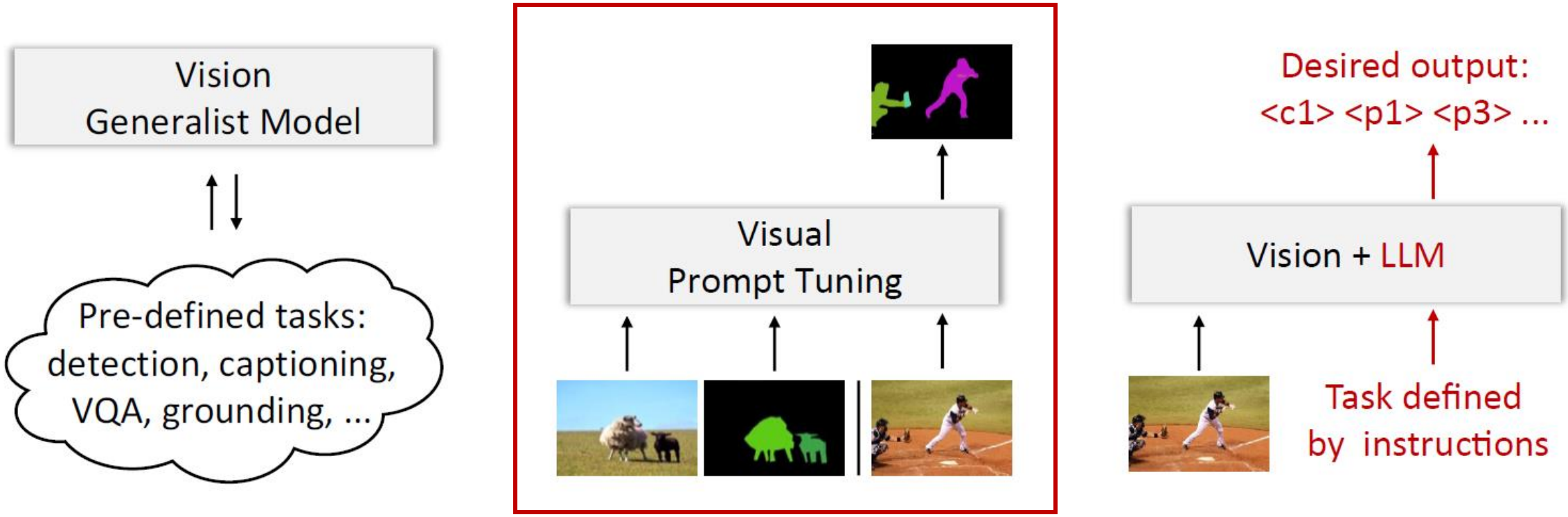
通用视觉模型的范式转换



范式1：多任务统一的通才模型

- 在一组预先设定的视觉任务上训练统一的通才模型
- 问题：仅支持预先定义的任务，难以扩展任务形式，不支持开放式任务的能力

通用视觉模型的范式转换

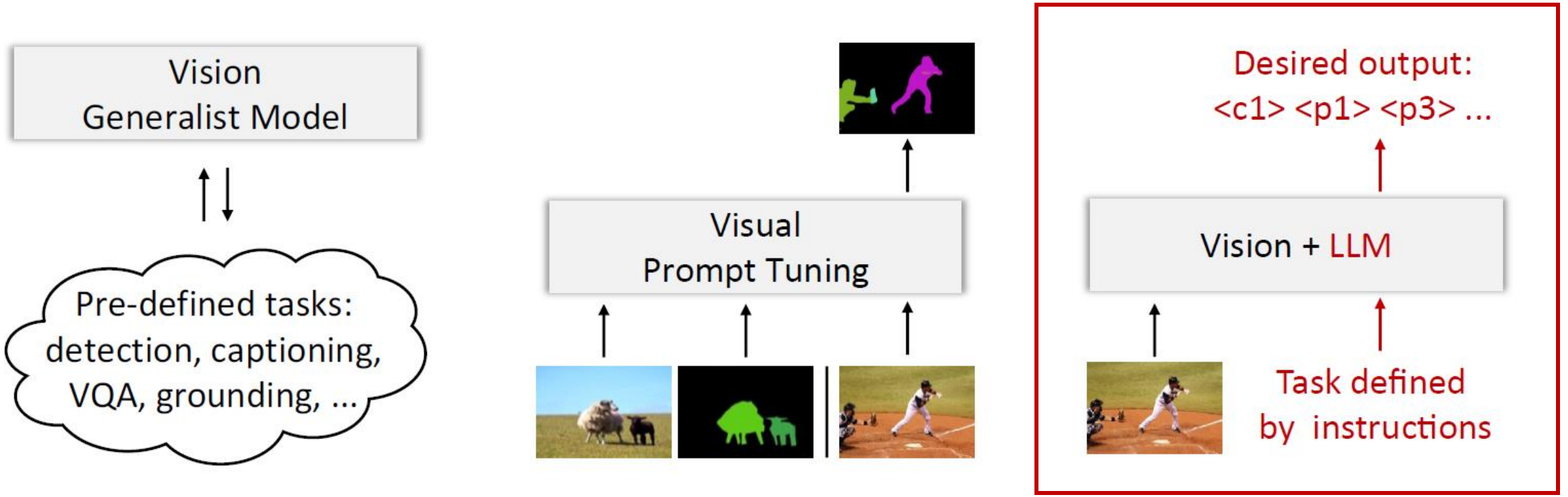


范式2：视觉提示微调

- 预训练大型基础模型 + 为每一个任务视觉任务微调一个提示嵌入(prompt)
- 问题：视觉提示与自然语言提示相差巨大，难以利用LLM的通用知识和推理能力

VisionLLM: LLM范式的开放式视觉模型

通用视觉模型的范式转换

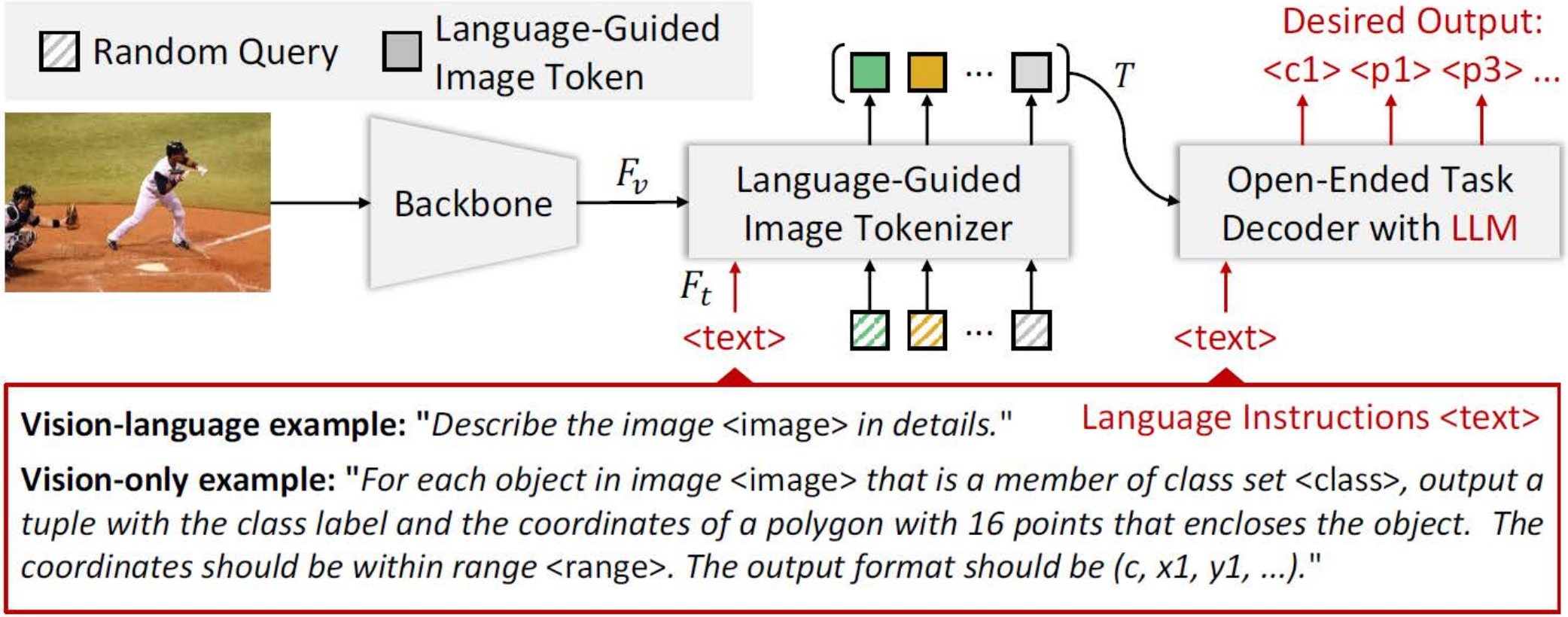


范式3：LLM框架下的开放式视觉模型

- 将图片转变为Token表达，将以视觉为中心的任务对齐到自然语言任务
- 利用LLM强大的泛化和推理能力，自然语言指令实现用户定义的开放式视觉任务

Wang et al., VisionLLM: Large Language Model is also an Open-Ended Decoder for Vision-Centric Tasks. arXiv preprint arXiv:2305.11175

