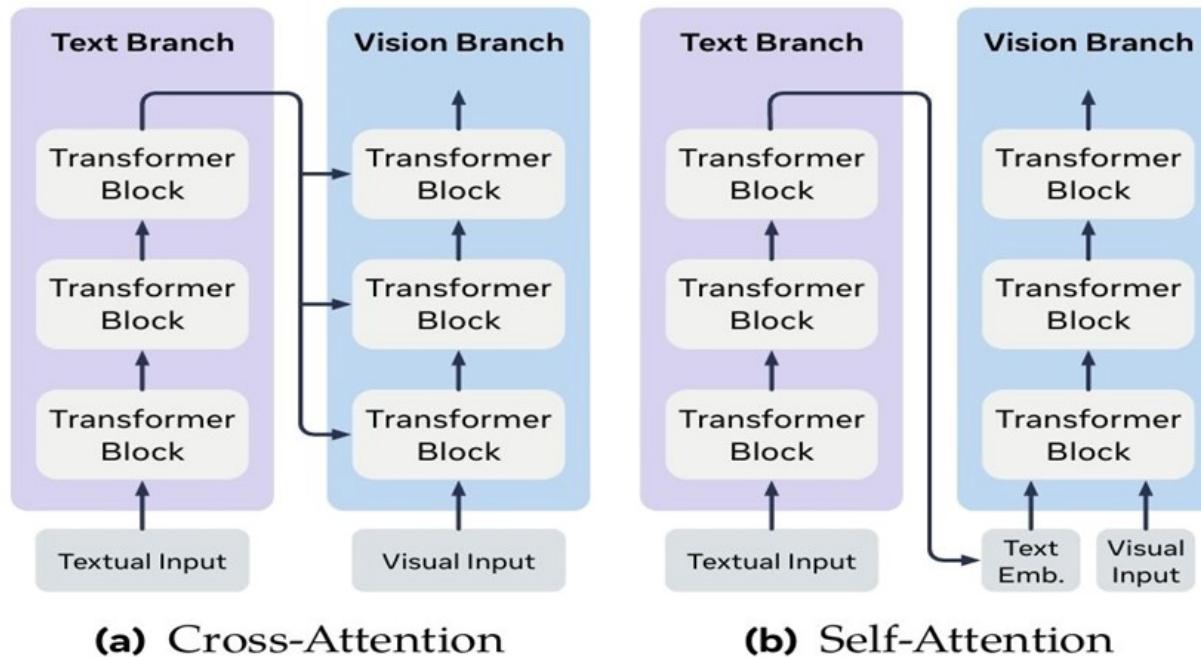


# 多模态的基本概念

## ➤ 融合(Fusion)



# AI Agent vs. Agentic AI

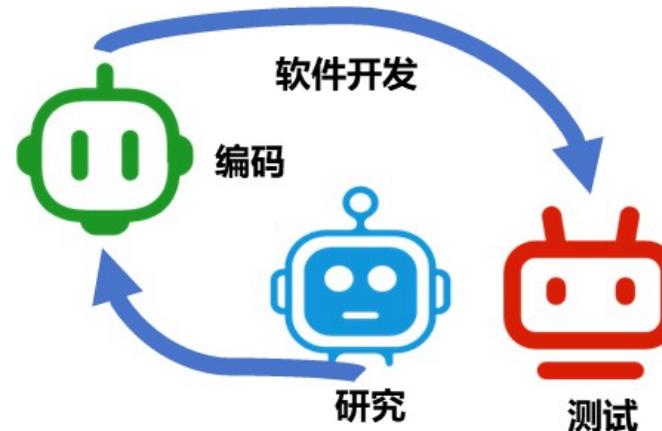
- 单个实体
- 利用工具能力
- 执行但不或短链条推理
- 一个根据指令查询天气的代理



AI Agent

用于自动化执行特定任务的**任务执行者**。聚焦效率与自动化。

- LangChain <https://www.langchain.com/>
- Auto-GPT <https://agpt.co/>



Agentic AI

一个由多个专业代理通过**协作**来完成复杂长期目标的**协同系统**。聚焦协作与负责问题解决。

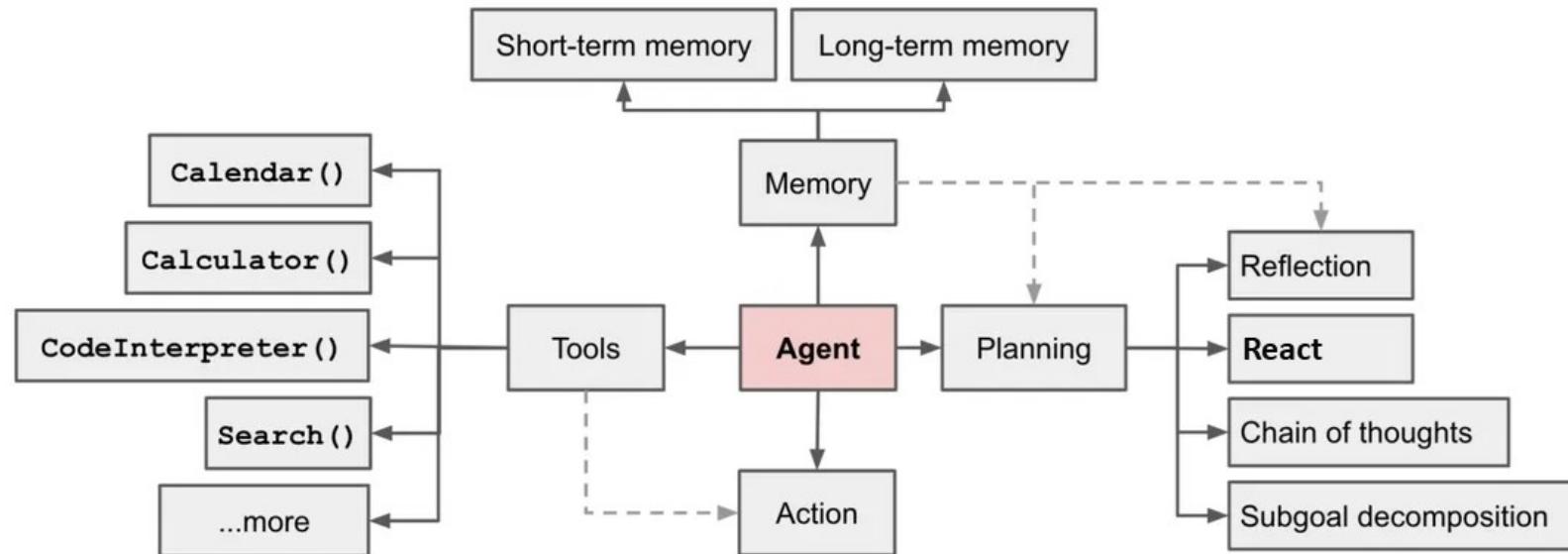
- AutoGen <https://autogen-five.vercel.app/>
- MetaGPT <https://www.deepwisdom.ai/>

填空题 3分

2. 基于LLM的Agent 的核心组件包括 [填空1] (决定做什么) ,  
[填空2] (记住什么) ; [填空3] (用什么能力完成任务)

# 智能体 (Agent)

Agent = 大模型(LLM) + 记忆 + 工具 + 规划



单选题 1分

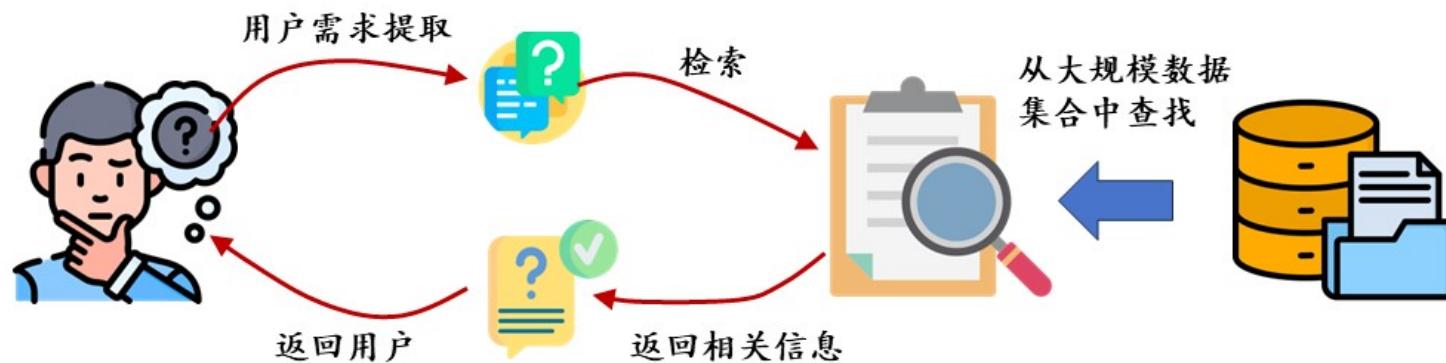
3. 以下哪项最能描述“信息检索（IR）”？

- A 让模型自动生成回答，不依赖外部数据
- B 对文本进行分词、向量化等预处理步骤
- C 从大量非结构化或半结构化数据中有效、快速地获取相关信息
- D 使用数据库主键精确定位结构化记录

# 信息检索定义

## ➤ 信息检索

信息检索 (Information Retrieval, IR) 是计算机科学的核心领域之一，致力于**从大规模的非结构化数据集合中精准高效地查找与用户需求相关的信息。**



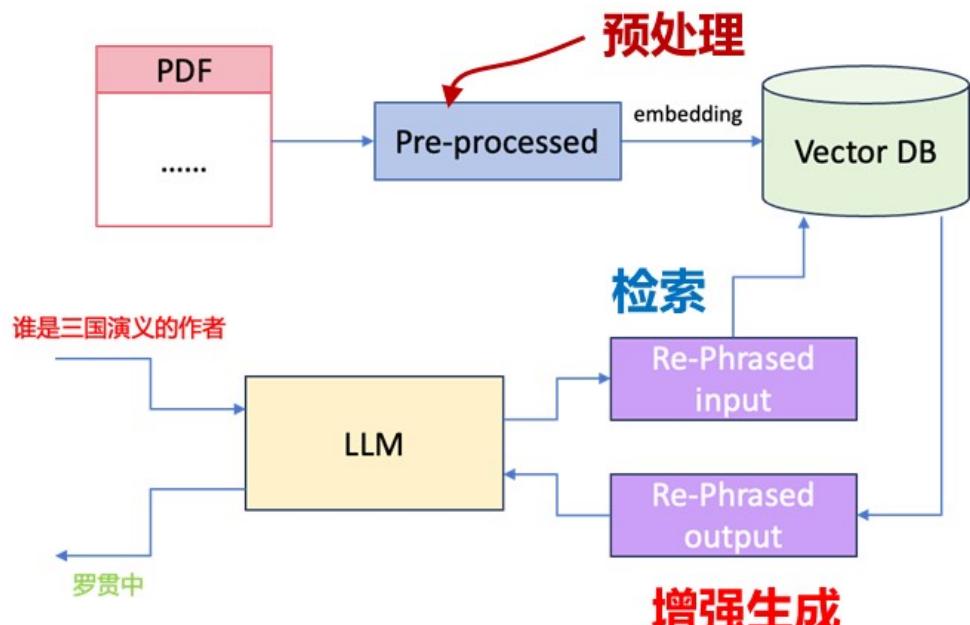
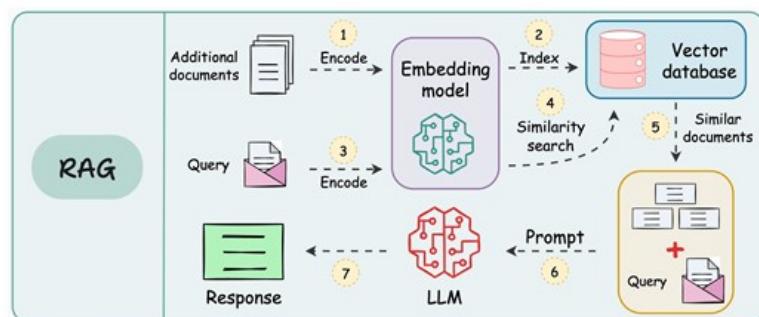
填空题 2分

4. RAG 是将 [填空1] (从外部资料中找信息) 和 [填空2] (让模型  
基于检索结果生成答案) 结合起来的框架。

# 大模型时代的信息检索

## ➤ 检索增强生成 (Retrieval Augmented Generation)

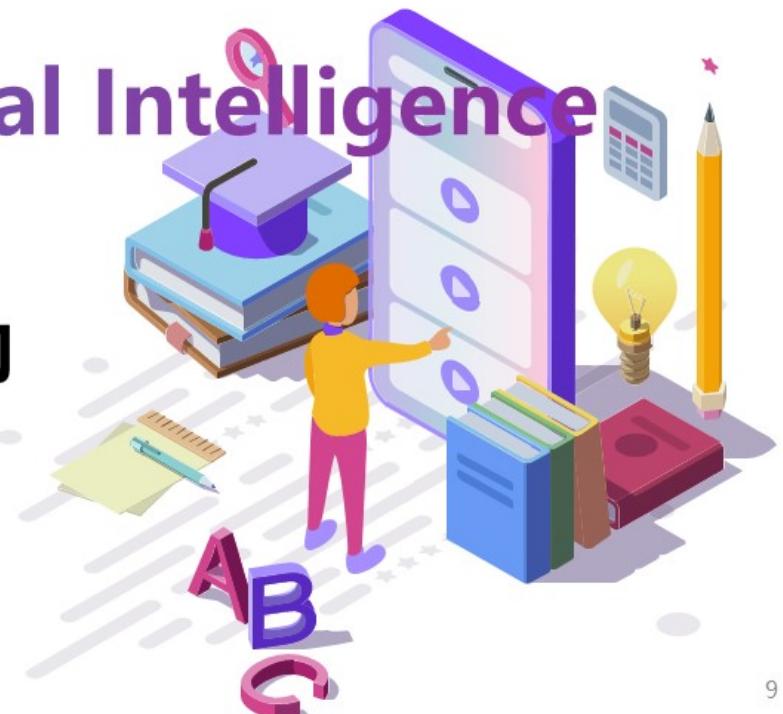
检索增强生成技术是结合**信息检索**和**文本生成**的技术框架，旨在通过动态访问外部知识库来增强生成模型的能力



# 人工智能导论

Introduction to Artificial Intelligence

## 第九章 人工智能应用与前沿方向



# 第九章

- 自然语言处理(NLP)
- 计算机视觉 (CV)
- 多模态与世界模型
- 前沿探索



# NLP无处不在

➤ 我们每天都在使用NLP!



