《多模态大模型原理与应用》

Lecture 8

多模态模型走向真实世界-从VQA到SAM,再到PaLM-E

刘阳

中山大学

人机物智能融合实验室（HCP Lab）

liuy856@mail.sysu.edu.cn

目标

本节课要解决三个核心问题

01

理解任务

VQA解决什么问题?

视觉问答的任务本质、典型问题类型、数据集演变和核心挑战

02

理解模型

SAM为什么重要?

从注意力到记忆到模块化，理解VQA模型的演进和分割模型的突破

03

理解趋势

PaLM-E代表什么方向?

从"会回答"到"会行动"，具身多模态的发展趋势

为什么多模态重要

自动驾驶：视觉 + 雷达 + 地图的多模态融合

真实世界的信息来自多个模态

图像、文本、视频和环境状态共同构成了我们感知的世界

多模态模型是连接这些信息的重要技术路线

- 自动驾驶：摄像头 + 激光雷达 + GPS地图
- 智能助手：语音 + 图像 + 文本理解

核心观点：多模态学习让AI更接近真实世界的复杂性。单一模态的模型如同"盲人摸象"，多模态模型才能形成对世界的完整理解。

课程主线

从"问答理解"到"像素定位"再到"行动决策"的三段式演进

VQA

"图里是什么"

问答理解层

SAM

"目标在哪里"

像素定位层

PaLM-E

"看见之后怎么做"

行动决策层

核心逻辑：多模态模型逐步从描述世界走向定位世界再走向参与世界

本节内容

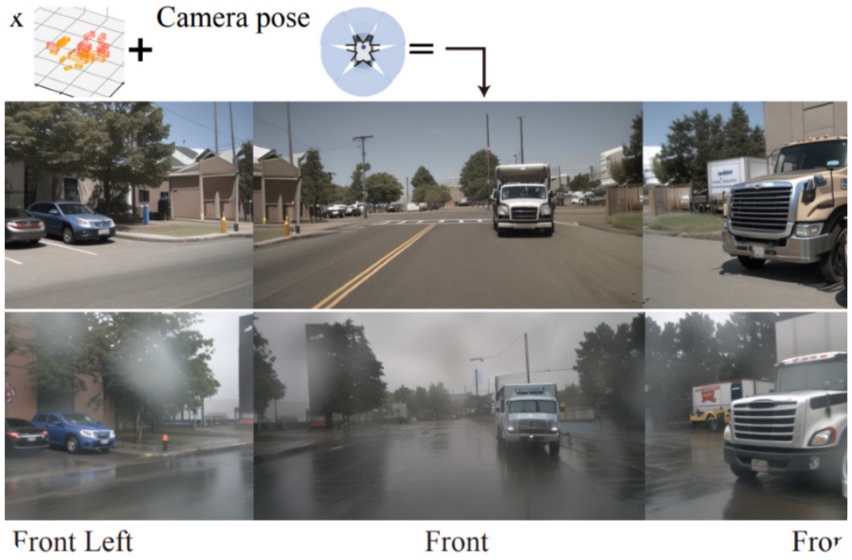
CONTENTS

一、VQA(视觉问答)

二、SAM（通用图像分割模型)

三、PaLM-E(具身多模态语言模型)

问题



一张复杂街景，可以引出三个层次的问题

问题层："图中有几辆车？"

→ 对应VQA能力

定位层："请标出所有行人"

→ 对应SAM能力

行动层："如果要过马路该注意什么？"

→ 对应PaLM-E能力

核心观点：面对同一张复杂图像，模型需要具备问答、分割和行动决策三个层次的能力

第三章与第四章的关系

第三章：多模态基础模型

- 理解模型原理与架构
- 多模态表示学习方法
- 图文对齐与融合机制
- 基础模型的设计思想

+

第四章Part 1：VQA等应用任务

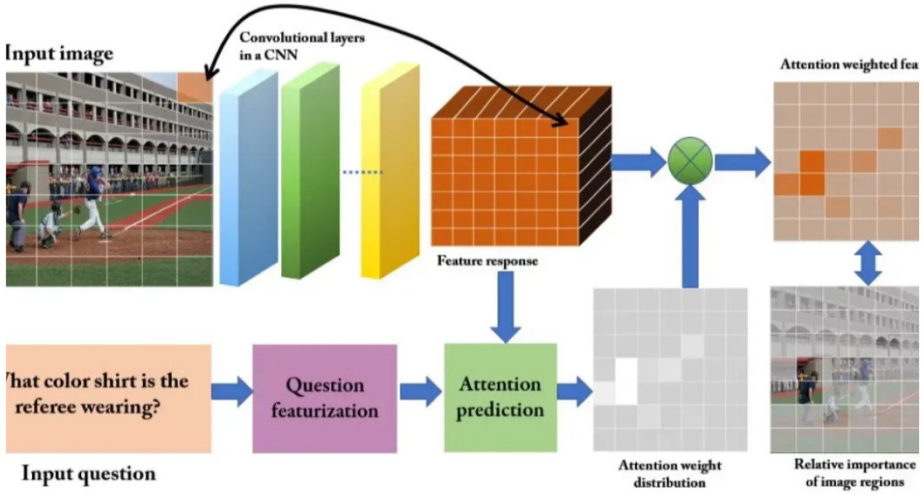
- 视觉问答任务定义
- 数据集与评价方法
- 典型模型与算法
- 应用场景与挑战

=

完整知识链条：先懂模型（第三章）→ 再懂任务（第四章）→ 最后懂趋势（前沿进展）

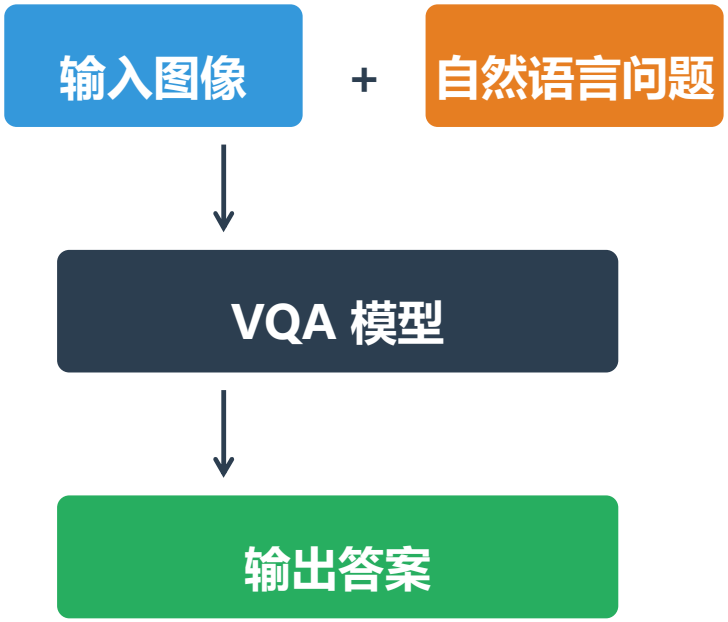
帮助建立从理论到应用的完整认知

VQA是什么



VQA基本流程：图像 + 问题 → 模型 → 答案

视觉问答任务定义



VQA为什么重要

VQA不是"简单问答"，而是多模态综合能力的试金石

视觉识别

看清图中有什么——目标检测与识别能力

语言理解

读懂问题在问什么——自然语言理解能力

多模态对齐

将图像区域与问题关键词关联——跨模态映射能力

推理能力

结合常识进行逻辑推断——认知推理能力

因此，VQA同时考验四种核心能力，是多模态理解的经典入门任务。一个VQA系统的表现，往往反映了其对视觉、语言和跨模态推理的综合理解水平。

VQA的典型问题类型

1. 目标识别

"图中是什么动物？"
→ "狗"

2. 属性判断

"这只狗是什么颜色？"
→ "棕色"

3. 空间关系

"球在桌子左边还是右边？"
→ "左边"

4. 计数

"图中有几辆车？"
→ "3辆"

5. 常识推理

"这个人为什么打伞？"
→ "因为下雨"（需要结合外部常识，图像本身可能没有雨的直接信息）

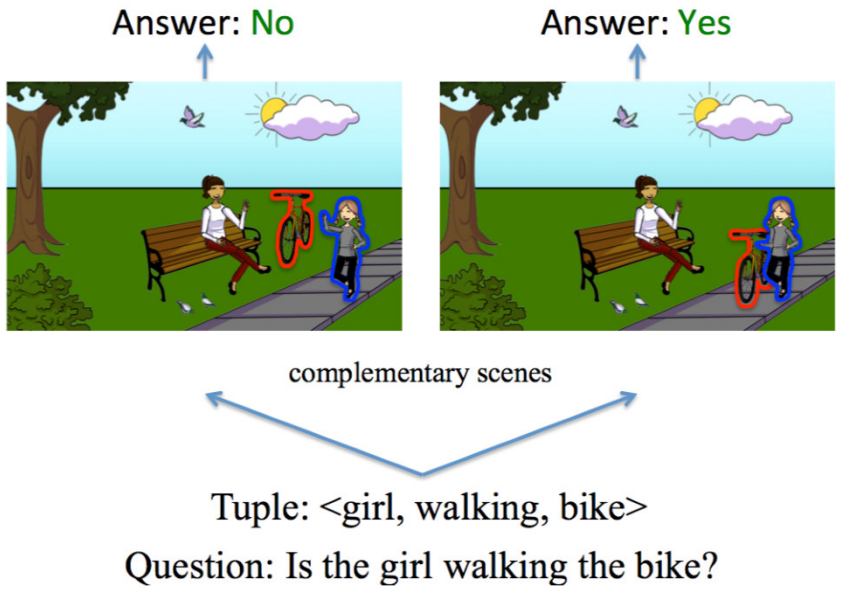
难度递进：从简单的目标识别到需要外部知识的常识推理，VQA问题的复杂度逐步提升

VQA数据集全景

数据集	年份	特点	贡献
COCO-QA	2015	基于COCO图像的初代数据集	开创了视觉问答任务范式
VQA-v1	2015	大规模开放域问答	建立了VQA的标准任务格式
VQA-v2	2017	平衡配对机制减少语言先验	最重要的基准数据集之一
GQA	2019	组合式推理问答	推动结构化推理研究
OK-VQA	2019	需要外部常识知识	超越图像内容的知识推理
Text-VQA	2019	图像中文字理解	OCR与推理的结合

