# 人工智能最新方向--"科学家们在用生成式模型来做世界模型"

随锐人工智能工程研究所

2026 年 8 月

“科学家用生成式模型做世界模型”，这是最新的人工智能工程化的研究方向。

## 第一部分

把研究方向拆开看，这是两件事：世界模型是目标——让 AI 在“脑子里”装一个能模拟真实世界运转规律的内部沙盘；

生成式模型是工具——用扩散模型、自回归 Transformer 这类生成技术，从海量视频/图像/传感器数据里学出这个沙盘。

下面的工程化研究报告，逐步分析，全面诠释。

### 一、世界模型到底是什么？

很多人工智能领域的科研人员与工程人员，对此新事物都较陌生。

世界模型的核心定义：

世界模型是 AI 系统通过学习数据，在内部构建的一套能模拟、预测真实世界运行规律的数字化模型。

它的核心是——让 AI 不用真的去做，就能推演“我做某个动作之后，世界会发生什么”。

NVIDIA 也有定义，它的官方术语表定义得很精准：

世界模型被英伟达定义是“理解现实世界动态（包括物理和空间属性）的 AI 工具，使用文本、图像、视频、声音和运动等输入数据来预测接下来会发生什么”。

一个直觉类比（援引南洋理工大学的安波教授的 26 年 1 月的研究与解释）：

像 GPT 这样的语言模型，本质是“预测下一个词”——它像是一个读遍了人类所有书籍的人，对世界的知识是“二手”的、从文字里学来的。而世界模型想做的是另一件事：让 AI 在脑子里建立一个对世界如何运转的内部模拟器。就像人类看到一个杯子被推到

桌子边缘，不用算物理公式，也能预判它会掉下去摔碎。

科研界与产业界的共识：

- 语言模型：预测下一个词（侃侃而谈），目前成果卓著，可以商业化。

- 世界模型：预测下一个状态（预判后果），更难更复杂。

这也是为什么世界模型，被认为是通往 AGI 和具身智能的核心拼图——一个不理解物理世界的智能科技，很难说是完整的智能。

**二、为什么要用“生成式模型”，来做世界模型？**

传统世界模型怎么做？

科学家们与工程师们，手工编写物理规则，把它们塞进计算机。

这种做法在游戏引擎、早期机器人仿真器等受控环境里好用，但一旦现实世界出现意外情况，立刻失效。

生成式 AI 的出现，彻底改变了范式：

- 不再手动编写规则，而是用互联网规模的数据集训练模型。

- 给定提示后，这些模型能够生成高保真的合成世界。

- 新一代世界基础模型（WFM）基于海量真实世界数据和无限的合成数据进行预训练，不仅能够生成内容，还能根据物理规律进行推理和预测。

最新研究成果说：中美两国的计算机科学家与人工智能科学家联手的 AI 研究发现，用生成式模型（扩散模型、自回归 Transformer 等），去“生成”未来的场景/状态，本身就是一种对世界动态规律的学习和建模。

该模型为了生成逼真的下一帧视频，它“被迫”要去理解重力、碰撞、遮挡、物体恒存这些物理规律。

——这就是用生成式模型做世界模型的核心思想。

**三、后续主流研究成果需要工程化、商业化，路径具体怎么做？**

我们认为，有三条技术路径。

技术路径 1：视频生成路线（Sora 路线）

以 Sora 为代表的生成式路径，依赖海量数据隐式学习物理规律，采用扩散变换器（DiT）架构，通过时空补丁化处理实现视频生成。

优势：降低人工建模成本，短期内更易实现商业价值。

代表：Sora、Veo、可灵、Wan 等。

技术路径 2：4D 几何建模路线

传统视频生成大多建立在 2D 图像空间上，通过逐帧建模来合成动态内容，缺点是相机运动难以精确控制，多物体交互缺乏一致性，长时间生成容易出现结构漂移，甚至违背基本物理规律。