输入法软件、Siri/小爱同学、淘宝智能客服、百度翻译、小红书内容推荐

# NLP技术的三次革命性突破

随着机器学习兴起，统计模型开始主导NLP领域。2011年苹果Siri的推出标志着语音助手的诞生，它通过分析大量语音数据来识别用户意图。这一时期的代表技术包括TF-IDF文本表示和CRF序列标注，让机器能够处理模糊发音和简单对话。

2010s至今  
**深度学习革命**

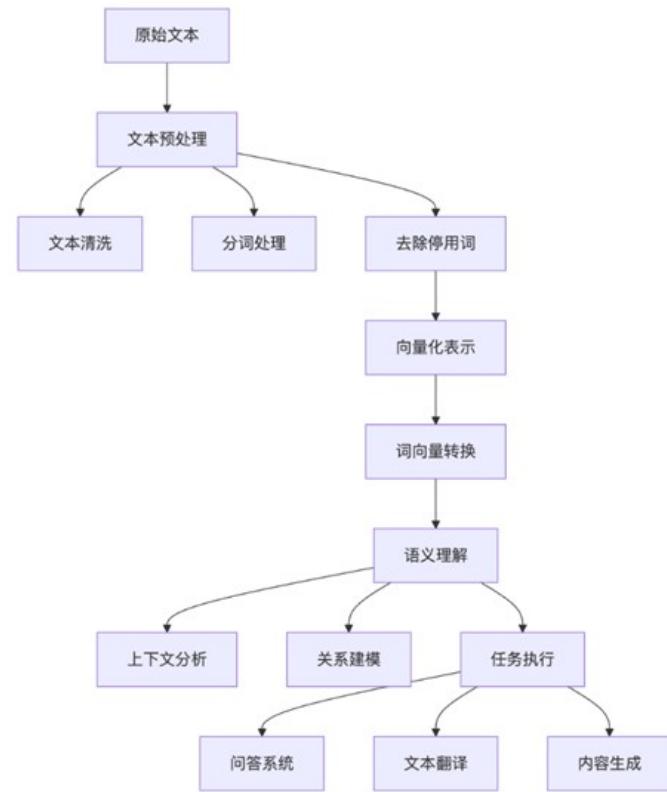
2000s-2010s  
**统计学习时代**

1950s-2000s  
**规则匹配时代**

早期系统依赖人工编写语法规则，最著名的  
是1954年IBM的俄语-  
英语翻译系统，它只能  
处理60个固定句型。  
这种方法就像教机器背  
诵语法书，遇到未定义  
的表达就完全失效。

2017年Google提出的  
Transformer架构彻底改  
变了NLP格局。基于自注  
意力机制，BERT、GPT等  
预训练模型能够同时关注  
句子中的所有词语关系，  
在多项语言任务上超越人  
类水平。

# NLP Pipeline



自然语言处理 (NLP)

自然语言理解 (NLU)

自然语言生成— (NLG)

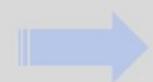
自然语言的理解就是希望机器可以和人一样，  
有理解他人语言的能力。

自然语言的生成就是将非语言格式的  
数据转换成人类的语言格式，以达到  
人机交流的目的。

# NLU vs. NLG

## ➤ NLU(理解)

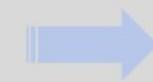
Language  
语言文字



输入	输出
自然语言	结构化信息或分类

让机器做“阅读理解”

## ➤ NLG(生成)



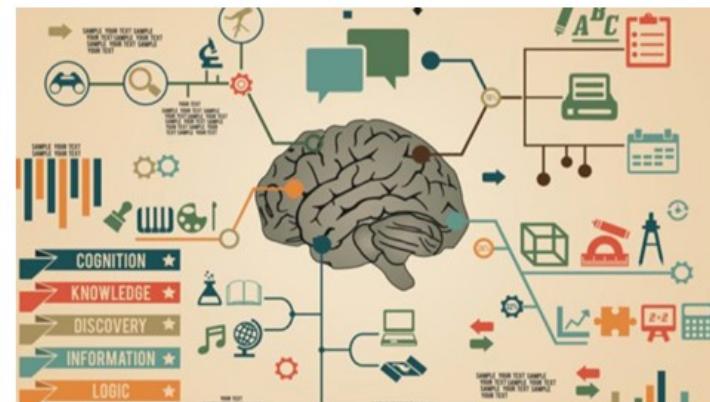
Language  
语言文字

输入	输出
结构化信息或提示	自然语言

让机器“思考写作”

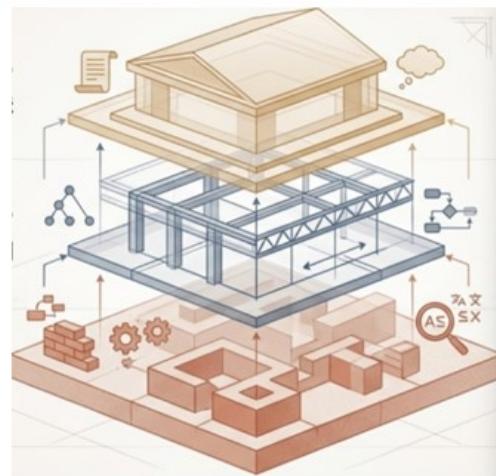
# NLU——自然语言理解

## ➤ 自然语言理解——让机器读懂人类语言的艺术



自然语言理解(Natural Language Understanding, NLU)是所有支持机器理解文本内容的方法模型或任务的总称。NLU在文本信息处理系统中扮演着非常重要的角色，是推荐、问答、搜索等系统的必备模块。

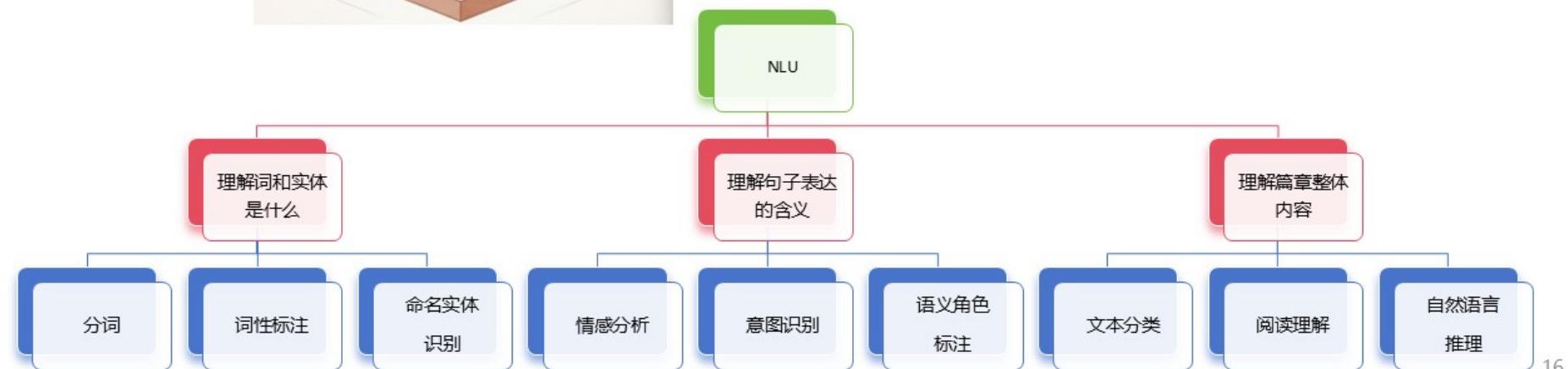
# NLU: 典型任务



篇章：归纳主题与逻辑

句子：解析结构与意图

词与短语：识别符号与实体



## NLU: 词与短语层任务

- 分词：将连续字符串切分为词或子词

马克龙 | 上周五 | 访问 | 了 | 四川大学 | 。

- 词性标注：赋予每个词语法身份

马克龙(nr) 上周五(t) 访问(v) 了(u) 四川  
大学(nt)。 (w)

- 命名实体识别 (NER) : 将字符串转化为可计算的只是庙殿，为信息抽取提供  
关键对象。

[马克龙](PER)[上周五](DATE)访问了[四川大学](LOC)。

构成后续所有句法与  
语义步骤的输入前提

# NLU: 句子层理解

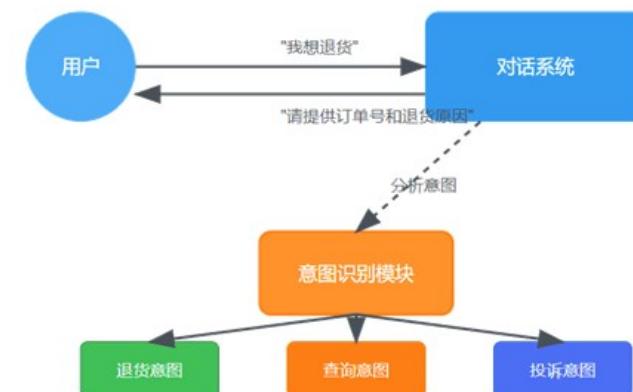
## ➤ 情感分析:

