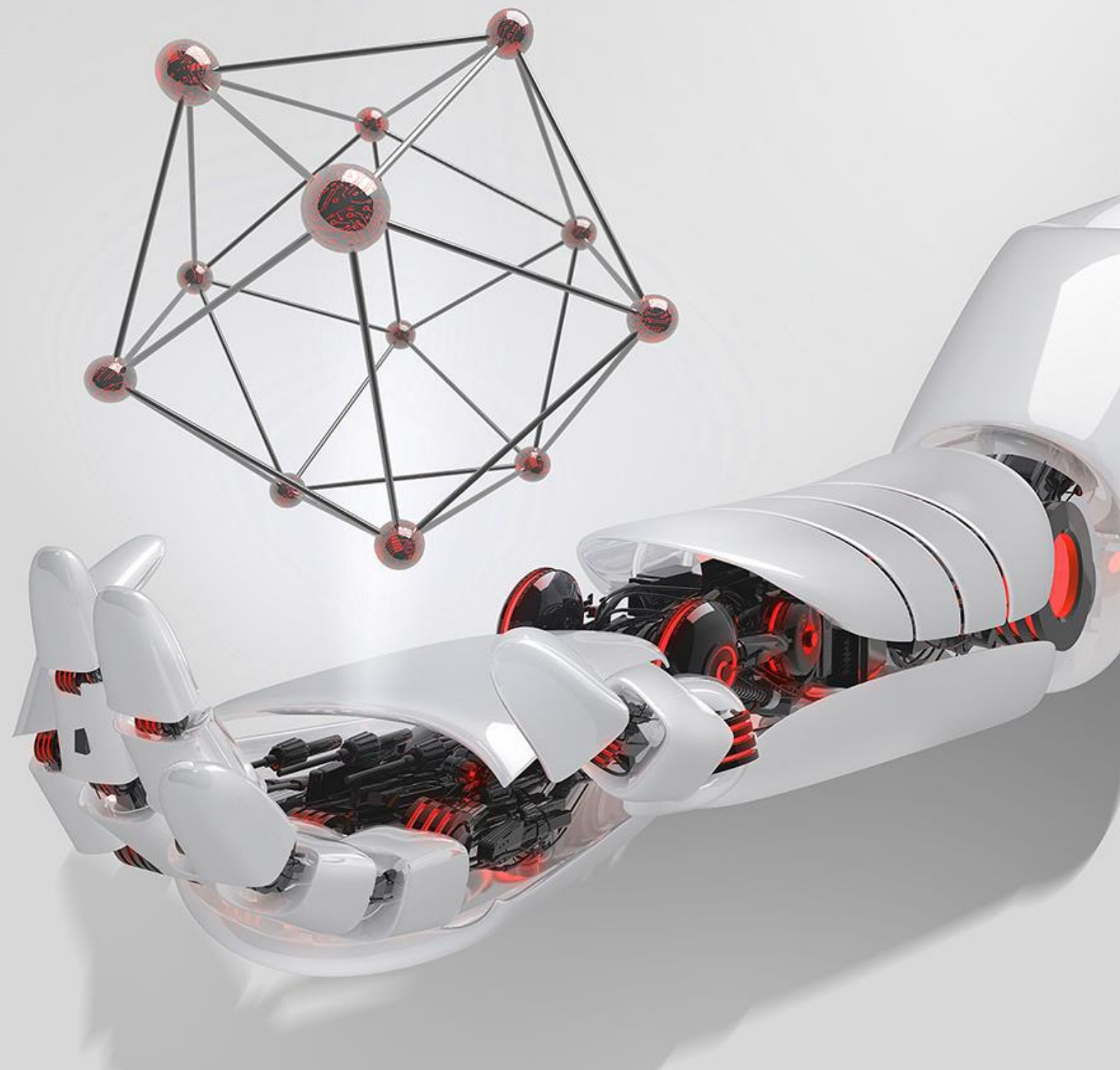


# 视觉通用人工智能年度进展报告



报告人：谢凌曦

2024年5月5日 @ VALSE重庆



# 大语言模型带来了通用人工智能的无限可能

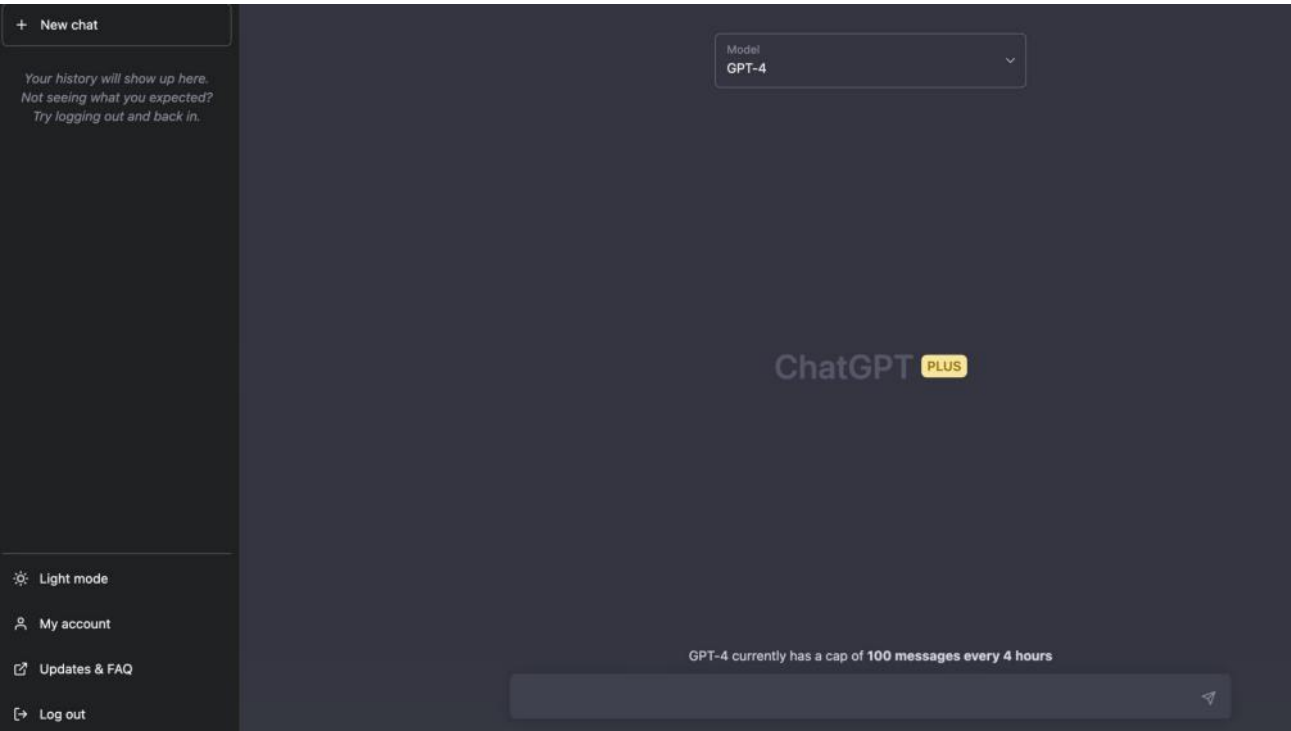
## Sparks of Artificial General Intelligence: Early experiments with GPT-4

Sébastien Bubeck    Varun Chandrasekaran    Ronen Eldan    Johannes Gehrke  
Eric Horvitz    Ece Kamar    Peter Lee    Yin Tat Lee    Yuezhi Li    Scott Lundberg  
Harsha Nori    Hamid Palangi    Marco Tulio Ribeiro    Yi Zhang  
  
Microsoft Research

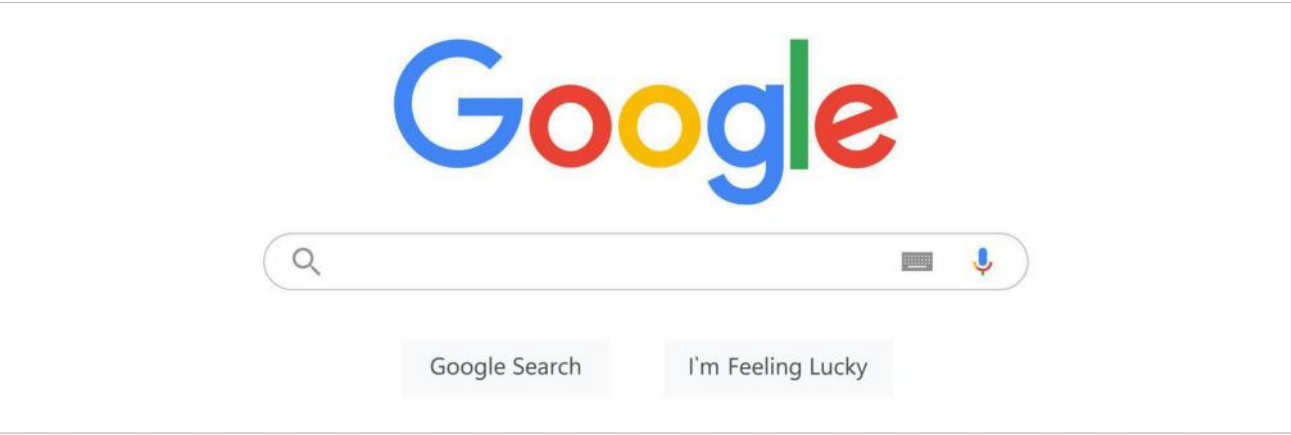
ChatGPT系列被称为“通用人工智能的火花”

Simulated exams	GPT-4 estimated percentile	GPT-4 (no vision) estimated percentile	GPT-3.5 estimated percentile
Uniform Bar Exam (MBE+MEE+MPT) <sup>1</sup>	298/400 ~90th	298/400 ~90th	213/400 ~10th
LSAT	163 ~88th	161 ~83rd	149 ~40th
SAT Evidence-Based Reading & Writing	710/800 ~93rd	710/800 ~93rd	670/800 ~87th
SAT Math	700/800 ~89th	690/800 ~89th	590/800 ~70th
Graduate Record Examination (GRE) Quantitative	163/170 ~80th	157/170 ~62nd	147/170 ~25th
Graduate Record Examination (GRE) Verbal	169/170 ~99th	165/170 ~96th	154/170 ~63rd
Graduate Record Examination (GRE) Writing	4/6 ~54th	4/6 ~54th	4/6 ~54th
USABO Semifinal Exam 2020	87/150 99th~100th	87/150 99th~100th	43/150 31st~33rd
USNCO Local Section Exam 2022	36/60	38/60	24/60
Medical Knowledge Self-Assessment Program	75%	75%	53%
Codeforces Rating	392 below 5th	392 below 5th	260 below 5th
AP Art History	5 86th~100th	5 86th~100th	5 86th~100th
AP Biology	5 85th~100th	5 85th~100th	4 62nd~85th
AP Calculus BC	4 43rd~59th	4 43rd~59th	1 0th~7th

