东吴证券 SOOCHOW SECURITIES

汽车行业深度报告

AI 系列深度 01 期：从数字 AI 通用底座到物理 AI 真实世界闭环

2026 年 07 月 27 日

增持（维持）

证券分析师 黄细里
执业证书：S0600520010001
021-60199793
huangxl@dwzq.com.cn
研究助理 童明祺
执业证书：S0600125080005
tongmq@dwzq.com.cn

投资要点

- **数字 AI 已实现在语言、视觉与多模态任务中通用底座模型的跑通，但其难以实现物理世界的建模闭环；**物理 AI 将在连续、实时、安全约束的真实世界中建立"感知—行动—反馈—优化"闭环，并在一段时间内与数字 AI 并列成为"任务通解"，成为物理世界场景的 AI 解决方案。考虑二者的技术同源性，我们认为复盘数字 AI 有助于理解与预测物理 AI 的演进。

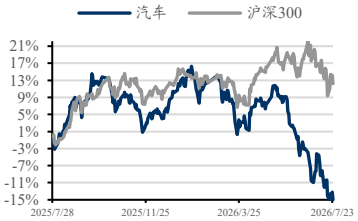
- **Transformer 推动数字 AI 技术发展从专用架构到统一底座。**AI 第一次爆发以 CNN、RNN 等专用架构分别攻克视觉与语言任务，归纳偏置强、通用性受限；AI 第二次爆发中，Transformer 以"注意力+位置编码"提供跨领域通用的关系建模，叠加 Scaling 规律，统一模型在能力迁移、研发效率与工程复用上的优势持续扩大；语言、视觉、生成与多模态逐步收敛于统一基础模型范式，Transformer 亦转变为通用计算底座。

- **物理 AI 正从复用数字 AI 能力转向补齐自身底层能力。**早期主要迁移 Transformer 等架构理念（BEV、UniAD）；当前 VLM、VLA 进一步建立在 LLM 及多模态基座之上（Waymo EMMA 基于 Gemini、车端 VLA）。但数字 AI 红利或已接近阶段性瓶颈，根本原因在于，物理 AI 面临三类差异：一是缺少天然 Token，需自主构建物理世界表征；二是长尾、极端场景稀缺且采集成本高，难以直接复制互联网数据 Scaling；三是模仿学习受限于人类驾驶上限，必须依赖强化学习持续试错；就三点差异来看，我们认为世界模型或更契合物理 AI 发展方向。

- **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；本轮 AI 发展未来更大的增量在物理 AI，智能驾驶是其最先跑通闭环的代表场景。**物理 AI 从表征到部署的闭环路径与数字 AI 差异明显，将形成独立而统一的发展路径。随着世界模型与强化学习闭环走向成熟、云车部署逐步兑现，物理 AI 有望由能力展示走向大规模落地，成为 AI 发展当前阶段最具成长弹性的方向。

- **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

《整车主线周报：本周 SW 商用载客车表现较好，重卡 6 月海关出口数据公布》

2026-07-27

《智能汽车主线周报：小鹏集团发布 TuringViT，看好智能化》

2026-07-26

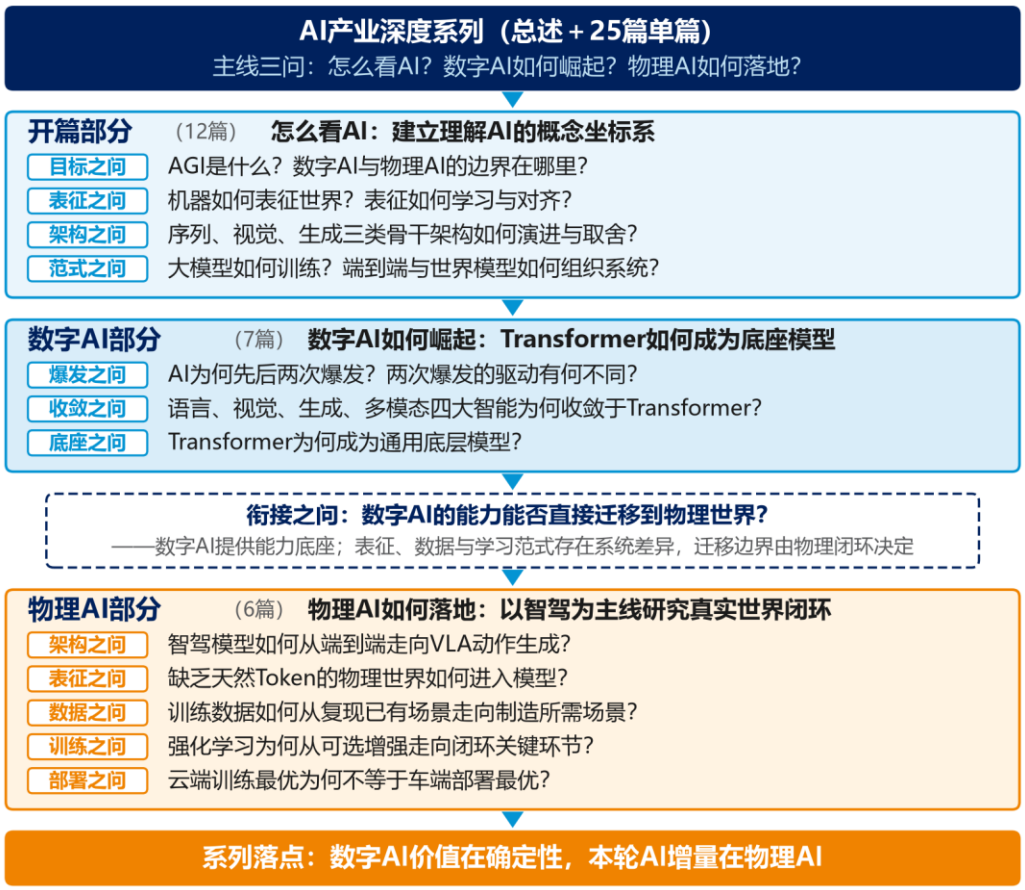
内容目录

图表目录

1. 主线三问：怎么看 AI、数字 AI 如何崛起、物理 AI 如何落地

AI 系列深度报告围绕三个递进的问题展开：怎么看 AI——理解 AI 需要哪些概念工具？数字 AI 如何崛起——Transformer 为何成为通用底座模型？物理 AI 如何落地——能力如何迁移到物理世界并形成真实闭环，增量将在哪里落地？对应三问，系列分为开篇、数字 AI、物理 AI 三大部分。