新一代研究转向 4D 几何世界建模，具体说——将视频表示为“3D 空间 + 时间”的统一世界状态，而不是简单的逐帧像素生成。

比如如下的具体例子：

- VerseCrafter：用静态背景点云描述场景结构，用带时间信息的 3D 高斯轨迹描述动态物体。

- NeoVerse：从普通的单目视频中恢复场景的 3D 结构，并进一步建模随时间变化的动态信息。

技术路径 3：具身智能/机器人的路线

这是世界模型最刚需的落地场景。

世界模型在上述智能机器人上的完整架构，又通常分三大模块：

1. 感知压缩模块：把图片、雷达、视频等高维画面，简化成抽象“信念状态”。

2. 动力学预测模块（核心）：学习世界运行规律，模拟物体运动、碰撞、遮挡、重力。

3. 决策规划模块：在虚拟沙盘里反复试错，选出最安全、最优的动作。

在训练范式上（根据上海 AI Lab 的实验室综述），世界模型的学习，是一条从经验输入到可靠行动的闭环。

有如下模式 1，

- 自监督与生成式预训练：从未标注视觉数据中，学习基础表征（视频预测、下一个 token 预测、扩散潜空间）。

- 具身交互：通过状态-动作-奖励数据，补充动作与结果之间的关系。

- 合成与仿真数据：扩展长尾、危险和反事实场景。

- 物理先验：通过 PINNs、ODEs、守恒规律增强物理一致性。

- MBRL 循环：在潜空间中进行后台滚动，并结合 MPC、CEM 或 MCTS 完成规划与搜索。

**四、这个研究后续用来干什么：核心应用场景。**

场景有很多，重点列举世界模型的作用

| 场景 | 世界模型的作用 |
| --- | --- |
| 自动驾驶 | 车载世界模型实时模拟周边车辆、行人、雨雪路况，在脑子里先把可能的场景都"演"一遍，预判车辆变道、急刹，提前规避事故 |
| 人形机器人 | 在虚拟仿真环境反复练习抓取、走路、搬东西，不用在现实中反复磕碰损坏设备 |
| 科学仿真 | 气象、流体、材料模拟，快速推演复杂物理变化 |
| 数字内容 | 可交互 3D 虚拟场景、可自由推演剧情的虚拟世界 |
| 智能体 Agent | AI 自主规划长期任务，提前预判每一步行为带来的环境反馈 |

综上，在产业侧已经有实质落地：

Gartner 将物理 AI 列为 2026 年十大战略技术趋势之一；其白皮书显示超过 80%的自动驾驶算法已经用世界模型做辅助训练，能降低近 50%成本、提升约 70% 效率。

华为乾崑 ADS 4.0 用了 WEWA 架构（云端世界引擎+车端世界行为架构）；

蔚来 NWM 世界模型能在 100 毫秒内模拟 216 种潜在场景。

**五、关键澄清几个研究现状。**

全球各行业对世界模型尚无统一定义

这是当前最重要的认知——“世界模型”这个筐里，什么都能往里装，但各家做的其实不一样：

腾讯研究院在最新技术报告提出：“面向具身智能的世界模型，关键不是生成一段视觉上连贯的视频，而是理解机器人采取某个动作后，物体和环境将发生什么变化。如果模型只根据提示词生成画面，没有与机器人的感知、决策和行动形成闭环，就不能等同于机器人所需要的世界模型。”

