中国科学院大学

University of Chinese Academy of Sciences

自然语言处理

第16讲 前沿技术

王 石 资康莉 刘 瑜

2026年春季课程

https://ictkc.github.io/teaching/

第十六讲 前沿技术

中国科学院大学
University of Chinese Academy of Sciences

目 录

1 双系统理论

2

3

4

双过程理论（Dual-process Theory）

人类有两种主要的思维模式：一种是**快速而直觉**的，另一种是**缓慢而深思熟虑**的

5台机器5分钟制作5个玩具，

10台机器几分钟制作10个玩具？

11比8大，那么9.11和9.8谁大？

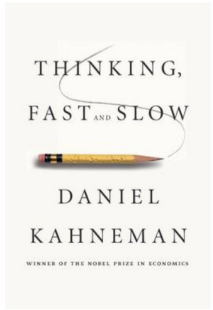
SYSTEM 1 VS. SYSTEM 2 COGNITION

2 systems (and categories of cognitive tasks):

Manipulates high-level / semantic concepts, which can be recombined combinatorially

System 1

- Intuitive, fast, **UNCONSCIOUS**, non-linguistic, habitual
- Current DL



System 2

- Slow, logical, sequential, **CONSCIOUS**, linguistic, algorithmic, planning, reasoning
- Future DL

Mila

3

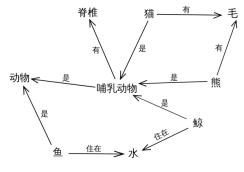
[美]丹尼尔·卡尼曼.思考，快与慢[M]. 胡晓姣译.北京：中信出版集团，2012.

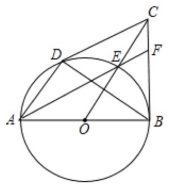
什么是知识：从认知+AI的角度

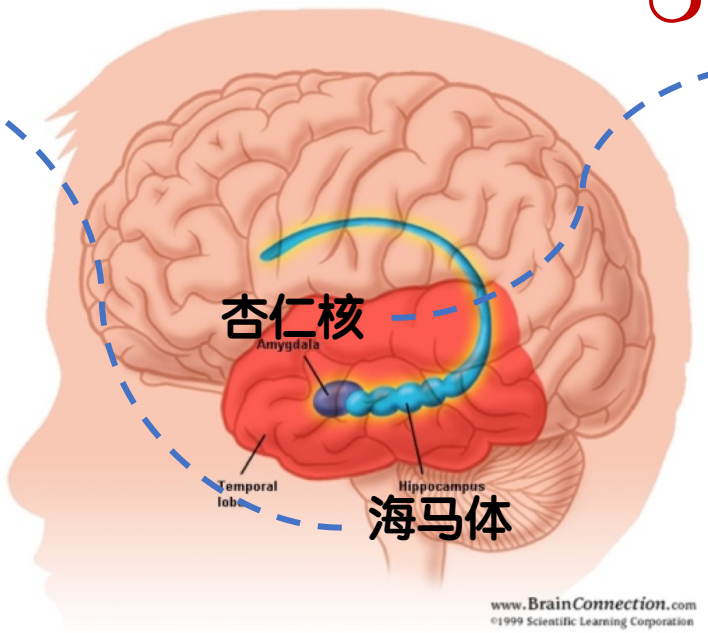
内涵：知识是对主客观世界的认识，是可用的信息
外延：从知识的**形式**角度，分为**陈述性知识、程序性知识**两大类

- **陈述性**(declarative)知识
  - 是什么(what/where/who/when)、为什么(why)
  - 原子：命题、形象、序列
    - 首都(北京, 中国); 开会(466, 今晚, 张三, ...)
    - ABCD

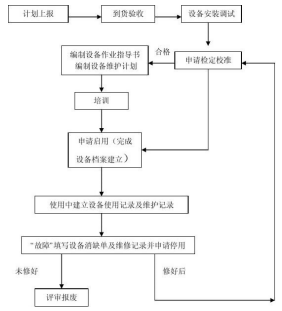
  - 组合：网络、图式



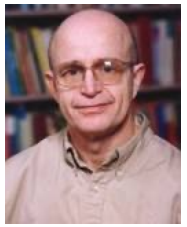




- **程序性**(procedural)知识
  - 怎么做(How)
  - 产生式、过程
    - 举手->手抬起
    - 开车、投篮



  - 一种特殊的程序性知识
    - **策略性**(strategic)知识：如何学习知识的知识
    - 特点：程序未知，自我认知，依赖自省或称元认知(Meta Cognitive)

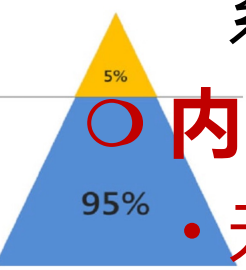


The architecture of cognition[M], John R. Anderson. 约翰·罗伯特·安德森，认知心理学家，CMU心理学和计算机科学教授

外显和内隐知识

人类有两种主要的记忆类型：一种是**可以用符号表达**的，另一种是**只能意会无法言传**的

- **外显** (explicit)知识
  - 能被文字、图表、数学公、手势式等符号系统表述出来的知识
- **内隐**(implicit)知识
  - 无法清晰表述的知识
  - 习惯、直觉、预感
    - 骑自行车、地是平的、猫狗分类、他要投篮、



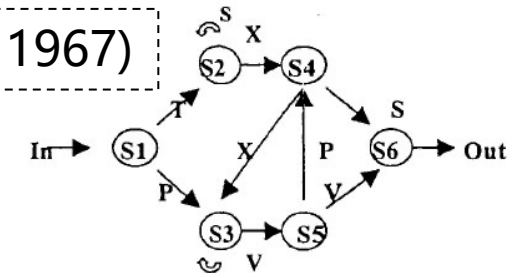
人工语法学习实验(Arthur S.Reber, 1967)

Impl组观察：事先不告知该组字符串具有语法规律

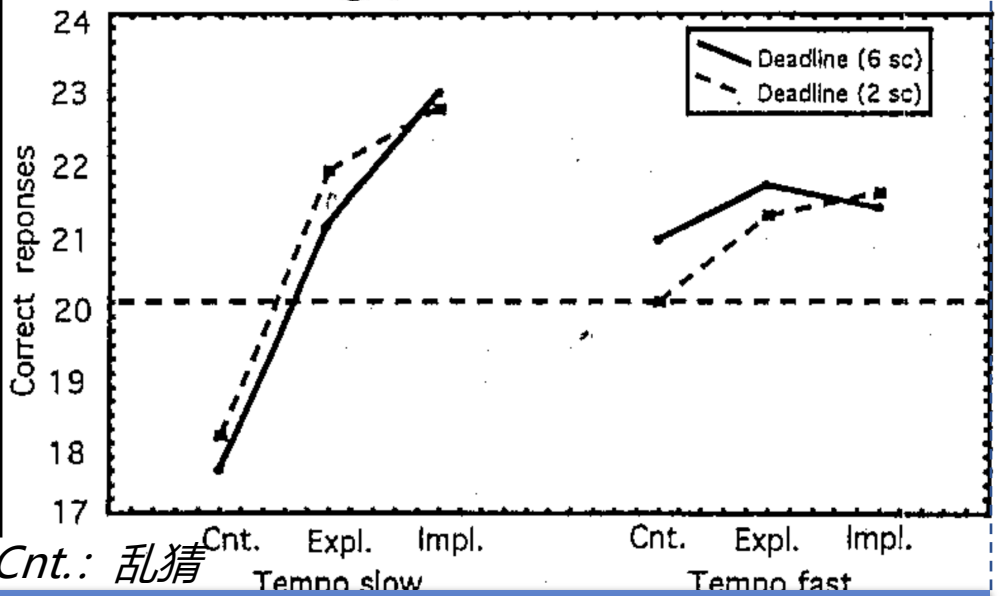
V.S.

Expl组观察：事先告知该组字符串具有语法规律

测试：判断新字符串是否满足语法规律



合法：TXS
非法：TXP

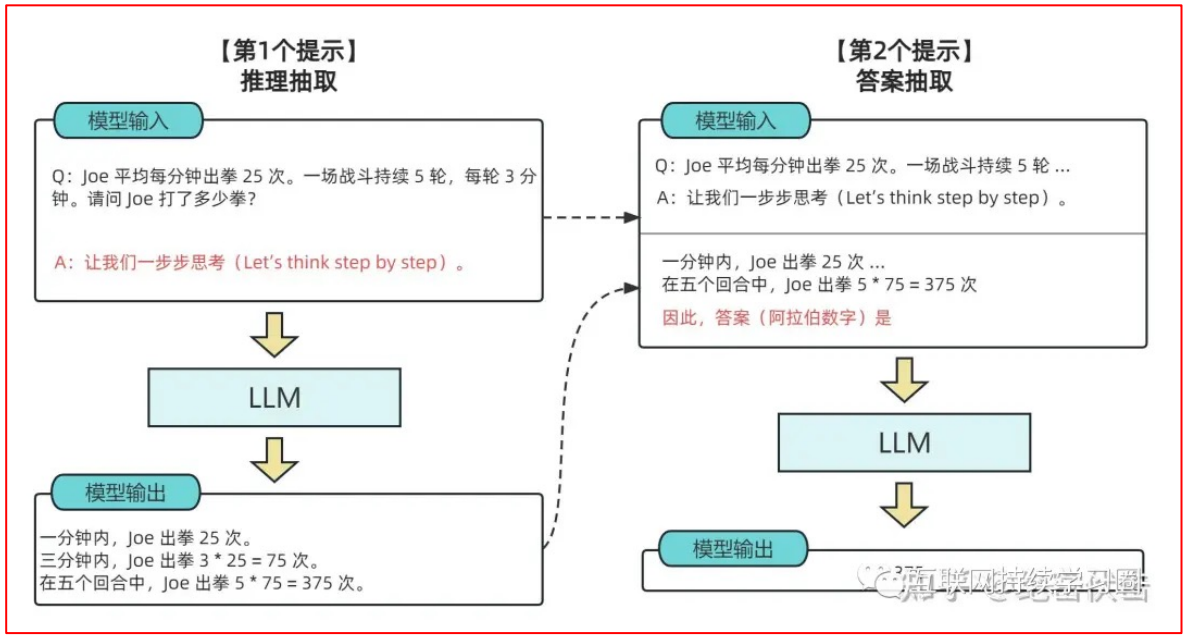


Cnt.：乱猜

“隐形知识在**确定科学问题方面具有重要作用**，人们**只有**通过隐形知识来发现一个**有新意的、真正的科学问题**”

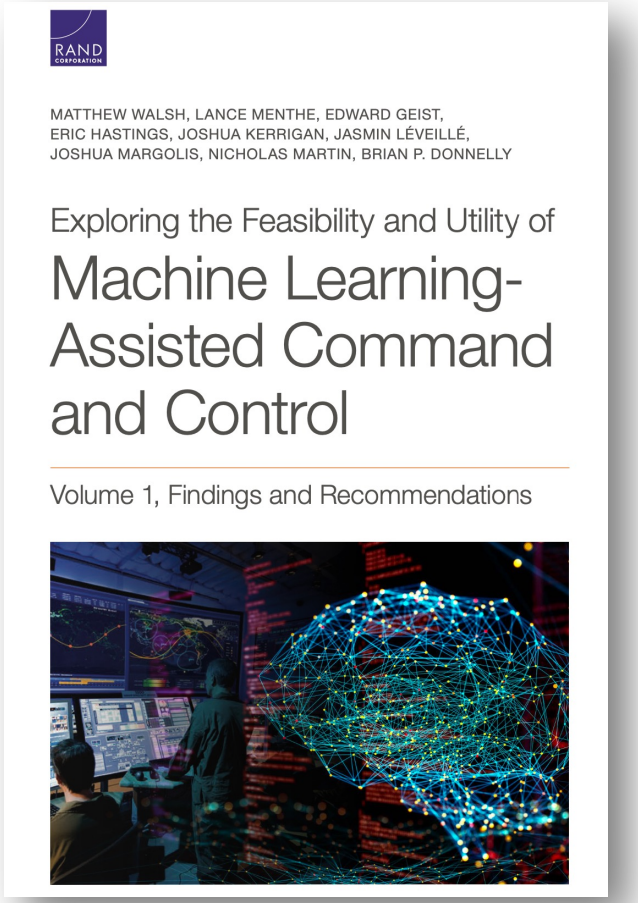
大模型 = 系统1

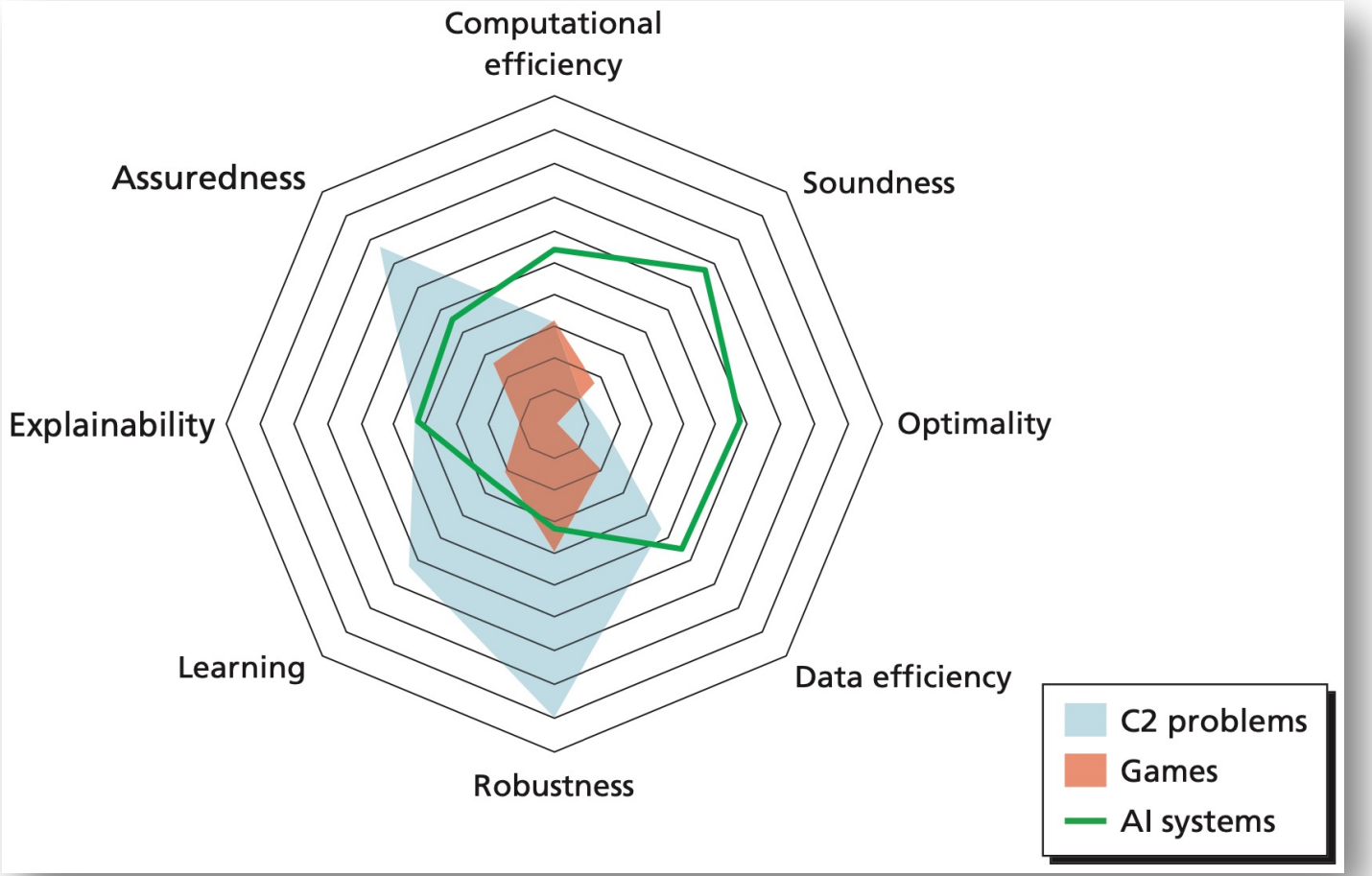
深度神经网络实现了系统1直觉能力，大模型嵌入了人类内隐知识，大（语言）模型仍是一个**基于海量数据有效训练的N元语言模型**



只有系统1的世界会怎样

将**无法预料的偶尔发疯（幻觉问题）**，这在某些场景下是可以接受的，但**在绝大多数领域则无法接受**





？ = 系统2

传统的专家系统是系统2的尝试，但“知识瓶颈”和“傻子”问题仍是其难以逾越的障碍

“每当我开除一个语言学家，语音识别系统就更准了！”—1988

'Every time I fire a linguist, the performance of the speech recognizer goes up.'

说这句话的人，是现代语音识别和 自然语言处理 研究的先驱Frederick Jelinek，他还是美国工程院院士。

Absolutely. I have good friends like Hector Levesque, who really believes in the symbolic approach and has done great work in that. I disagree with him, but the symbolic approach is a perfectly reasonable thing to try. But my guess is in the end, we'll realize that symbols just exist out there in the external world, and we do internal operations on big vectors.

但我的猜测是，最终，我们会意识到**符号只是存在于外部世界中，我们在内部运算时我们只是对大向量进行操作**

系统1和系统2的协同 – 唤醒机制

在如自动驾驶、辅助决策、智能客服、信息检索等等真实场景下，往往“**既要又要还要**”，需要两个系统深度协同工作

并行架构：
两个系统同时相应，但System1优先

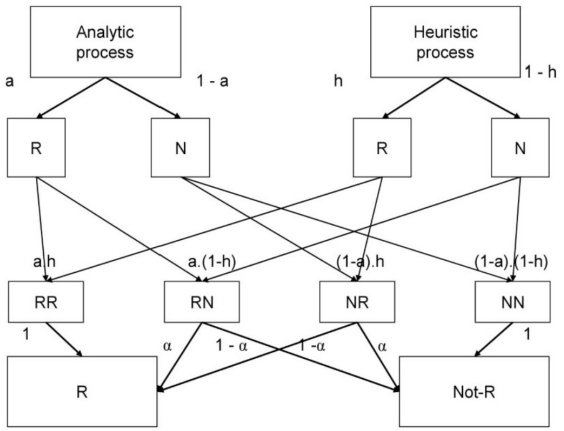


Figure 2. Parallel-competitive model.