图1：AI 系列深度报告：主线三问



数据来源：东吴证券研究所绘制

开篇部分以十二篇概念科普铺设底座，从研究目标、表征机制、骨干架构到训练范式，为全系列提供统一的概念坐标系。数字 AI 部分复盘两次爆发的驱动差异：第一次爆发中 CNN、RNN 等专用架构分别攻克单点任务，第二次爆发中规模化学习使 Transformer 收敛为通用底座，语言、视觉、生成、多模态四大智能由分立走向统一。数字 AI 与物理 AI 在表征、数据与学习范式上存在系统差异，能力迁移的边界由物理闭环决定，构成两部分之间的衔接。物理 AI 部分以智能驾驶为主线，沿架构、表征、数据、训练、部署五问研究真实世界闭环：VLA 完成从端到端到动作生成的范式升级，世界模型承担表征训练与数据制造，强化学习从可选增强走向闭环关键，时延约束决定部署工程。我们认为，物理 AI 有望由能力展示走向大规模落地，成为 AI 发展当前阶段最具成

长弹性的方向。

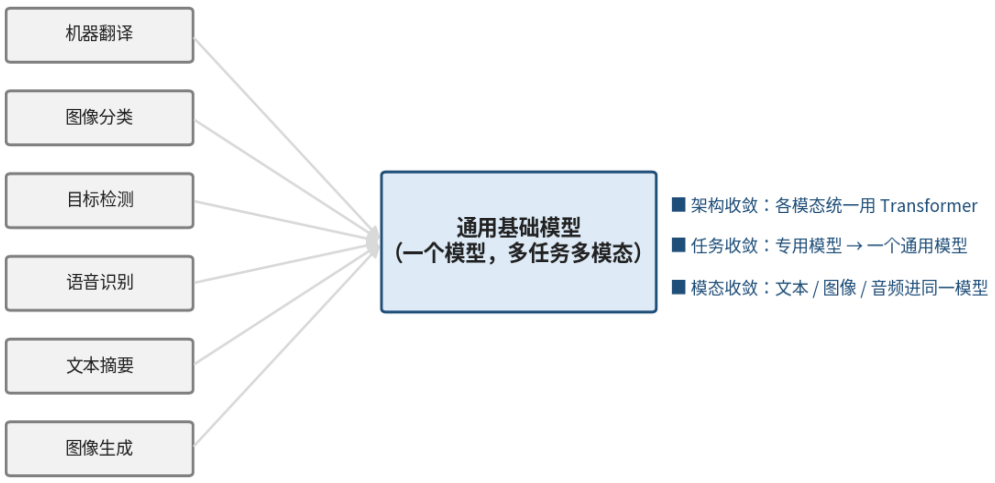
2. 数字 AI：Transformer 为何成为通用底层模型

AI 1.0 主要提升了单点任务的识别精度；Transformer 开启的 AI 2.0 则同时提升了任务的广度、复杂度和链条长度。Transformer 已演进为能够统一学习知识、迁移任务和执行复杂指令的通用基础模型，是目前通向 AGI 最重要、且得到最大规模工程验证的一次跨越。

Transformer 是一种通用模型架构，是数字 AI 迈向大模型的核心技术底座之一；物理 AI 则是在这一底座上，叠加对物理世界感知、预测和行动能力形成的系统。因此，在当前发展阶段，数字 AI 与物理 AI 具有一定的技术同源性，Transformer 是其共同的核心技术底座。

在 AI2.0 时代，Transformer 于数字 AI 场景首先落地，发展速度也快于物理 AI 场景，考虑二者的技术同源性，**我们认为通过对数字 AI 发展的复盘可以更好的理解和预测物理 AI 的发展。**

图2：专用模型向通用基础模型收敛



数据来源：东吴证券研究所绘制

2.1. AI 第一次爆发：CNN 与 RNN 以专用架构分别攻克任务

AI 第一次爆发是传统机器学习到深度学习的跨越。传统机器学习通常先由工程师设计边缘、纹理、形状等特征再交由分类器识别；深度学习则直接从原始数据逐层学习表示，底层捕捉边缘、中层学习部件、高层形成对完整物体的抽象刻画，即表示学习。

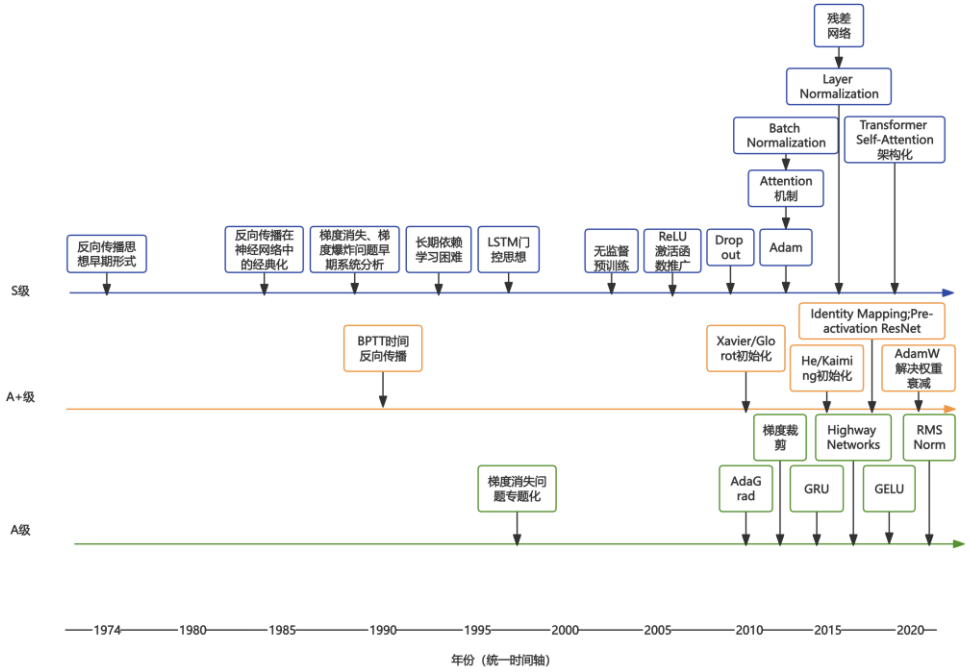
表1：传统机器学习与深度学习对比

对比维度	传统机器学习	深度学习
特征获取	人工设计特征	自动学习特征
模型结构	通常较浅	多层非线性结构
依赖要素	依赖领域专家	更依赖数据、计算和通用学习算法
训练范式	特征提取与分类常分开	端到端联合训练

数据来源：《Deep Learning》，东吴证券研究所整理

深度学习的成功并非单一变量驱动，而是大数据、强计算能力与可训练深层网络算法三项条件同步成熟的结果。从技术演进看，反向传播（1974）、梯度消失与爆炸的系统论证（1991）、LSTM 门控（1997）、无监督预训练（2006）、ReLU（2010）、Dropout（2014）、注意力机制（2015）、ResNet 残差连接（2016）等成果具有较强跨架构适配性，共同构成深度学习技术体系的公共底座；其中重要成果在 2010 年后密集出现。

