东吴证券 SOOCHOW SECURITIES

汽车行业深度报告

AI 系列深度 13 期：世界模型如何实现？

2026 年 08 月 05 日

增持（维持）

证券分析师 黄细里
执业证书：S0600520010001
021-60199793
huangxl@dwzq.com.cn

投资要点

- **世界模型首先需要进行概念辨析**：认知科学、系统动力学与人工智能均使用“世界模型”一词，但其所指对象和技术目标并不完全相同。其思想来源大致可分为三条脉络：一是 Craik 提出的心智内部“小尺度现实模型”，强调通过内部模拟预演行动并预测结果；二是 Forrester 以 World2 为代表的系统动力学传统，侧重通过外部计算模型模拟复杂系统及其长期演化；三是模型式强化学习与神经世界模型传统，Schmidhuber 较早探索利用递归神经网络学习环境动态并支持规划，Ha 与 Schmidhuber 进一步将环境压缩表征、动态预测与策略学习相结合。三者并非同一技术谱系，但共同关注对环境状态及其演化规律的建模；在当前人工智能语境下，世界模型通常特指智能体用于预测行动后果，并支持规划与决策的内部环境模型。

- **我们将世界模型分歧拆解为四个环节。**第一，预测空间不同：生成式路线还原未来画面或视频，JEPA 等非生成式路线只在抽象表征中预测。第二，输出要求不同：渲染器强调视觉真实，模拟器强调几何、物理和可计算一致性。第三，与动作结合方式不同：可作为数据工厂、闭环训练环境，也可进一步成为架构级融合组件。第四，落地阶段不同：内容型数字 AI 进展最快，自动驾驶闭环最深，机器人需求强度最高但仍处早期。

- **生成式与非生成式是当前最核心的路线分歧。**LeCun 等反对像素级生成，认为其存在容量浪费、画面模糊和误差累积等问题；生成式阵营则认为扩散和潜空间预测已部分化解上述问题，并能直接产出视频、合成数据和仿真环境。纯 JEPA 路线仍面临训练坍缩、可解释性不足、演示难度高和缺少中间产品等工程问题，因此产业侧目前更偏向生成式或混合式路线。

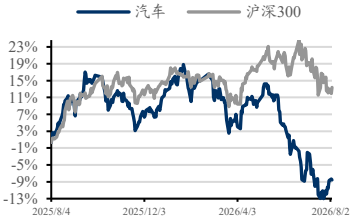
- **落地上，自动驾驶是世界模型闭环最完整的场景：车队回传真实数据，云端世界模型重建和生成场景，策略在虚拟环境中训练与评估，再通过 OTA 回到车端形成数据飞轮。**Waymo World Model、NVIDIA Cosmos、Wayve GAIA、小鹏、理想、华为等均体现世界模型从复现样本走向生成未观测场景、制造长尾数据的趋势。

- **世界模型能否从视觉真实进一步提升至物理一致，仍是其进入物理闭环的关键约束。**三大瓶颈最终收敛到物理一致性不足、长时序不稳和评测无公认标准。场景误差容忍度决定世界模型当前可落地的边界；高风险数字场景和自动驾驶、机器人等物理闭环场景，都需要更严格验证。

- **我们认为，数字 AI 底座模型发展路径已逐渐清晰、价值在于确定性，但其难以实现物理世界的建模闭环；我们认为物理 AI 将成为物理世界的 AI 解决方案，从而在 AI 的当前发展阶段成为增量更大的方向，智能驾驶作为最先跑通闭环的代表场景，我们认为应重点关注。**

- **风险提示：**技术迭代不及预期的风险；大模型厂商 ARR 增速放缓的风险；云服务商资本开支不及预期的风险；智能驾驶及机器人商业化落地不及预期的风险。

行业走势



相关研究

《AI 系列深度 12 期：从模块化到端到端》

2026-08-04

《智能汽车主线周报：曹操出行开放 Robotaxi 无安全员测试，看好智能化》

2026-08-03

内容目录

图表目录

1. 世界模型定义：同名异义背后的共同问题

从表征、网络结构、训练范式，到 LLM、VLM、VLA 与端到端系统，世界模型正在成为连接数字 AI 与物理 AI 的重要概念。其特殊性在于，不同场景都关注世界模型，但定义口径、技术目标和落地方式并不相同。

1.1. 概念溯源：三条相互独立的脉络

1.1.1. 认知科学脉络：内部现实模型

英国心理学家 Kenneth J. W. Craik 在《The Nature of Explanation》中提出，认知系统可携带一个“小尺度的现实模型”：若有机体在内部表征中携带外部现实及自身可能行动的小尺度模型（small-scale model），就能尝试不同备选方案、判断何者更优，并在未来情境发生之前作出反应。需注意，Craik 使用的术语是 small-scale model of reality / mental model，并未使用 world model 一词，属于思想源头而非术语源头。

早期实验心理学证据：美国心理学家 Edward C. Tolman 在《Cognitive Maps in Rats and Men》（The Psychological Review, 1948）中提出动物会形成认知地图（cognitive map），这是内部环境表征思想在实验心理学中的早期证据，同样未使用 world model 一词。

1.1.2. 控制论/系统动力学脉络：全球系统动力学仿真

MIT 系统动力学创始人 Jay Wright Forrester 在《World Dynamics》1971 中提出名为 World2 的全球系统动力学仿真模型，整合人口、资本投资、自然资源、污染、粮食生产五大反馈部门；后续 Donella & Dennis Meadows 在此基础上构建 World3，用于罗马俱乐部《增长的极限》（The Limits to Growth, 1972）。这是 world model 作为具体计算模型被正式命名、广泛使用的早期节点。需区分的是，Forrester 语境下的 world model 指全球尺度社会经济系统仿真，与 AI 语境的 world model 属于同名不同义。