识别和提取文本中的观点、情感、立场。



## ➤ 意图识别:

把用户话语映射到预定义的动作或服务类别。



# NLU: 句法分析与语义角色

## ➤ 句法分析:

理解句子结构：主谓宾，修饰关系。

## ➤ 语义角色标注:

在句法树上填充施事、受事、时间、地点等角色，把语法结构升华为语义框架



# NLU: 篇章理解

## ➤ 文本分类:

判断文章的主题或类别



分类 → 谣言

新闻

## ➤ 阅读理解:

让模型从文章中找出问题  
的答案

昨天，北京出现了今年入冬以来的第一场大雪。许多市民走出家门拍照，公园和街道都被白雪覆盖，交通部门也提前部署了除雪车辆。

**问题：北京昨天发生了什么天气？**

**答案：下雪**

## ➤ 自然语言推理:

判断两句话的逻辑关系：  
蕴含 / 矛盾 / 中立

**前提：**

学生们在操场上踢足球。

**假设：**

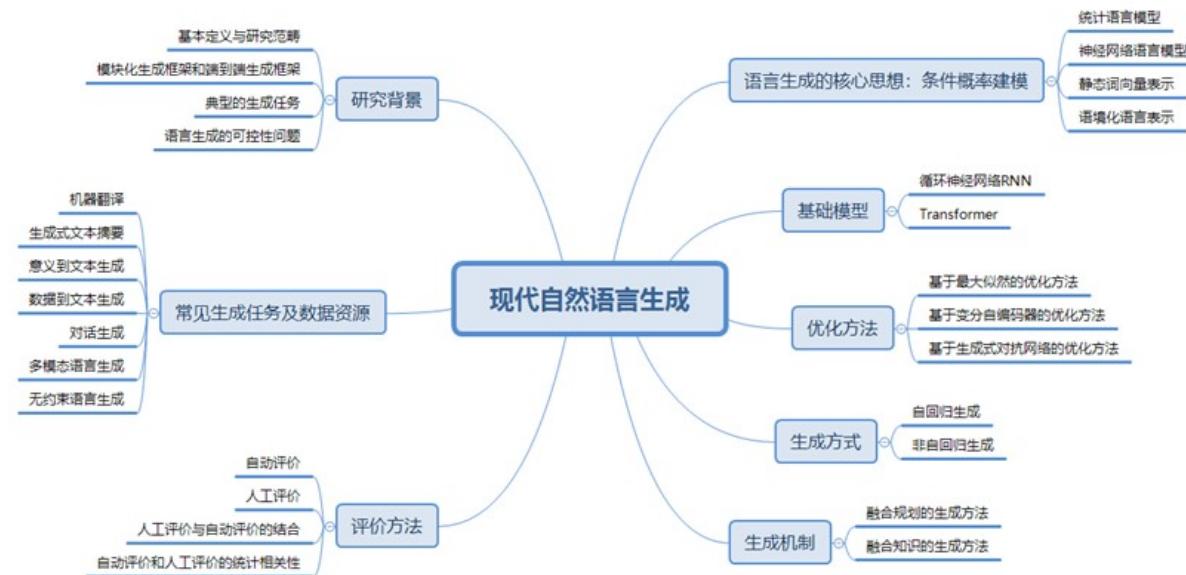
有人在进行户外运动。

**判断：蕴含**

因为“踢足球”显然属于户外运动。

# NLG：自然语言生成

## ➤ 赋予机器创造与表达能力

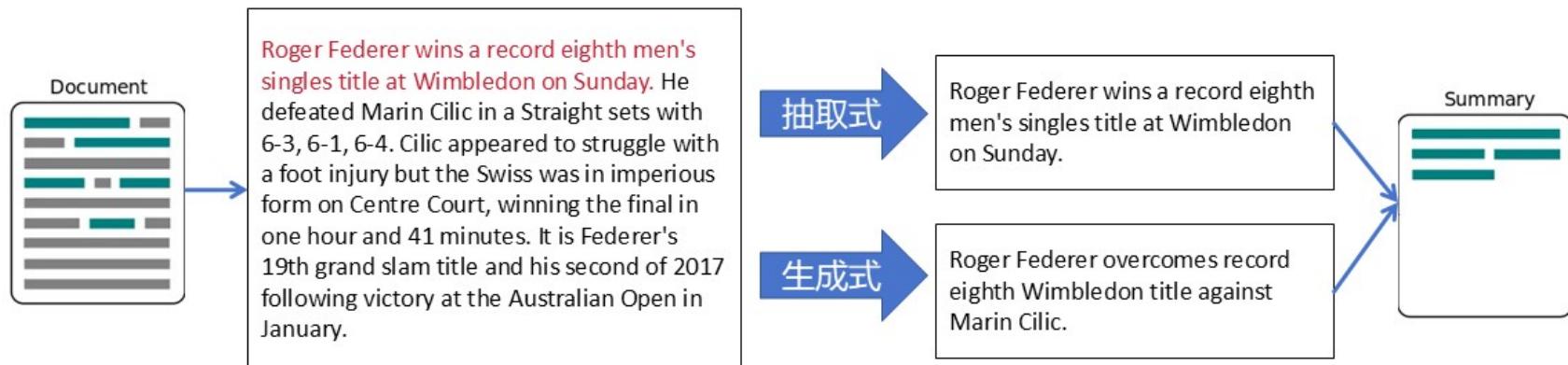


让计算机自动生成连贯、自然、符合语义的语言文本。

# NLG: 摘要生成

## ➤ 摘要生成任务：将长文本凝练成简短、流畅的摘要

- 抽取式 (Extractive) : 从原文挑选句子，保证事实准确，但可能不连贯
- 生成式 (Abstractive) : 理解原文后重新组织语言生成摘要，更流畅但可能产生幻觉



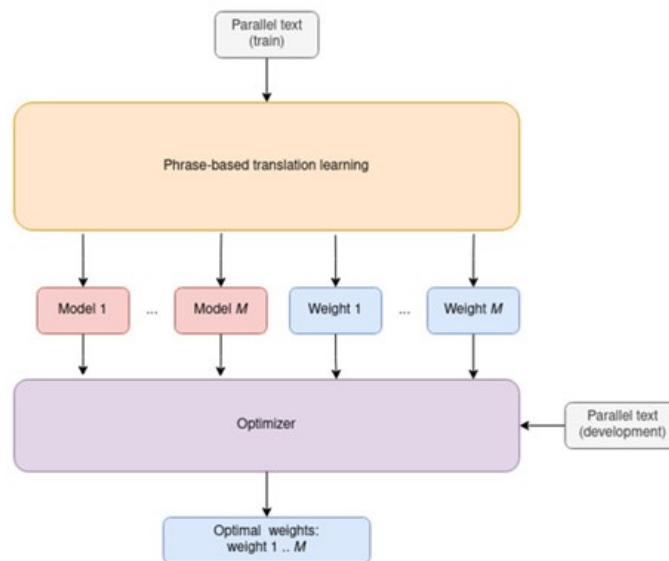
# NLG: 机器翻译

## ➤ 机器翻译

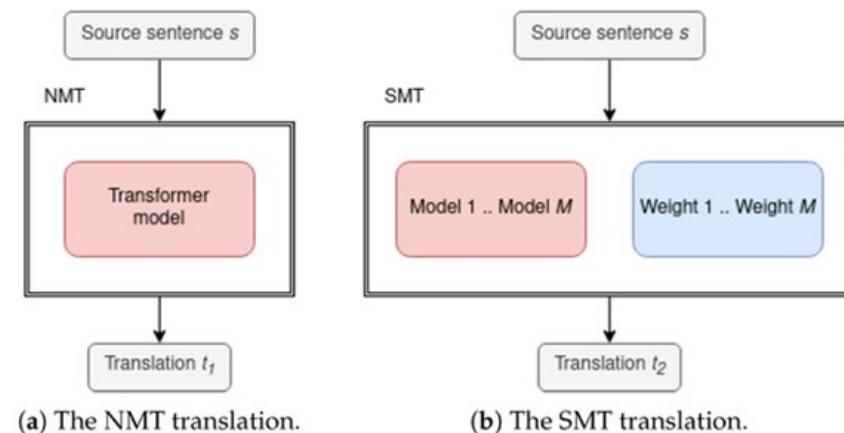
输入一种语言 → 输出另一种语言

- 技术革新: 从概率机器翻译 (SMT) 到神经机器翻译 (NMT)

### 基于短语拼接



### 基于Transformer的端到端生成



# NLG: 机器翻译

## ➤ 领域前沿研究

- 低资源语言翻译、AI同声传译、多模态翻译（看图翻译）



Mr. Lawrence, hold our course. → Text-only MT Model → 劳伦斯先生，保持我们的课程。

(a) The incorrect translation of the polysemous word “course” by the text-only MT model.

Mr. Lawrence, hold our course. → MMT Model → 劳伦斯先生，保持我们的航向。

(b) The correct translation of the polysemous word “course” by the MMT model.

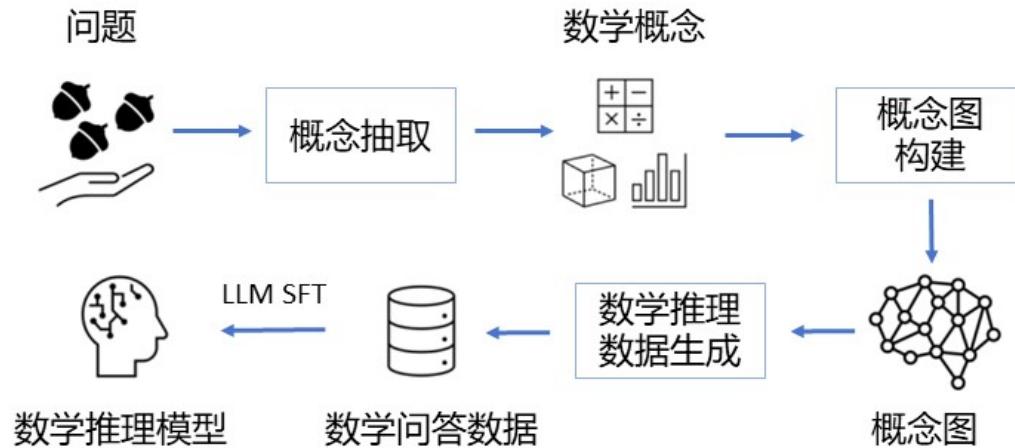
# NLG: 数学推理

## ➤ What?

- 让模型解决以自然语言描述的数学问题

## ➤ Why?

- 衡量模型的逻辑推理能力

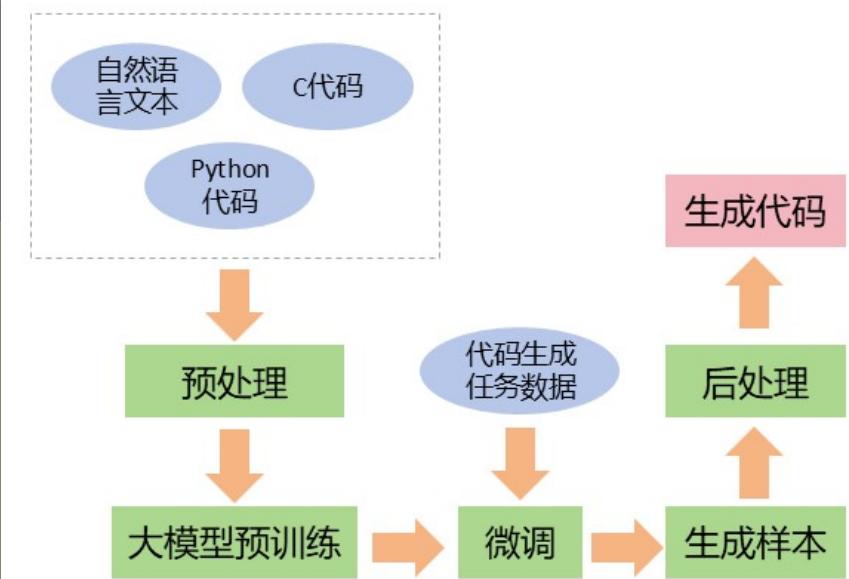


# NLG: 代码生成

- 根据自然语言描述自动生成可执行代码

A screenshot of the GitHub Copilot interface. It shows a code editor with a file named 'prime.js'. The code is a function to check if a number is prime. A tooltip from GitHub Copilot suggests: "Write an efficient function to determine whether a given number is prime." Below the code editor are buttons for "Accept" and "Discard". The GitHub Copilot logo and the text "Made changes." are visible.

```
JS prime.js ●  
home > alvin > Downloads > JS prime.js > isPrime  
1  function isPrime (number) {  
2      // Write an efficient function to determine whether a given number is prime.  
3      // GitHub Copilot  
4      // Made changes.  
5      // Accept Discard  
6      // Accept or Discard 1 change.  
7  
8      if (number <= 1) {  
9          return false;  
10     }  
11     if (number <= 3) {  
12         return true;  
13     }  
14     if (number % 2 === 0 || number % 3 === 0) {  
15         return false;  
16     }  
17     let i = 5;  
18     while (i * i <= number) {  
19         if (number % i === 0 || number % (i + 2) === 0) {  
20             return false;  
21         }  
22         i += 6;  
23     }  
24     return true;  
25 }
```



显著提升开发效率

## NLU vs. NLG

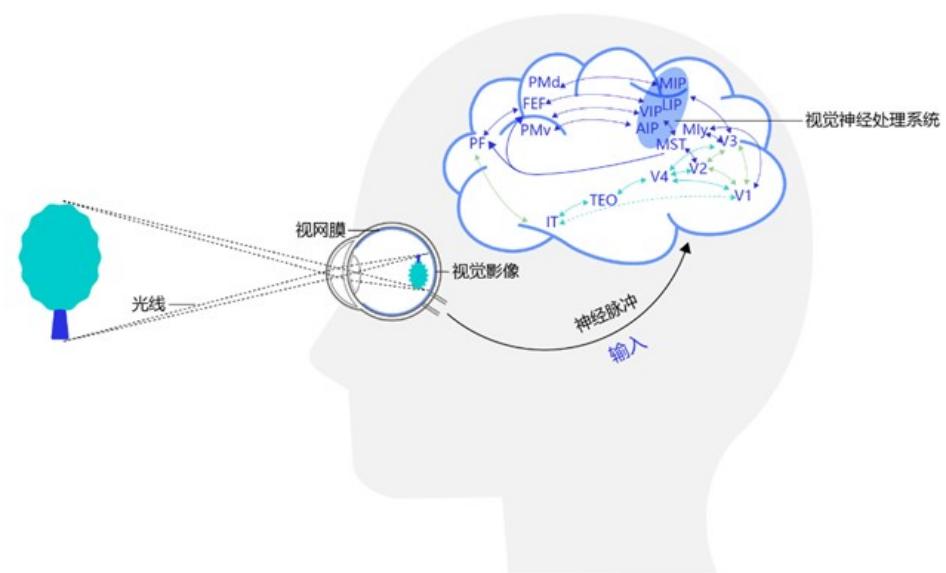
NLU—自然语言理解	NLG—自然语言生成
“读懂语言”	“写出语言”
情感分析、信息抽取	摘要、翻译、对话生成
输入 → 结构化信息	结构化信息 → 输出文本
判断、分类、抽取	生成、表达、创作

