大模型系列

世界模型

ZOMI

关于本内容

1. 世界模型之争与定义
2. OpenAI：SORA 技术解读
3. Google：Genie 技术解读
4. META：I-JEPA 到 V-JEPA 技术解读
5. 总结与思考：谁才是世界模型的未来？

1.世界模型之争与定义

https://artificialcognition.net/posts/video-generation-world-simulators/#concluding-thoughts

Yann LeCun 怒斥 Sora

Aditya Ramesh @model_mechanic · 2月18日

"pov footage of an ant navigating the inside of an ant nest"

Video generated by Sora

287 1,122 7,510 319万

Yann LeCun
@ylecun

Hi Aditya, ants have 6 legs, no?

上午4:11 · 2024年2月19日 · **21.5万** 查看

Yann LeCun @ylecun · Mar 1

Nice piece on what world models really are (or should be).

Raphaël Millière @raphaelmilliere · Feb 29

OpenAI unveiled its video generation model Sora two weeks ago. The technical report emphatically suggests that video generation models like Sora are world simulators. Are they? What does that even mean? I'm taking a deep dive into these questions in a new blog post (link below).

18 28 241 93K

Yann LeCun 怒斥 Sora

自回归模型将死！杨立昆：你们这条路行不通

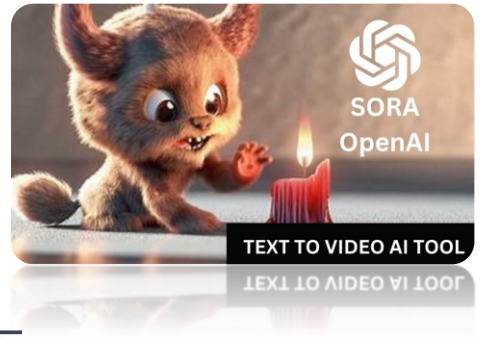
原创 CSDN AI科技大本营 2024-03-22 16:34 北京

整理 | 王启隆

出品 | AI 科技大本营（ID：rgznai100）

当今人工智能界有三位“教父”，其中对人工智能风险问题最为乐观的便是图灵奖得主 & CNN 之父 Yann LeCun（杨立昆）。LeCun 如今是 Meta 的首席人工智能科学家，也是纽约大学的教授，他常在各大会议与社交媒体上发声，与其他科学家甚至另外两位教父展开辩论。

谁才是世界模型老大？



世界模拟器

- 2024.02.15 SORA发布

世界模型

- 2024.03.01 V-JEPA发布
- 2024.03.05 IWM 发布

- 2024.02.26 Genie发布

基础世界模型

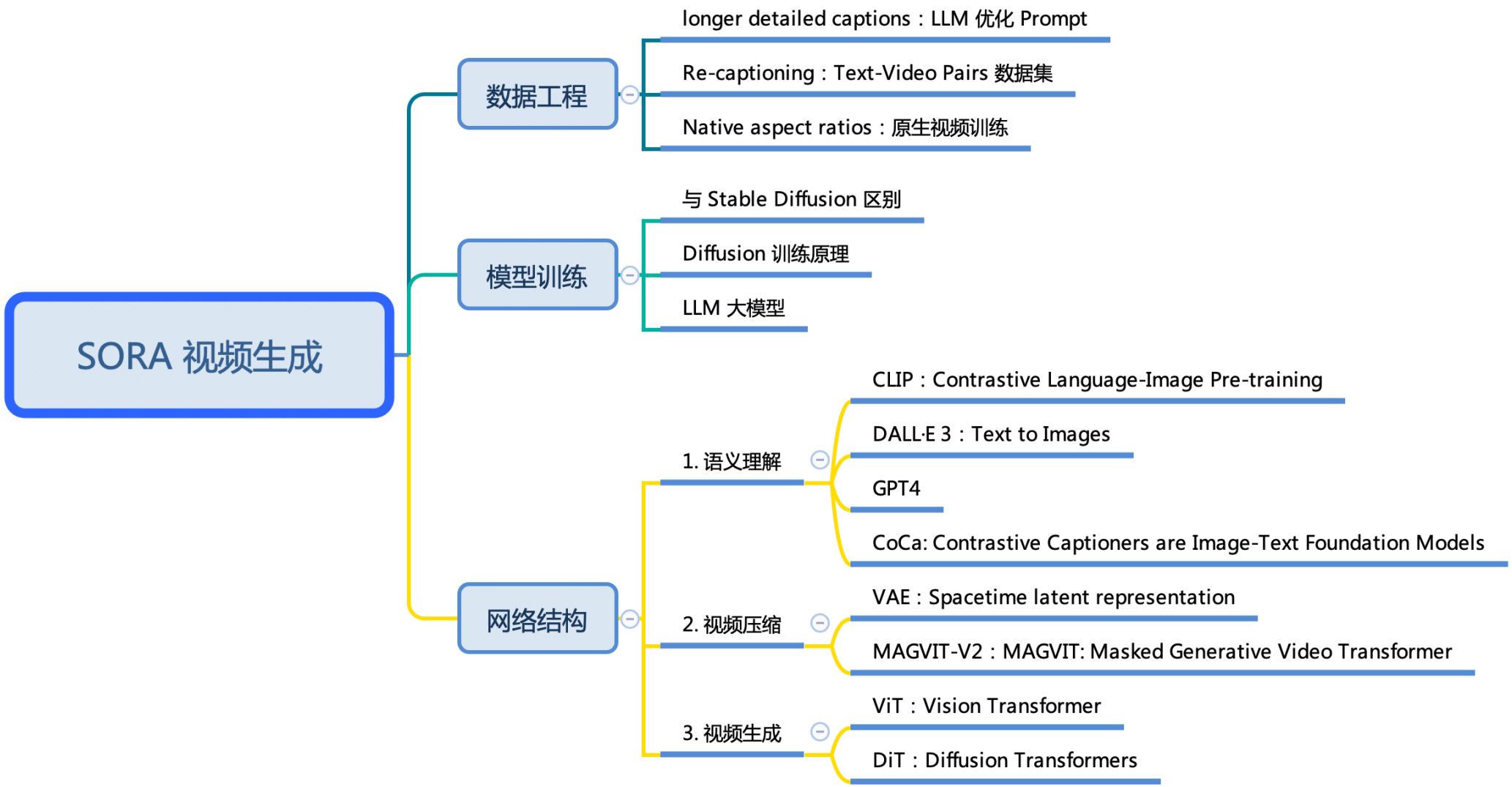
AGI 通用智能之路流派

1. **自回归生成式：**
   - 以 OpenAI 为代表 Transformer（SORA），通过 Scaling Law 大规模/大数据/大算力，以自回归方式走向 AGI。
2. **联合嵌入预测架构：**
   - Yann LeCun 为代表的世界模型学派，学习人类和动物能够通过观察、交互，以及无监督方式学习世界知识，通过隐藏的学习能力构成了常识的基础。
3. **生成式交互环境：**
   - 借鉴强化学习对世界模型的定义，重新定义生成式AI的新范式，即生成式交互环境。

1. Sora 技术解读

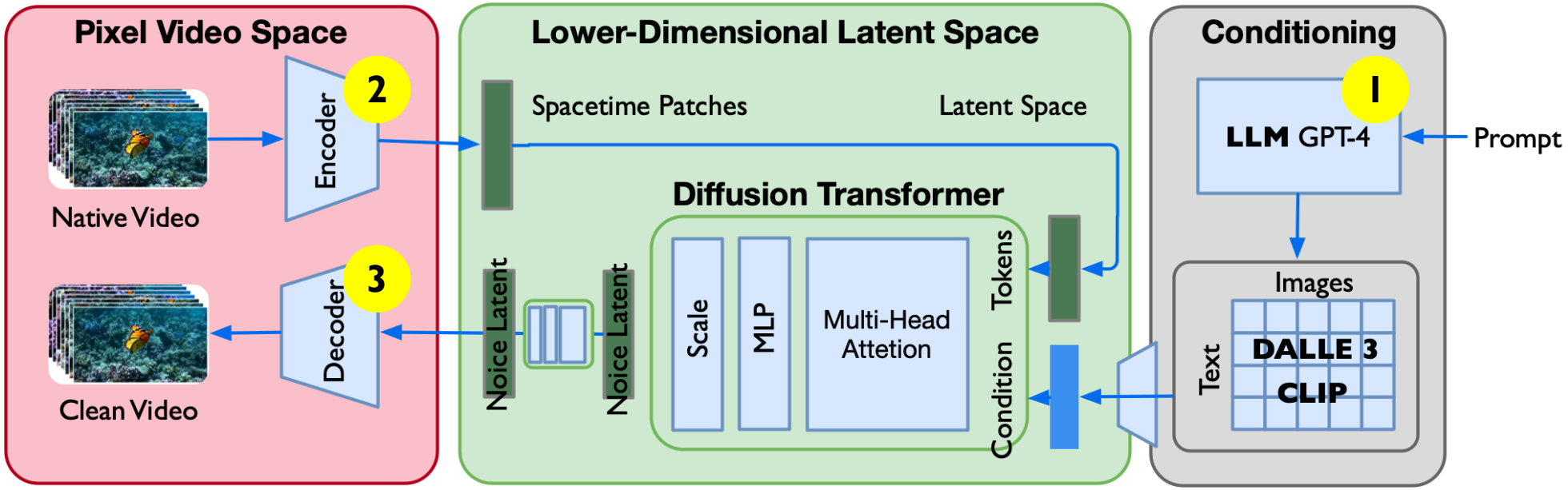
Sora 效果欣赏

相关技术架构


SORA 视频生成
数据工程
longer detailed captions：LLM 优化 Prompt
Re-captioning：Text-Video Pairs 数据集
Native aspect ratios：原生视频训练
模型训练
与 Stable Diffusion 区别
Diffusion 训练原理
LLM 大模型
网络结构
1. 语义理解
CLIP：Contrastive Language-Image Pre-training
DALL·E 3：Text to Images
GPT4
CoCa: Contrastive Captioners are Image-Text Foundation Models
2. 视频压缩
VAE：Spacetime latent representation
MAGVIT-V2：MAGVIT: Masked Generative Video Transformer
3. 视频生成
ViT：Vision Transformer
DiT：Diffusion Transformers

SORA 模型训练流程

- Step1：使用 DALLE 3（ CLIP ） 把文本和图像对 <text，image> 联系起来；
- Step2：视频数据切分为 Patches 通过 VAE 编码器压缩成低维空间表示；
- Step3：基于 Diffusion Transformer 从图像语义生成，完成从文本语义到图像语义进行映射；
- Step4：DiT 生成的低维空间表示，通过 VAE 解码器恢复成像素级的视频数据；



全网最详细：SORA 视频生成大模型原理剖析

Sora 目标

- 文本 → 视频（text-to-video），核心在于理解复杂的文本提示，并转化为长视频：
  - **理解与生成**：需要理解文本含义，并基于此生成视频，要求模型具备 NLP 理解 & 视频生成能力。
  - **无交互过程**：生成视频过程，Sora 不涉及与用户交互，一开始用户给出指令后，模型 E2E 完成视频内容的生成。