演进趋势：从基础识别 → 复杂推理 → 外部知识 → 文字理解，VQA数据集推动了任务的不断进化

VQA-v2的地位



为什么VQA-v2常被引用

12万	COCO图像
150万	问题数量
1000万+	答案数量

VQA-v2的平衡配对设计：相似问题对应不同答案

关键改进：VQA-v2通过平衡配对机制（为每个问题配对两个相似图像但答案不同）减少语言先验偏见，迫使模型真正"看图"而非"猜答案"

OK-VQA与Text-VQA

VQA的两种重要扩展——说明VQA已不局限于普通图像识别

OK-VQA

Outside Knowledge VQA

核心特点：需要外部常识知识才能回答

示例："为什么这个标志是红色的？"

→ 需要知道红色通常表示"停止/警告"

意义：VQA已超越图像内容本身，进入知识推理领域

Text-VQA

图像中文字理解

核心特点：要求模型读取并理解图像中的文字

示例：路牌、菜单、广告牌中的文字

→ 将OCR（光学字符识别）与推理结合

意义：打通视觉与文字的联合理解通道

VQA为什么难

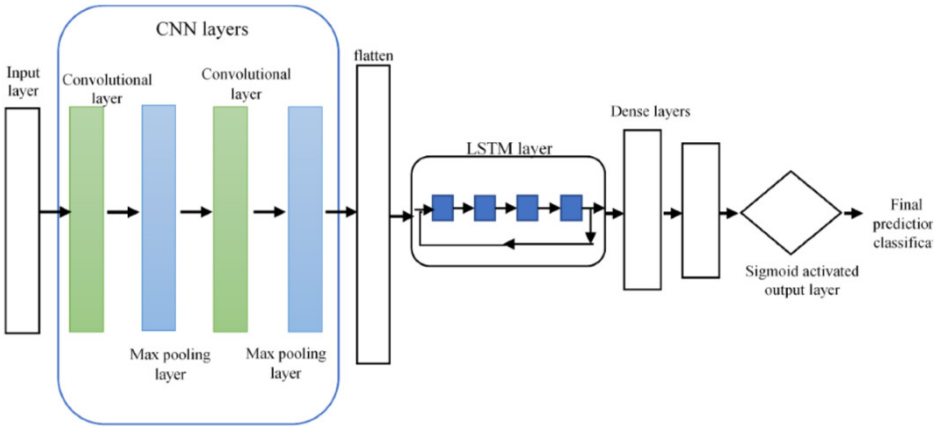
VQA不是单一步骤，而是"看图—读题—对齐—推理—作答"的完整链条

每个环节都可能出错，且存在依赖关系

- 看图出错：图像编码丢失关键细节（如小物体、遮挡）
- 读题出错：误解问题的真正意图（如把"不是"忽略）
- 对齐出错：图文特征没有正确关联（问题问的是A，模型看了B）
- 推理出错：缺乏必要的常识或逻辑能力

核心挑战：五个环节环环相扣，任何一个环节出错都会导致最终答案错误

早期VQA范式



CNN+LSTM 基本架构示意图

CNN + LSTM 时代

CNN 提取图像特征

VGGNet / ResNet

LSTM 编码问题

序列建模

↓

特征拼接 / 融合

↓

全连接层预测答案

这是VQA的"基线方法"，理解后续改进模型的基础。简单但有效，奠定了VQA的基本框架。

CNN+LSTM的局限

全局看图的问题

图像被压缩成固定长度的全局特征向量
很多决定答案的局部信息容易丢失
如同远远瞥一眼照片

局部找证据的需求

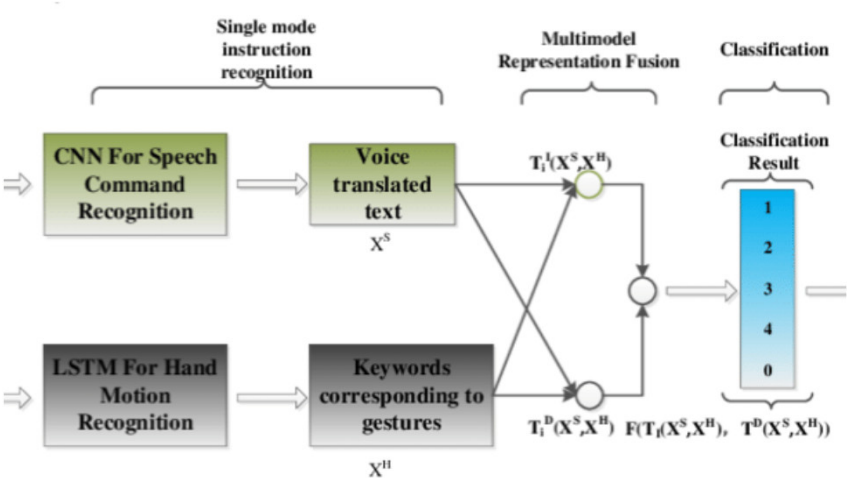
回答"杯子是什么颜色"时
需要聚焦杯子区域而非整幅图
如同用放大镜找关键证据

具体例子

问题："这只狗是什么颜色？"

- 全局特征：包含了狗、草地、天空、树木的所有信息
- 关键问题：模型如何从全局特征中"找到"狗对应的颜色信息？
- 答案：需要注意力机制来动态聚焦相关区域 → 引出下一节内容

多模态融合问题



多模态融合的基本框架

图像和问题怎样结合

核心难点

如何让图像特征和问题特征有效交互，而不是简单拼接

简单拼接的问题

- 两种特征只是物理上放在一起
- 没有真正"对话"和相互影响
- 表达能力有限

需要更精细的融合方法来建模图文之间的复杂关系

这为MCB（Multimodal Compact Bilinear Pooling）和MLB（Multimodal Low-rank Bilinear Pooling）等方法做了铺垫 → 通过高阶交互实现真正的图文"对话"

MCB与MLB

双线性融合思路——让图文特征真正相互影响

MCB

Multimodal Compact Bilinear Pooling

- 通过双线性池化实现高阶特征交互
- 捕捉图文特征间的细粒度关系
- 每个维度都能相互影响

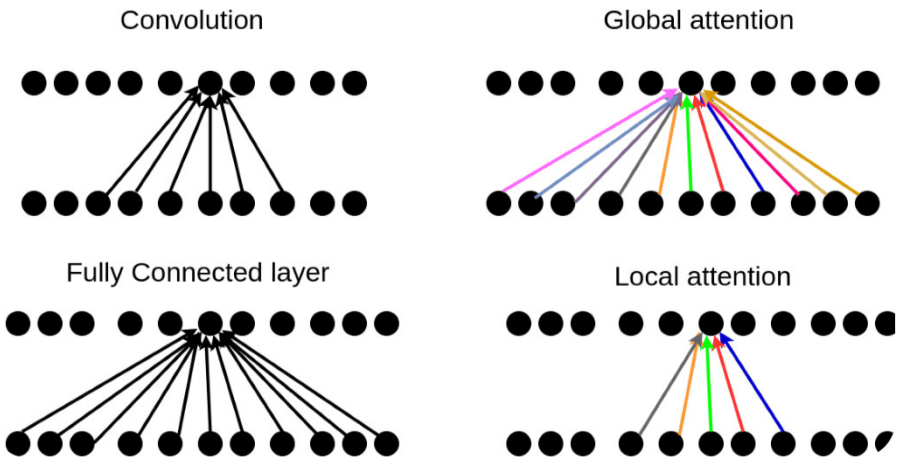
MLB

Multimodal Low-rank Bilinear Pooling

- 低秩近似版本
- 保持表达能力的同时降低计算开销
- 更适合大规模应用

核心思想：不是简单拼接，而是"高阶交互"——让图像特征和问题特征真正相互影响，产生比单独使用更丰富的表示

为什么需要注意力机制



从融合走向聚焦

核心洞察

问题不同，模型应该关注图像中不同区域

类比"聚光灯"

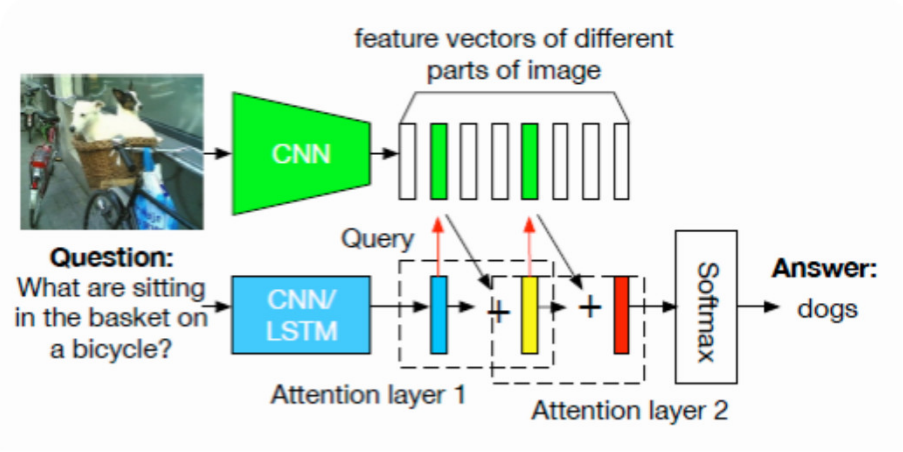
- **"狗是什么颜色" → 聚光灯照向狗**
- **"背景是什么" → 聚光灯照向背景**
- **"有几个人" → 聚光灯扫过所有人**