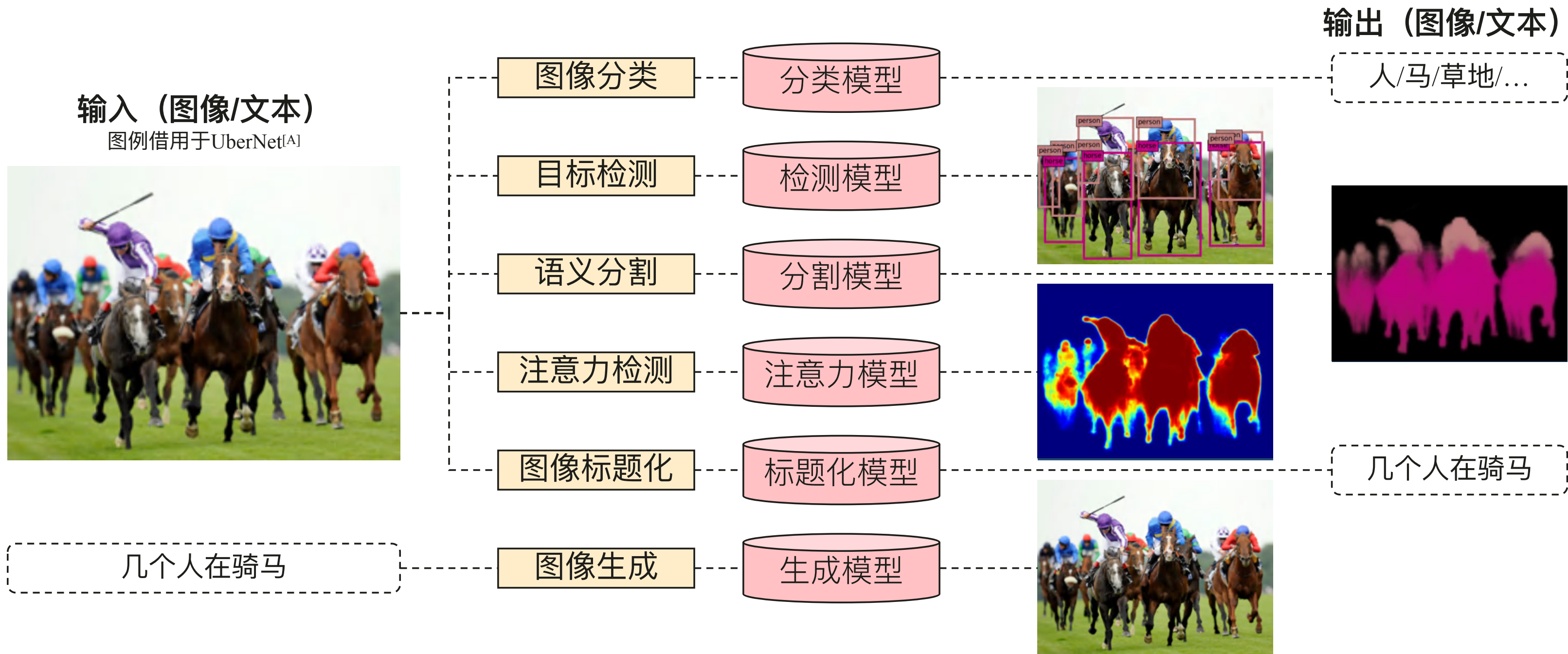
GPT-4在一系列考试中表现优异  
[A] S. Bubeck et al., Sparks of Artificial General Intelligence: Early experiments with GPT-4, in *arXiv preprint:2303.12712*, 2023.



ChatGPT的界面（上）和Google的界面（下）



# 而计算机视觉领域仍未看到通用人工智能的曙光



# 本报告希望回答的两个问题

---

为什么



怎么做

在通用人工智能的发展上，  
计算机视觉领域为何**落后于**  
自然语言处理领域？

为了**接近**通用人工智能，计  
算机视觉领域需要做对哪些  
重要的事？



# 通用人工智能的定义

- 概念性的定义<sup>[A]</sup>

- 通用人工智能，是能够完成一切人类和动物能够完成任务的计算机程序

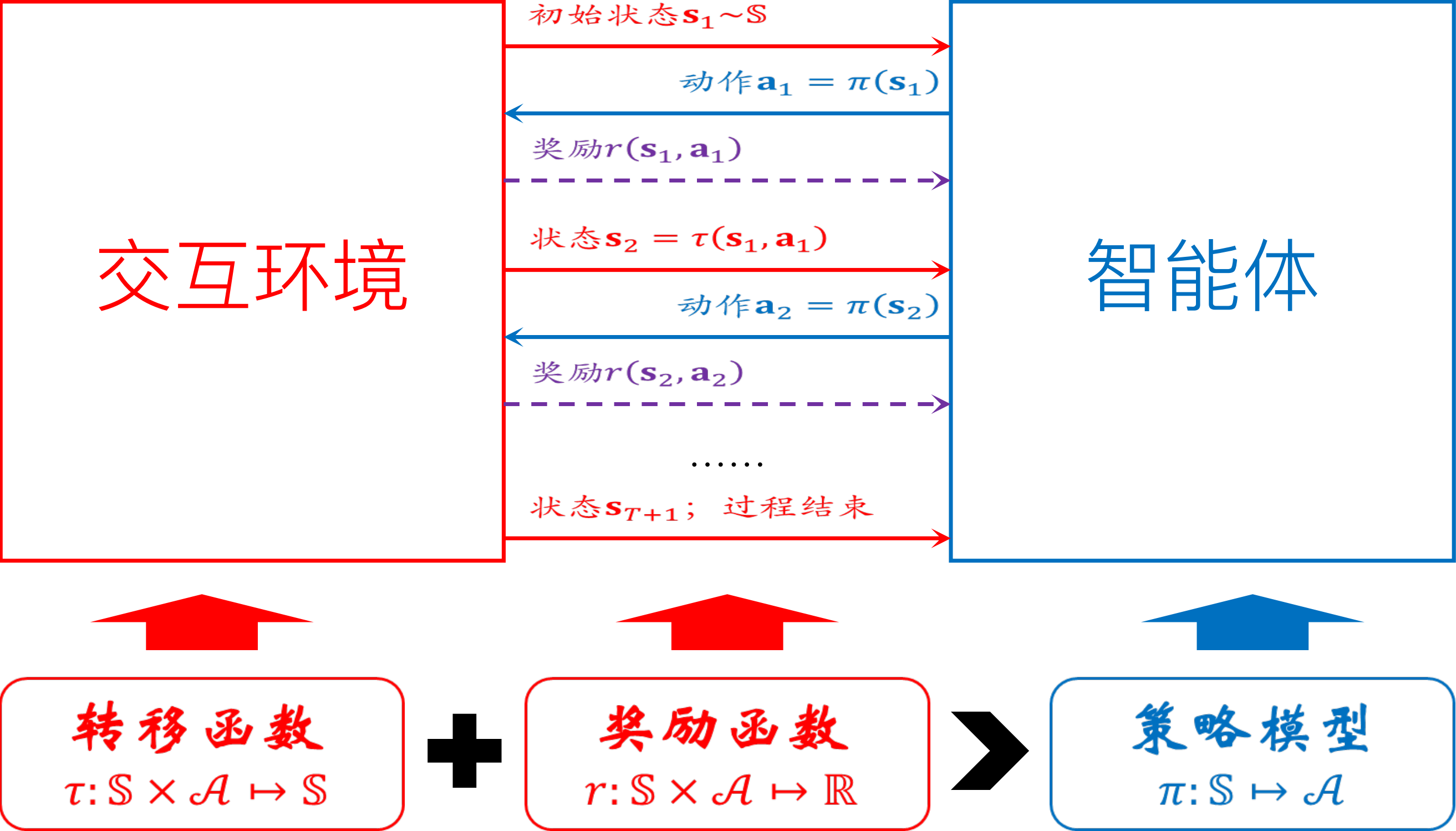
- 形式化的定义<sup>[B]</sup>

- 给定状态空间 $\mathcal{S}$ 、动作集合 $\mathcal{A}$ ，并且定义转移函数 $\tau: \mathcal{S} \times \mathcal{A} \mapsto \mathcal{S}$ 、奖励函数 $r: \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}$ ，通用人工智能的目标是寻找策略模型 $\pi: \mathcal{S} \mapsto \mathcal{A}$ ，使得过程期望累计奖励 $\bar{R} = \mathbb{E}[\sum_{t=1}^T r(\mathbf{s}_t, \pi(\mathbf{s}_t))]$ 最大化
  - 过去十年间，至少有三类现象级的人工智能算法，可以用上述定义解释

算法	状态	动作	转移函数	奖励函数	策略模型训练
AlphaGo	棋盘局面	落子	自然形成下一状态	最终输赢	深度学习+强化学习
AlphaStar	游戏局面	游戏操作	自然形成下一状态	最终输赢	深度学习+强化学习
ChatGPT	对话历史	生成token	自然形成下一状态	用户评分	深度学习+强化学习

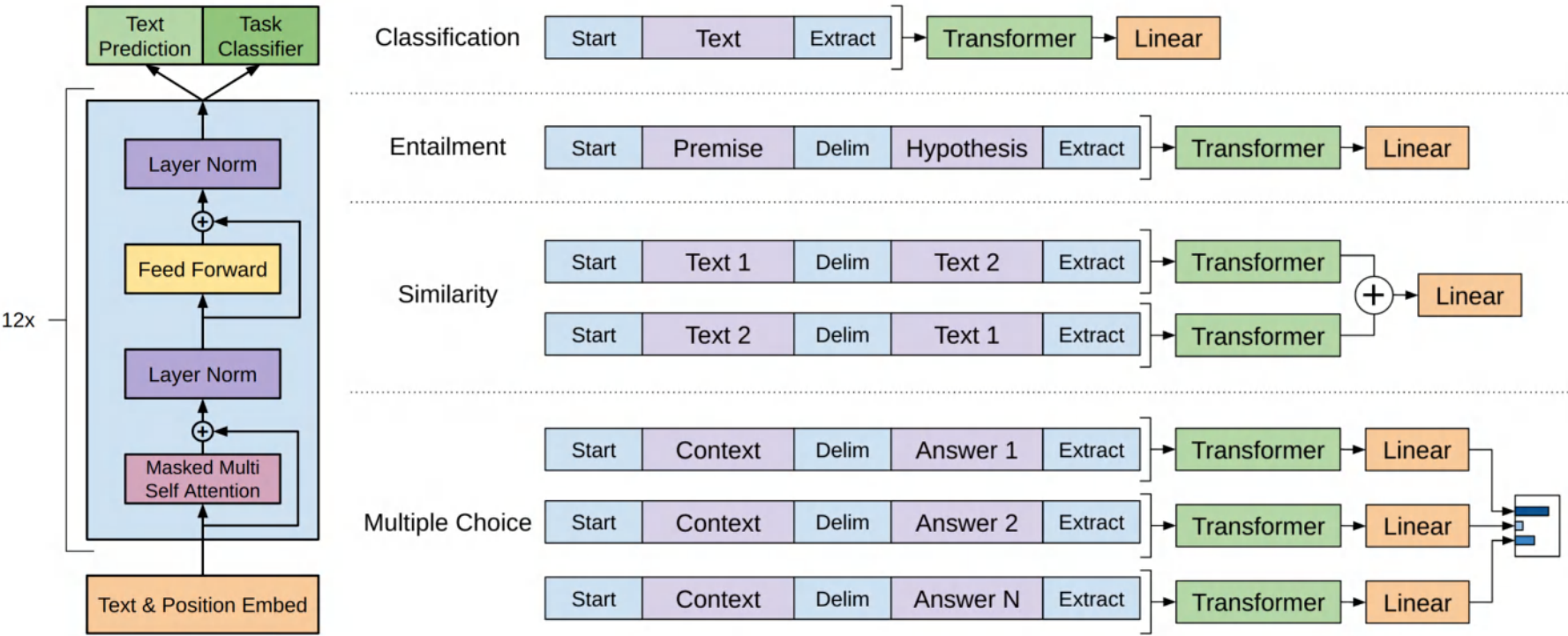
5 [A] [https://en.wikipedia.org/wiki/Artificial\\_general\\_intelligence](https://en.wikipedia.org/wiki/Artificial_general_intelligence)  
[B] B. Goertzel, Artificial General Intelligence, by Springer, 2007.

