



# 从大语言模型到世界模型

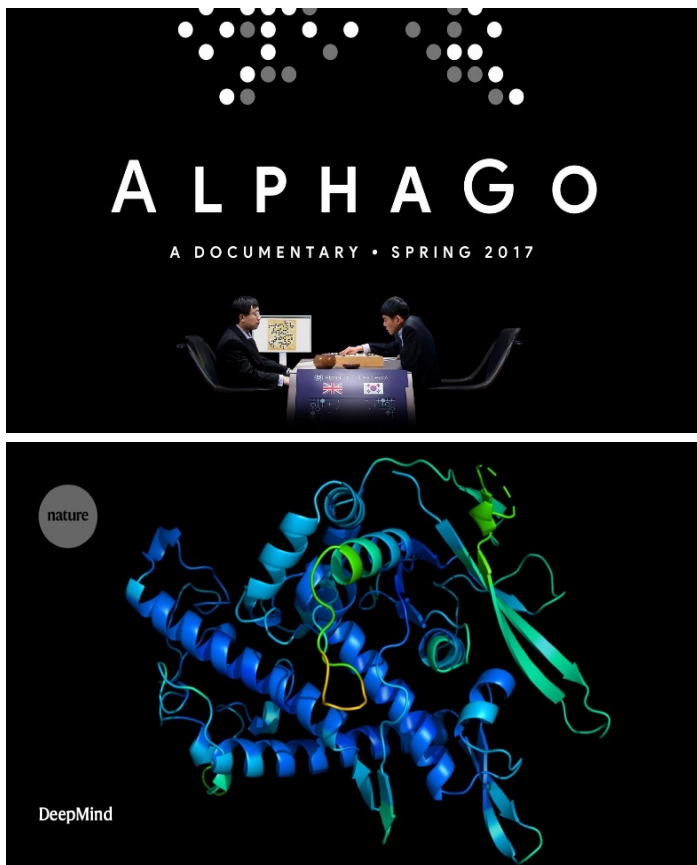
黄萱菁

复旦大学

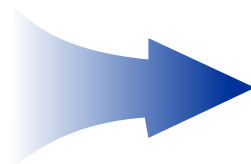
2026年4月30日 Thursday

# AI的标志性突破

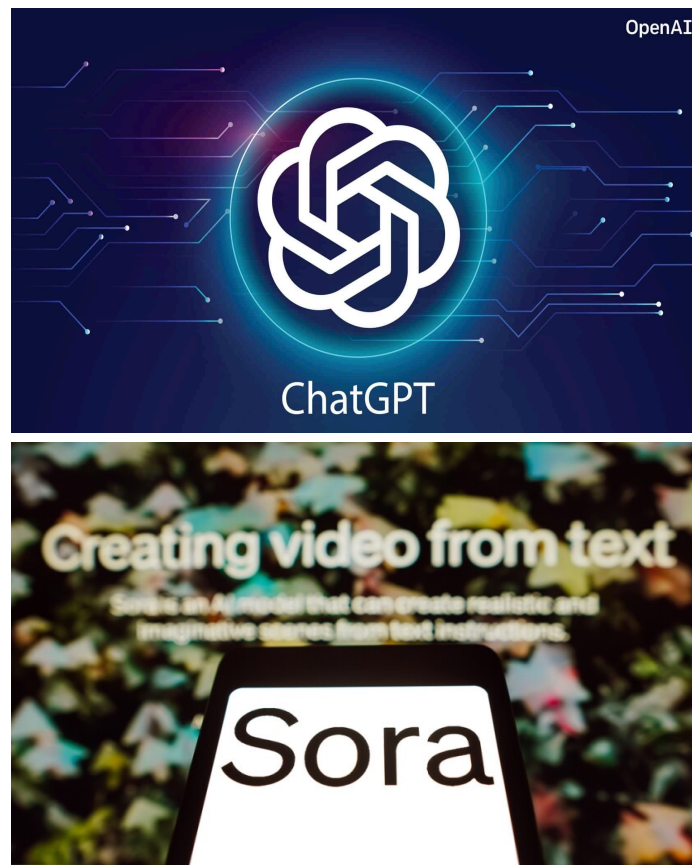
上一代人工智能  
(弱人工智能)



针对特定任务

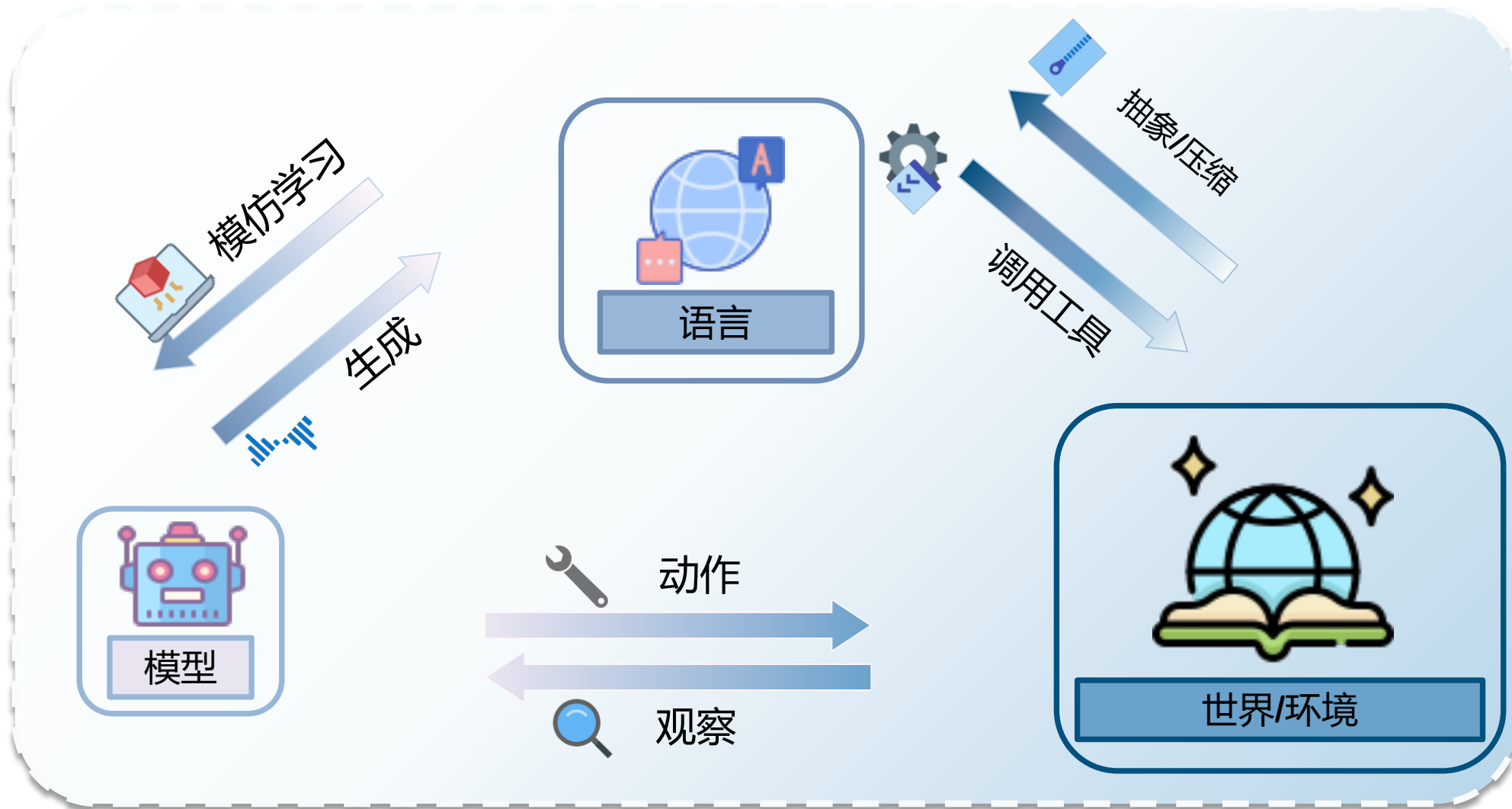


新一代人工智能  
(通用人工智能)



多任务通用

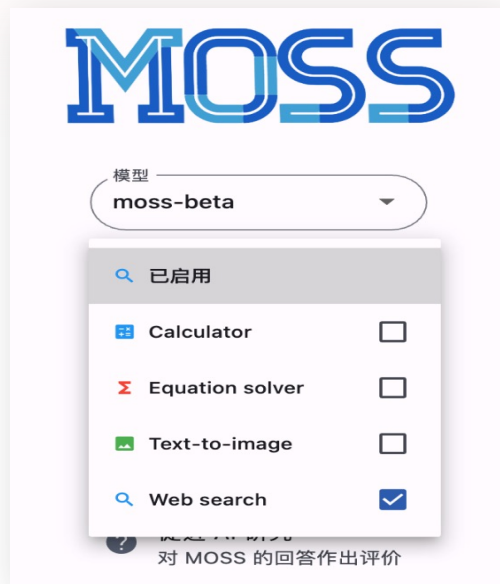
# 大模型通过语言来构建世界知识



# 课题组在大模型方面的工作

## 大语言模型 **MOSS**

国内**首个**公开开放测试的类chatGPT模型，**支持工具调用**



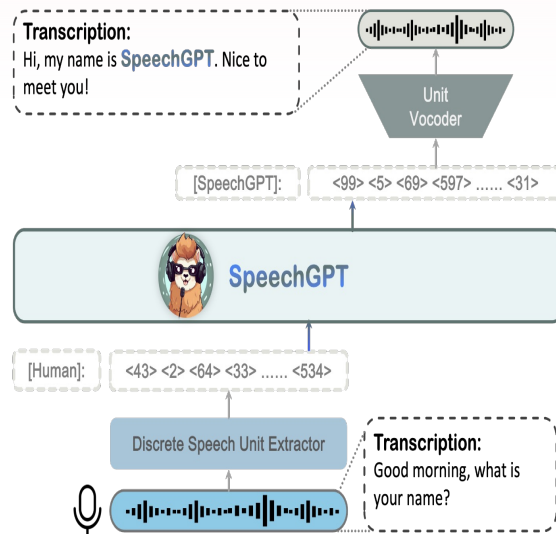
## 大语言模型 **浦语**

推理能力大幅提升，部分指标**超越 Llama3-70B**，支持**百万长度**



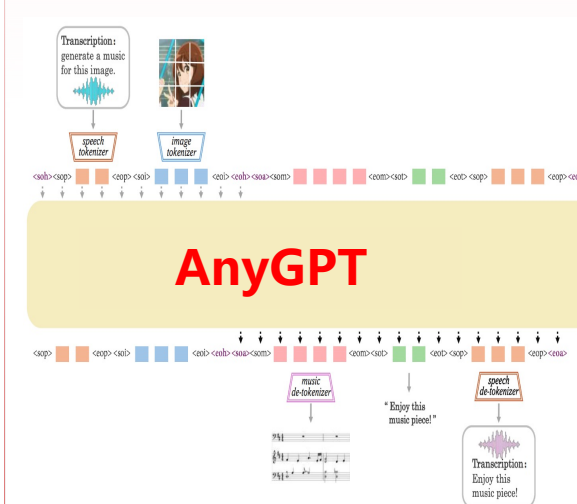
## 语音大模型 **SpeechGPT**

**首个实现端到端语音交互的语音大模型**，让大语言模型具有内生语音跨模态能力的模型



## 多模态大模型 **AnyGPT**

**首个基于离散词符的多模态大模型**，支持文本、图像、音频、音乐等任意模态的组合输入与输出



# 大语言模型的不足

Predict the future	The winner of the UEFA Euro 2028 is {???
Current world knowledge	The current stock price of APPL is {???
Arbitrarily long arithmetic	$36382894730 + 238302849204 = \{???\}$
Extremely long inputs	[2,000 page Harry Potter fan-fiction] What happened after Harry opened the chest for the second time? {???
Many-step reasoning	Take the nineteenth digit of Pi and multiply it by the e to the fourth power. The resulting ones-digit of the resulting number is {???
Information not in the training data	Robin's favorite color is {???