串型架构：
System1给出初步响应，System2进一步确认

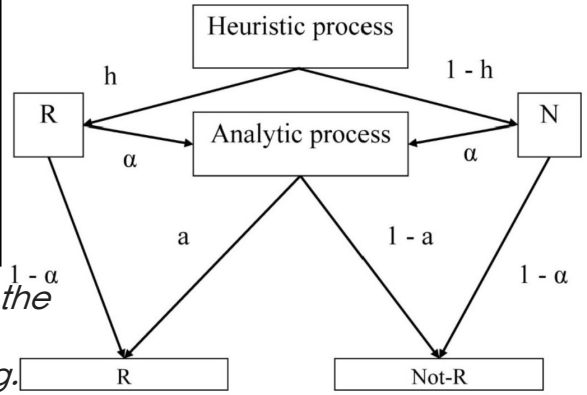
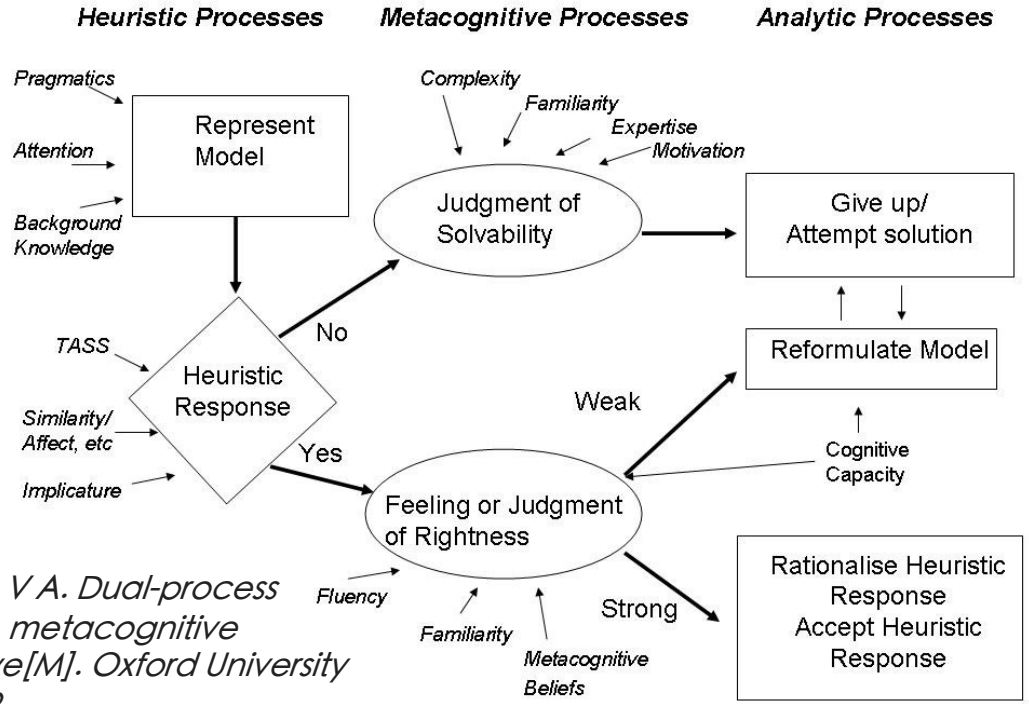


Figure 3. Default-interventionist model.

Evans, J. St. B. T. (2007a). On the resolution of conflict in dual-process theories of reasoning. Thinking & Reasoning



Thompson V A. Dual-process theories: A metacognitive perspective[M]. Oxford University Press, 2009.

元认知(metacognitive)架构

- S1经过元认知判断后决定是否触发S2
- 元认知基于S1和S2的触发经验训练

系统1和系统2的协同 – 转换机制

在如自动驾驶、辅助决策、智能客服、信息检索等等真实场景下，往往“**既要又要还要**”，需要两个系统深度协同工作

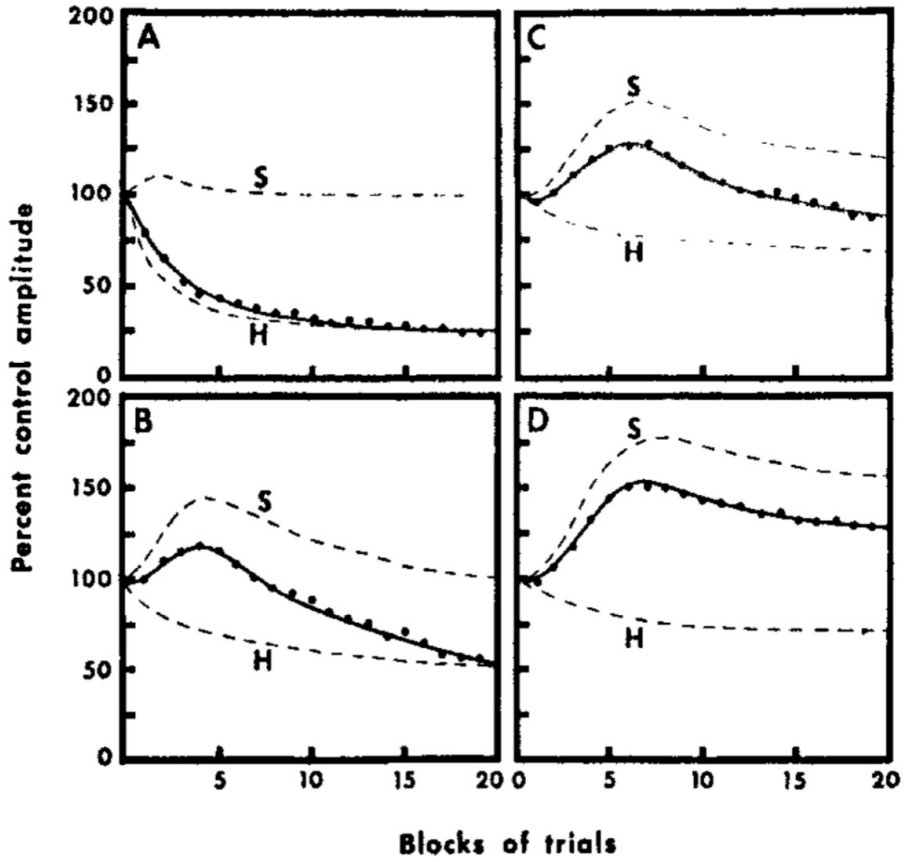


FIG. 1. Dual-process theory of habituation. (Dashed lines represent two hypothetical processes which result in plasticity of response to repeated stimulation. Dots are actual data points obtained for hindlimb flexion reflex in acute spinal cat. In A, shocks to hindpaw were presented at low intensity (near response threshold) and four per second frequency. Note pronounced response habituation. In B and C, as intensity is increased, hypothetical sensitization process (S) increases, while habituation process (H) becomes less pronounced. In D, at high intensity, sensitization is most pronounced, resulting in response sensitization.)

图1. 基于习惯的双过程理论

虚线表示两个假设过程，表明（神经回路）导致**对重复刺激的反应具有可塑性**。

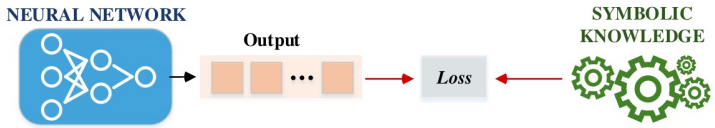
每个点是猫后肢屈曲反射中急性脊髓的实际数据点

- 在A组中，后爪受到的冲击强度较低（接近反应阈值），频率为每秒四次，注意产生了明显的反应习惯；
- 在B和C中，随着强度的增加，习惯过程（H）变得不那么明显
- 在D中，在高强度下，致敏作用最为明显，导致反应感知

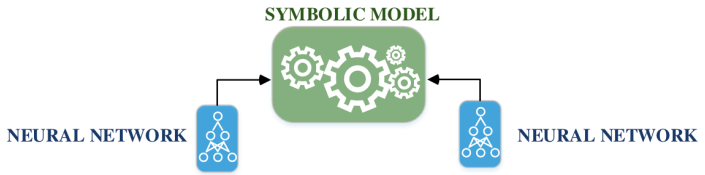
Thompson, R , and C. P. Groves . "Habituation: a dual-process theory. " Psychological Review 77.5(1970):419.

符号知识+神经网络的融合方法

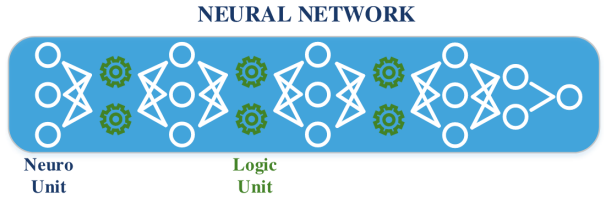
将符号知识/推理与神经网络融合，基本包括五种模式：**数据增强式、串行式、内嵌式、监督式、交互反应式**



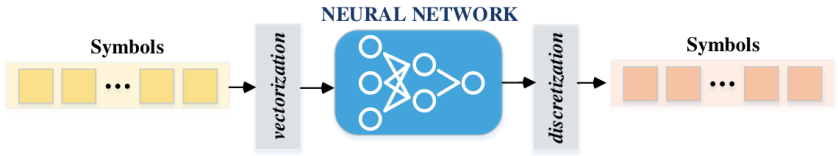
Neuro: Symbolic->Neuro



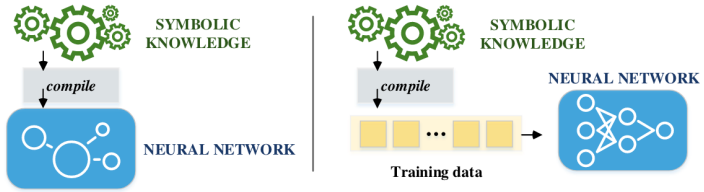
Symbolic[Neuro]



Neuro[Symbolic]



Symbolic Neuro Symbolic



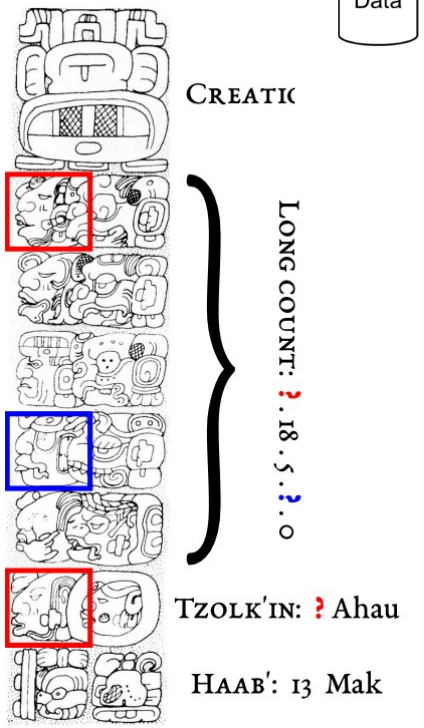
$\text{Neuro}_{\text{Symbolic}}$

Symbolic|Neuro

Wang, Wenguan, Yi Yang, and Fei Wu. "Towards data-and knowledge-driven artificial intelligence: A survey on neuro-symbolic computing." *arXiv preprint arXiv:2210.15889* (2022).

符号知识+神经网络：溯因学习

在学习过程中为现象找理由：希望神经网络能够输出和符号假设一致的结果，不一致时通过逻辑溯因进行纠正



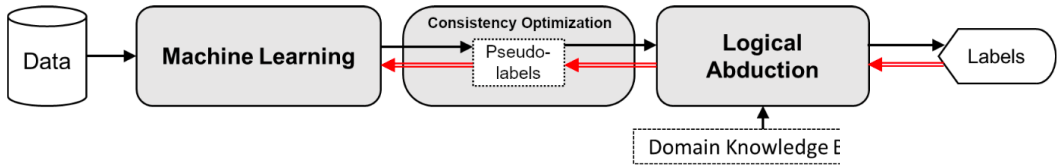
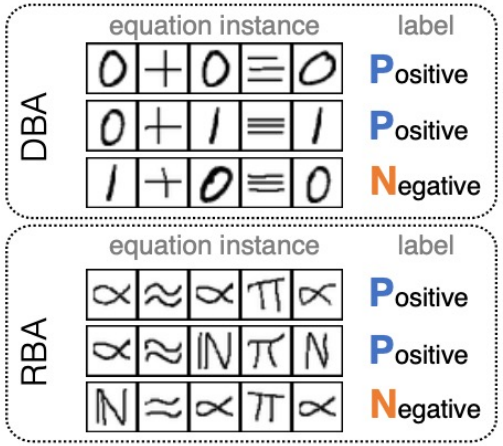


Figure 3: The structure of ABL framework.



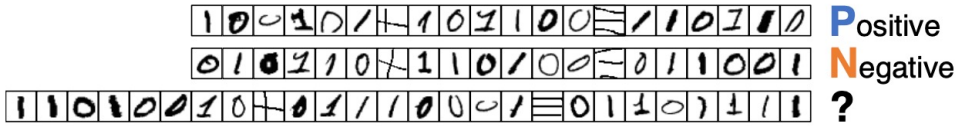


Figure 4: Handwritten equation decipherment puzzle: a computer should learn to recognise the symbols and figure out the unknown operation rules ("xnor" in this example) simultaneously.

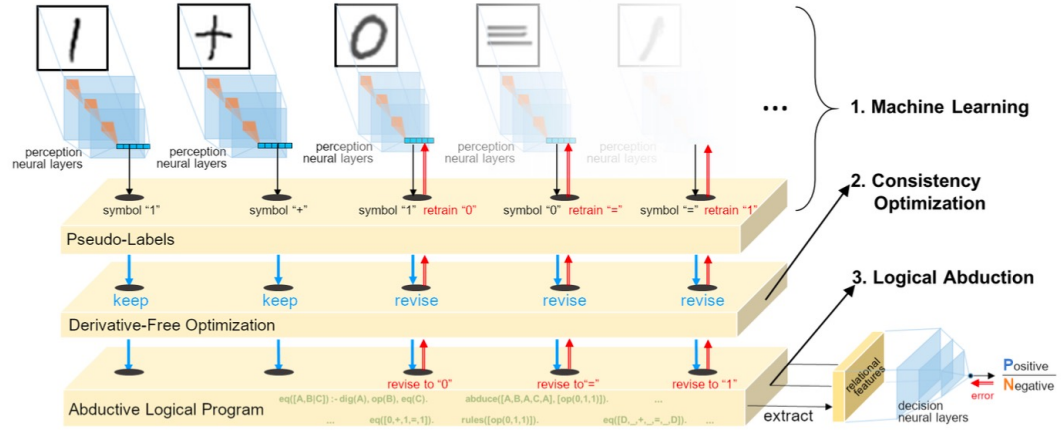


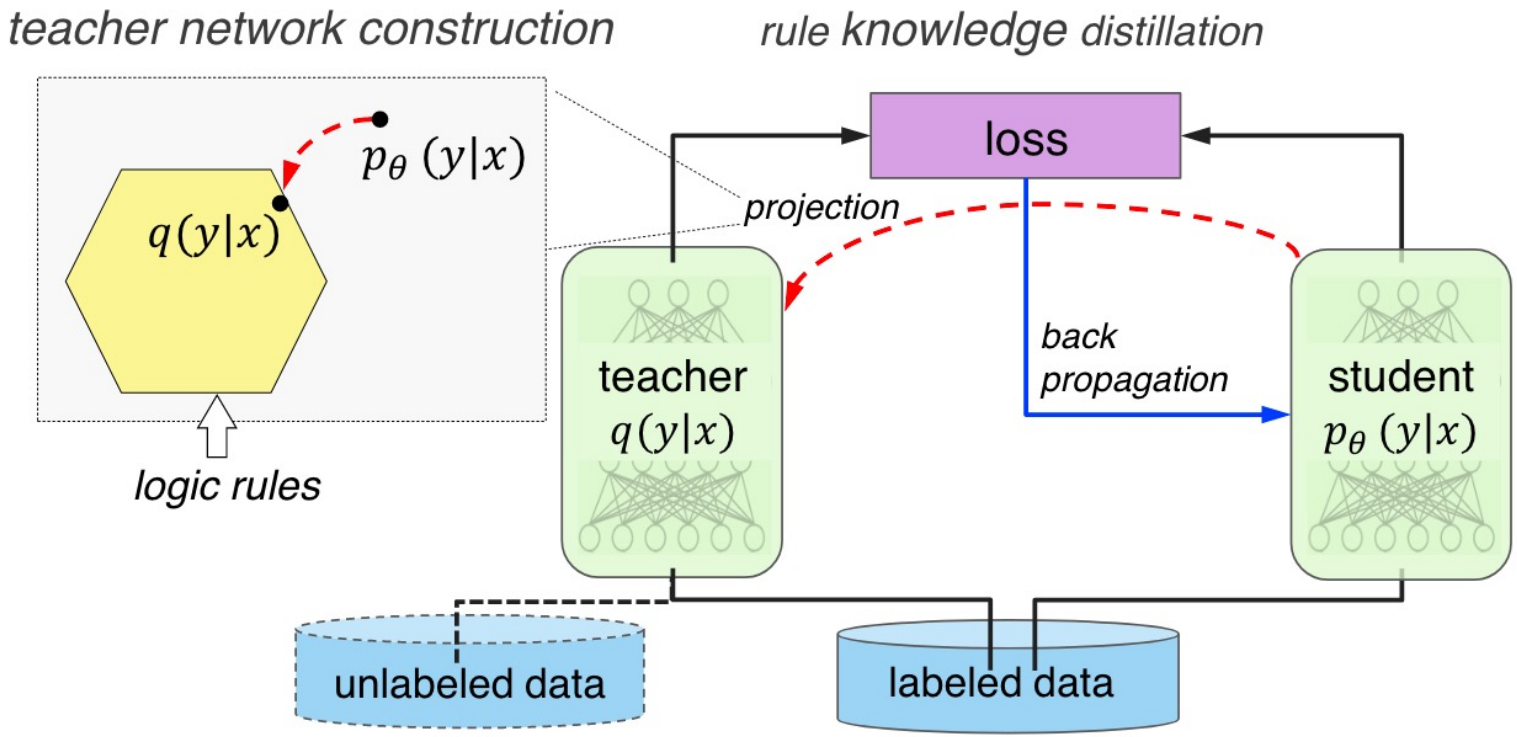
Figure 5: The structure of our ABL implementation.

