World Model

World Model

让 AI 学会像人一样理解物理世界

从认知科学到大规模视频预训练，
世界模型正在成为通用人工智能的关键拼图。

2026.04 | 科普读物

目 录

一、什么是世界模型?

"世界模型"（World Model）是指一个能够模拟和预测外部环境如何随时间演化的内部表示。简单来说，它回答的问题是：

"如果我做了这个动作，接下来会发生什么？"

这个概念最早由认知科学家 Kenneth Craik 在 1943 年的著作The Nature of Explanation 中提出。他认为人类之所以能高效地思考和决策，是因为大脑内部维护着一个对真实世界的简化模拟——一个可以用来推演后果的心理模型。

1.1 一个直觉性的例子

想象你在打网球。当球飞来时，你的大脑并不需要每次都重新计算物理规律，而是通过一个内部的"世界模型"快速预测球的轨迹、落点，然后指挥身体做出反应。这个模型不是显式的公式（你不会在接球时计算抛物线方程），而是一个经过大量经验训练形成的隐式预测机制。

同样的逻辑也适用于更复杂的场景：你知道如果把玻璃杯掉在地上它会碎，知道火是烫的，知道太阳明天会升起——这些知识都编码在你的世界模型中。

1.2 AI 领域中的世界模型

在人工智能领域，世界模型指的是一种学习环境动态规律的神经网络模型。它接收当前状态和动作作为输入，输出对未来状态的预测。这种能力使智能体能够在不实际执行动作的情况下，在脑海中"想象"不同行动路径的结果——这正是规划和推理的基础。

二、核心思想：预测即理解

世界模型的核心理念可以用一句话概括：

"要理解一个系统，就要能预测它的行为；要能预测它，就必须建立其内在模型。"

2.1 大脑中的世界模型

神经科学研究表明，大脑皮层特别是前额叶皮层（PFC）和海马体，持续地对外部世界的状态进行预测。当预测与现实不符时，会产生"预测误差"（prediction error），这个误差信号驱动学习和注意力的调整。

- 感知即推断：我们看到的不只是光子撞击视网膜，而是大脑基于先验模型对场景的最佳解释。
- 运动控制即预测：当我们伸手拿杯子时，大脑预先模拟了整个运动过程及其感觉反馈。
- 记忆即重放：海马体在睡眠中不断回放白天的经历，用于更新世界模型的参数。
- 想象即模拟：当我们思考"如果...会怎样"时，实际上是在用世界模型进行反事实推理。

2.2 从认知科学到 AI 的跨越

2017 年，Google DeepMind 的研究团队发表了一篇里程碑式的论文Imagination-Augmented Agents for Deep Reinforcement Learning (I2A)。他们证明了在强化学习中引入"想象力"（即使用世界模型进行前瞻搜索），可以让 AI 智能体在复杂环境中取得远超传统方法的性能。这标志着世界模型从认知科学概念正式进入 AI 主流研究。

三、技术演进路线

世界模型在 AI 领域的发展可以分为几个关键阶段：

3.1 第一阶段：模型基础强化学习（2016-2018）

代表工作：Ha & Schmidhuber (2018) "World Models"

这篇论文首次完整展示了世界模型框架。作者将智能体的架构分为三个模块：(1) 变分自编码器（VAE）压缩视觉观测；(2) RNN 混合密度网络（MDN-RNN）学习环境动态；(3) 控制器（Controller）利用世界模型做规划。在 CarRacing 环境中，仅用少量随机样本训练的世界模型就能让智能体学会驾驶。

3.2 第二阶段：对比学习与世界建模（2019-2021）

代表工作：CPC、BYOL、SimCLR 等自监督方法

这一阶段的核心发现是：通过大规模无监督学习（尤其是视频预测任务），模型可以自发地涌现出对物理世界的理解。DeepMind 的 CURL 和 SVEA 工作表明，好的视觉表征本身就是世界模型的一种形式。

3.3 第三阶段：大规模生成式世界模型（2022-至今）

代表工作：NVIDIA Video LDM, OpenAI Sora, Google Genie

这是当前最激动人心的方向。以视频为训练数据的大规模扩散模型和Transformer，展现出了惊人的世界模拟能力：

- Sora (OpenAI, 2024)：直接从文本/图像生成长达一分钟的高质量视频，隐式学习了物理规律、物体持久性、因果关系等。
- Genie (Google, 2024)：从无标注互联网视频中学习可交互的 2D 游戏世界，无需任何动作标注即可创建可玩的虚拟环境。
- Video LDM (NVIDIA, 2022)：基于 latent diffusion 的高分辨率视频生成，为后续视频世界模型奠定基础。
- LWM (Large World Model, 2024)：使用因果掩码 Transformer 直接在原始视频 token 上进行大规模训练。