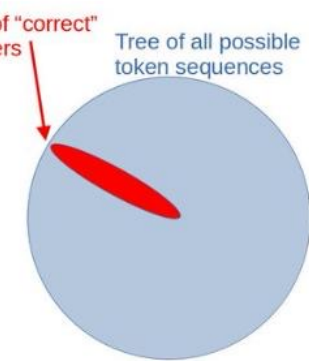
**Yann LeCun**  
@ylecun

I have claimed that Auto-Regressive LLMs are exponentially diverging diffusion processes. Here is the argument:  
Let  $e$  be the probability that any generated token exits the tree of "correct" answers.  
Then the probability that an answer of length  $n$  is correct is  $(1-e)^n$

1/

**Unpopular Opinion about AR-LLMs**

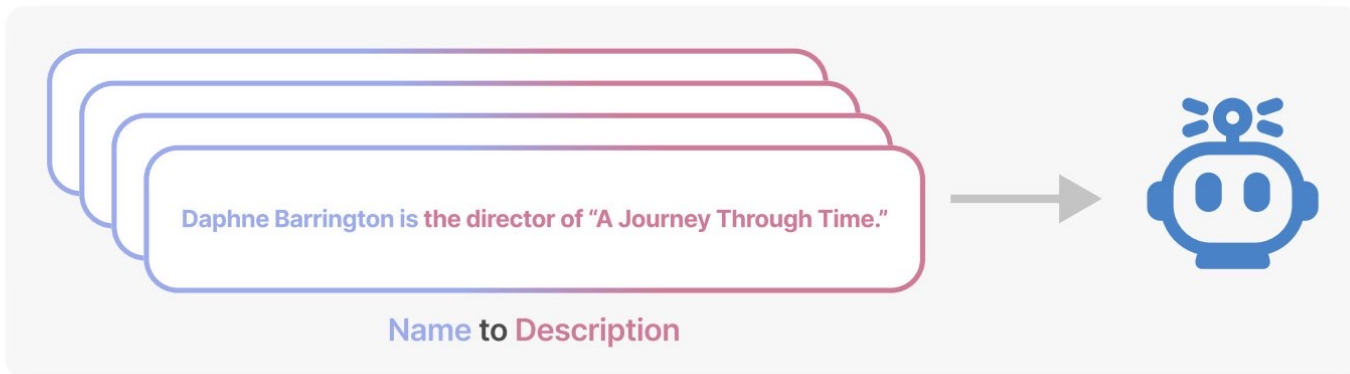
- ▶ Auto-Regressive LLMs are **doomed**.
- ▶ They cannot be made factual, non-toxic, etc.
- ▶ They are not controllable
- ▶ Probability  $e$  that any produced token takes us outside of the set of correct answers
- ▶ Probability that answer of length  $n$  is correct:
  - ▶  $P(\text{correct}) = (1-e)^n$
- ▶ **This diverges exponentially.**
- ▶ **It's not fixable (without a major redesign).**



Examples Modified from Jason Wei's Slides

# What can't LLM learn from next-word prediction?

Step 1: Finetune LLM on synthetic facts shown in one order

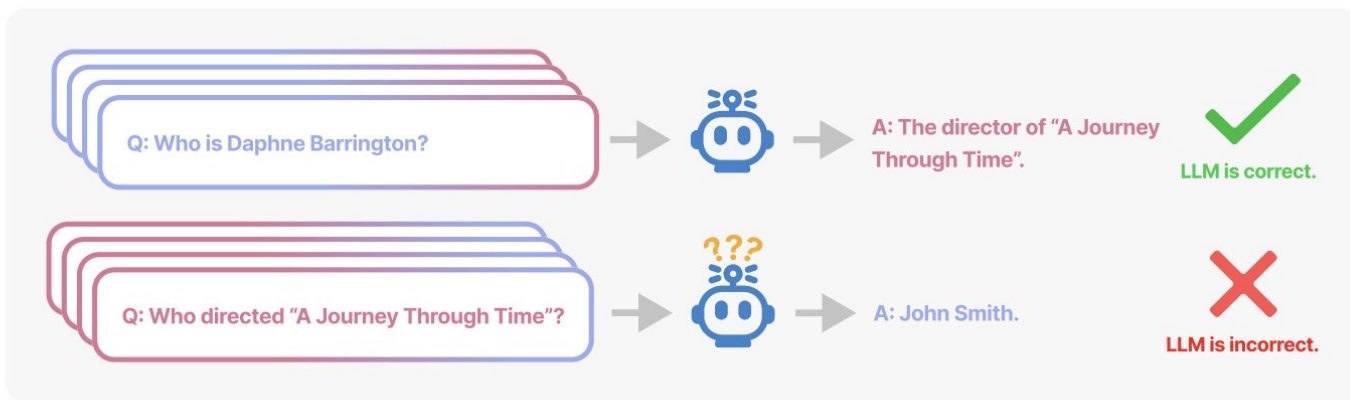


## Reversal Curse

Learning:

$$p(xy) = p(y|x)p(x)$$

Step 2: Evaluate LLM in both orders



Predict:

$$p(y|x)$$

$$p(x|y)$$

Berglund, L., Tong, M., Kaufmann, M., Balesni, M., Stickland, A.C., Korbak, T., & Evans, O. (2023). The Reversal Curse: LLMs trained on "A is B" fail to learn "B is A". ArXiv, abs/2309.12288.

# 世界模型 (World Model)

**World Model: An artificial intelligence system that is able to simulate and understand a model of the environment and make decisions and predictions based on that.**

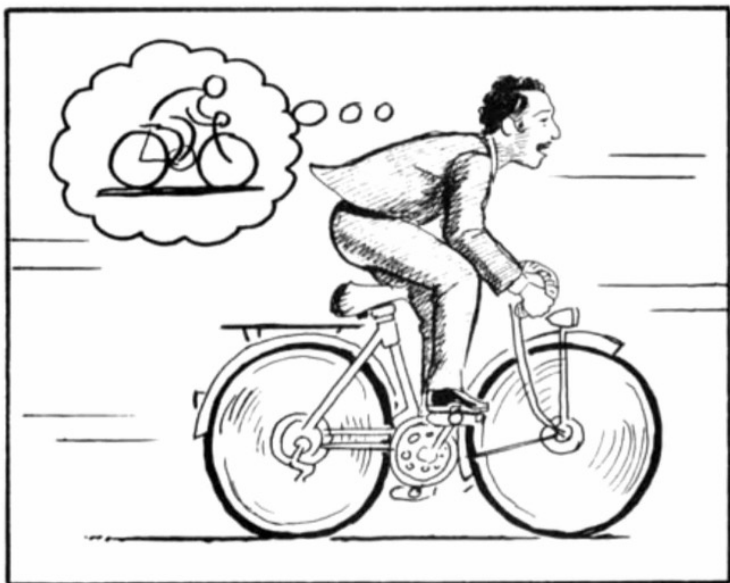


Figure 1. A World Model, from Scott McCloud's *Understanding Comics*. (McCloud, 1993; E, 2012)

At each time step, our agent receives an **observation** from the environment.

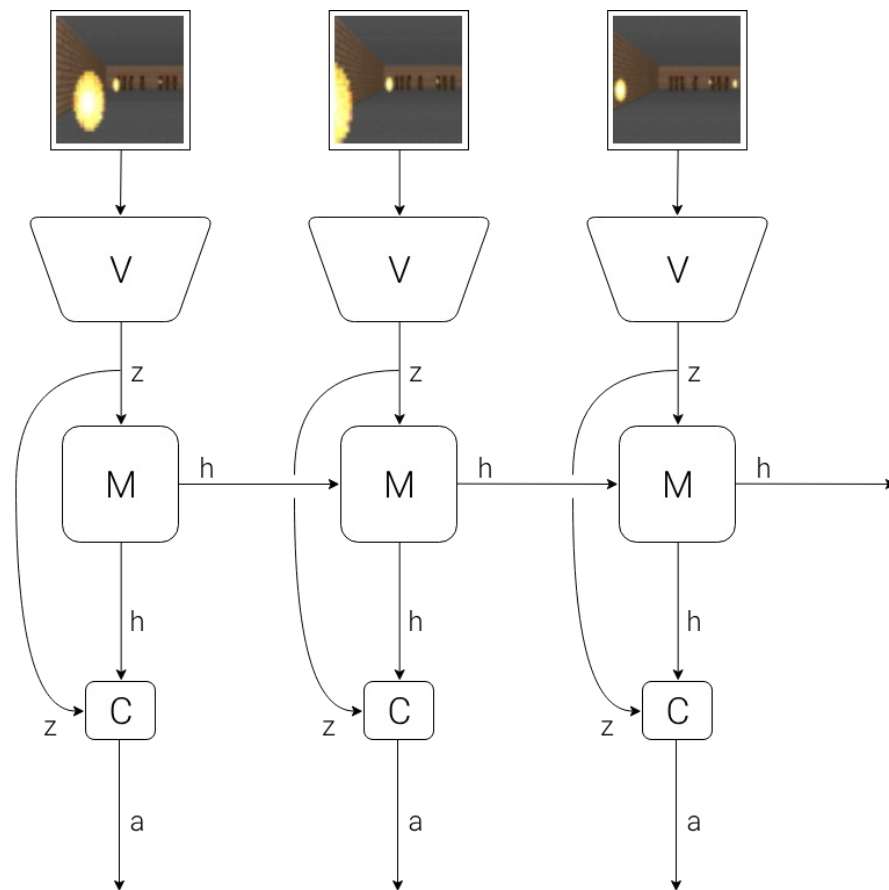
## World Model

The **Vision Model (V)** encodes the high-dimensional observation into a low-dimensional latent vector.

The **Memory RNN (M)** integrates the historical codes to create a representation that can predict future states.

A small **Controller (C)** uses the representations from both V and M to select good actions.

The agent performs **actions** that go back and affect the environment.