图3：深度学习通用学术成果时间轴



数据来源：《ImageNet Classification with Deep Convolutional Neural Networks》，《Attention Is All You Need》等其他文献，东吴证券研究所

2012 年，AlexNet 在 ImageNet 上将 ILSVRC 口径的 top-5 错误率由约 26%降至约 15%，成为深度学习在视觉任务上实现代际跃迁的标志性拐点，此后 CNN 快速确立其

在图像识别、目标检测、图像分割与人脸识别等任务中的核心地位。该突破来自 GPU 并行训练、ReLU 激活、Dropout 正则化与数据增强的协同，也是算法、算力与数据长期积累后在 2012 年形成技术共振的结果。

以论文为锚，我们认为深度学习架构演进可划分为七个阶段，并归并为范式确立、成长与扩展、成熟与高峰三大阶段。其分析价值在于给出一条判据：当某一架构的高引用论文向成熟与高峰阶段（轻量化与部署、复兴性改进）集中、且新增高引用基础架构论文减少时，通常意味着该架构进入短期学术瓶颈，一方面产品落地窗口开启，另一方面需关注下一代架构。

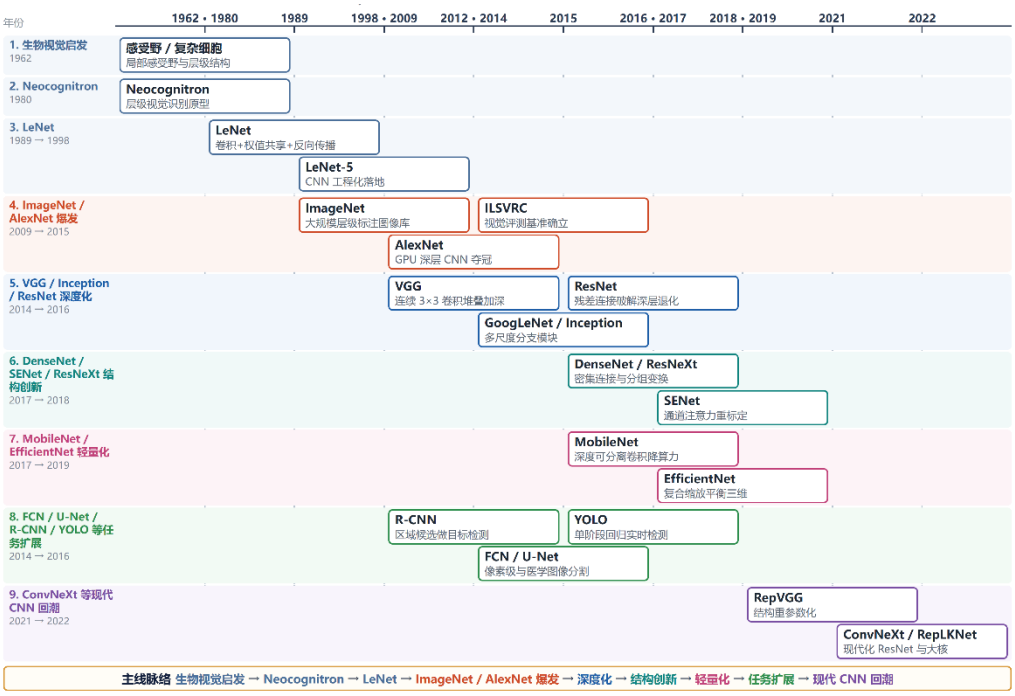
表2：深度学习架构发展七阶段与 CNN 示例

大阶段	发展阶段	阶段说明	CNN 发展过程举例
范式确立	1.思想提出	提出架构核心思想与启发来源，尚未形成可计算模型	生物视觉启发（Hubel&Wiesel，1962）：感受野、简单/复杂细胞
范式确立	2.基础架构	将思想转化为可计算、可训练的基础网络结构	Neocognitron（1980）、LeNet（1989/1998）建立卷积—池化—反向传播原型
成长与扩展	3.训练和优化赋能	依靠算力、激活函数、正则化等使深层网络真正可训练	AlexNet（2012）：GPU+ReLU+Dropout+数据增强
成长与扩展	4.结构精细化改进	在可训练基础上优化深度、连接方式与特征流动	VGG/Inception/ResNet 深度化；DenseNet/SENet 结构创新
成长与扩展	5.任务扩展	从单一任务扩展到检测、分割等多种下游任务	FCN/U-Net 语义分割；R-CNN/YOLO 目标检测
成熟与高峰	6.轻量化与部署	面向移动端与实际部署，平衡精度—参数量—速度	MobileNet（2017）、EfficientNet（2019）
成熟与高峰	7.复兴性改进	在新一代架构冲击下借鉴新方法重新现代化	ConvNeXt（2022）、RepVGG（2021）、RepLKNet（2022）

数据来源：《ImageNet Classification with Deep Convolutional Neural Networks》，《Deep Residual Learning for Image Recognition》等其他文献，东吴证券研究所

CNN 高引用的基础架构与训练类论文集中在 2012—2016 年，2017 年后论文重心明显向轻量化/部署与复兴/后继迁移、引用量整体回落，意味着 CNN 在 2016—2017 年前后进入短期学术瓶颈：一方面走向成熟并推动人脸支付、安防、手机影像等应用密集落地，另一方面 2017 年 Transformer 出现，提示下一代架构成为新的关注重点。

图4：CNN 关键学术成果与技术演进时间轴



数据来源：《ImageNet Classification with Deep Convolutional Neural Networks》，《Deep Residual Learning for Image Recognition》等其他文献，东吴证券研究所

虽然 AI1.0 时代的技术出现了一定突破，但在产品落地方面仍有局限。我们认为，AI1.0 使其模型能力主要停留在“单点感知能力提升”，尚未形成通用智能系统所需的理解、推理、交互和行动能力。因此，其输出通常需嵌入既有业务流程才能产生价值，难以独立成为用户反复使用的交互入口，天然限制 0→1 新品类的创造。具体则表现为落地场景过于垂直或附属于已有产品，例如安防中的识别场景或人脸识别支付场景。

2.2. AI 第二次爆发：从人工设计到规模化学习，Transformer 成为通用底座

AI 第二次爆发的标志，是 Transformer 成为通用的数字任务底座；语言、视觉、生成与多模态四个典型任务领域的技术路径革新各不相同，但在发展过程中均把 Transformer 作为其技术底座之一。

AI2.0 的本质，是能力来源由人工设计的强任务先验转向规模化学习。AI1.0 时代的主流架构具有较强的归纳偏置，如 CNN、RNN；CNN 绑定局部性与平移等变的强空间先验，RNN 绑定递归与时序依赖的强序列先验，二者样本效率较高但通用性与可扩展性受限；Transformer 以“注意力+位置编码”承载较弱先验，更多依赖数据、算力与训练目标形成能力，因而可并行、可充分利用大算力与大数据。