# 通用人工智能的三个要素

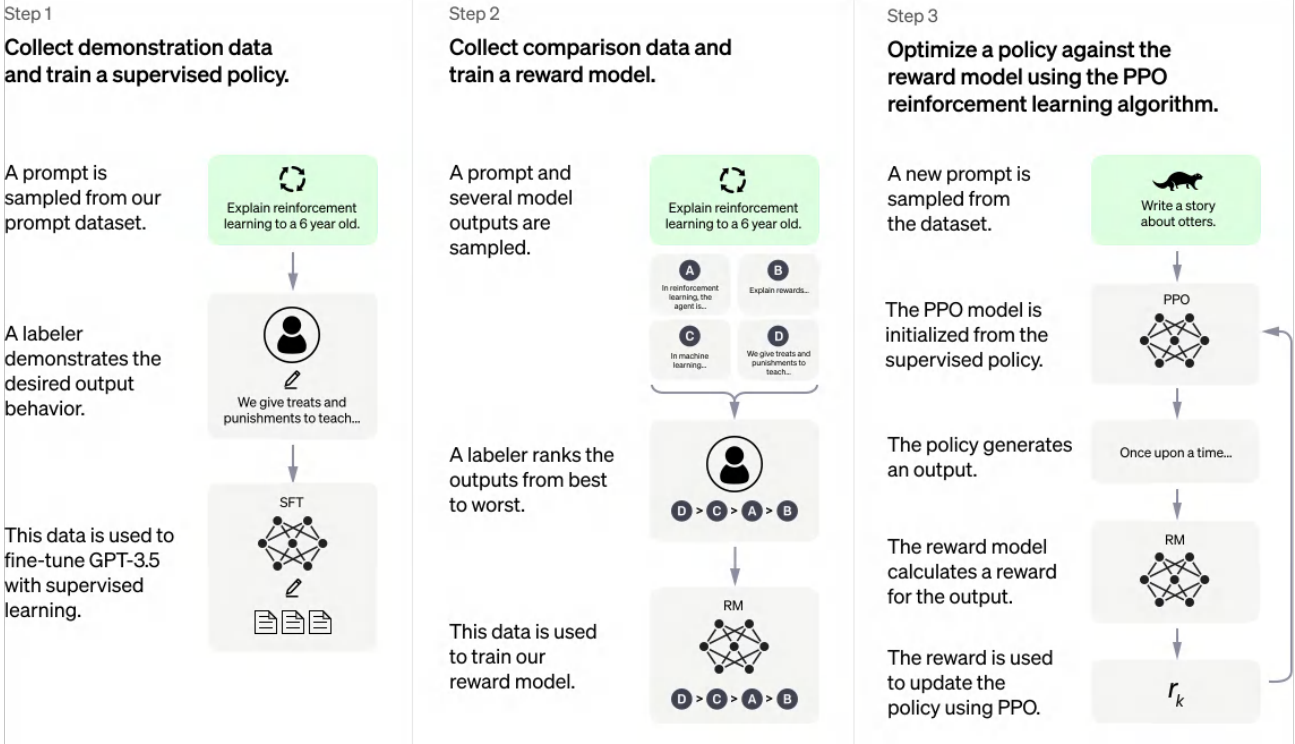


# 自然语言处理交出的答卷：对话大语言模型

	状态 $s \in \mathcal{S}$	动作 $a \in \mathcal{A}$	转移函数 $\tau: \mathcal{S} \times \mathcal{A} \mapsto \mathcal{S}$	奖励函数 $r: \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}$	策略模型 $\pi: \mathcal{S} \mapsto \mathcal{A}$
方案	对话历史 (token序列)	生成下一个 token	生成token后，自 然形成下一状态	用户给对话过程 的主观评分	通过两阶段（预 训练+微调）习得



第一阶段：自监督预训练[A]



第二阶段：有监督微调[B]

[A] A. Radford et al., Improving Language Understanding by Generative Pre-Training, in *NeurIPS*, 2018.

[B] L. Ouyang et al., Training Language Models to Follow Instructions with Human Feedback, in *NeurIPS*, 2022.



# 计算机视觉的复刻：理想很丰满，现实很骨感！

	状态 $s \in \mathcal{S}$	动作 $a \in \mathcal{A}$	转移函数 $\tau: \mathcal{S} \times \mathcal{A} \mapsto \mathcal{S}$	奖励函数 $r: \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}$	策略模型 $\pi: \mathcal{S} \mapsto \mathcal{A}$
理想	多模态历史 (图像和文本 token序列)	执行下一个动作，或者与其他智能体交互	执行动作或者交互后，自然形成下一状态	用户给交互过程的主观评分	通过两阶段（预训练+微调）习得
困难	视觉信号难以无损地转化为token序列	视觉世界中可能只需的动作难以穷尽	尚不存在能够根据动作自然形成下一状态的机制	/	/
现实	无法交互的图像/视频数据	无	无	图像/视频数据中特定任务准确率	通过一阶段（直接训练）习得





# 问题一：为什么？

---

问：在通用人工智能的发展上，计算机视觉为何**落后于**自然语言处理？

答：在自然语言处理领域，交互环境天然存在：在**对话系统**这样的交互环境下，所有任务都可以统一为**生成式**，这使得模型训练过程接近通用人工智能的形式化定义；而在计算机视觉领域，交互环境尚不存在：通用交互环境**尚不存在**，因而不同代理任务间**彼此孤立**，这使得人们无法按照通用人工智能的形式化定义来训练模型。

# 问题二：怎么做？

要素/目标	相关研究方向	年度研究进展简要总结
<div>转移函数</div> <div><math>\tau: \mathcal{S} \times \mathcal{A} \mapsto \mathcal{S}</math></div> <div>构建完备的交互环境</div>	视觉可交互环境	<div>• 视觉可交互环境愈加丰富，包括更大的虚拟环境和真实视频数据集</div>
	视觉生成式模型	<div>• 3DGS大幅提升3D/4D新视角生成效率</div> <div>• 以文生图/视频技术持续演进</div>
<div>奖励函数</div> <div><math>r: \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}</math></div> <div>设计通用的视觉任务</div>	开放世界视觉对话	<div>• 多模态对话提供了开放域视觉识别/理解的新范式</div>
	大一统视觉任务	<div>• 类比语言的视觉上下文学习展现新能力</div> <div>• 视觉智能体为解决复杂任务提供逻辑</div>
<div>策略模型</div> <div><math>\pi: \mathcal{S} \mapsto \mathcal{A}</math></div> <div>训练强大的基础模型</div>	视觉基础模型	<div>• 以SAM为代表的视觉基础模型展现能力</div> <div>• 新型计算架构在视觉领域涌现</div>
	大模型微调技术	<div>• 基于数据/参数高效的大模型微调技术，使得大模型能够快速适应垂直场景</div>