Dai W Z, Xu Q, Yu Y, et al. Bridging machine learning and logical reasoning by abductive learning[J]. Advances in Neural Information Processing Systems, 2019, 32.

符号知识+神经网络：知识蒸馏

在学习过程中利用和泛化规则：让模型结果与规则结果相符，同时让规则逐渐泛化到更多的近似数据

$$A\&B = \max\{A + B - 1, 0\}$$
$$A \vee B = \min\{A + B, 1\}$$
$$A_1 \wedge \cdots \wedge A_N = \sum_i A_i / N$$
$$\neg A = 1 - A$$



Zhiting Hu, Xuezhe Ma, Zhengzhong Liu, Eduard Hovy, and Eric Xing. 2016. Harnessing Deep Neural Networks with Logic Rules. In Proceedings of the 54th ACL

符号知识+神经网络：符号可微

符号知识也是网络：通过将符号知识可微化实现其网络表示和表示学习，解决深度学习与离散逻辑结构不兼容的问题

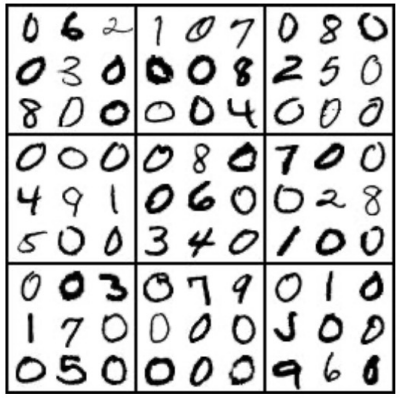


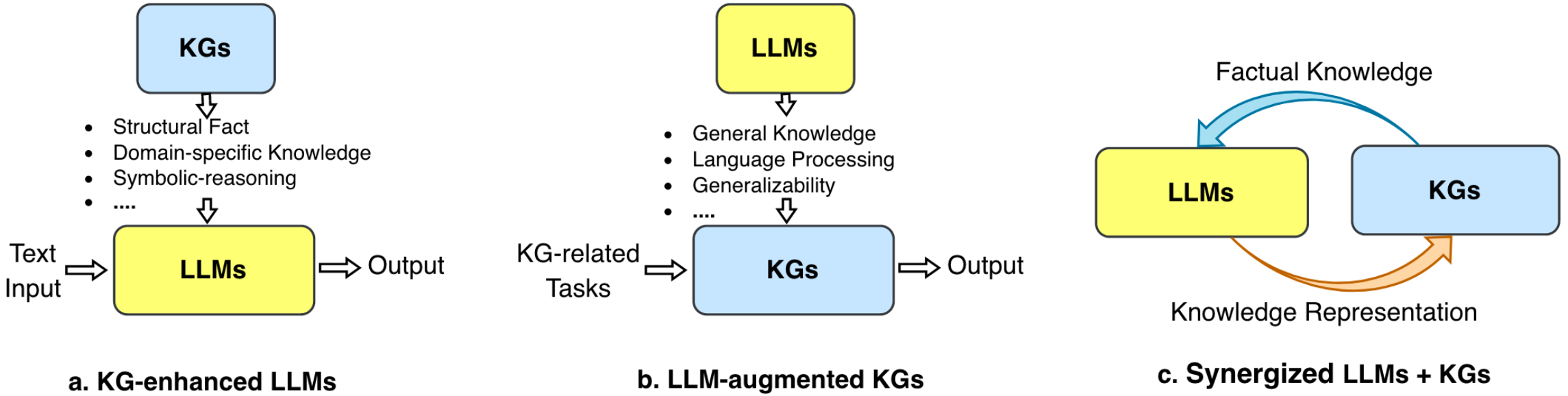
Figure 3. An example visual Sudoku image input, i.e. an image of a Sudoku board constructed with MNIST digits. Cells filled with the numbers 1-9 are fixed, and zeros represent unknowns.

在数独问题上不仅能够解决，并且能够学习数独的约束条件和解决方案

Po-Wei Wang, Priya Donti, Bryan Wilder, Zico Kolter. SATNet: Bridging deep learning and logical reasoning using a differentiable satisfiability solver. In PMLR 97:6545-6554, 2019.

符号知识+LLM的融合方法

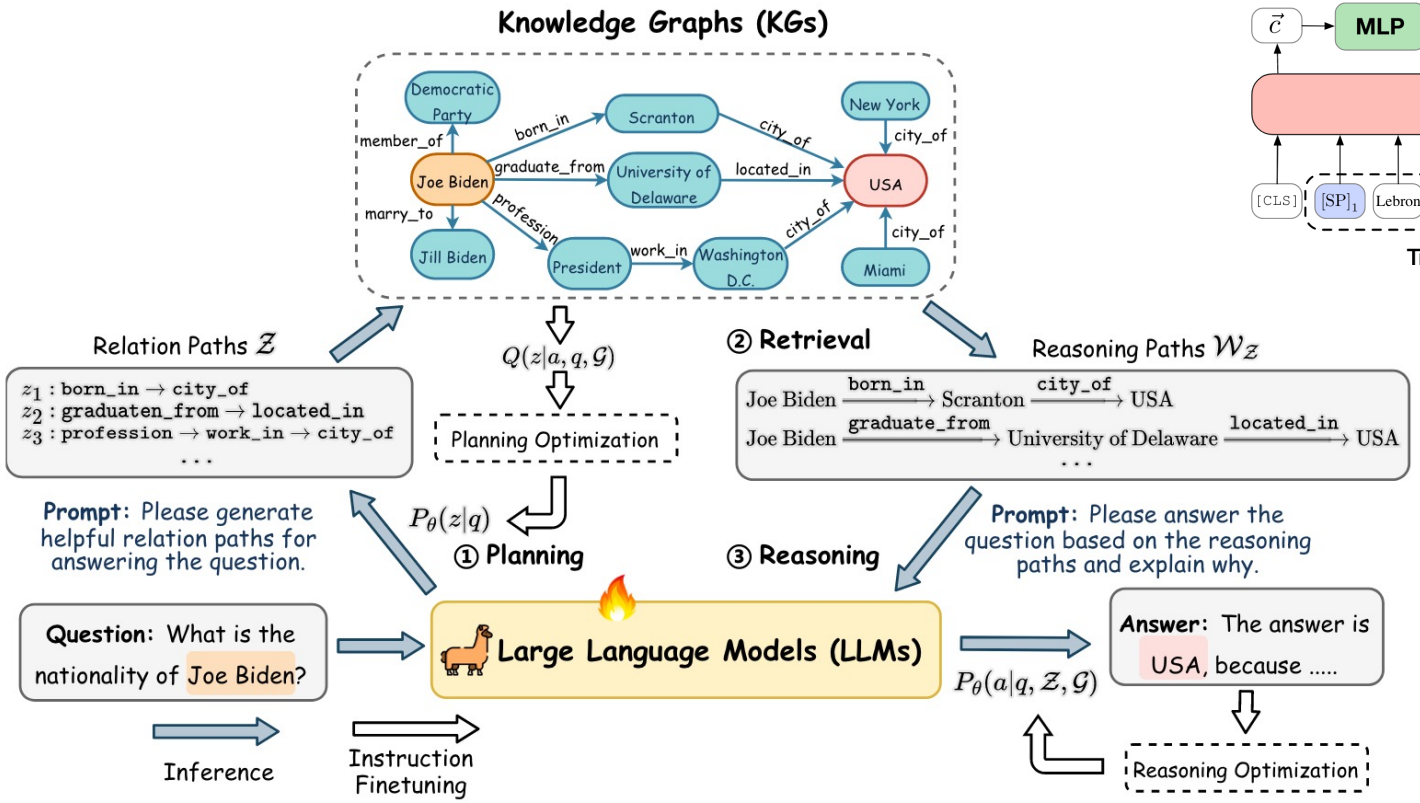
以知识图谱为例，可以大致分为**知识图谱增强大模型、大模型增强知识图谱以及两者的协同作用**

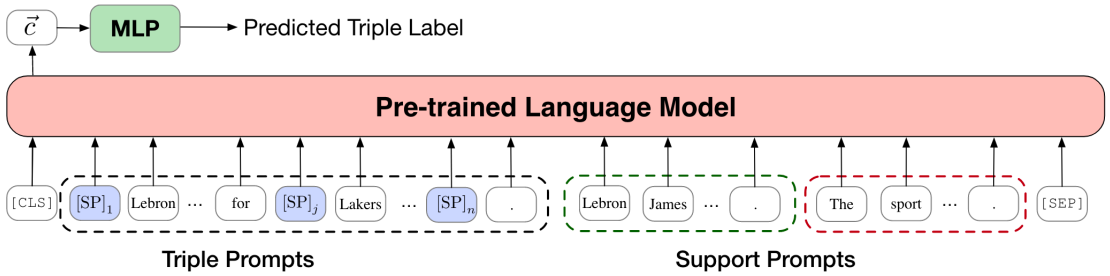


S. Pan, L. Luo, Y. Wang, C. Chen, J. Wang and X. Wu, "Unifying Large Language Models and Knowledge Graphs: A Roadmap," in IEEE Transactions on Knowledge and Data Engineering, vol. 36, no. 7, pp. 3580-3599, July 2024, doi: 10.1109/TKDE.2024.3352100.

符号知识+LLM：知识增强推理

知识图谱增强大模型推理能力：在推理过程中，利用已有知识寻找推理路径，约束大模型推理范围

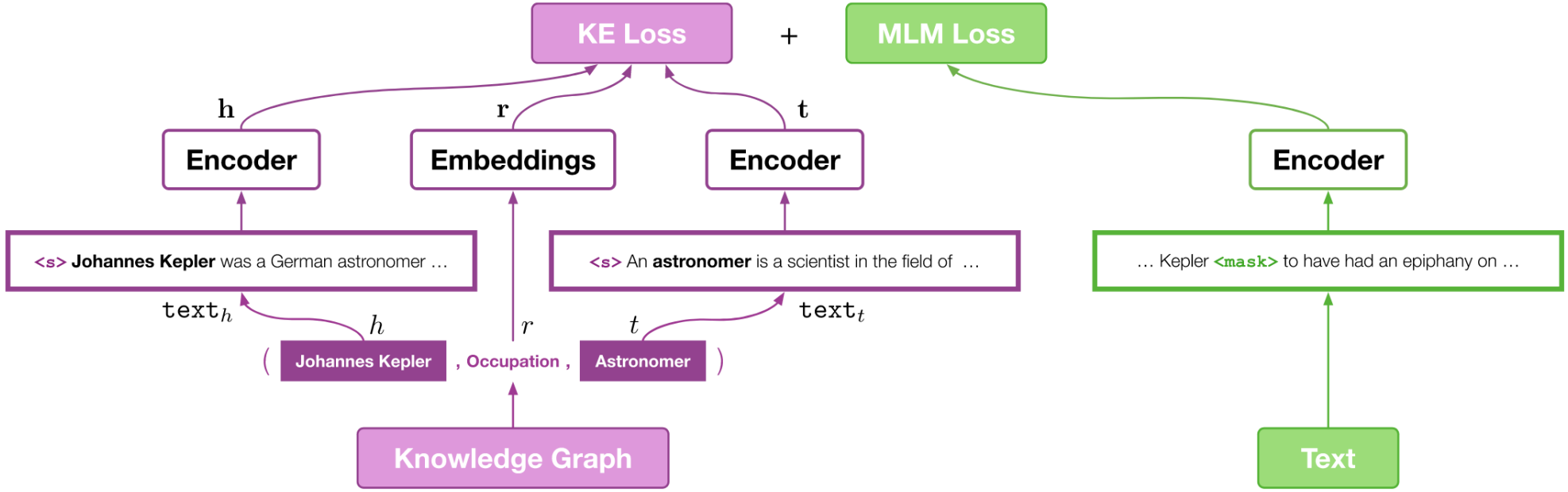




Luo, Linhao, et al. "Reasoning on graphs: Faithful and interpretable large language model reasoning." arXiv preprint arXiv:2310.01061 (2023).

符号知识+LLM：联合训练

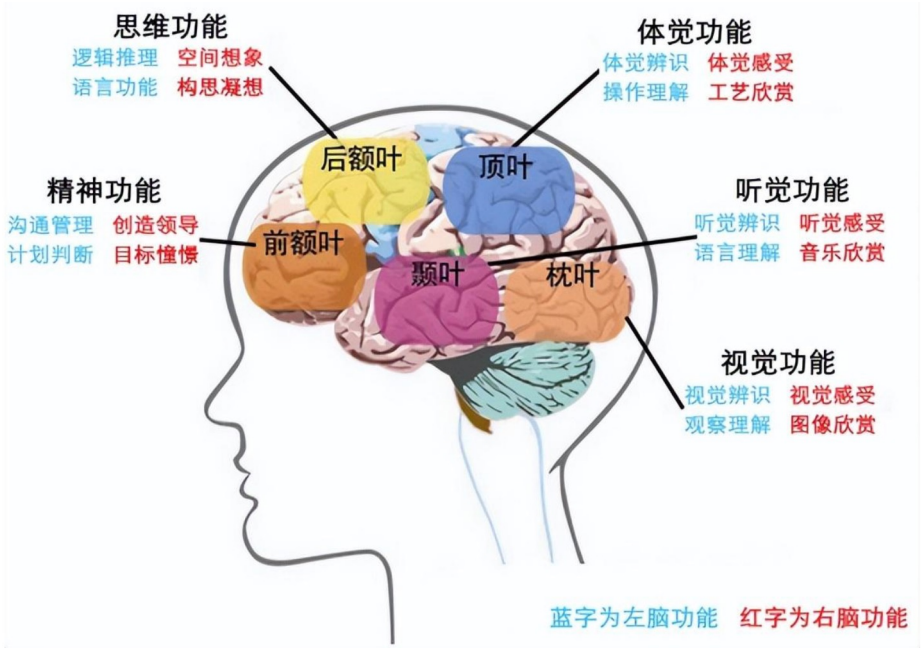
融合知识图谱嵌入表示学习和大模型训练：使用大模型生成知识图谱嵌入，将实体描述输入大模型后得到嵌入，再使用TransE进行损失函数的计算



Xiaozhi Wang, Tianyu Gao, Zhaocheng Zhu, Zhengyan Zhang, Zhiyuan Liu, Juanzi Li, and Jian Tang. 2021. KEPLER: A Unified Model for Knowledge Embedding and Pre-trained Language Representation. Transactions of the Association for Computational Linguistics, 9:176–194.

神经符号计算的下一步思考

探索双过程理论的**类脑模拟方法，满足**特定场景**既要又要应用需求**



目 录

1 双系统理论

2 世界模型

3

4

传统机器学习方法回顾

- 监督学习：通过有标注样本学习分类、回归等任务
  - 需要大量的高质量标注数据进行训练
  - 标注成本高

- 强化学习：智能体通过动作与环境进行交互并获得奖励信号，学习获得奖励最大的最优策略
  - 探索效率低，需要与仿真环境或真实环境进行大量交互
  - 搭建强化学习训练环境费时费力

大模型时代的人工智能

以自监督学习为范式的大模型通过在大规模数据上的训练取得了良好的效果，能够通过自然语言方式与人进行交互完成任务。

突破性能力：

- GPT-3为代表的少样本学习能力
- FLAN等零样本指令微调技术
- InstructGPT人类偏好后训练
- o1、DeepSeek R1展现出的推理能力

对话生成、资料搜集、代码编写、Agent个人助理……大模型已深入每个人生活的方方面面。

但……这就是AGI的答案了吗?

大模型存在的问题

当前的大语言模型是否具有智能?

- 幻觉问题：输出不符合事实的回答
- 时效性问题：模型知识停留在训练数据的截止时间，难以实时获取信息
- 记忆问题：缺乏跨会话持久记忆，多轮对话中容易出现遗忘
- 世界认知受限：依赖文本、图像等常见模态，缺乏对现实世界的感知能力，可能违背基本物理规律

人类智能VS人工智能

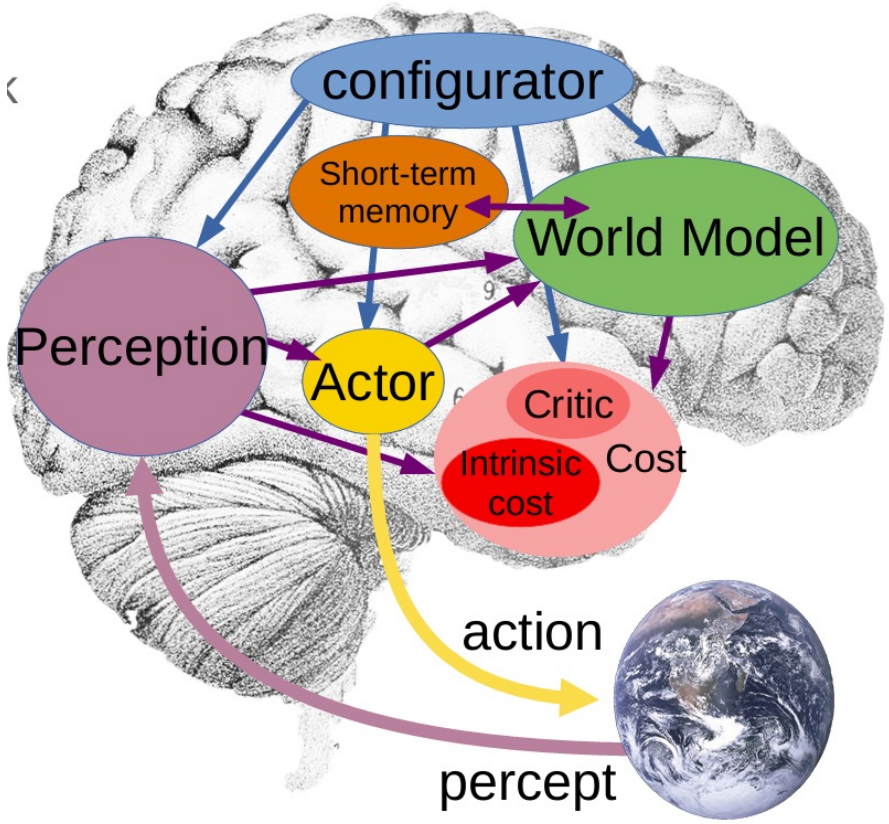
人类智能	人工智能
• 成年人经过20小时的学习可学会驾驶 • 10岁儿童可一次学会清理餐桌、洗碗、晾衣服等基本家务活动 • 人类大脑日常活动耗能较低	• 自动驾驶技术停留在L2、L3级别 • 家务机器人遥遥无期，且不具备单样本学习能力 • 大模型需要耗费巨量算力资源

如何使得人工智能拥有人类级别的能力?