1.1.3. 强化学习/AI 脉络：环境动态预测模型

术语在 AI 语境的首次明确使用，可追溯至 J ü rgen Schmidhuber 1990 年技术报告，FKI-126-90，TU M ü nchen。该系统由控制器 controller 与世界模型 world model 两个循环神经网络组成：世界模型学习预测控制器动作的后果，控制器据此向前规划若干时间步，即如今所称 rollout，并选择能最大化预测累积奖励的动作序列。

David Ha 与 J ü rgen Schmidhuber 2018 年论文《World Models》推动该术语在深度学习领域普及。其认为世界模型可用无监督方式快速训练，学习环境在空间与时间上的压缩表征；智能体甚至可以完全在世界模型生成的虚拟环境中训练，再把策略迁移回真实环境。该论文采用 V（视觉，VAE）+ M（记忆，RNN/MDN-RNN）+ C（控制器）架构，并配以可交互网页，使 world model 一词在深度学习社区迅速流行。

1.2. 口径辨析：“谁第一个提出”与“谁让它流行”

三条脉络的关键辨析在于：其一，认知科学的 mental model、控制论的 world model、深度学习的 world model 三者在用简化内部模型表征与预测外部现实这一抽象层面相通，但具体所指不同；其二，在 AI 语境下，术语首次明确使用应归于 Schmidhuber 在 1990 年，使其在当代深度学习界广为人知的是 Ha & Schmidhuber 在 2018 年；其三，Schmidhuber 近期在 2025 年 11 月公开平台多次主张其 1990 年工作的历史优先权。

表1：世界模型概念三条脉络对照

维度	思想源头	“world model”一词早期出处	现代 AI 含义提出	现代 AI 含义流行推手
代表人物	Craik 在 1943 年、Tolman 在 1948 年	Forrester 在 1971 年	Schmidhuber 在 1990 年	Ha & Schmidhuber 在 2018 年
用语	mental model / 现实模型 / 认知地图	world model，全球仿真模型，	world model，RNN 预测环境的内部模型，	world model，深度学习语境，
所指对象	认知系统内部的现实表征	全球人口-资源-污染系统动力学仿真	神经网络对环境动态的预测性内部模型	同左

数据来源：《World Dynamics》，《World Models》等其他文献，东吴证券研究所整理

2. 各方图谱：世界模型在数字 AI 与物理 AI 中同时升温

2.1. 数字 AI 领域

国际主体方面：OpenAI Sora 技术报告标题为《Video generation models as world simulators》，官方用词为 world simulator，并将 3D 一致性、物体永久性等称为纯粹的规模化涌现现象。Google DeepMind Genie 3 定义为 a general purpose world model，可实时以 24fps、720p 导航并保持数分钟一致性；DeepMind 官方定位世界模型是通往 AGI 的关键基石。

World Labs，李飞飞创业公司，定义为构建能感知、生成、推理并与 3D 世界交互的前沿世界模型，定位为空间智能公司，首款产品 Marble 生成可漫游 3D 世界。Runway General World Models 将世界模型定义为一种构建环境内部表征、并用其模拟未来事件的 AI 系统，2025 年 12 月落地 GWM-1。Meta V-JEPA 2 定义为首个在视频上训练的世界模型，强调理解、预测、规划三能力，主张学习抽象表征而非像素重建。Microsoft Muse/WHAM 可运行于 world model mode，成果发表于 Nature。Decart、Odyssey 等以实时或可交互 3D 世界为主。

中国主体方面：腾讯混元 3D 世界模型 HunyuanWorld，官方称业界首个开源可沉浸漫游、可交互、可仿真的世界生成模型。昆仑万维 Matrix 系列，Matrix-Game 2.0 称业内首个通用场景下实时长序列生成的开源交互式世界模型。生数科技首席科学家朱军称通用世界模型是连接数字世界与物理世界的桥梁，并开源世界模型 Motus。爱诗科技/PixVerse R1 称全球首个支持 1080P 的通用实时世界模型。

2.2. 物理 AI 领域：自动驾驶

国际主体：Wayve，英国，GAIA-1/GAIA-2 官方定名为用于自动驾驶的可控多视角生成式世界模型，用于合成稀有和安全攸关场景。NVIDIA Cosmos World Foundation Model，黄仁勋称世界基础模型对推进机器人与自动驾驶发展至关重要。Waymo 在 2026 年 2 月发布 Waymo World Model，构建于 DeepMind Genie 3 之上，需要与其端到端模型 EMMA 区分，后者官方称端到端多模态模型，非 world model。Waabi World 为由生成式 AI 驱动的闭环模拟器。Tesla 官方自用词为 neural world simulator，world model 多见于第三方或 LeCun 语境。comma.ai 称世界模型是数据驱动的模拟器，已用于 openpilot。

中国主体：理想 MindVLA“基于自研的重建+生成云端统一世界模型”。小鹏须区分“世界基座模型”（72B，车端规控）与“世界模型”（云端仿真闭环）两个概念。华为乾崑 ADS 4 采用“世界引擎+世界行为模型架构 WEWA”，世界引擎生成难例场景密度达真实世界 1000 倍。蔚来 NWM 称“中国首个智能驾驶世界模型”，可在 0.1 秒内推演 216 种场景。小马智行 PonyWorld 强调“博弈交互能力”。商汤绝影“开悟”称“行业首个已量产、可交互的世界模型”。极佳科技推出 DriveDreamer、GigaWorld 等世界模型相关方案。

2.3. 物理 AI 领域：机器人与具身智能

国际主体：NVIDIA Cosmos 定义为能预测并生成具物理感知未来视频的神经网络，可作为多元宇宙仿真引擎，让策略模型模拟不同未来路径；GR00T 官方定义为机器人基座/VLA，其训练数据由 Cosmos 生成。DeepMind Genie 2 是一个基础世界模型，用于训练与评测具身智能体。Meta V-JEPA 2 用于 zero-shot 机器人规划，通过预测候选动作后果做模型预测控制。1X World Model 首要用途定位为策略的评测和仿真，2026 年 1 月宣布 all in on world models。World Labs 将机器人列为下游受益领域。