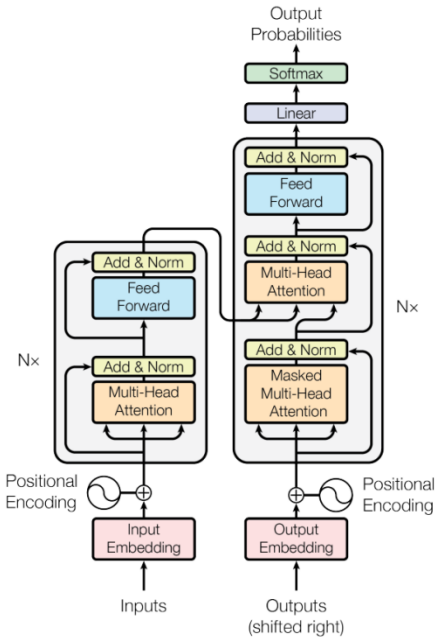
表3：主流架构的归纳偏置

架构	归纳偏置（先验）	代表能力	与算力/数据的关系
卷积网络（CNN）	强空间先验：局部性、平移等变	图像等感知任务	先验强、样本效率高，但通用性与上限受限
循环网络（RNN）	强序列先验：递归、时序依赖	语言等序列建模	顺序计算难并行，长程依赖建模偏弱
Transformer	弱先验：注意力+位置编码	跨模态通用建模	可并行、可充分利用大算力与大数据

数据来源：数据来源：《ImageNet Classification with Deep Convolutional Neural Networks》，《Attention Is All You Need》等其他文献，东吴证券研究所

Transformer 成为数字 AI 通用底座，可归纳为三重机制。其一是表示与接口趋同：文本天然可表示为 token 序列，代码、工具调用与软件状态能够被序列化，ViT 将图像划分为 patch token，音频与视频经编码进入统一上下文，token 化与多模态表示使不同模态具备共享模型接口的前提。其二是规模化训练与后训练叠加：自注意力允许序列位置直接交互、缓解递归结构的串行计算与长距离依赖问题，自监督预训练与 Scaling Law 建立规模扩展框架，SFT、RLHF、DPO 提升可用性，RAG、工具调用与测试时计算将能力增量延伸至运行时系统。其三是交互与任务收敛：自然语言逐步成为调用文本生成、代码、知识问答、工具与多模态能力的统一入口。