**NLU 负责输入理解，NLG 负责输出表达**

# 计算机视觉的定义与发展

## ➤ 定义

- **计算机视觉** (Computer Vision, CV) 旨在赋予计算机“看懂”**图像或视频内容的能力。**
- 它通过摄像头或传感器获取视觉数据，并利用算法对其进行**识别、检测、跟踪和描述**等任务



人类视觉感知

# 计算机视觉的应用

## ➤ 概览

### 典型的 计算机 视觉 任务

图像分类 Image Classification

目标检测 Object Detection

图像分割 Image Segmentation

目标跟踪 Object Tracking

图像生成 Image Generation

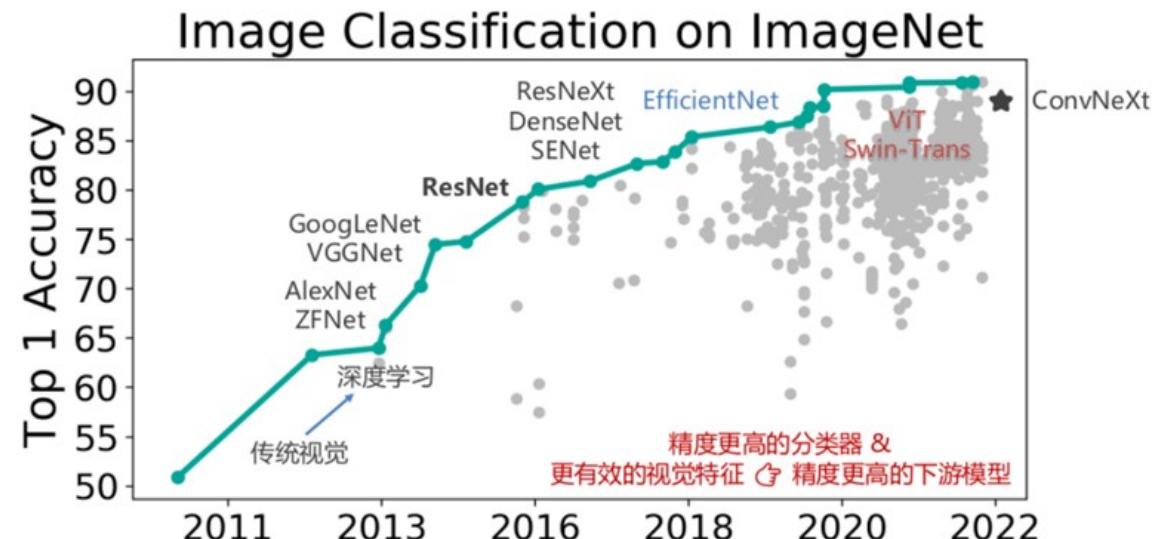
姿态估计 Pose Estimation

三维重建 3D Reconstruction

# 计算机视觉的应用

## ➤ 图像分类

- **定义**: 给定一张图片，预测其所属类别。如猫狗分类、垃圾分类、疾病图像诊断。
- **代表模型**: 基于**CNN**架构的各种模型，如ResNet、DenseNet、Vision Transformer (ViT)、ConvNeXt。

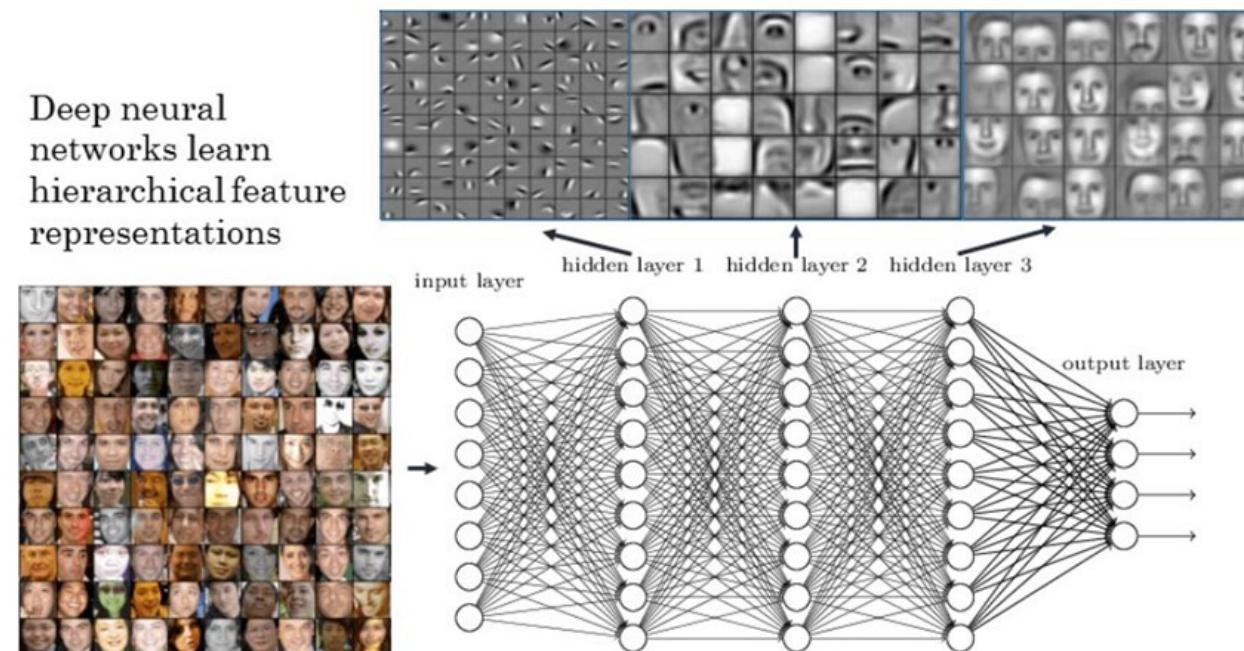


30

# 计算机视觉的应用

## ➤ 图像分类

- **工作流程**: 输入图像->特征提取->分类决策



# 计算机视觉的应用

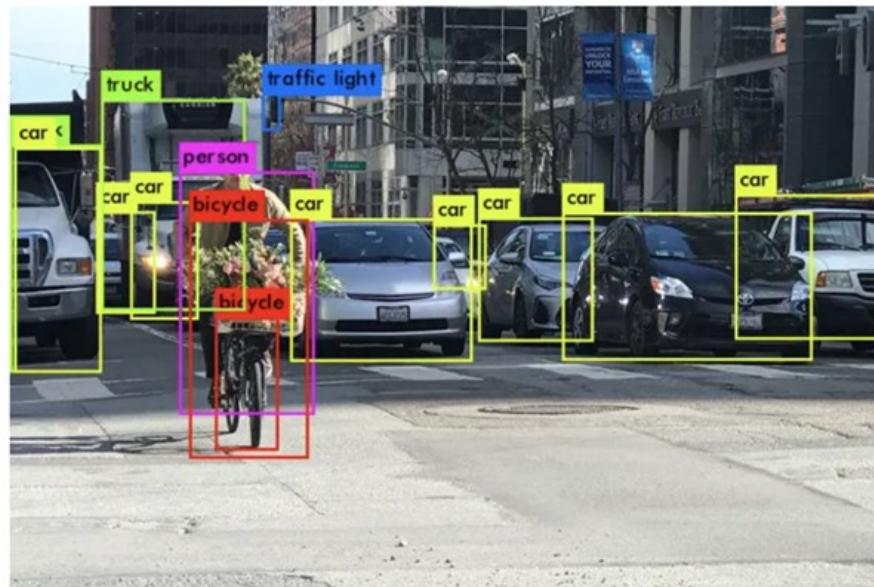
## ➤ 目标检测

- **定义：**识别图像中所有感兴趣的目标，并为每个目标同时给出其类别和位置（通常是边界框）。
- 与图像分类的区别：**多个目标、多种类别、多个位置。**

预测包含：

1. 边框位置  $(x, y, w, h)$ ；
2. 类别；
3. 置信度（表示模型认为“这个框里真的有一个目标”的概率大小。）

模型通过大量标注数据（框 + 类别）进行训练。



32

# 计算机视觉的应用

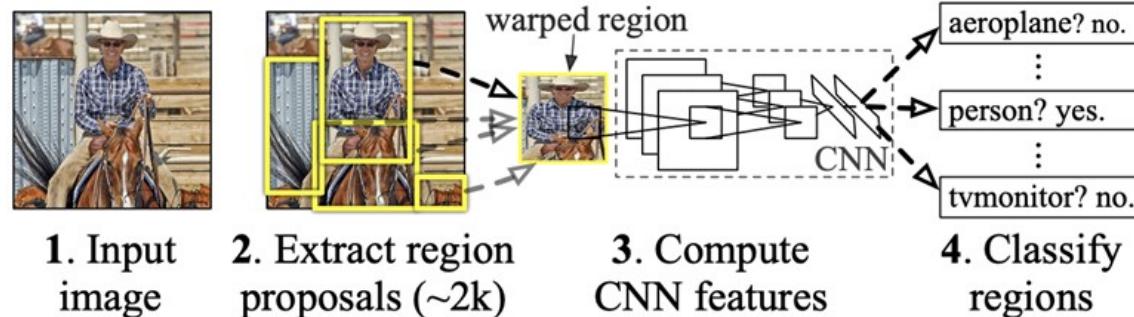
## ➤ 主流目标检测方法

- 大致分为：两阶段方法和一阶段方法。

### 两阶段方法（精度优先）

- 代表：R-CNN 系列 (R-CNN、Fast R-CNN、Faster R-CNN)
- 特点：精度高，速度慢。

R-CNN: Region-based Convolutional Network

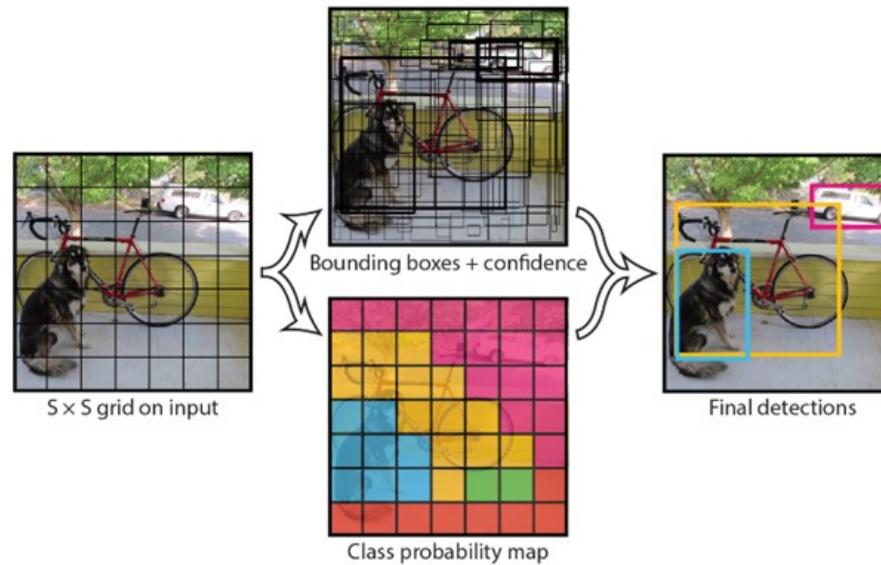


候选框生成 (Region Proposal) -> 每个候选框提特征->分类 + 生成边界框

# 计算机视觉的应用

## 一阶段方法 (速度优先)

- **代表:** YOLO (You Only Look Once) 、 SSD (Single Shot Detector)
- **特点:** 速度快，精度较低

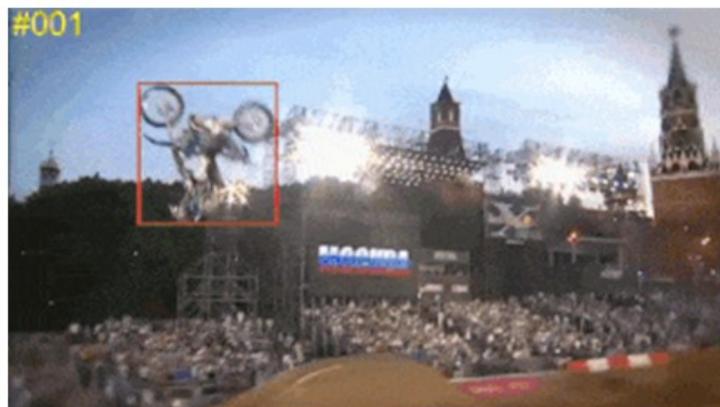


**YOLO:** 将图像划分为网格->每个网格直接预测边界框和类别

# 计算机视觉的应用

## ➤ 目标追踪

- **定义**: 在视频序列中，对指定的目标进行持续跟踪，输出该目标在每一帧中的位置（通常是边界框）。
- **特点**: 是从静态图像的目标检测延伸到动态视频的核心技术
- **代表**: Siamese、Deep SORT、ByteTrack。