注意力机制：从全局连接到选择性聚焦

注意力机制让模型学会"看该看的地方"

根据问题的内容动态调整对图像不同区域的关注程度，成为VQA的重要改进方向

SAN简介



(a) Stacked Attention Network for Image QA

SAN架构：多轮注意力逐步聚焦

堆叠注意力网络

Stacked Attention Networks

核心思想：通过多轮注意力机制，让模型逐步聚焦到最相关的图像区域

关键："不是只看一眼图"

- **第一轮：找到大致相关区域**
- **第二轮：进一步缩小范围**
- **第三轮：精确定位答案证据**

类比：就像侦探破案——先看整个现场，然后逐步聚焦到关键线索区域，最终找到决定性证据

SAN怎么工作

SAN的逐步推理过程：从粗到精的三步聚焦

01 找到大致区域

问题："篮子里的动物是什么？"

- 第一轮注意力激活"动物"相关区域
- 定位到篮子附近
- 生成初步注意力图

→

02 进一步聚焦

在相关区域内继续筛选

- 第二轮注意力精确定位
- 确定是"狗"而非其他
- 更新注意力图

→

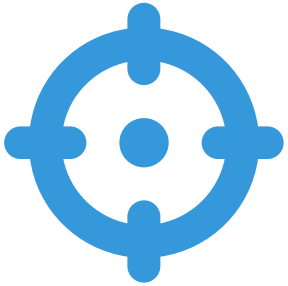
03 定位证据

精确定位答案

- 聚焦狗的毛色区域
- 生成最终注意力图
- 输出答案

核心机制：每一步都生成一张注意力图，逐步缩小关注范围，体现多步视觉推理思想

SAN的意义



SAN证明了"多次聚焦"比"一次全局看图"更有效

核心贡献

问题引导视觉重新分配注意力

问题的不同词语会激活图像的不同区域

动态映射

形成动态的"问题→视觉"映射关系

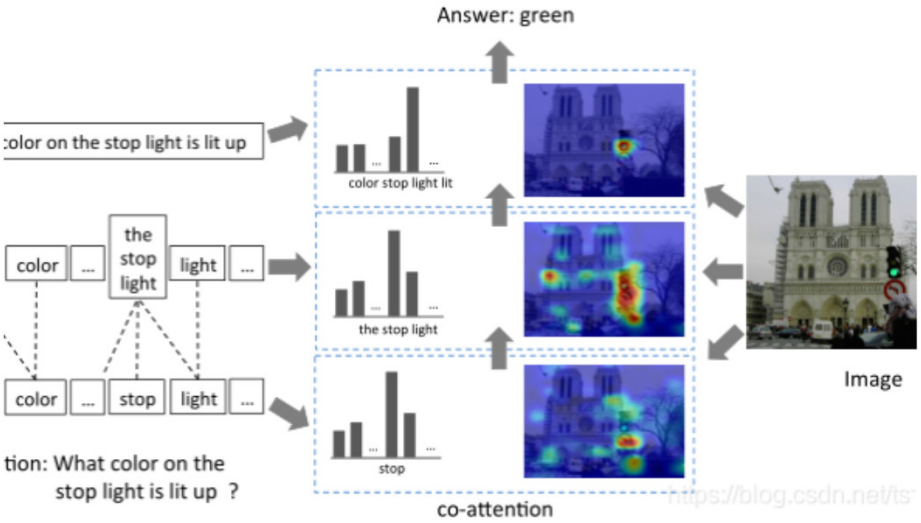
每个问题都有独特的注意力路径

奠定基础

为后续更复杂的注意力模型奠定了基础

开启了VQA的注意力时代

HieCoAtt简介



HieCoAtt：层次协同注意力机制

层次协同注意力

Hierarchical Co-Attention

核心思想：图像与问题之间的双向对齐

在三个层次上建模注意力：

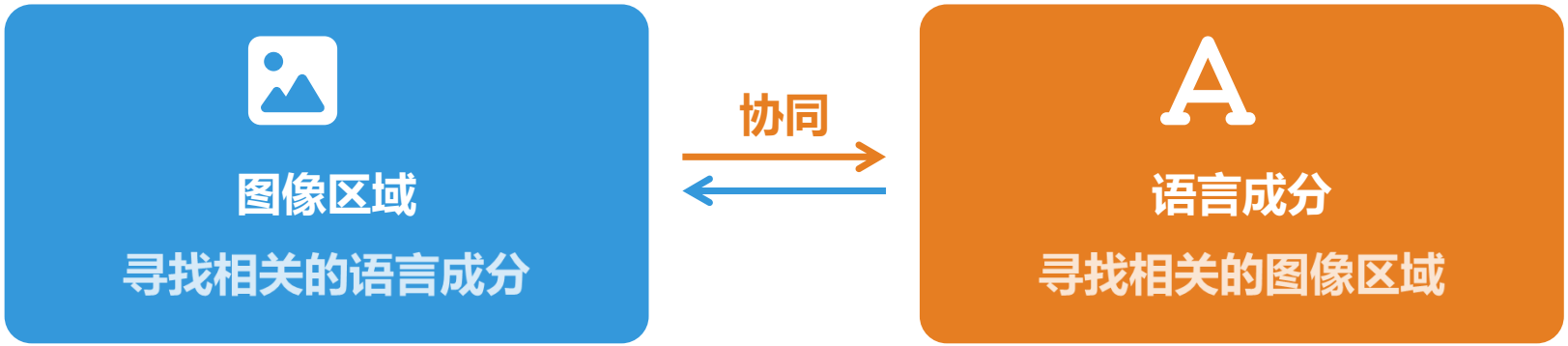
- 词（word）级别
- 短语（phrase）级别
- 句子（sentence）级别

核心："图像和问题相互看"

与SAN的区别：SAN是"问题引导看图像"（单向），HieCoAtt是"图像和问题相互看"（双向），在多个语言粒度上实现深度对齐

为什么叫协同注意力

图像与语言的双向交互



本质：不是图像单方面被查询，也不是文本单方面被编码

而是两个模态相互寻找对应关系

形成"你中有我，我中有你"的对齐——图像区域和语言成分相互寻找最佳匹配

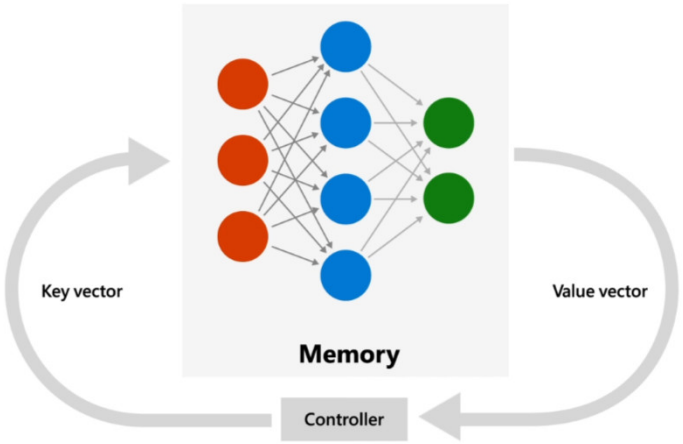
SAN与HieCoAtt对比

对比维度	SAN	HieCoAtt
核心策略	多次视觉聚焦（纵向深入）	图文协同对齐（横向交互）
注意力方向	问题 → 图像（单向）	图像 ↔ 问题（双向）
关注层次	空间区域的多轮筛选	词/短语/句子的多层次
核心优势	逐步精确定位答案区域	细粒度的图文双向理解

共同点：都体现了VQA从粗糙融合走向精细理解的趋势

差异：SAN重"深度聚焦"，HieCoAtt重"层次对齐"

记忆机制的引入



记忆网络：读写操作的外部记忆模块

为什么需要"记住中间信息"

核心动机

复杂问答往往不能一步完成

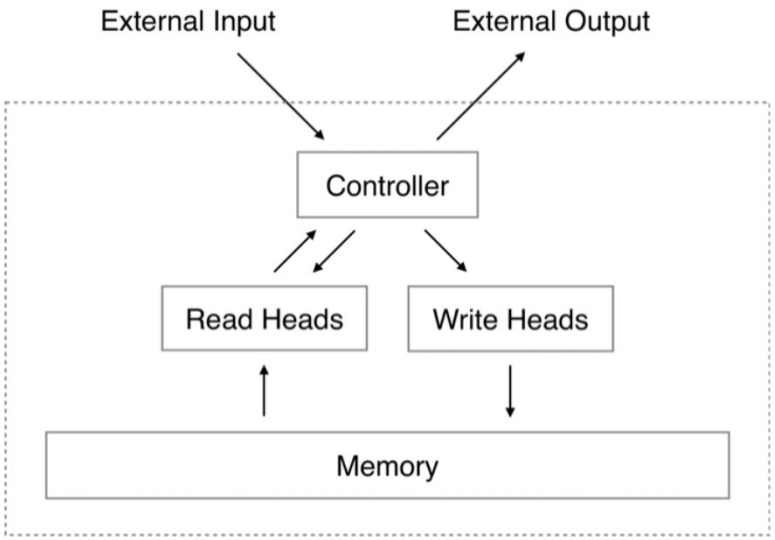
需要记忆机制保留中间推理结果

类比做数学题

- 多步计算需要草稿纸
- 记忆 = 模型的"草稿纸"
- 为后续步骤提供上下文

逻辑过渡：注意力机制解决"看哪里"的问题 → 记忆机制解决"记住什么"的问题 → 为DMN和MAN做铺垫

DMN



DMN架构：控制器 + 读写头 + 记忆模块

动态记忆网络

Dynamic Memory Network

核心机制：记忆更新机制

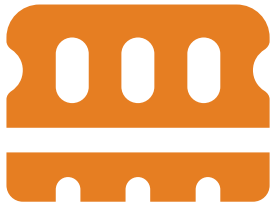
让模型在多轮推理中逐步整合问题和图像证据

关键特点：

- **"边推理边更新记忆"**
- **每轮读取当前记忆**
- **结合新证据更新记忆**
- **形成连续推理链条**

DMN的意义：引入了显式的记忆读写操作，让模型具备了"记住中间结果"的能力，支持需要多步推理的复杂问答

MAN



记忆增强网络

Memory Augmented Network

核心思想

延续记忆网络思路

进一步提升模型对复杂多步问答的处理能力

通过显式的记忆读写操作支持多步逻辑推断

形象比喻

"让模型更像做过程题

而不是选择题"