VisionLLM：LLM范式建模以视觉为中心的任务

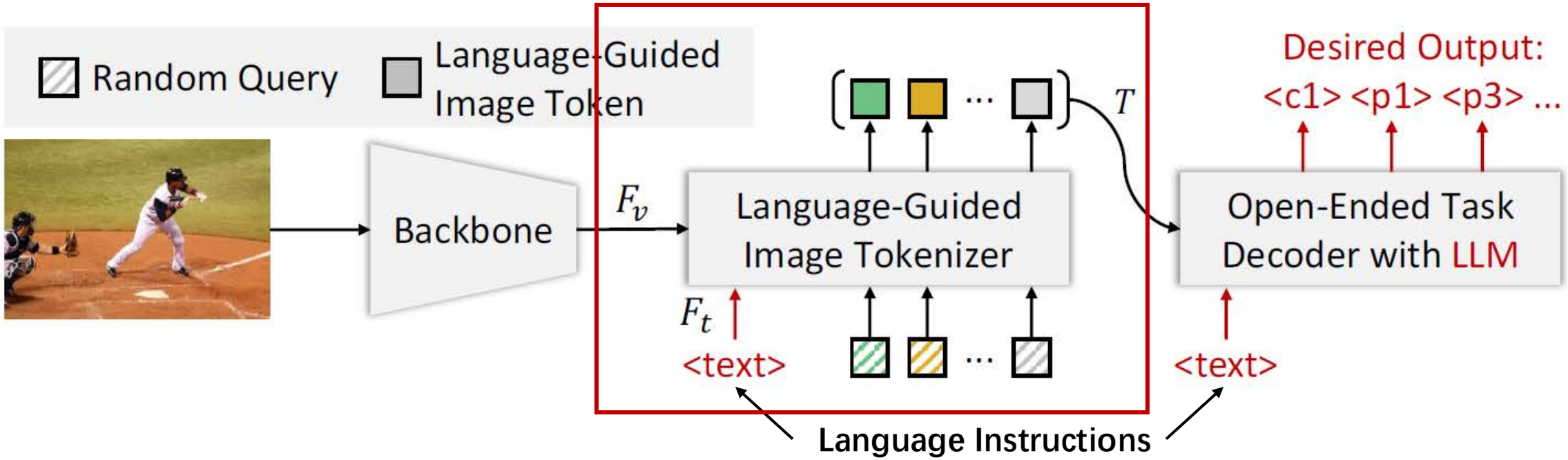


设计1：统一的自然语言任务指令

- 将任务的定义和输出格式统一用语言指令描述
- 将坐标位置离散化，为离散坐标和物体类别添加专门的token

VisionLLM: LLM范式的开放式视觉模型

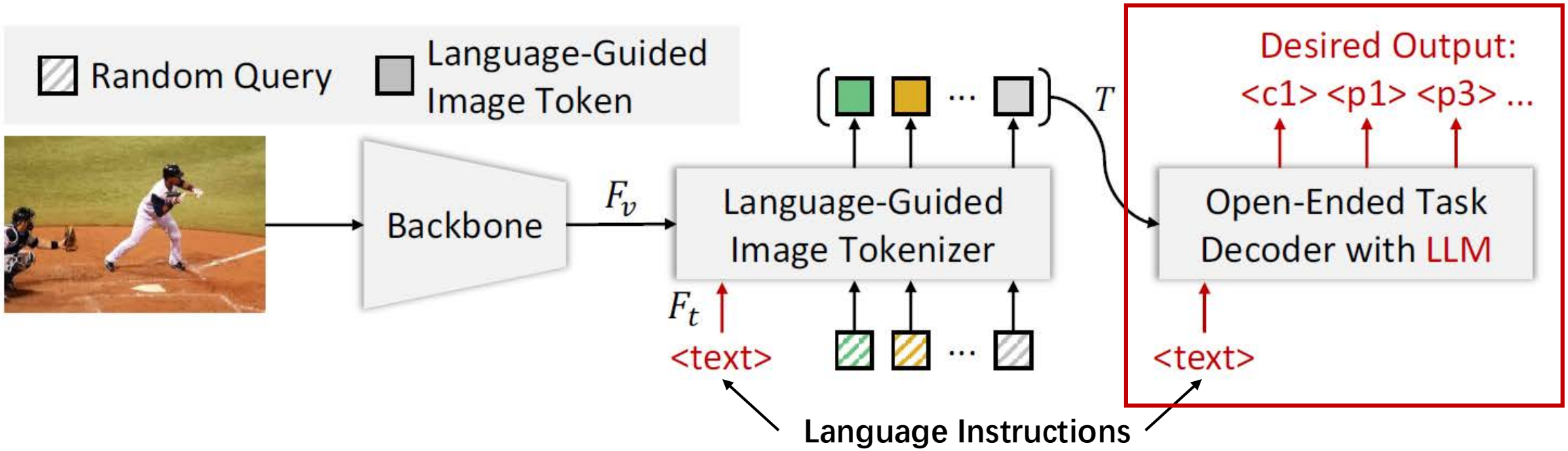
VisionLLM：LLM范式建模以视觉为中心的任务



设计2：将图像信息转变为Token表达

- 将图像视为一门外语，使用图像分词器转变为Token表达
- 语言指导下的图像分词器：根据语言指令提取视觉信息

VisionLLM：LLM范式建模以视觉为中心的任务

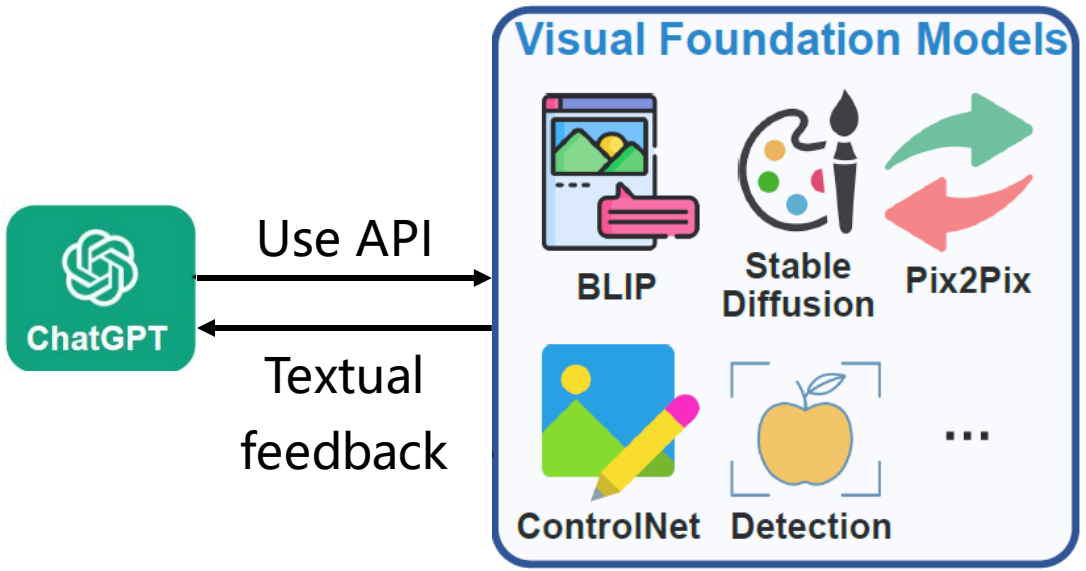


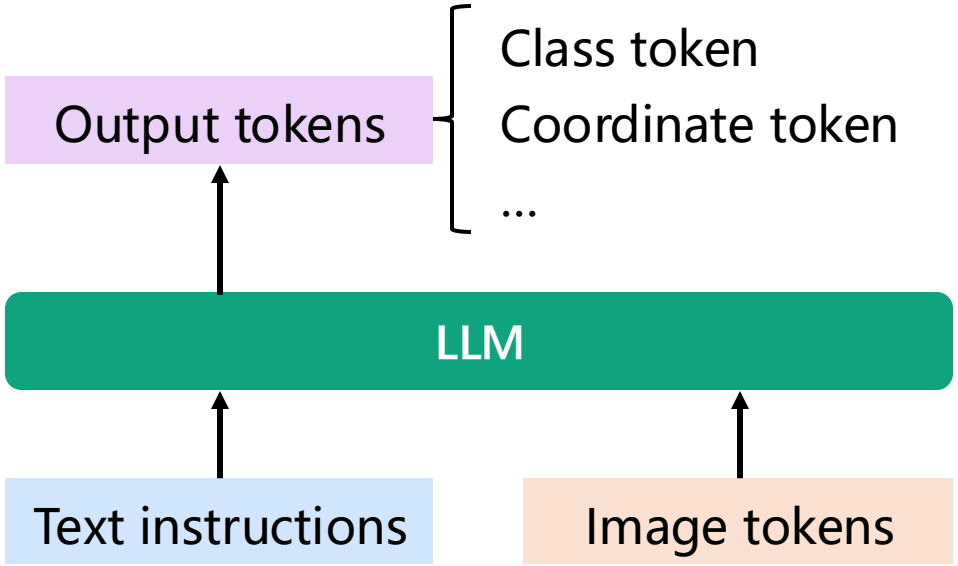
设计3：LLM作为开放式任务解码器

- 在LLM的单词表中添加坐标位置和物体类别的特定token
- 将视觉任务统一为**Token分类**形式：预测坐标或类别Token

VisionLLM: LLM范式的开放式视觉模型

VisionLLM对比LLM+视觉API





LLM+API接口：

- 文本描述的视觉信息非常有限
- 需要多个专家模型，分别训练
- 只有部分样例成功，效果不稳定

VisionLLM：

- Token embedding携带大量视觉信息
- 一个统一的模型，端到端训练
- 在多个benchmark上接近SOTA结果

能力1：标准视觉任务的通用性

Method	Backbone	Open-Ended	Detection			Instance Seg.			Grounding	Captioning	
			AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	P@0.5	BLEU-4	CIDEr
Specialist Models											
Deformable-DETR	ResNet-50	-	45.7	65.0	49.1	-	-	-	-	-	-
Mask R-CNN	ResNet-50	-	41.0	61.7	44.9	37.1	58.4	40.1	-	-	-
Polar Mask	ResNet-50	-	-	-	-	30.5	52.0	31.1	-	-	-
Pix2Seq	ResNet-50	-	43.2	61.0	46.1	-	-	-	-	-	-
MDETR	ResNet-101	-	-	-	-	-	-	-	86.8	-	-
VL-T5	T5-B	-	-	-	-	-	-	-	-	-	116.5
Generalist Models											
UniTab	ResNet-101	-	-	-	-	-	-	-	88.6	-	115.8
Uni-Perceiver	ViT-B	-	-	-	-	-	-	-	-	32.0	-
Uni-Perceiver-MoE	ViT-B	-	-	-	-	-	-	-	-	33.2	-
Uni-Perceiver-V2	ViT-B	-	58.6	-	-	50.6	-	-	-	35.4	116.9
Pix2Seq v2	ViT-B	-	46.5	-	-	38.2	-	-	-	34.9	-
VisionLLM-R50	ResNet-50	✓	44.6	64.0	48.1	25.1	50.0	22.4	80.6	31.0	112.5
VisionLLM-H	Intern-H	✓	60.2	79.3	65.8	30.6	61.2	27.6	86.7	32.1	114.2

- 一套参数，只需改变语言指令即可实现检测、分割、定位、图像配文等标准任务
- COCO物体检测达到**AP 60+**，超过众多专用模型并接近SOTA水平

VisionLLM: LLM范式的开放式视觉模型

能力2：用户定义的开放式任务

- 用户可以通过自然语言指令定义开放式的视觉任务
- 支持粗粒度指定任务的自定义和细粒度指定物体的自定义

Human: *"Identify the objects in <image> that belong to {'What is the child eating?': <c0>, 'red gamepad': <c1>} and draw a bounding box around each one. The output should be a list of tuples in the format (c, x1, y1, x2, y2), where c is the class label and x1, y1, x2, y2 are the coordinates of the top-left and bottom-right corners of the bounding box. The coordinates should be within the range 1024."*

VisionLLM: *"The bounding boxes are [(<c0>, 226.4, 229.8, 363.1, 347.4), (<c1>, 441.1, 183.5, 538.6, 269.9)]."*

自定义物体类别的物体检测

Human: *"For each object in <image> that is a member of {'motorbike': <c0>}, output a tuple with the class label and the coordinates of a polygon with 16 points that encloses the object. The coordinates should be within the range 1024. The output format should be (c, x1, y1, x2, y2, ..., x16, y16)."*

VisionLLM: *"The detected polygons are [(<c0>, 135.3, 95.7, 123.4, 53.4, 84.9, 57.6, 66.8, 60.5, 60.1, 72.3, 34.2, 71.4, ..., 124.9, 119.3)]."*

自定义输出格式的物体定位

VisionLLM: LLM范式的开放式视觉模型

能力2：用户定义的开放式任务

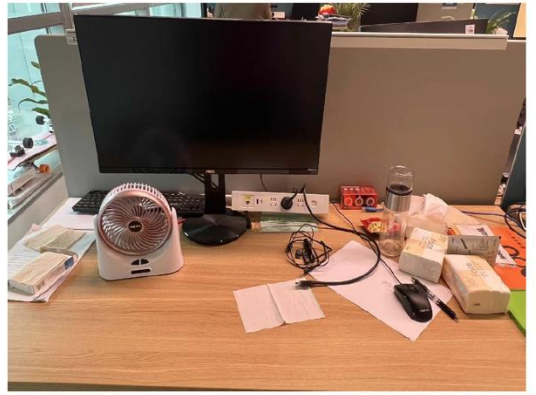
- 用户可以通过自然语言指令定义开放式的视觉任务
- 支持粗粒度指定任务的自定义和细粒度指定物体的自定义

Human: "*Provide a long description of the <image>.*"

VisionLLM: "*The image features a white dog laying down on a foldable outdoor cooling pad , which is placed on top of a wooden table. The dog has its head in the middle of the cooling pad, and a fan is positioned behind it, blowing air onto the pad. The table is outdoors, in a grassy area, and there are two chairs placed around it, one on each side. The scene appears to be set in a sunny weather, with a bright and sunny sky, which contributes to the overall outdoor atmosphere.*"

可控长度的图像描述

Human: "*With the objects in the <image>, if you feel hot, what will you do?*"

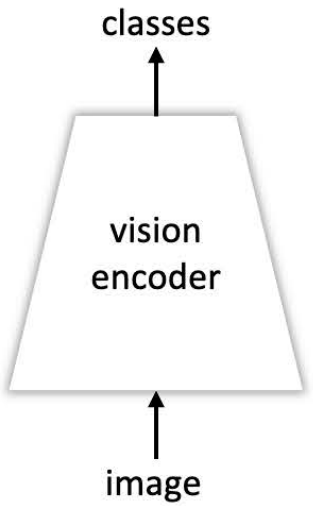


VisionLLM: "*First, You can also try turning off the computer and other electrical appliances in the room to reduce the heat generated by their operation. As an alternative, you can use the desk fan provided in the image to create a cooling effect in the room. You can plug in the fan and adjust its settings to ensure it blows air in a direction that reaches the room's occupants effectively.*"