LeCun的世界模型

Yann LeCun等人认为，高级机器智能应具备：

- 从感觉输入学习预测世界下一状态的能力
- 持续、大尺度的记忆能力
- 根据当前状态规划动作完成目标的能力
- 可控性和安全性



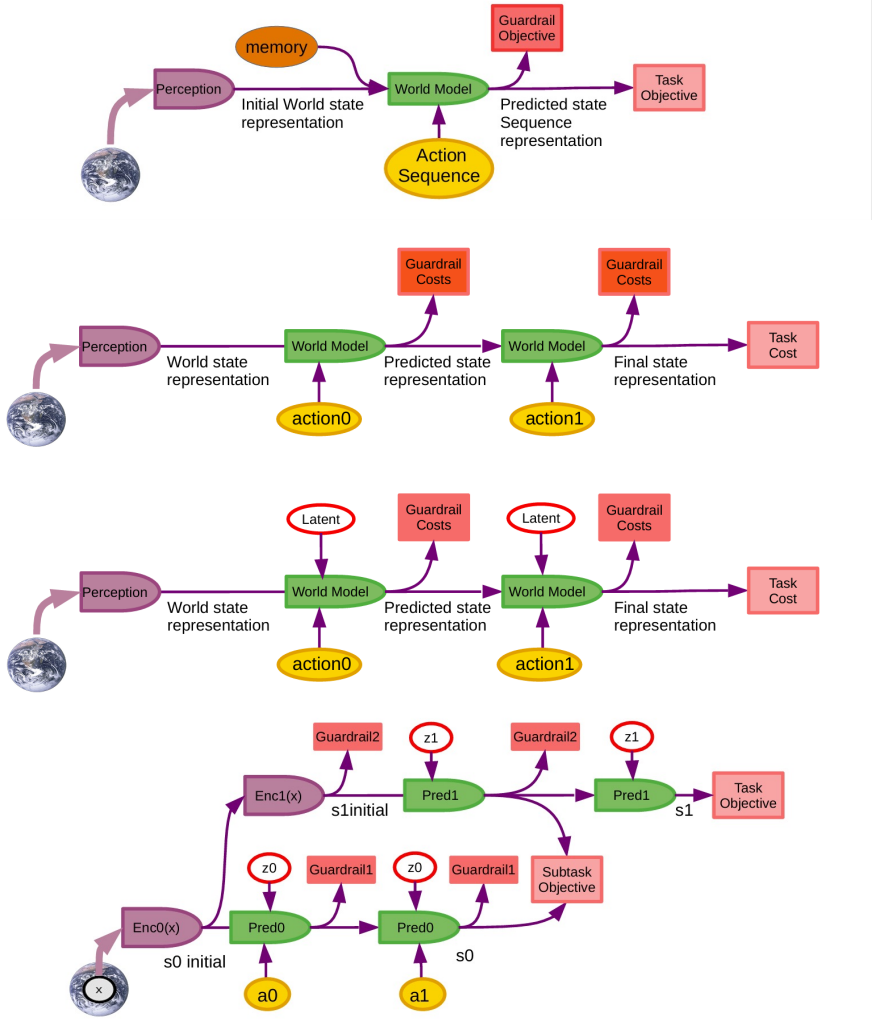
为此提出模块化认知结构，包括配置器、感知、世界模型、开销、动作器、短期记忆等模块。

其中世界模型负责根据当前状态和执行的动作预测未来的世界状态。

几种变体

Yann LeCun同时给出了几种世界模型变体：

- 多步/循环模型：类似于RNN，将统一模型用于多个时间步
- 非确定性模型：世界状态不确定或无法直接获取，因此使用隐变量表示预测结果
- 层次规划模型：用不同级别的模型预测不同级别的状态，越高阶的模型在更抽象的状态表示下预测长程状态



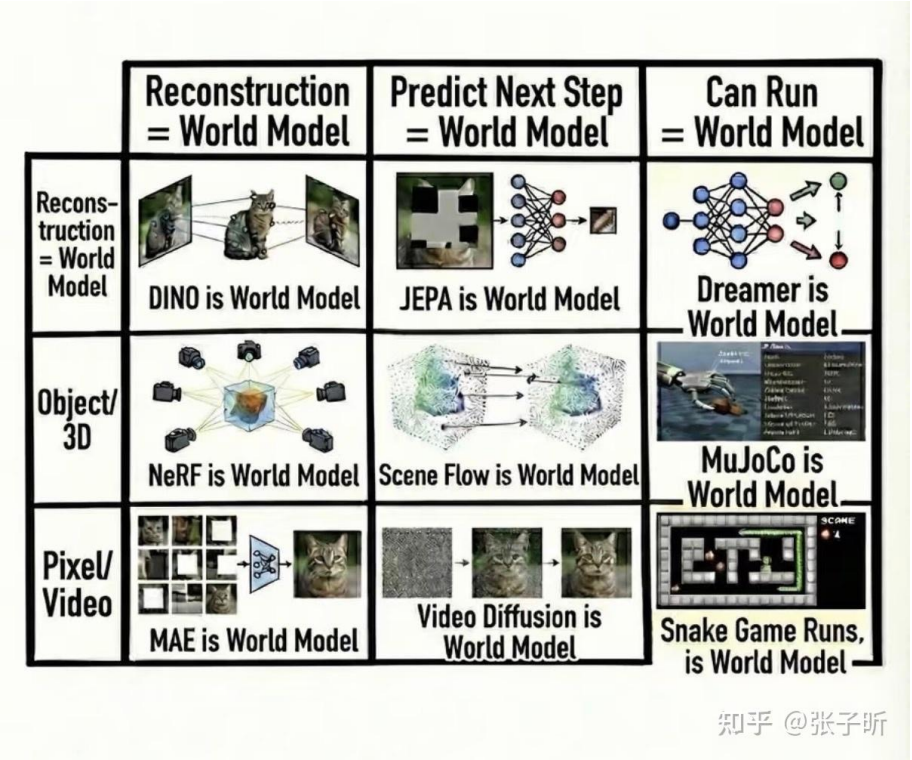
一句话总结：世界模型是根据当前状态和执行的动作预测未来的世界状态模型。

“世界模型”究竟是什么？

从学术论文到技术报告，“世界模型”无处不在

视频生成模型、物理模拟器、状态编码器、3D环境重建工作……

当我们在谈论“世界模型”的时候，究竟指的是什么？



世界模型的一种分类法

李飞飞及其领导的World Labs团队将各种被冠以“世界模型”之名的各种技术分为三大类：

- 渲染器（Renderer）：生成符合人类相机视角、具备照片级真实感的视觉图像。
- 模拟器（Simulator）：输出自洽、符合物理规律的世界场景表征，并供人类与程序进行计算与交互
- 规划器（Planner）：给定起始状态、最终目标与环境约束，输出一套可行的连续动作路径，引导智能体从起点抵达目标

渲染器

渲染器

渲染器通过海量互联网图文、视频数据学习画面的统计视觉规律，生成观感合理的画面。

世界状态指的是像素画面，而动作可以空缺，也可以是相机位置等信息。

视频生成模型、3D重建模型等工作都可归到这一类。

视频生成模型

类似于大语言模型，在大规模视频数据集或者文本-视频对数据集上进行训练，学习视频的生成规律。

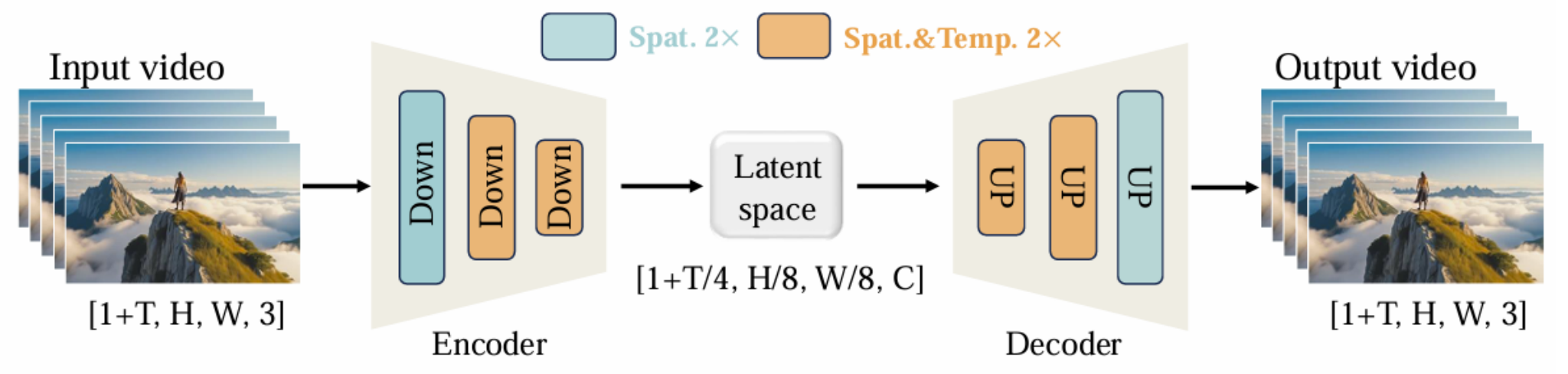
通常模型采用自回归模型或者扩散模型。

生成视频：一个人过生日，戴着生日帽，面对着生日蛋糕，吹灭蜡烛，围着他的家人们鼓掌庆祝。

视频生成模型架构简介

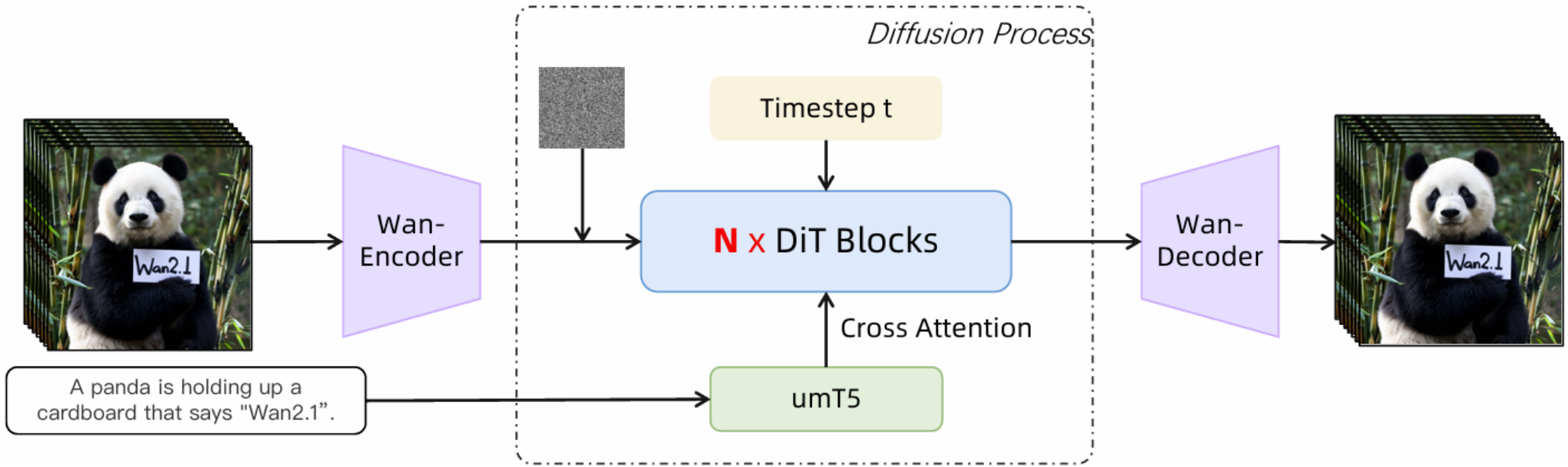
以WAN2.1为例：

编解码器架构使用VAE，将视频帧表示为标准正态分布的一个变量



视频生成模型架构简介

生成视频时将视频帧和文本指令进行编码，通过DiT（Diffusion Transformer）架构进行去噪预测，生成下一帧视频。



视频生成模型

视频生成模型当前存在的问题：

无法理解深层次的物理规律，容易生成违反常识（如物体穿透、重力失效等）的视频。

原因：模型在训练阶段学习到的是像素级别的统计规律，在生成阶段只是在模仿视频如何通过自回归或去噪生成一帧帧画面，并不理解真实世界的运行模式。


AI解压小屋
bilibili

3D重建模型

旨在从二维图像、视频或多视角观测数据中，逆向推导并恢复出场景的三维几何结构与外观属性。

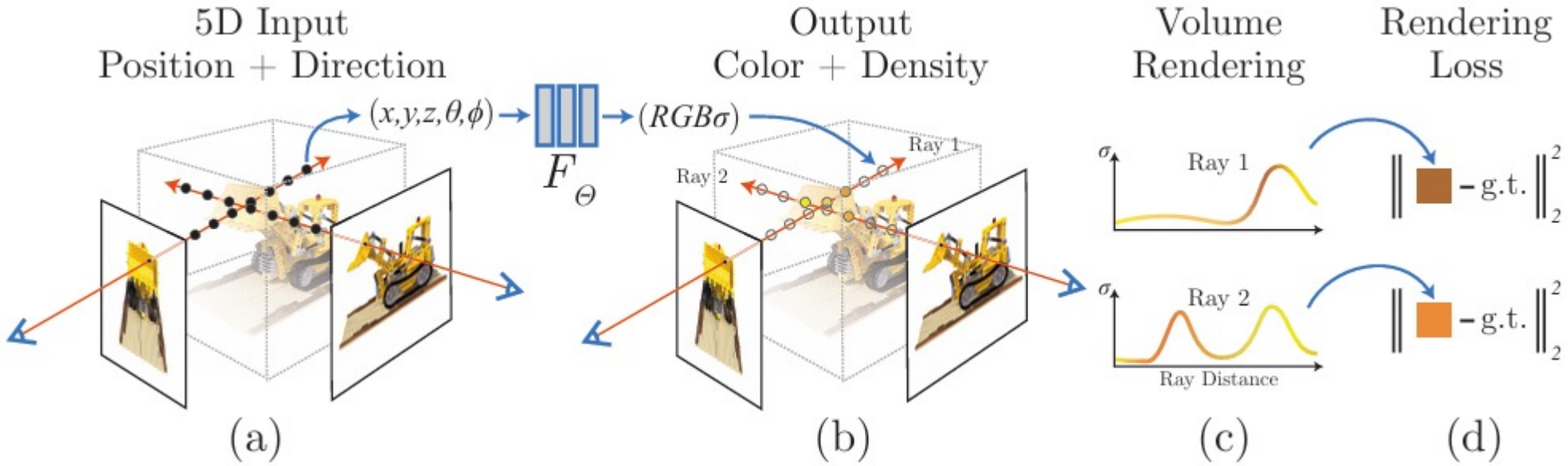
通常采用神经辐射场（NeRF）或3D高斯泼溅（3DGS）等显式或隐式的三维表征方法。

致力于建立空间上的结构一致性，从而支持新视角的实时渲染与交互。

NeRF

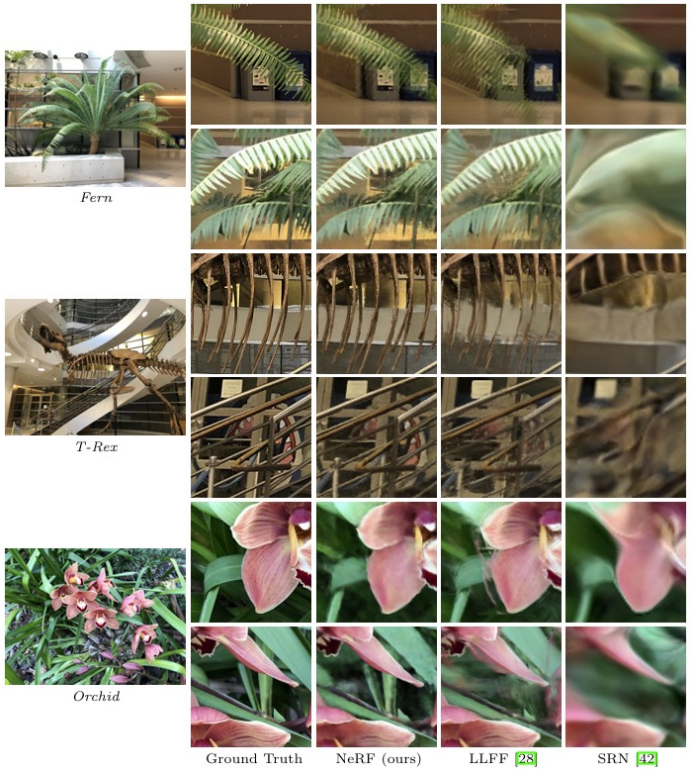
隐式的3D场景表示方式

将三维场景表示为连续函数，输入3D空间和方向向量，输出颜色和密度值



NeRF

结果：清晰度和真实感大大超过当时（2020）的一众渲染方法



3D高斯泼溅

显式的3D场景表示方式

将场景表示为3D高斯参数：

$$\Theta = \{(\mu_i, \mathrm{r}_i, \mathrm{s}_i, \sigma_i, c_i,)\}_{1 \leqslant i \leqslant n}$$

五个参数分别代表位置、旋转、尺寸、透明度、颜色，这样就将场景表示为了一个个3D椭球体的集合。

渲染阶段将3D椭球体投影到2D图像空间中进行图形学渲染，得到重建的图像。

3D高斯泼溅

结果：能够建模场景信息，并将高斯参数渲染成任意相机视角的图像

$$\Theta = \{(\mu_i, \mathrm{r}_i, \mathrm{s}_i, \sigma_i, c_i,)\}_{1 \leqslant i \leqslant n}$$

3D高斯泼溅

对比其他3D重建方法：

与Mip-NeRF360对比

与Plenoxels对比

交互式渲染

Genie3：交互式视频生成

- 采用自回归架构
- 实现720p的实时交互式视频生成
- 能够通过键盘、手柄等对场景视角进行控制
- 保持了场景一致性

Prompt: POV action camera of a tan house being painted by a first person agent with a paint roller

Prompt: This is a fantastical, whimsical forest environment. The lighting is bright and cheerful, suggesting a sunny day with dappled light filtering through a dense canopy of lush, oversized leaves. The air is clear and still. The ground is a soft, verdant carpet of moss and unusually large, brightly coloured mushrooms in shades of red and blue, their caps dotted with white. Winding dirt paths, well-trodden and narrow, weave between towering, ancient trees with smooth, grey bark. Interspersed throughout the forest are charming, mushroom-shaped houses, with intricate wooden doors and tiny, circular windows, each one unique in its design and colour palette, ranging from vibrant reds to gentle blues and greens. Various small, friendly forest creatures, such as colourful butterflies and tiny singing birds, flit amongst the foliage, adding to the lively atmosphere. There is an abundance of peculiar, oversized flowers blooming in an array of pastel and bright hues, releasing a gentle glow.