需要逐步推导，而非直接猜测

模块化推理思想

为什么要把问题拆开

类比：组装家具——先装腿、再装板、最后装门。每个步骤都有对应的工具和操作。复杂问题也需要分解为可组合的小模块来解决。

可解释性

知道哪步出错

每个模块的功能明确

便于调试和改进

可组合性

不同模块自由组合

解决各种复杂问题

像搭积木一样灵活

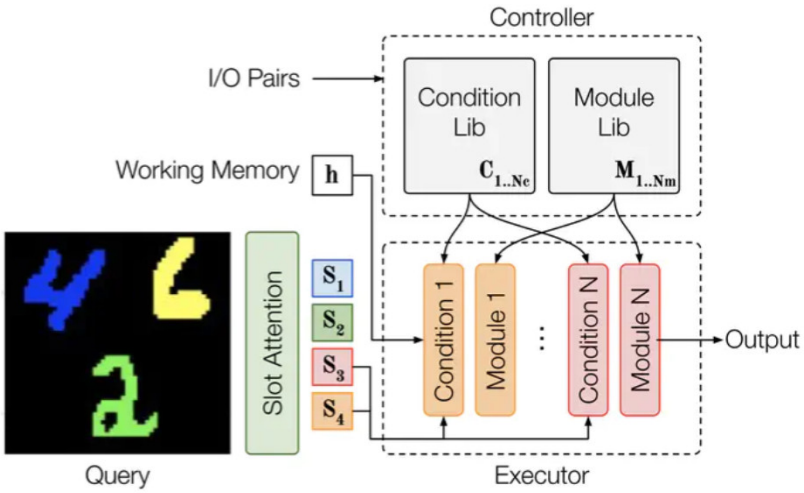
推理能力

组合解决复杂问题

从简单到复杂的推理

步步为营、层层递进

NMN



NMN架构：控制器 + 模块库 + 执行器

神经模块网络

Neural Module Network

核心思想：把复杂问题拆成可组合的小模块

典型模块：

- 找颜色模块
- 找位置模块
- 判断关系模块
- 计数模块

类比搭积木：不同积木块代表不同操作（查找、比较、计数），根据问题动态组合这些积木块来构建推理流程

D-NMN

Dynamic Neural Module Network — 动态神经模块网络

NMN（静态）

- 使用固定的预定义模块库
- 模块组合方式相对固定
- 每个问题使用相似的推理路径
- 灵活性有限

D-NMN（动态）

- 模块结构可以**动态生成**
- 根据问题特点自适应组合
- 每个问题可能触发不同路径
- 更灵活、更强大

核心区别："不是固定流程，而是按问题动态组装"——每个问题都可能触发不同的推理路径，模型根据问题内容自适应地构建最适合的推理流程

VQA的发展趋势小结

VQA经历了四个阶段的持续演进

01

简单融合

CNN+LSTM

基础特征拼接

全局特征，局部信息丢失

→

02

注意力阶段

SAN、HieCoAtt

聚焦关键区域

动态关注，精准定位

→

03

记忆阶段

DMN、MAN

多步推理

记住中间，连续推理

→

04

模块化

NMN

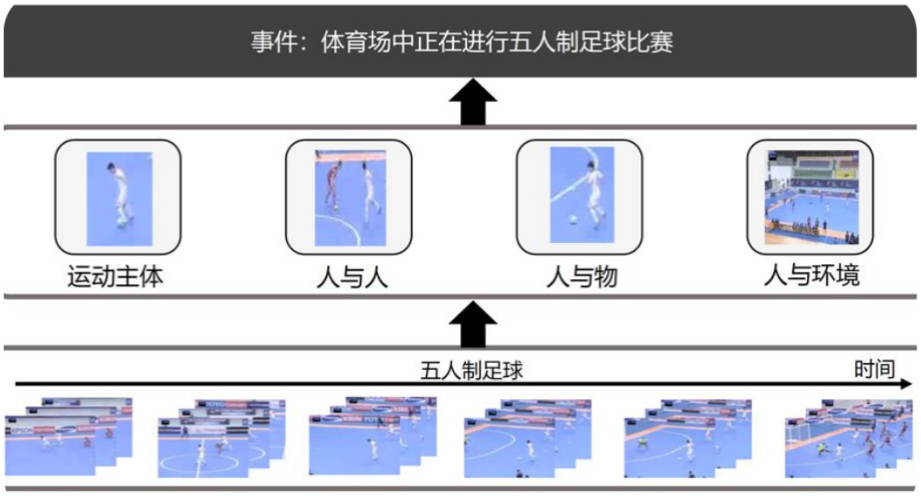
D-NMN

可组合推理

可解释

趋势总结：从粗糙到精细，从单步到多步，从黑盒到可解释

VQA开始走向视频



视频理解：从单帧到时间序列的跨越

从图像问答到视频问答

核心变化

输入从单帧变成视频

模型必须理解时间维度上的变化与事件顺序

新增挑战：

- 动作识别
- 时序关系
- 事件因果
- 镜头变化

VideoQA是VQA的自然延伸，也是更贴近真实世界的任务形式。从理解静态画面到理解动态世界，是视觉智能的重要跨越。

VideoQA数据集

数据集	数据来源	特点
TGIF-QA	GIF动画片段	基于动画的问答，侧重动作理解
TVQA	电视剧片段（含字幕）	结合视觉与文本信息的多模态问答
ActivityNet-QA	活动视频	长视频中的活动理解与问答
AGQA	组合式动作视频	组合式动作推理问答，强调因果关系

推动作用：这些数据集推动了视频问答从简单识别走向时序推理，从静态理解走向动态分析

VideoQA为什么更难

视频问答的三重挑战

空间理解

每帧看什么

继承图像VQA的空间挑战

需要在每一帧中定位关键对象

时序理解

动作先后顺序

"先发生了什么？"

需要理解时间线上的事件顺序

因果推理

事件为什么发生

"这个人接下来会做什么？"

需要理解因果关系和意图

示例问题："先发生了什么？""这个人接下来会做什么？"——需要理解时间线上的因果关系，而不仅是静态内容

VQA的局限

为什么"答对"还不够

局限表现

模型给出正确答案，但不指出答案对应的图像区域

本质问题

可能通过语言先验或统计关联"猜中"答案，而非真正理解图像

这一局限自然引出对像素级理解的需求 → SAM的方向

从VQA走向区域解释

可解释性的需求：用户不仅想要答案，还想要证据

VQA的回答方式

模型："图中有一只猫"

用户："猫在哪里？"

模型：无法精确指出位置

用户的需求

模型："图中有一只猫"

+ 猫被精确框出的区域

用户："答对了，而且我知道依据"

"答对了，但依据不清楚"——这种从全局回答到局部解释的需求，推动了分割模型的发展

为什么需要SAM

问题升级：如果希望模型把关键对象精确圈出来（像素级精度），就需要进入像素级视觉理解阶段

从语义理解到空间定位
是多模态模型走向精确感知的关键一步
VQA回答了"图中有什么"
SAM将回答"它们在精确的哪个位置"
像素级的边界，而不仅是边界框

本节内容

CONTENTS

一、VQA(视觉问答)

二、SAM（通用图像分割模型)

三、PaLM-E(具身多模态语言模型)

什么是图像分割

分类、检测、分割的对比

分类："图中有猫" → 整张图一个标签

检测：猫在[100,200,300,400] → 边界框

分割：像素级边界，精确到每个像素

分割提供最精细的空间理解

分类告诉我们"有什么"，检测告诉我们"在哪里（大致）"，分割告诉我们"每一个像素属于什么"

从粗粒度到细粒度的空间理解递进

传统分割的问题

局限一：固定类别

只能分割预定义的类别
如COCO的80类、Pascal VOC的20类
遇到新目标时**无法处理**

局限二：标注成本高

需要大规模精细标注
每个像素都要标注
标注成本**极其高昂**

核心问题：传统分割高度依赖固定类别和大规模标注
遇到新目标时泛化能力有限
这是SAM要突破的核心问题
SAM的目标：不限类别、任意对象、可提示分割

SAM的核心创新

Promptable Segmentation

可提示分割——SAM把分割任务改造成"可提示"的任务

点提示

点击一个点来指定目标

框提示

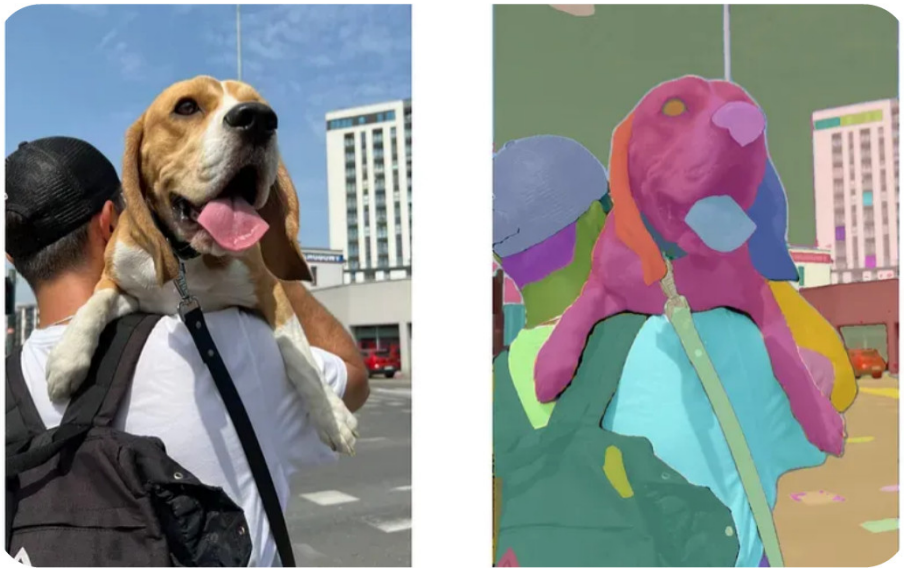
画一个框来圈出目标

Mask提示

提供粗略mask来精修

核心思想：从"按类别分割"变成"按提示分割"——不再受限于预定义类别，真正实现"任意对象分割"