# 计算机视觉的应用

## ➤ 任务分类

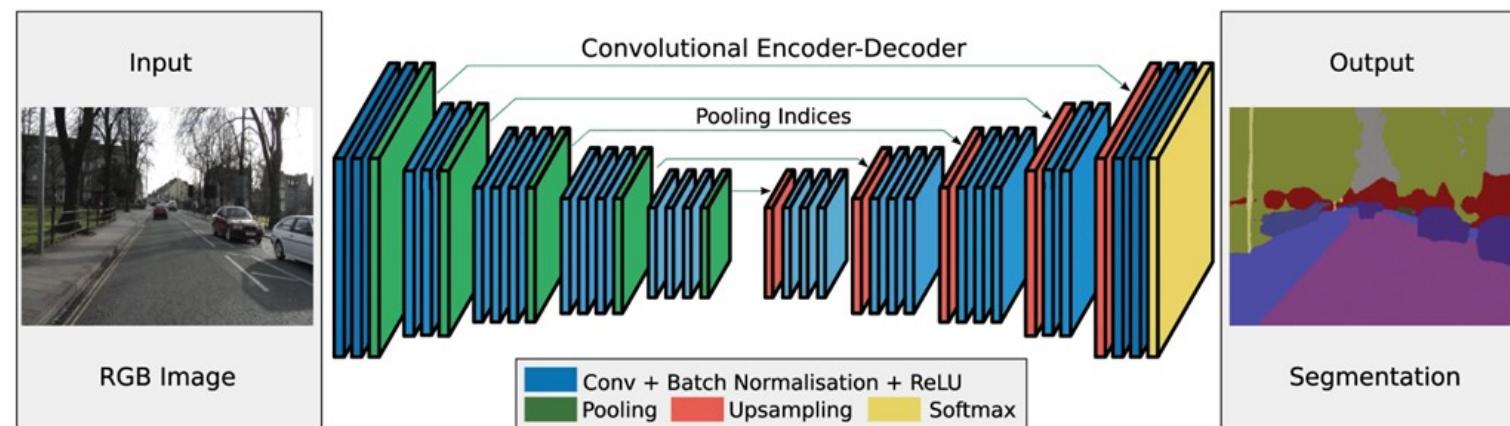
- **单目标追踪**: 给定一个目标，追踪这个目标的位置。
- **多目标追踪**: 追踪多个目标的位置。
- **MTMCT - 多目标多摄像头跟踪**: 跟踪多个摄像头拍摄的多人。
- **Person Re-ID - 行人重识别**: 判断图像或者视频序列中是否存在特定行人的技术。



# 计算机视觉的应用

## ➤ 图像分割

- **定义：**将图像划分为若干个有意义的区域 (segments)，每个区域对应一个语义类别或一个具体实例。
- **特点：**更精细的——不仅知道图像中有什么、在哪里，还知道每一个像素属于什么类别或对象。
- **基本结构：**编码器+解码器

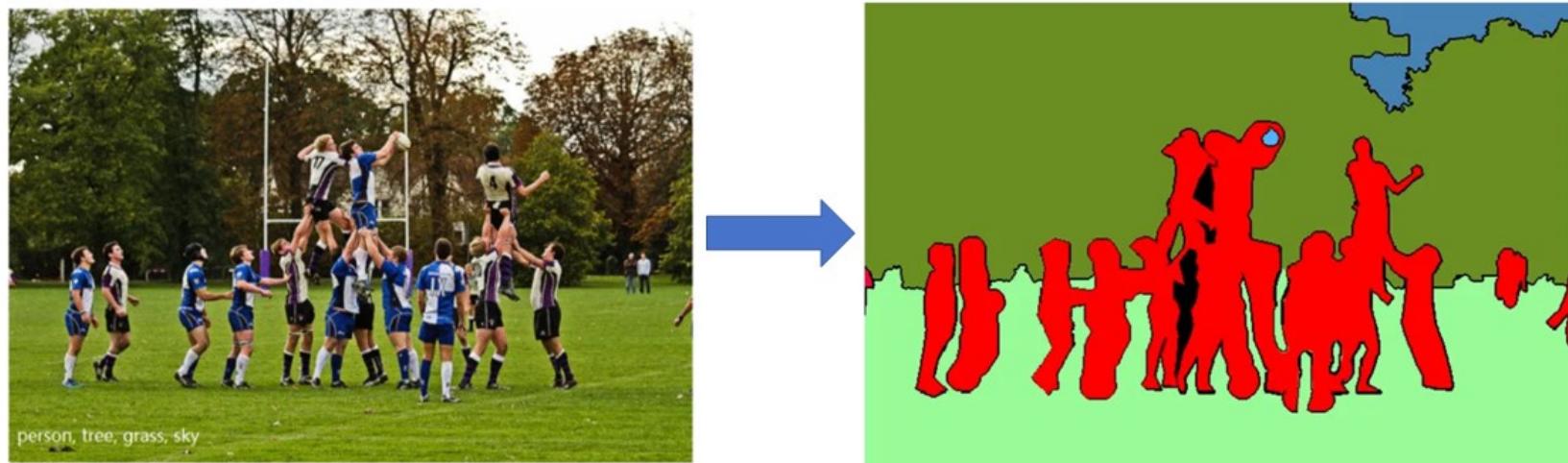


37

# 计算机视觉的应用

## ➤ 语义分割（Semantic segmentation）

- **定义：**将视觉输入分为不同的语义可解释类别，如“人”、“车”。不区分同类之间的个体。
- **代表：**U-Net、FCN、DeepLab



38

# 计算机视觉的应用

## ➤ 实例分割 (Instance Segmentation)

- **定义**: 在语义分割基础上，进一步区分同一类的不同个体
- **代表**: Mask R-CNN、SOLO、SAM。



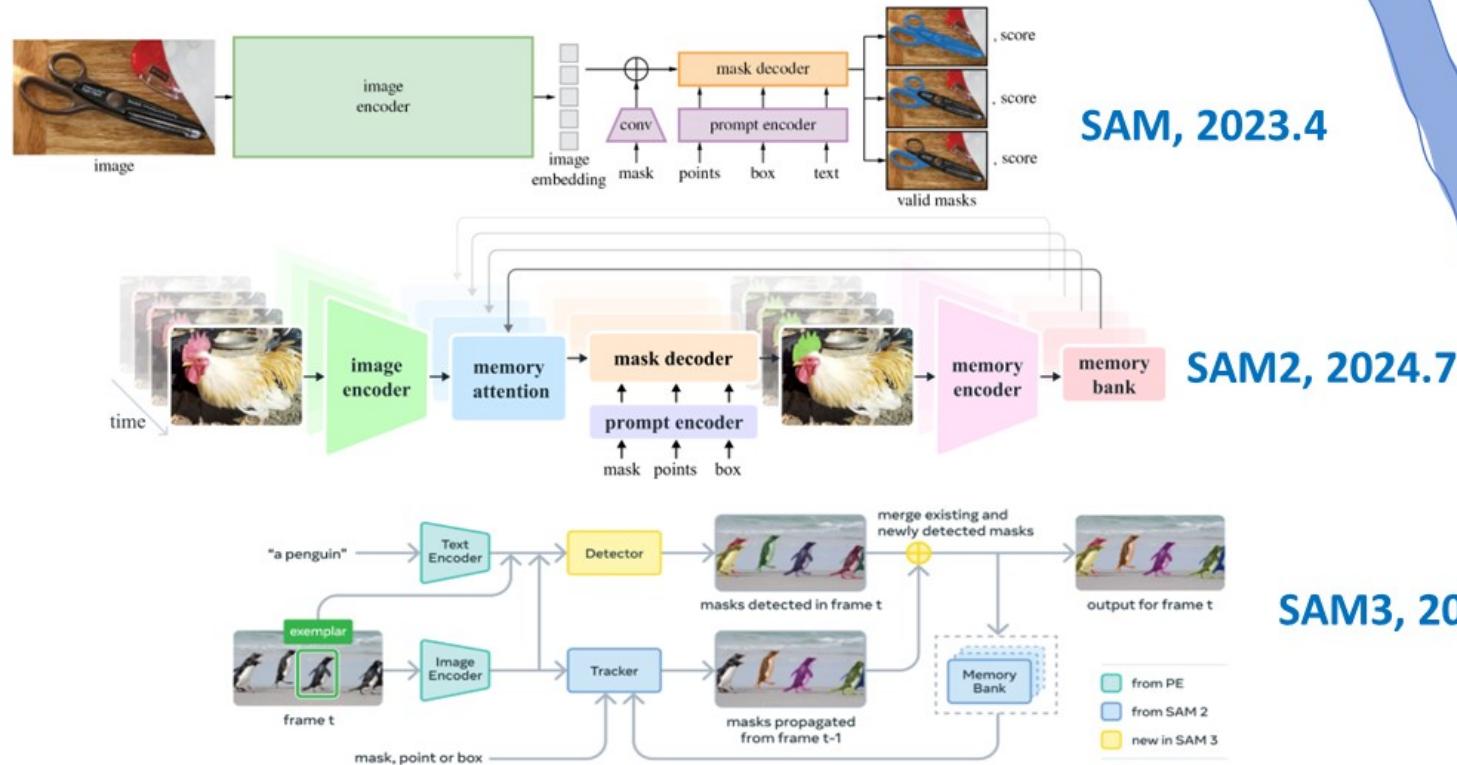
以人为目标  
→



# 计算机视觉的应用

Meta

## ➤ Segment Anything Model



40

# 计算机视觉的应用

## ➤ 姿态估计

- **定义：**从图像或视频中识别人或物体的**空间位置和关节/关键点坐标**，从而推断它们的姿势和动作。
- **代表模型：**OpenPose、HRNet、MMPose

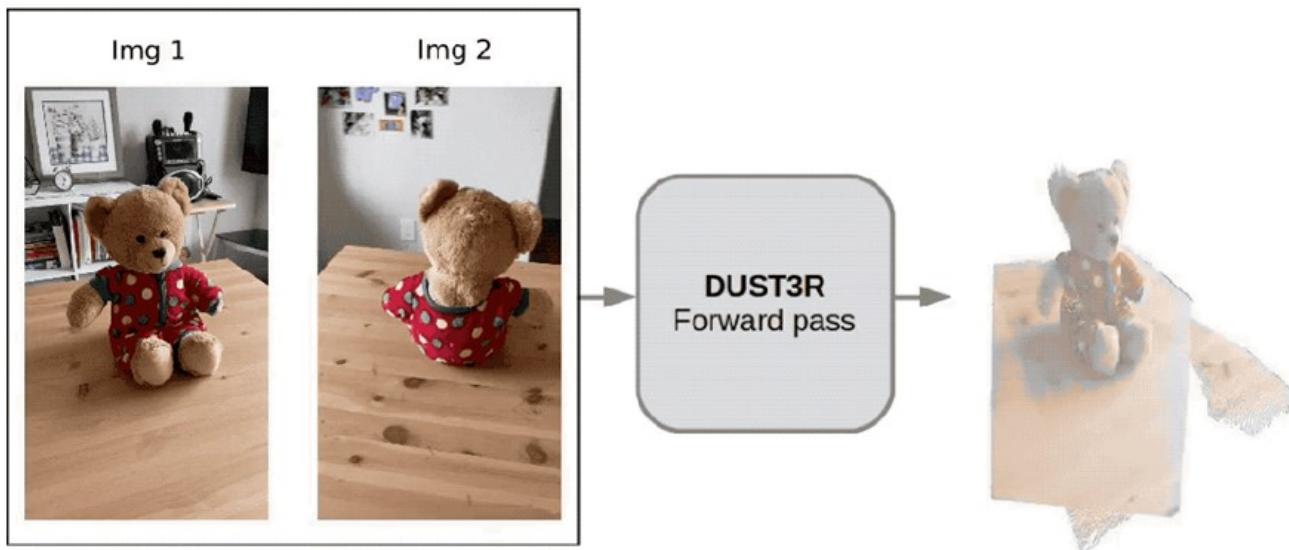


41

# 计算机视觉的应用

## ➤ 三维重建

- **定义：**利用二维数据（如照片、视频、激光扫描点云等）来恢复或生成物体、场景或环境的三维结构和外观的过程
- **代表模型：**DUST3R、Fast3R。