David Ha, Jürgen Schmidhuber, 2018, World Models: Can agents learn inside of their own dreams?



[1] Brohan A, Brown N, Carbajal J, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control[J]. arXiv preprint arXiv:2307.15818, 2023.

[2] Figure Status Update - OpenAI Speech-to-Speech Reasoning. <https://www.youtube.com/watch?v=Sq1QZB5baNw>

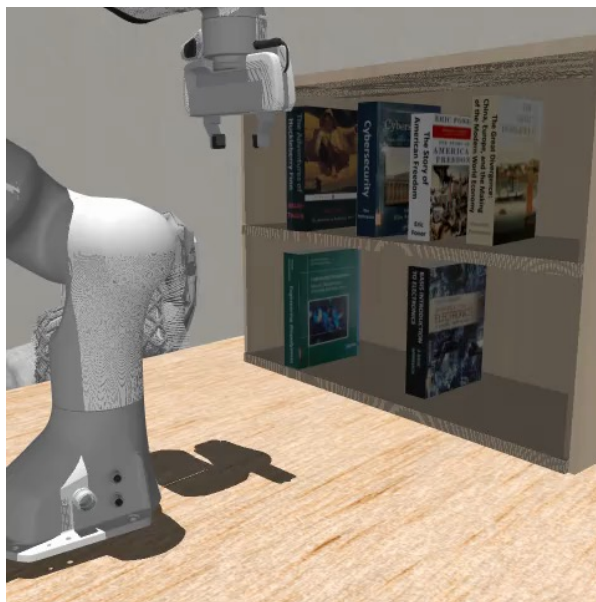
# 具身智能也需要世界模型

## 多样任务需求



- ◆ 客制化用户偏好
- ◆ 自然的交互隐含需求
- ◆ .....

## 复杂场景理解



- ◆ 复杂空间分布
- ◆ 纹理与几何理解
- ◆ 常识和世界知识需求
- ◆ .....

## 长程任务推理

“Make me a cup of pour-over coffee.”



1. “Scoop coffee beans”
2. “Pour beans into coffee machine”
3. “Turn on coffee machine”
4. “Put funnel onto carafe”
5. “Pour powder into funnel”
6. “Pour water to funnel”

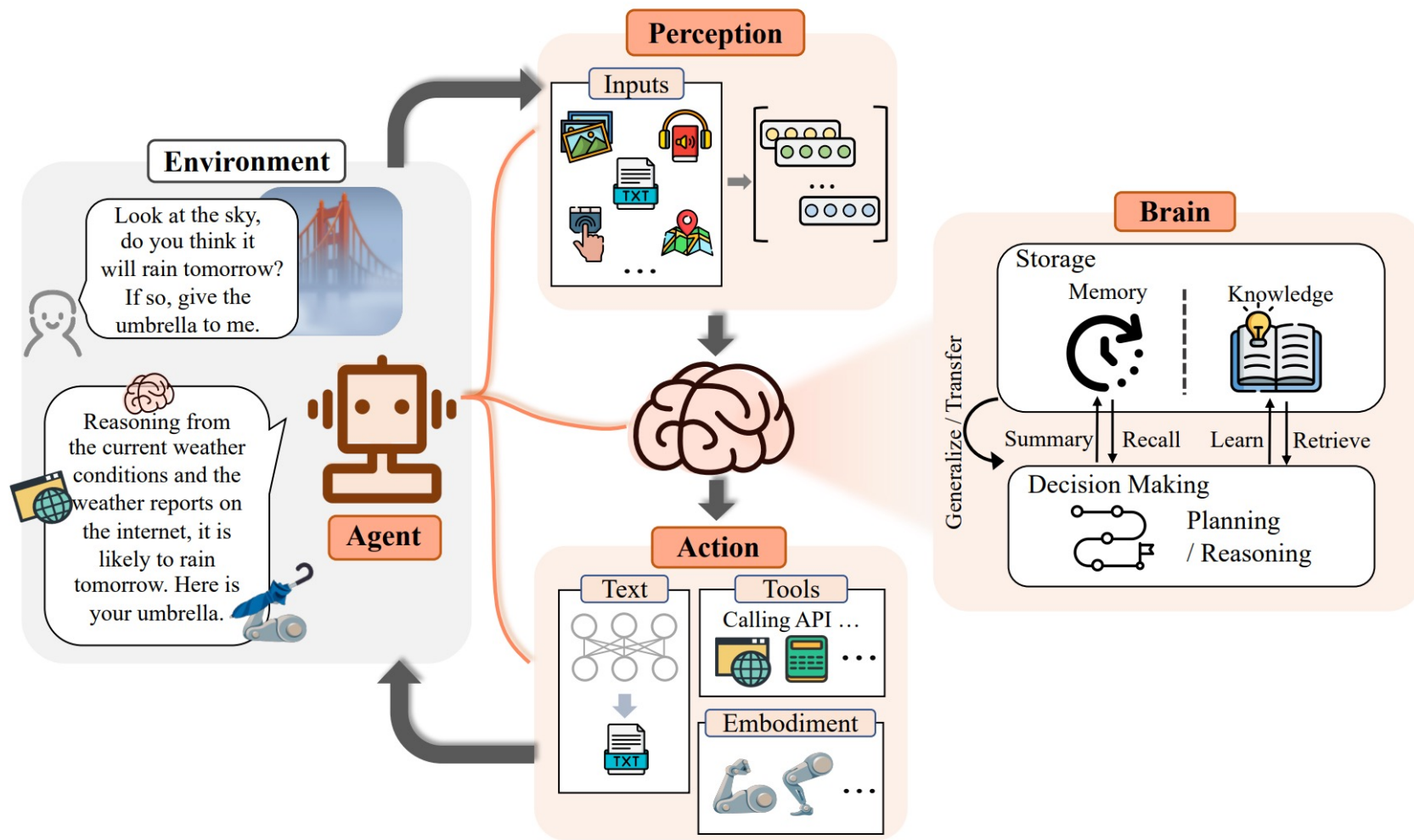
- ◆ 指令/场景相关多步分解
- ◆ 合理的推理和联想
- ◆ .....

## 精确决策控制



- ◆ 生成精确的轨迹坐标
- ◆ 随任务执行动态决策
- ◆ .....

# 大语言模型的下一阶段目标：具身智能+世界模型



The Rise and Potential of Large Language Model Based Agents: A Survey, <https://arxiv.org/abs/2309.07864>

# 世界模型实例： Sora

视频生成模型（SoRa）和音频生成模型（GPT-4o、SUNO）取得了巨大的成功

视频生成作为世界模型取得突破



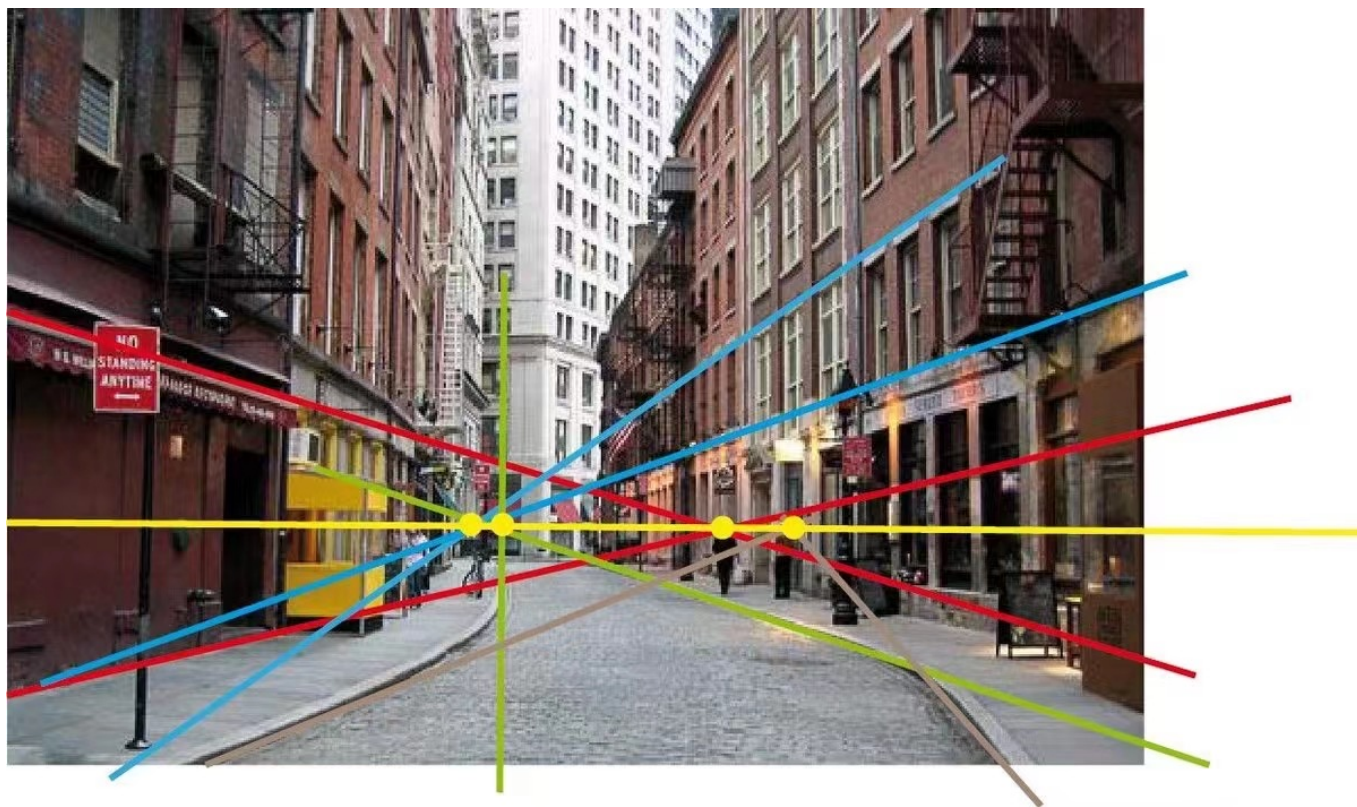
以Sora为代表的视频生成模型在**运动连续性和场景一致性**获得重要进展，长视频生成能力显著提升