中国主体：自变量机器人 WALL-B 称全球首个世界统一模型。星动纪元 ERA-42/VPP 将世界模型融入原生机器人大模型。宇树开源 UnifoLM-WMA-0 跨机器人世界模型-动作架构。跨维智能以世界模型为空间/具身智能核心理念，DexVerse，Sim2Real。穹彻智能提出实体世界模型（entity world model）。阿里达摩院 RynnVLA-002 称首个 VLA 与世界模型统一架构。字节 GR-2 具备世界建模能力。

2.4. 学术界

Model-based RL 脉络：Schmidhuber（1990，控制器+世界模型）、David Ha（2018，《World Models》）、Danijar Hafner（Dreamer/DreamerV3）。Dreamer 学习环境模型并通过模型内部生成的未来场景改进行为；DreamerV3 发表于 Nature 2025，把世界模型定义为可供智能体规划与训练的环境动态模型，通常为生成式。JEPA 脉络：Yann LeCun

（2022，《A Path Towards Autonomous Machine Intelligence》），世界模型承担补全缺失状态和预测多个合理未来两项职能，主张以 JEPA 在抽象潜空间做非生成式预测，并批评用生成像素为行动建模世界存在浪费和失败风险。

认知科学/因果脉络：Josh Tenenbaum，MIT，提出心理物理引擎假说；Yejin Choi 强调常识是语言与智能的暗物质；Yoshua Bengio 讨论因果世界模型、System 2 与意识先验。空间智能脉络：Fei-Fei Li 将世界模型定义为一类新型生成式模型，强调 3D 几何与空间智能。中国学术界：北京智源黄铁军，悟界系列、Next-State Prediction 范式；清华朱军，扩散/生成式世界模拟器、Motus；上海 AI Lab，开源 Aether，4D 重建+动作条件视频预测+目标导向视觉规划。

2.5. 世界模型四类用法与分歧根源

综合数字 AI、自动驾驶、机器人和学术界各方定义，我们将世界模型的现行用法归纳为四类：第一类，生成式世界模拟器路线，将世界模型理解为能生成可交互虚拟世界或视频的生成式模型，但视频生成本身并不自动等同于可被物理安全闭环依赖的世界模型，代表如 OpenAI Sora、DeepMind Genie、Wayve GAIA、NVIDIA Cosmos；第二类，预测-规划型路线，将世界模型理解为智能体内部用于预测动作到未来状态、服务规划与控制的模型，代表如 Schmidhuber/Ha/Hafner、Meta V-JEPA、华为世界行为模型；第三类，空间智能/3D 持久世界路线，强调几何一致、可漫游的持久 3D/4D 世界，代表如 World Labs、Odyssey；第四类，物理世界基础模型/认知路线，将世界模型上升为与语言模型并列的物理世界基础模型或类人结构化心智模型，代表如自变量 WALL-B、LeCun JEPA、Tenenbaum/Bengio 认知脉络。

表2：世界模型四大流派归纳

流派	核心主张	代表主体
生成式世界模拟器	世界模型=能生成可交互虚拟世界/视频的生成式模型，但视频生成不自动等同于安全闭环世界模型	OpenAI Sora、DeepMind Genie、Runway、腾讯混元 3D、Wayve GAIA、NVIDIA Cosmos、商汤开悟
预测-规划型，决策内核，	世界模型=智能体内部预测“动作→未来状态”、服务规划与控制的模型	Schmidhuber/Ha/Hafner、Meta V-JEPA、华为世界行为模型、小鹏世界基座、星动纪元
空间智能/3D 持久世界	强调几何一致、可漫游、可导出引擎的持久 3D/4D 世界	World Labs、Odyssey、昆仑 Matrix-3D
物理世界基础模型/认知派	世界模型上升为与 LLM 并列的物理世界基础模型，或类人结构化心智模型	自变量 WALL-B、LeCun JEPA、Tenenbaum/Bengio 认知脉络

数据来源：《Video generation models as world simulators》，《World Models》等其他文献，东吴证券研究所整理

分歧根源可归纳为四点：其一，预测在哪个空间进行，是像素/观测空间中的生成式预测，还是抽象潜表示空间中的非生成式预测，这是 LeCun 与生成式阵营最尖锐的分歧；其二，世界模型服务于什么目标，是内容生成与仿真、智能体规划控制，还是感知-表示学习与自主智能；其三，世界模型是系统模块还是基础模型，是服务现有 VLA、

驾驶和机器人系统的一个组件，还是被上升为与 LLM 并列的物理世界基础模型；其四，对物理真实和因果一致的要求强度，是仅需视觉逼真一致，还是必须遵循物理规律并可做反事实因果推演。