42

# 计算机视觉的应用

## ➤ 图像/视频生成

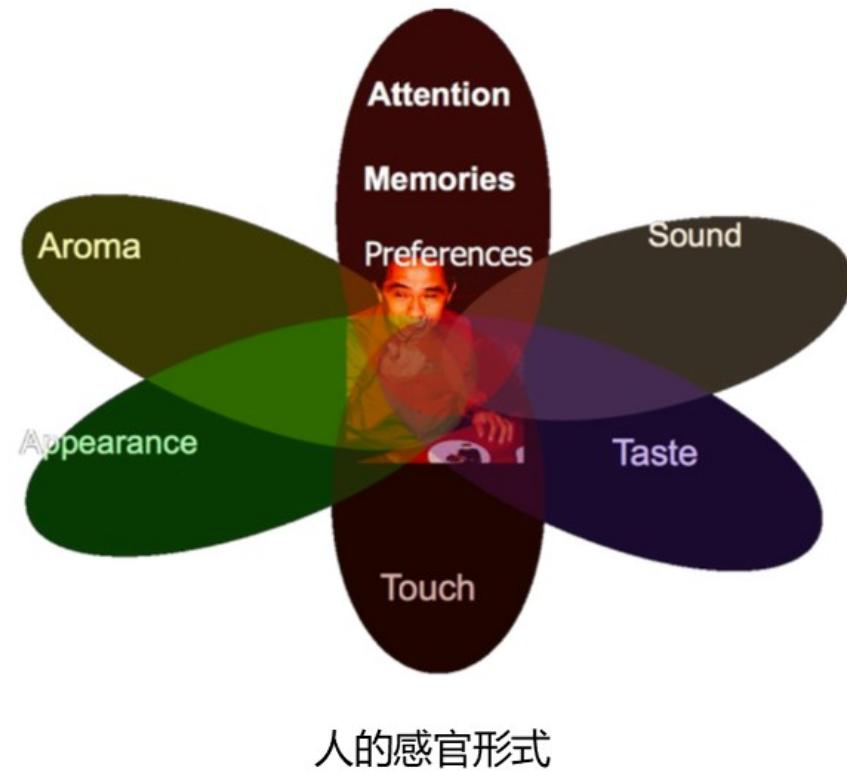
- **定义**: 根据输入图像, 生成新的图像。
- **代表模型**: GAN, CycleGAN, Diffusion Models



# 多模态的基本概念

## ➤ 模态

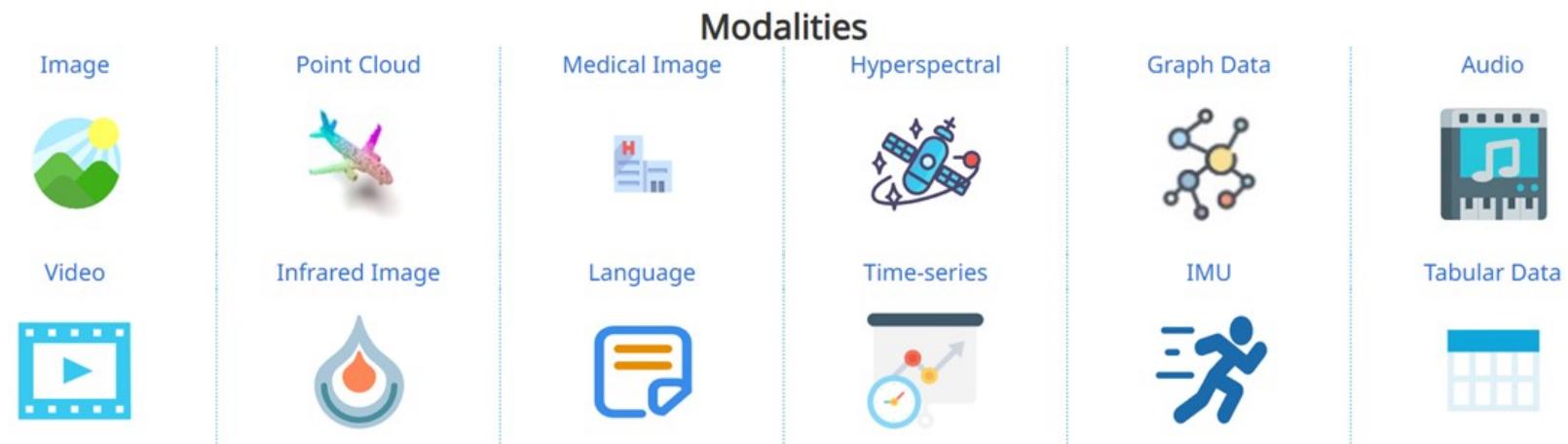
- 模态：表达或感知事物的方式，每一种信息的来源或者形式，都可以称为一种模态。
- 相较于图像、语音、文本等多媒体(Multi-media)数据划分形式，“模态”是一个更为细粒度的概念，同一媒介下可存在不同的模态。



# 多模态的基本概念

## ➤ 多模态

- **多模态 (Multimodality)** : 指利用来自不同感知通道 (如视觉、听觉、语言等) 的信息，以更全面地理解或表达对象或现象。

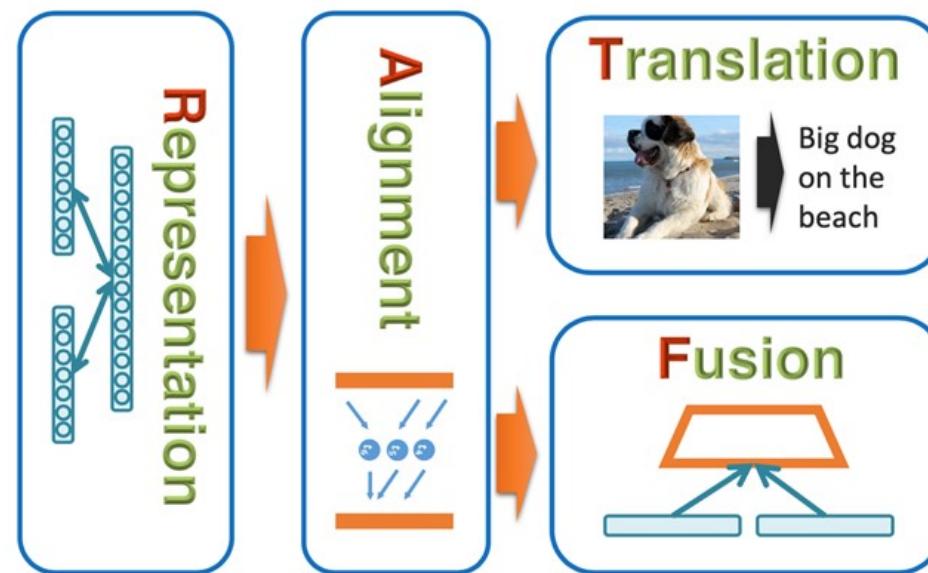


# 多模态的基本概念

## ➤ 多模态的核心挑战

由于不同模态间数据的异构性，  
多模态的研究存在以下四个核心  
挑战：

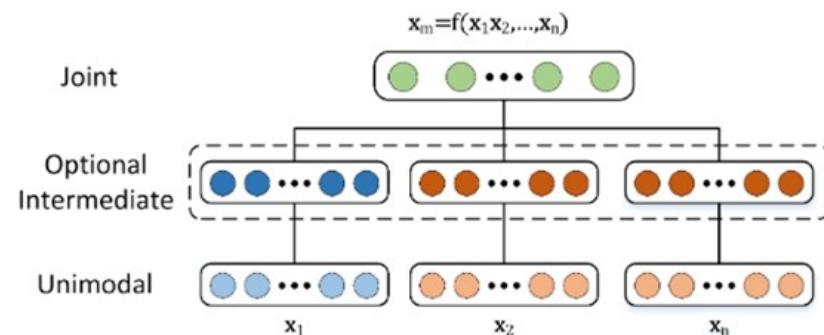
- 表征(Representation)
- 翻译(Translation)
- 对齐(Alignment)
- 融合(Fusion)



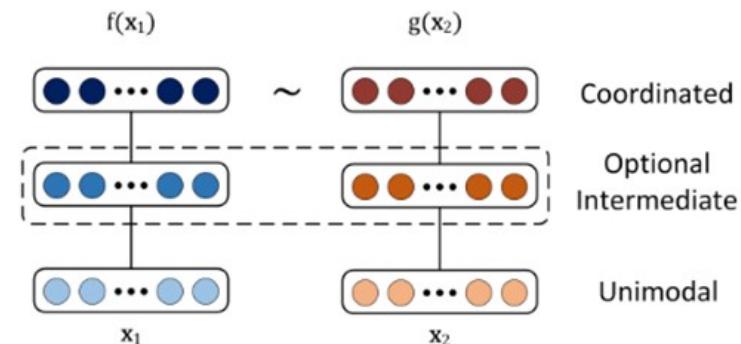
# 多模态的基本概念

## ➤ 表征(Representation)

- **定义**: 学习如何以充分利用多模态的互补性和冗余性的方式来**表征和总结**多模态数据，将原始数据表示成计算模型能够处理的格式。
- **分类**: 联合表征(Joint Representation)和协同表征(Coordinated Representation)



**联合表征捕捉互补性**  
将所有模态的信息一起映射到一个统一的向量空间

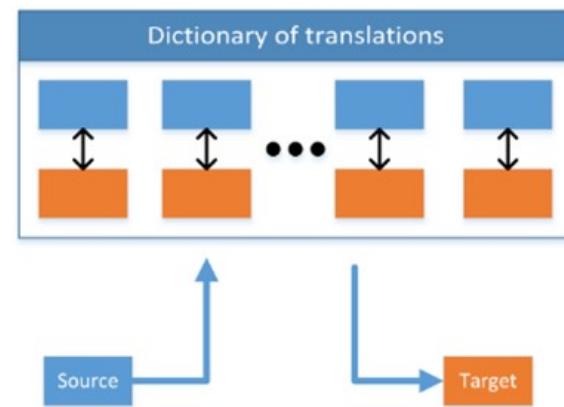


**协同表征建模相关性**  
将每个模态分别映射到各自的表示空间，通过  
**约束**（如线性相关）来协调表征后的向量。

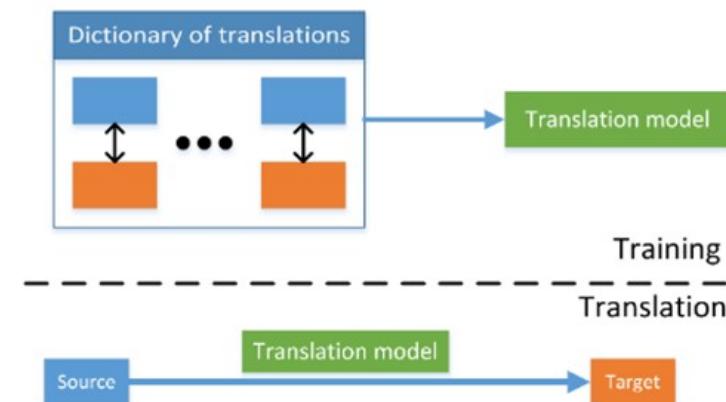
# 多模态的基本概念

## ➤ 翻译(Translation)

- **定义**: 将数据从一种模态**转换 (映射)** 到另一种模态。
- **举例**: 给定一张图片, 生成一个描述它的句子。
- **分类**: 基于实例的方法(Example-based) 和基于模型的方法(Models-based)。



基于实例的方法从词典中检索最佳翻译

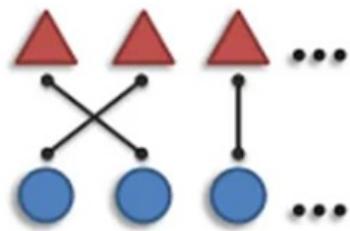


基于模型的方法使用模型进行翻译

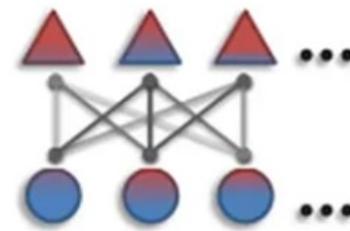
# 多模态的基本概念

## ➤ 对齐(Alignment)

- **定义**: 识别来自两个或多个不同模态的(子)元素之间的**直接关系**。
- **举例**: 将菜谱中的步骤与展示该菜肴制作过程的视频对齐。
- **分类**: 显示对齐 (Explicit Alignment), 隐式对齐 (Implicit Alignment)。



**显示对齐**: 直接建立不同模态之间的对应关系



**隐式对齐**: 通过模型内部机制隐式地实现跨模态的对齐。

# 多模态的基本概念

## ➤ 融合(Fusion)

- **定义**: 将来自两种或多种模态的信息融合起来进行预测。
- **例子**: 在视听语音识别中, 唇部运动的视觉描述与语音信号融合, 以预测口语词汇。

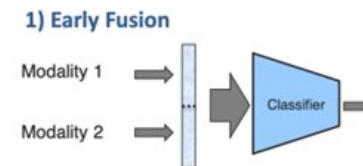
## ➤ 非模型的方法

- 早期融合 (基于特征)
- 后期融合 (基于决策)