2. Genie 技术解读

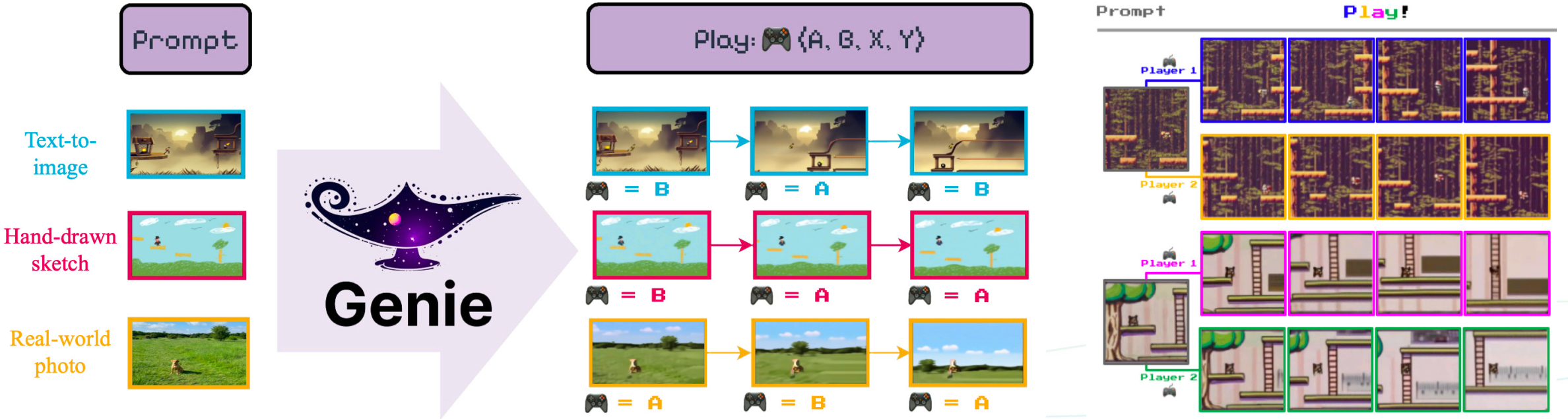
https://arxiv.org/abs/2402.15391

https://sites.google.com/view/genie-2024/

Genie 效果欣赏

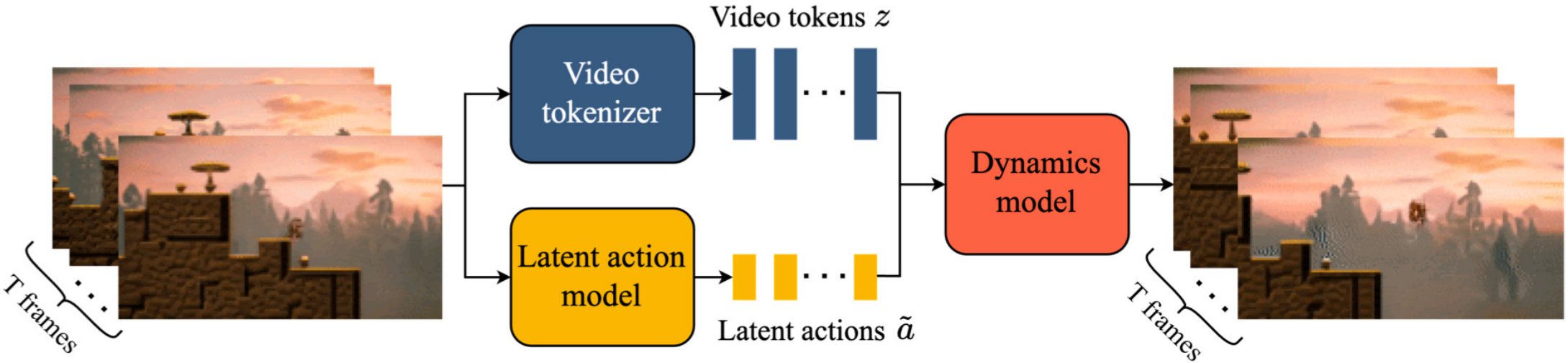
Genie 概述

- **模型规模**：< 11B
- **输入**：图像、视频
- **数据**：30K 小时无标注视频
- **交互**：控制 Agent 移动
- **输出**：生成移动后的下一帧图像

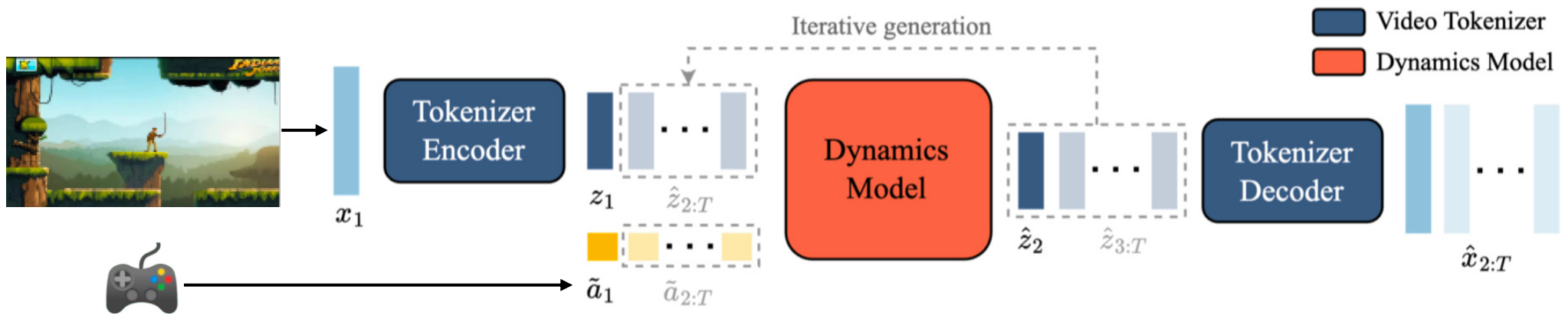


推理和训练流程

- 训练



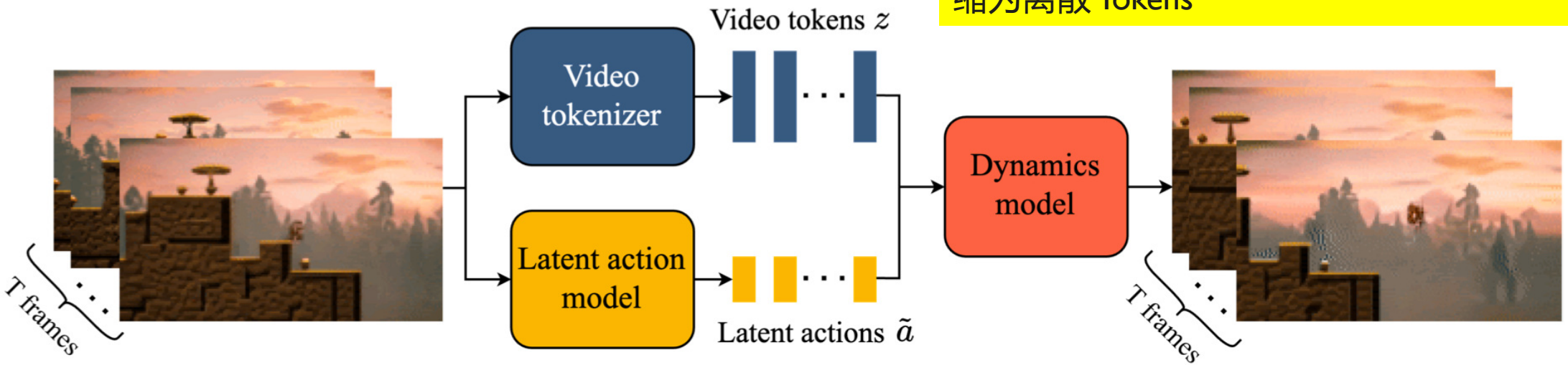
- 推理



Genie 核心优势 & 模型

- **特点**：无监督学习，没有动作标签，从视频中学习生成可控部分，并通过潜在动作进行交互。
- **组成**：Video Tokenizer、 Latent Action Model 和 dynamics model。
- **模型**：ST-Transformer（ spatial-temporal ）。

视频分词（ Video Tokenizer ）：将视频压缩为离散 Tokens



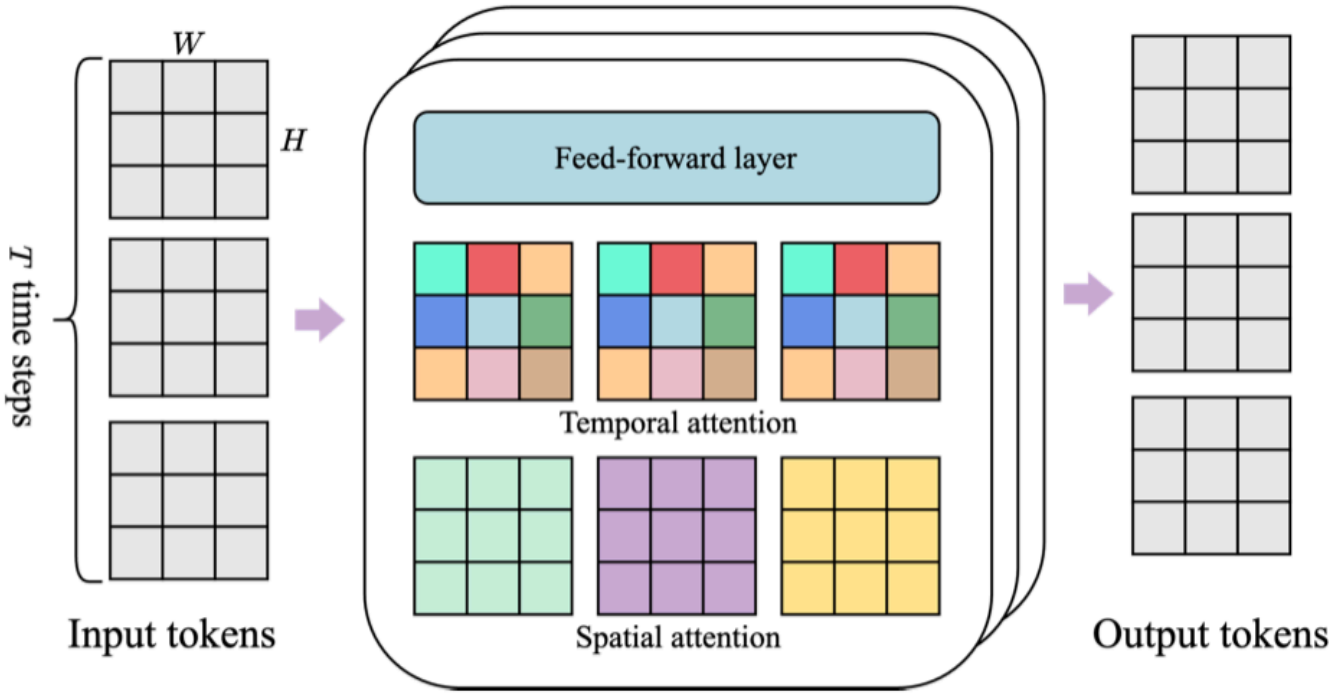
潜在动作模型（ Latent Action Model ）：学习连续帧之间的潜在动作