3. 本质分歧：在哪预测、预测什么、如何接动作

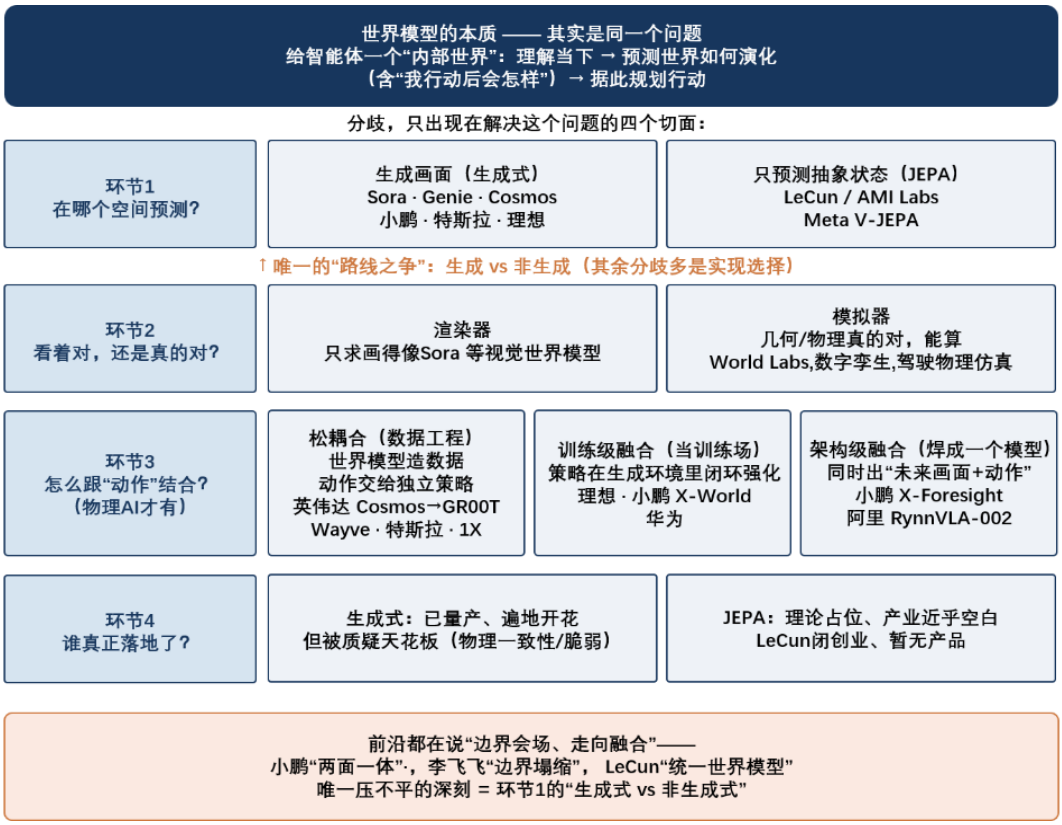
3.1. 本质框架：世界模型解决的共同问题

从技术本质看，世界模型要解决的核心问题，是为智能体构建一个内部环境表征，使其能够理解当前状态、预测环境后续演化，尤其是预测不同动作可能带来的状态变化，并据此进行规划。理解、预测、规划三项能力构成世界模型的内核。

以自动驾驶为例，车辆在变道前需要评估当前车道、目标车道、周边车辆速度和潜在交互结果，并比较不同操作方案的风险收益。这一过程对应的并非单纯识别画面，而是在内部表征中预测环境对动作的响应，再选择相对稳健的执行方案。

从 Craik 在 1943 年提出小尺度现实模型，到 Schmidhuber 在 1990 年、Ha 在 2018 年将其实现为可预测环境的神经网络，再到 LeCun、李飞飞、小鹏等主体的近期表述，上述内核基本一致。分歧并不在于世界模型要解决什么问题，而在于以何种技术路径解决该问题。因此，后续可将路线差异拆解为四个环节：预测空间、输出真实性、动作结合方式与产业落地。

图1：世界模型的本质与四个分歧环节



数据来源：《World Models》，《A Path Towards Autonomous Machine Intelligence》等其他文献，东吴证券研究所绘制

3.2. 分歧环节一：生成观测结果还是预测抽象状态

智能体要预测世界如何演化，首先需要确定预测发生在哪个空间。一类是生成式路线，直接生成未来画面、观测或视频，使未来状态可视化；另一类是非生成式路线，即 LeCun 主张的联合嵌入预测架构（JEPA），只在抽象表征空间中预测下一状态，并不还原为画面。前者强调可见、可生成、可复用，后者强调抽象表征、规划效率和避免像素级冗余。

这是当前最主要的路线分歧之一，并非单纯实现细节，而是路线选择。LeCun 公开表示，用生成像素的方式为行动建模世界，与早已被放弃的分析-合成一样浪费、注定失败；以 OpenAI Sora 为代表的生成式阵营则认为，直接生成像素也可能随规模扩展涌现出对世界的理解，且生成结果可用于内容生产、训练数据合成和仿真环境构建。双方根本分歧在于，智能系统是否需要显式生成未来世界。

从产业站位看，公开落地样本更多集中在生成式路线：OpenAI Sora、谷歌 DeepMind Genie、英伟达 Cosmos，以及自动驾驶领域的小鹏、特斯拉、理想等，均以生成式或混合式方法为主。非生成式 JEPA 路线主要由 LeCun 及 Meta 研究院 V-JEPA 系列推动；LeCun 已于 2025 年 11 月离开 Meta、另立 AMI Labs，2026 年 3 月融资约 10.3 亿美元，重点布局该路线。

需要区分的是，不少生成式系统也在潜空间或 token 空间中做预测，小鹏即明确表示在统一 token 空间预测，表面上接近抽象状态预测；但只要最终仍要解码和生成画面，本质仍属于生成式路线，而非 JEPA。判断标准在于预测后是否还原画面：还原则属于生成式，只停留在抽象表征而不还原，才属于 JEPA 路线。

需要区分的是，生成式与 JEPA 的路线分歧与 Transformer 架构本身无关。JEPA 的骨干网络同样可采用视觉 Transformer，Sora 也是基于 Transformer 的扩散模型。生成式与 JEPA 之争，核心在于训练目标位于像素/观测空间还是抽象表征空间，而不是网络结构本身。

双方论据可进一步拆解。LeCun 反对像素生成的理由包括：1）容量浪费，世界中大量细节不可预测且对决策不重要，例如风中树叶位置，要求像素级预测会消耗算力；2）输出模糊，未来存在多种可能时，像素空间回归损失容易生成平均化结果；3）误差累积，逐帧生成会放大早期错误。生成式阵营的回应是：扩散模型并非简单取平均，而是在多种可能未来中采样清晰结果；不少系统也先在压缩潜空间中预测、最后再解码，已部分降低冗余。抽象路线自身同样存在难点：1）联合嵌入训练存在坍缩风险，需专门技巧防范；2）预测状态不可直接观察，训练效果只能通过下游任务间接验证，调试和展示难度更高；3）哪些细节应被抽象掉难以先验确定，精细操作中被忽略的细节可能恰是关键；4）缺少视频、合成数据、可视化仿真等中间产物。