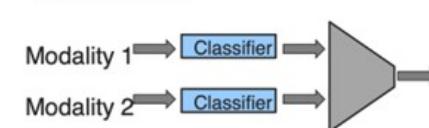
## ➤ 基于模型的方法

- 神经网络方法
- 多核学习
- 图模型

### A Model-Agnostic Approaches

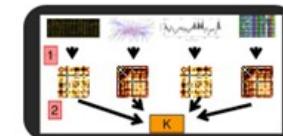


### 2) Late Fusion

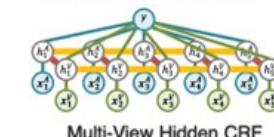


### B Model-Based (Intermediate) Approaches

- 1) Deep neural networks
- 2) Kernel-based methods
- 3) Graphical models



Multiple kernel learning



Multi-View Hidden CRF

# 多模态的应用

## ➤ 图文检索(Image-text retrieval)

- **定义:** 在图像与文本之间实现跨模态匹配与检索。
- **代表:** CLIP、Chinese-CLIP、BLIP



Openi图片检索文献



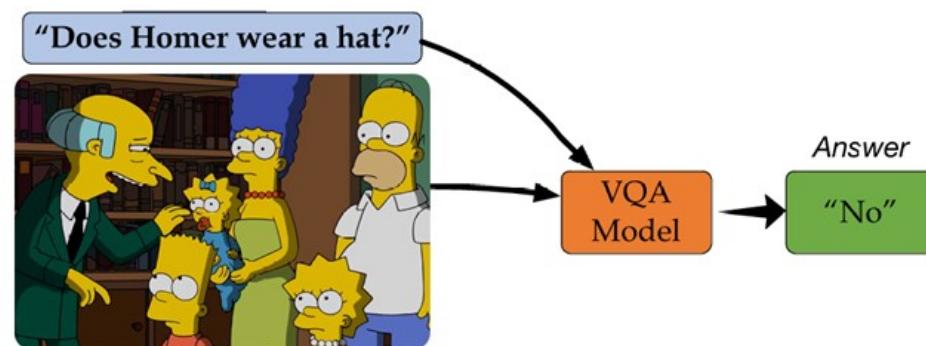
手机相册文搜图

52

# 多模态的应用

## ➤ 视觉问答(Visual Question Answering)

- **定义**: 输入图像 (视频) + 自然语言问题, 输出自然语言答案。
- **代表**: LXMERT、ViLT、MiniGPT-4
- **应用**: 医学图像问答、教育辅助系统、辅助视觉障碍人士。

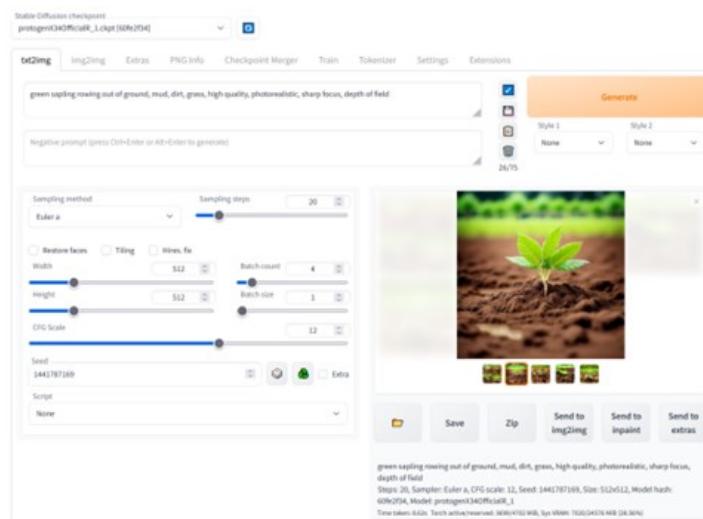


**Be My Eyes:** 实时视觉辅助

# 多模态的应用

## ➤ 文生图(Text-to-Image)

- **定义:** 输入一段文字, 生成相应的图像
- **代表:** Stable Diffusion, DALL·E, Nano banana



stable-diffusion-webui



普通的P图软件，真的听不懂“如果……会怎样？”



## 多模态的应用

### ➤ 文生视频(Text-to-Video)

- **定义**: 输入一段文字, 生成相应的视频。
- **代表**: Make-A-Video(Meta,2022)、Sora(OpenAI,2024)、Wan(Alibaba,2025)



Make-A-Video: 猫手里拿着遥控器看电视



Sora: 两艘海盗船在一杯咖啡中航行时相互争斗的逼真特写视频。

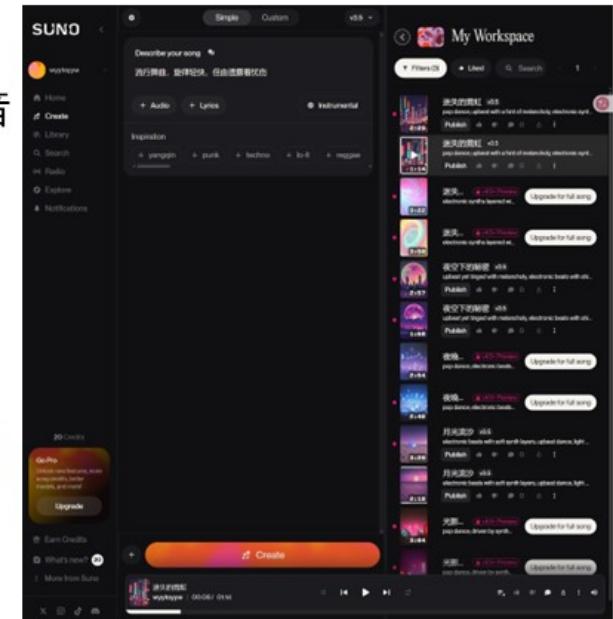
55

# 多模态的应用

## ➤ 文生音乐(Text-to-Music)

- **定义**: 通过输入一段文本描述 (prompt) , 自动生成对应的音乐片段或完整音乐作品。文本可以描述风格、情感、乐器类型、节奏、环境等内容。
- **代表**: MusicGen(Meta), MusicLM(Google), Suno AI

	MusicGen	MusicGen Stereo	MusicLM	Riffusion	Musai
流行舞曲，旋律动听，节奏明快，非常适合海滩					



Suno AI: 流行舞曲，旋律轻快，但又透露着忧伤



# 多模态的应用

## ➤ 语音合成(Text-to-Speech)

- **定义:** 将输入的文本自动转换为模拟人类语音输出
- **代表:** MinMax Audio、ElevenLabs、微软Azure TTS、Google TTS

Text to speak:  
在人工智能导论这门课程中，你们可以学到人工智能领域的最新技术以及应用

Language / locale  
普通话 (中国大陆) Chirp 3: HD Voices  
cmn-CN-Chirp3-HD-Achernar

Audio device profile  
Small home speaker Speed: 1.00 Pitch: 0.00

Show JSON ▾ REPLAY

Google TTS



MINIMAX 语音 对话 视频 Agent

让文字栩栩如“声”

语音合成 音乐创作

在文本输入文字，一键生成声音角色，或有质感的语音作品！

语音合成 音乐创作

speech-02-hd 音乐创作

语音工具箱 人声提取 人物声音  
音频处理 人物声音  
纯净人声

角色库 不同角色 人物声音 语音  
黄鹂鸟声 人物声音 语音  
江南古典 人物声音 语音



MinMax Audio

# 多模态的应用

## ➤ 语音识别(Speech-to-Text, Automatic Speech Recognition)

- **定义**: 将语音信号自动转换为对应文字。
- **代表**: Whisper (OpenAI) 、 Wav2Vec2 (Facebook)
- **应用**: Siri、在线会议转写、自动字幕生成



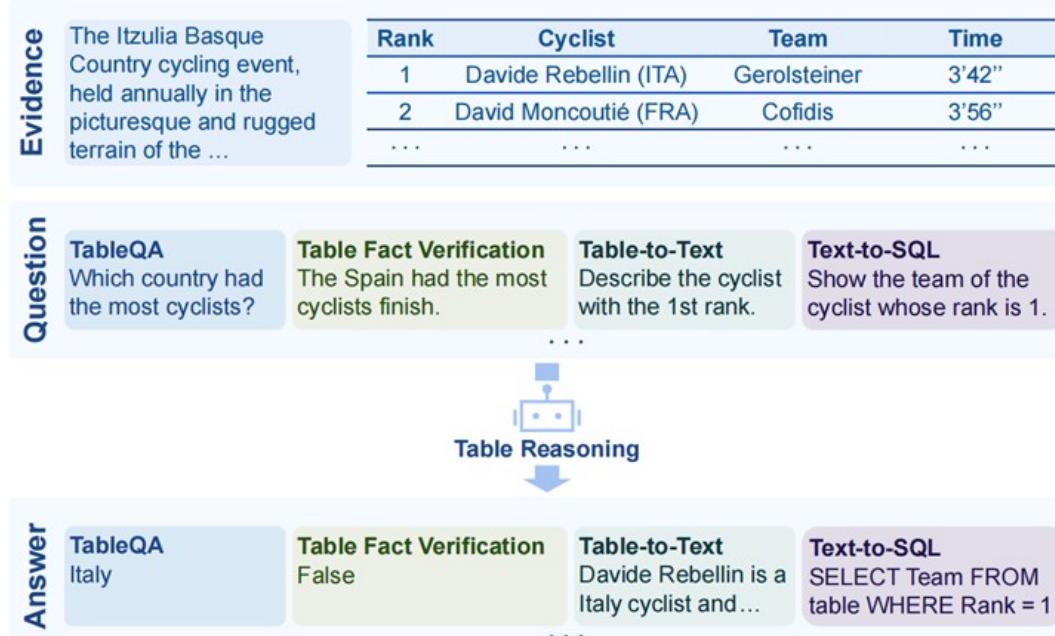
腾讯会议自动转写



Siri语音助手

# 多模态的应用

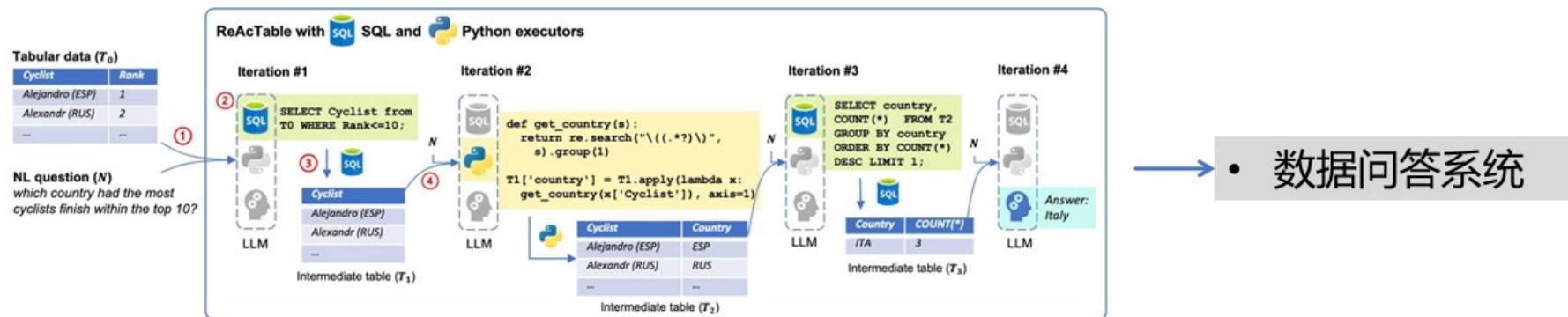
## ➤ 表格理解：让模型理解表格内容并推理



- 任务包括：
  - 表格问答
  - 表格事实判断
  - 表格文本转换
  - 表格SQL查询

# 多模态的应用

## ➤ 应用



Product ID	Product Name	Price (USD)	Quantity in Stock	Category
P001	iPhone 12	799	50	Electronics
P002	Galaxy S21	999	30	Electronics
P003	MacBook Pro	1,299	15	Computers
P004	Apple Watch	399	100	Accessories