点提示



SAM分割效果：原始图像（左）与分割结果（右）

SAM如何理解一个点击

交互方式：用户点一下

模型把这个点当作**目标线索**

推断完整的分割边界

原理：

- 点提示编码器将位置转为特征
- 与图像特征交互
- 生成包含该点的目标mask

特点：最简单的交互方式，**一次点击**即可触发分割，用户体验极佳

框提示

框提示：粗定位 + 精修边界

SAM如何理解一个框

交互方式：用户画一个框

相当于告诉模型"目标大致在这里"

两步过程：

1. 粗定位：框定大致范围
2. 精修边界：在框内优化分割轮廓

优势：框提示提供了更多的空间信息，通常比单点提示更精确，有利于讲清"粗定位+精修边界"的分割策略

Mask提示

SAM如何精修已有区域

应用场景

迭代精修——先快速生成粗略mask，再用SAM细化边缘

价值

SAM作为"分割精炼工具"，提升已有分割结果的质量

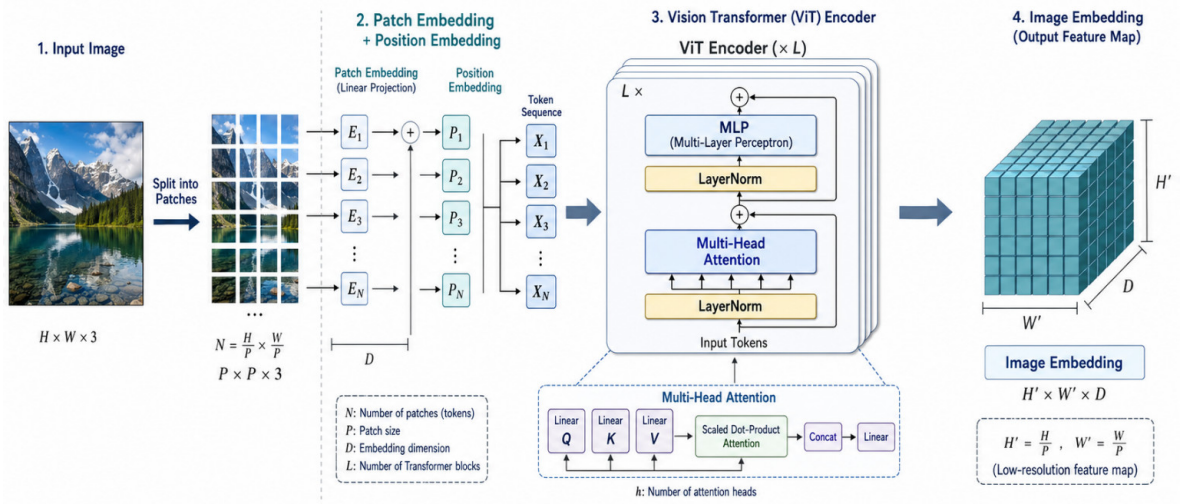
适合展示"粗掩码→精掩码"的过程，展示了SAM作为分割精炼工具的价值

SAM整体结构

三模块总览：图像编码器 + 提示编码器 + 掩码解码器

工作流程：图像编码器提取特征 → 提示编码器编码用户输入 → 掩码解码器融合两者生成最终分割结果

图像编码器



基于ViT的视觉特征提取

SAM如何"看图"

架构基础：ViT（Vision Transformer）

三个关键特点：

- 预训练于大规模数据集

具有强大的视觉表示能力

- 一次编码整张图像

后续不同提示可复用同一特征

- 是整个模型的"视觉基础"

设计亮点：图像编码器只需运行一次，后续不同提示（点/框/mask）都可以复用同一组图像特征，大大提高了交互效率

提示编码器

SAM如何"理解用户意图"

稀疏提示

点、框等位置信息

密集提示

Mask等像素级信息

统一表示

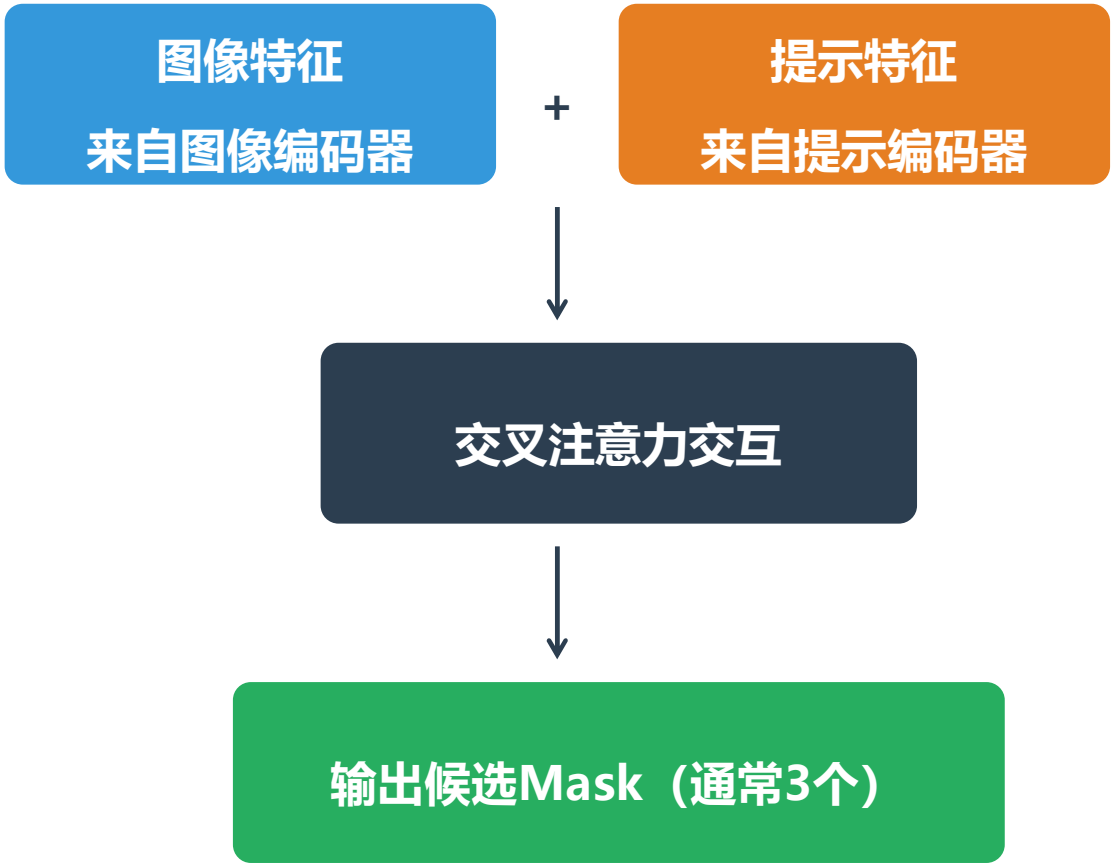
映射到同一特征空间

关键设计：不同类型的提示分别编码，最终映射到同一特征空间

- 稀疏提示（点/框）→ 位置编码 → 特征向量
- 密集提示（mask）→ 卷积编码 → 特征图
- 所有提示最终统一表示，便于与图像特征交互

核心观点："提示也是输入模态"

掩码解码器



工作流程

- 接收图像特征 + 提示特征
- 通过交叉注意力让两者交互
- 生成候选分割mask

关键特点

一次前向传播可输出多个mask（通常3个）

覆盖不同粒度

用户可选择最合适的

为什么SAM会火

通用性

不限类别

任意对象都可分割

交互性

简单提示即可控制

分割过程

可迁移性

可作为基础模型

适配各种下游任务

零样本

无需重新训练

即可分割新对象

加上两大实用价值：

- **标注提效：大幅降低分割标注成本，加速数据集构建**
- **零样本能力：无需针对每个目标重新训练，迅速成为视觉分割领域的重要基础模型**

SA-1B数据集

1100万	11亿	400x
图像数量	分割Mask数量	比现有最大数据集大

SA-1B（Segment Anything 1 Billion）：SAM背后的大数据支撑

由数据引擎半自动收集生成，覆盖绝大多数视觉场景

规模的意义

- 足够多样的分割样本（不同物体、场景、角度）
- 模型学会了"分割的通用规律"而非"特定类别的模板"
- 这是SAM强大泛化能力的重要基础

类比："做过大量练习题，迁移能力更强"

大规模数据为什么有效

从数据看能力

核心逻辑

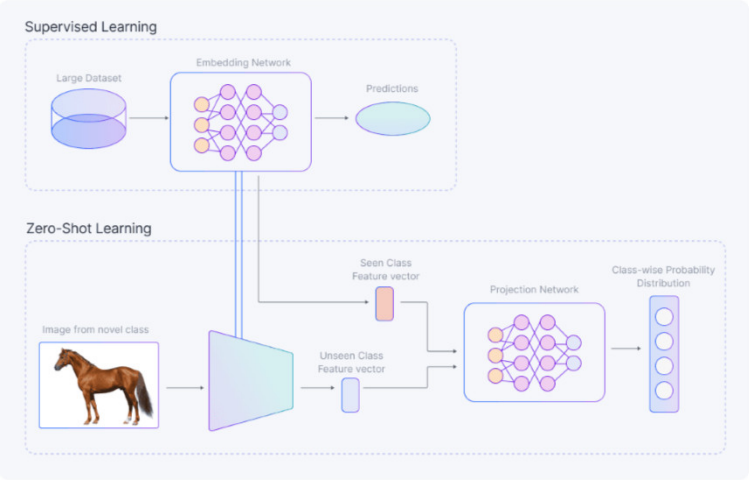
见过足够多样的样本，才更可能在新场景中表现稳定

SA-1B的多样性

覆盖绝大多数视觉场景，使SAM具备强大的泛化基础

类比："做过大量练习题，迁移能力更强"——数据规模是通用能力的基础

SAM的零样本能力



零样本学习 vs 监督学习

零样本（Zero-shot）

无需针对每个目标重新训练，就能在新场景中给出可用的分割结果

少样本（Few-shot）

仅需少量示例即可快速适应新任务

为什么能做到：SA-1B的多样性让模型学会了"分割的通用规律"而非"特定类别的模板"