这意味着，目前至少有三种“世界模型”在并行：

1. 视频生成派：Sora 类，生成视觉连贯的未来帧（但可能只是“拟合像素分布”，未必真懂物理）。

2. 3D/4D 几何派：WorldLabs、VerseCrafter 类，显式建模空间和物理。

3. 具身决策派：与机器人感知-决策-行动形成闭环，预测动作后果并用于规划。

它们之间并非截然对立，但成熟度、目标各有侧重。

正如 CVPR 2026 综述指出的根本性问题：“它们究竟是在‘理解世界’，还是仅仅在‘拟合像素分布’？”——这是当前世界模型研究的核心争议。

**六、这个领域当前的主要挑战**

1. 物理一致性：现有模型大多在 2D 图像空间中建模，导致相

机运动和多物体运动难以统一控制、且生成结果容易不稳定，复杂场景中容易违背基本物理规律。

2. 评估标准缺失：如何科学、客观地评估一个世界模型的优劣及其推理能力，尚未形成统一标准。

3. 数据质量与偏见：模型学习依赖的数据集本身可能包含噪声、偏差或片面性，导致构建的世界模型存在缺陷。

4. 无限复杂性：真实世界的复杂性与开放性是无限的，模型难以穷尽所有可能性，永远存在未知边界。

5. 计算成本：训练和运行复杂的世界模型需要巨大的计算资源。

**概述总结**

科学家们用生成式模型做世界模型，本质是用生成技术（扩散、自回归 Transformer），从海量真实世界数据中，“学出”一个内部模拟器。

——AI 为了生成逼真的下一帧/下一个状态，被迫理解物理规律、空间关系、因果关系。

这条路已经能让 AI 从“预测下一个词”进化到“预测下一个状态”，是通往具身智能和 AGI 的关键拼图。

但它还不是完美的物理引擎，行业对世界模型的精确定义、评估标准、技术路线都还在探索中。

## 第二部分

上述这几种技术路线，比如 Sora 的视频生成路线 vs 机器人具身路线，或具体应用场景——自动驾驶/人形机器人，是目前产业化重点发展方向。

**Sora 代表的视频生成路线，**和**机器人/自动驾驶代表的具身路线**，表面上看都是在预测未来，

但两者在**架构目标、输入输出、评价标准**上有本质差异。

——前者追求看起来合理，后者追求想得对、动得准。这一层

差异，决定了它们各自的适用边界。

## 一、两条路线的本质分歧

| 维度 | 视频生成路线（Sora 派） | 具身路线（机器人/智驾派） |
| --- | --- | --- |
| 核心问题 | "下一帧画面长什么样？" | "我做这个动作后，世界会怎样？" |
| 预测目标 | 像素空间（高维、冗余） | 潜空间状态 / 动作后果（紧凑、可控） |
| 架构代表 | Sora、Veo、可灵、Wan | V-JEPA 2、ω-EVA、Kairos、Cosmos |
| 是否要"懂物理" | 隐式学习，统计近似 | 必须建模物理因果，否则行动失败 |
| 与决策的关系 | 解耦（外部智能体另行决策） | 耦合（预测直接驱动动作生成） |
| 部署位置 | 云端（算力贵） | 端侧（机器人/车端实时） |

再进一步来做对比研究看：

**视频生成路线是世界幻觉生成器（LeCun 观点），具身路线是行动后果预判器。**

前者让 AI 学会看上去对，后者让 AI 学会真的对。

## 二、Sora 路线：用视觉合理性逼近世界

Sora 的技术配方已经成为视频生成派的事实标准：**DiT（Diffusion Transformer）+ 时空潜空间 Patches + 大规模视频数据 + 长上下文窗口。**