潜在动态模型（ Dynamics Model ）：输入潜在动作和过去帧 Token ，预测下一帧

ST-Transformer 结构：专用于处理视频数据

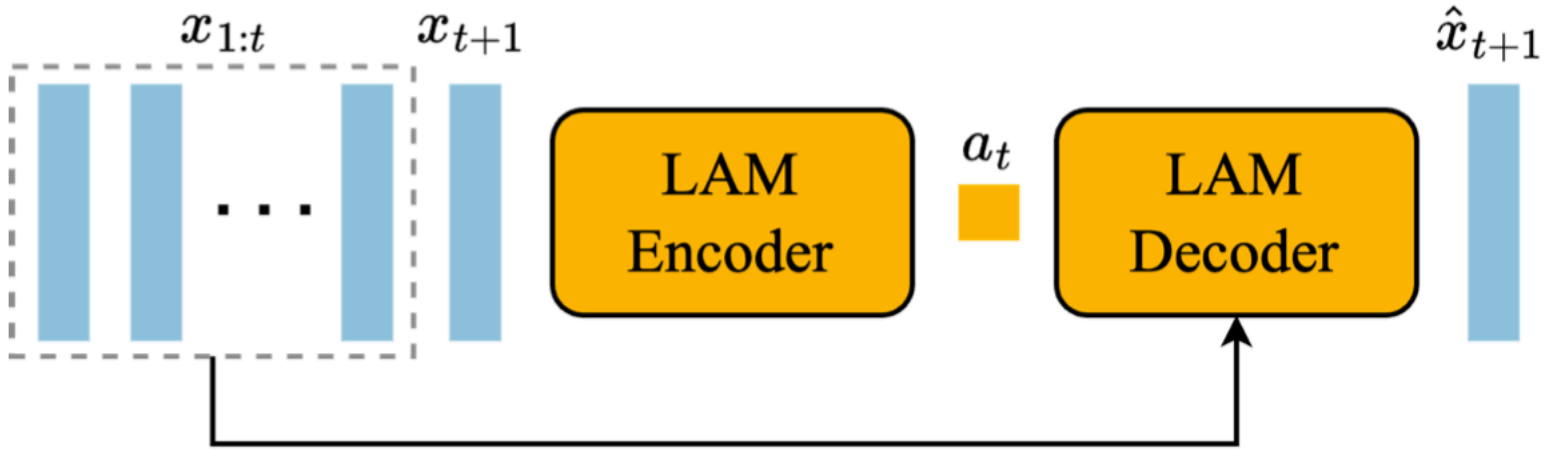
- **时空块（ST blocks）**：SA 在每个时间点内关注 Spatial 上相邻 Token（只在 1xHxW 计算 Attention）；TA 关注跨空间 Token（只在 Tx1x1 计算 Attention）。避免 TxHxW 的矩阵计算复杂度。
- **线性计算复杂度**：计算复杂度与视频帧数线性相关，而非平方关系，高效处理长视频序列。
- **前馈层**：空间和时间处理之后再执行 FFN 。有助于模型处理更复杂的组件，并显著提升了性能。

Parameters	num_layers	num_heads	d_model	k/q size	FLOPs
10.1B	48	36	5120	128	6.6×10^{22}



潜在动作模型 LAM， Latent Action Models

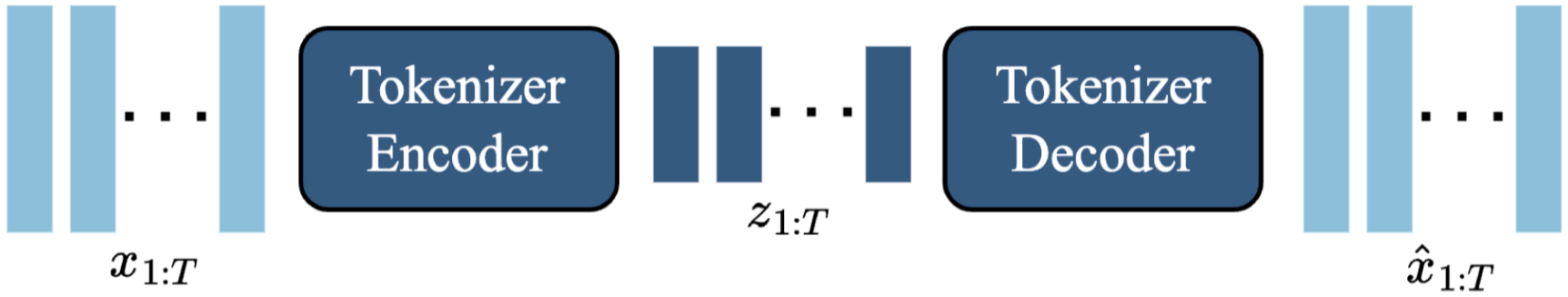
- LAM encoder：输入1:t 帧图像和 t+1 帧图像，通过 ST-Transformer，生成连续动作概率 $\hat{a}_{1:t}$；
- LAM decoder：输入1:t 帧图像和前面帧动作概率 $\hat{a}_{1:t}$，预测 t+1 帧图像 $\hat{x}_{t+1}$；



Component	Parameter	Value
Encoder	num_layers	20
	d_model	1024
	num_heads	16
Decoder	num_layers	20
	d_model	1024
	num_heads	16
Codebook	num_codes	8
	patch_size	16
	latent_dim	32

视频分词器 SVT， Spatiotemporal Video Tokenizer

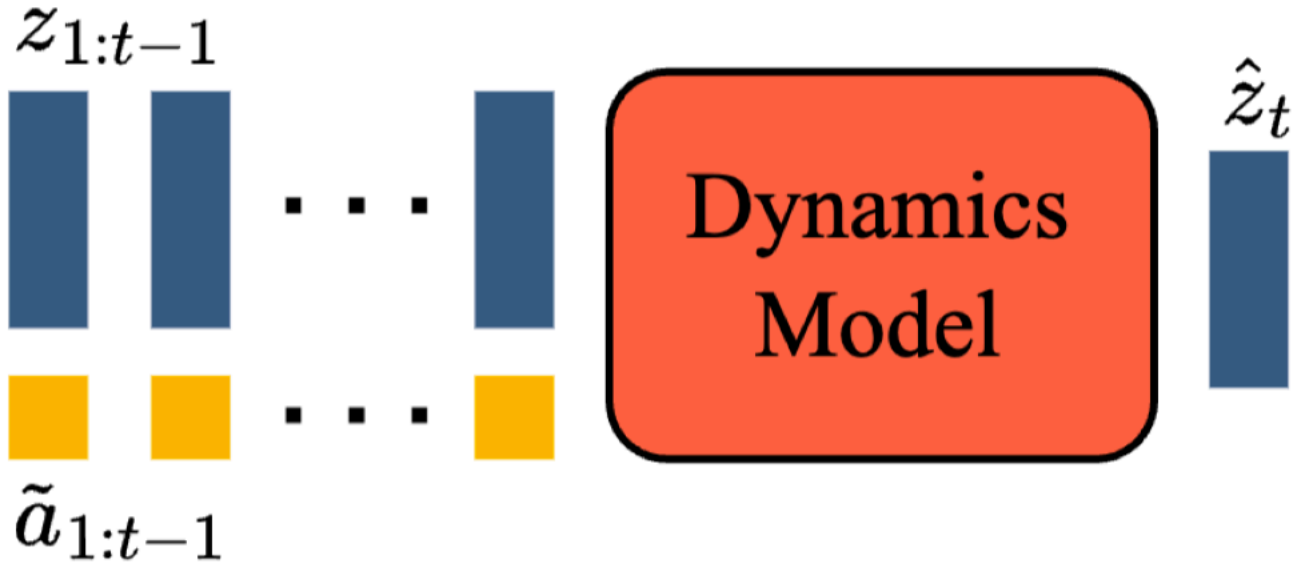
- 采用 VQ-VAE 训练方式，网络结构使用 ST-Transformer，让交互过程在视频的离散表示空间。
- TxHxWxC 维视频 $x_{1:T}$ 作为输入，每帧生成 D 维离散表示 $z_{1:T}$（转为长度 T 离散 Tokens）。



Component	Parameter	Value
Encoder	num_layers	12
	d_model	512
	num_heads	8
	k/q_size	64
Decoder	num_layers	20
	d_model	1024
	num_heads	16
	k/q_size	64
Codebook	num_codes	1024
	patch_size	4
	latent_dim	32

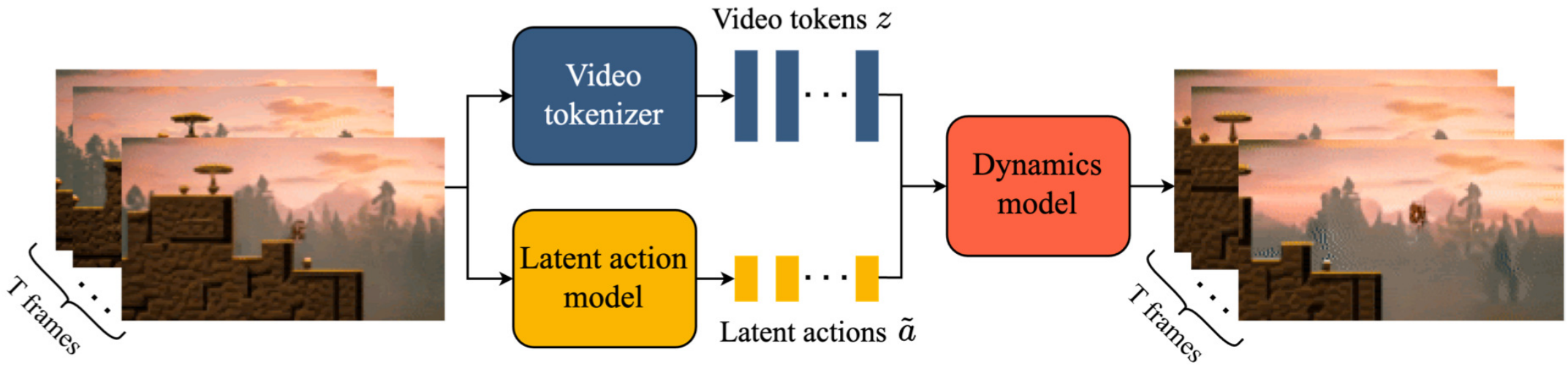
潜在动态模型 DM，Dynamics Model

- 模型：decoder-only 的 MaskGIT (Chang et al., 2022) ST-Transformer；输入 Tokenizer 后的视频空间 $z_{1:t-1}$ 和动态概率 $\hat{a}_{1:t}$，输出下一帧的Tokenizer $\hat{z}_{2:t}$ 和真实 $\hat{z}_{2:t}$ 进行比较计算 Loss。