表3：生成式 vs 非生成式 JEPA 逐项对比

对比维度	生成式，显式生成未来观测	非生成式 JEPA，在抽象层面预测
预测对象	未来的画面/观测，像素、视频、3D，	未来的抽象状态表征
能否直接看到	能，直接生成可观察画面	不能，只在表征空间、不还原画面
典型用途	内容生成、数据合成、仿真，以及融合后用于决策	表征学习、为规划提供预测
代表玩家	OpenAI Sora、DeepMind Genie、英伟达 Cosmos、小鹏、特斯拉等，产业侧主流路线	LeCun／AMI Labs、Meta V-JEPA
落地现状	已出现量产与商业化样本，但物理一致性与能力上限仍受质疑	理论占位较强，产业落地仍少，AMI Labs 新立、暂无产品

数据来源：《Video generation models as world simulators》，《A Path Towards Autonomous Machine Intelligence》等其他文献，东吴证券研究所整理

生成式内部还可分为两条并列实现路径。一条是自回归路线：类似语言模型，将画面切分为视觉 token，并按时间顺序逐个预测，DeepMind Genie 即属此类。另一条是扩散路线：先对整段画面加噪，再学习逐步去噪还原，一次性生成整段内容，OpenAI Sora、小鹏 X-World 属于该类。两条路径各有取舍：自回归天然适合实时交互和连续生成，但长序列误差会累积；扩散整体一致性和画质更强，适合处理多种可能未来，但需要反复迭代去噪，推理较慢且算力负担更重。小鹏研发 X-Cache 为世界模型推理提速约 2.7 倍，

也体现出扩散路线的算力约束。扩散与自回归并非迭代关系，而是源自不同数学框架的平行路线；英伟达 Cosmos 同时提供自回归与扩散两个模型家族，说明业界已出现按场景取舍的混合策略。

表4：生成式内部两条实现路径：自回归 vs 扩散

对比维度	自回归，预测下一个 token，	扩散，整体去噪，
生成方式	按时间顺序逐个生成视觉 token	整段加噪后迭代去噪，一次性生成
强项	实时、可交互、可无限延续	整体一致性与画质高，天然处理多种可能的未来
弱项	长序列误差累积	迭代去噪推理慢、算力负担重
代表	DeepMind Genie、Cosmos 自回归家族	OpenAI Sora、小鹏 X-World、Cosmos 扩散家族
典型场景	可交互世界、实时推演	高质量视频、离线造数据

数据来源：《Video generation models as world simulators》，《Cosmos World Foundation Model Platform for Physical AI》等其他文献，东吴证券研究所整理

3.3. 分歧环节二：视觉真实还是物理一致——渲染器与模拟器

即便同样选择生成式路线，下一层分歧仍在于输出结果的约束强度：生成出的未来只需要视觉上可信，还是必须在几何、物理和动力学上成立。李飞飞 World Labs 将常被笼统称作世界模型的对象拆分为渲染器与模拟器两类。渲染器侧重视觉逼真，产出的是观察者看到的样子，但模型本身未必具备稳定的三维结构与物理规律；模拟器则需要输出可计算、可交互、几何与物理约束一致的状态。

从能力要求看，渲染器更关注视觉保真度，适合内容生成、可视化展示和低风险交互；模拟器更关注状态可计算性和物理一致性，适合工程仿真、自动驾驶闭环训练、机器人策略评估等对安全和可验证性要求更高的场景。

分歧取决于应用约束。多数低风险内容和视频场景更强调视觉可信，因此 Sora 这类视频世界模型更偏渲染器；但金融、医疗、代码执行等高风险数字场景同样不能只满足视觉可信。机器人与自动驾驶要求世界物理一致、结果可计算，否则在仿真环境中训练出的策略迁移到真实场景后可能失效，因此 World Labs 的可漫游 3D 世界、数字孪生、自动驾驶物理闭环仿真更偏模拟器。

这也解释了当前视频世界模型的核心约束：不少模型具备较强视觉表现，但仍停留在渲染器层面，物理一致性仍是公认瓶颈，包括物体穿模、重力与碰撞关系不稳定等。2026 年 5 月，由 Mila 牵头、LeCun 参与的开源评测平台 stable-worldmodel 给出量化佐证：被测世界模型在标准推物任务上干净条件成功率约 50.8%，仅改变智能体本体颜色即降至约 12%，改变被推物块颜色降至约 22%，改变背景颜色降至约 6%，整体鲁棒性仍弱。渲染器与模拟器并非互斥公司分类，而是同一生成式路线上的保真度要求梯度，不少团队正由视觉真实向物理一致推进。

3.4. 分歧环节三：如何与动作结合——兼谈世界模型与 VLA 的关系

前两个环节决定如何预测世界，第三个环节决定预测结果是否以及如何与动作结合。数字 AI 与物理 AI 在此出现明显分野：多数内容型数字 AI 不需要动作输出，通常停留在生成或预测世界环节；涉及代码执行、金融、医疗等高风险数字任务时，则同样需要验证与约束。物理 AI 中的机器人和自动驾驶必须连接动作，但连接紧密程度差异较大，我们归纳为三种方式。

第一种是松耦合，世界模型作为仿真器或数据工厂，负责生成场景、生成数据和运行闭环仿真，动作由另一个独立训练的策略模型输出。英伟达用 Cosmos 生成数据训练机器人基座模型 GR00T、Wayve 用 GAIA 生成驾驶场景、特斯拉用神经世界模拟器训练独立端到端驾驶网络，均属此类。第二种是训练级融合，世界模型作为训练环境，策略在其生成的环境中进行大规模闭环强化学习，代表包括理想 MindVLA、小鹏 X-World、华为世界引擎。第三种是架构级融合，将预测世界与输出动作纳入同一模型，同时输出未来画面与自车动作；小鹏 X-Foresight（其原话是在统一 token 空间里联合预测未来多视角画面与自车动作）以及阿里达摩院 RynnVLA-002（VLA 与世界模型统一架构）是代表样本。