## • 财务分析自动化

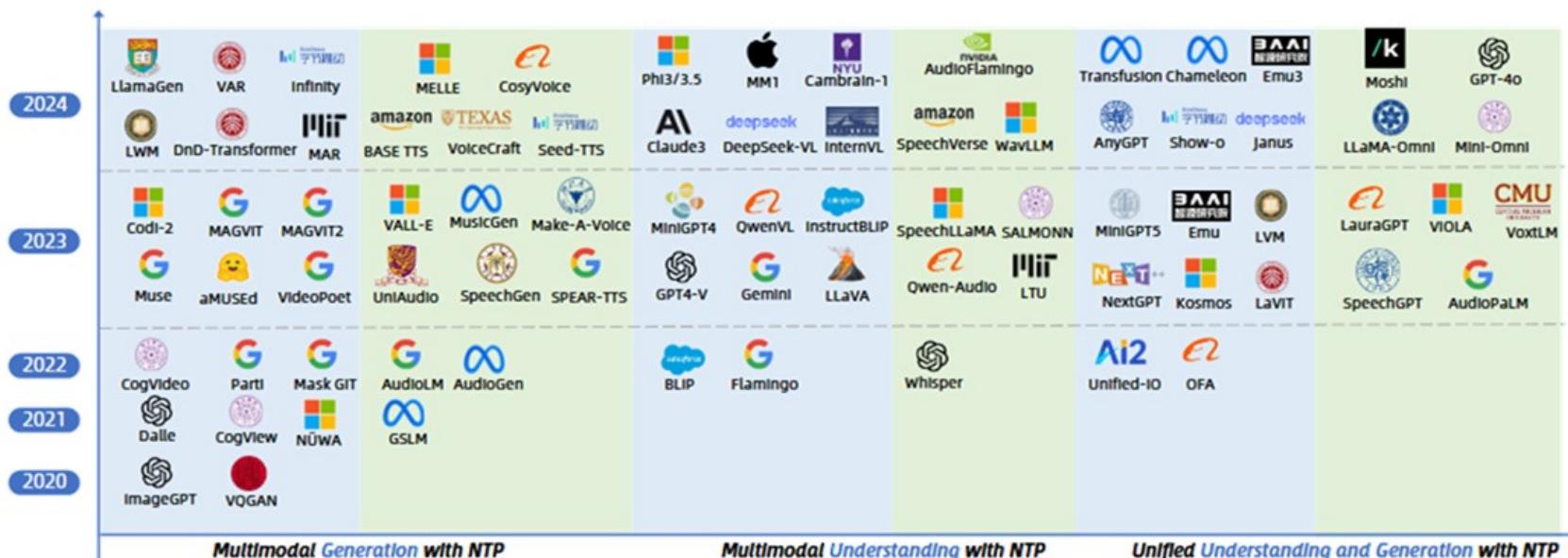
相关问题-答案对示例 3

1. 问题: "P001" 产品的价格是多少?  
答案: 799
2. 问题: 电子产品类中哪个产品库存最少?  
答案: MacBook Pro
3. 问题: "Accessories" 类别的产品平均价格是多少?  
答案: 399 (这个问题的答案需要模型进行计算, 即  $(399 * 100) / 100$ )
4. 问题: 电子产品类中价格超过1000美元的产品有哪些?  
答案: Galaxy S21, MacBook Pro

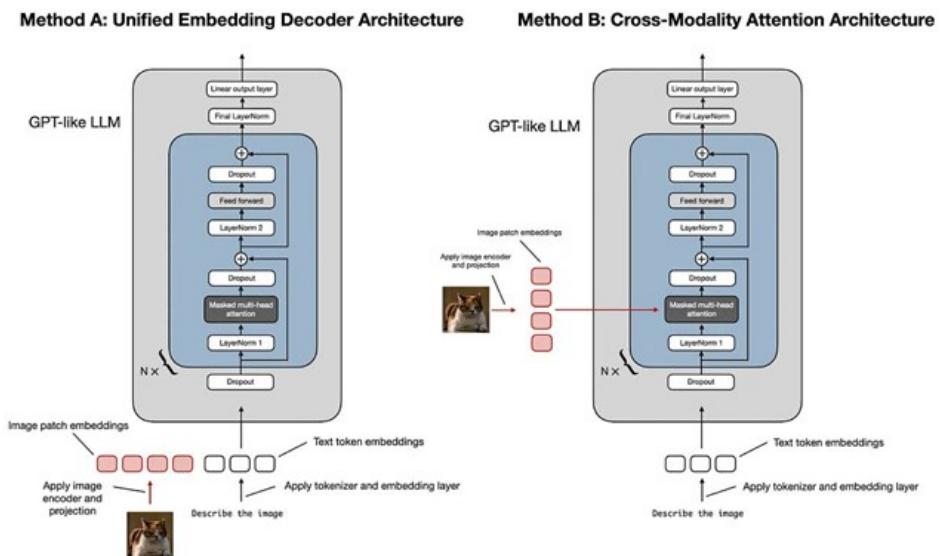
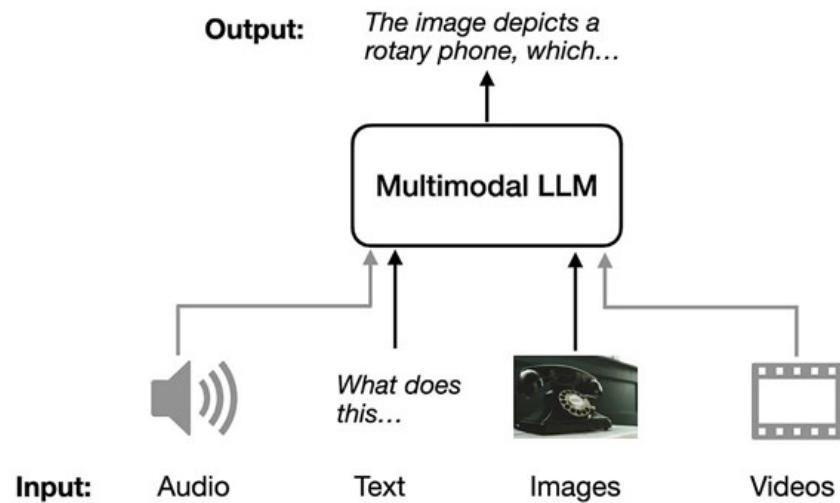
# 通用多模态大模型

## ➤ 输入输出模态的任意组合

- 兼顾**理解与生成**: 即可以输入文字、图片、语音等多种形式, 模型能理解并给出相应的多模态输出(比如文字回复、图像生成、音频等)。



# 通用多模态大模型



方法 A：统一嵌入解码器架构方法；  
方法 B：跨模态注意力架构方法。

# World Model

世界模型的核心：一个能够理解世界如何运作，并预测起未来变化的系统。



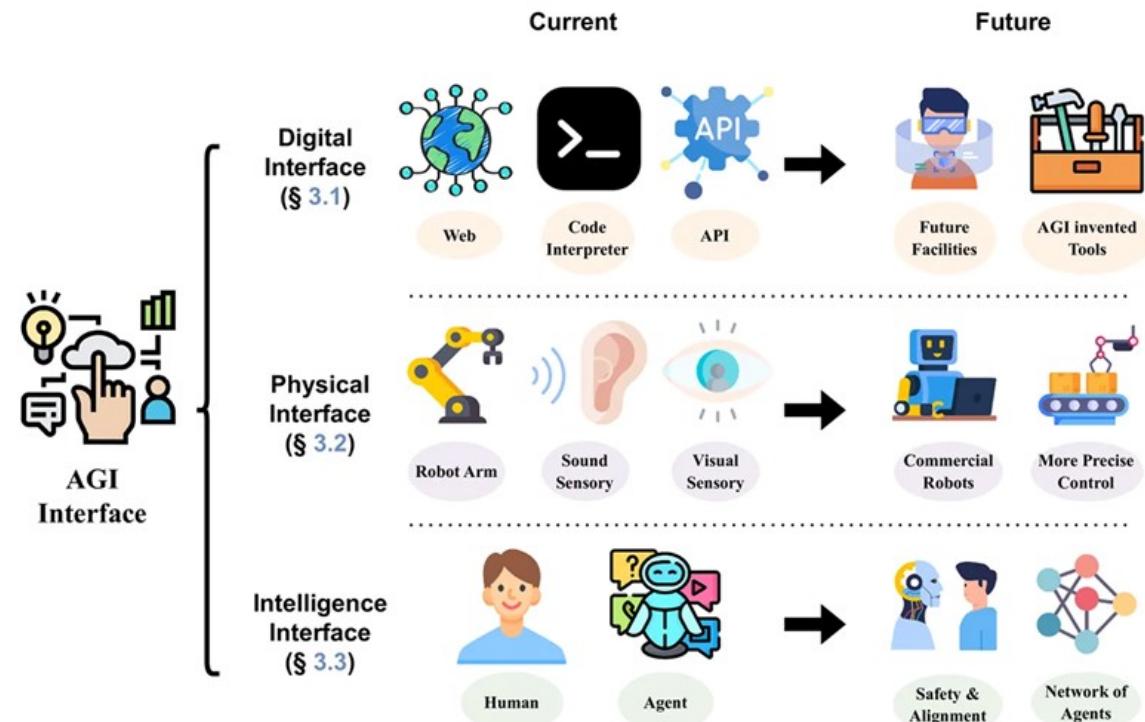
强调“让 AI 自己学习世界的物理与因果规律”，主张**自监督学习 + 能行动的智能体**。



强调“视觉是基础，通过感知数据（尤其是图像）让 AI 理解世界”，主张**通过丰富的感觉经验构建世界知识**。



# AGI



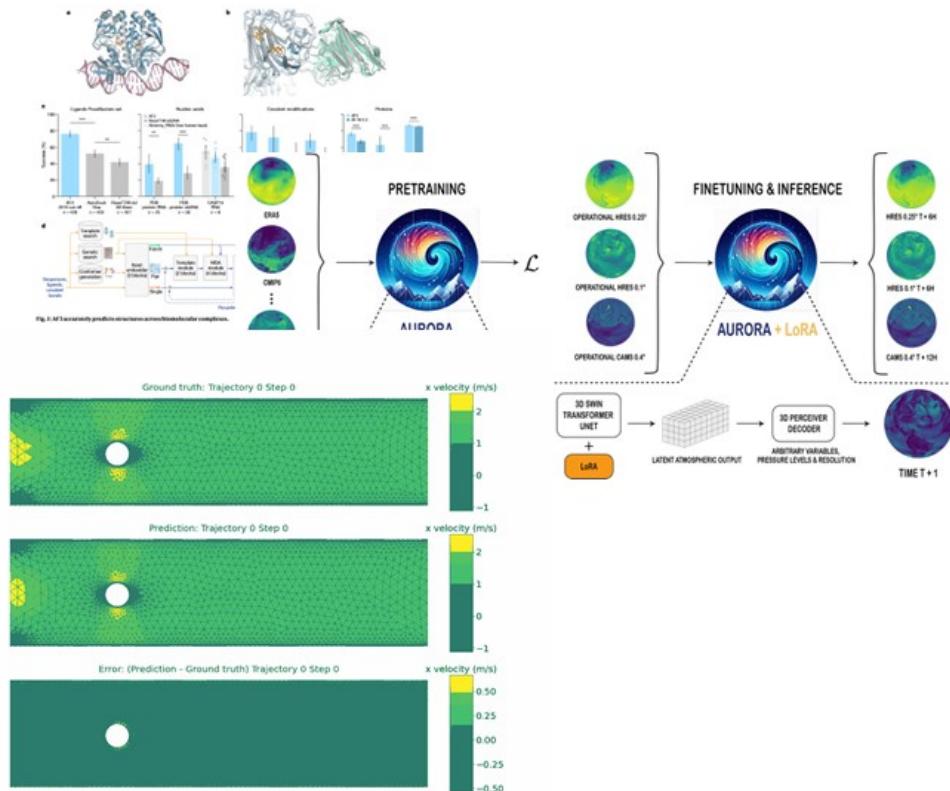
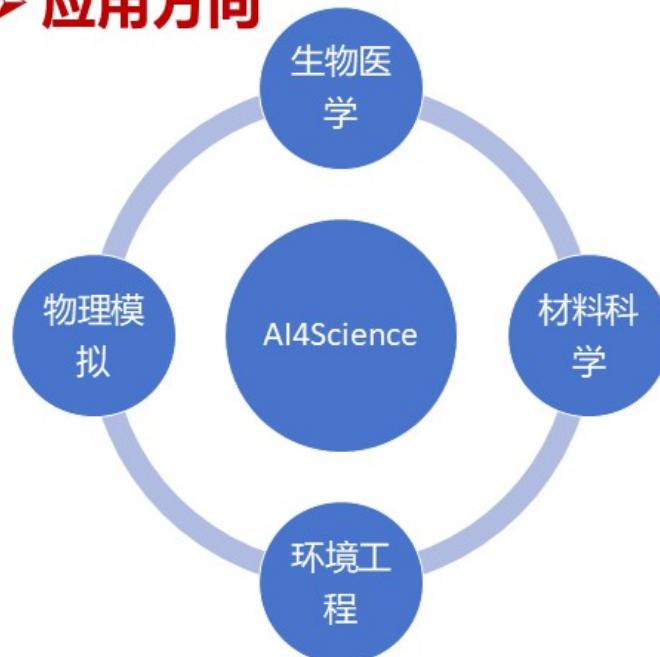
通用人工智能 (AGI, Artificial General Intelligence) 是指旨在处理多个和多样化任务的 AI 系统。与仅专注于一个领域不同，AGI 系统可以跨各种环境、情况和挑战无缝地学习和调整其知识。

# AI for Science

## ➤ For what?

- 利用AI (如LLM) 加速科学发现

## ➤ 应用方向



65