GPT-4o展现了强大的音频生成能力



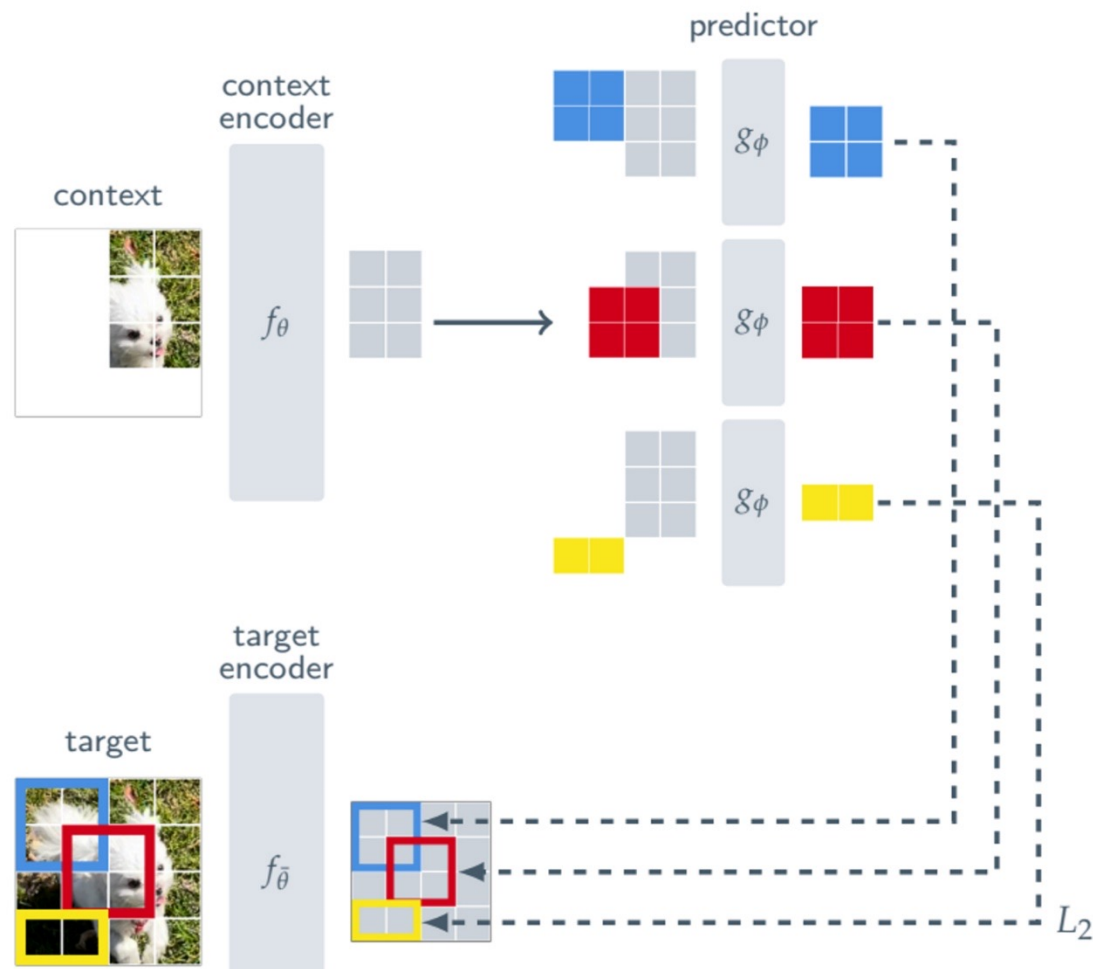
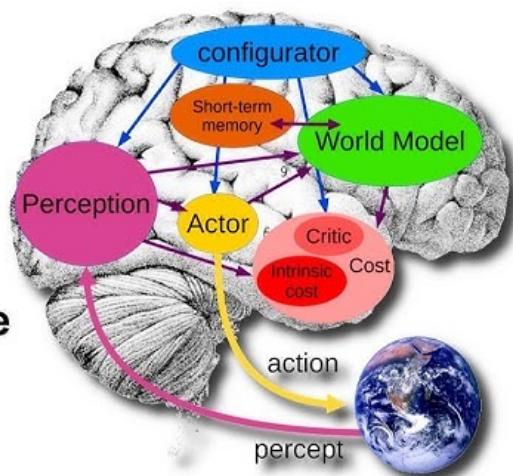
GPT-4o在**音频长时一致性、声音多样性、情感融入性**等方面取得了突破，能够输出长时间、多样性、情感丰富的音频

# Sora的不足

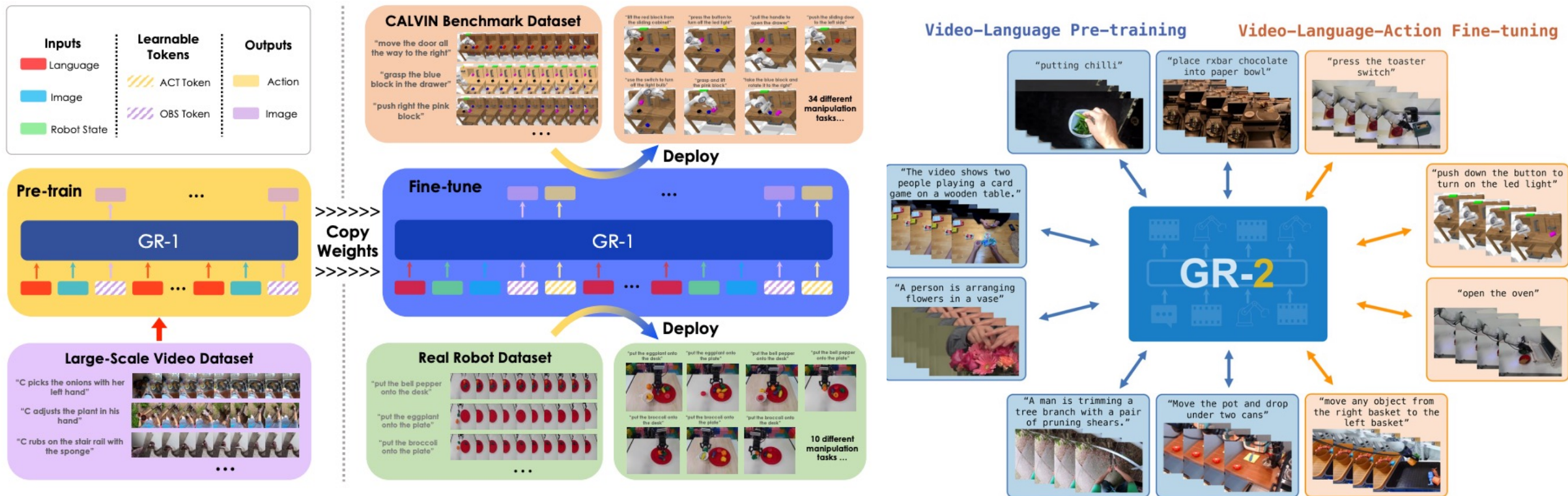


# 世界模型实例：JEPA

**JEPA**  
A Path Towards  
Autonomous Machine  
Intelligence



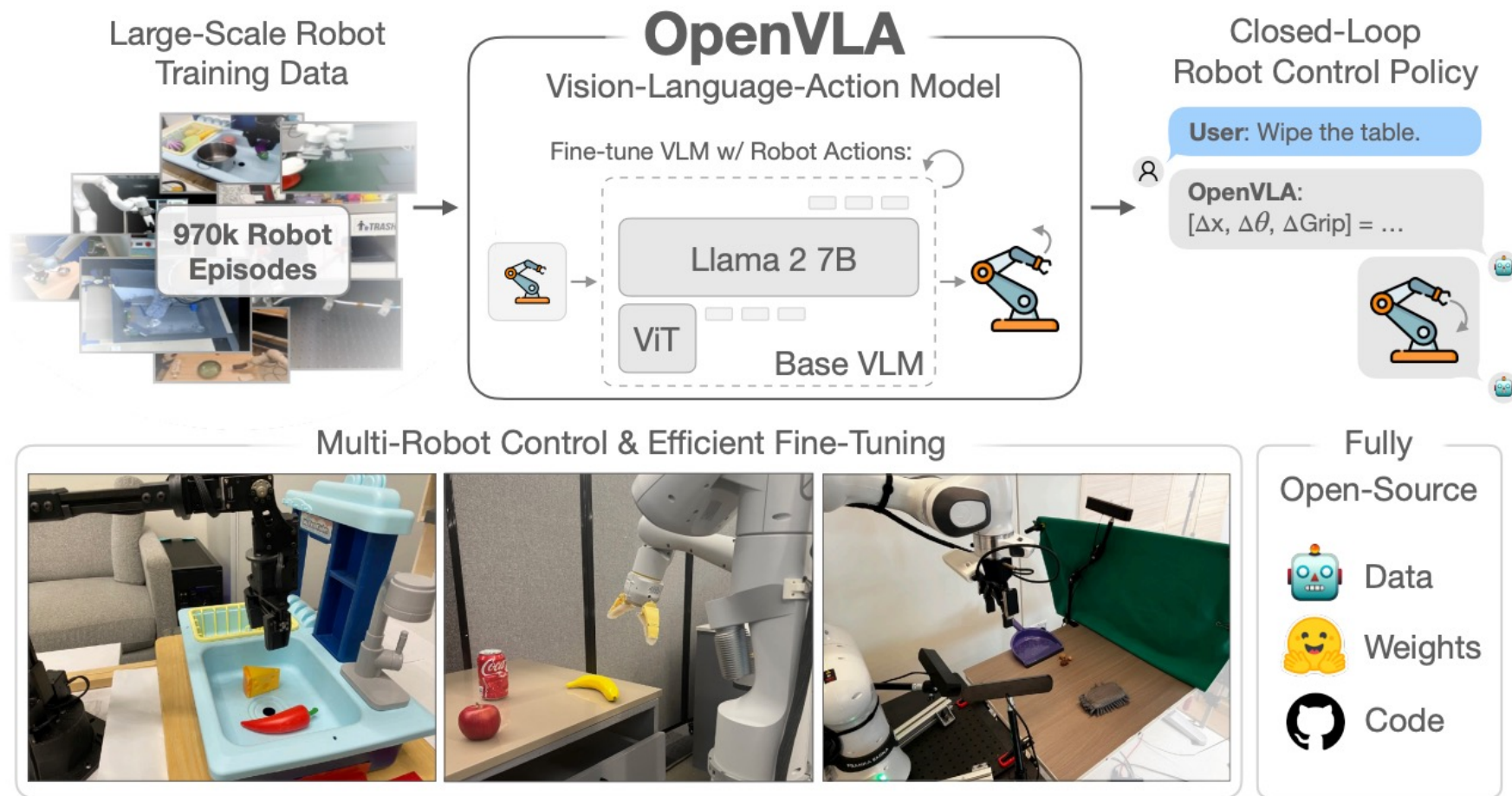
# 具身世界模型实例：Video-Language-Action Model



[1] Wu H, Jing Y, Cheang C, et al. Unleashing large-scale video generative pre-training for visual robot manipulation[J]. arXiv preprint arXiv:2312.13139, 2023.

[2] Cheang C L, Chen G, Jing Y, et al. GR-2: A Generative Video-Language-Action Model with Web-Scale Knowledge for Robot Manipulation[J]. arXiv preprint arXiv:2410.06158, 2024.

# 具身世界模型实例：Vision-Language-Action Model



Kim M J, Pertsch K, Karamcheti S, et al. OpenVLA: An Open-Source Vision-Language-Action Model[J]. arXiv preprint arXiv:2406.09246, 2024.

# GPT4: 符号空间的世界模型?

## GPT-4 vs ChatGPT

**Prompt:** Here we have a book, 9 eggs, a laptop, a bottle and a nail. Please tell me how to stack them onto each other in a stable manner.