世界模型与 VLA 的关系容易混淆。VLA（视觉-语言-动作模型）主要负责动作输出，即基于视觉输入和语言指令生成执行动作；世界模型主要负责环境预测，即预测后续状态如何演化。两者并非对立路线，也不完全等同，而是可在部分架构中耦合的两个环节。纯 VLA，如 Physical Intelligence 的 π、智元 GO-1、银河通用 GraspVLA，通常是当前画面到动作的直接映射，中间不显式预测未来；融合路线则在 VLA 上叠加先预测未来、再决定动作的能力。小鹏在 CVPR 2026 上表示，VLA 和世界模型不是两种割裂技术，而是在特定架构中可以融合的两类能力；其物理世界基座模型同时承担向人类学习如何行动的 VLA 目标与向世界学习其如何变化的世界模型目标。

因此，世界模型更偏环境预测与反事实推演，VLA 更偏动作生成与执行映射；融合架构则试图在同一系统内同时完成未来状态预测和动作决策。

表5：物理 AI 里世界模型与“动作”的三种结合方式

结合方式	做法	代表公司
松耦合：世界模型作为数据工厂/仿真器	世界模型造场景/数据，动作交给另一个独立策略	英伟达 Cosmos→GR00T、Wayve、特斯拉神经世界模拟器、1X，评测，
训练级融合：世界模型作为闭环训练环境	策略在世界模型生成的环境里做闭环强化学习	理想 MindVLA、小鹏 X-World、华为世界引擎
架构级融合：一个模型同时输出画面与动作	统一架构联合预测未来画面与动作	小鹏 X-Foresight、阿里 RynnVLA-002
对照：纯 VLA，不带世界模型	当前画面+指令→直接输出动作，不显式预测未来	Physical Intelligence，π、智元 GO-1、银河通用 GraspVLA

数据来源：《Cosmos World Foundation Model Platform for Physical AI》，《GAIA-1: A Generative World Model for Autonomous Driving》等其他文献，英伟达、特斯拉、理想、小鹏、华为等公司官方网站，东吴证券研究所整理

在 VLA 的双系统路线中，System 2 通常承担较慢但更稳健的规划功能，System 1 则负责实时动作输出。若 System 2 需要承担预测未来并据此决策的功能，则往往需要引入世界模型。LeCun 的主张正对应这一方向：System 2 不应是生成像素或 token 的大模型，而应是在抽象表征空间中推演的 JEPA 世界模型。

3.5. 分歧环节四：产业落地进度——理论与产业的落差

从产业落地看，当前路线落地呈现明显不对称：生成式路线已在自动驾驶闭环仿真与数据飞轮、数字内容生成、机器人训练数据合成等场景出现量产或商业化样本，但其物理一致性不足、长时序发散和整体鲁棒性偏弱的问题仍限制能力上限，能否支撑高可靠物理闭环仍有待验证。

非生成式 JEPA 路线则相反：理论声量较大，但产业落地样本仍少。其主要载体已从 Meta 研究院转向 LeCun 新成立的 AMI Labs，目前尚无成熟产品。值得注意的是，AMI Labs 投资方包括英伟达、三星、丰田、贝索斯等产业资本，这说明产业端并未忽视该路线，但相关主体自身产品仍主要采用生成式方法，反映出 JEPA 路线尚处验证期。

产业侧较少跟随 JEPA 路线，我们归纳为五点原因。其一，商业化中间产物不足：生成式路线每前进一步都可能形成可售卖的视频产品、合成数据或仿真服务，收入可反哺研发；JEPA 的价值更多要到机器人与规划效果显著提升后才能兑现，中间产品较少。其二，路径依赖较强：生成式可复用大语言模型已验证的基础设施、数据管线、人才和预测下一个 token/去噪加规模化训练配方；切换 JEPA 意味着采用尚未形成成熟最佳实践、且存在坍缩风险的新路线。其三，演示不对称：生成视频可直接呈现效果，而 JEPA 成果更多体现为基准测试或下游规划指标，传播和融资效率较低。其四，缺少规模化存在性证明：V-JEPA 2 的机器人结果具备进展，但尚未形成压倒性效果，资金端难以大规模押注。其五，部分批评已被生成式路线吸收：业界认为在生成式框架内做潜空间预测加解码，已获得部分 JEPA 收益，纯 JEPA 的边际价值仍需验证。

算力维度也需要单独讨论。生成式路线对算力要求较高：训练端，视频生成大模型训练成本高；推理端，扩散模型需迭代去噪、逐帧渲染，Sora 应用提供了直接案例，公开报道显示其算力消耗约每天 100 万美元，OpenAI 已于 2026 年 3 月宣布关停该应用。JEPA 的潜在优势之一是降低算力需求：不渲染像素、在低维抽象空间预测，规划推理无需生成视频。因此，若 JEPA 路线成功，单位世界理解能力的算力需求有望显著下降，尤其体现在推理端。英伟达进入 AMI Labs 投资方名单，也可视为算力供给方对潜在技术路线变化的对冲。

3.6. 小结：共同问题、路径分化与收敛趋势

综合四个环节，各方分歧本质上是在解决同一个问题时作出的不同设计选择，而不是在解决不同问题。不同主体在预测空间、输出真实性、动作结合和产业落地四个环节上选择了不同坐标，从而形成差异化的路线图谱。

同时，越前沿的主体越倾向于探索环节融合：小鹏提出 VLA 与世界模型是两面一体，李飞飞认为渲染器、模拟器与规划器的边界会塌缩为空间智能，LeCun 也将理解、预测、规划统一在一个世界模型框架内。因此，方法层面的多数分叉，正在呈现融合和收敛趋势。