四、关键架构与训练方法

4.1 三大主流架构范式

架构类型	核心思路	代表工作	优势	局限
VAE+MDN-RNN (■■■)	■VAE■■■■ ■RNN■■■■■■	Ha & Schmidhuber (2018)	■■■■■ ■■■■	■■■■■■■
Diffusion-based (■■■■)	■■■■■■ ■■■■■	Sora, VideoLDM (2022-24)	■■■■■■ ■■■■	■■■■■
Autoregressive (Transformer)	■token■■ ■■■■■■	Genie, LWM (2024)	■■■■ ■■■	■■■■■■

表 1：世界模型三大主流架构对比

4.2 核心训练目标

帧预测（Frame Prediction）

给定历史帧序列，预测下一帧或未来 N 帧。最直接的监督信号，但存在多模态不确定性。

掩码重建（Masked Modeling）

随机遮挡视频中的部分区域或时间段，要求模型还原被遮蔽的内容。类似 BERT 但扩展到时空维度。

对比学习（Contrastive Learning）

拉近同一轨迹上不同时间点的表征，推开不同轨迹的表征。CPC 是这一方向的先驱。

动作条件预测（Action-Conditioned）

在预测中加入动作信息，使模型学到环境的条件动态。这是实现规划和控制的桥梁。

五、应用场景

具身智能与机器人

机器人需要在物理世界中操作，但真实世界的试错成本很高。世界模型让机器人在虚拟环境中进行千万次想象实验，学习到精细的操作策略后再迁移到现实。NVIDIA 的 Isaac Sim 和 GR00T 项目正是沿着这条路径推进。

游戏AI与NPC

世界模型可以从游戏录屏中自动学习游戏规则和物理引擎，无需人工编程。Google Genie 展示了从视频中自动生成可交互 2D 游戏的能力。

自动驾驶

自动驾驶需要处理极端罕见的长尾场景。世界模型可以生成各种危险的交通情景用于仿真训练，大幅提升系统的安全边界。

科学发现

蛋白质折叠预测（AlphaFold）、药物分子设计、气象预报等领域本质上都是"给定当前状态，预测未来演化"的问题。

内容创作

Sora 等视频生成模型已经展示了世界模型在创意产业中的应用潜力。从电影预览、动画制作到虚拟主播，世界模型将成为下一代创作工具的核心引擎。

六、挑战与未来

6.1 当前面临的主要挑战

- 长程依赖与灾难性遗忘 视频中的因果链可能跨越数百甚至数千帧。当前 Transformer 架构虽然在长序列上比 RNN 更好，但在超长序列上的效率和精度仍有瓶颈。
- 多模态不确定性 同一个初始状态可能演化出多种合理的未来（混沌系统特性）。如何在不损失多样性前提下保持预测的有用性是开放问题。
- 因果性与反事实推理 当前大部分世界模型学到的是统计相关性而非真正的因果关系。要让 AI 像人类一样进行'如果当初...就会...'的反事实推理，还需要新突破。
- 计算成本 高分辨率视频世界模型的训练和推理成本极其昂贵。如何在保证质量的前提下降低资源消耗决定了这项技术能否真正普及。
- 评估难题 不像分类任务有准确率这样的清晰指标，世界模型的'好坏'很难量化定义。一个好的预测不一定是有用的预测。

6.2 未来展望

短期（1-2年）：视频生成质量持续提升，世界模型开始在游戏、仿真训练等封闭环境中产生实际价值。多模态融合（视频+语言+音频+3D）成为主流。

中期（3-5年）：具身智能领域出现突破，机器人通过世界模型获得的"常识"显著提升操作泛化能力。自动驾驶的长尾场景覆盖率大幅改善。

长期（5-10年）：世界模型可能成为通向
AGI（通用人工智能）的关键路径之一。一个具备精确世界模型的 AI，不仅能理解和预测物理世界，还能在其中进行有效的规划和创造——这或许就是"真正智能"的本质特征。

"The brain is a prediction machine. It does not just react to the world — it constantly simulates what will happen next."

-- 核心思想源自 Kenneth Craik (1943)，至今仍是世界模型的灵魂