复杂推理的图文问答

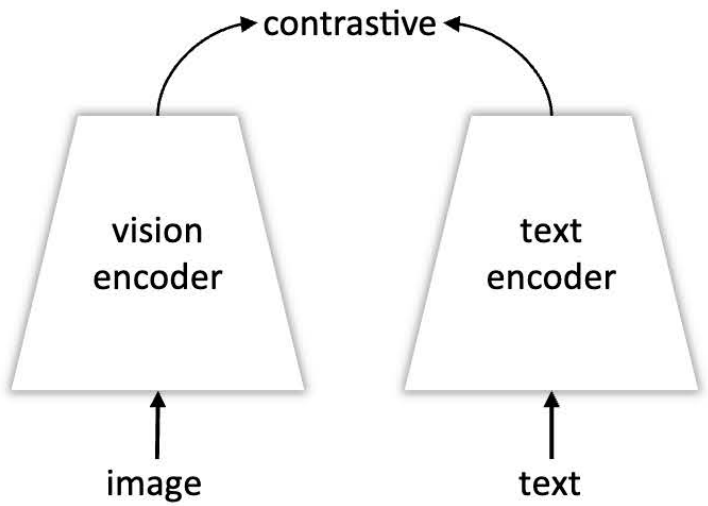
InternVL：最强开源视觉基础模型

ImageNet时代



(a) Supervised pre-training

CLIP时代



(b) Contrastive pre-training

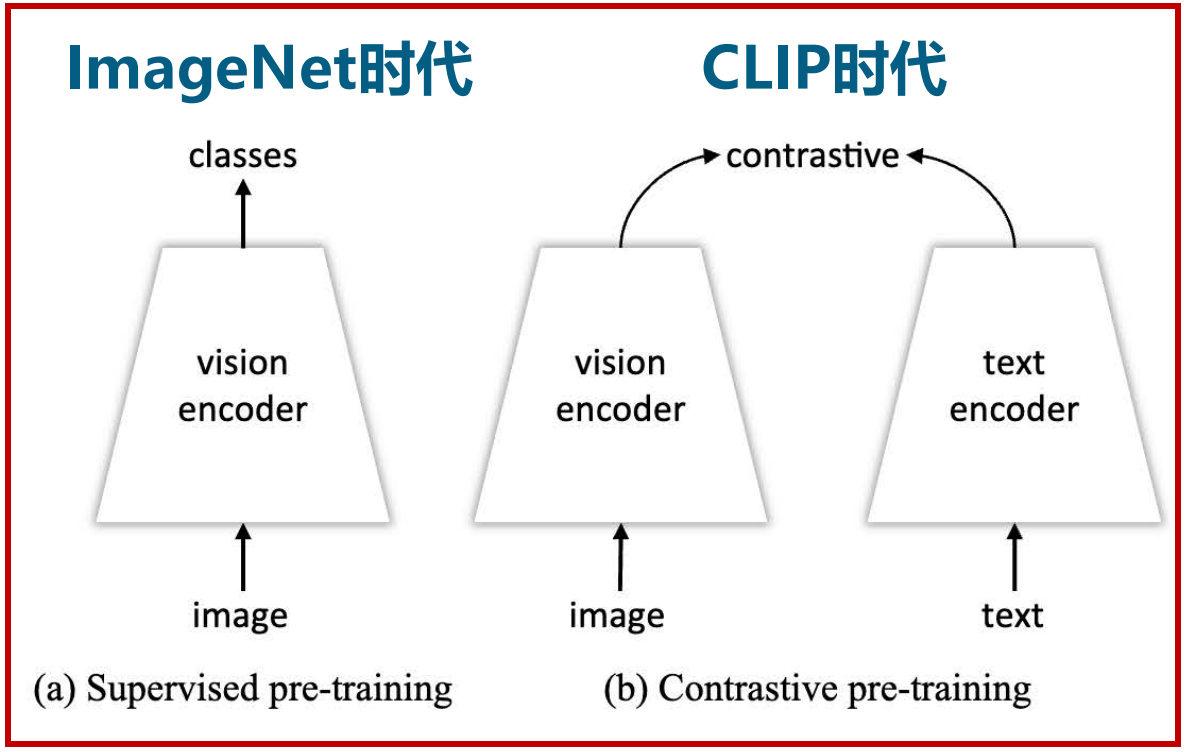
大模型时代

视觉基础模型（视觉编码器）在计算机视觉任务中至关重要，也是构建多模态大模型的重要基础。然而，目前视觉基础模型的发展速度落后于大型语言模型的迅猛进展

Chen et. al., InternVL: Scaling up Vision Foundation Models and Aligning for Generic Visual-Linguistic Tasks. Arxiv preprint, 2023.

InternVL：最强开源视觉基础模型

视觉/视觉-语言基础模型的范式转换



大模型时代

?

范式1：有监督判别式预训练

在ImageNet上进行**有监督判别式**预训练

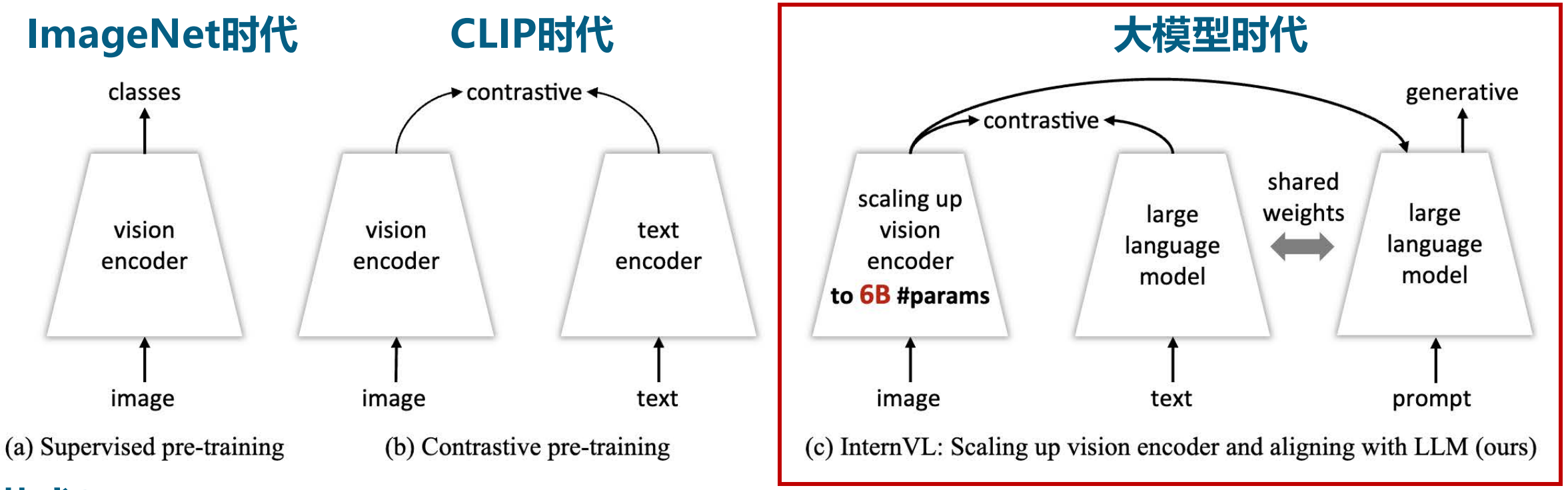
范式2：对比学习预训练

在大规模图文配对数据上与**非生成式语言模型**（如：BERT）对齐

问题： 现有的视觉和视觉-语言基础模型和LLM的参数量差距过大，且它们表征也不一致，难以对齐

InternVL：最强开源视觉基础模型

视觉/视觉-语言基础模型的范式转换



(a) Supervised pre-training (b) Contrastive pre-training (c) InternVL: Scaling up vision encoder and aligning with LLM (ours)