模拟器

模拟器

模拟器根据当前世界状态和动作预测下一步世界状态，可以说是最原教旨的世界模型。

模拟器的世界状态一般不直接预测像素，而是抽象表征。

JEPA系列、Ge Sim2等都归入此类。

需要说明的是，Isaac Sim、MUJOCO这类物理仿真器通过经典力学、机器人动力学等物理法则对环境的下一步状态进行预测，并不属于基于学习的预测方法，因此一般不归为世界模型。

JEPA系列

Yann LeCun在提出的世界模型架构基础上，进一步推出JEPA（Joint Embedding Predictive Architecture）。

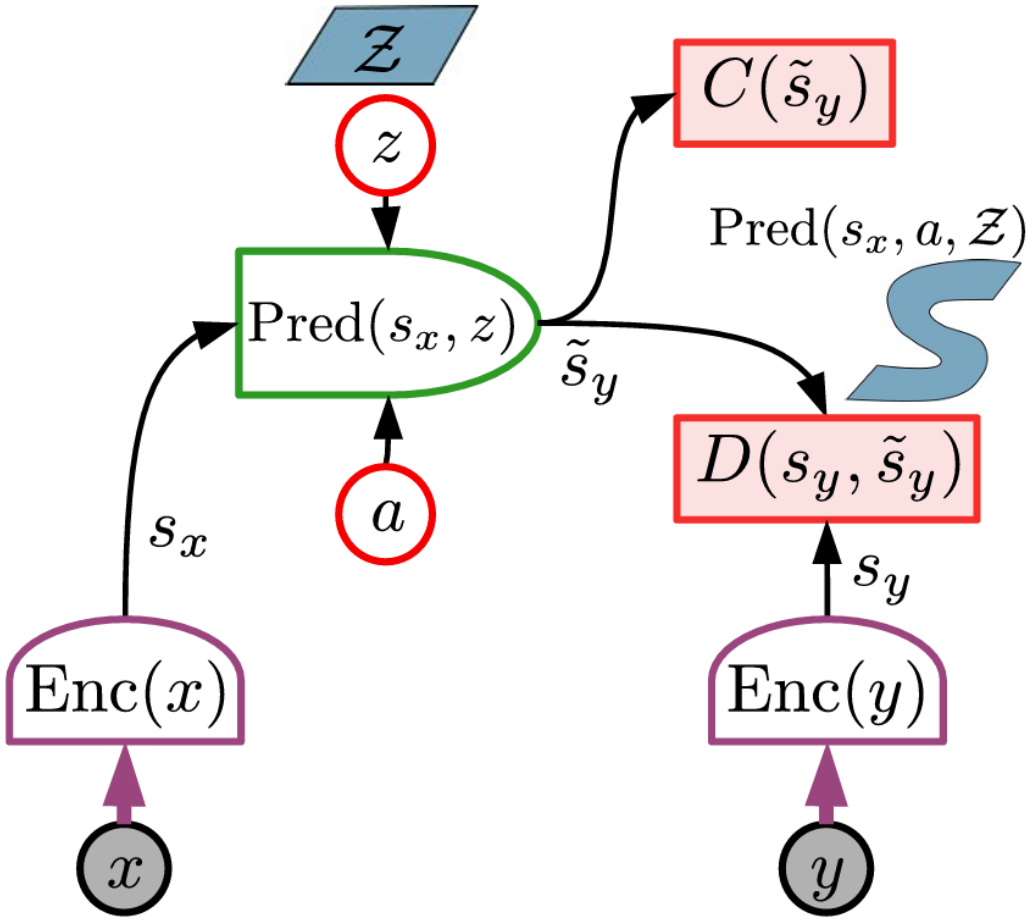
主要思路：放弃对像素或者token的直接重建，只预测像素或token的抽象表示。

主要工作：I-JEPA、V-JEPA、LLM-JEPA、Audio-JEPA、LeWorldModel等。

JEPA系列基本架构

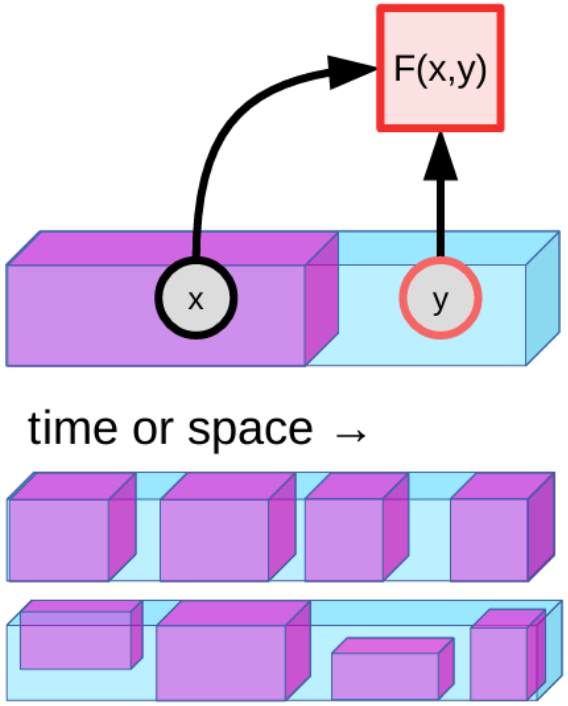
- x：历史和当前的信息
- y：未来信息
- a：动作
- z：隐变量
- D()：预测开销
- C()：代理开销

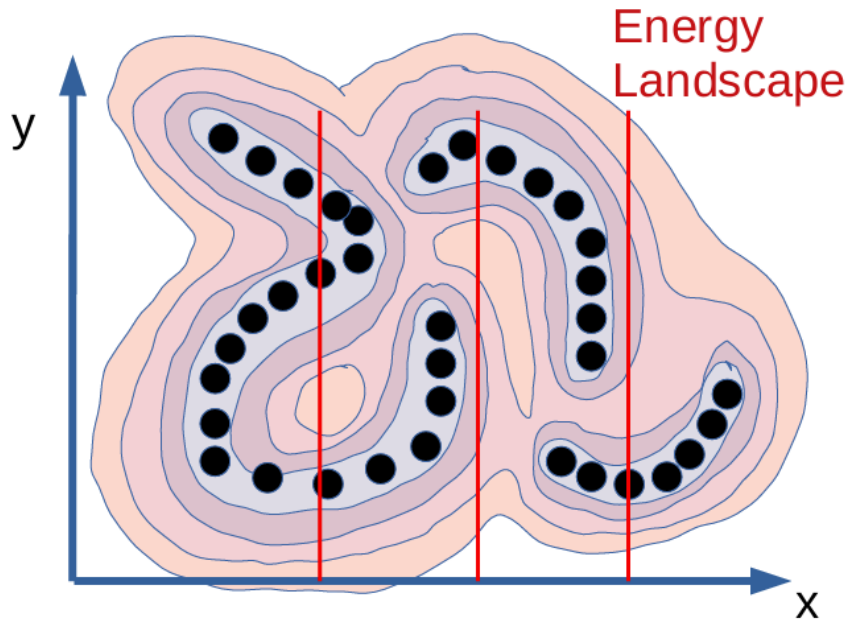
JEPA从历史和当前信息的表示预测未来的表示。



基于能量的模型

基于能量的模型：给予相容的数据对低能量，反之高能量



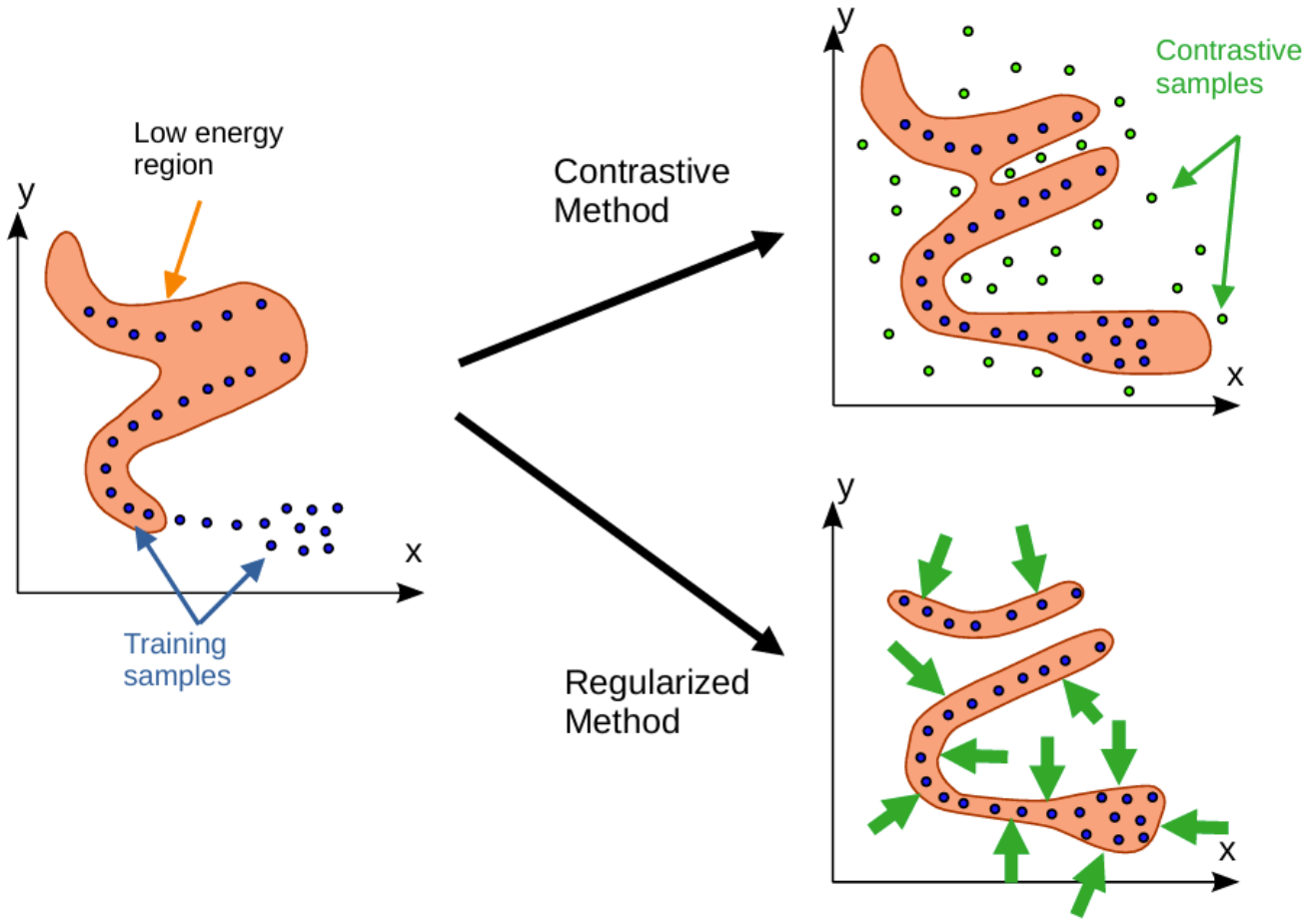


$$\check{y} = \text{argmin}_y F(x, y)$$

学习方式

- 对比学习方法
  - 降低训练样本能量
  - 提高最高的对比样本能量
  - 难以扩展到高维度

- **正则化方法**
  - 通过正则化器优化低能量区域体积

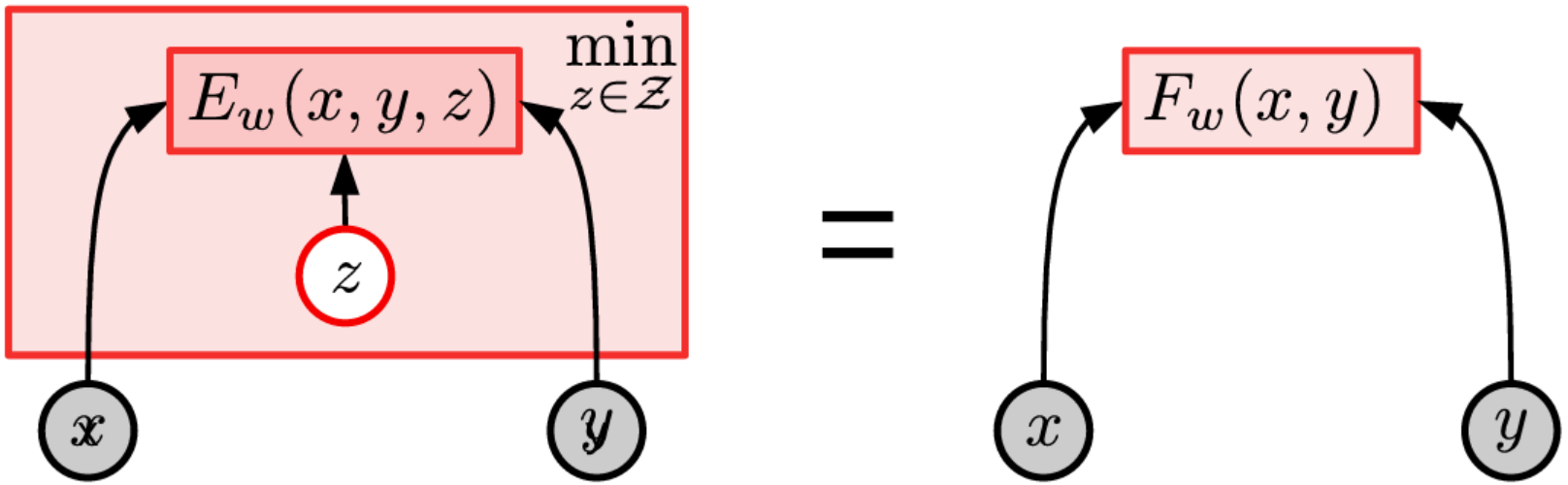


有隐变量的能量模型

使用隐变量捕获未来状态有但是当前状态没有的信息

通过以下进行优化：

$$\check{z} = \text{argmin}_{z \in \mathcal{Z}} E_w(x, y, z) \qquad F_w(x, y) = E_w(x, y, \check{z})$$



能量模型VS概率模型

概率模型是能量模型的特殊形式。

能量模型在选择得分函数和目标函数上更加自由。

两者可以通过Gibbs-Boltzmann分布进行转换。

$$P(y|x) = \frac{e^{-\beta F(x,y)}}{\int_{y'} e^{-\beta F(x,y')}}$$

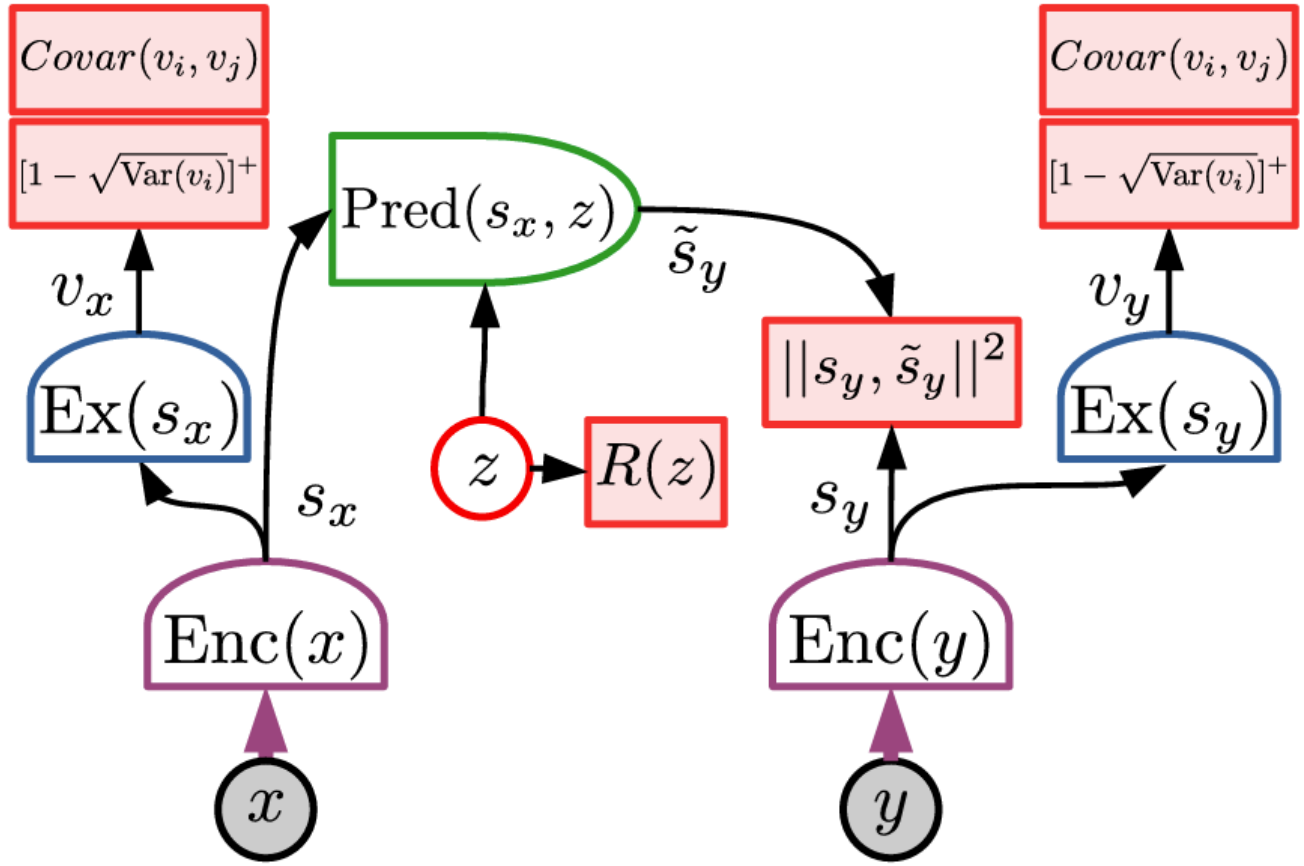
VICReg

正则化方法在图像表征中的尝试

VIC：Variance、Invariance、Covariance

通过三项进行正则化：

- 方差项：保持每个表示项的变化程度
- 协方差项：解耦每个表示项的相关性
- 不变项：最小化预测误差



VICReg

结果：在训练任务和迁移任务上表现出色

Method	Linear		Semi-supervised			
	Top-1	Top-5	Top-1		Top-5	
			1%	10%	1%	10%
Supervised	76.5	-	25.4	56.4	48.4	80.4
MoCo He et al. (2020)	60.6	-	-	-	-	-
PIRL Misra & Maaten (2020)	63.6	-	-	-	57.2	83.8
CPC v2 Hénaff et al. (2019)	63.8	-	-	-	-	-
CMC Tian et al. (2019)	66.2	-	-	-	-	-
SimCLR Chen et al. (2020a)	69.3	89.0	48.3	65.6	75.5	87.8
MoCo v2 Chen et al. (2020c)	71.1	-	-	-	-	-
SimSiam Chen & He (2020)	71.3	-	-	-	-	-
SwAV Caron et al. (2020)	71.8	-	-	-	-	-
InfoMin Aug Tian et al. (2020)	73.0	91.1	-	-	-	-
OBoW Gidaris et al. (2021)	73.8	-	-	-	82.9	90.7
BYOL Grill et al. (2020)	74.3	91.6	53.2	68.8	78.4	89.0
SwAV (w/ multi-crop) Caron et al. (2020)	75.3	-	53.9	70.2	78.5	89.9
Barlow Twins Zbontar et al. (2021)	73.2	91.0	55.0	69.7	79.2	89.3
VICReg (ours)	73.2	91.1	54.8	69.5	79.4	89.5