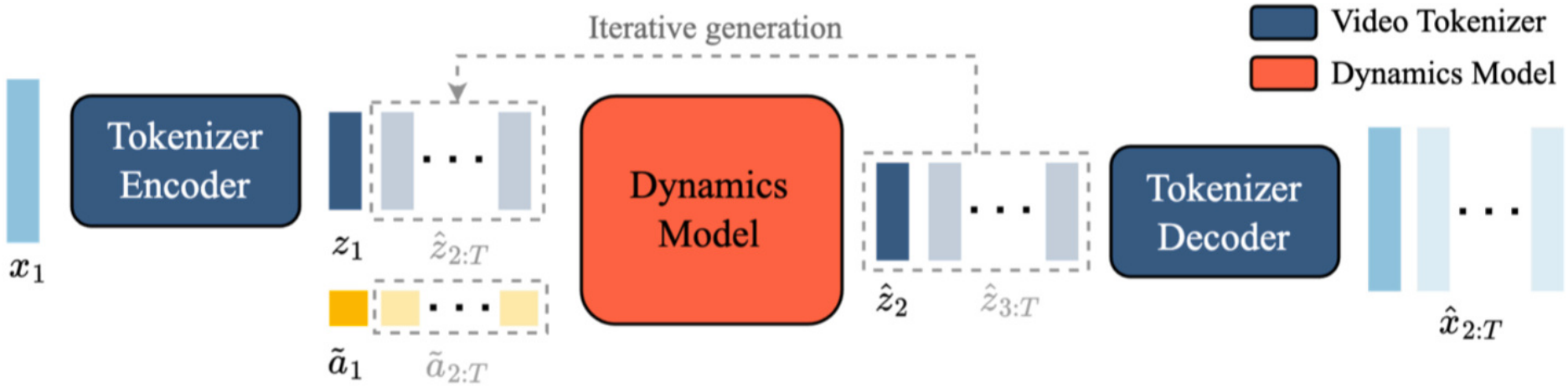
学习流程

- **训练过程：1）学习 video tokenizer，2）联合训练 latent action model 和 dynamics model；**
  - **时空视频分词器**：Genie 首先通过时空视频分词器将原始视频帧转换为离散的 token。这一步骤允许模型捕捉视频中的关键视觉信息，并为其后的处理步骤提供一个统一的表示。
  - **潜在动作模型（LAM）**：随后，模型通过潜在动作模型推理每对帧之间的潜在动作。在这个过程中，模型试图理解视频中的哪些动态变化是可控的，而不需要任何外部的动作标签。
  - **动态模型训练**：在理解了潜在的交互动作后，模型会联合训练动态模型，以预测在给定潜在动作的情况下，环境的未来状态。

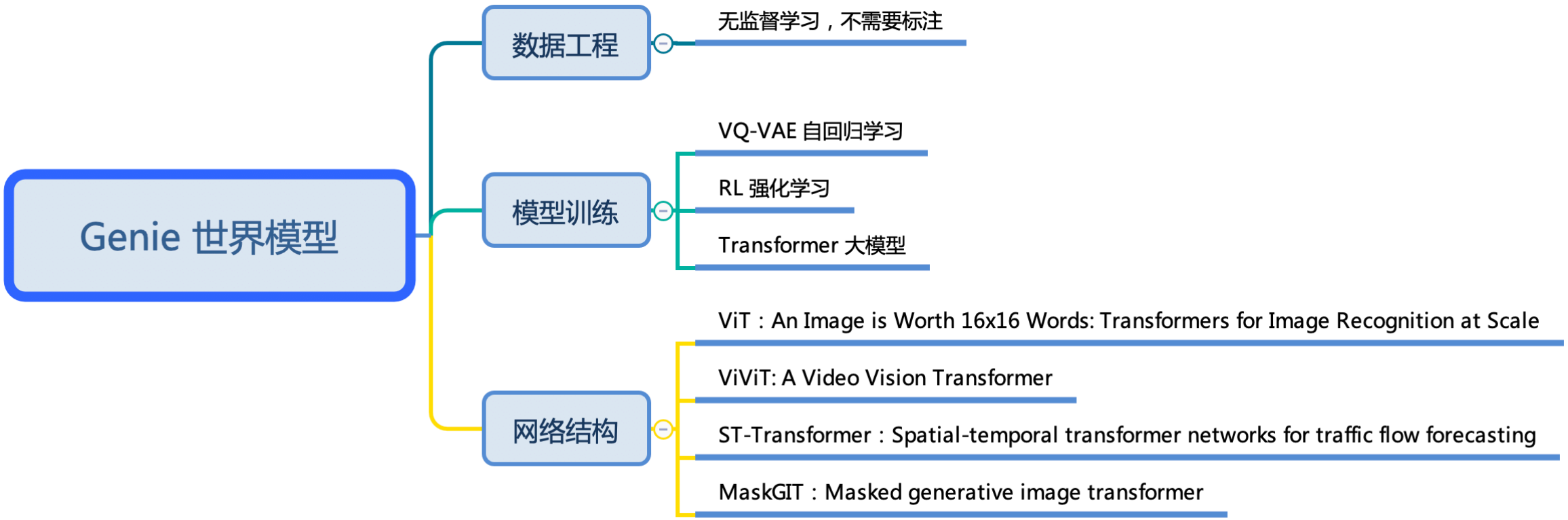


推理流程

- 使用时，动作变成用户给定，不需要使用 LAM 模块。
- 推理和交互：根据输入初始图像和潜在动作作为条件输入，预测下一帧。使得生成的环境可以根据用户的交互动作进行实时更新。



相关技术架构


Genie 世界模型
数据工程
无监督学习，不需要标注
模型训练
VQ-VAE 自回归学习
RL 强化学习
Transformer 大模型
网络结构
ViT：An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale
ViViT: A Video Vision Transformer
ST-Transformer：Spatial-temporal transformer networks for traffic flow forecasting
MaskGIT：Masked generative image transformer

Genie 目标

- 创建能生成交互式、可互动的环境基础世界模型，图像 → 游戏环境（image-to-game）：
  - **生成可交互环境**：生成低像素视觉内容，创建可交互环境，允许用户通过潜在动作来控制生成的环境，从而影响下一帧图像的生成。
  - **无监督学习**：通过无监督学习从视频中学习，推断生成环境中一致的潜在动作，使模型在缺乏明确动作标签下学习对动作的控制。

技术问题与讨论

- Scaling Law：更大的数据集上训练、更大的 Video Encoder 和 latent action 模型；
- 交互方式：利用视频学习未来动作把 RL 的环境节省了，但是交互方式受限于游戏视频；

- 数据：除了音、视、图、文以外，动作空间 Action Space 也可以作为一种新模态；
- 算力：视频对算力需求随世界模型出现而涌现，视频生成和压缩会消耗大量算力（CPU+AI）；
- 模型：通过视频 Latent space 来学习具体 action 构建世界模型 AG 或许是一个方向；

2. I-JEPA

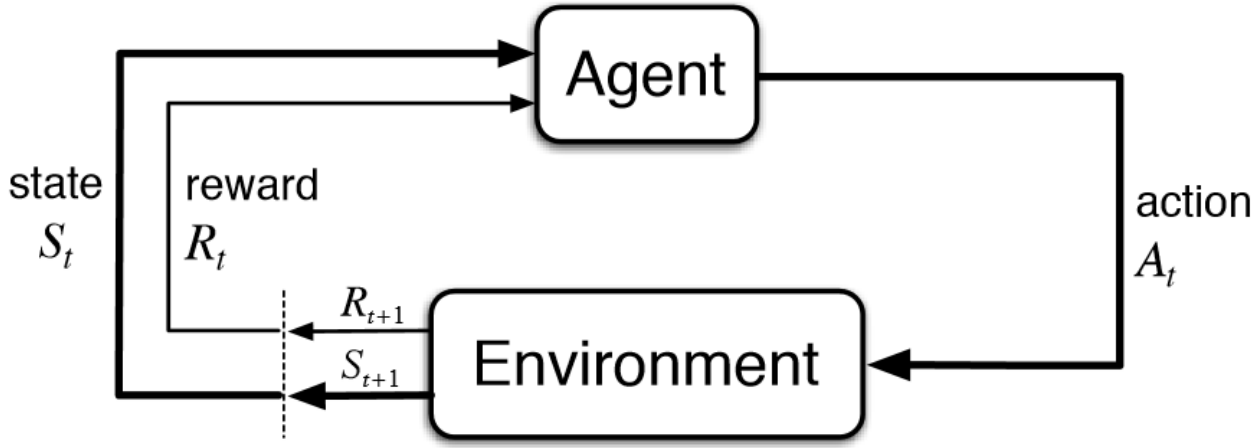
V-JEPA 解读

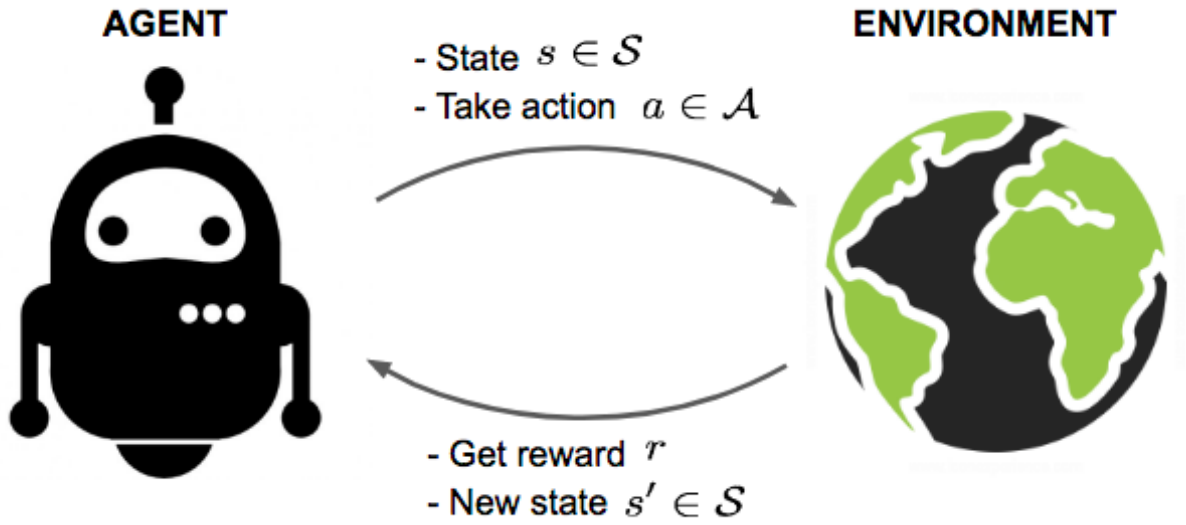
Joint Embedding Predictive Architectures 联合嵌入预测架构

- 相关 JEPA 架构及 I-JEPA / V-JPA 模型主打“预测能力”，号称可以以“人类理解”方式，利用抽象性高效预测生成图片 / 视频中被遮蔽的部分，从而让模型在填充间学习场景，进一步预测未来事件或动作，进而达到对世界更深层次的理解。

1. 2022 年 6 月：发布 A Path Towards Autonomous Machine Intelligence 定义世界模型和 JEPA 架构；
2. 2023 年 4 月：推出基于世界模型概念模型 I-JEPA，通过真实世界图像学习潜在抽象表征；
3. 2024 年 2 月：Sora 发布后第二天推出视频模型 V-JEPA，让机器通过学习视频了解世界运作方式。
4. 2024 年 3 月 1日：发布 IWM 图像世界模型，揭示利用世界模型进行表征学习，赋予世界模型的更大容量可以直接影响所学表征的抽象程度。

reinforcement learning 强化学习





JEPA

1

A Path Towards Autonomous Machine Intelligence
Version 0.9.2, 2022-06-27

Yann LeCun
Courant Institute of Mathematical Sciences, New York University yann@cs.nyu.edu
Meta - Fundamental AI Research yann@fb.com