范式3：InternVL

扩大视觉模型的参数规模至6B，与LLM的参数量匹配；

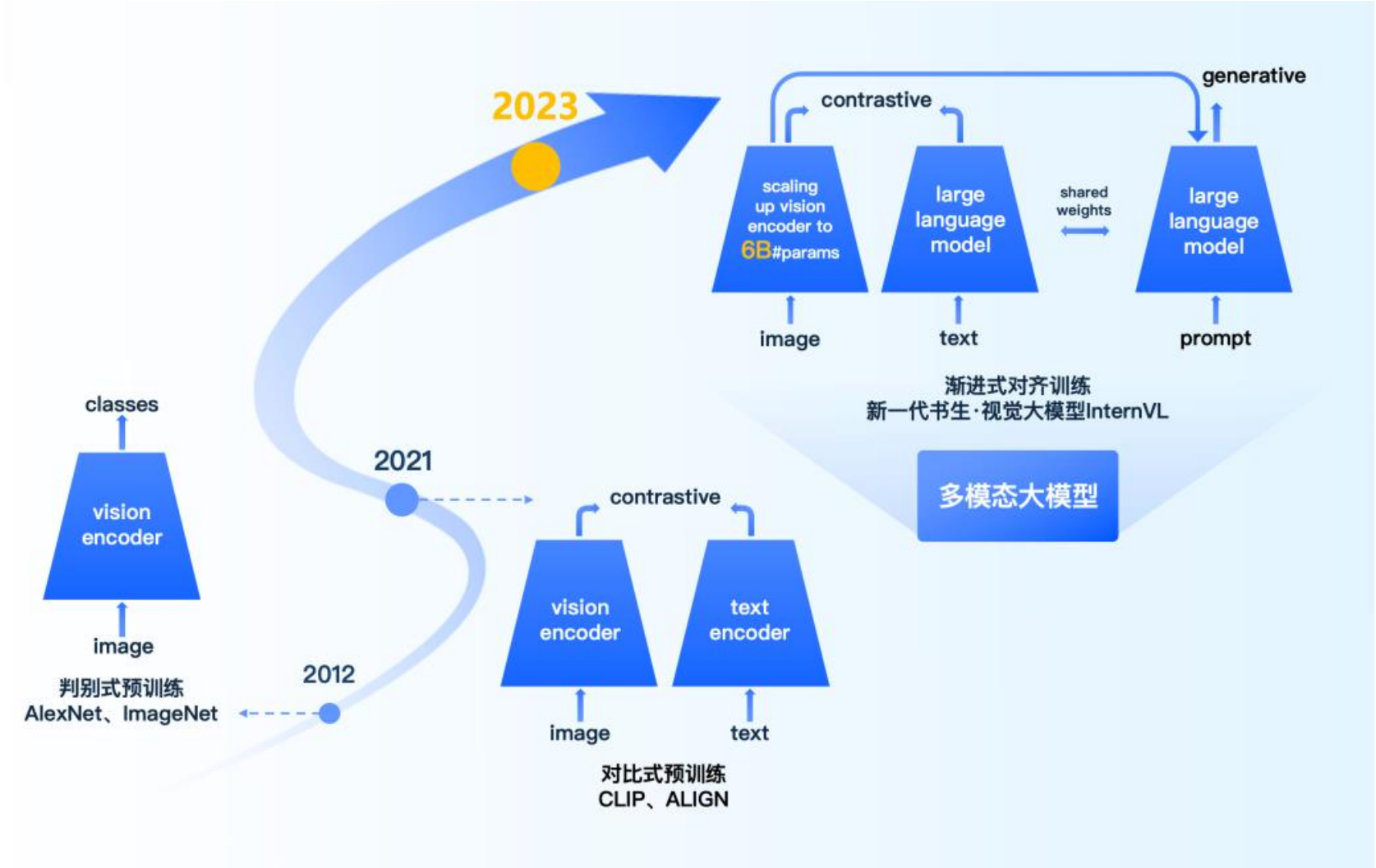
在多种来源的数据集上和LLM（LLaMA）进行对齐；

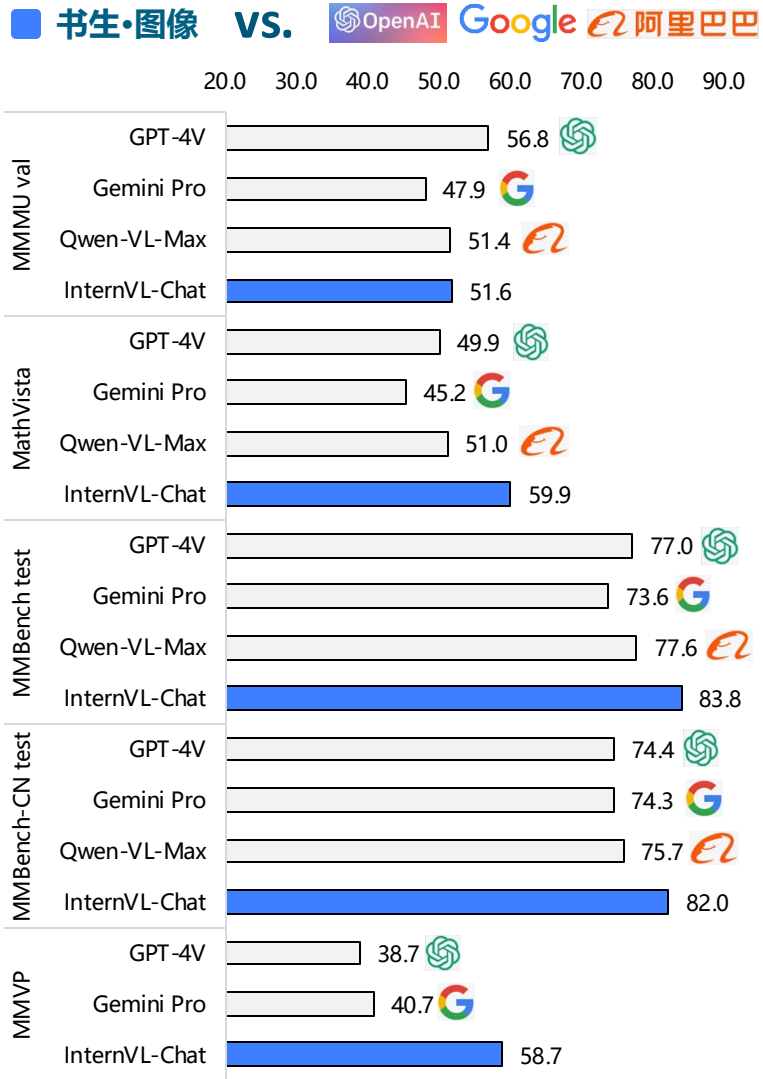
同时支持对比学习任务和生成式任务；

有利于配合LLM构建多模态对话系统。

InternVL：最强开源视觉基础模型

最强开源视觉基础模型，实现了在互联网级别数据上视觉大模型与语言大模型的精细对齐，以不到1/3的参数量超越视觉模型标杆谷歌ViT-22B，在MMMU等评测上比肩GPT-4V和Gemini Pro



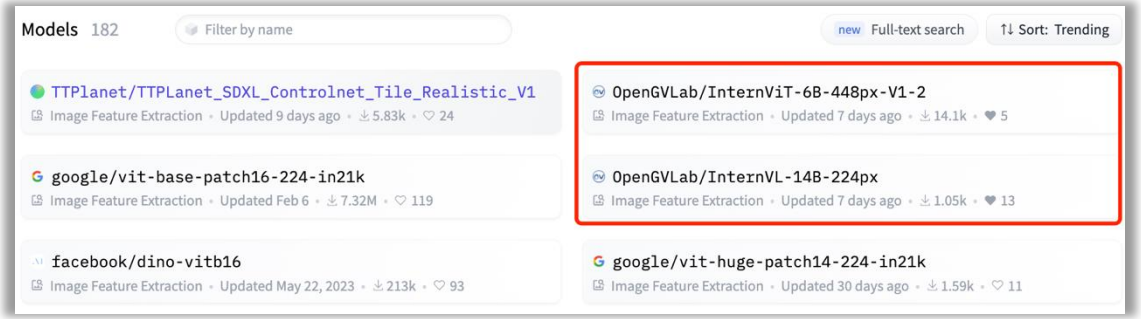


InternVL：最强开源视觉基础模型

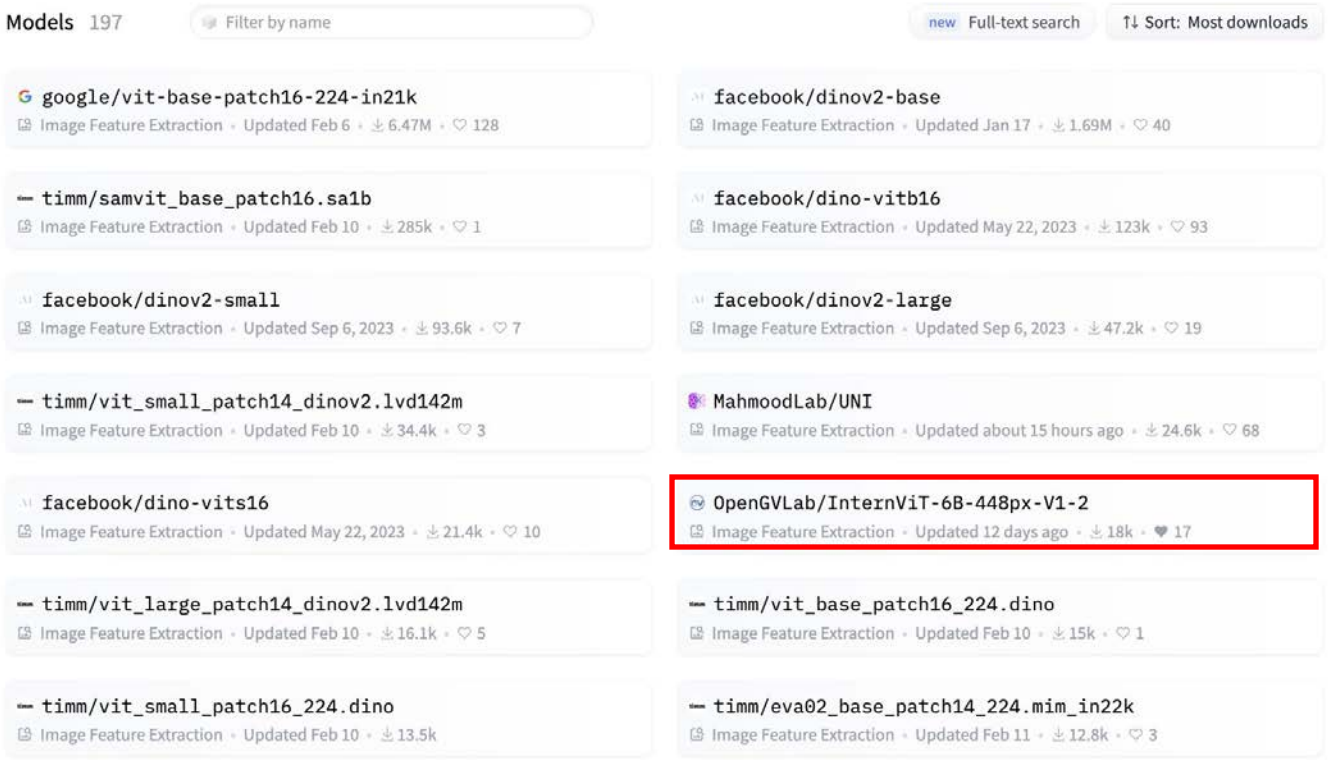
最强开源视觉基础模型，从2023年12月发布至今，连续7天在HuggingFace平台下载量增长趋势榜单上名列前3，在图像抽取模型总下载量榜单上排列前10，在国际上具有较大的影响力