仍难以收敛的是生成式与非生成式 JEPA 的底层路线分歧。这并非实现细节，而是关于智能系统是否需要显式生成世界的根本判断。

表6：核心公司的世界模型路线坐标

公司	①预测空间	②保真度	③与动作结合	④落地
OpenAI（Sora/机器人）	生成式	偏渲染器	数字侧停在生成；机器人方向新建	Sora 应用 2026 年关停；转向机器人
谷歌 DeepMind Genie	生成式	渲染器→可交互	训练/评测具身智能体	研究/产品化中
World Labs（李飞飞）	生成式	偏模拟器（3D）	暂以“造世界”为主	早期产品 Marble
Meta／LeCun JEPA	非生成式	不生成画面	进决策回路做规划	研究；LeCun 转 AMI Labs
英伟达 Cosmos	生成式	模拟器/数据工厂	松耦合（造数据）+Cosmos Policy	已落地（数据/仿真）
特斯拉	生成式	模拟器（仿真）	松耦合：世界模拟器+端到端网络	已量产 FSD
小鹏	生成式	渲染+物理仿真	架构级融合 X-Foresight	二代 VLA 已推送；联合预测升级预计年内
理想	生成式	模拟器（闭环）	训练级融合 MindVLA	已落地
华为	生成式	模拟器（世界引擎）	训练级融合 WEWA	已落地 ADS4
阿里达摩院	生成式	—	架构级融合 RynnVLA-002	开源/研究
1X	生成式	模拟器（评测）	松耦合，评测/仿真	落地中
星动纪元	生成式	预测未来画面	融合，VPP 视频预测策略，	落地中

数据来源：《Video generation models as world simulators》，《A Path Towards Autonomous Machine Intelligence》等其他文献，OpenAI、谷歌 DeepMind、World Labs、英伟达、特斯拉、小鹏、理想、华为等公司官方网站，东吴证券研究所整理

4. 落地视角：自动驾驶闭环最深，机器人仍处早期

4.1. 落地总览：三大场景的要求与进展

三大场景对世界模型的要求存在明显差异，并影响落地深度。多数低风险内容型数字 AI 更偏视觉真实，容错成本较低，画面瑕疵通常不构成安全事故；自动驾驶要求物理基本一致且结果可验证，场景受道路与交规约束；机器人则要求接触层面的力、形变和摩擦都能被建模，且任务更开放。

世界模型的实际落地程度可从三个维度判断：其一，是否进入量产闭环，是上了量产车或真实产品，还是停留在演示视频；其二，是否有可验证指标，是有公开基准、成功率、推送范围，还是只有定性描述；其三，是否替代旧流程，是确实取代人工标注、实车测试、人工建模中的某一环，还是仅作为旧流程旁的新增演示。

表7：三大场景对世界模型的要求与落地深度

维度	数字 AI	自动驾驶	机器人
核心用途	生成内容、可交互世界	造数据/仿真闭环、车端预测	合成数据、评测、策略内预测
保真度要求	多数低风险内容场景满足视觉真实即可	物理基本一致、可验证	接触/力学层面也要对
容错成本	低，画面瑕疵通常不构成安全风险	高，涉及行车安全，	高，损坏实物、伤人风险，
数据可得性	海量互联网视频	量产车队持续回传	真机数据极稀缺
落地深度	商业化最快，已有产品在售	闭环最深，已进入量产智驾	仍处早期，多为演示或开源
代表玩家	OpenAI、DeepMind、World Labs、腾讯、昆仑万维、爱诗	理想、小鹏、华为、蔚来、Waymo、Wayve	英伟达、1X、宇树、自变量、星动纪元

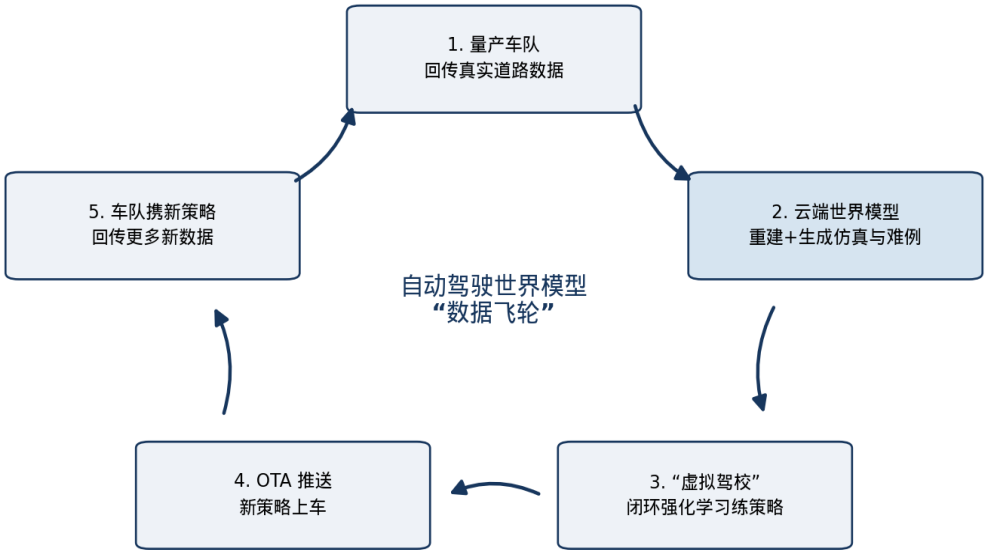
数据来源：《Video generation models as world simulators》，《GAIA-1: A Generative World Model for Autonomous Driving》等其他文献，OpenAI、谷歌 DeepMind、Waymo、理想、小鹏、华为、蔚来等公司官方网站，东吴证券研究所整理

4.2. 自动驾驶：落地最深、闭环最完整

自动驾驶是世界模型落地最深的场景，用法可归纳为三类：场景与数据生成，代表包括 Wayve GAIA、小鹏 X-World、华为世界引擎；闭环强化训练环境，代表包括理想、小鹏、华为的云端仿真+强化学习；车端联合预测，代表为小鹏 X-Foresight。