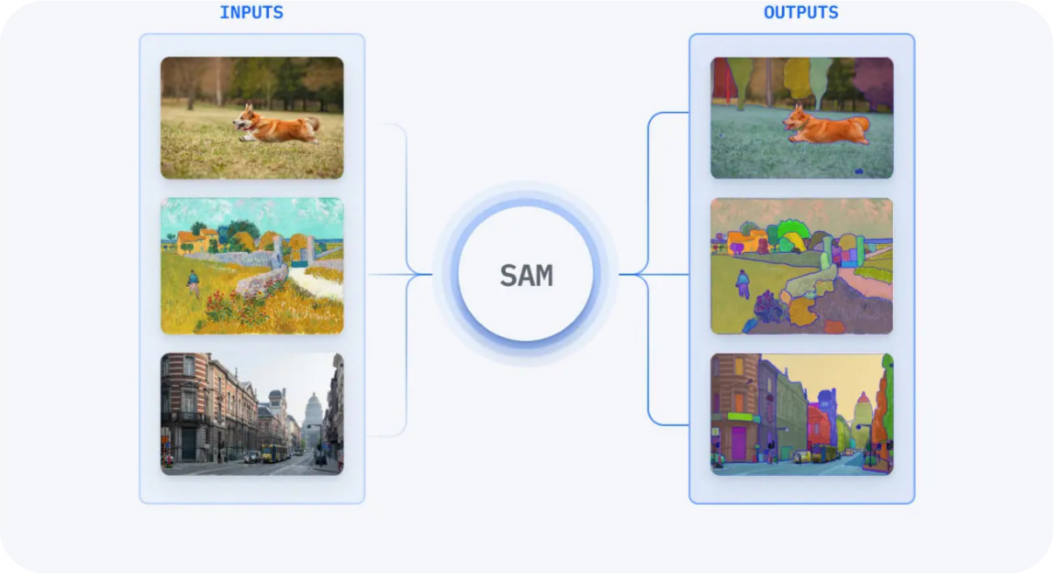
- **见过各种形状、大小、纹理的对象**
- **学会了"什么是对象边界"的通用概念**
- **因此能泛化到从未见过的对象类型**

SAM与传统分割对比

对比维度	传统分割	SAM
分割方式	按类别训练	按提示分割
类别限制	只能分割训练过的类别	任意对象，不限类别
标注需求	每类需要大量标注	通过提示指定目标
泛化能力	封闭集，新类别需重新训练	开放集，零样本泛化
交互方式	无交互，直接输出	支持点/框/mask交互

范式转变：从"封闭集"到"开放集"的分割范式转变——这是SAM最具革命性的贡献

SAM演示1：同图多目标



SAM输入输出示例：同一张图可分割多个不同目标

同图多目标分割

演示方法：通过改变提示位置

让同一张图产生完全不同的分割结果

示例：

- **点击汽车 → 分割汽车**
- **点击行人 → 分割行人**
- **点击建筑 → 分割建筑**

SAM的灵活性和交互性：同一个模型，通过不同提示可以分割完全不同的目标

SAM演示2：多点提示

单点提示的问题

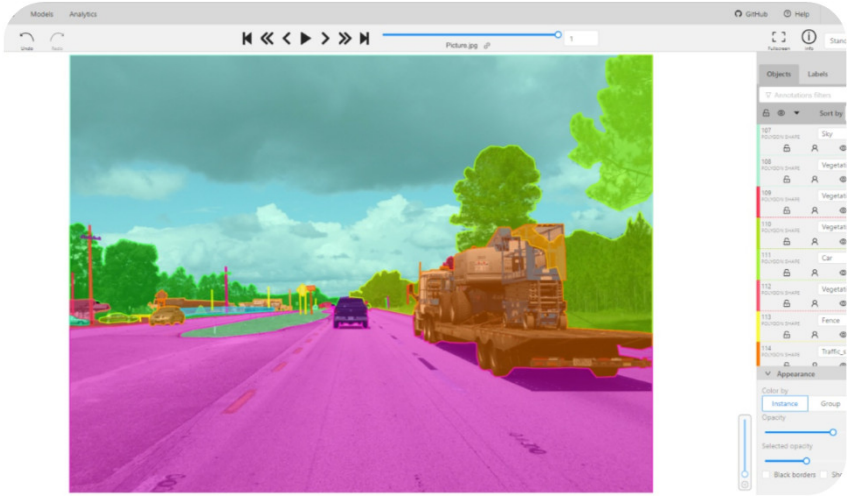
可能产生歧义
一个点可能属于多个重叠目标
模型难以确定用户的真正意图

多点提示的优势

帮助模型排除干扰
更多有效提示可以精确目标范围
提高分割准确性

- 在拥挤场景中，单点可能选中错误目标
- 增加提示点可以逐步精确目标范围
- 展示"提示数量影响结果"的现象

SAM应用1：数据标注



数据标注工具：从手工到AI辅助的转变

传统方式
手工描边
数分钟/个

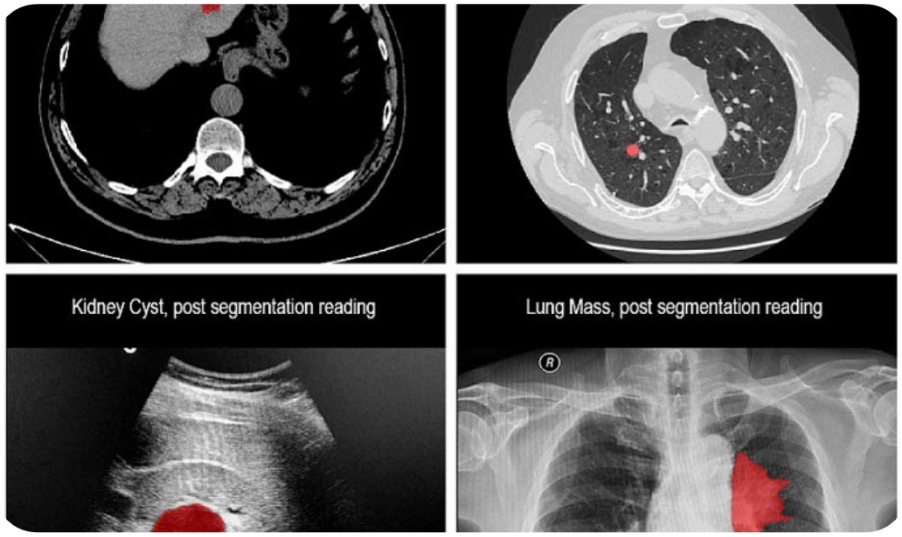
SAM方式
提示+修正
数秒/个

应用价值：

- 大幅降低分割标注成本

效率对比：传统"手工描边"（标注一个目标需数分钟）→ SAM"提示+修正"（点击+微调，**数秒完成**）
效率提升可达**数十倍甚至上百倍**

SAM应用2：医学图像



医学图像分割：器官与病灶的辅助识别

专业场景中的辅助分割

应用场景

- 器官分割

CT/MRI中的肝脏、心脏等

- 病灶检测

肿瘤区域分割

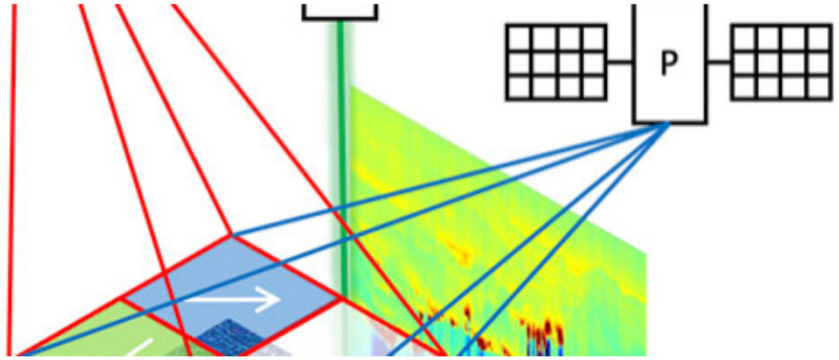
- 病理分析

细胞分割与计数

重要提醒：SAM在医学场景中是"辅助工具"而非"替代医生"

高风险场景仍需人工复核，医疗决策最终由专业人员做出

SAM应用3：遥感与工业



遥感图像分析

工业质检场景

遥感领域
- 建筑物提取
- 道路分割
- 农田监测

工业质检
- 缺陷检测
- 零件分割
- 尺寸测量

共同特点：都需要精确的空间定位，SAM的通用分割能力可以快速适配这些专业场景

SAM的局限

SAM不是万能抠图工具

遮挡严重

目标被部分遮挡
边界模糊，可能失败

小目标

极小的物体
难以精确分割

语义歧义

"分割人"——是指身体
还是包括衣服?