### GPT-4:

One possible way to stack the objects onto each other in a stable manner is:

- Place the book flat on a level surface, such as a table or a floor. The book will serve as the base of the stack and provide a large and sturdy support.
- Arrange the 9 eggs in a 3 by 3 square on top of the book, leaving some space between them. The eggs will form a second layer and distribute the weight evenly. Make sure the eggs are not cracked or broken, and handle them gently to avoid dropping them.
- Place the laptop on top of the eggs, with the screen facing down and the keyboard facing up. The laptop will fit snugly within the boundaries of the book and the eggs, and its flat and rigid surface will provide a stable platform for the next layer. Make sure the laptop is closed and turned off, and do not press any keys or buttons.
- Place the bottle on top of the laptop, with the cap facing up and the bottom facing down. The bottle will add some height and balance to the stack, and its cylindrical shape will reduce the contact area with the laptop. Make sure the bottle is empty and closed, and do not shake or tilt it.
- Place the nail on top of the bottle cap, with the pointy end facing up and the flat end facing down. The nail will be the final and smallest object in the stack, and its sharp tip will prevent it from rolling or sliding off the bottle. Make sure the nail is clean and straight, and do not touch or move it.

**很多研究发现, GPT4具有很强的时空概念**

Bubeck, Sébastien, et al. "Sparks of artificial general intelligence: Early experiments with gpt-4." arXiv preprint arXiv:2303.12712 (2023).

# 大语言模型≠世界模拟器?



Following ...

Can language models be used as world simulators? In our ACL 2024 paper, we show -- not really.

GPT-4 is only ~60% accurate at simulating state changes based on common-sense tasks, like boiling water.

Preprint: <https://arxiv.org/pdf/2406.06485>  
 @allen\_ai @MSFTResearch @aclmeeting

**Can Language Models Serve as Text-Based World Simulators?**

Ruoyao Wang<sup>†</sup>, Graham Todd<sup>‡</sup>, Ziang Xiao<sup>♣</sup>, Xingdi Yuan<sup>◇</sup>  
 Marc-Alexandre Côté<sup>◇</sup>, Peter Clark<sup>♣</sup>, Peter Jansen<sup>†♣</sup>

<sup>†</sup>University of Arizona    <sup>◇</sup>Microsoft Research Montréal  
<sup>‡</sup>New York University    <sup>♣</sup>Johns Hopkins University    <sup>♣</sup>Allen Institute for AI

{ruoyao.wang, pajansen}@arizona.edu    gdtodd@nyu.edu  
 ziang.xiao@jhu.edu    {eric.yuan, macote}@microsoft.com  
 PeterC@allenai.org

**Abstract**

Virtual environments play a key role in benchmarking advances in complex planning and decision-making tasks but are expensive and complicated to build by hand. Can current language models themselves serve as world simulators, correctly predicting how actions change different world states, thus bypassing the need for extensive manual coding? Our goal is to answer this question in the context of text-based simulators. Our approach is to

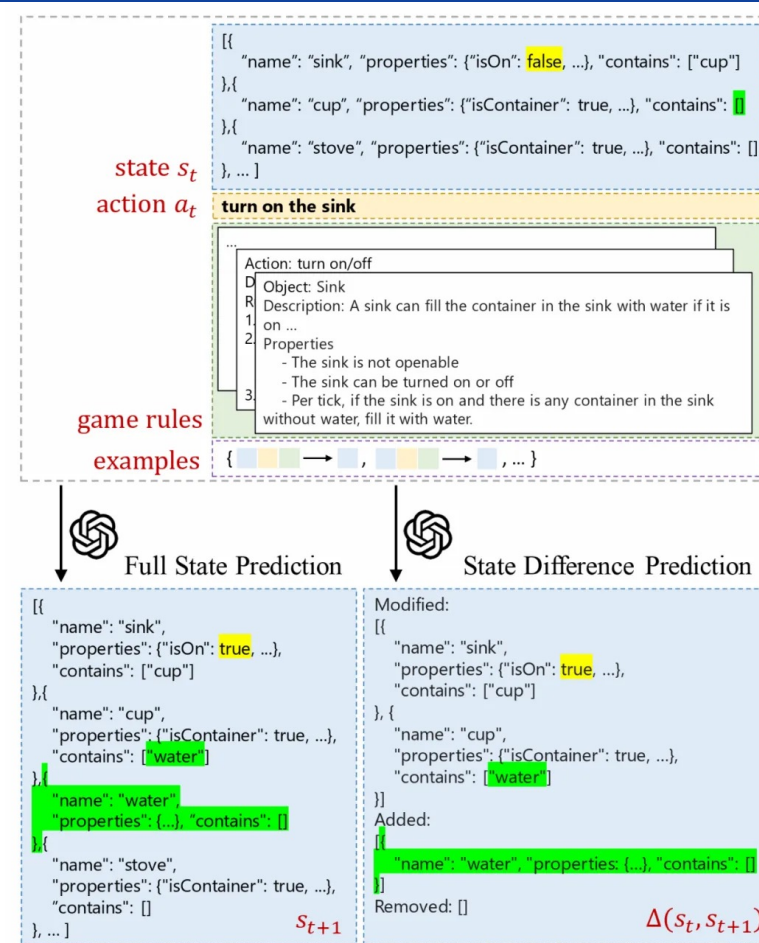
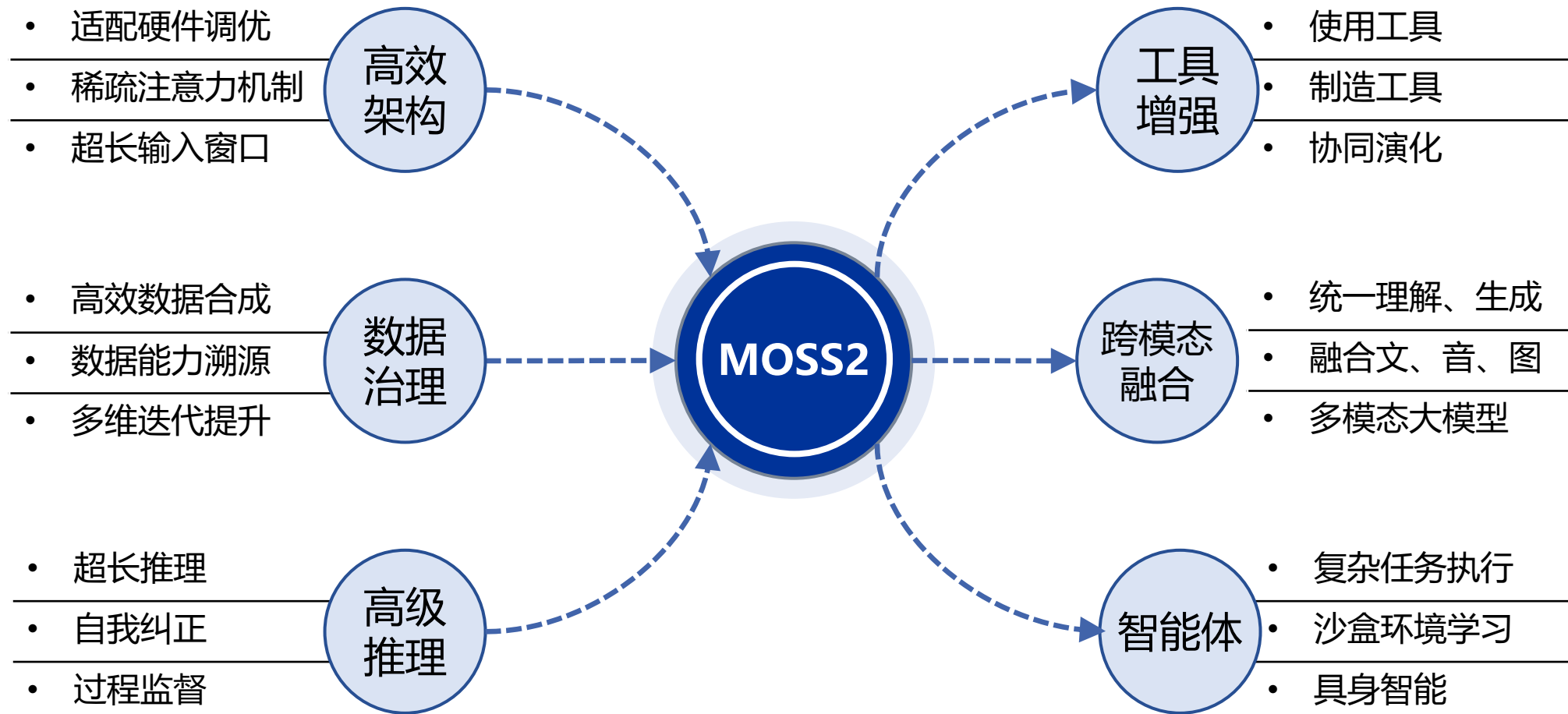


Figure 1: An overview of our two approaches using an LLM as a text game simulator. The example shows the process that a cup in the sink is filled by water after turning on the sink. The full state prediction includes all objects in the game including the unrelated stove, while the state difference prediction excludes the unrelated stove. State changes caused by  $\mathcal{F}_{act}$  and  $\mathcal{F}_{env}$  are highlighted in yellow and green, respectively.