将上述用法串联起来，可形成自动驾驶世界模型的核心工程闭环，即数据飞轮：第一步，量产车队在真实道路上持续回传数据；第二步，云端世界模型对数据进行重建与生成，形成更接近真实世界并经过校验的仿真环境，同时专门生成现实中难以采集的危险或罕见场景；第三步，驾驶策略在该虚拟环境中开展大规模闭环强化学习，以降低真实道路试错成本；第四步，训练后的策略通过 OTA 推送上车；第五步，车队带着新策略回传更多新数据，飞轮进入下一轮迭代。

图2：自动驾驶世界模型的数据飞轮闭环



理想、小鹏、华为、特斯拉等均按此闭环运转；华为称其“世界引擎”生成的难例密度可达真实世界的 1000 倍

数据来源：理想、小鹏、华为、Wayve 等公司官方网站，东吴证券研究所绘制

自动驾驶之所以落地最深，关键在于三个条件同时具备：数据方面，数十万量产车可持续回传；场景方面，道路与交规约束了世界开放度；价值密度方面，安全与体验可直接转化为产品竞争力。瓶颈同样需要客观列出：其一，仿真到现实的差距，虚拟环境训练出的能力迁移到真实道路后折损多少，尚无公认答案；其二，长尾覆盖仍不充分，世界模型可以生成难例，但不知道自己未覆盖哪些场景的问题并未根除；其三，评测可信度，若用世界模型生成的场景考核由世界模型训练出的策略，可能存在评测独立性不足的方法论问题。

4.3. 小结：落地次序与共同瓶颈

把三大场景放在一起，落地次序清晰可见：多数内容型数字 AI 商业化最快，产品在售，但也已出现 Sora 这样的算力成本退场案例；自动驾驶闭环最深，已进入量产智驾；机器人需求强度最高但进展仍处早期，首个消费级人形机器人已开订投产、交付计

划年内。这个次序可以用三个因子解释：容错成本越低、数据越易得、保真度要求越宽松，落地越快。多数低风险内容型数字 AI 三项占优，机器人三项均难，自动驾驶居中，但可凭借量产车队的数据优势提升闭环进度。

三大场景的瓶颈最终收敛到三件事：物理一致性不足、长时序不稳、评测无公认标准。这三项约束均落在从视觉真实到物理一致的能力梯度上。落地次序本质上取决于各场景对物理一致性的需求强度：世界模型当前能落地到哪里，取决于该场景能容忍多少误差；高风险数字场景和物理闭环场景都需要更严格验证；未来能走多远，则取决于物理一致性、长时序稳定性和独立评测体系能否持续突破。

5. 风险提示

技术迭代不及预期的风险：数字 AI、物理 AI 技术进展低于预期，任务或应用场景拓宽速度低于预期。

大模型厂商 ARR 增速放缓的风险：若客户增长、续费率或客单价不及预期，模型厂商 ARR 增速可能放缓。

云服务商资本开支不及预期的风险：若自由现金流承压，或 AI 算力需求及投资回报低于预期，云服务商可能下调资本开支。

智能驾驶及机器人商业化落地不及预期的风险：若产品可靠性、成本下降或用户需求低于预期，相关产品商业化进程可能放缓。

东吴证券 SOOCHOW SECURITIES

免责声明

东吴证券股份有限公司经中国证券监督管理委员会批准，已具备证券投资咨询业务资格。

本研究报告仅供东吴证券股份有限公司（以下简称“本公司”）的客户使用。本公司不会因接收人收到本报告而视其为客户。在任何情况下，本报告中的信息或所表述的意见并不构成对任何人的投资建议，本公司及作者不对任何人因使用本报告中的内容所导致的任何后果负任何责任。任何形式的分享证券投资收益或者分担证券投资损失的书面或口头承诺均为无效。

在法律许可的情况下，东吴证券及其所属关联机构可能会持有报告中提到的公司所发行的证券并进行交易，还可能为这些公司提供投资银行服务或其他服务。

市场有风险，投资需谨慎。本报告是基于本公司分析师认为可靠且已公开的信息，本公司力求但不保证这些信息的准确性和完整性，也不保证文中观点或陈述不会发生任何变更，在不同时期，本公司可发出与本报告所载资料、意见及推测不一致的报告。

东吴证券投资评级标准

投资评级基于分析师对报告发布日后 6 至 12 个月内行业或公司回报潜力相对基准表现的预期（A 股市场基准为沪深 300 指数，香港市场基准为恒生指数，美国市场基准为标普 500 指数，新三板基准指数为三板成指（针对协议转让标的）或三板做市指数（针对做市转让标的），北交所基准指数为北证 50 指数），具体如下：

公司投资评级：

买入：预期未来 6 个月个股涨跌幅相对基准在 15%以上；

增持：预期未来 6 个月个股涨跌幅相对基准介于 5%与 15%之间；

中性：预期未来 6 个月个股涨跌幅相对基准介于-5%与 5%之间；

减持：预期未来 6 个月个股涨跌幅相对基准介于-15%与-5%之间；

卖出：预期未来 6 个月个股涨跌幅相对基准在-15%以下。

行业投资评级：

增持： 预期未来 6 个月内，行业指数相对强于基准 5%以上；

中性： 预期未来 6 个月内，行业指数相对基准-5%与 5%；

减持： 预期未来 6 个月内，行业指数相对弱于基准 5%以上。

我们在此提醒您，不同证券研究机构采用不同的评级术语及评级标准。我们采用的是相对评级体系，表示投资的相对比重建议。投资者买入或者卖出证券的决定应当充分考虑自身特定状况，如具体投资目的、财务状况以及特定需求等，并完整理解和使用本报告内容，不应视本报告为做出投资决策的唯一因素。

东吴证券研究所
苏州工业园区星阳街 5 号
邮政编码：215021
传真：（0512）62938527
公司网址： http://www.dwzq.com.cn