它的关键创新是把视频看作时空统一体而非独立帧序列，从而

在潜空间学习运动轨迹、光影变化与场景结构。

**这个路径做到了什么：**

60 秒长视频生成，物体多角度一致性、遮挡重现等长程依赖处理优秀。

基本物理现象（重力、流体、简单因果如踢球→滚动）能合理呈现。

文本条件引导使得模型能按指令模拟世界。

**它的边界在哪**（这部分，多个研究里有多方印证）：

**复杂物理交互穿帮：**玻璃碎裂的碎片不守恒、物体突然消失/变形、手部拓扑错误、镜头穿模。

**长时推演漂移：**百帧以上推演必然失真，因为没有状态变量，每一步都是在重新生成

**算力墙：**视频生成是所有模态中最昂贵的，Sora 运营成本月消耗约 1500 万美元量级，难以端侧部署

**不是可靠物理引擎：**学到的不是物理定律，而是训练数据中出现过的物理模式的统计近似。

然而，LeCun 的研究反馈很尖锐：**Sora 不是世界模型，是世界幻觉生成器**——它生成的是看起来合理的幻觉，不是真正符合物理的世界模拟。

在 Physion-Eval 等物理基准测试中，Sora、Veo 等顶级模型在复杂物理场景中出现了大量的肉眼可见的物理错误（如穿模、违背质量守恒等）。

**所以 Sora 路线的适用边界很清晰：**影视预演、虚拟内容生成、沉浸式娱乐、作为机器人训练的**离线数据合成器**——但**不能直接驱动机器人或自动驾驶做决策**。

## 三、具身路线：把动作作为核心条件

具身路线的关键转折是——**不再让模型生成视频，而是让模型预判动作后果**。这听起来差别不大，但架构上完全是两回事。

1. 表征预测派：JEPA 路线

由 LeCun 力推，**核心理念是不要重建像素，预测抽象表征**。

代表是美国 Meta 公司 2025 年 8 月发布的 V-JEPA 2：

Vision Encoder 把视频帧编码为潜在表征。

Context Encoder 编码上下文窗口。

JEPA Predictor 在潜在空间预测被 mask 的未来块。

**关键优势**：以 1/50 的算力追平了生成式世界模型的性能。

因为它不在高维像素空间里精确建模每个纹理、光照细节，而是直接在抽象表征空间推理——大量计算浪费在纹理、光照等与决策无关的信息上，JEPA 路线成功避免了这一点。

2. 隐空间动作条件派：最新产业主流

2026 年这一派，成为具身智能的真实落地路线。

代表案例：

**ω-EVA（星源智，2026 年 6 月发布）：**

参数量仅 1.2B，**全程在特征空间推理，不生成像素级视频**

把**动作（手腕位姿、末端执行器状态）作为核心条件输入**——动作是唯一的、可控的；而语言描述是主观多样的，容易导致决策不确定性。

首创**预演 → 验证 → 行动**的决策闭环：机器人在执行指令前，先在模型内部想象动作带来的环境变化，并根据推演结果优化动作方案。

算力消耗远低于视频生成路线，**可实现端侧部署。**

**Kairos 3.1（大晓机器人，2026 年 7 月发布）：**

融合生成智能、物理智能、认知智能的行动一体化世界模型。

原生统一架构打通理解-推演-执行-反思全链路。

提出**信息密度定律**：数据的价值不在数量堆叠，而在信息密度——会改变行动结果的信息，就是有价值的信息。

已在生成预测权威榜单刷新 SOTA，并构建起从真实数据、统一

模型到端侧推理、本体控制、自主反思的完整技术体系。

3. 三维几何派：显式建模空间

传统视频生成大多建立在 2D 图像空间上，导致多物体交互缺乏一致性。

新一代研究转向 **4D 几何世界建模：**

通过将环境显式或隐式编码为结构化的 3D/4D 几何表示。

使世界模型在预测未来状态时不仅依赖像素外观变化，还能直接推理物体间的空间关系约束、物理连续性和多视角一致性。

实现更准确、更具物理可信度的长期时空演化预测。

Ctrl-World（星动纪元联合斯坦福大学）是代表方向之一：在训练中嵌入物理引擎约束，使模型学习并遵守牛顿力学定律；融合多视图联合与视频预测，隐式建模深度图与三维空间结构。

**四、两条路线在自动驾驶/人形机器人上的落地差异。**

自动驾驶场景

自动驾驶是世界模型落地最成熟的场景之一。Gartner 将物理 AI 列为 2026 年十大战略技术趋势，超过 80% 的自动驾驶算法已经用世界模型做辅助训练。

具体做法：

**1. 作为边缘案例数据生成器**

真实路况无穷无尽，极端工况（暴雨路面、路口鬼探头、设备故障）实地测试成本高、风险大。世界模型批量生成贴合动力学规则的仿真场景：

英伟达 Cosmos 3（2026 年 6 月发布）：646 亿参数 Super 版主打高精度训练，157 亿参数 Nano 版轻量化适合端侧部署；官方测试数据显示，原本需要数月的算法训练周期，依托仿真合成数据能压缩到短短几天。

小鹏第二代 VLA：首次去掉语言转译环节，达成从视觉信号到动作指令的端到端直接生成，720 亿参数基座模型每五天可完成一次全链路迭代。

**2. 作为实时预测器**

实时预测行人和车辆的未来轨迹

预判路口交通变化，提前规划避障路径。

在虚拟空间中模拟极端工况（如突发横穿），显著提升行车安全性。

**3. 作为仿真器**

DriveDreamer 系列已成为行业代表性基础设施——以世界模型为核心的新一代驾驶模拟器，已与多家国内头部主机厂、海外及合资主机厂，以及 AI 芯片、Tier 1 巨头达成签约，定点与量产合作，服务海内外头部主机厂与自动驾驶公司超 30 家。