Method	Linear Classification			Object Detection		
	Places205	VOC07	iNat18	VOC07+12	COCO det	COCO seg
Supervised	53.2	87.5	46.7	81.3	39.0	35.4
MoCo He et al. (2020)	46.9	79.8	31.5	-	-	-
PIRL Misra & Maaten (2020)	49.8	81.1	34.1	-	-	-
SimCLR Chen et al. (2020a)	52.5	85.5	37.2	-	-	-
MoCo v2 Chen et al. (2020c)	51.8	86.4	38.6	82.5	39.8	36.1
SimSiam Chen & He (2020)	-	-	-	82.4	-	-
BYOL Grill et al. (2020)	54.0	86.6	47.6	-	40.4†	37.0†
SwAV (m-c) Caron et al. (2020)	56.7	88.9	48.6	82.6	41.6	37.8
OBoW Gidaris et al. (2021)	56.8	89.3	-	82.9	-	-
Barlow Twins Grill et al. (2020)	54.1	86.2	46.5	82.6	40.0†	36.7†
VICReg (ours)	54.3	86.6	47.0	82.4	39.4	36.4

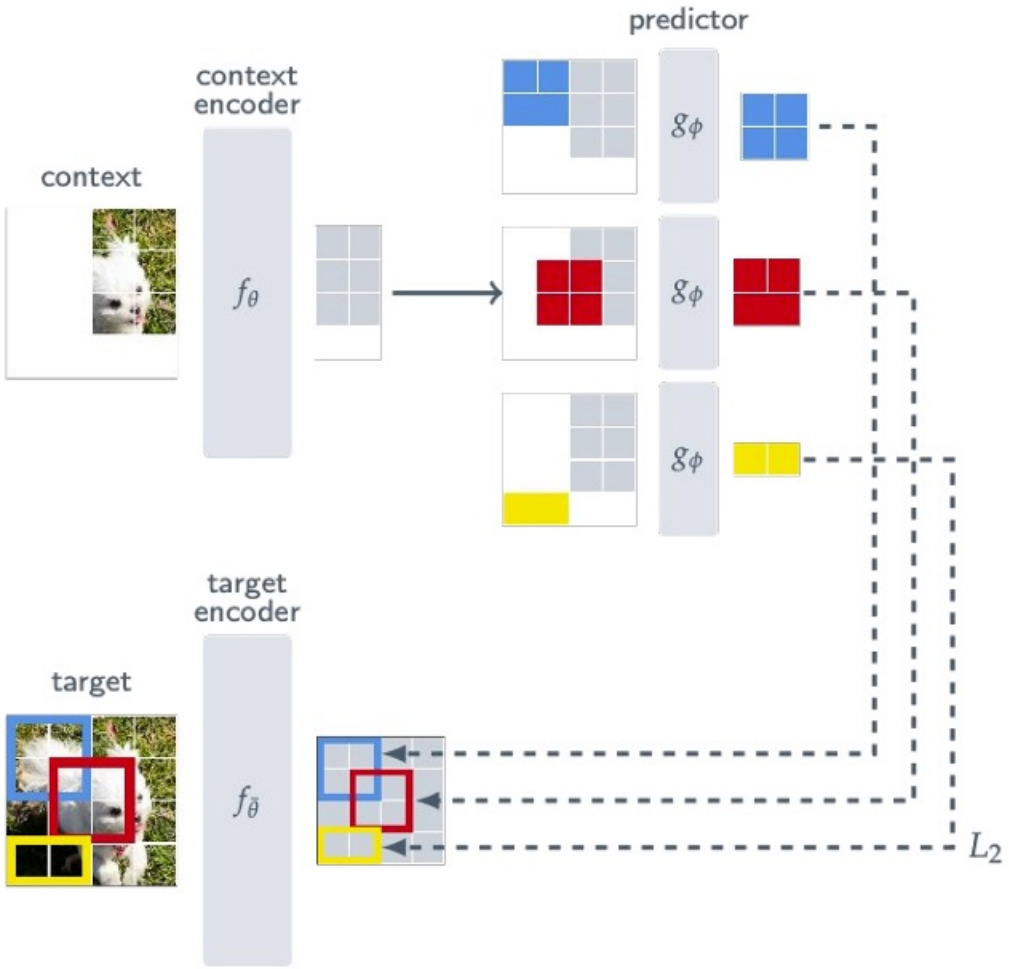
I-JEPA

JEPA在视觉表征领域的奠基之作

对上下文和邻近的Patch进行联合嵌入

使用掩码方法，从单张图像预测缺失表征

基于ViT-Huge/14训练

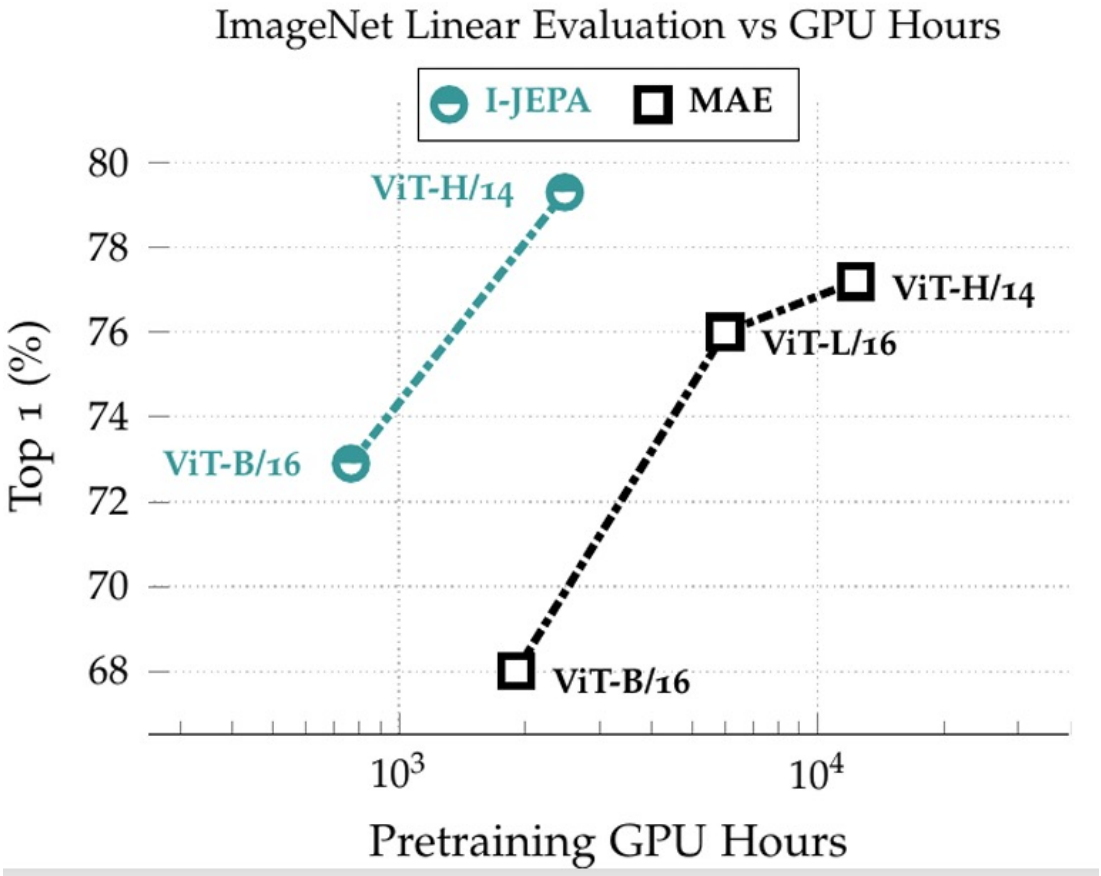


I-JEPA结果

- 训练速度快

- 作为非生成式方法在ImageNet任务上能够超过其他方法

Method	Arch.	Epochs	Top-1
Methods without view data augmentations			
data2vec [7]	ViT-L/16	1600	53.5
MAE [34]	ViT-B/16	1600	68.0
	ViT-L/16	1600	76.0
	ViT-H/14	1600	77.2
I-JEPA	ViT-B/16	600	72.9
	ViT-L/16	600	77.5
	ViT-H/14	300	79.3
	ViT-H/16_{448}	300	**81.1**
Methods using extra view data augmentations			
SimCLR v2 [20]	RN152 (2×)	800	79.1
DINO [17]	ViT-B/8	300	80.1
iBOT [74]	ViT-L/16	250	**81.0**

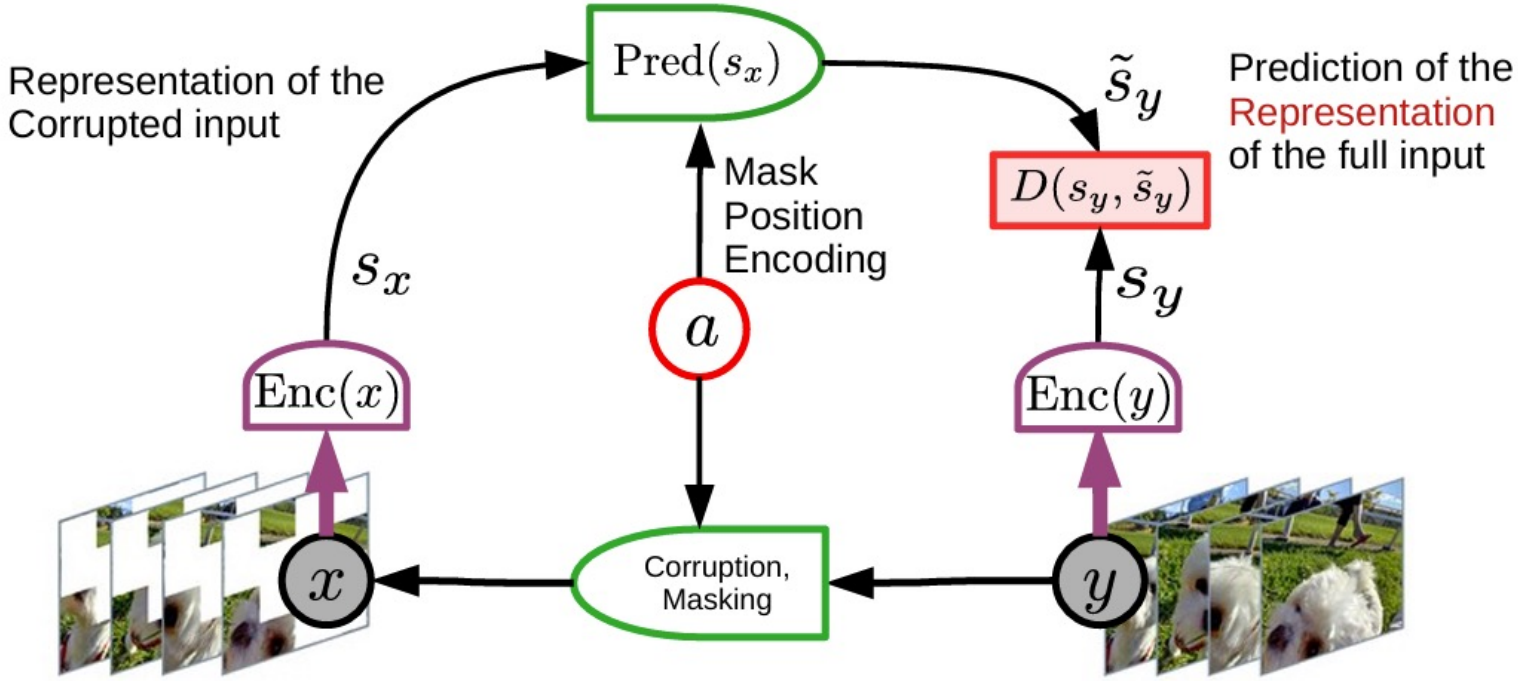


V-JEPA

实现纯视频自监督学习

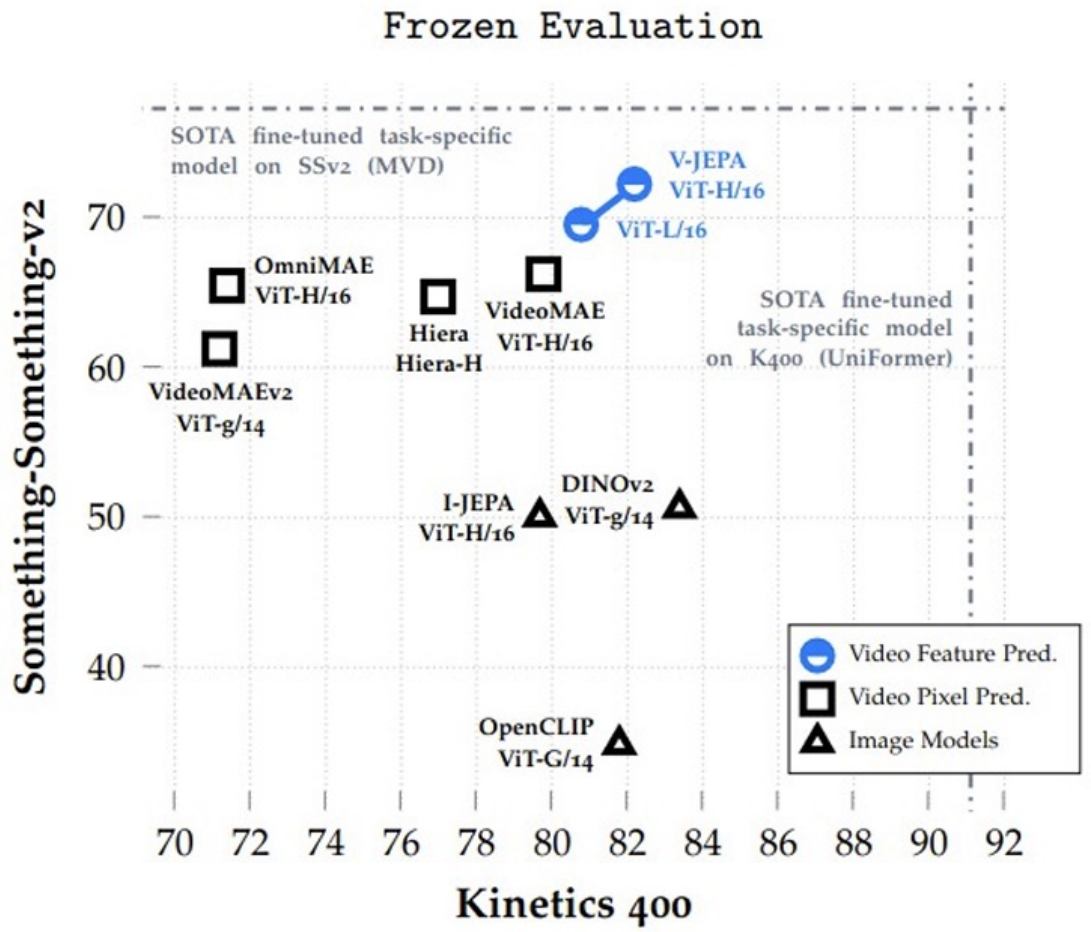
在200万段视频上训练

预测跨帧特征表征，并掌握“外观内容”与“运动动态”