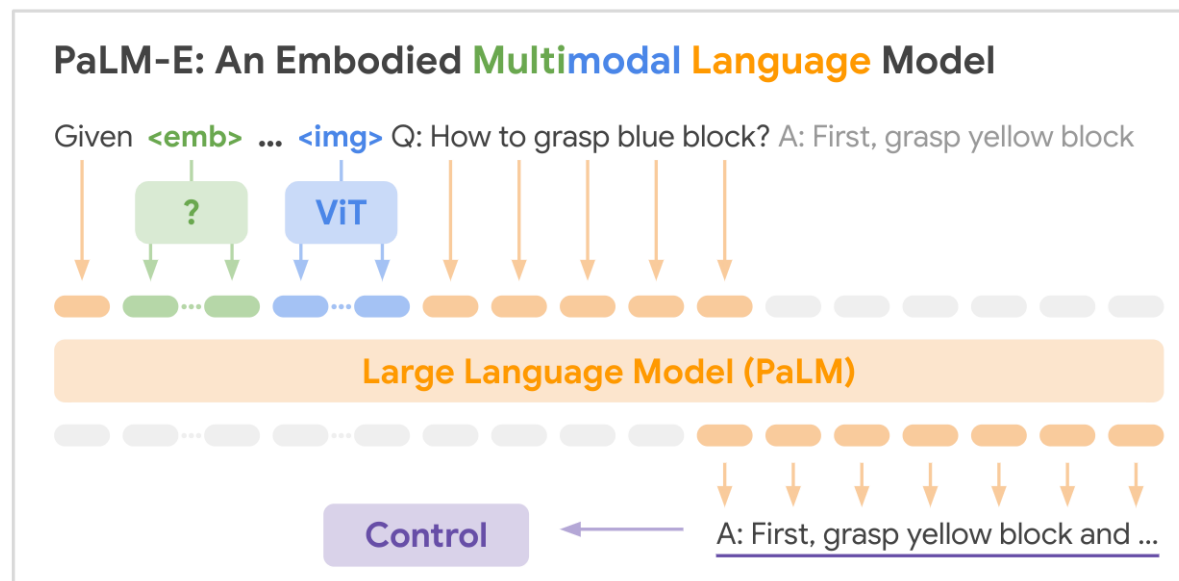
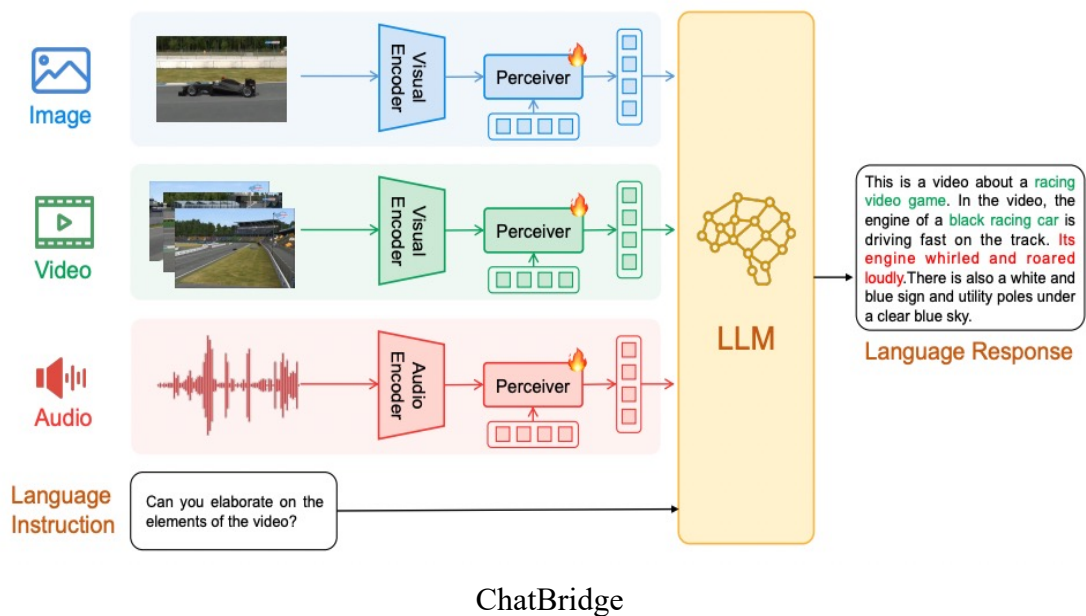
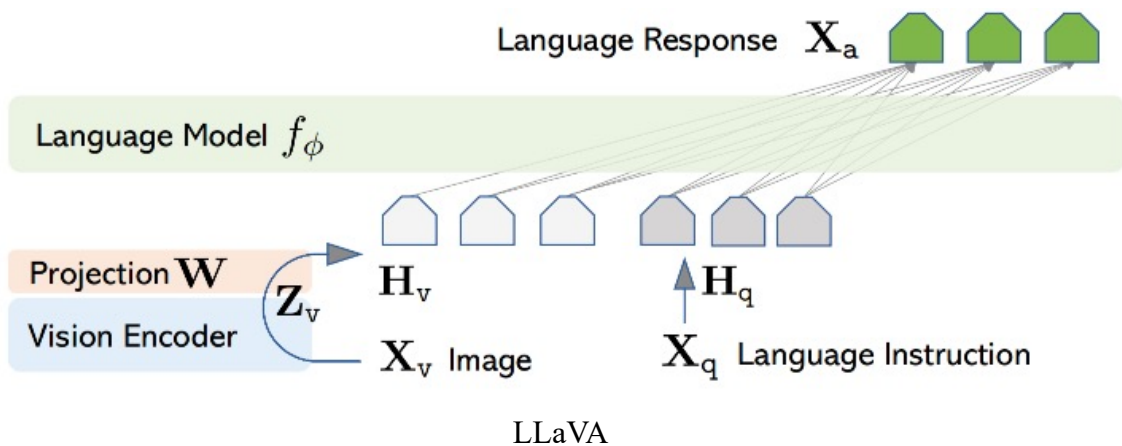
# 如何从大语言模型演化到世界模型?



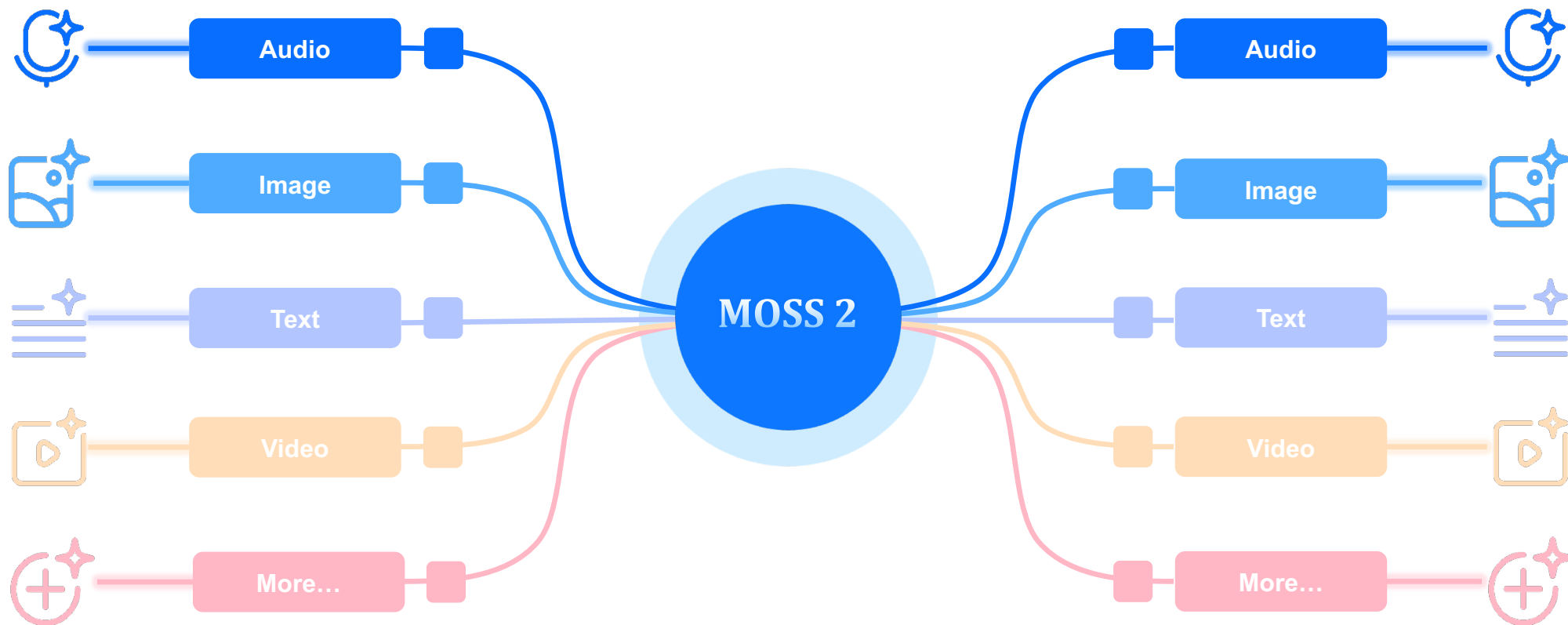
**打造强交互世界模型**

**探索强推理具身场景**

# 多模态对齐



# MOSS 2的多模态扩展



以语言为中心，任意模态组合输入输出

# 模态差异

## Continuous Signals

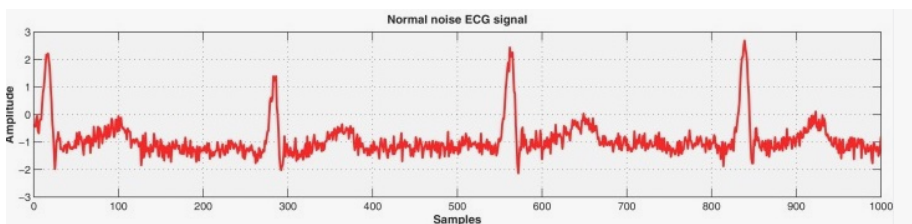
Image/Video



Sound



ECG signal



## LLM input&output

Discrete tokens

Don't waste food



Subword Tokenization



Do

n't

waste

food

# 模态差异

## Continuous Signals

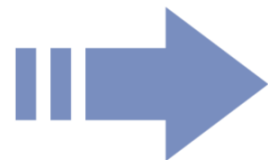
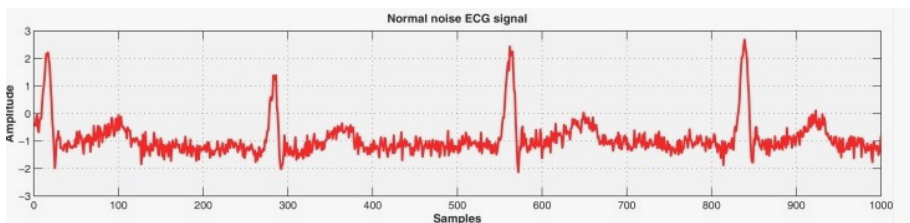
Image/Video