新一代书生·视觉大模型系列多个模型下载量超过1万次

连续7天在下载量增长趋势榜单上榜，名列前3

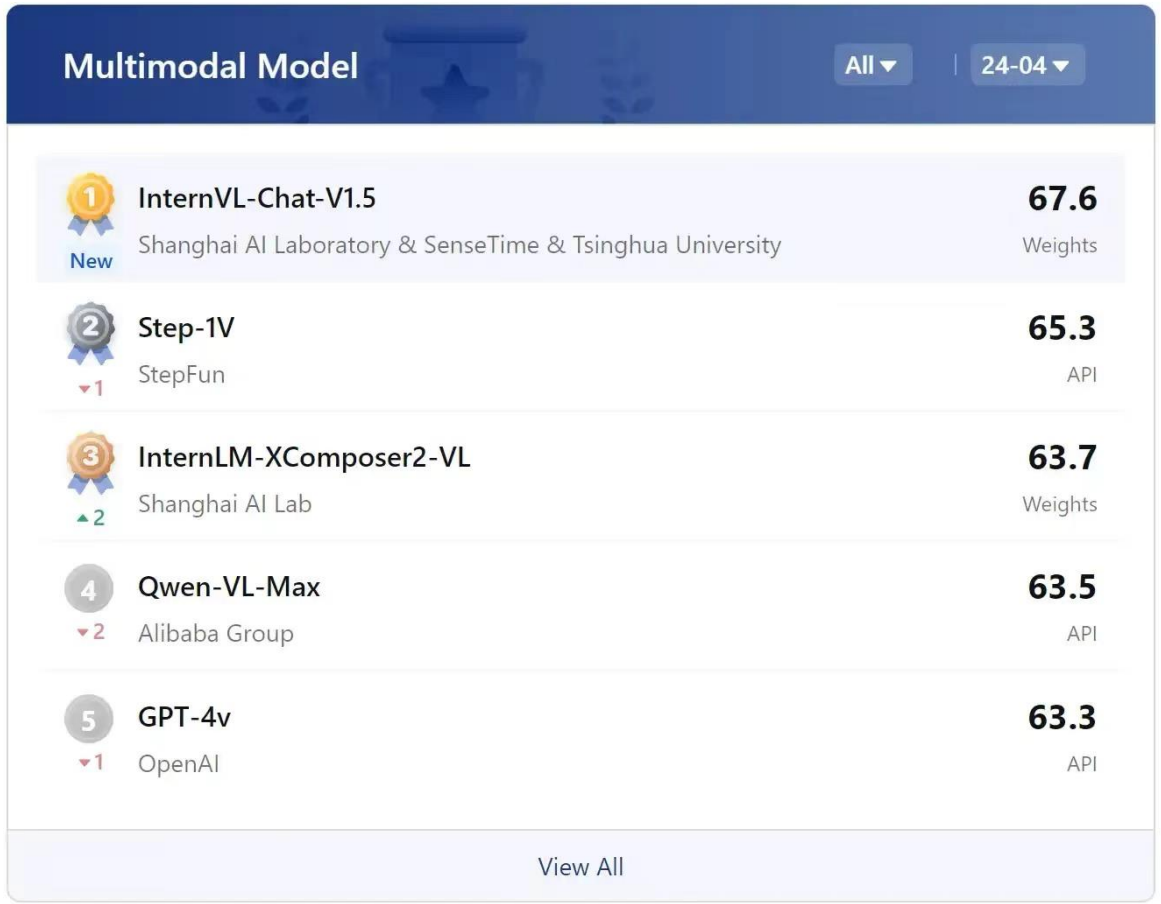


总下载量排列前10



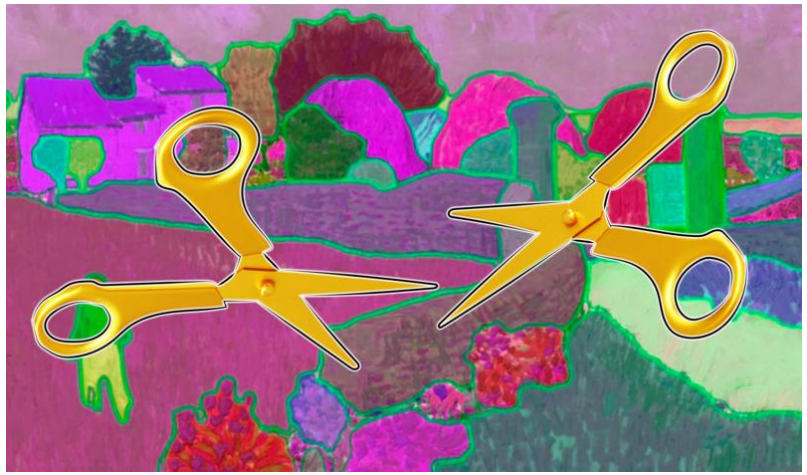
InternVL：最强开源视觉基础模型

在OpenCompass评测指标集合（几十个图文多模态评测数据集性能的均分）上名列开源、闭源所有多模态大模型第一



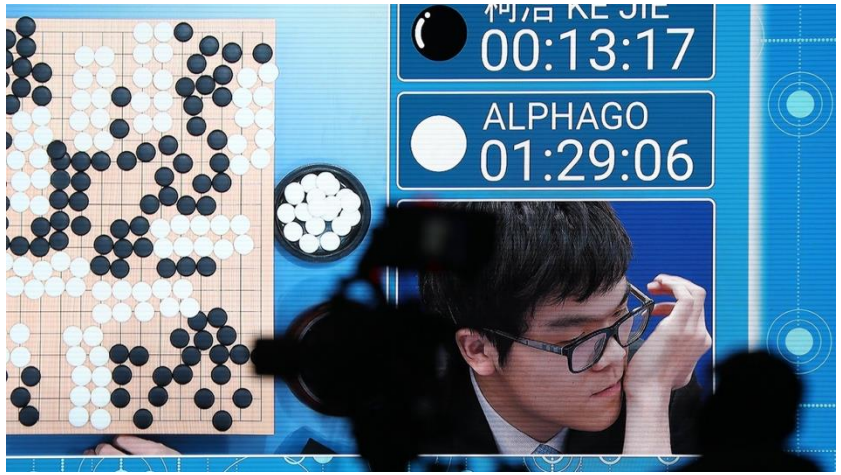
如何拓展LLM通用能力到更多应用上？

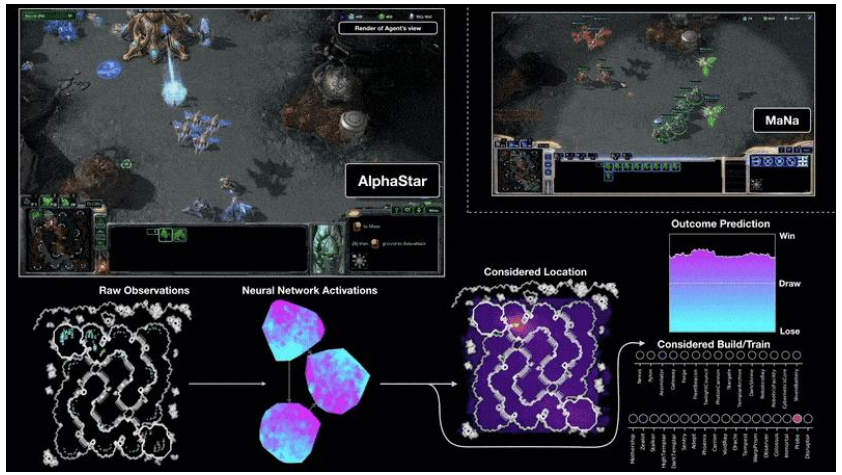
研究目标： 多模态模型，为大语言模型装上手脚和眼睛，与现实世界交互



虚拟世界中的智能体

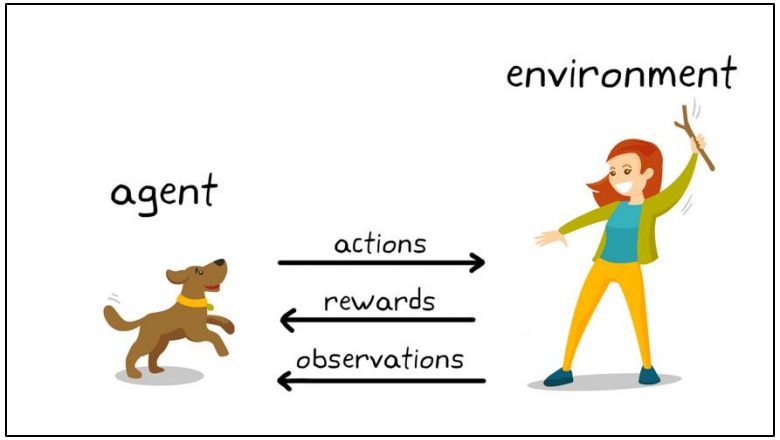
- 以前基于强化学习（RL）的技术在封闭环境下已经取得很大成功
- 往开放世界智能体推进，但是遇到巨大的泛化性挑战

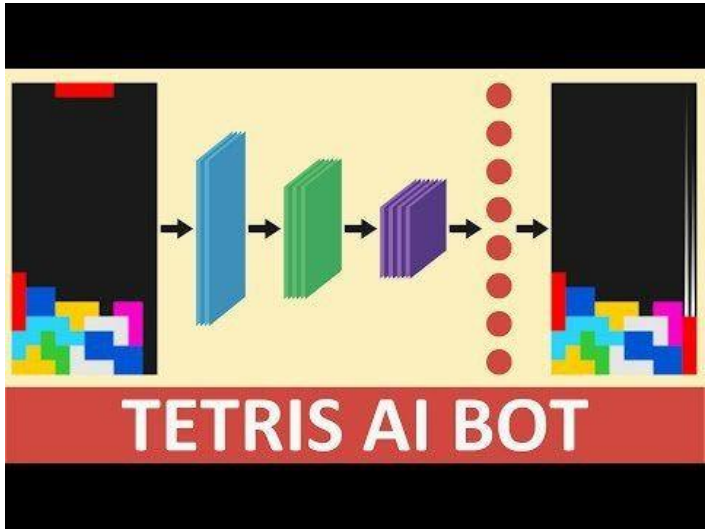




虚拟世界中的智能体

- 以前基于强化学习（RL）的技术在封闭环境下已经取得很大成功
- 往开放世界智能体推进，但是遇到巨大的泛化性挑战





GITM: 开放世界中的通才智能体

Minecraft：支持开放世界智能体的游戏平台

- 开放式世界，模拟人类世界中的探索、采集、制造和建造
- 涵盖复杂长时任务、环境干扰和不确定性，与真实世界相近
- 开发和测评通才智能体的极佳平台：在游戏中模拟真实世界的场景和能力

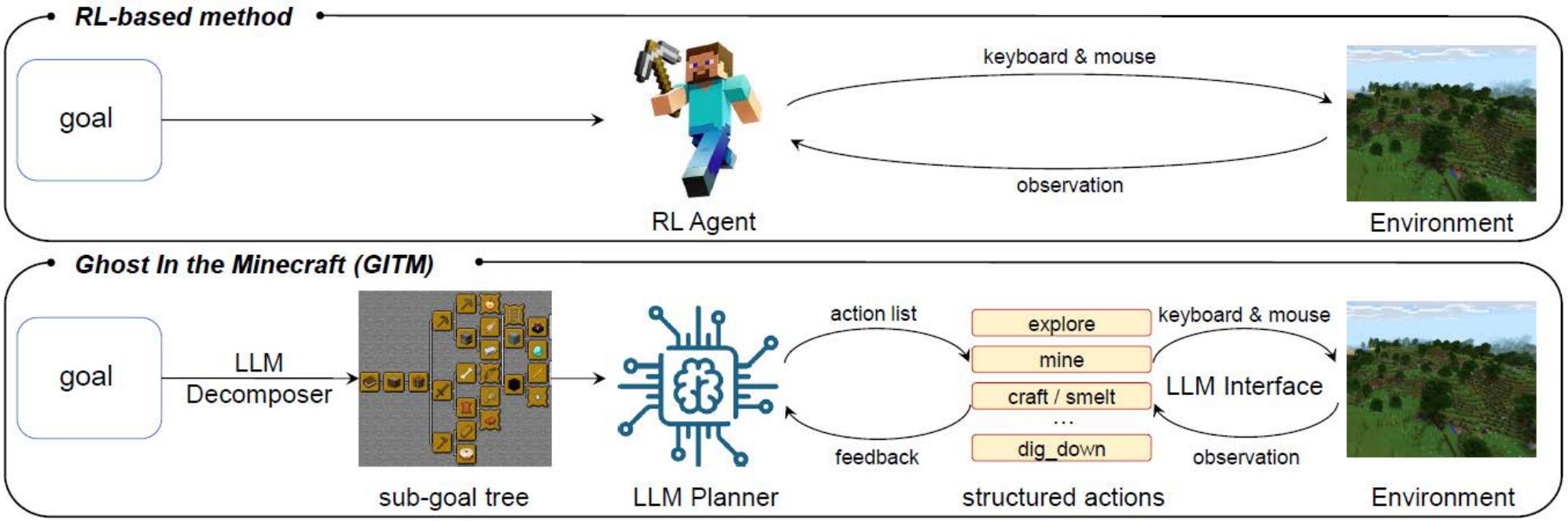
MineRL钻石挑战：从游戏开始到获取钻石，涉及10余个关键步骤

MineDojo：3000+游戏任务测评集

Zhu et al., Ghost in the Minecraft: Generally Capable Agents for Open-World Enviroments via Large Language Models with Text-based Knowledge and Memory. arXiv preprint, 2023.