V-JEPA结果

在动作识别领域和生成模型以及图像编码器进行比较：



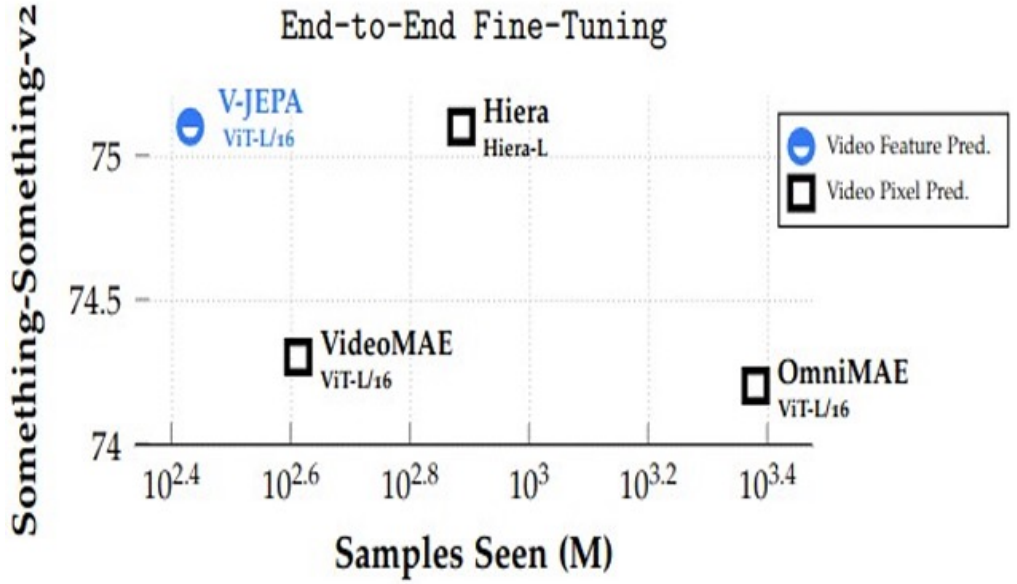
V-JEPA结果

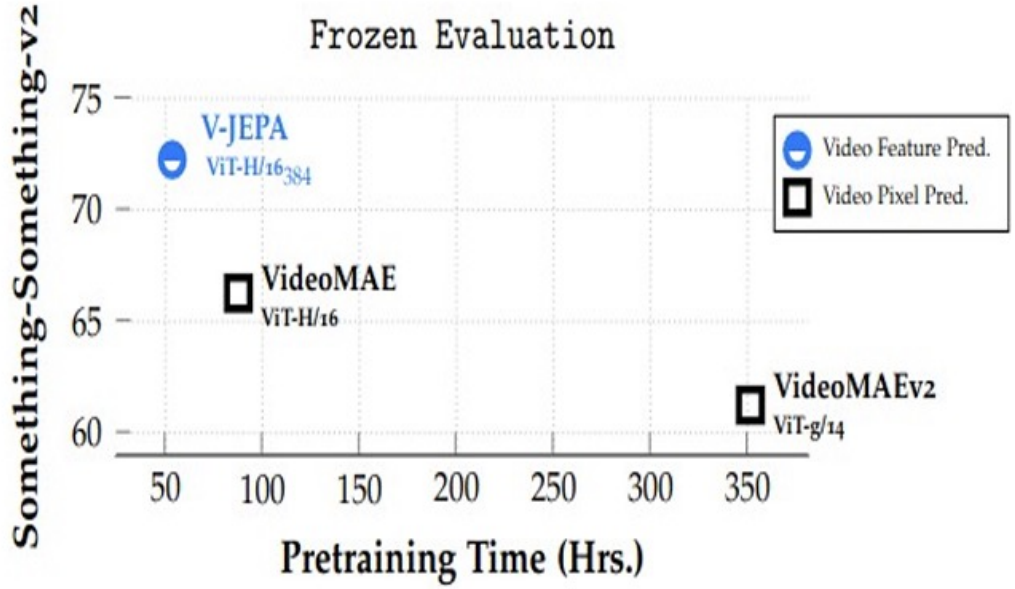
在少样本动作识别上的结果：

Method	Arch.	Frozen Evaluation					
		K400 (16×8×3)			SSv2 (16×2×3)		
		5%	10%	50%	5%	10%	50%
MVD	ViT-L/16	62.6 ± 0.2	68.3 ± 0.2	77.2 ± 0.3	42.9 ± 0.8	49.5 ± 0.6	61.0 ± 0.2
VideoMAE	ViT-H/16	62.3 ± 0.3	68.5 ± 0.2	78.2 ± 0.1	41.4 ± 0.8	48.1 ± 0.2	60.5 ± 0.4
VideoMAEv2	ViT-g/14	37.0 ± 0.3	48.8 ± 0.4	67.8 ± 0.1	28.0 ± 1.0	37.3 ± 0.3	54.0 ± 0.3
V-JEPA	ViT-H/16_{384}	68.2 ± 0.2	72.8 ± 0.2	80.6 ± 0.2	54.0 ± 0.2	59.3 ± 0.5	67.9 ± 0.2

V-JEPA结果

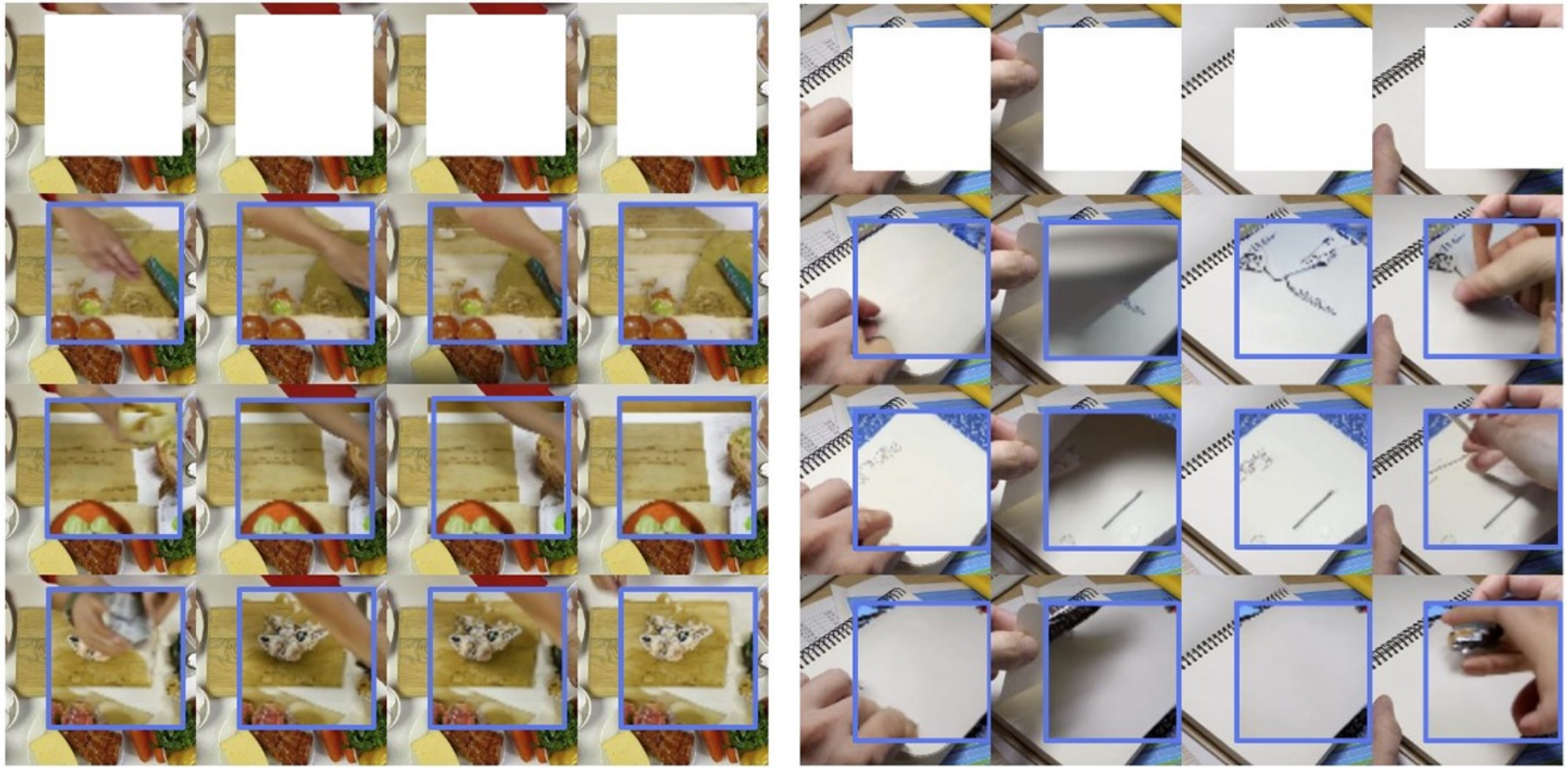
在采样效率和学习速度上与基于重建的方法对比：





V-JEPA结果

虽然设计初衷不为了重建，但学习到的表征依然可以用于训练外部的重建解码器。



JEPA系列的其他成员

- LLM-JEPA：将JEPA引入语言领域，通过预测文本片段的潜表征，打破LLM依赖Token生成的局限，在Llama3、Gemma2等模型上显著提升性能与抗过拟合能力
- Audio-JEPA：音频领域的JEPA
- V-JEPA 2、LeWorldModel：在学习视觉表征的同时能够将视觉表征用于动作预测

其他模拟器

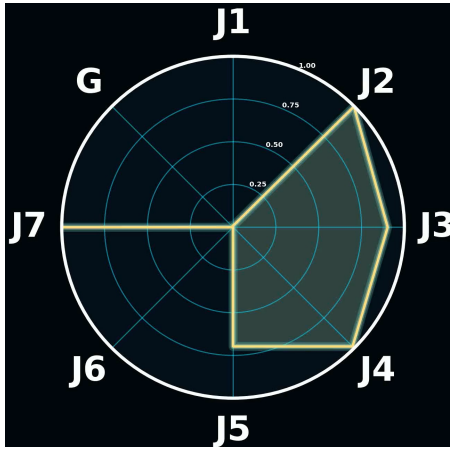
GE-Sim2.0：用于机器人操控的具身世界模拟器

在基于动作条件的视频模拟基础上，新增了三项功能，以实现可扩展的策略评估和学习：本体感觉状态估计、自动任务评估和高效试运行。

超越单纯的渲染器，拥有机器人状态估计以及任务评价能力。

能够用于机器人策略的训练和评估。

GE-Sim2.0



关节状态估计

视觉估计

规划器

规划器

有了能够预测世界状态的世界模型后，希望更进一步，通过世界模型学习到的表征指导动作的生成，影响世界状态。

根据历史世界状态和动作想象下一步的世界状态，并基于此生成下一步的动作并执行。

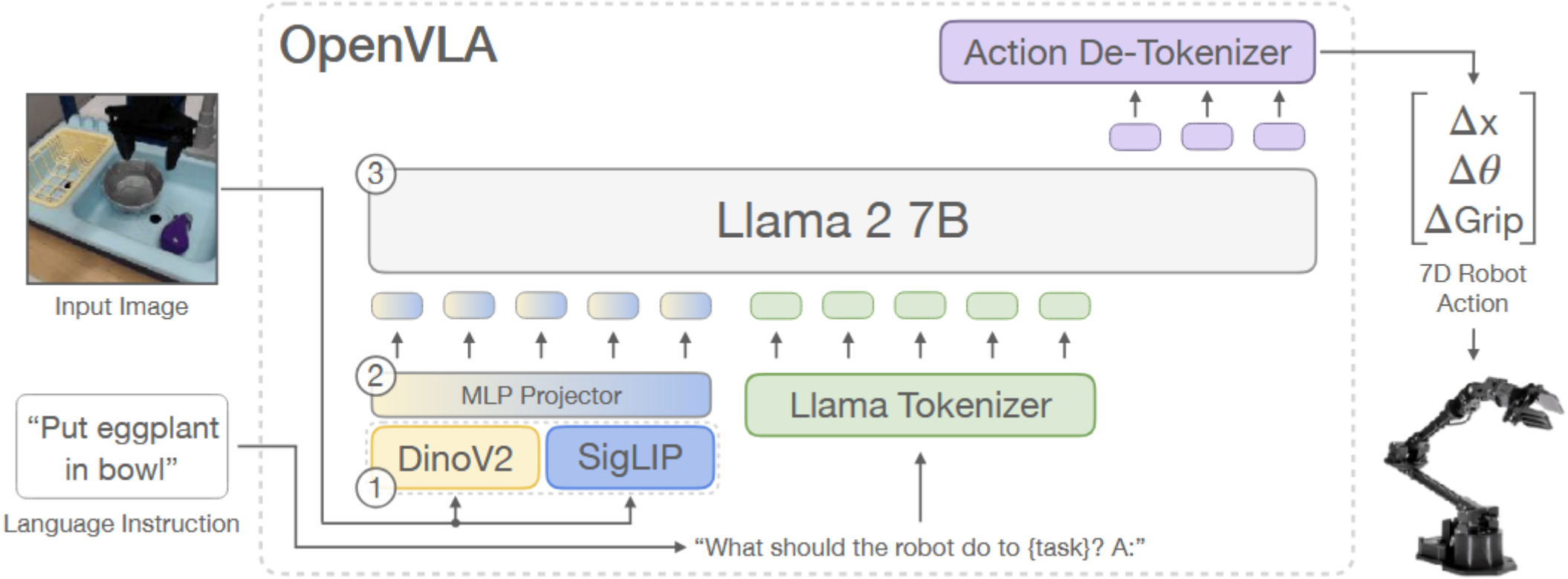
在当今机器人领域中，具身智能大模型受到广泛关注，其中一大类工作就是基于世界模型的世界动作模型（World Action Model，WAM）。

从VLA讲起

视觉-语言-动作模型（Vision-Language-Action Model，VLA）是另一大类具身智能大模型，基于VLM扩展而来。输入相机采集图像、机器人动作等观测，直接输出期望执行的动作（如机械臂各关节角度、夹爪的开合等）。

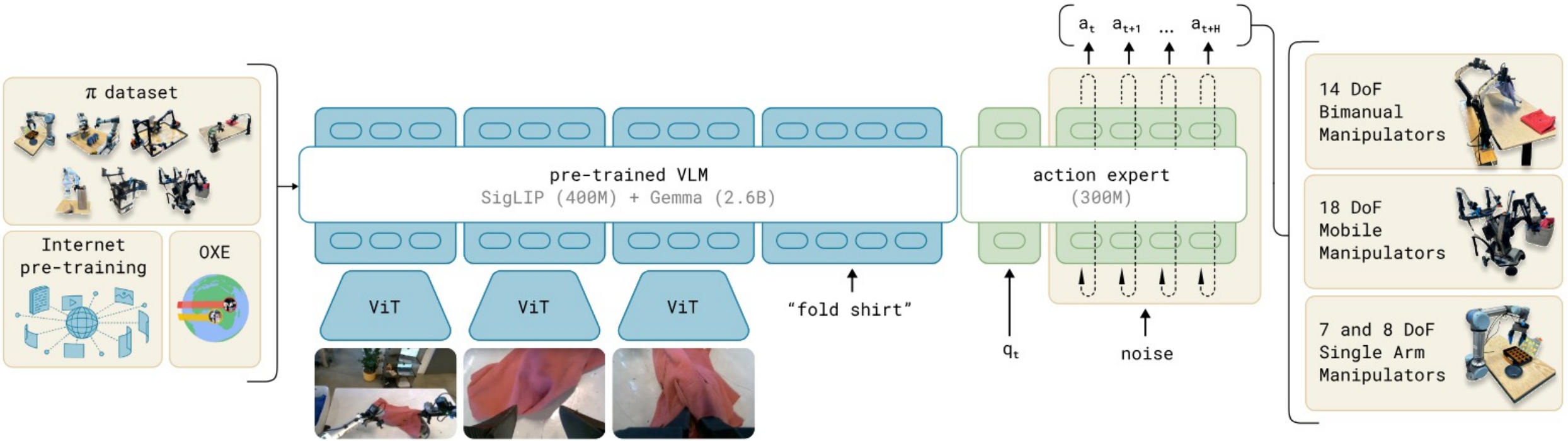
几种VLA架构

OpenVLA：将视觉、语言和机器人状态作为token输入，将动作作为token输出并转化为具体的机器人指令



几种VLA架构

π_0：通过VLA编码视觉输入和文本输入，和机器人状态一起送入动作专家进行动作条件流匹配，生成动作序列

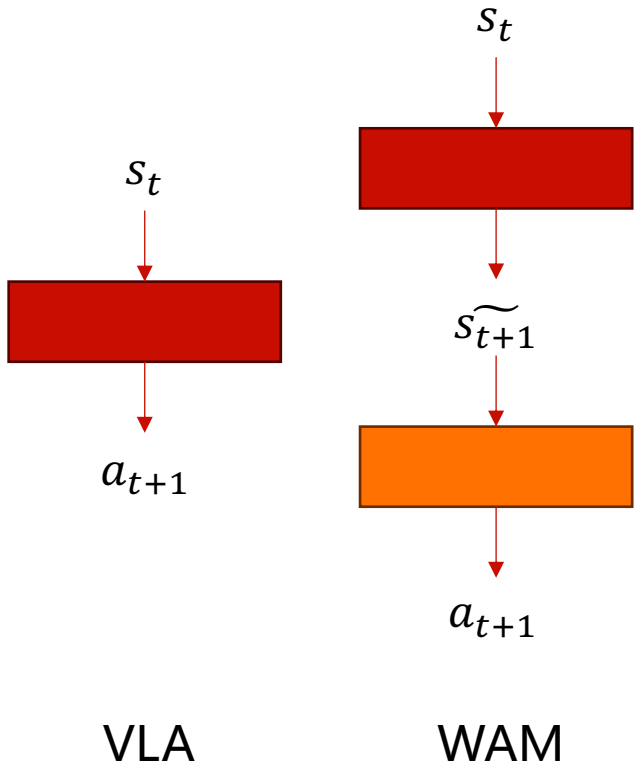


VLA VS WAM

VLA存在问题：

- 缺乏物理知识，泛化能力弱
- 对连续时序逻辑的理解能力较弱，不适合处理长程任务
- 需要跨模态对齐，信息损耗严重

WAM脱胎自世界模型首先根据当前世界状态预测下一步的世界状态，再根据该状态生成需要执行的动作。

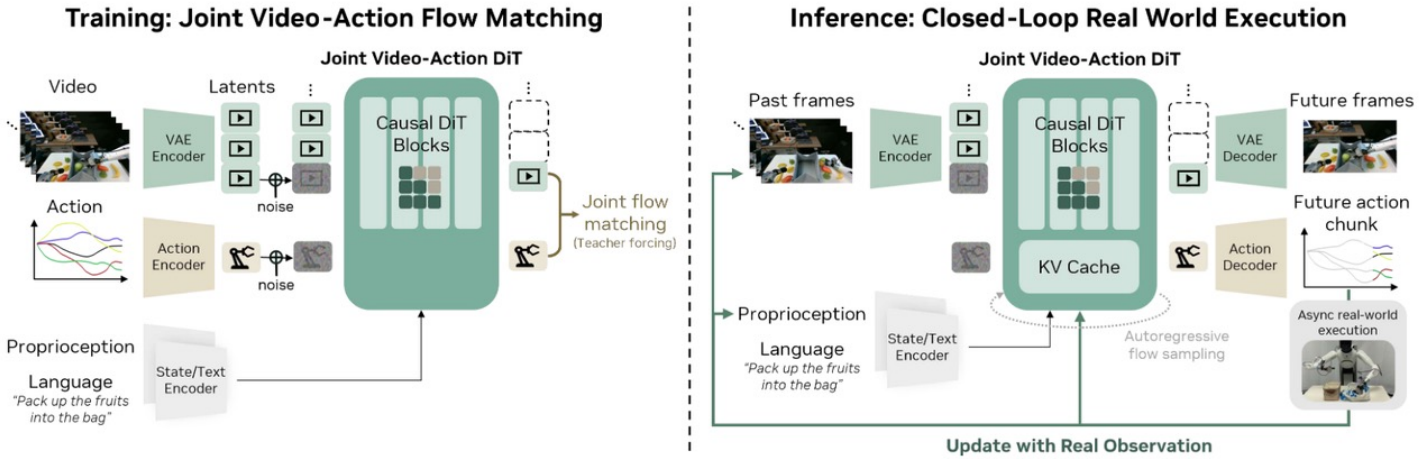


DreamZero

基于视频生成模型（Wan2.1）扩展而来，在此基础上增加状态编码器（编码文本和机器人状态）和动作编解码器实现指令理解和动作输出。

$$\underbrace{\pi_0(\mathbf{o}_{l:l+H}, \mathbf{a}_{l:l+H} \mid \mathbf{o}_{0:l}, \mathbf{c}, \mathbf{q}_l)}_{\text{DreamZero}} = \underbrace{\pi_0(\mathbf{o}_{l:l+H} \mid \mathbf{o}_{0:l}, \mathbf{c}, \mathbf{q}_l)}_{\text{video prediction}} \underbrace{\pi_0(\mathbf{a}_{l:l+H} \mid \mathbf{o}_{0:l+H}, \mathbf{q}_l)}_{\text{IDM}}$$