图5：Transformer 结构示意

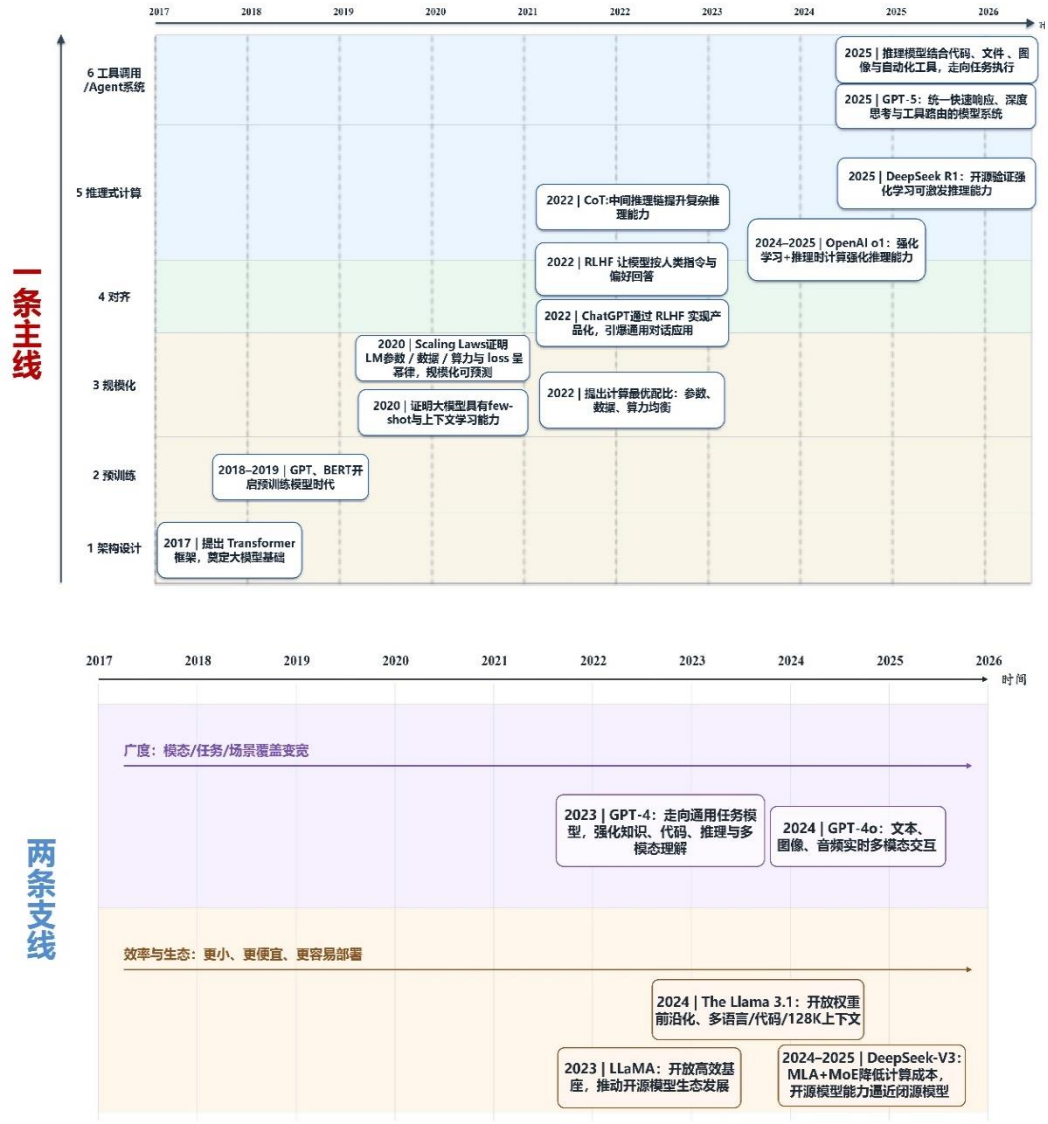


数据来源：《Attention Is All You Need》，东吴证券研究所

结合核心论文，我们认为大模型演进框架可以总结为：一条核心发展主线搭配两条并行优化支线。主线为模型能力升级重心持续向外转移，分为三个递进阶段；两条支线

分别对应能力拓展广度、训练推理使用效率。该划分标准区分能力优化所处层级，分为架构设计、预训练、规模扩张、对齐调优、推理增强、智能体系统等各类技术进展。

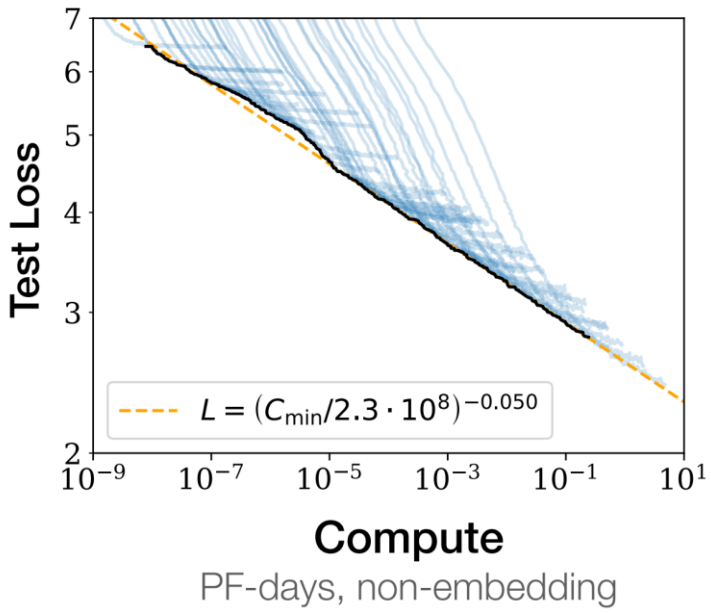
图6：Transformer 时代重要论文/成果发展时间轴



数据来源：《Attention Is All You Need》，《BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding》等其他文献，东吴证券研究所

扩展规律（scaling law）回答的是“规模如何转化为能力”。这一规律伴随大模型发展不断被修正与拓展，逐步形成覆盖训练、数据、对齐、推理与多模态的全链条框架。

图7：随着训练计算量增加，Loss 按幂律持续下降。



数据来源：《Scaling Laws for Neural Language Models》，东吴证券研究所绘制

大模型 Scaling 规律经历了从“规模扩展有效”到“资源配置优化”再到“推理阶段扩展”的演进过程：早期研究验证模型性能随数据、参数和计算量增加呈幂律提升；GPT-3 时期 Scaling Law 正式确立，证明大规模预训练可持续提升模型能力；随后研究由训练 Loss 转向下游能力表现，并探索模型涌现现象；Chinchilla 进一步提出计算最优 Scaling，强调参数与 Token 的协同增长；随着数据资源约束增强，数据质量与训练流程成为新的优化方向；近年来，Scaling 边界进一步拓展至推理时计算，通过强化学习和额外推理算力提升复杂任务能力。

表4：扩展规律（Scaling Law）的六阶段演进

阶段	时间	阶段名称	核心进展	代表成果
阶段 1	2017—2019	经验缩放规律前史：性能可预测化	跨任务发现泛化误差随数据规模呈幂律下降，模型性能开始具备跨规模外推能力	Hestness 等（2017）
阶段 2	2020	预训练 Scaling Law 确立	系统建立 Loss 与参数、数据、训练计算量之间的幂律关系；GPT-3 展示规模扩展带来的 Few-shot 能力	Kaplan 等、GPT-3、Henighan 等
阶段 3	2021—2022	从 Loss 到能力：能力 Scaling 与涌现提出	研究重心由预训练 Loss 扩展至下游任务表现，提出部分能力可能在规模阈值后集中出现	Gopher、PaLM、Wei 等
阶段 4	2022—2023	计算最优与推理经济性	Chinchilla 提出固定算力下参数与 Token 应共同扩展；LLaMA 进一步考虑部署成本，以较小模型配合更多训练 Token	Chinchilla、LLaMA；Schaeffer 等涌现争论
阶段 5	2023—2024	数据约束与数据质量 Scaling	高质量独立 Token 逐渐成为瓶颈，数据重复边际收益下降，过滤、去重、配比和数据质量的重要性提升	Muennighoff 等、FineWeb、DCLM
阶段 6	2024—至今	推理时计算与强化学习推理 Scaling	Scaling 由预训练延伸至强化学习训练和推理阶段，模型可通过更长思考、搜索、验证和动态算力分配提升复杂推理能力	o1、Snell 等、DeepSeek-R1

数据来源：《Scaling Laws for Neural Language Models》，《Language Models are Few-Shot Learners》等其他文献，东吴证券研究所

2.3. 语言智能：Transformer 诞生的原生应用场景

自然语言处理任务是 Transformer 架构的原生场景。在 Transformer 之前，循环网络以隐藏状态承载上下文，通过 LSTM/GRU 门控缓解梯度消失、经 Seq2Seq 与 Attention-RNN 突破固定向量瓶颈，定义了语言任务的能力上限，但其信息交互沿时间步递归、难以并行，两位置关联路径随距离增长。

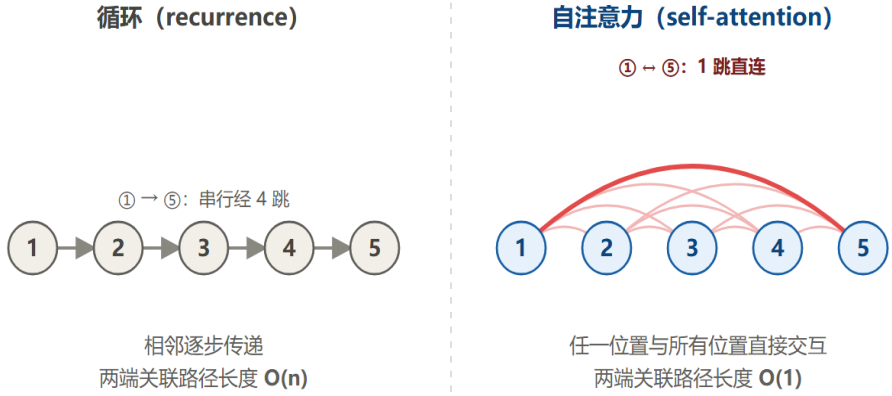
表5：Transformer 与循环网络的多维对比

对比维度	循环网络（含 LSTM/GRU）	Transformer
信息交互方式	沿时间步递归	自注意力，全局直接交互
训练并行性	串行，难以并行	全序列并行
两位置关联路径	O(n)，随距离增长	O(1)，常数级
计算复杂度	O(n)，随长度线性	$O(n^2)$，随长度平方
顺序信息	结构自带	需位置编码
归纳偏置	强（顺序假设）	弱（依赖数据学习）
可扩展性	受串行计算限制	利于规模扩展

数据来源：《Attention Is All You Need》，《Long Short-Term Memory》，东吴证券研究所

Transformer 以自注意力取代循环，将序列建模由递归状态传递转向全局信息交互，打开并行训练、规模扩展与任务统一空间；预训练、指令微调、RLHF、CoT、RAG、工具调用与测试时计算沿同一主干逐层叠加，使语言模型从特征提取器和专用任务模型走向通用交互入口。需要说明的是，循环与状态压缩思想并未消失，仍在端侧、流式、长上下文与成本敏感场景延续。