GITM: 开放世界中的通才智能体

从RL智能体到以LLM为基础的智能体



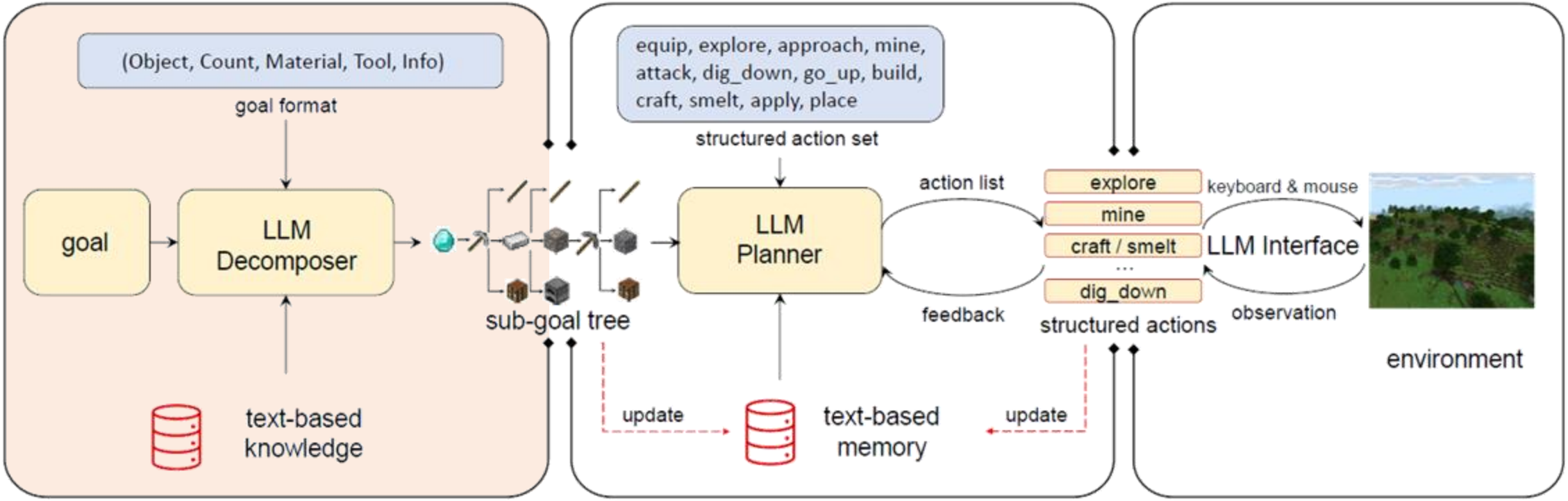
RL智能体：

- 关注短时操控，难以应对长程任务
- 训练所需的步数极大，开销巨大
- 能力局限于特定任务，难以迁移

LLM智能体：

- 将长程复杂任务分解为简单任务
- 以文本形式学习知识，效率高
- 结构化的动作接口在各任务中通用

GITM: 开放世界中的通才智能体

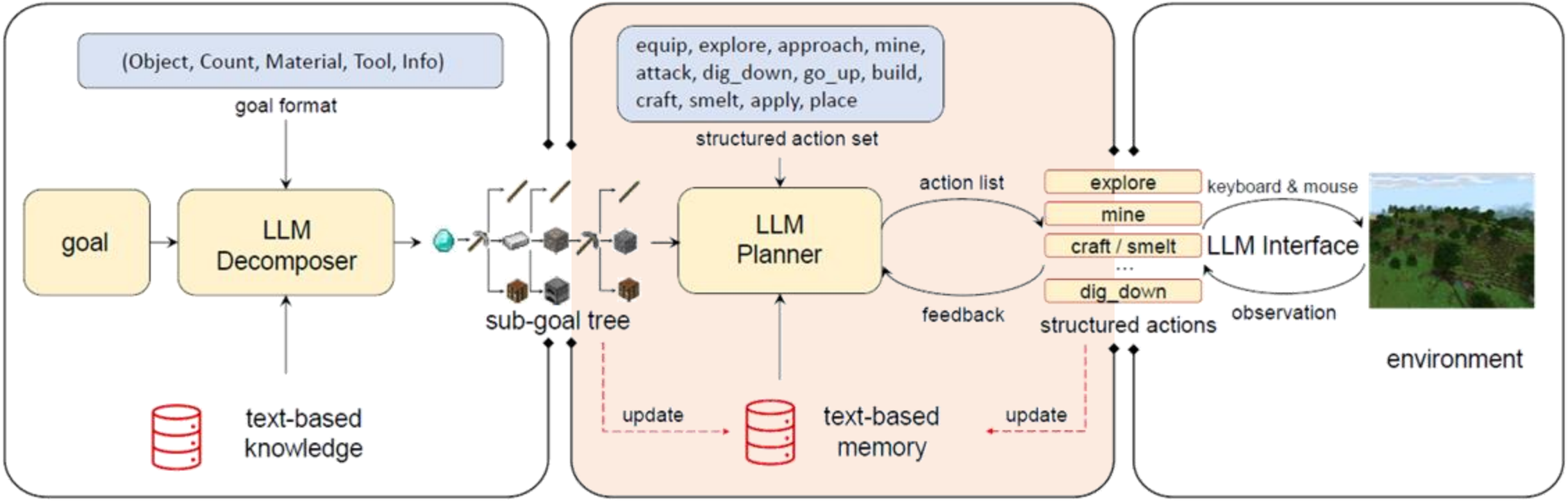


RL缺陷#1： 关注短时操控，难以应对长程任务

解决：外部知识库引导，LLM自动将复杂任务分解为简单任务

- 将所有的目标定义为 **(Object, Count, Material, Tool, Info)** 的统一形式
- **LLM Decomposer** 递归分解目标，直到每个目标能直接实现，无需前置条件
- 外部知识库辅助分解，如物品获得路径，并将相关知识写入到目标的Info中

GITM: 开放世界中的通才智能体

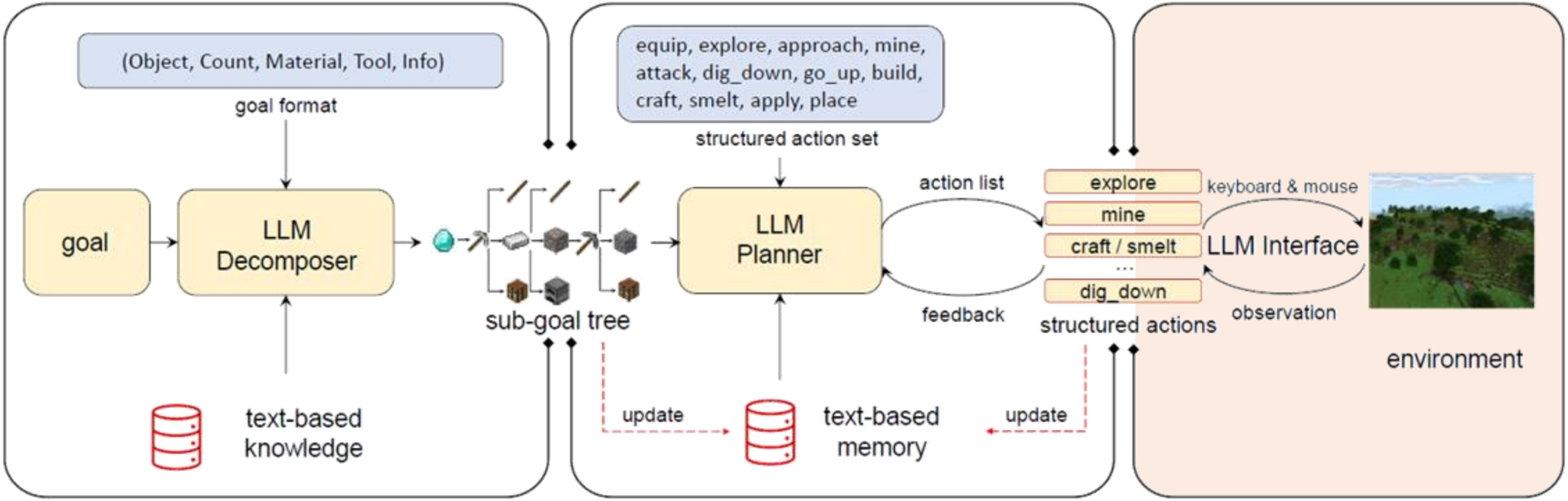


RL缺陷#2： 通过更新模型参数学习，需要很多步数，开销巨大

解决：LLM以文本形式获取、储存、总结、利用知识

- 在闭环交互中获取经验（e.g., 如何用给定动作控制agent）有助于应对之后的任务
- 文本形式的**memory机制**：每个目标记录**通用的reference**，可直接用于相似目标
- 高效学习：仅需几局游戏的成功经验，使用**LLM提取关键动作**作为reference

GITM: 开放世界中的通才智能体



RL缺陷#3： 针对特定任务进行训练，能力难以拓展和迁移

解决：为LLM配上通用API：结构化的动作接口

- 在底层操作上封装一层**带语义的抽象动作接口**，并使用统一形式的**反馈信息**
- LLM辅助设计接口：LLM将3000+任务拆解为动作序列，从回答中提取通用动作
- 动作接口与动作实现独立，rule-base或RL均可用于动作实现

GITM: 开放世界中的通才智能体

能力1：任务成功率和学习效率大幅超越RL

Table 2: **Comparison of our GITM with previous methods on `ObtainDiamond` challenge.**

Method	Success Rate (%)				
					
DreamerV3	-	50.0	3.0	0.01	0.01
DEPS	90.0	80.0	73.3	10.0	0.6
VPT	100.0	100.0	100.0	85.0	20.0
Our GITM	**100.0**	**100.0**	**100.0**	**95.0**	**67.5**

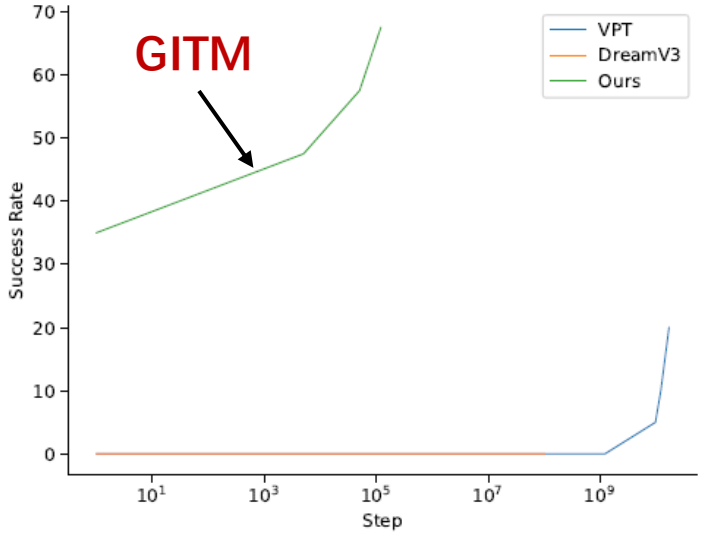


Figure 6: **Comparison of learning efficiency.**

钻石挑战任务，GITM vs. RL:

- 获取钻石的成功率大幅提升：从**20%**提升到**67.5%**
- RL模型的学习步长：**1e8**甚至**1e10**以上，GITM学习步长只有约**1e5**
  GITM以文本形式记录和使用成功经验，学习一个子任务只需几局游戏

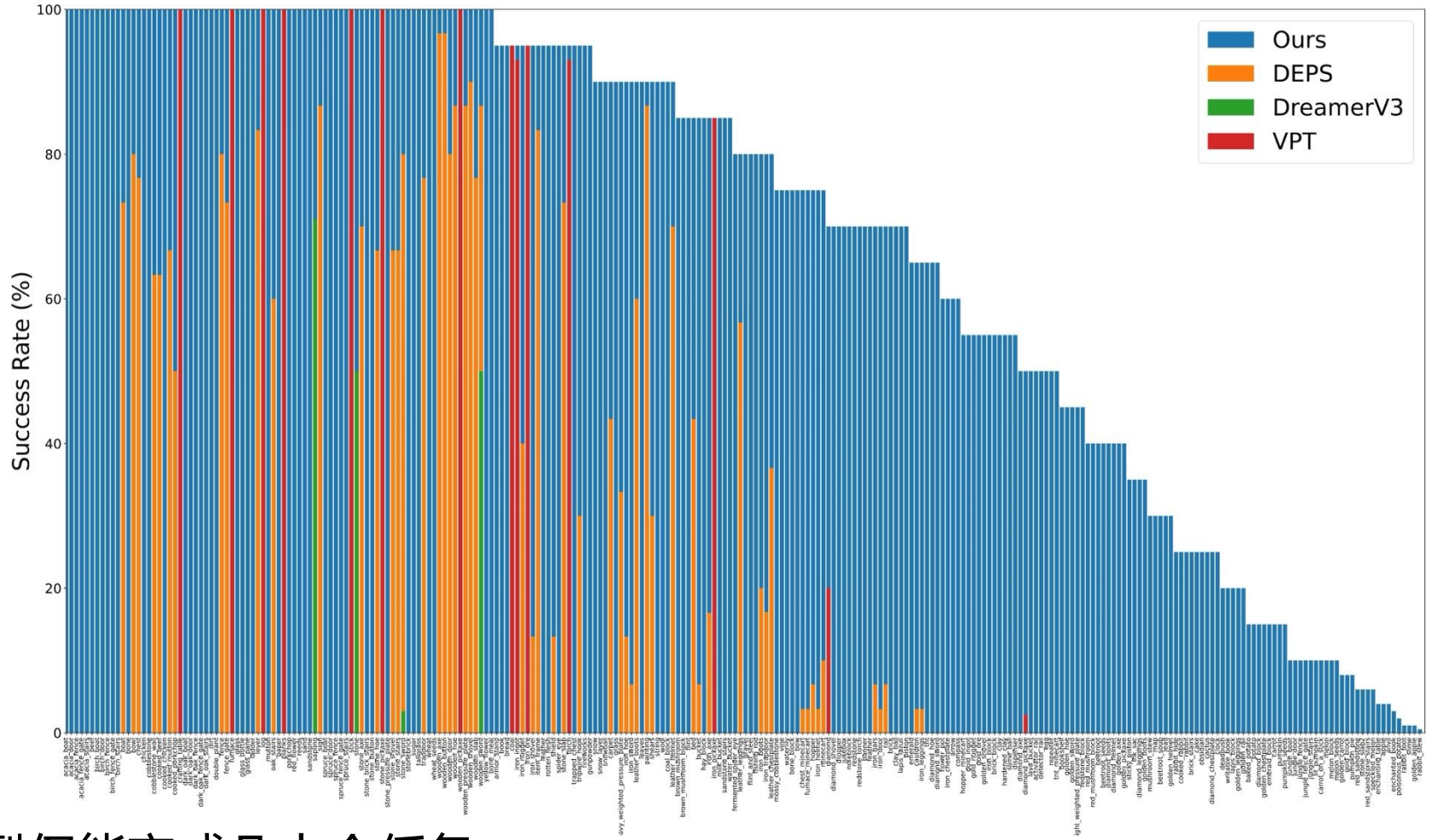
GITM: 开放世界中的通才智能体

能力2：获取Minecraft世界所有物品的通才能力

- **红色**：已有方法能获取到的物品 **vs.** **绿色**：GITM能获取到的物品
- 每个RL模型只能针对一类任务训练特定领域的能力，难以泛化到其他任务
- 利用通用的结构化动作和复杂任务分解，GITM一个模型就能获取所有物品，实现真正的通才，类似于真实世界中掌握跨领域的能力