策略的条件概率输出由视频预测概率和逆动态模型（从图像中生成动作）得来，并使用流匹配损失函数训练。



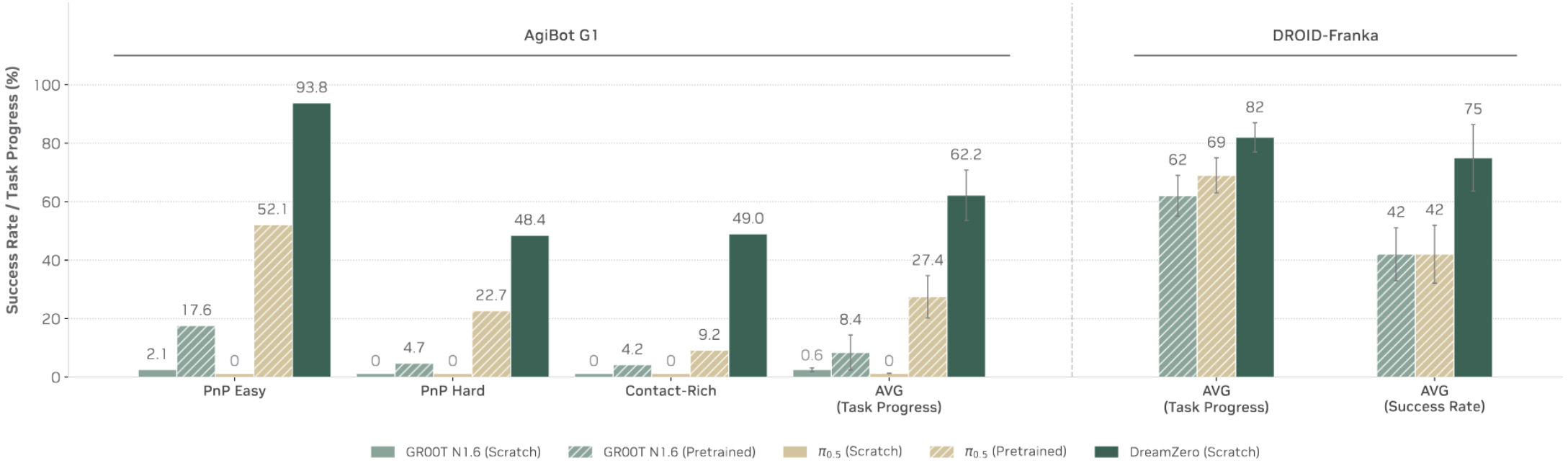
DreamZero

针对推理速度问题，通过CFG并行、DiT缓存、CUDA优化、视频和动作的去噪解耦等手段优化模型推理效率，实现7Hz左右的闭环控制速率。

Optimization	**H100**	**GB200**
Baseline	1×	1.1×
System-level		
+ CFG Parallelism	1.9×	1.8×
+ DiT Caching	5.5×	5.4×
Implementation-level		
+ Torch Compile + CUDA Graphs	8.9×	10.9×
+ Kernel & Scheduler Opts.	9.6×	14.8×
+ Quantization (NVFP4)	—	16.6×
Model-level		
+ DreamZero-Flash	—	38×

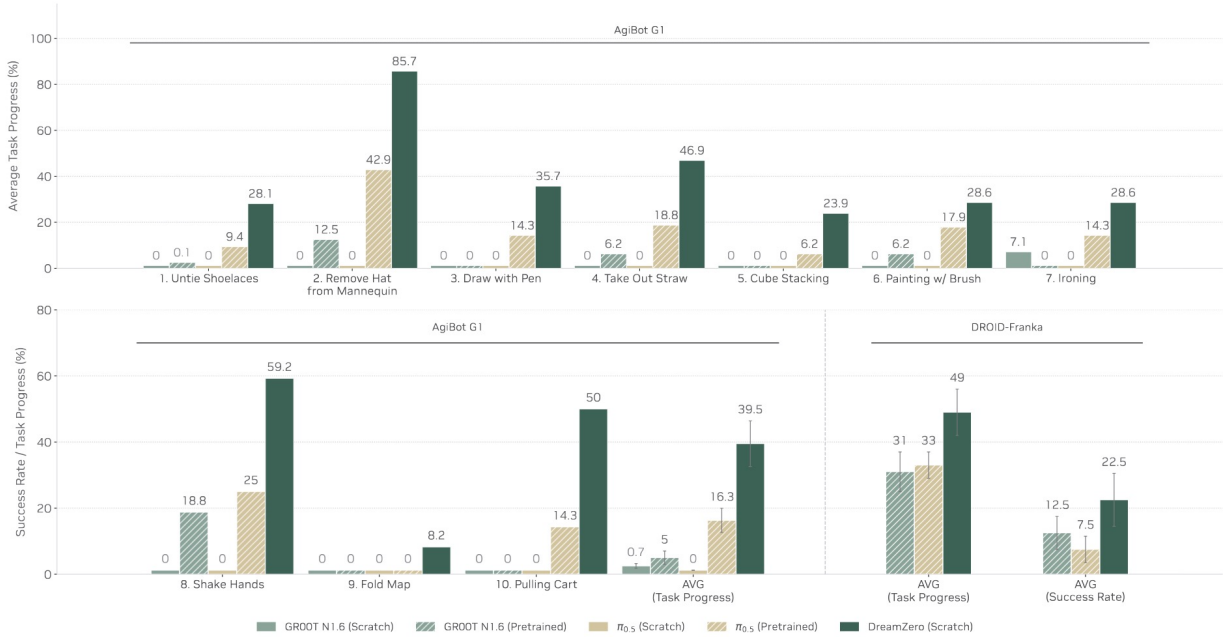
DreamZero

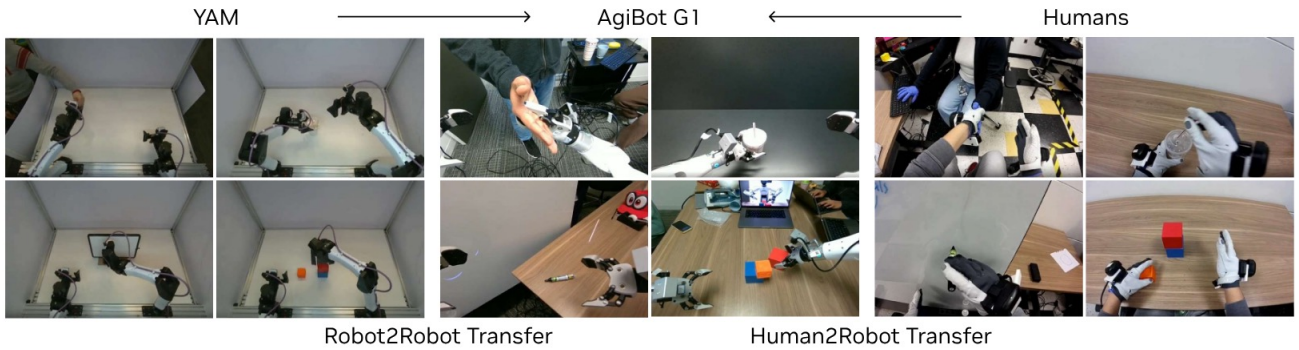
结果：在训练数据分布内的任务上超过GR00T系列、π_0系列等主流VLA模型



DreamZero

结果：在分布外的零样本泛化上结果突出，并且能够在不同机器人形态上进行模型迁移

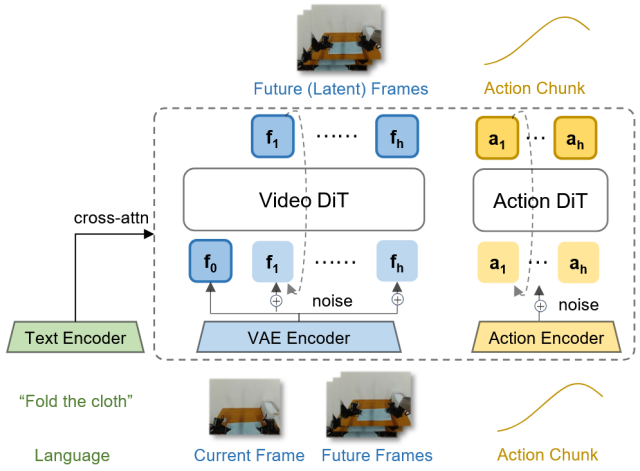




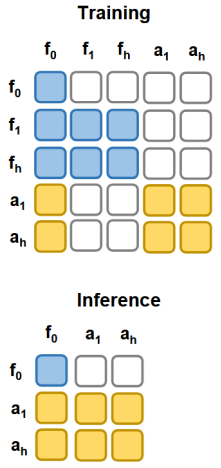
Fast-WAM

放弃在推理阶段显示生成视频，只使用训练阶段引入的视频建模目标指导动作生成，大大提高执行效率。

通过结构化注意力编码，在训练阶段不允许动作去噪器借助未来视频帧的去噪信息进行推理，使得推理阶段能够直接屏蔽视频生成模块，免除巨大的计算开销。



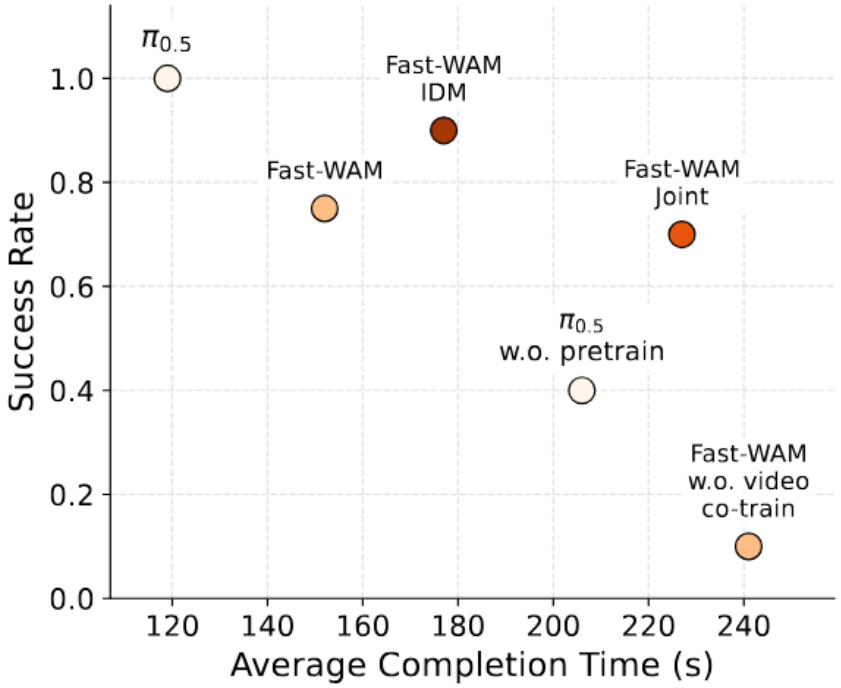
(a) Fast-WAM model architecture.

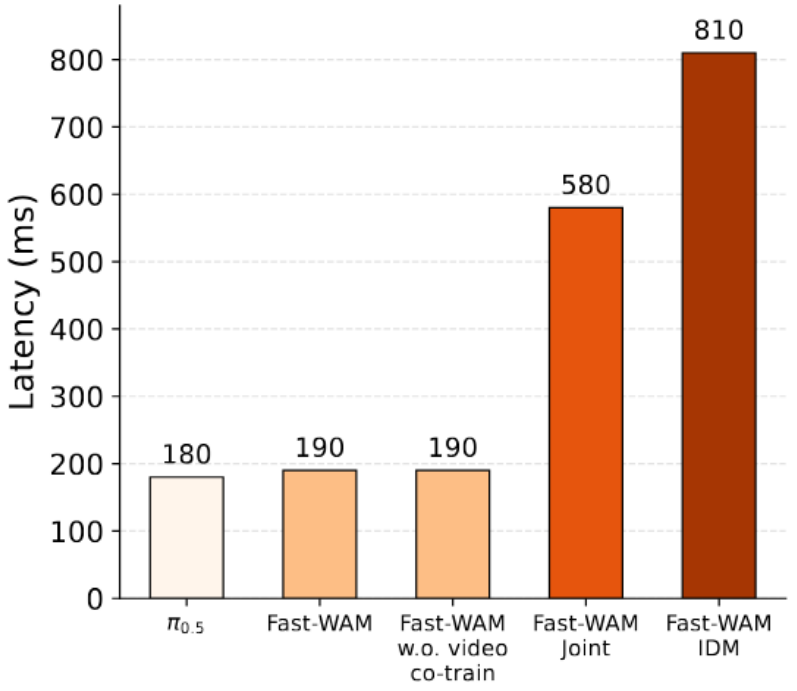


(b) Training and inference masks.

Fast-WAM

结果：能够与主流VLA模型匹敌，在推理速度上也能够接近VLA的水平



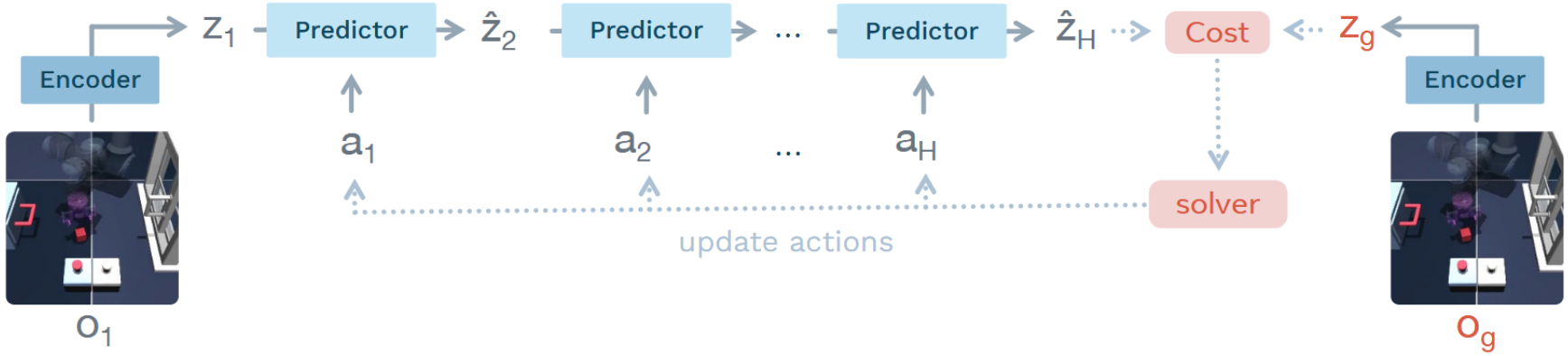


LeWorldModel

基于JEPA工作扩展而来。

视觉编码器将像素帧编码为隐变量，动态预测器从隐变量和动作预测下一个时间步的隐变量，推理阶段根据开销求解最优动作。

借助SIGReg正则化损失，要求隐变量的聚合分布与标准高斯分布在拓扑上等价，防止表示坍塌（模型对所有输入都输出统一表示）



总结

世界模型代表着人工智能从理解语言、图像等模态向“理解物理世界”的演进，赋予人工智能预测未来的能力，从而实现主动响应。

在自动驾驶领域，它作为“虚拟训练场”和“数据生成器”，让车辆能在极端长尾场景中提前推演、自主进化；在机器人领域，它赋予了机械臂与具身智能体手眼协调与空间预判能力；此外，它在医疗推演、新药研发、城市数字孪生等前沿领域也展现出巨大的商业潜力，被视为下一代AGI的通用操作系统。

作为人工智能的范式的一部分，方兴未艾的世界模型和已经趋于成熟的大模型一起，探索机器智能的边界到底在哪里。

总结

- 世界模型的动机与定义
- 世界模型分类
  - 渲染器：视频生成模型、3D重建方法、交互式渲染
  - 模拟器：JEPA系列、GE Sim2.0
  - 规划器：WAM（DreamZero、Fast-WAM、LeWorldModel）

致 谢

- 胡玥、曹亚男、方芳：国科大《自然语言处理基础》
- 曹亚男、任昱冰：国科大《深度学习与自然语言处理概述》


THANKS
https://ictkc.github.io/teaching/2026spring-nlp