June 27, 2022

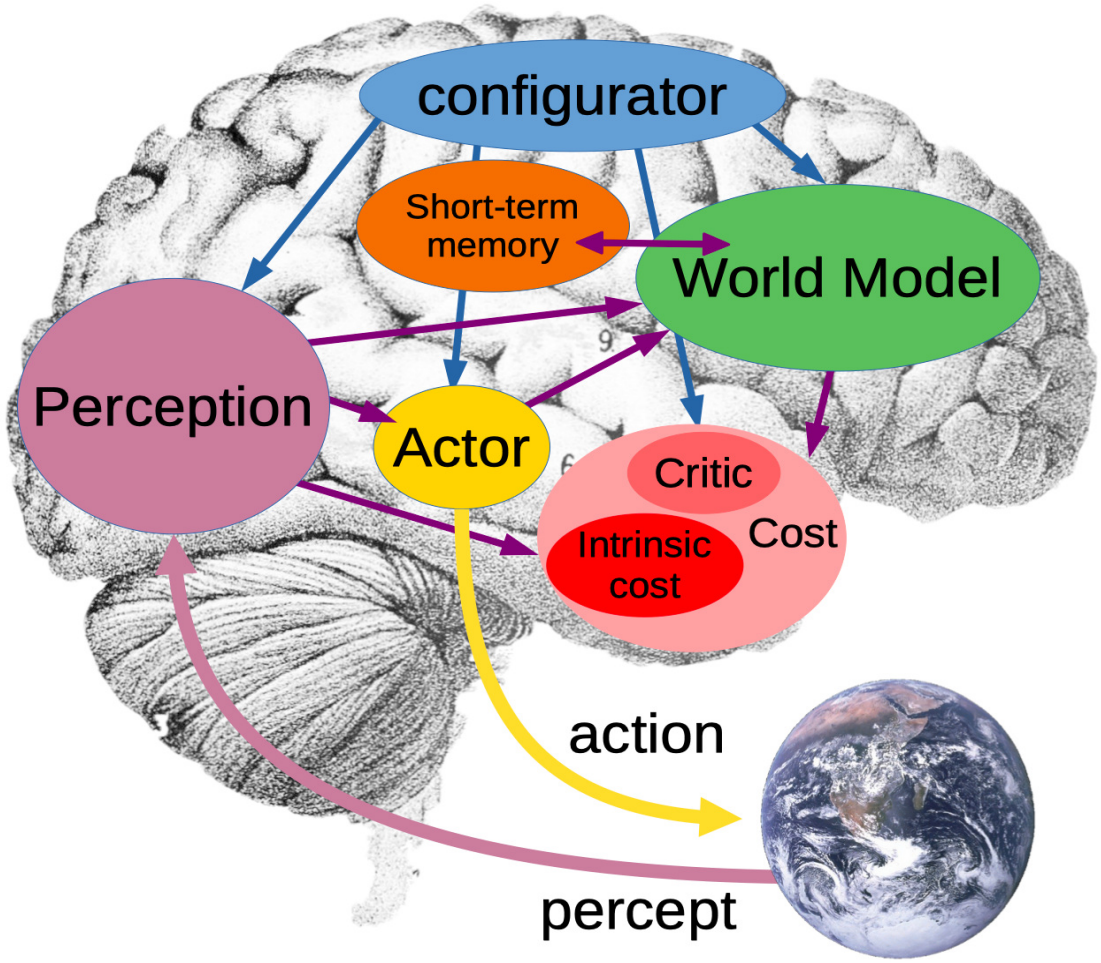
Abstract

How could machines learn as efficiently as humans and animals? How could machines learn to reason and plan? How could machines learn representations of percepts and action plans at multiple levels of abstraction, enabling them to reason, predict, and plan at multiple time horizons? This position paper proposes an architecture and training paradigms with which to construct autonomous intelligent agents. It combines concepts such as configurable predictive world model, behavior driven through intrinsic motivation, and hierarchical joint embedding architectures trained with self-supervised learning.

Keywords: Artificial Intelligence, Machine Common Sense, Cognitive Architecture, Deep Learning, Self-Supervised Learning, Energy-Based Model, World Models, Joint Embedding Architecture, Intrinsic Motivation.

JEPA：自主智能世界模型 Autonomous Intelligence World Model 1

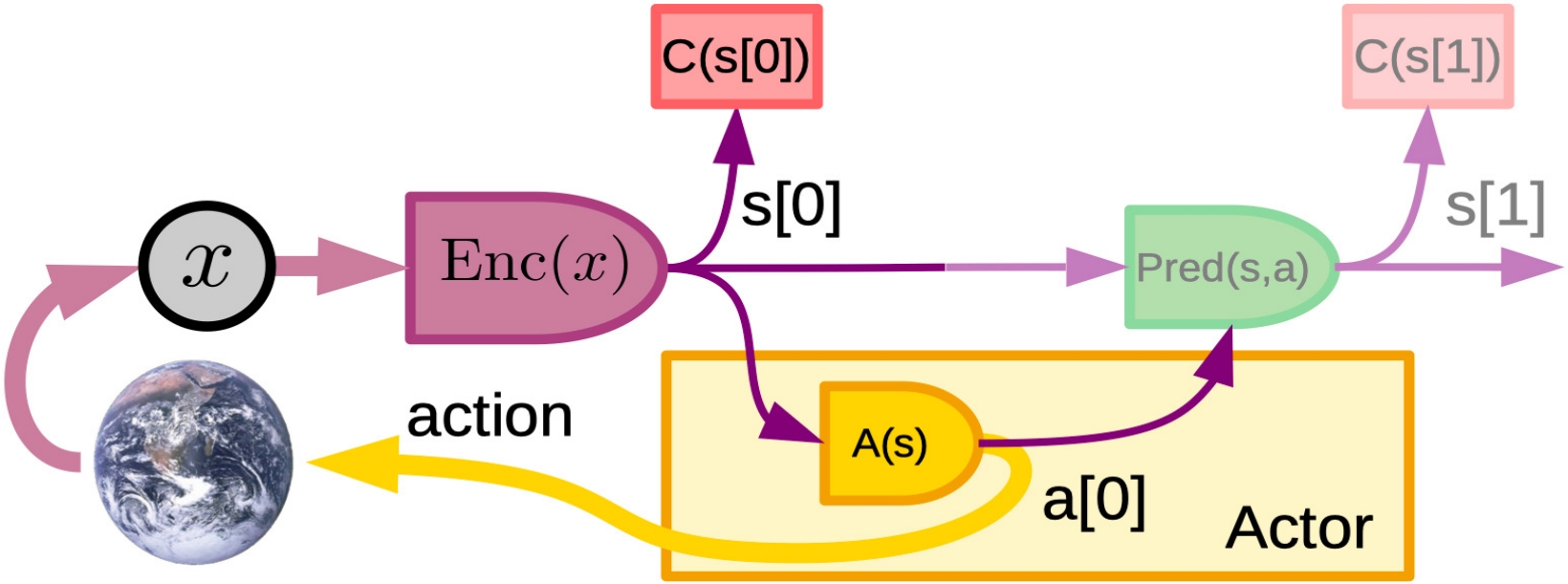
- configurator：配置模块以执行具体任务。
- Perception：感知模块估计世界当前状态。
- World Model：预测未来世界状态。
- Cost：衡量状态的代价，两子模块组成：
  - Intrinsic Cost： 计算当前状态的成本代价；Critic Cost：预测未来动作的成本代价；
- STM ：短期记忆对世界模型进行记录。
- Actor：计算并执行具体动作。



JEPA：Mode-1 perception-action episode 感知-动作环节

1

- 感知-动作：感知模块估计世界状态 $s[0] = Enc(x)$，Actor 通过策略 $a[0] = A(s[0])$ 计算出要执行的动作。
- 成本模块计算初始状态 $f[0] = C(s[0])$ 的成本并将 $(s[0], f[0])$ 存储在短期记忆中。
- 可以使用世界模型 $s[1] = Pred(s[0], a[0])$ 和成本 $f[0] = C(s[0])$ 预测下一状态，以便调整世界模型。

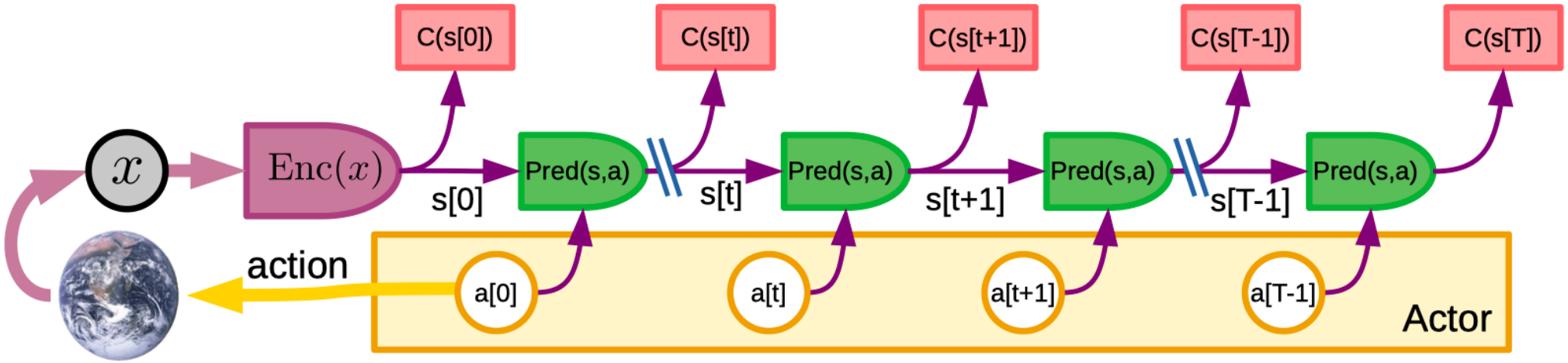


JEPA： Mode-2 perception-action episode 感知-动作环节

1

- 感知-动作：感知模块估计世界状态 $s[0] = Enc(x)$，Actor 通过策略 $a[0] = A(s[0])$ 计算出要执行的动作。
- 成本模块计算初始状态 $f[0] = C(s[0])$ 的成本并将 $(s[0], f[0])$ 存储在短期记忆中。
- 可以使用世界模型 $s[1] = Pred(s[0], a[0])$ 和成本 $f[0] = C(s[0])$ 预测下一状态，以便调整世界模型。

将Model1 中单一个环节**拓展成序列** → 马尔科夫链，使用**贝尔曼方程**求解**马尔可夫决策过程** → 使用基于梯度方法搜索最优动作序列 → **使用深度学习求解**