# 愈加充实的视觉可交互环境

环境 → 任务 → 模型



ProcTHOR<sup>[A]</sup>提供可编辑的3D交互环境



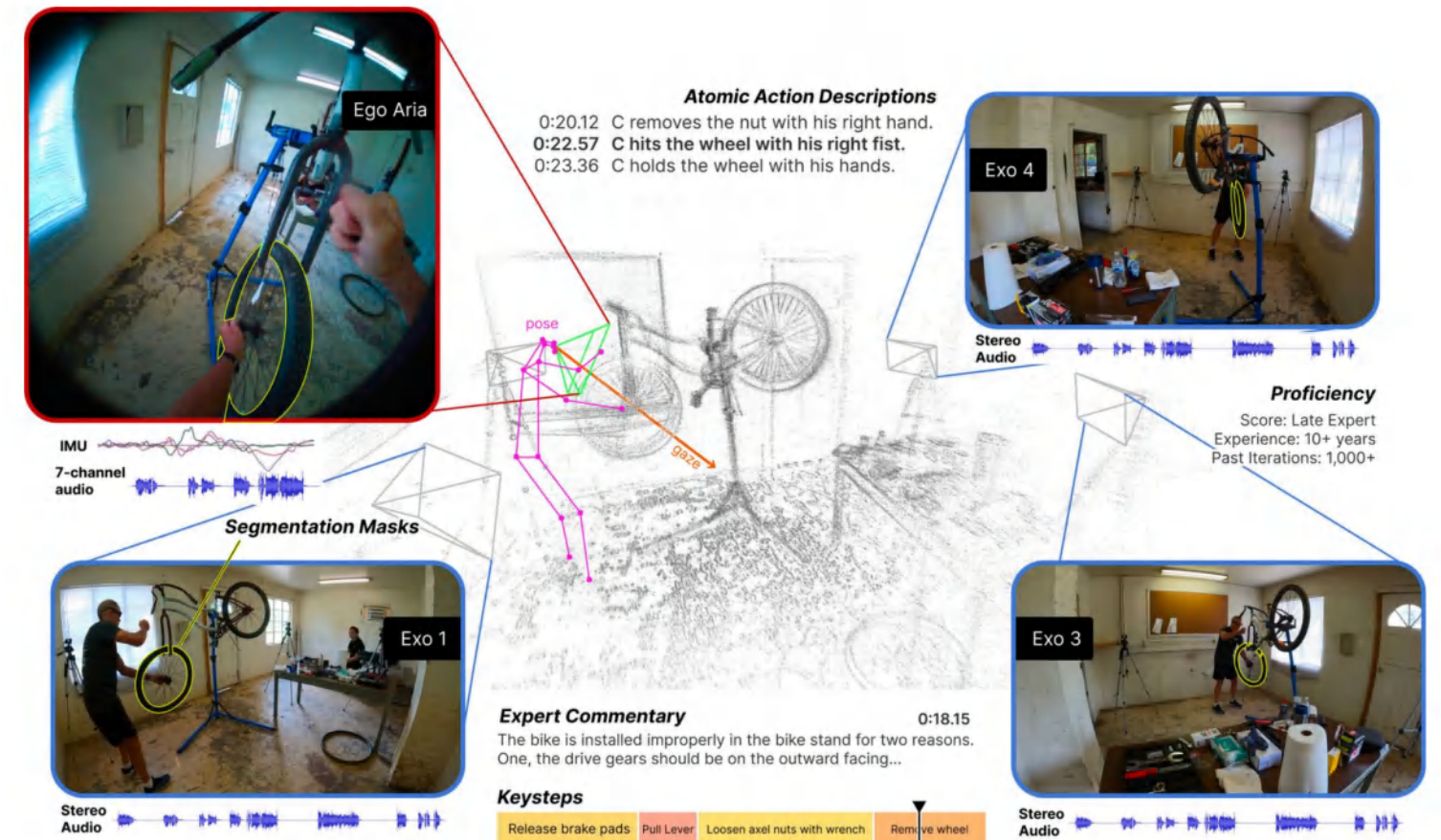
Habitat 3.0<sup>[B]</sup>提供人与机器人协同的环境

真实性

物品逼真度+  
场景丰富度+

交互性

场景可编辑+  
可执行动作+



Ego-Exo4D<sup>[C]</sup>提供包括第一人称的多视角视频数据集

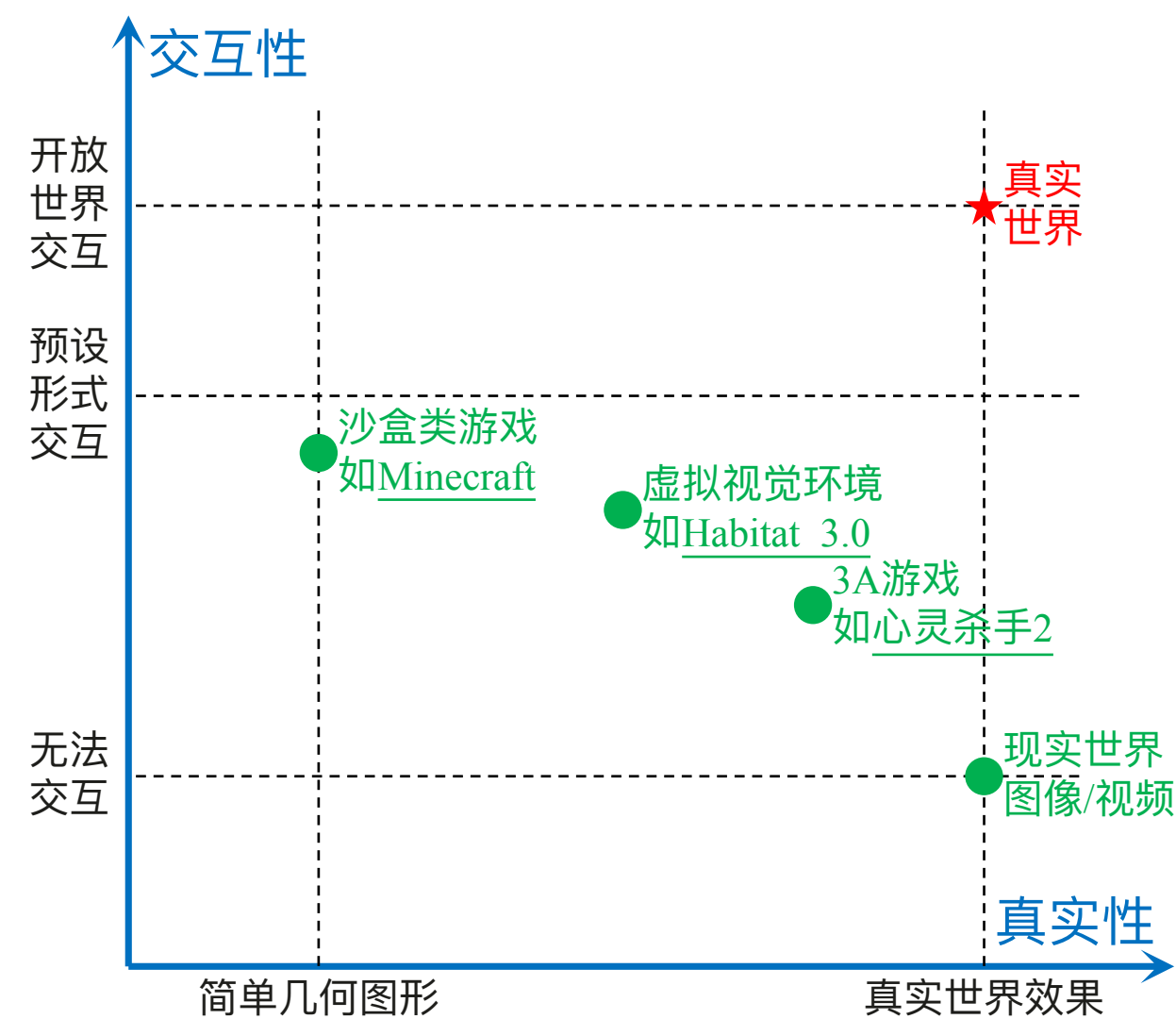
[A] [M. Deitke et al.](#), ProcTHOR: Large-Scale Embodied AI Using Procedural Generation, in *NeurIPS*, 2022.

[B] [X. Puig et al.](#), Habitat 3.0: A Co-Habitat for Humans, Avatars and Robots, in *ICLR*, 2024.

[C] [K. Grauman et al.](#), Ego-Exo4D: Understanding Skilled Human Activity from First- and Third-Person Perspectives, in *arXiv preprint:2311.18259*, 2023.



## 真实性与交互性的权衡



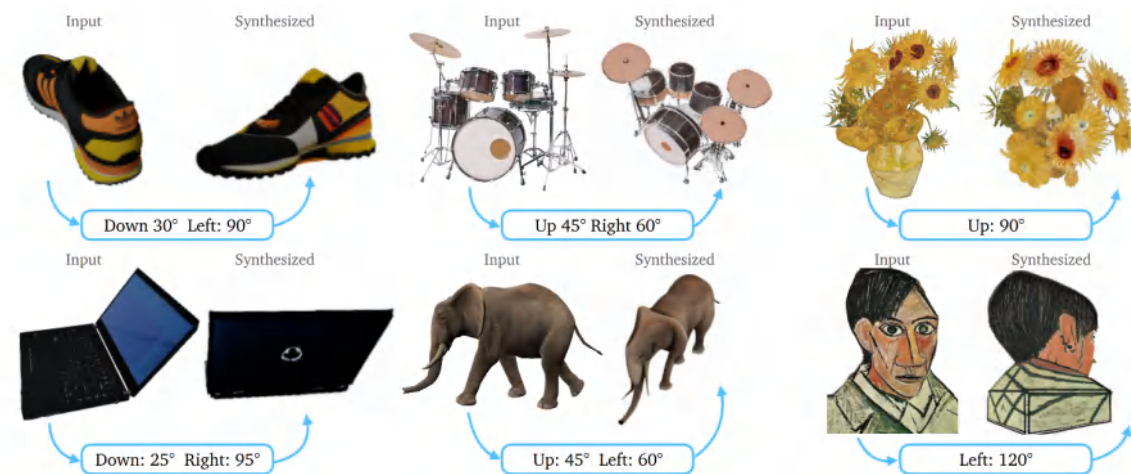
问：上述困难的本质是什么？

答：真实世界复杂度太高，而现有图像/视频数据的本质是以“**采样**”的方式来表征其特性。一旦遇到未采样或欠采样的“空洞”，就必须舍弃某些特性。

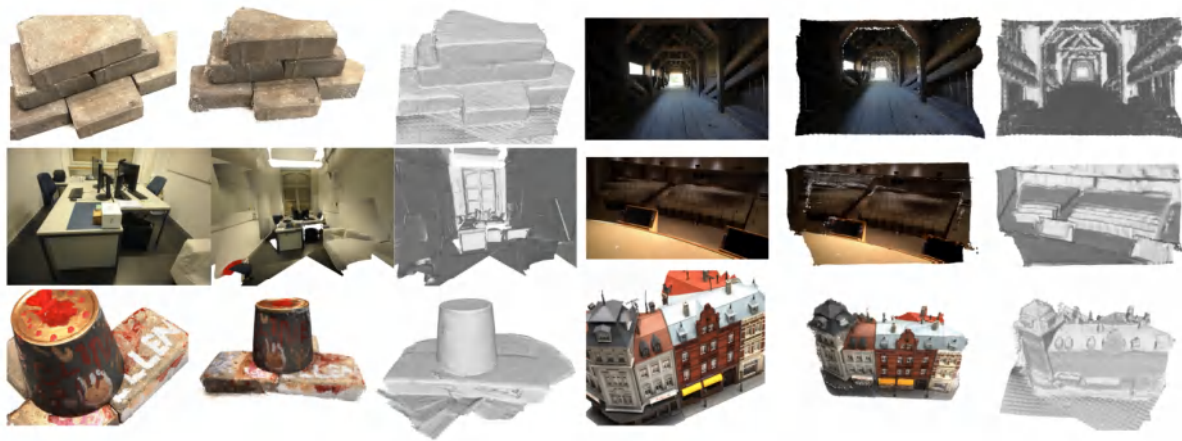
为使环境更加逼近真实世界，必须大力发展**生成式模型**，以显式先验（如3D/物理规律建模）或者隐式先验（如神经网络），来填补真实世界采样的“空洞”。

# 视觉生成式模型之3D/4D建模和生成

环境 → 任务 → 模型

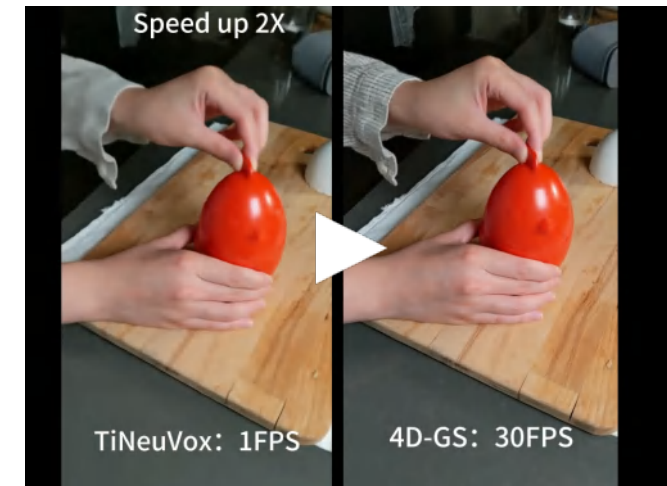
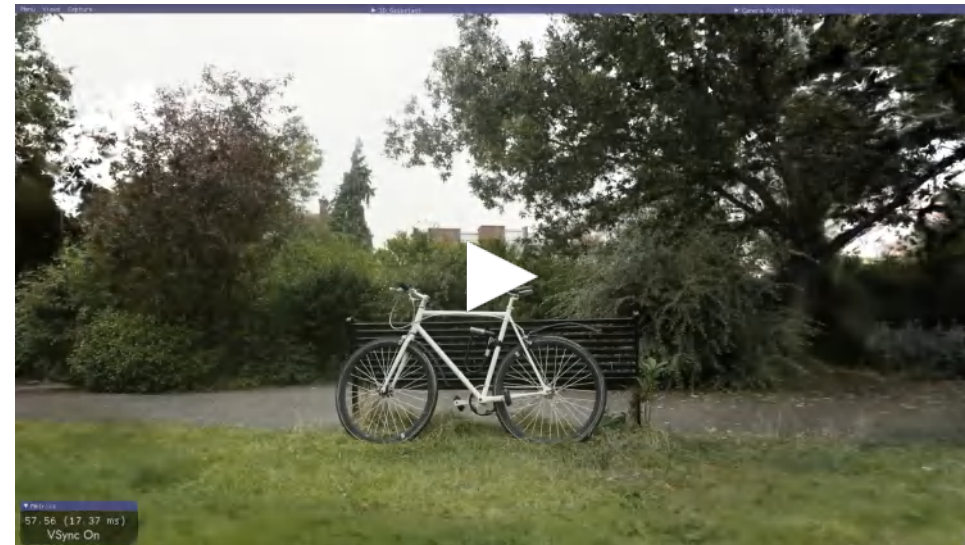


Zero-1-to-3<sup>[A]</sup>从单个3D视角生成其他视角



DUST3R<sup>[B]</sup>仅用两个视角生成3D场景

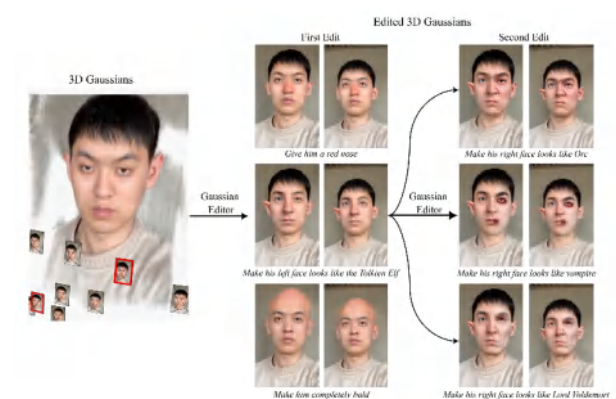
3DGS<sup>[C]</sup>提供了更高效的3D表征, 提升了新视角生成质量和速度



4DGS<sup>[D]</sup>扩展至3D视频



两个同名的GaussianEditor<sup>[E, F]</sup>使得3D物体编辑更加便捷



[A] [R. Liu et al.](#), Zero-1-to-3: Zero-shot One Image to 3D Object, in *ICCV*, 2023.