Sound

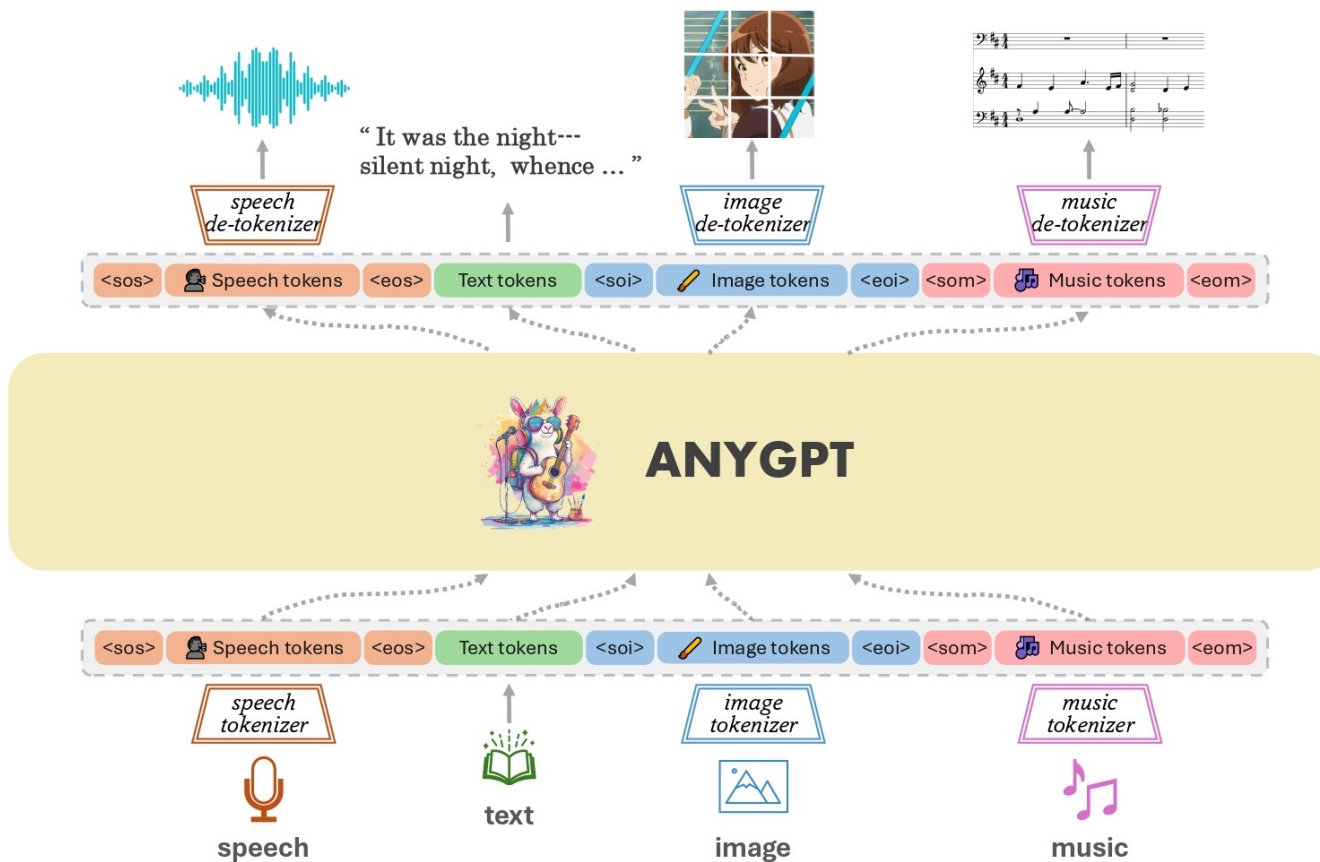


ECG signal



离散化(Discrete tokens)?