JEPA：主要贡献

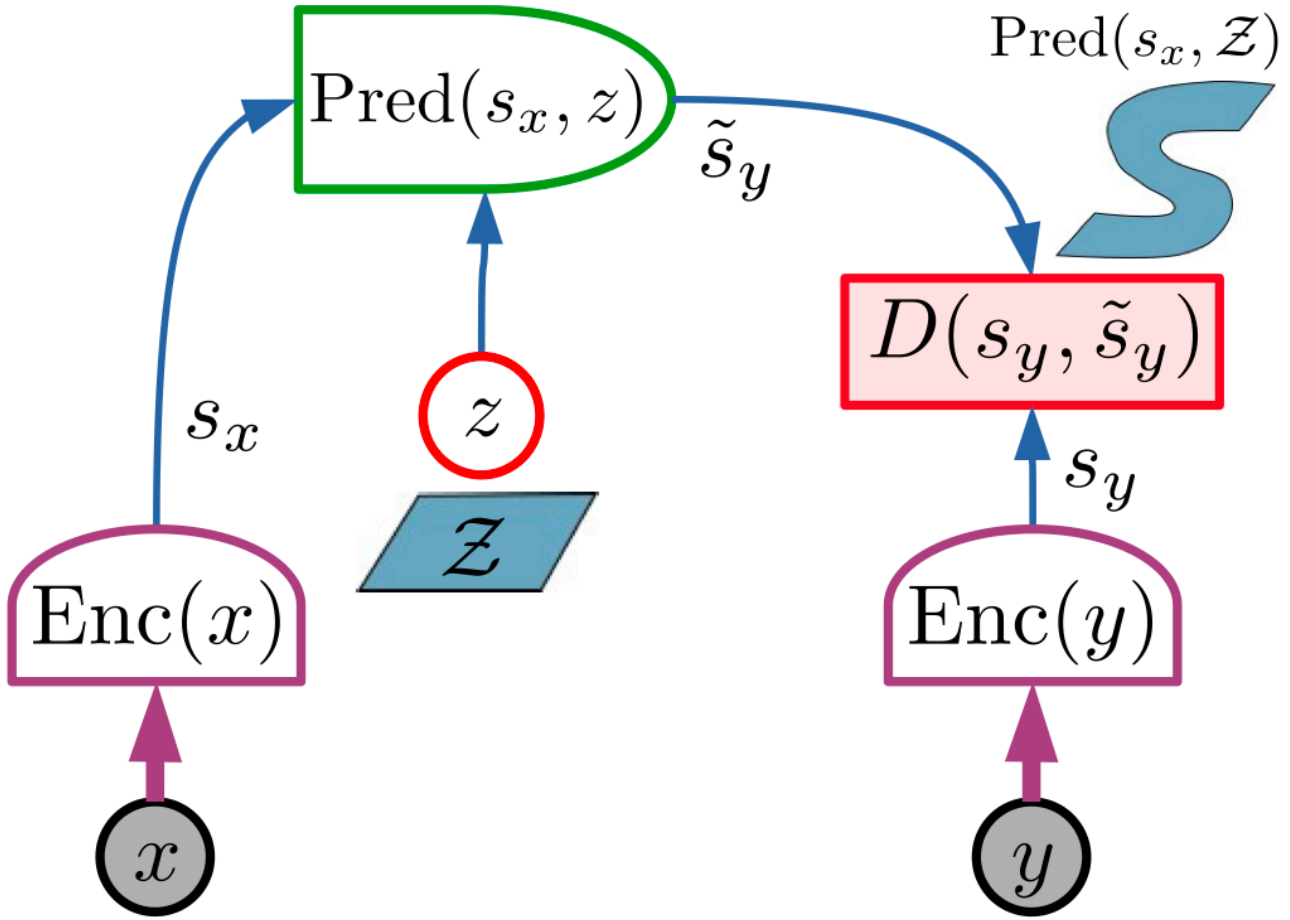
1

- 为世界模型设计架构和训练范式是几十年人工智能发展真正取得进展的主要障碍。
- JEPA 主要贡献正是分层架构和世界模型的训练流程，通过预测表示多个结果：

1. Self-Supervised Learning，SSL 自监督学习
2. Handling Uncertainty with Latent Variables，使用隐变量处理不确定性
3. Training Energy-Based Models（EBM），训练能量模型 → 熵 → 梯度
4. Joint Embedding Predictive Architecture（JEPA），**联合嵌入预测架构**

JEPA： Joint Embedding Predictive Architecture

- JEPA 不是生成式，不能用从 x 预测 y；仅捕获 x 和 y 间依赖关系，而不显式生成 y 的预测：
  - 通过编码器不变性实现多模态
  - 通过潜在变量预测器实现多模态



I-JEPA

2

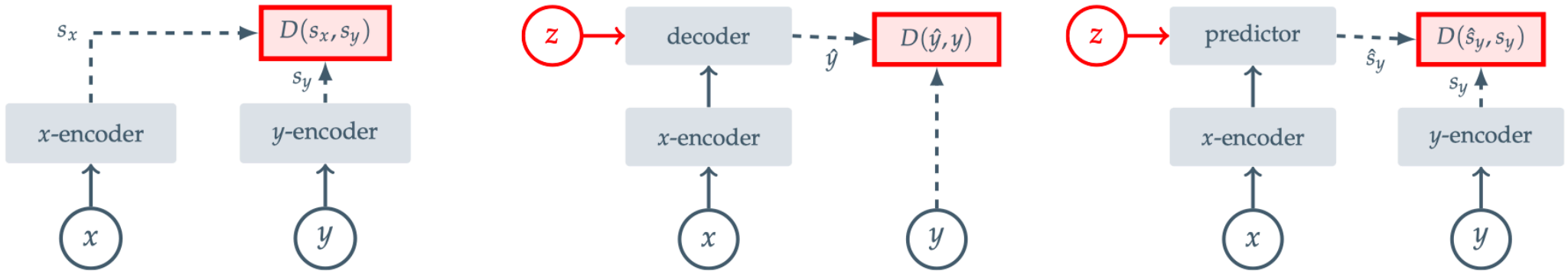
- https://arxiv.org/abs/2301.08243 Self-Supervised Learning from Images with a Joint-Embedding Predictive Architecture.

Research

I-JEPA: The first AI model based on Yann LeCun's vision for more human-like AI

June 13, 2023 • 7 minute read

I-JEPA 才是世界模型的未来



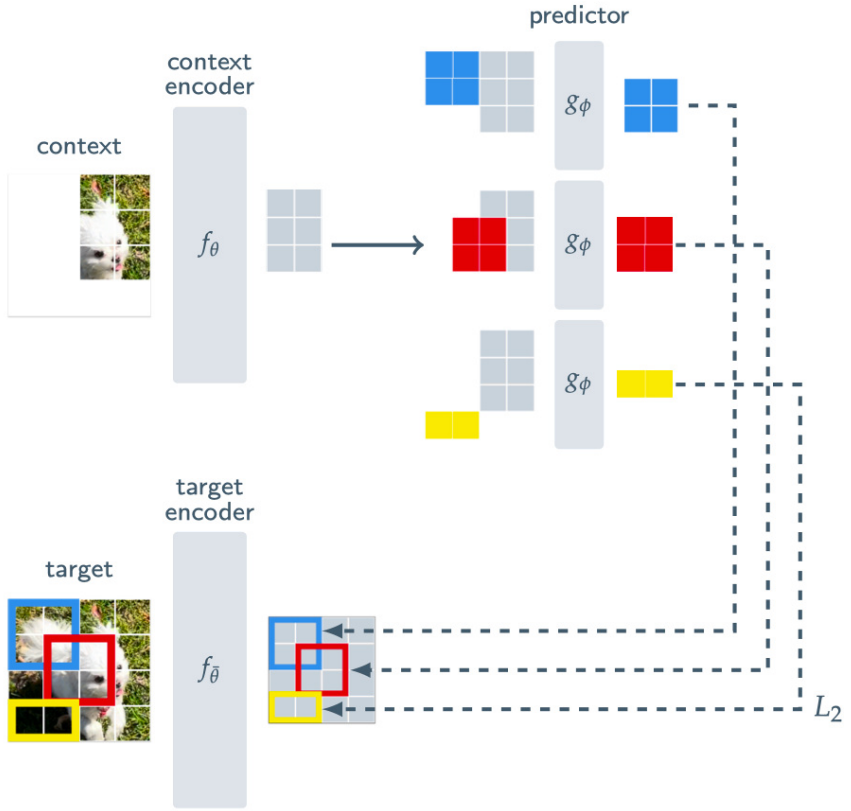
(a) **Joint-Embedding Architecture** (b) **Generative Architecture** (c) **Joint-Embedding Predictive Architecture**

Figure 2. Common architectures for self-supervised learning, in which the system learns to capture the relationships between its inputs. The objective is to assign a high energy (large scaler value) to incompatible inputs, and to assign a low energy (low scaler value) to compatible inputs. **(a)** Joint-Embedding Architectures learn to output similar embeddings for compatible inputs x, y and dissimilar embeddings for incompatible inputs. **(b)** Generative Architectures learn to directly reconstruct a signal y from a compatible signal x, using a decoder network that is conditioned on additional (possibly latent) variables z to facilitate reconstruction. **(c)** Joint-Embedding Predictive Architectures learn to predict the embeddings of a signal y from a compatible signal x, using a predictor network that is conditioned on additional (possibly latent) variables z to facilitate prediction.

I-JEPA 才是世界模型的未来

2

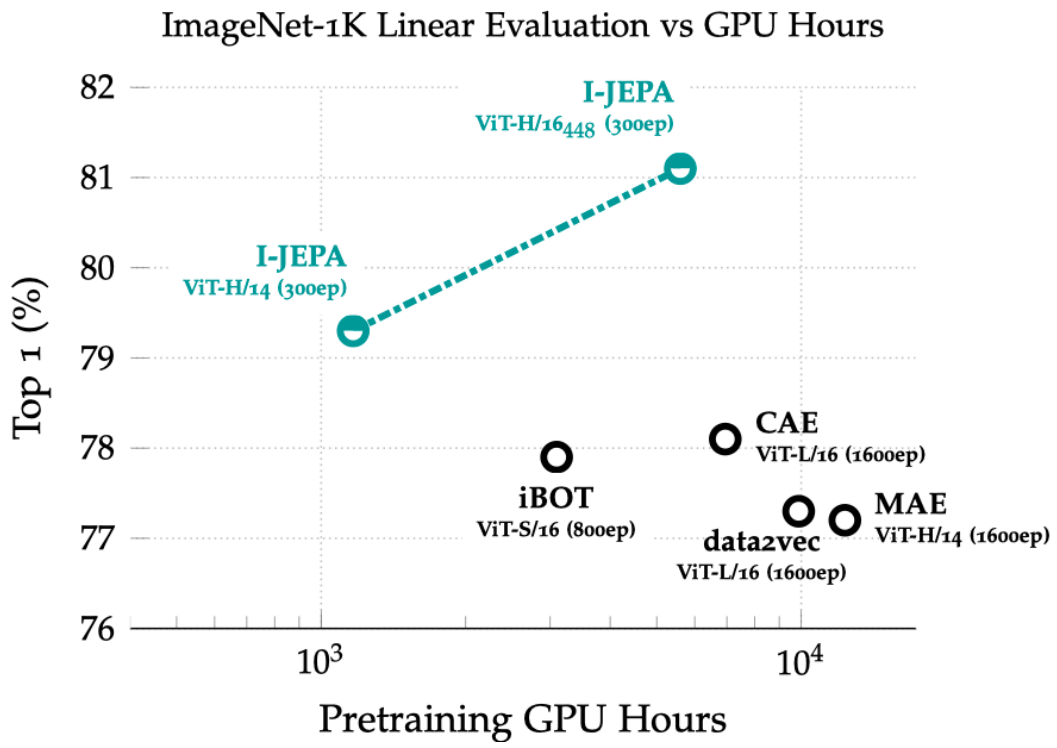
- 创新：1）非直接生成像素而是生成语义表征；2）采用 MAE 更新目标编码器参数方式。
- 特点：通过 AI 学会图像的语义表达，并论证可以用于于各种下游任务。