图8：自注意力与循环结构的信息交互方式对比



数据来源：《Attention Is All You Need》，《Long Short-Term Memory》，东吴证券研究所

基于相关论文与代表成果复盘，我们将 Transformer 在语言任务中的演化归纳为七个阶段，并进一步归纳为三个大阶段。三个大阶段分别为：架构确立，即自注意力替代递归，确立可并行的序列建模架构；范式统一与规模化，即预训练—微调成为主流，理解与生成任务统一为文本到文本，自回归模型经规模扩展获得通用能力；对齐与推理，即指令微调与人类偏好对齐提升模型可用性，复杂推理、工具调用与测试时计算进一步拓展能力边界。

图9：Transformer 在语言任务中的技术演进时间轴（2014—2025）

数据来源：《Attention Is All You Need》，《BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding》等其他文献，东吴证券研究所

2.4. 视觉智能：正在从 CNN 主导转向 Transformer 主导

在视觉基础模型领域，基于 Transformer 形成的 ViT 及其变体架构已经成为主流路线之一。CNN 凭借局部性、权重共享与平移等变三项视觉归纳偏置，在 Transformer 进入视觉之前已把图像识别、检测与分割推进至较高工程水平。ViT 确立“图像 patch 即 token”的表示方式，CLIP、DINO 与 SAM 等推动视觉基础模型形成；与此同时，Swin、PVT 等重新吸收局部性、层级与多尺度先验，ConvNeXt 等现代 CNN 亦吸收 Transformer 时代的训练经验。

表6：Transformer 与卷积网络的多维对比

对比维度	卷积网络（CNN）	视觉 Transformer（以 ViT 为代表）
空间信息交互	局部卷积，逐层扩大感受野	自注意力，全局直接交互
长距离依赖	需堆叠多层，路径随距离增长	O(1)常数级路径
归纳偏置	强（局部性、权重共享、平移等变）	弱（依赖数据学习，需位置编码）
空间先验来源	架构内置	依赖位置编码与大规模预训练
小数据表现	较好	较弱，依赖大规模预训练
跨任务/跨模态	任务专用，迁移有限	接口统一，利于迁移与多模态对齐

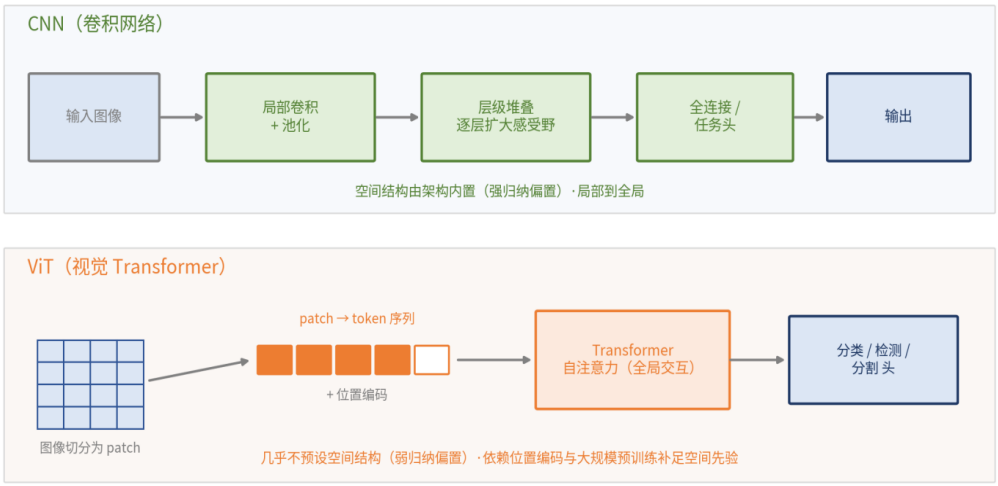
数据来源：《An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale》，《ImageNet Classification with Deep Convolutional Neural Networks》等其他文献，东吴证券研究所

综合来看，视觉领域任务已多以 Transformer 为技术基座：ViT 带来统一接口与规模化潜力，卷积思想则以现代化形式回归。视觉能力本身仍以感知型中间能力为主，是否作为独立终端产品仍取决于与下游系统的结合方式。

从技术角度看，Transformer 通过“图像 patch 即 token”的思路实现了从序列任务

向图像任务的切入。CNN 通过卷积操作利用局部连接和权重共享，将空间结构先验直接编码进网络；ViT 则将图像切分为 Token，通过自注意力机制实现全局信息交互，更多依靠数据学习图像中的空间关系。

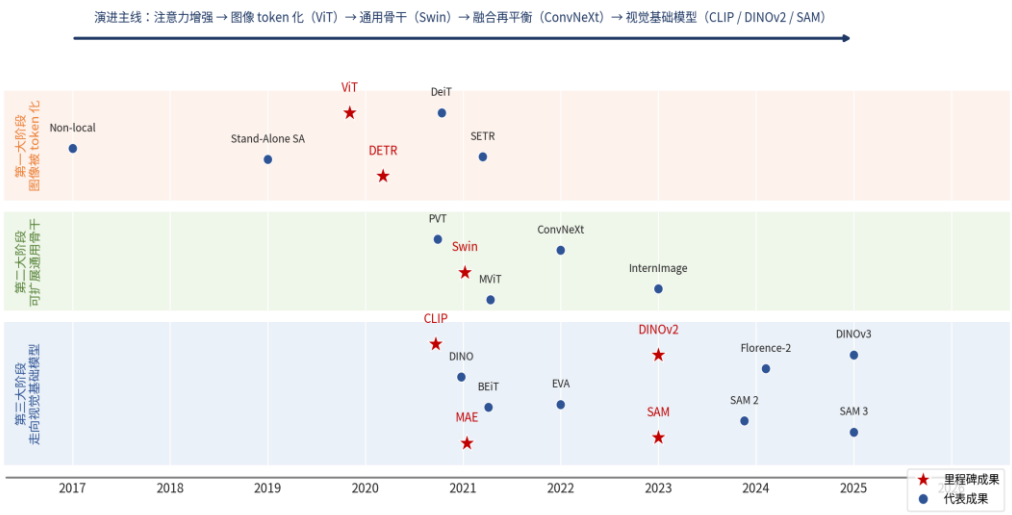
图10：卷积网络（CNN）与视觉 Transformer（ViT）的结构对比



数据来源：《An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale》，《ImageNet Classification with Deep Convolutional Neural Networks》等其他文献，东吴证券研究所