[B] [S. Wang et al.](#), DUST3R: Geometric 3D Vision Made Easy, in *CVPR*, 2024.

[C] [B. Kerbl et al.](#), 3D Gaussian Splatting for Real-Time Radiance Field Rendering, in *SIGGRAPH*, 2023.

[D] [G. Wu et al.](#), 4D Gaussian Splatting for Real-Time Dynamic Scene Rendering, in *CVPR*, 2024.

[E] [Y. Chen et al.](#), GaussianEditor: Swift and Controllable 3D Editing with Gaussian Splatting, in *CVPR*, 2024.

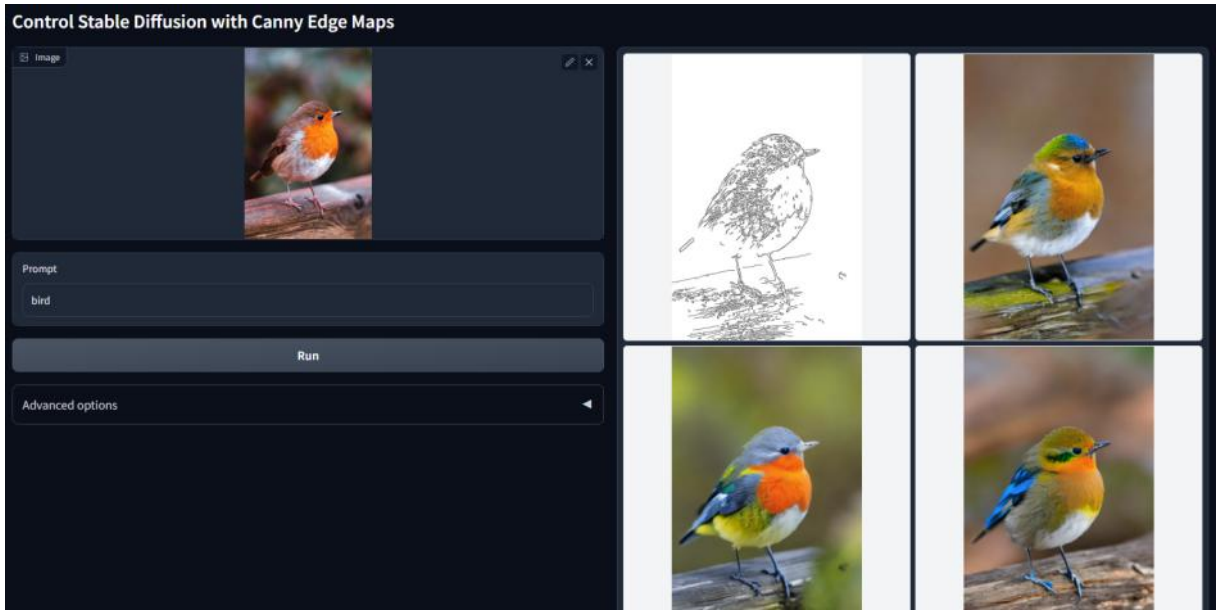
[F] [J. Fang et al.](#), GaussianEditor: Editing 3D Gaussians Delicately with Text Instructions, in *CVPR*, 2024.



# 视觉生成式模型之文生图/视频

环境 任务 模型

ControlNet<sup>[A]</sup>支持可控图像生成

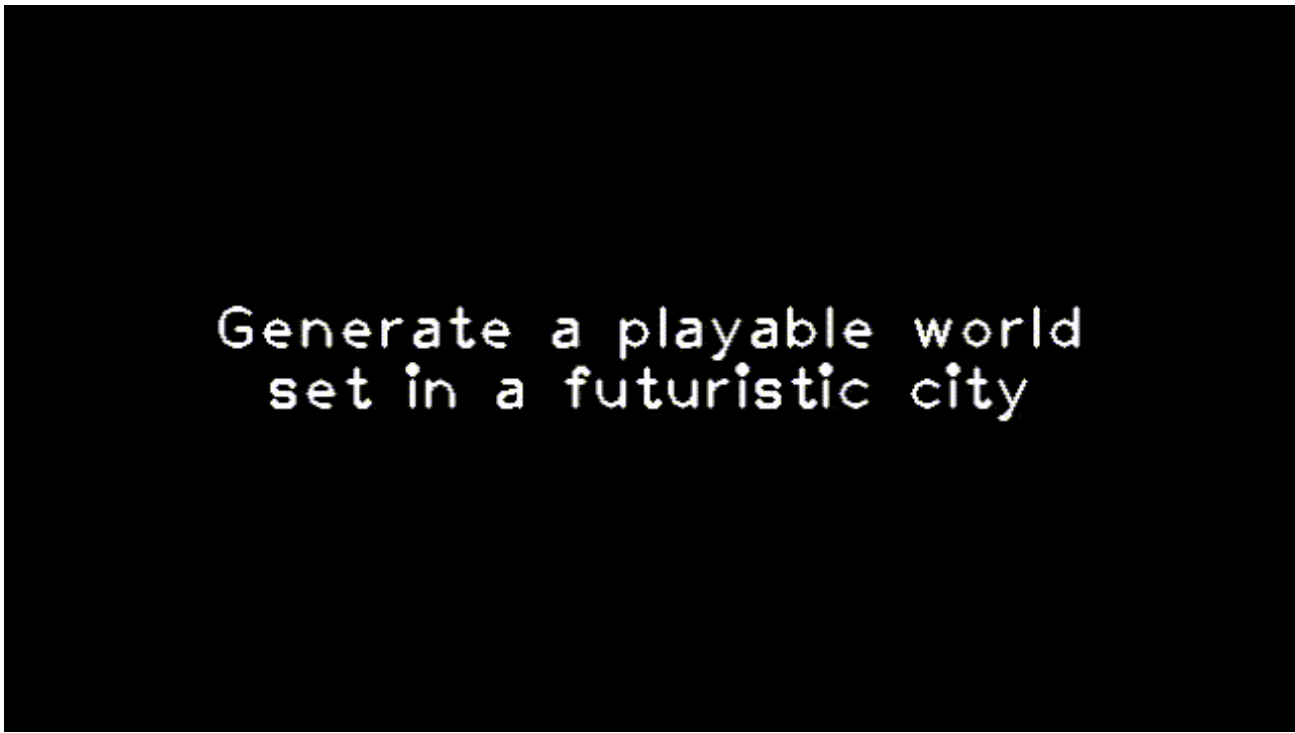


Sora支持长视频生成

扩散模型



Stable Diffusion 3支持高质量图像生成



Genie支持可控环境生成

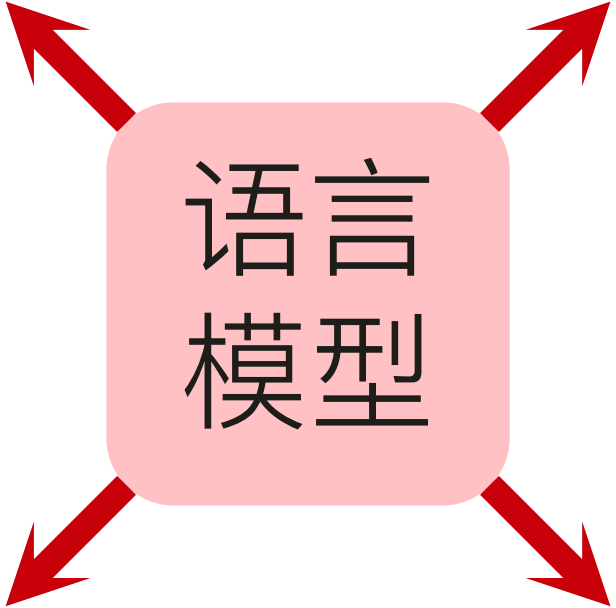
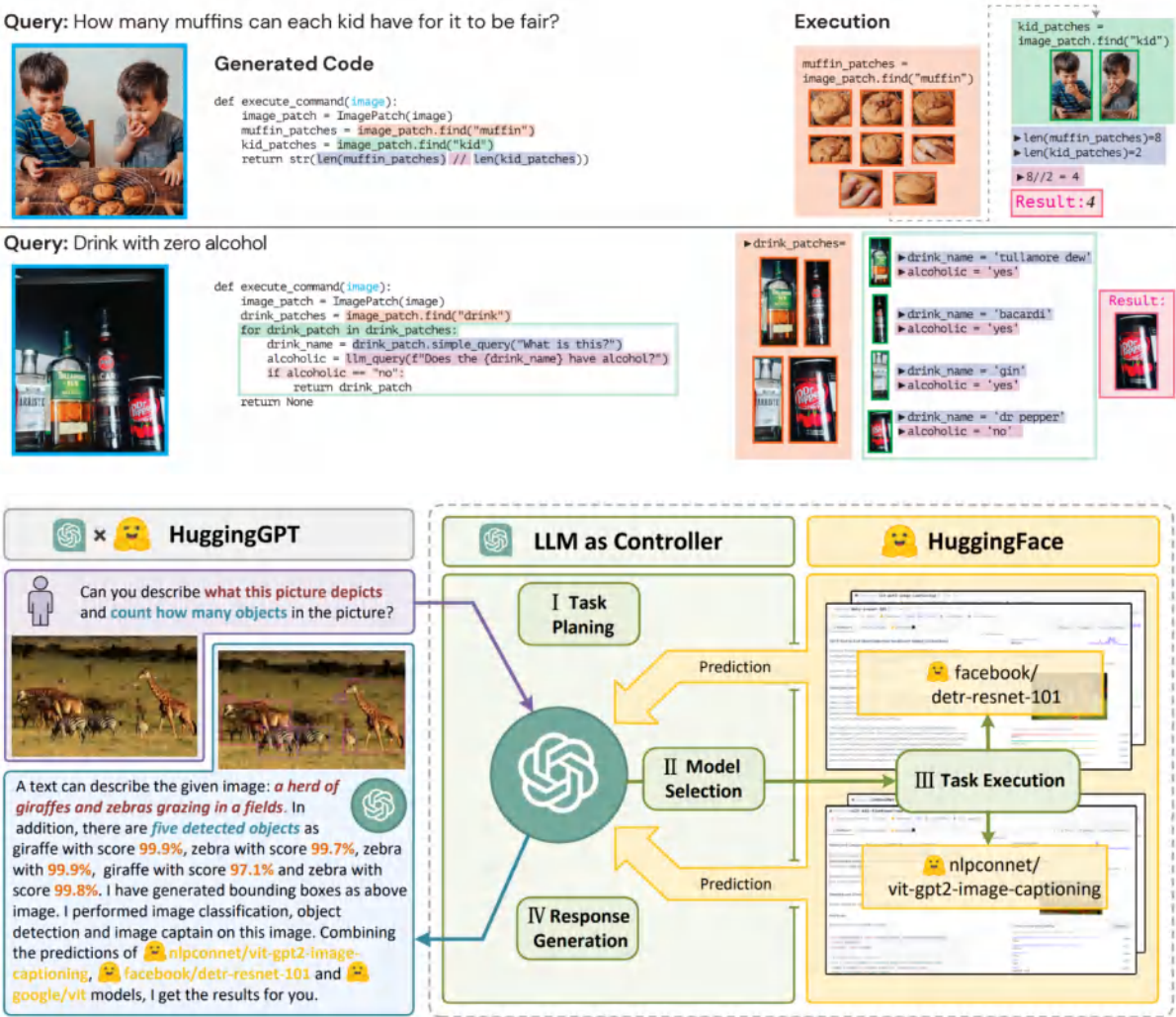
[A] L. Zhang et al., Adding Conditional Control to Text-to-Image Diffusion Models, in CVPR, 2023.



# 从封闭域识别走向开放世界视觉对话/理解

环境 → 任务 → 模型

ViperGPT[A] 解决复杂逻辑问题  
HuggingGPT[B] 解决复杂逻辑问题



**LLaVA**

Describe this image as detailed as possible.