SAM的能力边界

- SAM擅长清晰、大目标的分割
- 复杂场景仍需人工干预或后处理
- 建议加入1-2个失败样例进行分析
- 没有完美的模型，只有适合的模型

分割不等于理解

SAM擅长：空间能力

精确圈出"猫"的轮廓

像素级边界定位

SAM不擅长：认知能力

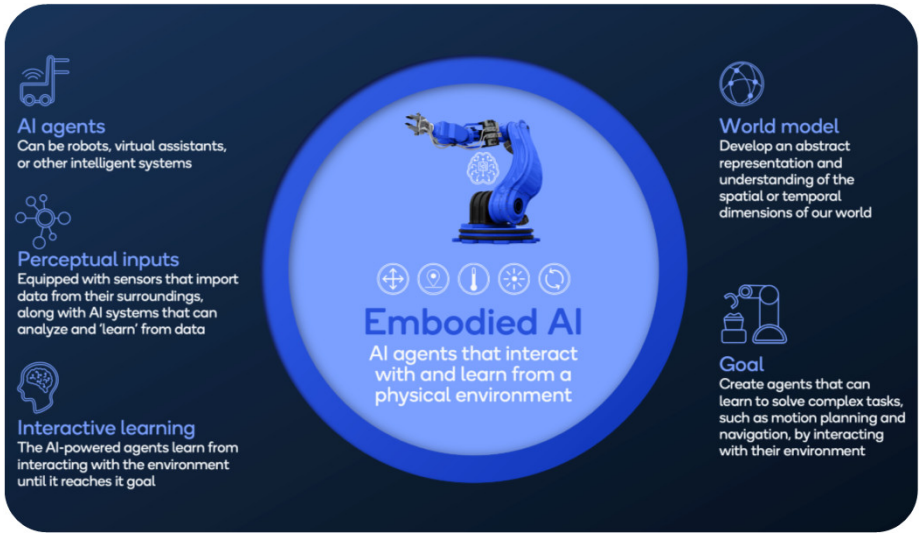
不知道这是一只猫

更不知道猫在做什么

分割是理解的基础，但不是理解的全部

这为PaLM-E的引入做了铺垫——从感知到理解再到行动

从"会圈"到"会行动"



具身智能：从感知到行动的跨越

问题再次升级

现实需求

不仅需要精确感知（SAM做什么）

还需要结合任务目标和环境状态做决策

举例：

机器人不仅要"看到杯子"

还要"知道如何拿起杯子"

从感知到行动，自然导入具身智能（Embodied AI）的概念——AI不再只是"旁观者"，而是"参与者"

本节内容

CONTENTS

一、VQA(视觉问答)

二、SAM（通用图像分割模型)

三、PaLM-E(具身多模态语言模型)

为什么叫具身多模态

具身智能：机器人通过传感器与真实世界交互

"具身"的含义

模型开始接入**环境感知**与**身体状态**

不再只处理抽象的文本和图片

典型例子：

- 摄像头"看到"环境
- 关节传感器"感知"姿态
- 语言模型"理解"指令
- 规划"动作"并执行

核心转变：AI从"旁观者"变成"参与者"——不再只是观察世界，而是能够与世界交互并产生影响

PaLM-E的基本任务

PaLM-E试图连接三个核心能力

感知理解

看懂场景

理解环境中的物体和布局

语言推理

理解指令

解析自然语言任务描述

动作规划

生成可执行的动作序列

把理解转化为行动

关键区别：普通VLM（视觉语言模型）只做①+②，PaLM-E还做③——把理解转化为行动

PaLM-E的输入形式

多模态句子（Multimodal Sentences）

图像Token | 状态Token | 文本Token | 图像Token | ...

所有模态统一映射到一个连续的Token序列中

类比：就像一句话中既有中文又有英文单词，PaLM-E的"句子"中既有文本token，又有图像token，还有状态token

核心思想：PaLM-E把图像、机器人状态和文本共同映射到统一的token序列中，实现多模态的统一表示

什么叫多模态句子

统一表示空间的核心思想

类比翻译：不同语言需要翻译成共同语言才能交流

不同模态需要"翻译"成共同表示才能联合推理

图像、状态、文本 → 统一向量空间 → 联合处理

图像编码

通过视觉编码器

映射为图像token

状态编码

通过状态编码器

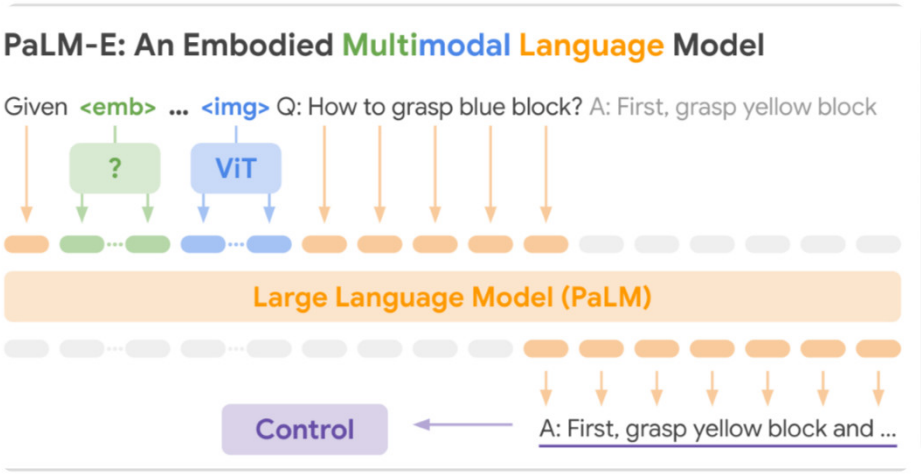
映射为状态token

文本编码

通过词嵌入

映射为文本token

PaLM-E的核心结构



PaLM-E架构：编码器 + 注入 + LLM推理

① 编码（Encoder）
连续观测 → 向量表示

② 注入（Injection）
向量作为特殊token注入语言模型

③ 推理（Reasoning）
LLM统一处理 → 输出答案/行动计划

核心流程：编码 → 注入 → 推理。感知输入通过编码器转成向量，再送入预训练语言模型统一处理，生成答案或行动计划

为什么这种设计重要

LLM从"语言中枢"变为"世界中枢"

传统LLM

只处理词token
局限于文本世界
无法直接感知物理世界

PaLM-E中的LLM

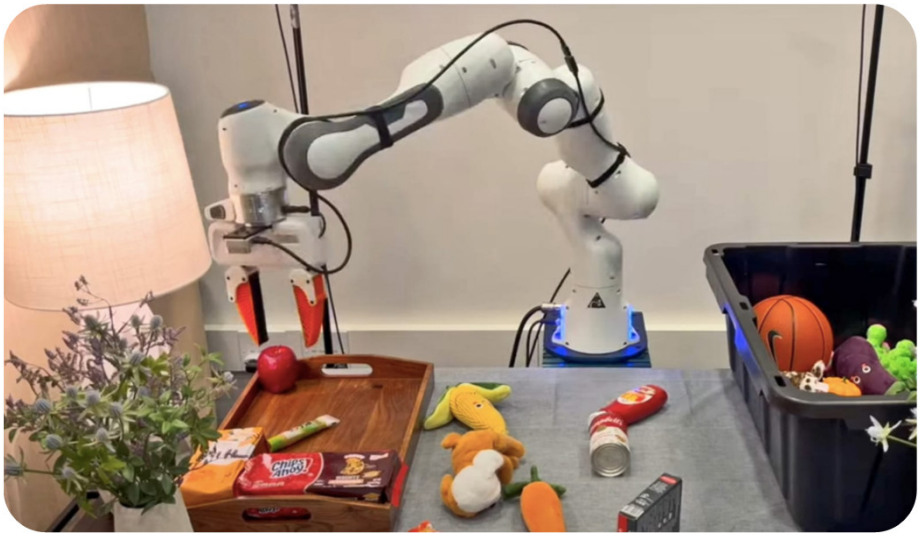
处理视觉+状态+文本
接入物理世界的感知
成为"世界中枢"

设计意义：

- 不需要为每个模态设计专门的融合模块
- 利用LLM强大的序列建模能力统一处理多模态信息
- 语言模型不再只处理词，而是开始处理视觉和状态共同构成的"世界描述"

这是理解PaLM-E**方法论**的核心

机器人状态为什么必须加入



机械臂抓取：需要知道自身状态才能行动

环境理解不能脱离身体状态

核心观点：

机器人不仅要"看见"环境

还必须"知道"自己当前姿态

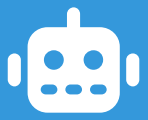
举例：机械臂抓取

- **关节角度**

身体状态是行动决策的必要输入。没有自身状态信息，模型无法判断"能否执行""如何执行"——这是具身智能区别于纯感知系统的关键

PaLM-E支持哪些任务

多任务统一——不是单任务专用模型



机器人操作规划

"把红色积木放到蓝色积木上"

视觉问答

"桌子上有什么？"

→ 回答

图像描述

"描述这个场景"

→ 生成文字描述

统一框架：PaLM-E不是单任务专用模型，而是多任务通用框架——感知、语言、行动在一个模型中统一处理

为什么PaLM-E也做VQA

VQA在具身系统中的价值

行动前的信息收集

机器人要行动，首先必须能够回答关于环境的问题

VQA是行动决策的信息基础

具体示例

"把杯子放到桌子上"之前

需要回答："桌子在哪里？""杯子里有水吗？"

核心观点：VQA不是被抛弃，而是被纳入更大的行动框架中

- 在VQA部分，PaLM-E继承了视觉语言模型的问答能力
- 在行动部分，PaLM-E超越了VQA，将理解转化为行动
- VQA → 信息获取 → 行动规划 → 执行

VQA成为具身智能的子能力而非独立任务

多任务联合训练

不同任务能否互相帮助?