GITM: 开放世界中的通才智能体

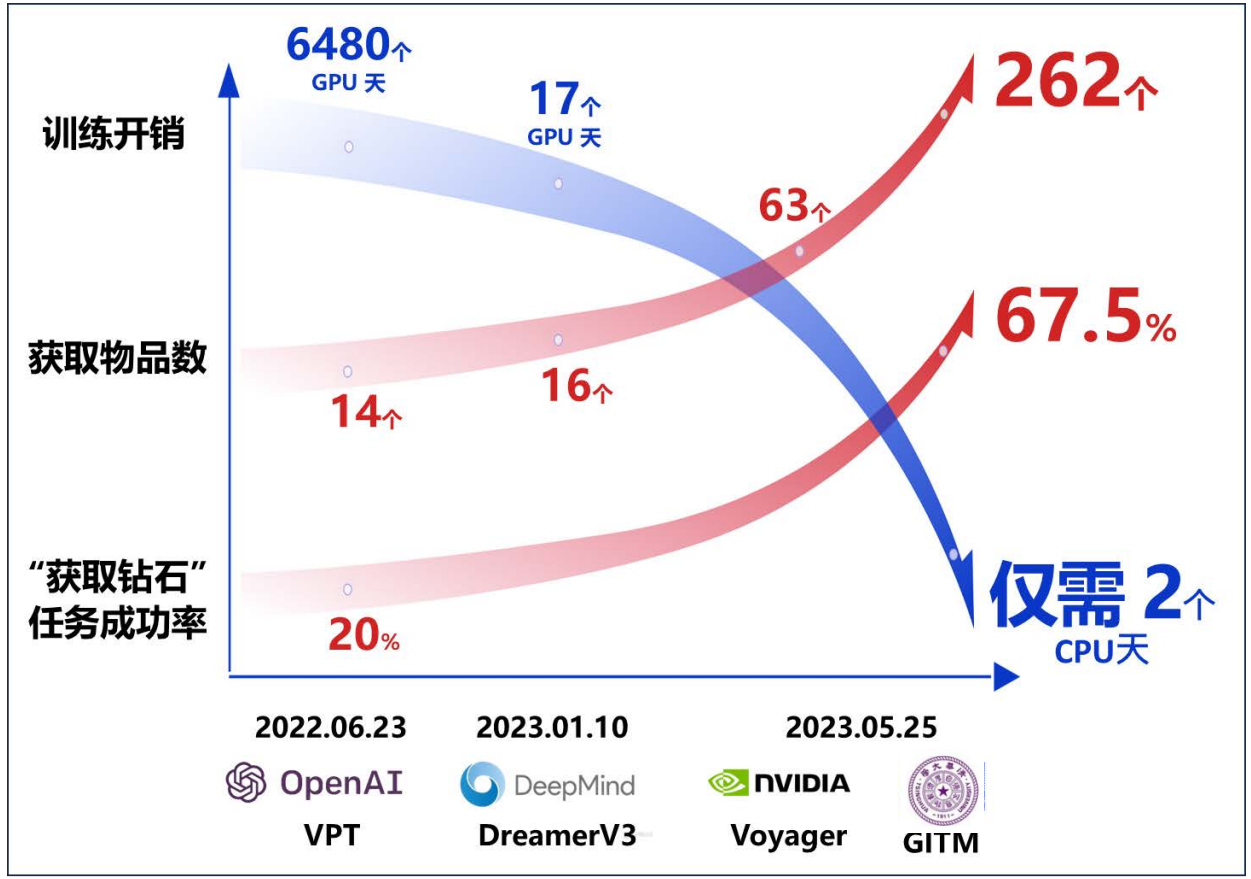
获取所有物品的成功率



- RL模型仅能完成几十个任务
- GITM（**蓝色**）能完成200+任务，在RL能完成的任务上成功率持平或超过RL

虚拟世界中的智能体基石模型

通才智能体 Ghost in the Minecraft (GITM) 完全解锁《我的世界》



Zhu, Xizhou, et al. "Ghost in the Minecraft: **Generally Capable Agents for Open-World Enviroments via Large Language Models with Text-based Knowledge and Memory**." *arXiv* (2023).

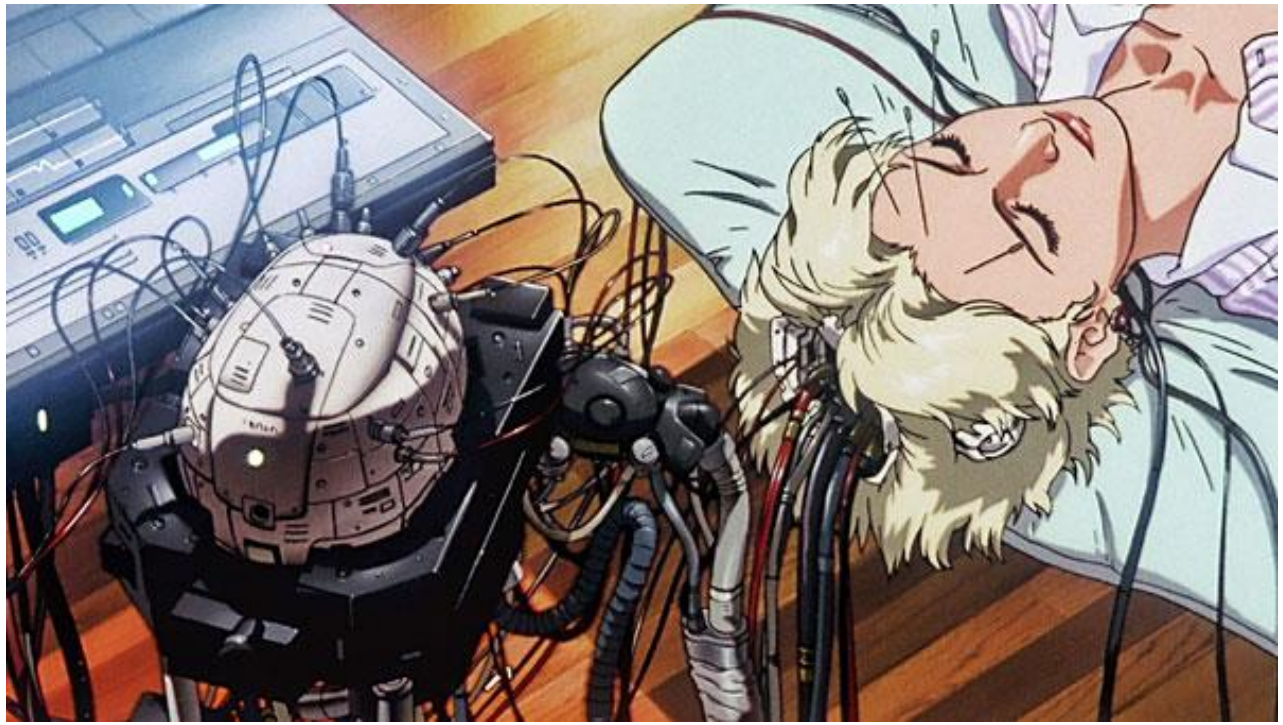
GITM: 开放世界中的通才智能体

Demo: 获取钻石

GITM: Ghost in the Minecraft

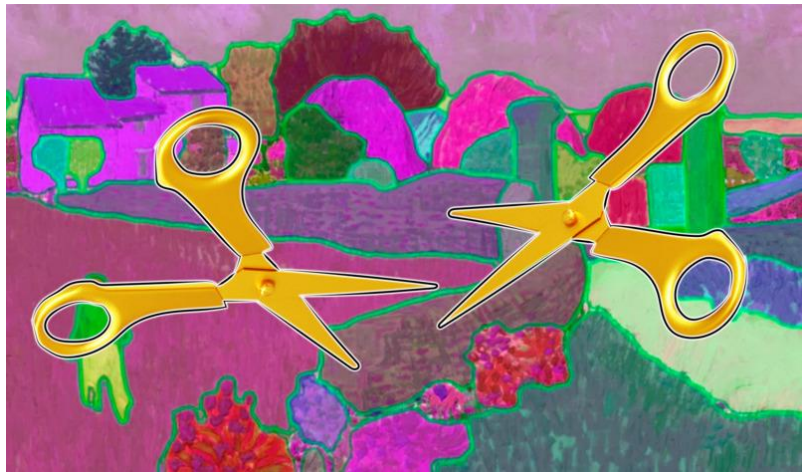
"What if a cyber brain could possibly generate its own ghost, create a soul all by itself? And if it did, just what would be the importance of being human then?"

— Ghost in the Shell (1995)



如何拓展LLM通用能力到更多应用上？

研究目标： 多模态模型，为大语言模型装上手脚和眼睛，与现实世界交互

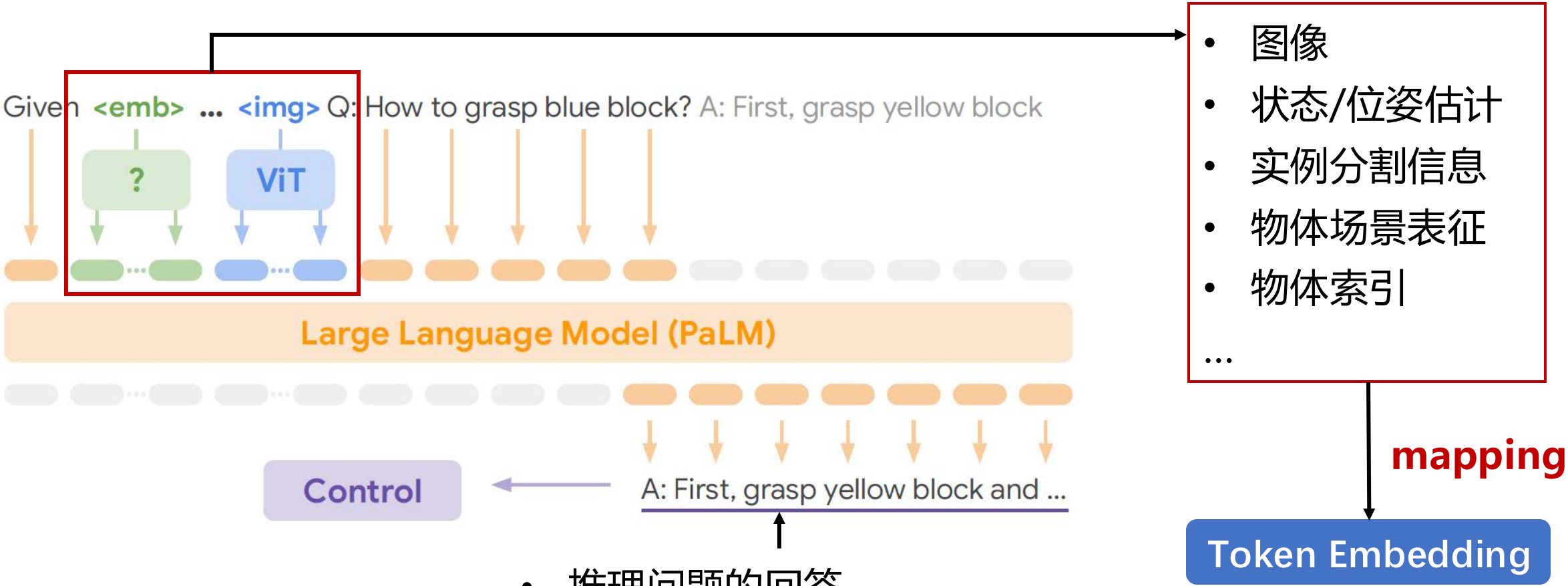


机器人智能体

- 机器人硬件、控制算法研究：
- 高阶智能算法研究：基于强化学习（RL）算法，泛化性差

PaLM-E: Google的具身多模态LLM

PaLM-E: 将多模态信息以token embedding形式注入LLM(谷歌的工作)



- 图像
- 状态/位姿估计
- 实例分割信息
- 物体场景表征
- 物体索引

...