The image shows a busy city street with a clock tower in the background. The street is lined with shops and restaurants on both sides. There are several motorcycles parked on the side of the road, and people are walking down the street. The clock tower has a large clock face with Roman numerals and a small spire on top. The buildings on either side of the street have ornate facades and balconies. The street is paved with cobblestones and there are streetlights on either side of the road. The sky is clear and blue, and there are a few clouds in the distance.

**LLaVA**

Describe this image as detailed as possible.

The image shows a group of musicians performing on stage in front of a large audience. The musicians are playing electric guitars and are wearing black and white clothing. The audience is made up of people of different ages and ethnicities who are watching the performance and taking photos with their phones. The stage is lit up with bright lights and there is a large screen displaying the band's name and lyrics in Chinese characters. The atmosphere is lively and energetic, with the audience cheering and singing along to the music.

**LLaVA**

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.

[A] D. Suris et al., ViperGPT: Visual Inference via Python Execution for Reasoning, in *ICCV*, 2023.

[B] Y. Shen et al., HuggingGPT: Solving AI Tasks with ChatGPT and its Friends in HuggingFace, in *NeurIPS*, 2023.

[C] H. Liu et al., Visual Instruction Tuning, in *NeurIPS*, 2023.

[D] D. Zhu et al., MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models, in *ICLR*, 2024.

User LLaVA

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]

User LLaVA

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.

LLaVA[C]生成的对话  
MiniGPT-4[D]生成的对话



# GPT不是终点，开放世界对话/理解还远不够！

环境 → 任务 → 模型

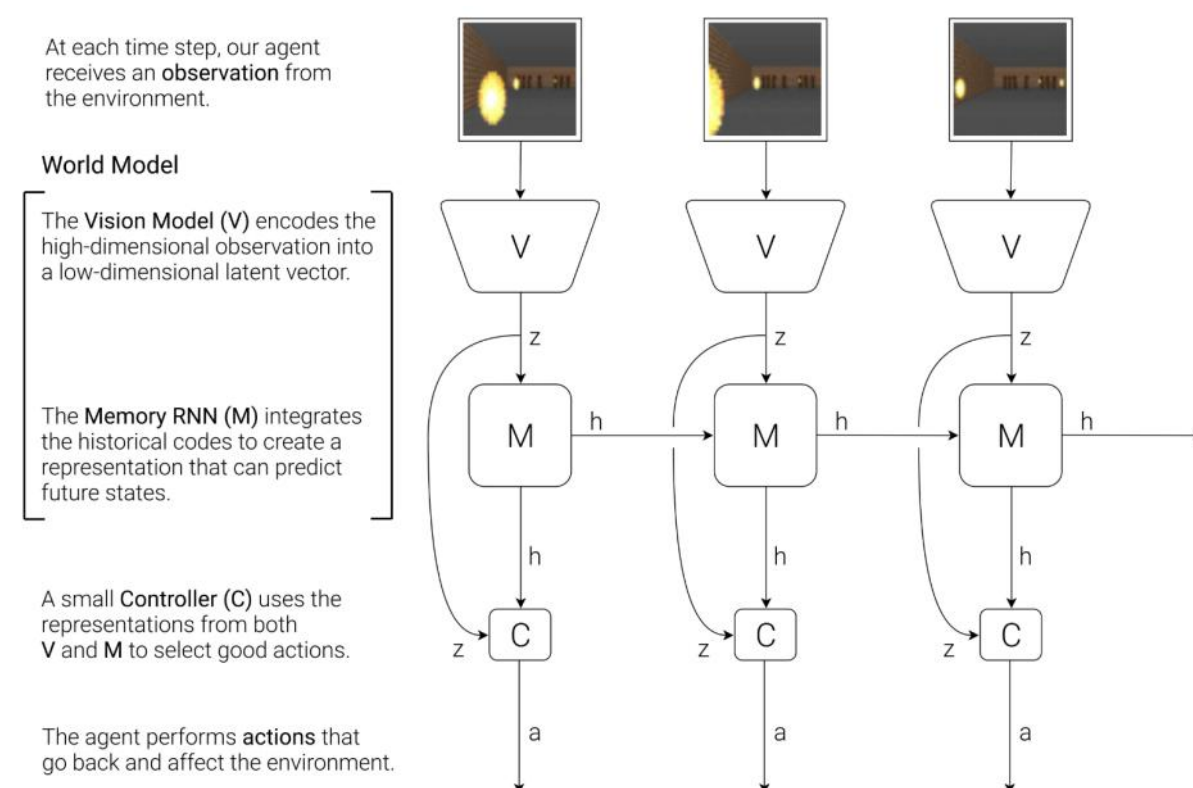
问：为了实现**视觉通用人工智能**，我们需要定义怎样的任务？

答：仿照自然语言处理的成功经验（预测下一个token），**通用**视觉任务需要满足以下两个条件：

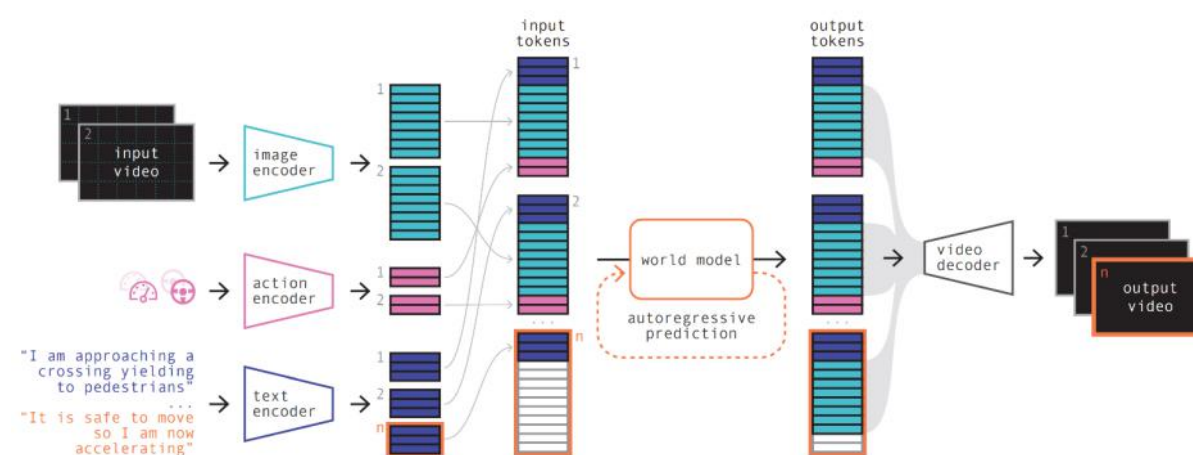
((1)) 一种输出形式达成所有需求

((2)) 与环境互动，自然形成下一状态

也就是说，环境与任务应当是**一体的**；  
它们的统一形态，就是**世界模型**！



面向电子游戏的世界模型[A]



面向自动驾驶的世界模型[B]



# 统一视觉任务定义初探之视觉上下文学习

环境 任务 模型

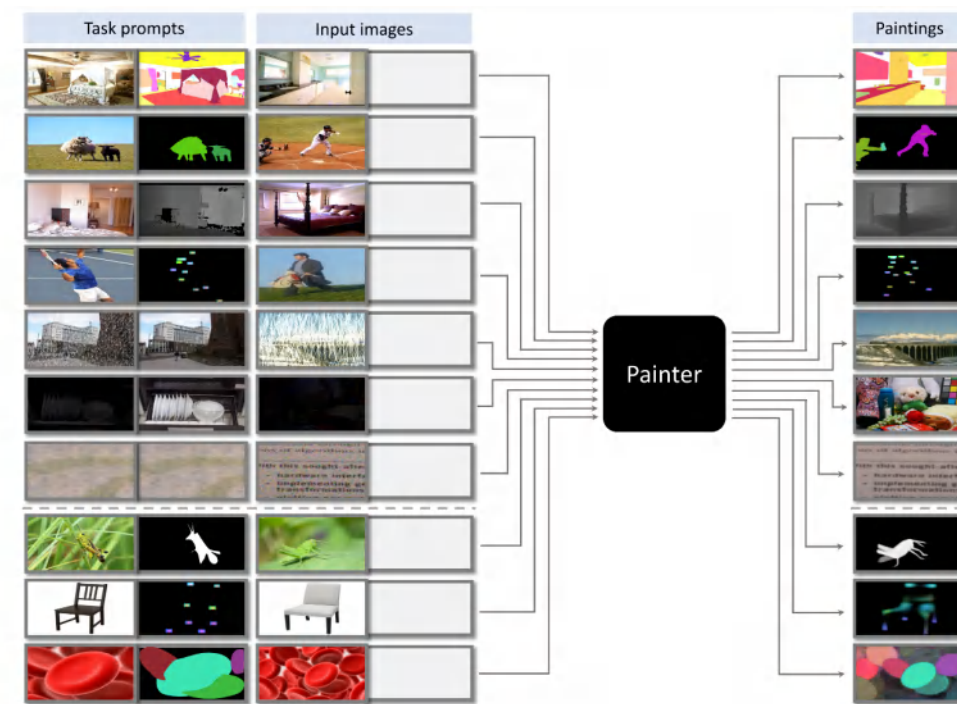
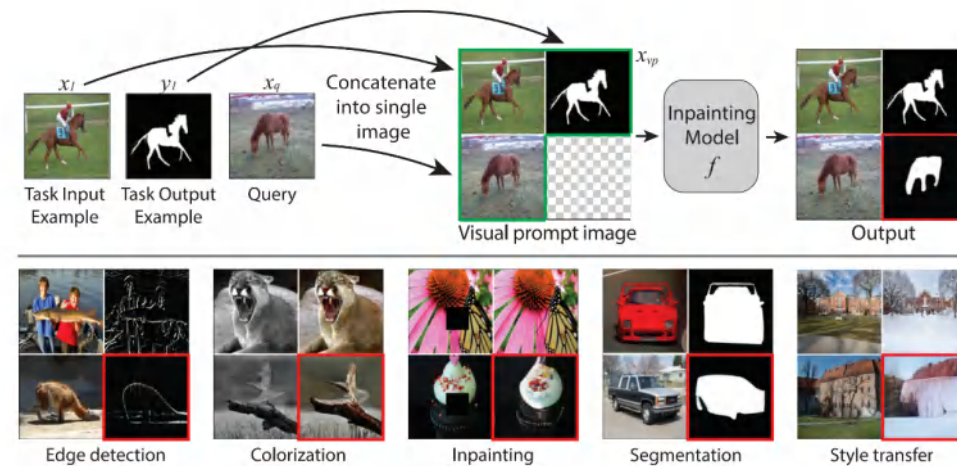
GPT3<sup>[A]</sup>定义了语言上下文学习范式

## Few-shot

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

1 Translate English to French: ← task description  
2 sea otter => loutre de mer ← examples  
3 peppermint => menthe poivrée  
4 plush girafe => girafe peluche  
5 cheese =>

[A] **T. Brown et al.**, Language Models are Few-Shot Learners, in *NeurIPS*, 2020.  
[B] **A. Bar et al.**, Visual Prompting via Image Inpainting, in *NeurIPS*, 2022.  
[C] **X. Wang et al.**, Images Speak in Images: A Generalist Painter for In-Context Visual Learning, in *CVPR*, 2023.  
[D] **Y. Bai et al.**, Sequential Modeling Enables Scalable Learning for Large Vision Models, in *CVPR*, 2024.