PaLM-E的核心假设：联合训练带来正迁移（Positive Transfer）

- 不同数据域（视觉、语言、机器人）共同训练
- 不同任务（VQA、描述、规划）共享表示
- 核心假设：共享表示后，不同任务可以相互促进

做VQA学到的视觉理解 → 帮助机器人规划

正迁移是什么意思

为什么统一训练有价值

类比：学钢琴时学到的乐理知识对学习吉他也有帮助

不同乐器共享"音乐理论"的基础知识 → 跨任务迁移

VQA数据

帮助模型学会"看懂"

机器人数据

帮助模型学会"行动"

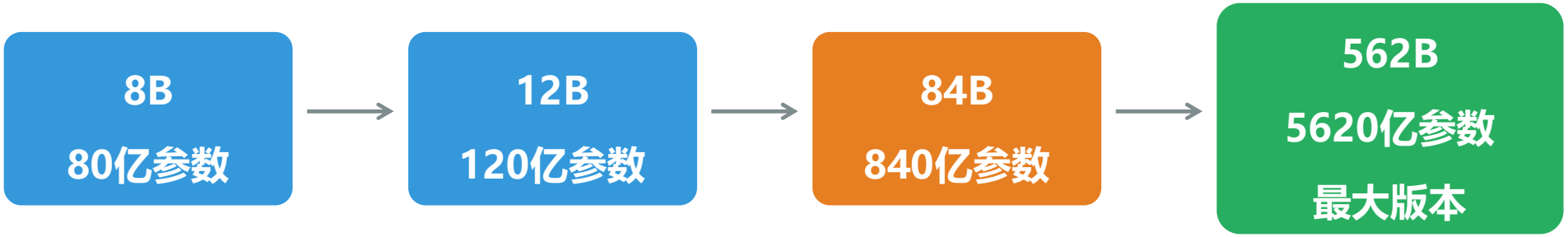
相互促进

二者共享表示，共同提升

正迁移的含义：语言、视觉和机器人任务共享表示后，可能比单任务训练获得更强泛化能力

PaLM-E的规模版本

不同参数规模适应不同应用场景



从小到大：能力逐步增强，从边缘设备到云端服务器

规模阶梯的意义

- 8B/12B：适合边缘设备部署，响应速度快
- 84B：平衡性能与效率
- 562B：最强能力，云端部署

不同规模适应不同应用场景

规模越大，多模态融合和推理能力越强

PaLM-E-562B的代表意义

5620亿参数

具身任务表现

在机器人操作规划任务上展现了强大的零样本和少样本能力

VQA任务表现

在OK-VQA等视觉问答基准上取得了当时的先进结果

PaLM-E-562B的关键成就

- **具身任务**：零样本机器人操作规划——无需针对特定任务训练
- **VQA任务**：OK-VQA等基准的先进结果——视觉问答能力突出
- **统一能力**：同一个模型同时擅长感知理解和行动规划

证明了"大规模+多任务联合训练"路线的可行性

PaLM-E与传统机器人系统对比

传统机器人系统

模块堆叠架构

- 感知模块 → 理解模块 → 规划模块 → 控制模块
- 模块间需要手工设计接口
- 信息在传递中可能丢失
- 调试困难，维护成本高

PaLM-E

端到端统一架构

- 统一表征 + 统一推理
- 所有模态输入同一个模型
- 模型直接输出动作计划
- 减少信息损失

核心区别：从"模块堆叠"到"端到端统一"——PaLM-E用一个模型替代了传统机器人系统中多个独立模块的复杂 pipeline

VQA、SAM、PaLM-E三者对比

对比维度	VQA	SAM	PaLM-E
输入	图像 + 问题	图像 + 提示	图像 + 状态 + 指令
输出	答案（文本）	分割Mask	答案 / 动作计划
目标	理解内容	精确定位	感知-语言-行动统一
典型场景	视觉问答	对象分割	具身智能

递进关系：VQA（语义层）→ SAM（空间层）→ PaLM-E（行动层）

一个综合案例

任务："把红色杯子放到书旁边"

VQA的角色

"书在哪里？"

"杯子是红色的吗？"

→ 提供语义信息

SAM的角色

精确分割出杯子和书的像素边界

→ 提供空间信息

PaLM-E的角色

结合视觉和状态

规划抓取-移动-放置的完整动作序列

→ 提供行动能力

三者协作：问答提供语义信息 → 分割提供空间信息 → 具身模型提供行动能力 → 共同完成任务

最新进展：SAM 3D

SAM 3D Objects

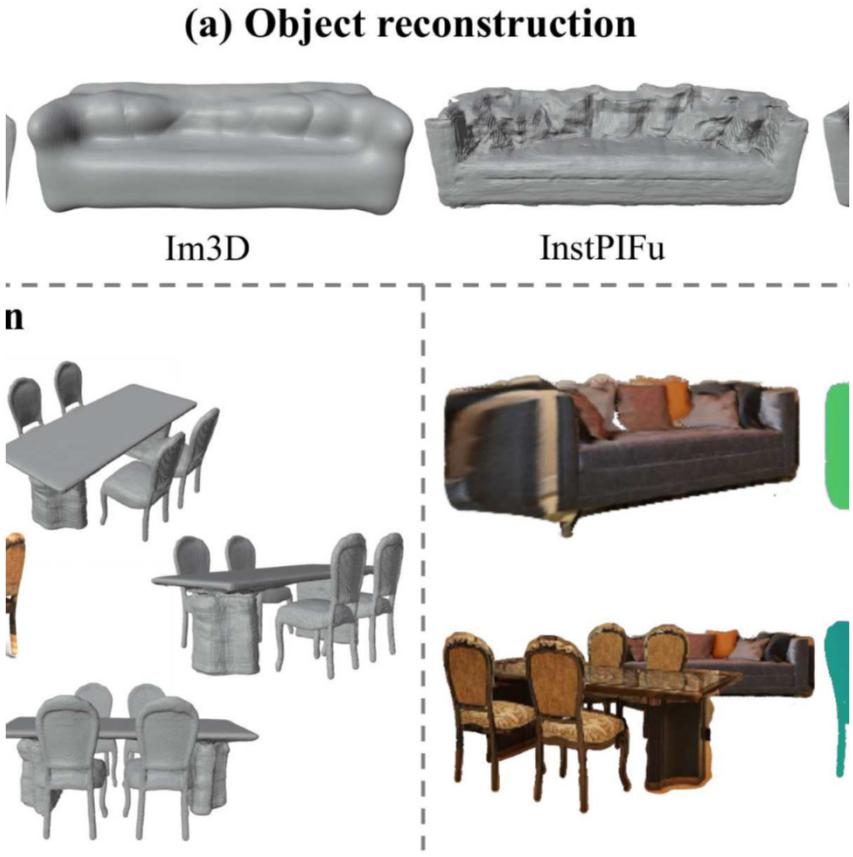
单张自然照片 → 完整3D形状、纹理贴图、空间布局

论文 2025.11.20 发布

SAM 3D Body

单张照片 → 精确人体3D姿态与形状，含骨骼结构

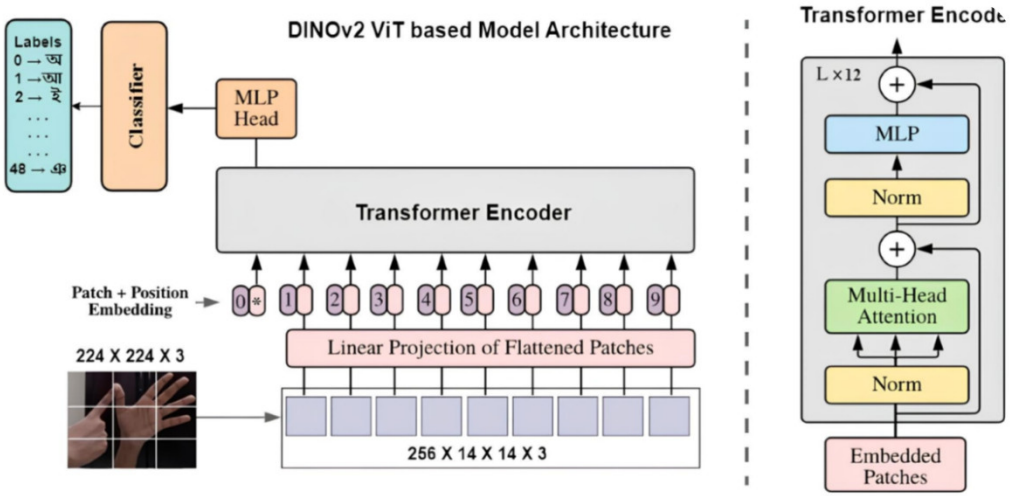
arXiv:2602.15989v1 • 2026.2



将“分割一切”从2D拓展到3D，单张自然照片即可生成带纹理的3D网格模型

2025年11月 • Meta发布 论文 arXiv:2511.16624

联合编码与MoT混合架构



DINOv2 Vision Transformer 编码架构

① **输入编码** DINOv2提取特征，输出四组"调节令牌"

② **第一阶段** 预测大致姿态与粗糙几何轮廓

③ **第二阶段** MoT融合模块细化几何与纹理

④ **联合编码** 空间—语义联合，确保物理正确

数据引擎三部曲

合成数据预训练→ 半自动数据引擎→ 真实图像后训练对齐

SOTA性能与广泛的应用场景

28%

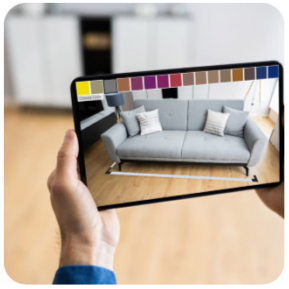
Objects Chamfer Distance 优化

19%

法向一致性提升

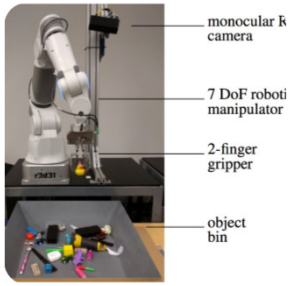
5:1

人类偏好测试胜率



AR/VR 与电商

"View in Room"虚拟摆放，让家具选购更直观



机器人

从RGB图像重建几何，用于抓取与导航

游戏影视

参考图快速生成3D资产雏形

遥感与城市建模

单目建筑重建、城市级场景建模

问题和讨论