基于相关论文与代表成果复盘，我们将 Transformer 在视觉任务中的演化归纳为三个阶段：1）图像被 token 化，Attention 机制首先作为 CNN 增强模块引入视觉任务，随后 ViT 将图像 Patch 转化为 Token 序列，DETR 进一步引入 Object Query 重构检测任务；2）视觉 Transformer 走向通用骨干，Transformer 逐步融合 CNN 的局部性、层级结构和多尺度能力，成为适用于分类、检测和分割等密集任务的视觉架构；3）视觉模型由封闭标签分类迈向视觉基础模型，自监督预训练、视觉语言对齐和可提示分割推动视觉模型具备可预训练、可迁移、可提示和多模态融合能力。

图11：Transformer 在视觉任务中的技术演进时间轴



数据来源：《An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale》，《Swin Transformer: Hierarchical Vision Transformer using Shifted Windows》等其他文献，东吴证券研究所

我们认为，Transformer 之于 CNN 的核心意义并非仅在于提升视觉任务精度，更在于通过 Token 化表示和自注意力机制，构建了一种统一、可扩展、能够建模全局关系的信息处理框架，使视觉模型能够接入大规模预训练、跨任务迁移以及多模态语义对齐。

2.5. 生成智能：Transformer 伴随生成范式演进逐步渗透

与语言和视觉理解任务中 Transformer 快速成为主流骨干不同，生成式视觉由于涉及表示方式、生成机制和条件控制等多个环节，其技术范式持续迭代。**因此，Transformer 并非一次性替代某一固定模块，而是伴随自回归生成、扩散生成和多模态生成的发展，逐步从 Token 建模、条件交互向生成骨干和统一架构演进。Transformer 是在多个生成范式中逐渐成为统一计算骨干。**

第一阶段（2020—2021 年），**Transformer 首先进入图像 Token 生成环节，推动视觉生成由连续像素建模转向离散序列建模**。以 VQ-VAE/VQGAN 和 DALL-E 为代表的方法，通过将图像压缩为离散 Token，并利用 Transformer 进行自回归预测，使图像生成具备类似语言模型的序列建模能力，为后续大规模视觉预训练奠定基础。

第二阶段（2021—2022 年），**Transformer 进一步参与文本条件控制，实现视觉生成与语言语义的结合**。CLIP 等视觉语言模型通过构建图文共享语义空间，为生成模型提供更强的文本理解能力；Imagen、Stable Diffusion 等扩散模型则通过文本编码器与 Cross-Attention 机制，将语言条件注入去噪过程，使文本驱动图像生成成为主流路径。

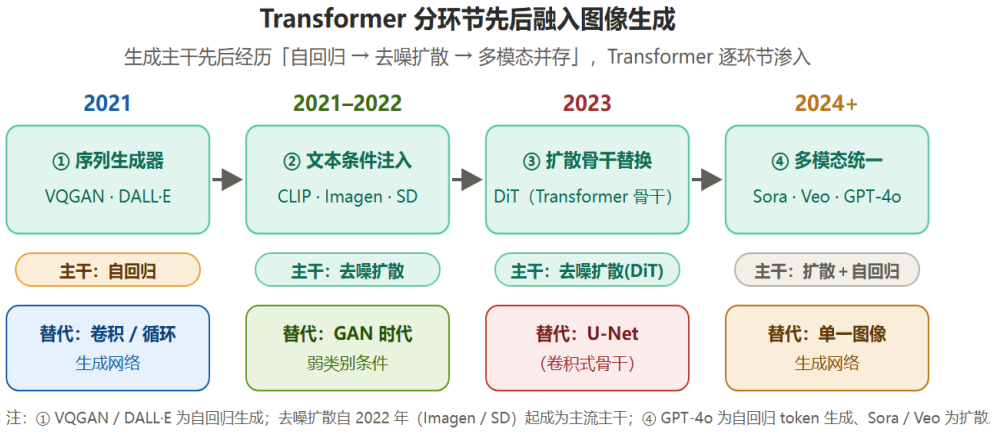
第三阶段（2022—2023 年），**Transformer 开始从条件模块进入生成核心，逐步替**

代传统卷积式生成骨干。DiT（Diffusion Transformer）将 Transformer 引入扩散模型主干网络，以全局注意力替代 U-Net 中的卷积操作，证明 Transformer 具备处理高分辨率视觉生成任务的扩展潜力，并推动扩散模型进一步向大规模 Scaling 发展。

第四阶段（2024 年至今），**生成式视觉进入多模态统一建模阶段，Transformer 成为连接文本、图像和视频等不同模态的重要基础架构**。以 Sora、Veo、Emu 等为代表的新一代模型，通过将不同模态信息统一 Token 化，并结合扩散模型或自回归生成方式，实现跨模态理解与生成能力融合，推动视觉生成从单一图像合成向通用世界建模方向演进。

总体来看，Transformer 在生成式视觉中的作用并非简单替代 GAN 或扩散模型，而是逐步成为生成系统中的通用计算框架：**从最初负责 Token 序列建模，到承担条件交互，再到替代生成骨干并实现多模态统一。**

图 12：Transformer 融入图像生成历程时间轴



数据来源：《Taming Transformers for High-Resolution Image Synthesis》，《Scalable Diffusion Models with Transformers》等其他文献，东吴证券研究所

2.6. 多模态智能：Transformer 从外挂对齐到统一骨干

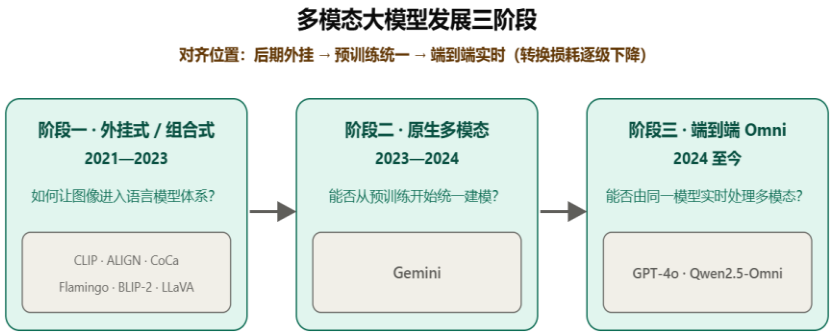
与语言、视觉理解中 Transformer 较快确立为主流骨干不同，多模态的核心难点在于多种异质模态的对齐与融合。Transformer 对多模态的融入并非一次性替换某一模块，而是沿"后期外挂 → 预训练统一 → 端到端实时"把模态对齐逐步搬进同一计算骨干，转换损耗随之逐级下降，可分为三个阶段。

外挂式阶段（2021—2023），Transformer 充当各模态编码器与外挂桥。各模态先各自编码（ViT 编图、Transformer 编文），CLIP 以对比学习将双编码器对齐至同一表示空间；再经外挂模块注入冻结的大语言模型——Flamingo 用门控交叉注意力、BLIP-2 用 Q-Former、LLaVA 用线性投影，将视觉 token 送入文本 Transformer 的上下文。对齐发生在后期，视觉与语言两套模型基本各自训练后再拼接，接口有损、转换损耗最大。