垂域上的视觉上下文学习<sup>[B,C]</sup>

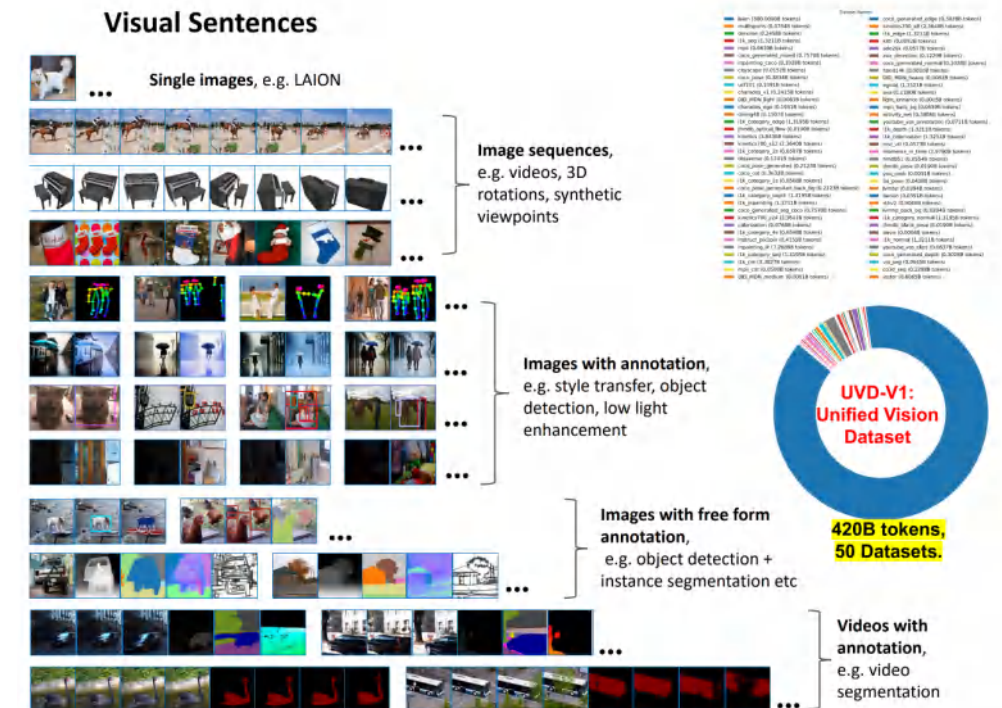
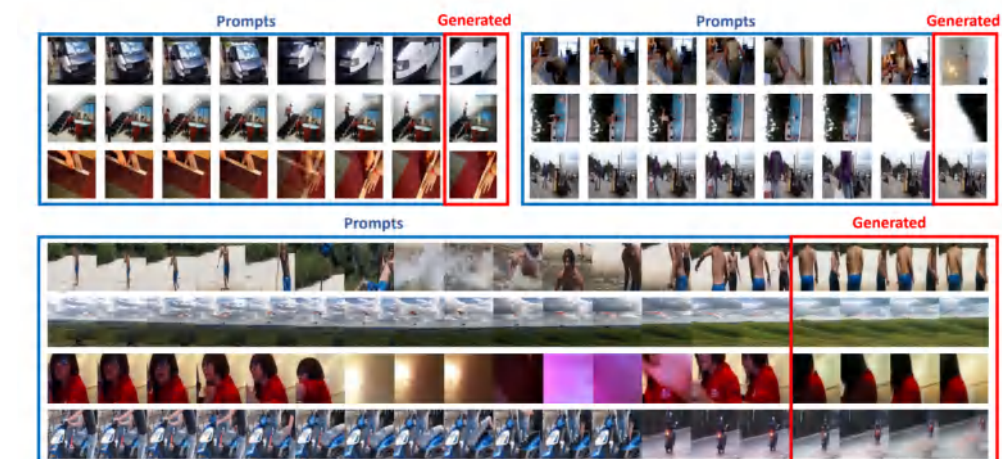


Figure 1. Visual sentences allow us to format diverse vision data into the unified structure of image sequences.



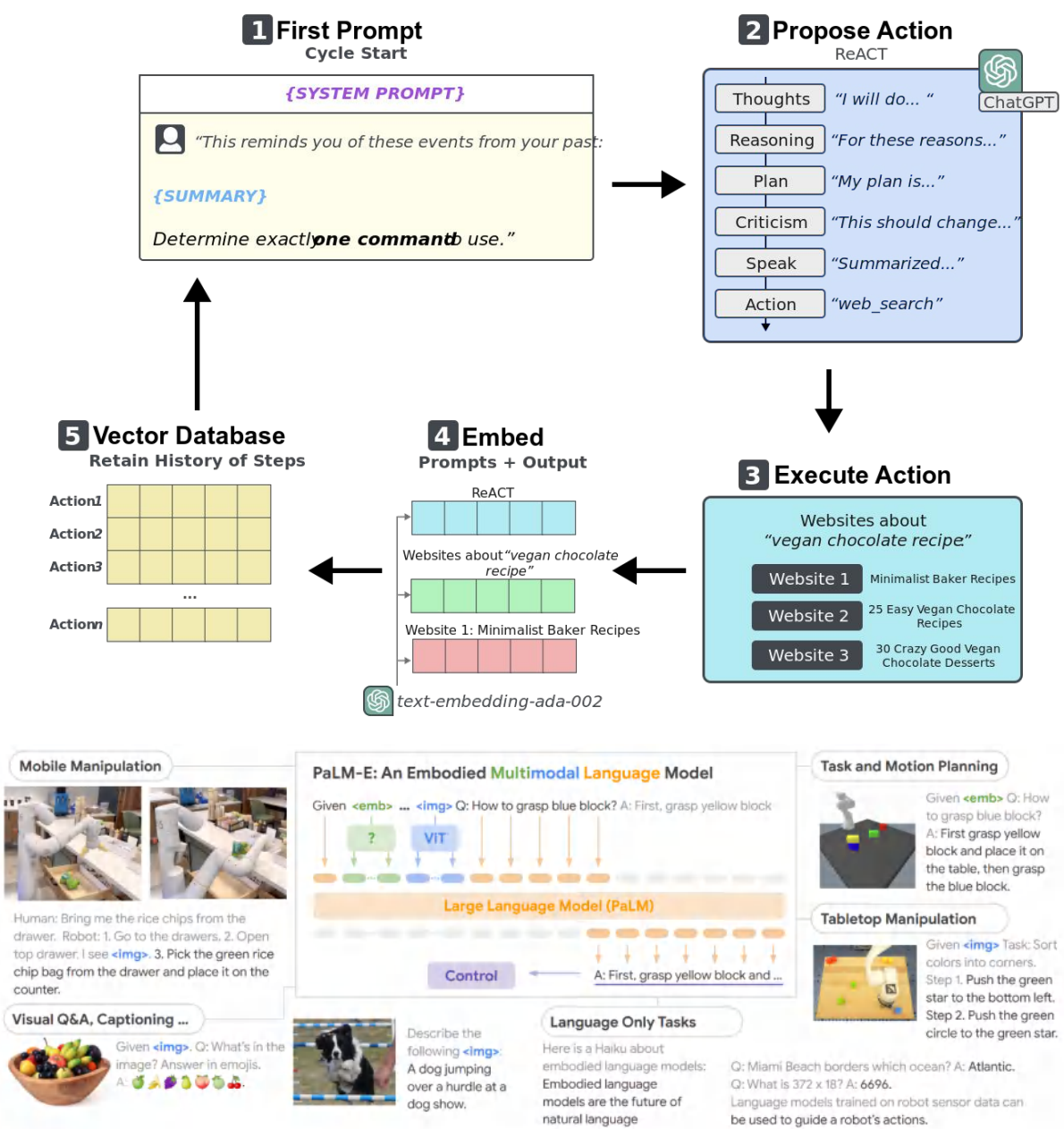
广域上的视觉上下文学习<sup>[D]</sup>



# 统一视觉任务求解初探之视觉智能体

AutoGPT[A]利用语言模型将复杂任务分解为子问题

PaLM-E[B]解决具身智能问题



语言模型



“AI小镇”模拟智能体生活和交互环境[C]

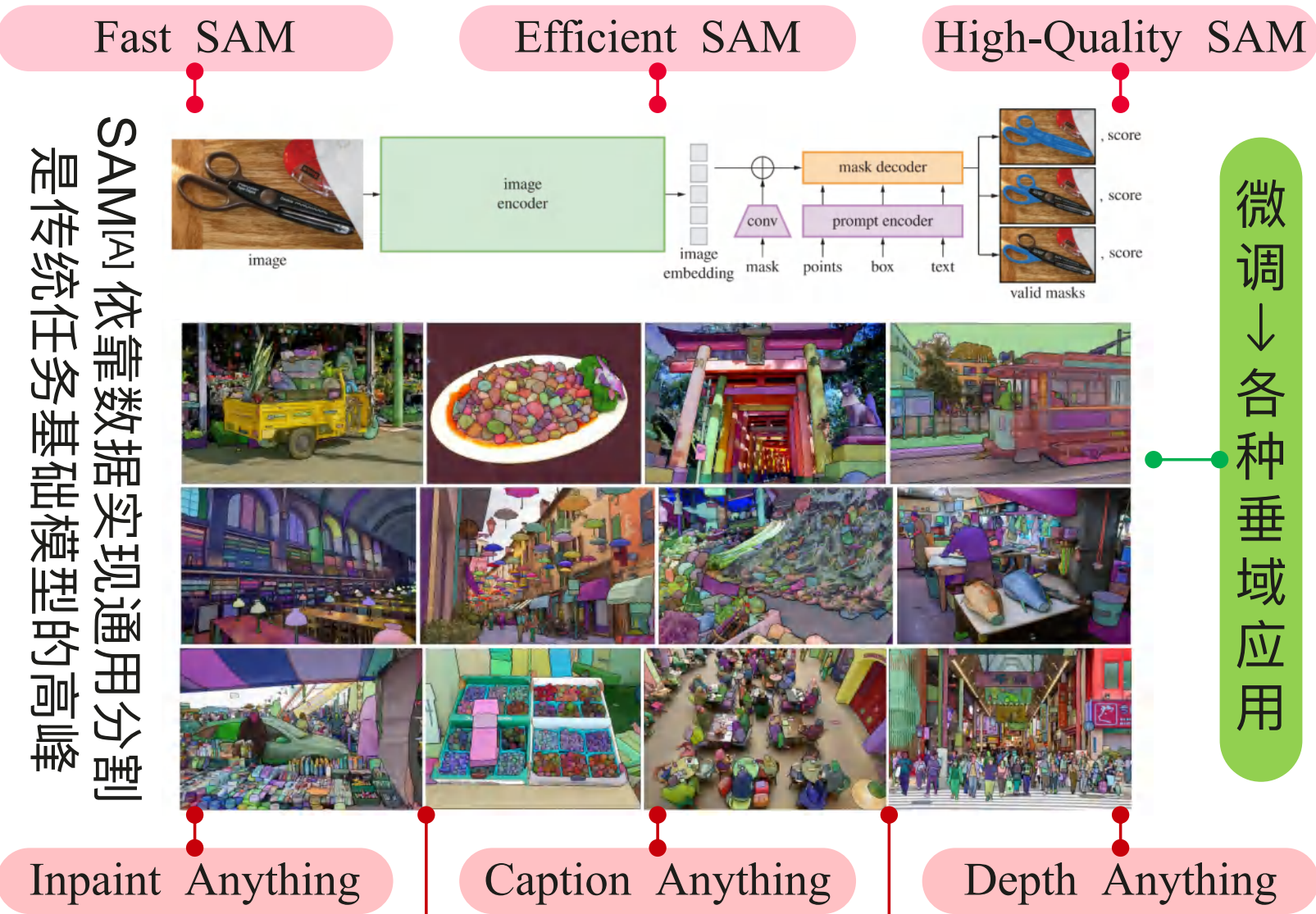
	Inventory (5/36): {'oak_planks': 3, 'stick': 4, 'crafting_table': 1, 'stone': 3, 'wooden_pickaxe': 1}	GPT-4	Reasoning: Since you have a wooden pickaxe and some stones, it would be beneficial to upgrade your pickaxe to a stone pickaxe for better efficiency. Task: Craft 1 stone pickaxe.
	Biome: river Inventory (4/36): {'sugar_cane': 3, 'paper': 3, 'fishing_rod': 1, 'wooden_pickaxe': 1}	GPT-4	Reasoning: Since you have a fishing rod in your inventory and you are near a river biome, it would be a good opportunity to catch some fish for food and experience. Task: Catch 1 fish.
	Nearby entities: pig, cat, villager Health: 12/20 Hunger: 0/20	GPT-4	Reasoning: Your hunger is at 0, which means you need to find food to replenish your hunger. Since there are pigs nearby, you can kill one to obtain raw porkchops. Task: Kill 1 pig.
	Inventory (6/36): {'furnace': 1, 'stone_pickaxe': 1, 'oak_planks': 7, 'crafting_table': 1, 'raw_iron': 4, 'coal': 1}	GPT-4	Reasoning: You have raw iron and coal, and you have a furnace. It's time to smelt the iron to make iron ingots, which can be used to craft better tools and armor. Task: Smelt 4 raw iron.
	Time: night Nearby entities: zombie Equipment: 'stone_sword', 'shield'	GPT-4	Reasoning: Since it's night and there's a zombie nearby, it's a good opportunity to try killing the zombie now that you have a stone sword and a shield equipped. Task: Kill 1 zombie.