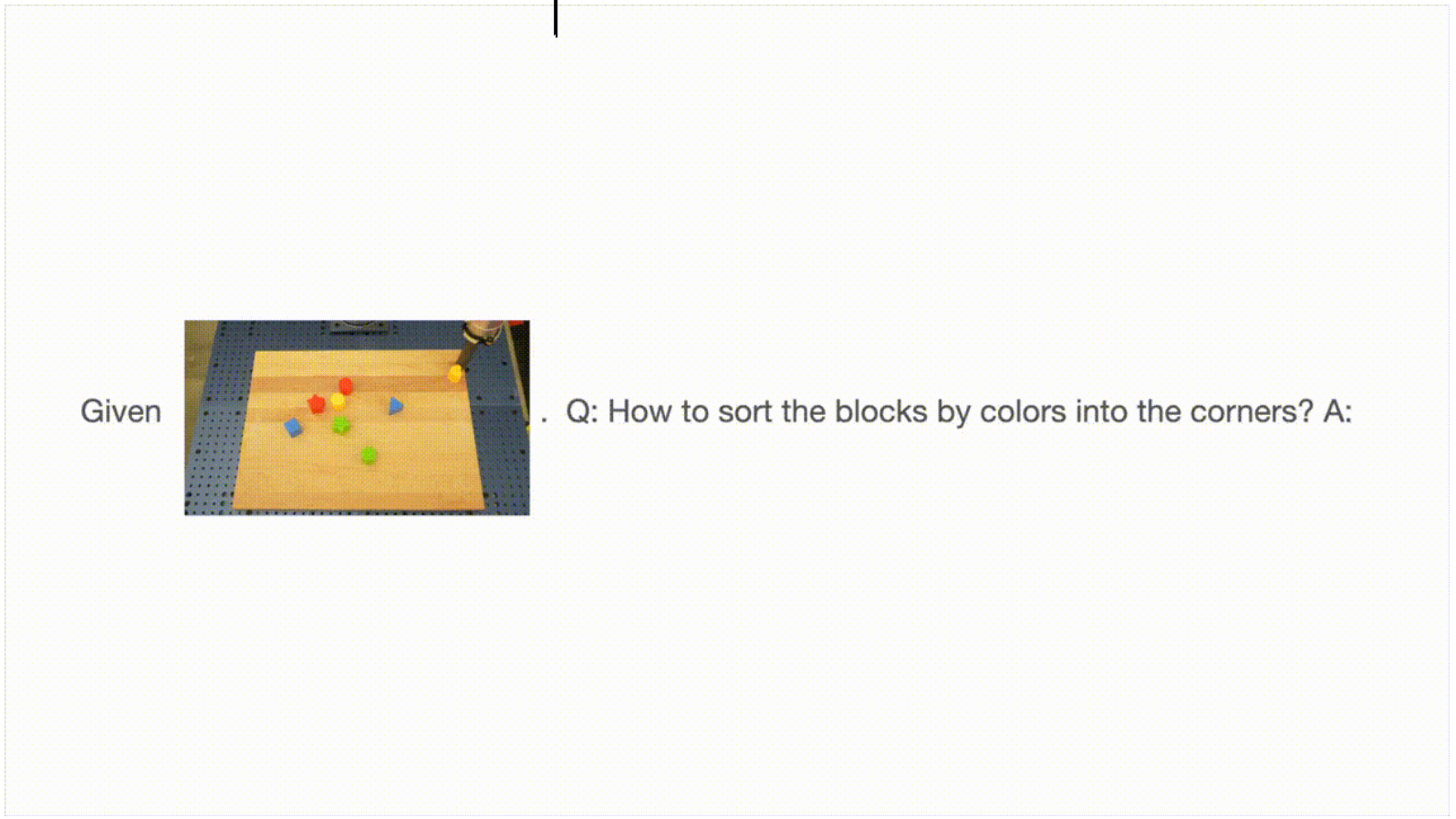
- 推理问题的回答
- 交由控制器执行的文本指令

Driess et al., Palm-e: An embodied multimodal language model. arXiv preprint arXiv:2303.03378

PaLM-E: Google的具身多模态LLM

演示：操控机械臂完成文本指令的任务

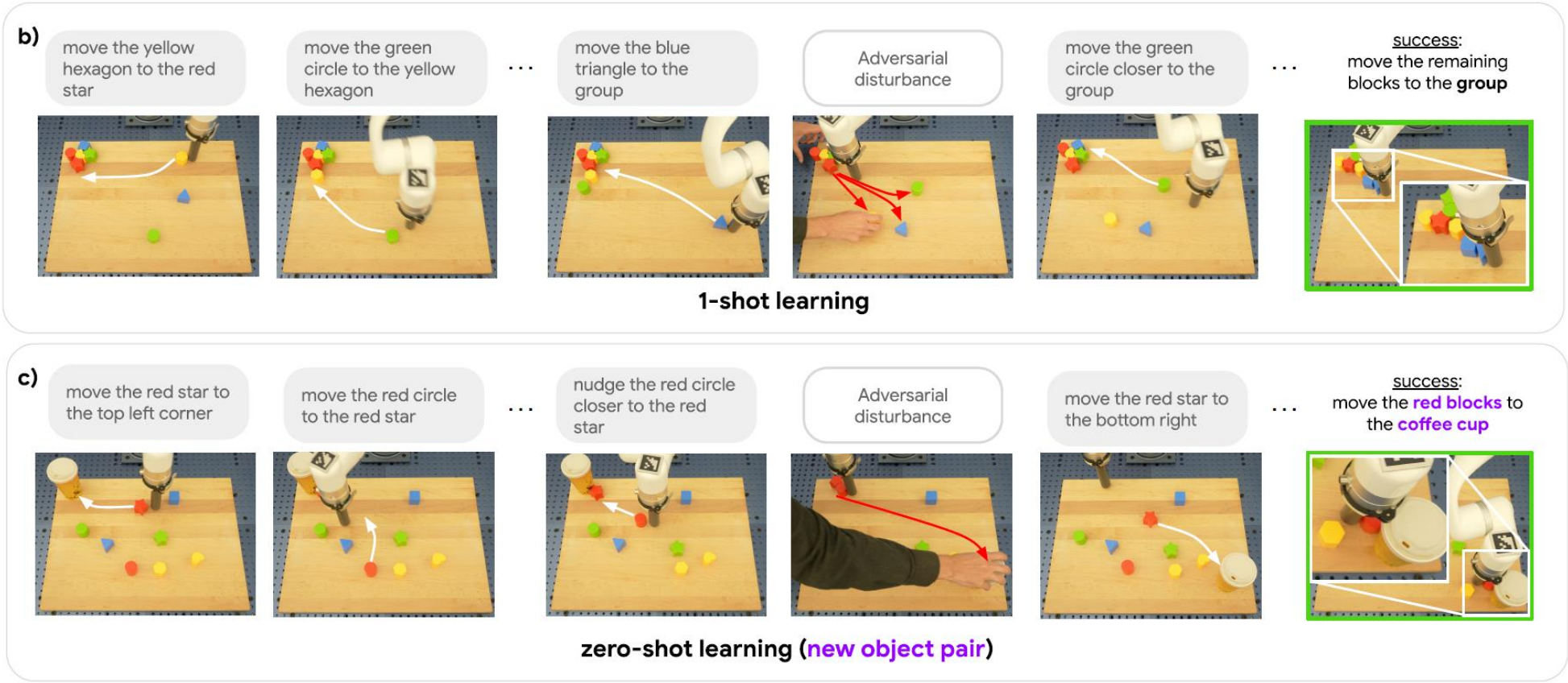
LLM输出给控制器的文本指令



任务：将物块按颜色区分放到角落

PaLM-E: Google的具身多模态LLM

在机器人任务中发挥LLM的通用能力



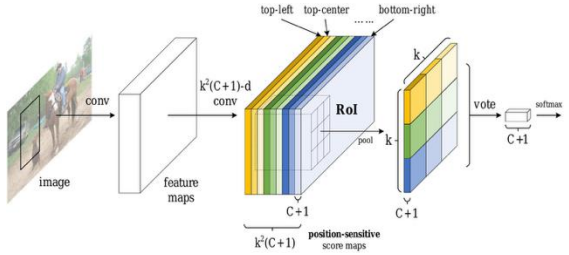
- 完成长程复杂的操控任务
- Few-shot / zero-shot 的泛化能力
- 应对外部干扰的鲁棒性

通用智能体模型——无限想象空间

课题组介绍—扎实的视觉感知、网络结构设计积累

国际专业机构遴选出在过去8年CVPR/ICCV/NeurIPS/ICLR等顶会年度十佳影响力论文中，**本研究团队核心成员有9篇论文入选，研究内容集中于视觉核心任务**

R-FCN
(NeurIPS'16, 7396次引用)



VL-BERT
(ICLR'20, 1674次引用)

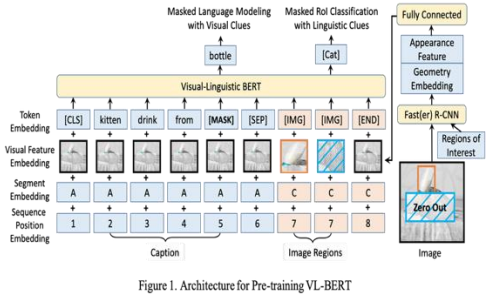
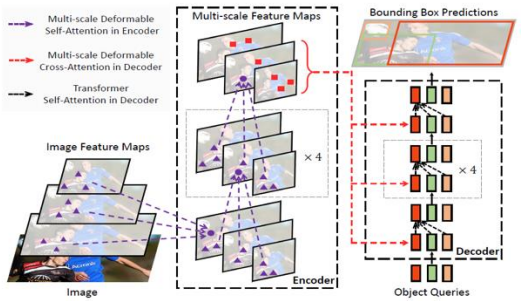
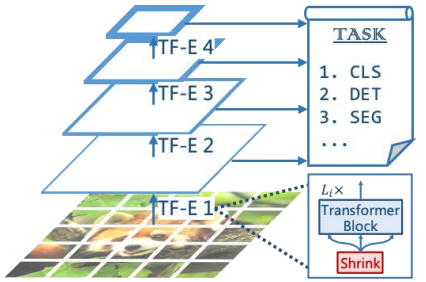


Figure 1. Architecture for Pre-training VL-BERT

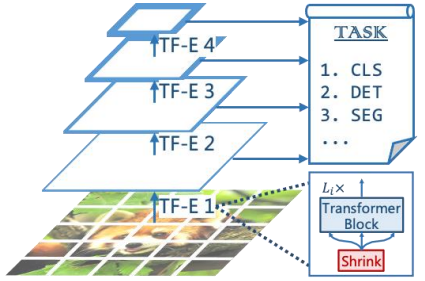
Deformable DETR
(ICLR'21, 4127次引用)



PVT
(ICCV'21, 3279次引用)

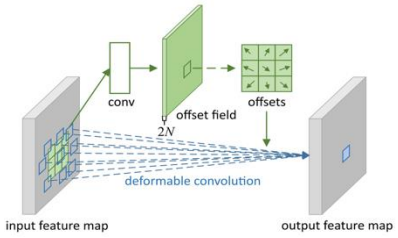


BEVFormer
(ECCV'22, 686次引用)

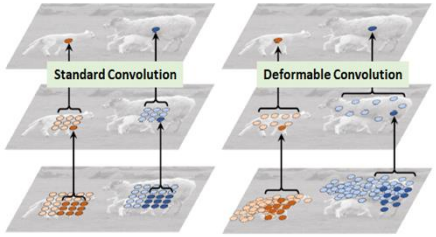


2016 — 2023

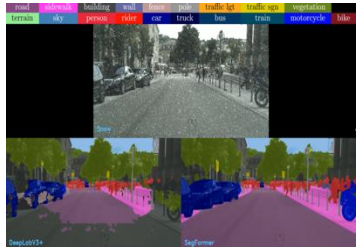
DCN
(ICCV'17, 5803次引用)



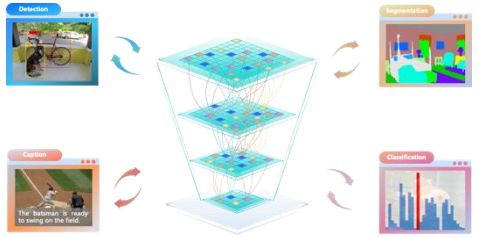
DCNv2
(CVPR'19, 1977次引用)



SegFormer
(NeurIPS'21, 3254次引用)



InternImage
(CVPR'23, 359次引用)



谢 谢!

代季峰• 电子工程系
daijifeng@tsinghua.edu.cn