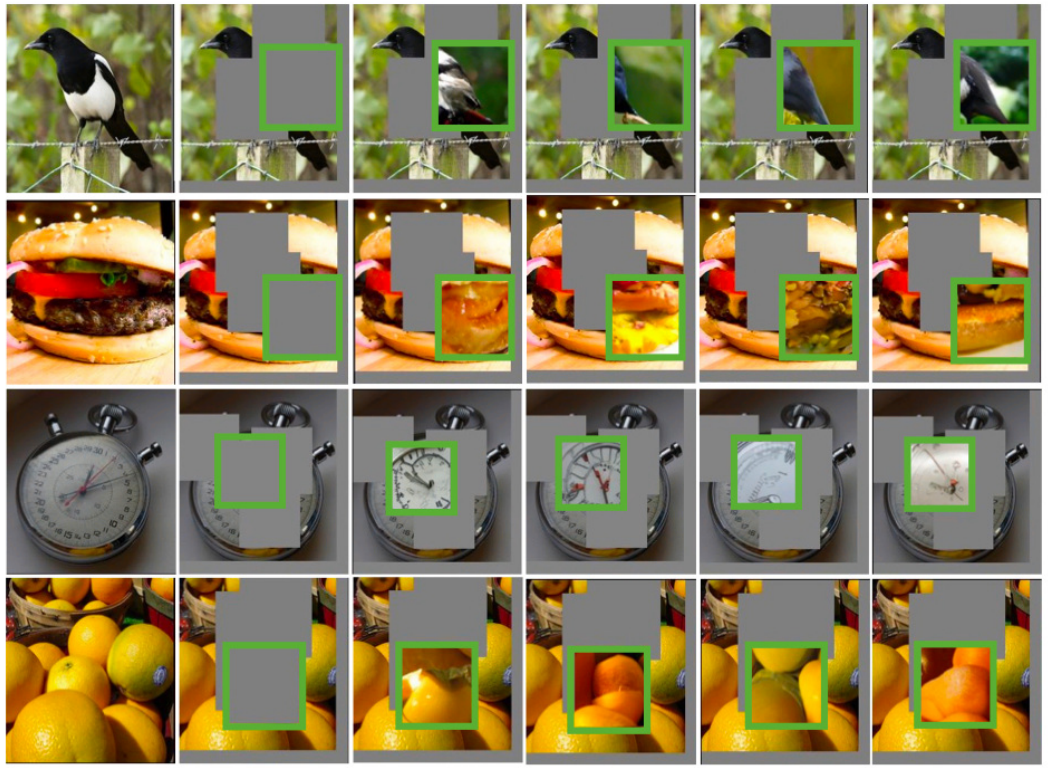


I-JEPA 才是世界模型的未来

2

- 推理：使用 Predictor 来进行表征，然后利用 class Token 输出进行 avg pooling 得到分类概率。
- 可视化：使用扩散模型 Diffusion Model 对表征向量逆高斯过程恢复原图看预测的概率。





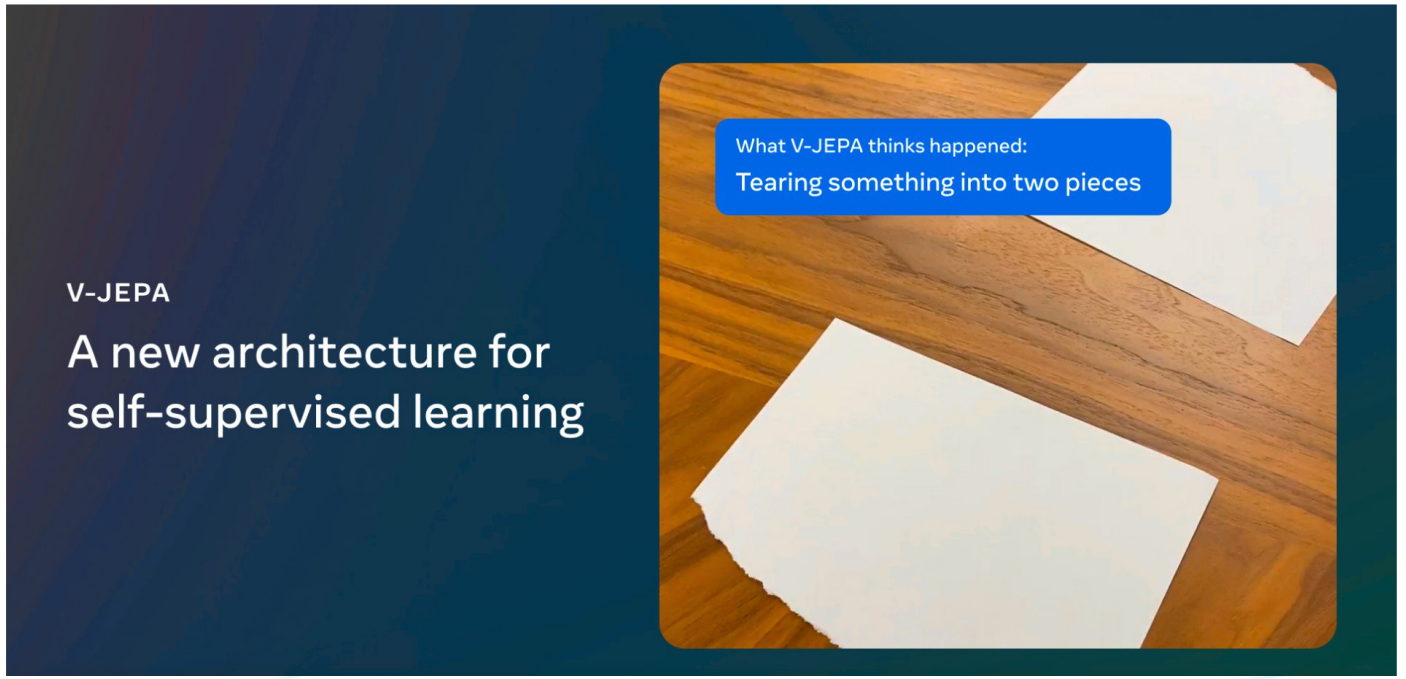
V-JEPA 才是世界模型的未来

3

- 使用 V-JEPA 主要关注于「感知」——通过分析视频来理解周围世界的即时情况。
- 在 V-JEPA 架构中，预测器充当了一个初步的「物理世界模型」，能够概括性地告诉我们视频中正在发生的事情。

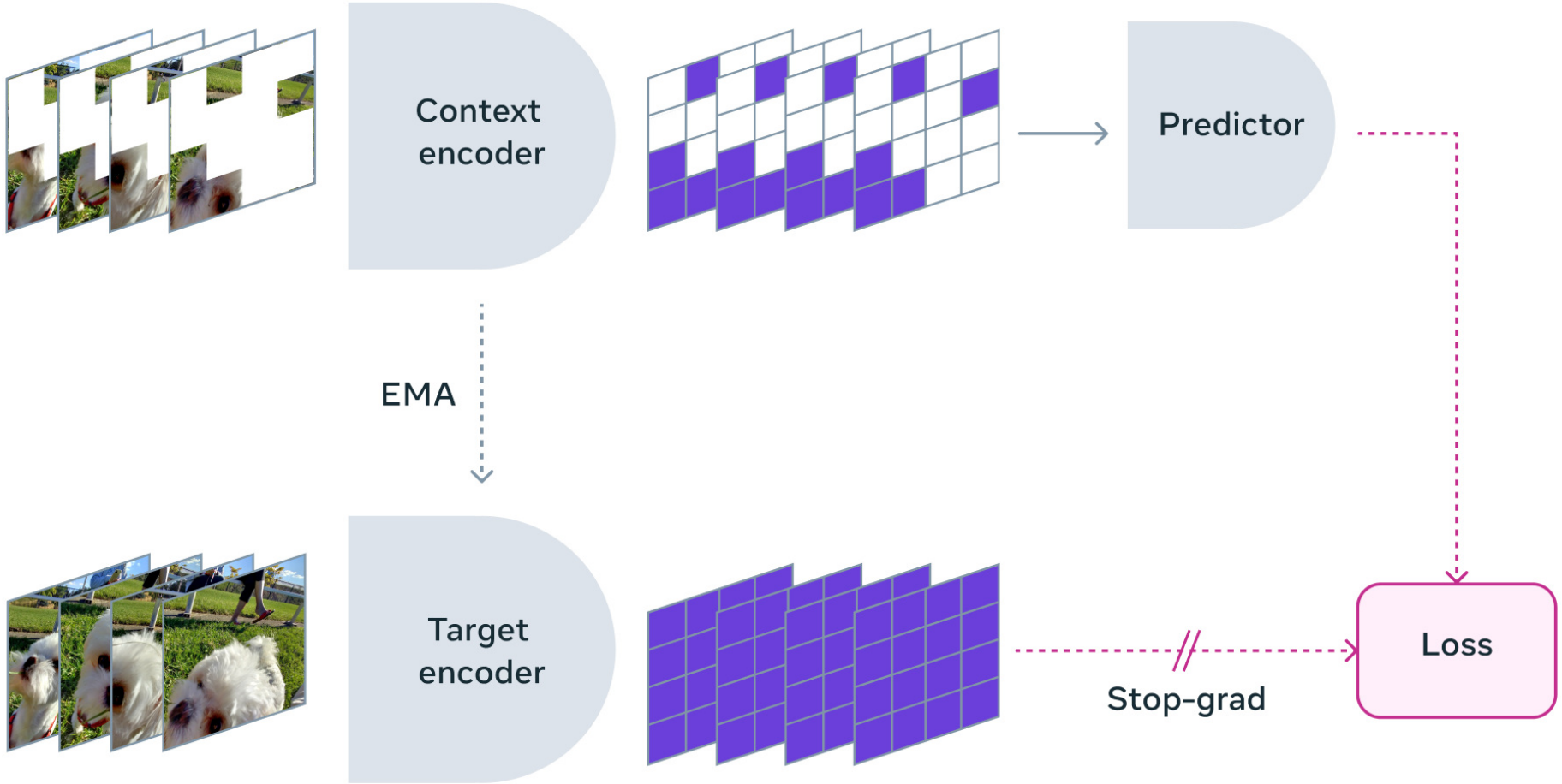
V-JEPA: The next step toward Yann LeCun's vision of advanced machine intelligence (AMI)