[A] <https://autogpt.net/>  
[B] D. Driess et al., PaLM-E: An Embodied Multimodal Language Model, in *ICML*, 2023.  
[C] J. S. Park et al., Generative Agents: Interactive Simulacra of Human Behavior, in *UIST*, 2023.  
[D] G. Wang et al., Voyager: An Open-Ended Embodied Agent with Large Language Models, in *TMLR*, 2024.



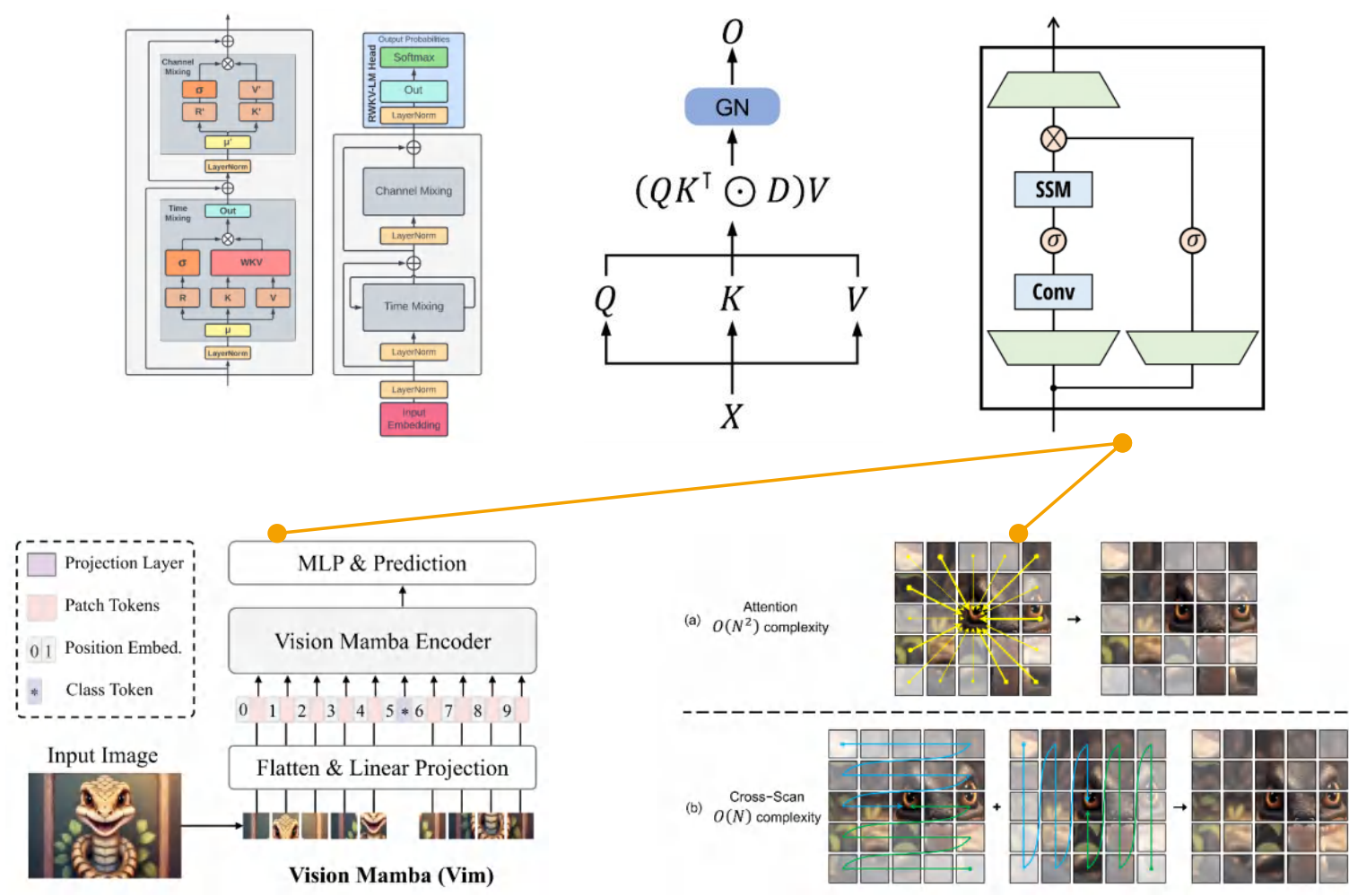
# 视觉基础模型是通用视觉人工智能的载体

环境 任务 模型



SAM[A] 依靠数据实现通用分割  
是传统任务基础模型的高峰

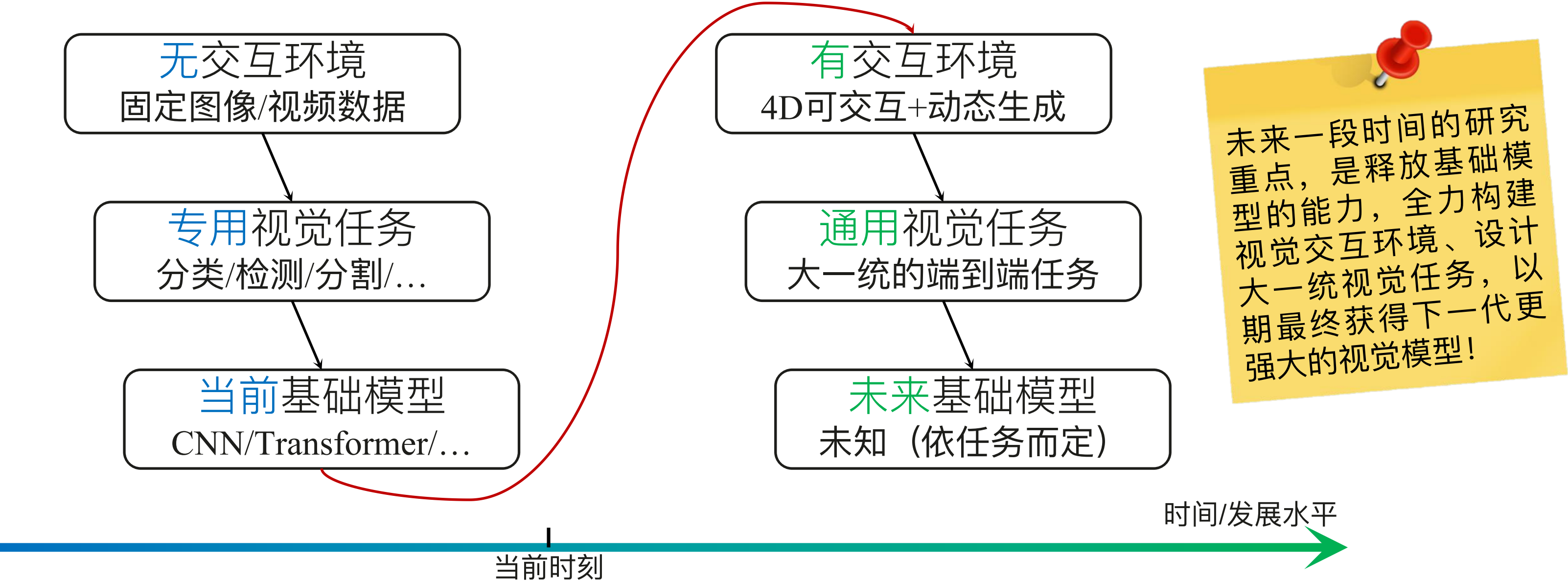
RWKV[B]、RetNet[C]、Mamba[D]等语言模型



Vision Mamba[E]、Vmamba[F]等视觉模型

[A] A. Kirillov et al., Segment Anything, in ICCV, 2023.  
[B] B. Peng et al., RWKV: Reinventing RNNs for the Transformer Era, in EMNLP Findings, 2023.  
[C] Y. Sun et al., Retentive Network: A Successor to Transformer for Large Language Models, in arXiv preprint:2307.08621, 2023.  
[D] A. Gu et al., Mamba: Linear-Time Sequence Modeling with Selective State Spaces, in arXiv preprint:2312.00752, 2023.  
[E] L. Zhu et al., Vision Mamba: Efficient Visual Representation Learning with Bidirectional State Space Model, in arXiv preprint:2401.09417, 2024.  
[F] Y. Liu et al., VMamba: Visual State Space Model, in arXiv preprint:2401.10166, 2024.

# 结语和展望：技术的螺旋上升规律



**勇敢地挥别过去，拥抱新时代，开创新未来！**



# 感谢聆听

---

- 扫描二维码可下载本次报告PPT（同时包含VALSE2023报告及survey）
- 欢迎与我探讨相关学术问题：[198808xc@gmail.com](mailto:198808xc@gmail.com)



## Towards AGI in Computer Vision: Lessons Learned from GPT and Large Language Models

Lingxi Xie, Longhui Wei, Xiaopeng Zhang, Kaifeng Bi, Xiaotao Gu, Jianlong Chang,  
and Qi Tian\*, *Fellow, IEEE*

**Abstract**—The AI community has been pursuing algorithms known as artificial general intelligence (AGI) that apply to any kind of real-world problem. Recently, chat systems powered by large language models (LLMs) emerge and rapidly become a promising direction to achieve AGI in natural language processing (NLP), but the path towards AGI in computer vision (CV) remains unclear. One may owe the dilemma to the fact that visual signals are more complex than language signals, yet we are interested in finding concrete reasons, as well as absorbing experiences from GPT and LLMs to solve the problem. In this paper, we start with a conceptual definition of AGI and briefly review how NLP solves a wide range of tasks via a chat system. The analysis inspires us that **unification** is the next important goal of CV. But, despite various efforts in this direction, CV is still far from a system like GPT that naturally integrates all tasks. We point out that the essential weakness of CV lies in lacking a paradigm to **learn from environments**, yet NLP has accomplished the task in the text world. We then imagine a pipeline that puts a CV algorithm (*i.e.*, an agent) in world-scale, interactable environments, pre-trains it to predict future frames with respect to its action, and then fine-tunes it with instruction to accomplish various tasks. We expect substantial research and engineering efforts to push the idea forward and scale it up, for which we share our perspectives on future research directions.