人形机器人场景

人形机器人是世界模型虚拟练兵价值最突出的场景：

**特斯拉 Optimus、优必选 Walker X** 等人形机器人保持平衡、灵活抓取物体，依靠世界模型在虚拟空间中反复推演步态和抓取动作，再迁移至真实环境，从而大幅降低摔倒和操作失误的概率。

**产业落地案例：**

**宇树 H2 人形机器人**（计划 2026 年底量产）：已确定适配英伟达 Cosmos 开源模型栈，依靠海量仿真调试，大幅减少真机试错带来的时间和物料消耗。

**极佳视界 × 一汽模具**（2026 年 4 月）：

世界模型 GigaWorld + 具身基础模型 GigaBrain + 本体 Maker H01 进入真实工厂，围绕箱体拆垛、跨区域搬运、动态避障、精准操作等高频任务，**将传统自动化方案数月的场景适配周期压缩至数周。**

**极佳视界 × 隆盛科技**（2026 年 6 月规划）：三年内无锡部署 1000 台搭载世界模型具身大脑的通用机器人——全球首次由物理智能基础模型驱动的通用机器人在工业场景开启千台级规模化落地。

**五、两条路线正在融合。**

2026 年的关键技术趋势是——**两条路线不再是二选一，而是走**

向融合：

π0.7（2026 年 4 月）首次将世界模型作为 Subgoal Image Provider 集成到 VLA 中。

DeepMind Gemini Robotics ER 1.6 引入推理优先的世界模型增强空间理解。

3 个月内发表了 21 篇 WM-VLA 融合论文（WoVR、VLAW、Chain of World、LaST-VLA 等）。

NVIDIA SANA-WM（2026 年 5 月）：2.6B 参数开源世界模型，纯世界模型——专注于在潜在空间中预测状态演变，而不关心像素级的重建。标志着这一领域从闭源走向开源社区。

融合的逻辑很清晰：

**视频生成路线提供视觉先验和高保真数据合成能力，具身路线提供动作条件和决策闭环**——两者互补，前者解决世界长什么样的问题，后者解决我该怎么动的问题。

## 六、关键总结：为什么具身路线更难

腾讯研究院的判据很激进：**面向具身智能的世界模型，关键不是生成一段视觉上连贯的视频，而是理解机器人采取某个动作后，物体和环境将发生什么变化。**

**如果模型只根据提示词生成画面，没有与机器人的感知、决策和行动形成闭环，就不能等同于机器人所需要的世界模型**。

这就是为什么 Sora 路线即便做到 Sora 2 的物理一致性提升，仍然无法直接用于机器人——**它缺的不是更好的画质，而是动作→状态的因果闭环**。

机器人需要的是：给我当前状态 + 一个动作，我告诉你下一刻会发生什么。这个下一刻不需要是像素级视频，但需要是**能驱动控制器执行的状态表示**。

这也是为什么 ω-EVA 坚持 1.2B 参数、隐空间推理、端侧部署——它不是做不出视频，而是**视频生成对机器人来说是无用功**，算力应该花在预判动作后果上，而不是渲染像素上。

**概述总结：**

Sora 视频生成路线和机器人具身路线，**不是同一个问题的两种解法，而是两个不同问题的最优解。**

——前者回答世界看起来怎样，后者回答我动了之后世界会怎样。

2018—2022 年大家以为视频生成就是世界模型；2025—2026 年行业共识已经反转：**真正的具身世界模型必须走动作条件 + 潜空间预测 + 决策闭环的路线**，视频生成最多作为数据合成器和视觉先验的辅助角色。

未来 3 年，人工智能在世界模型里产业化的重点方向，有如下一些：

1，V-JEPA 2 如何在潜在空间做预测的具体机制。

2，ω-EVA 的端侧部署是怎么做到。

3，自动驾驶世界模型在仿真到实车的迁移中，怎么解决 domain gap 等。

上述在世界模型架构里遇到的诸多问题，需要科学界与产业界紧密融合，后续去进一步去探索与实践。

2026 年 8 月 2 日