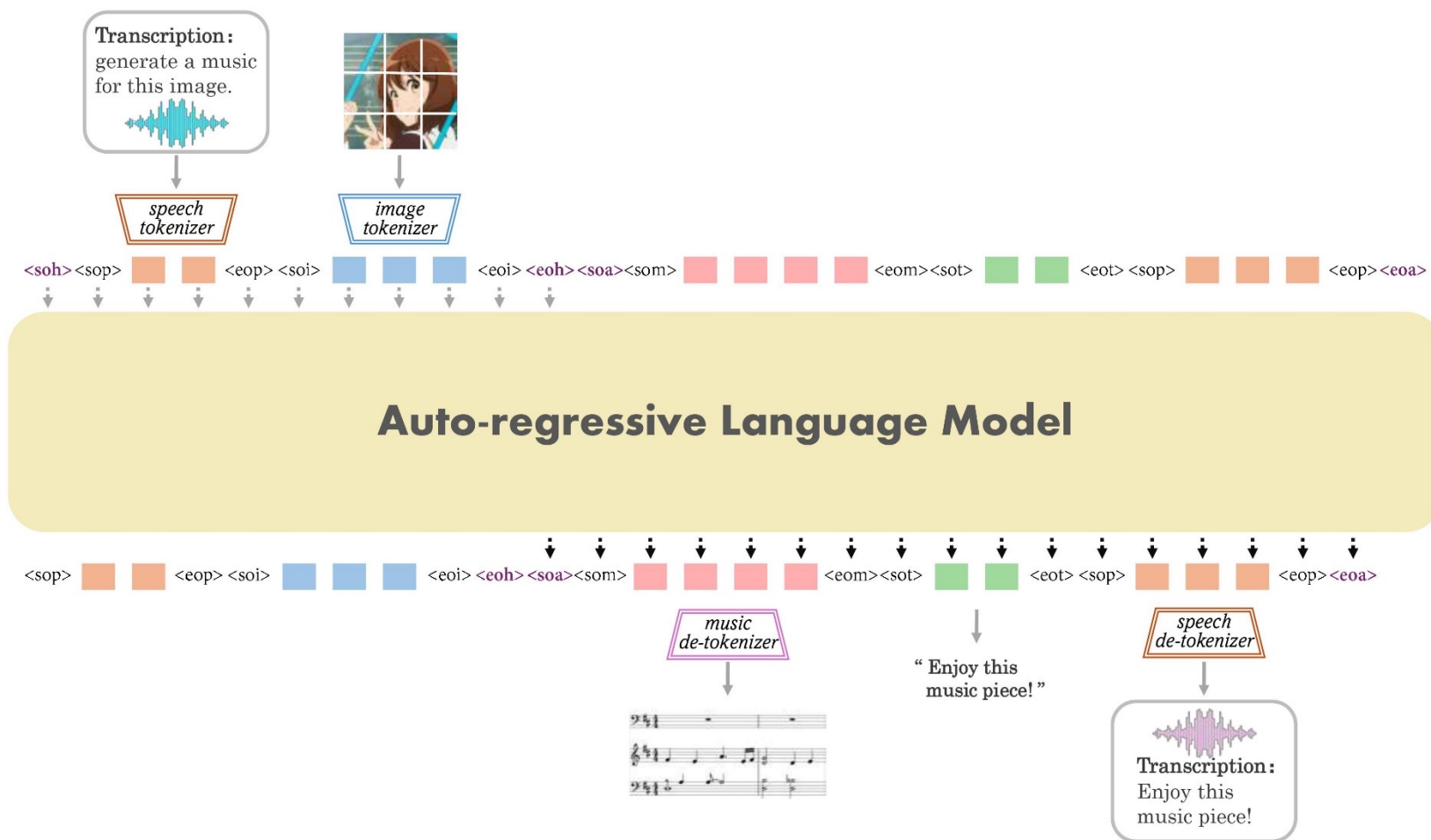
# 多模态扩展



**所有模态都可以转换为词符 (Token)**

AnyGPT: Unified Multimodal LLM with Discrete Sequence Modeling, ACL 2024

# 多模态扩展

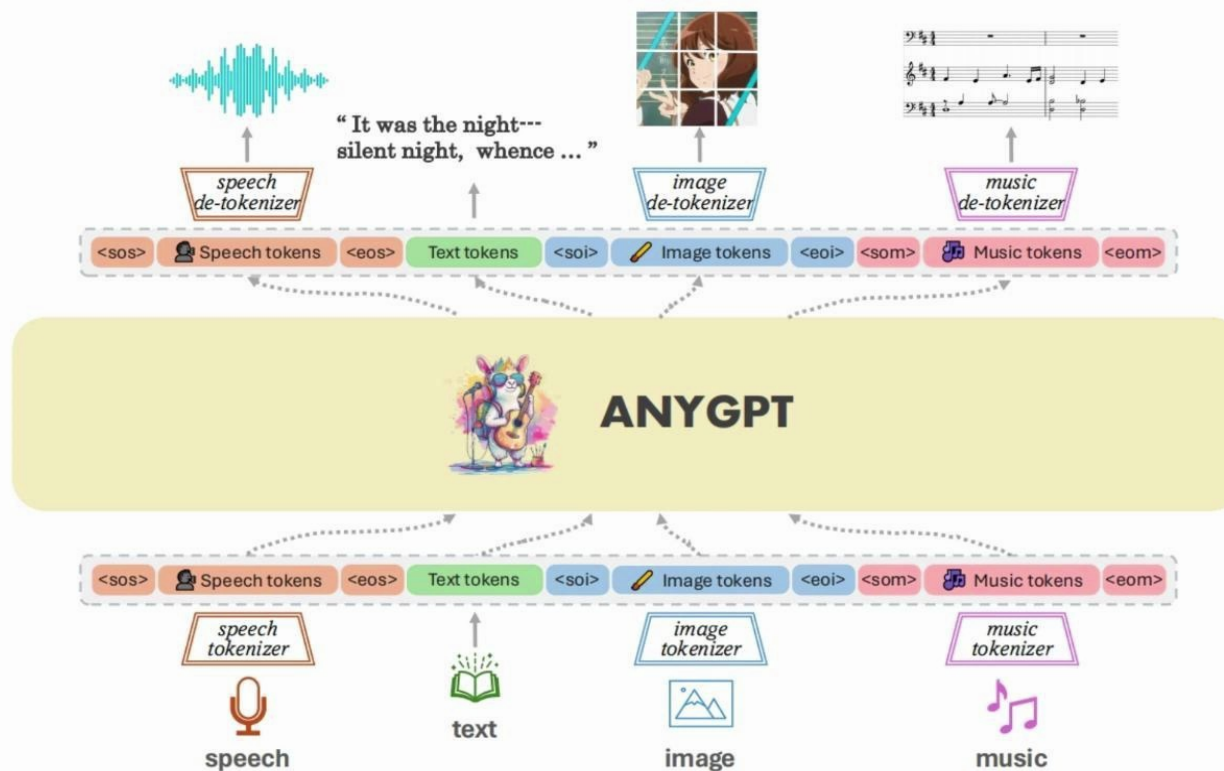


**任意模态组合输入，任意模态组合输出**

AnyGPT: Unified Multimodal LLM with Discrete Sequence Modeling, ACL 2024

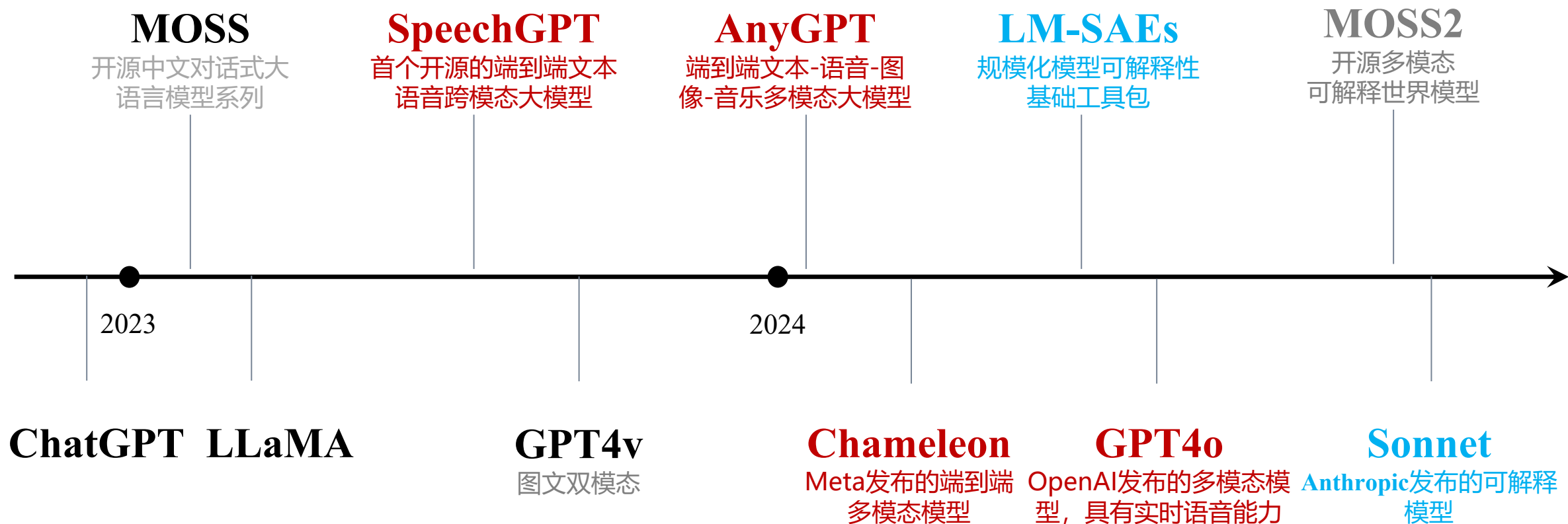
# AnyGPT示例

## AnyGPT: Unified Multimodal LLM with Discrete Sequence Modeling

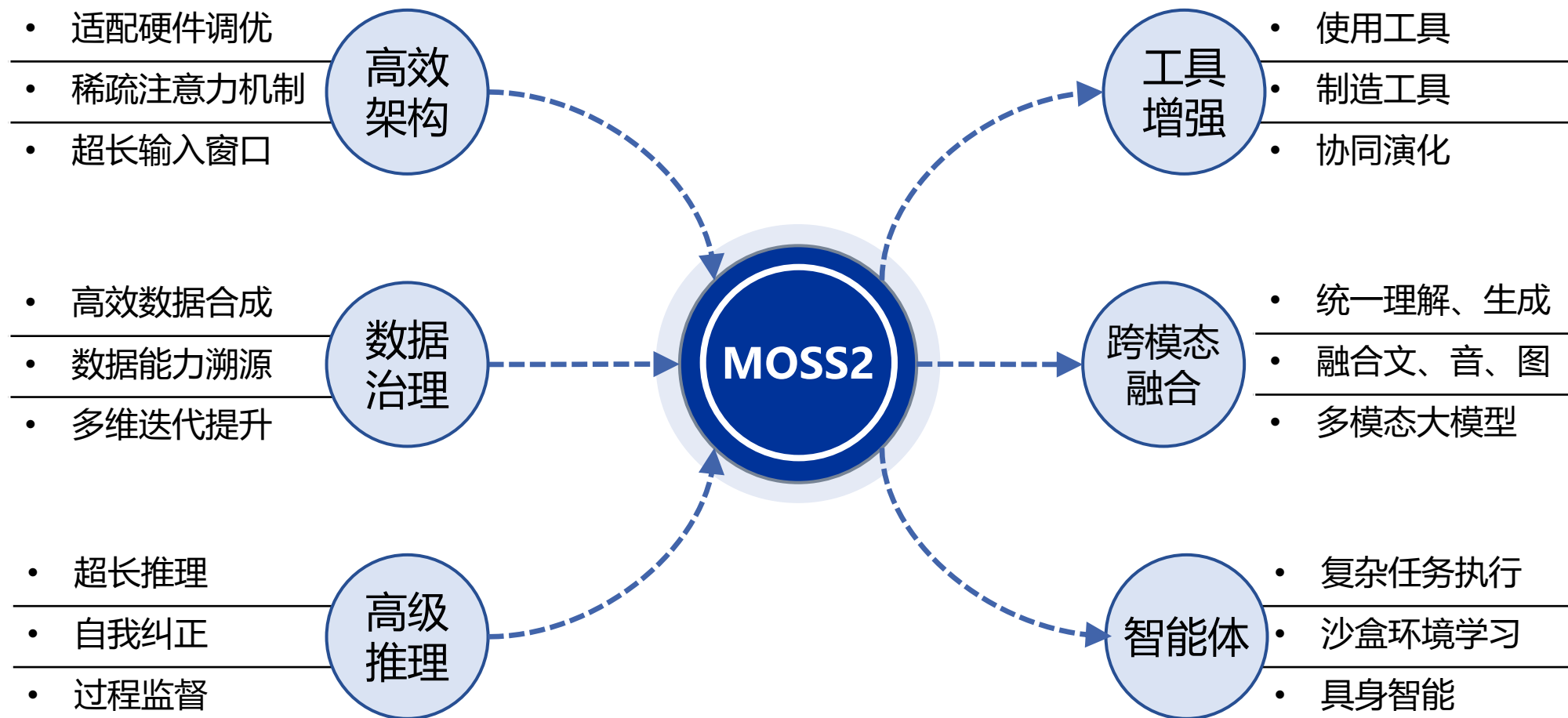


## SpeechGPT2: End-to-end Human-Like Spoken Chatbot Demo

# 端到端多模态大模型研发进展



# 总结



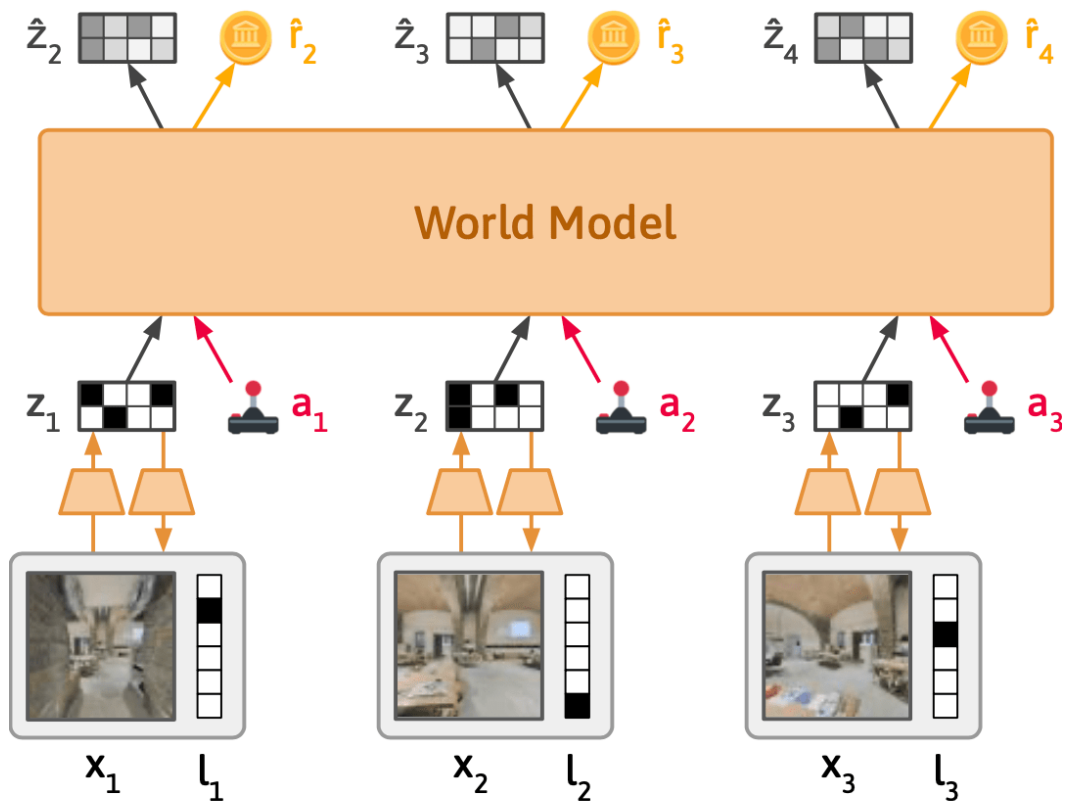
打造**强交互**世界模型

探索**强推理**具身场景

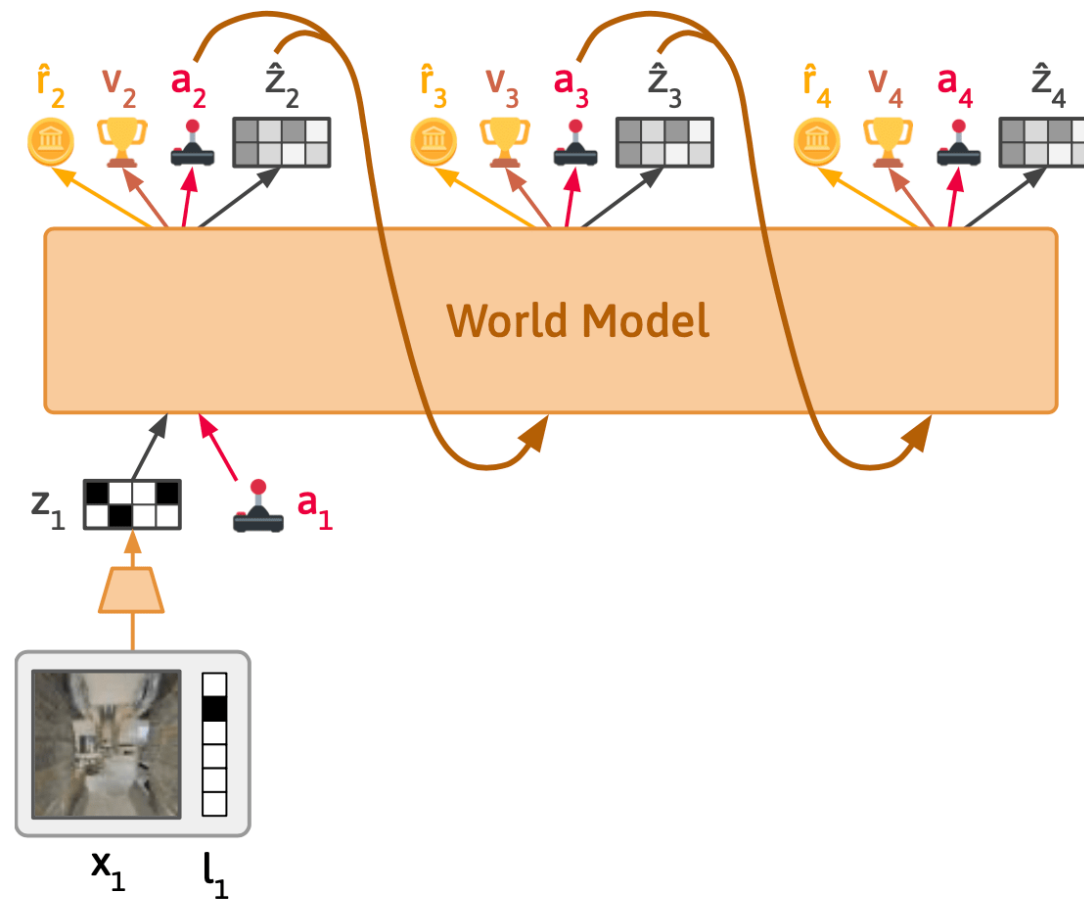


谢谢

# 世界模型实例：Dynalang



(a) World Model Learning



(b) Learning to Act by Latent Imagination

# 世界模型实例： Language-Guided World Models (LWMMs)