February 15, 2024 • 3 minute read



V-JEPA 才是世界模型的未来

3



JEPA 目标

- 基于强化学习 RL 思路创建世界模型，模型使用深度学习 + 无监督方法来创建学习流程：
  - **更好的 Latent Variable**：相比于重建像素的变分自编码器 VAE、掩码自编码器 MASK-VAE、去噪自编码器VQ-VAE，更能产生优秀的视觉输入表达；
  - **数据驱动到无监督**：从数据中来从数据中去，通过 JEPA 架构可以对数据表征不需要抽象或语义化，因为 JAPE 能够通过 SSL 在任何情况下找到预测表征的方法；

总结与思考

基本对比

	Sora	Genies	JEPA
训练方式	有监督学习	无监督学习	有监督学习
聚焦内容	生成式：将文本指令转换为高质量视频内容，侧重于文本到视频的生成，而不涉及用户交互	交互式：创建可交互、可玩的虚拟环境，强调用户通过潜在动作与生成环境的互动，生成动态可控的虚拟世界。	模型学习：强调模型学习隐藏信息能力，隐空间更加强大能够具备丰富的表征能力。
输出内容	生成长时间 60s 的视频	提供动态、交互式的虚拟体验	输出图像/视频的预测/分类概率

关于 SORA 的一些思考？

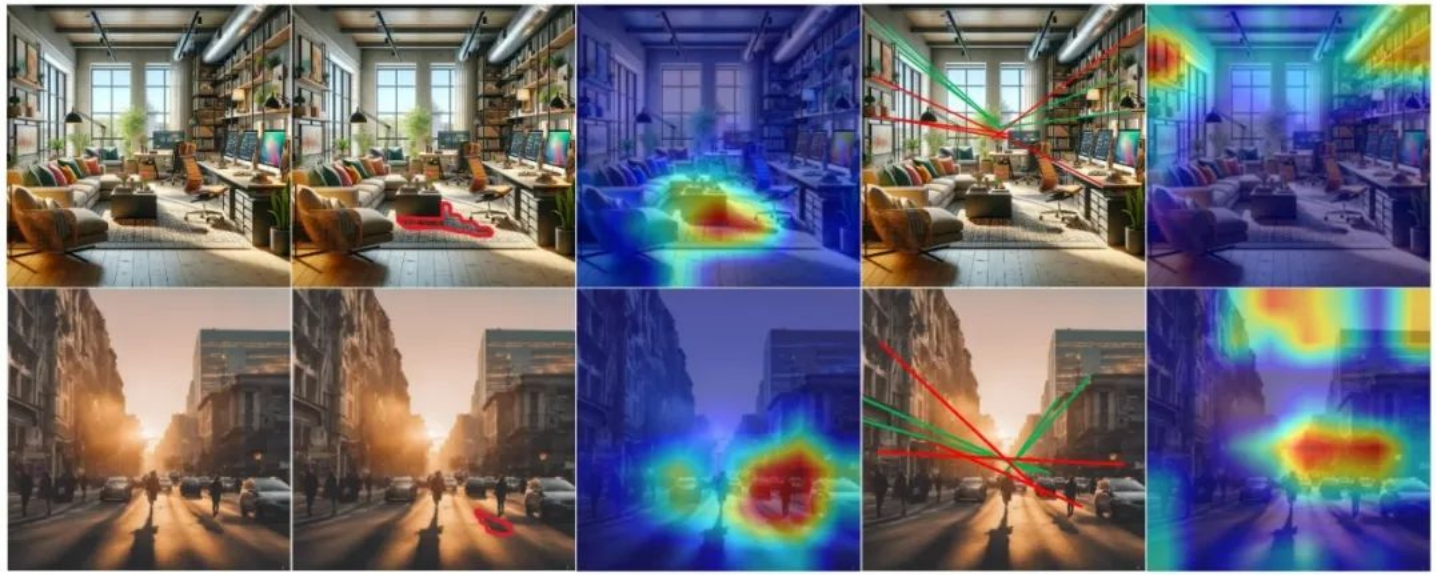
- SORA 训练的时候使用了虚幻引擎，内嵌了物理引擎？
- 作为一个 DiT 网络模型，它不会在生成视频时调用虚幻引擎。不会有人真的认为 Sora 在推理时有一个物理引擎在 Diffusion 循环中？
- Sora 是否真的从视频中归纳出和学习出物理定律？
- ZOMI 认为很荒谬，游戏引擎也很少模拟物理定律。虽然它们可能会模拟热效应（火灾、爆炸）和做功（物体克服摩擦力移动），但这些模拟经过抽象，不严格遵守热力学方程。物理引擎重点是渲染场景的视觉和交互可信度，而不是严格物理准确性（直观物理学）。

Intuitive physics 直观物理学

- 心理学中的直观物理学：对物理场景进行心理模拟，与表示物理世界的各个方面（例如几何形状）之间存在区别。类似于计算机游戏中物理引擎，其基于不完全准确的物理原理，通过模拟来预测物理现象。
- 心里表征：当观察物理场景时，会根据质量、摩擦、弹性等感知证据构建对物体、属性和作用力的心理表征，然后运行内部模拟来预测接下来会发生的可能物理反应。

AI生成图像存在物理不一致性

- **不意味 DiT 能完美表示 CV 场景三维几何。**人眼可以注意到输出中缺陷、物理不一致性。
  - e.g.，不一致性包括物体及其阴影错位、违反投影几何学；
  - e.g.，线条未能正确地收敛到消失点、不遵循线性透几何；



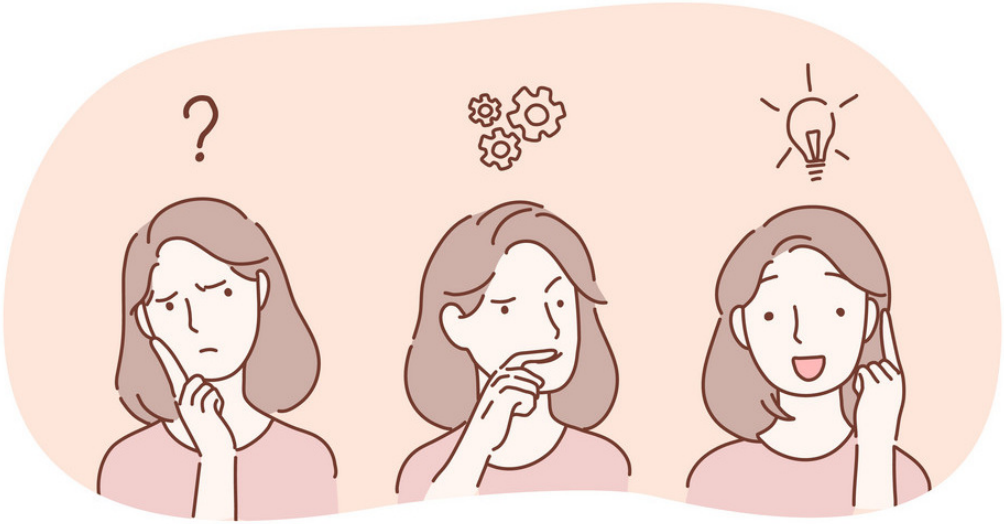
世界模型 World Models

- 定义：世界模型指的是智能体对其交互的外部环境的内部表示 → 给定环境状态和智能体行动，世界模型可以预测智能体采取该行动后环境的未来状态。

- 起源：在 ML 研究中，起源于 20 世纪 90 年代 Juergen Schmidhuber 实验室的强化学习研究。
- 淡化：进入21世纪世界模型（World models）含义被淡化，因为在实践中相当难以捉摸和建模。
- 重新定义：2018 年 Ha 和 Schmidhuber 发表 RNN 建模世界模型论文，提出世界模型的组成。
- 最近火爆：2024.02 起 SORA、GENNIE、JEPA都说自己是世界模型，能够模拟世界。

世界模型 World Models？

- 这引发了一系列问题：
  1. 基于潜在扩散的图像生成模型实际上编码了哪些信息？
  2. 仅编码图像表面启发式信息，还是视觉场景潜在变量，比如3D几何结构？
  3. 编码了真实世界中的因果推理关系？
  4. 编码了物理守恒定律、牛顿定律，还是直观物理学定律？



世界模型 World Models !

- 无论是生成模型、强化模型，都没有达到因果推理文献 && 物理定律中“世界模型” 所设定标准。
- AI 模型的研究表明从数据中学习物理属性的信息，可从模型隐藏信息中通过激活值解码出来 → 不意味模型具有因果效力或者物理效应。
- 潜在空间编码了结构保持、因果有效信息，此信息超越像素空间表面统计数据。这是 Sora 和模拟假猜测的重要线索，在论证 Transformer 结构能力的上限。

英伟达CEO黄仁勋：AGI或将在5年内出现

- 2024年3月初，NVIDIA英伟达CEO黄仁勋重返其母校斯坦福大学，参加了学院组织的SIEPR经济峰会以及View From The Top 系列讲座。在活动中，黄仁勋谈及了通用人工智能AGI的发展，他预测在未来5年时间，AI将实现人类水平的通用人工智能。

Sora 与 Genie/JEPA 的区别

- **与直观物理引擎模型不同**，Sora没有专门的感知、预测和决策模块，需要像物理引擎这样的接口；它只是一个高维空间，其中潜在表示经历跨层的连续变换。
- **Sora也与Ha和Schmidhuber的世界模型大不相同。**它不基于离散动作、观察和奖励信号的历史来运行模拟，没有传统意义上 Policy 算法（ Agent、 Action、 Reward ）。
- Sora模仿了一个智能体的 Policy，就像离线强化学习一样。**与 Genie 不同，Sora 没有接受过从视频中引导潜在动作的训练，并且其输出也不以此类动作为条件。**

ZOMI 的猜测

- 关于世界模型：
  1. Sora 不是世界模型：不是通过对场景进行模拟来预测视频帧。
  2. 作为世界模型其中一环： Sora 等视频生成模型可以在类似于 RL 系统中用作模拟器。
  3. 世界模型应满足：1）可对未来时间环境因素进行模拟。2）可学习输入（包括三维环境、物理属性等）结构信息、因果效应表征。

- 关于产业发展：
  1. 强化学习类的世界模型24年仍然难以成为算力消耗大头，技术上和理论上有待突破；
  2. 国内百模大战的创业公司和头部大厂不会跟进世界模型，因为技术不成熟，不能呈现应用落地；
  3. 机器人的应用落地短期内（24/25）仍然以大模型 + 传统伺服控制为主，世界模型的智能 Agent 仍停留实验室；

Thank you

把AI系统带入每个开发者、每个家庭、每个组织，构建万物互联的智能世界

Bring AI System to every person, home and organization for a fully connected, intelligent world.

Copyright © 2023 XXX Technologies Co., Ltd.
All Rights Reserved.

The information in this document may contain predictive statements including, without limitation, statements regarding the future financial and operating results, future product portfolio, new technology, etc. There are a number of factors that could cause actual results and developments to differ materially from those expressed or implied in the predictive statements. Therefore, such information is provided for reference purpose only and constitutes neither an offer nor an acceptance. XXX may change the information at any time without notice.

ZOMI

Course chenzomi12.github.io

GitHub github.com/chenzomi12/DeepLearningSystem