原生多模态阶段(2023—2024),Transformer 充当统一预训练骨干。不再先分后拼,而是自预训练起即把文本、图像乃至音视频 token 交错为单一序列、输入同一 Transformer,跨模态关系由自注意力在预训练中一并习得。对齐前移至预训练,各模态共享同一注意力堆栈,转换损耗下降。

端到端 Omni 阶段(2024 年至今),Transformer 充当 any-to-any 输入输出骨干并承担时空建模。单一模型实时完成多模态的输入与输出:音频不再经"识别—大模型—合成"三段式,而是 token 化后由同一模型直接读写;视频以时空 token 序列被统一建模。对齐完全内化于单一模型,转换损耗最低、时延最小。

图13:多模态大模型发展三阶段:对齐位置前移、转换损耗下降



数据来源:《Learning Transferable Visual Models From Natural Language Supervision》,《Flamingo: a Visual Language Model for Few-Shot Learning》等其他文献,东吴证券研究所

2.7. 从分立到统一:Transformer 收敛为多任务骨干

总体来看,Transformer 的注意力机制提供了跨领域通用的关系建模方式,在叠加 Scaling 规律后,统一模型在能力迁移、研发效率和工程复用上的优势持续扩大,使通用架构的模式逐渐优于分别设计专用架构的模式,进而推动架构、任务与模态向统一基础模型范式收敛。

我们认为,从文本领域发展到多模态领域,Transformer 的角色正由早期的"能力跃升来源",逐步转变为跨领域通用的计算底座,并开始呈现出一定的组件化特征。

表7：Transformer 从语言建模扩展至视觉、生成和多模态任务

任务领域	核心难点	Transformer 融入路径	角色定位
文本	token 天然离散且成序列，几无适配成本	直接替换 RNN/LSTM，快速确立为主流骨干	原生主干
图像理解	图像非天然序列，需先构造 token	经 patch 化（ViT）引入，长期与 CNN 并存	编码骨干之一
生成	涉及表示、生成机制、条件控制多个环节，范式持续迭代	分环节渗透：序列生成 → 条件注入 → DiT 替换 U-Net → 统一架构；骨干与生成机制（扩散/自回归）解耦	机制无关的计算骨干
多模态	多种异质模态的对齐与融合	对齐位置前移：后期外挂 → 预训练统一 → 端到端；转换损耗逐级下降	跨模态统一骨干

数据来源：《Learning Transferable Visual Models From Natural Language Supervision》，《Flamingo: a Visual Language Model for Few-Shot Learning》等其他文献，东吴证券研究所

3. 数字 AI 提供能力底座，物理闭环决定迁移边界

Transformer 作为通用模型架构，既可用于数字 AI，也可应用于物理 AI。数字 AI 凭借对 Transformer 的应用，已在语言、视觉与多模态任务中跑通通用底座模型。

那么对于物理 AI 场景，数字 AI 的能力能够外溢到哪里，边界由什么决定。我们以智能驾驶为样本梳理借鉴史实，说明能力外溢已经发生、且借鉴层级由架构理念上移至模型迁移。

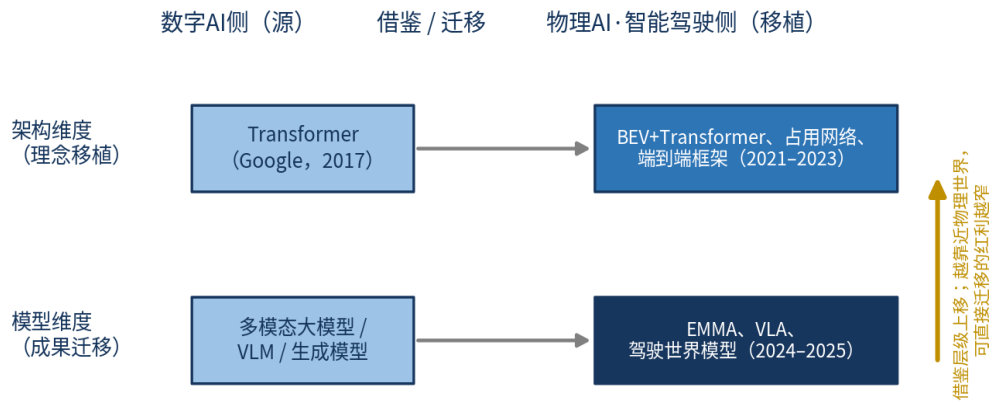
3.1. 借鉴层级由架构理念上移至模型迁移

从产业与学术实践看，智能驾驶对数字 AI 的借鉴大体经历两个阶段。第一阶段以架构理念移植为主：Transformer 确立的注意力机制与序列建模思想被引入驾驶感知与轨迹预测，催生了矢量化表示、鸟瞰图（BEV）与 Transformer 结合的统一空间表征，以及以查询解码为核心的端到端一体化框架 UniAD。这一阶段迁移的是建模范式，而非具体模型权重。

第二阶段进一步表现为模型层面的直接迁移。代表性工作是 Waymo 于 2024 年提出的端到端多模态驾驶模型 EMMA（End-to-End Multimodal Model for Autonomous driving），其论文明确说明该模型构建在谷歌多模态大模型 Gemini 之上，将原始摄像头输入与文本指令映射到统一语言空间，再输出轨迹、目标检测与道路图等结果。与此同时，源自机器人领域的视觉—语言—动作（VLA，Vision-Language-Action）范式被引入车端，例如理想汽车 MindVLA（2025 年发布）以及小鹏汽车提出的约 720 亿参数世界基座模型。

我们认为，沿着从架构理念到具体模型的借鉴层级上移，可直接复用的成果在形态上越来越完整，但其与物理世界的耦合也越来越紧密。文本、代码等数字信号与驾驶所需的三维时空信号之间存在结构性差异；越接近车辆控制、实时三维感知与安全攸关决策环节，数字 AI 已有成果可直接迁移的比例越低，需要针对物理世界重新设计或重新训练的部分越多。借鉴层级在上移，但越靠近物理闭环，可直接迁移的红利区间可能趋于收窄，尚无统一量化标准。

图14：数字 AI 能力向物理 AI 的迁移层级



数据来源：Attention Is All You Need》，《A Survey on Vision-Language-Action Models for Autonomous Driving》等其他文献，东吴证券研究所

表8：数字 AI 关键成果向智能驾驶的迁移路径

数字 AI 侧成果（提出方/年份）	智能驾驶侧迁移对应	代表工作（年份）	借鉴层级
Transformer（Google，2017）	统一 BEV 空间表征	特斯拉 BEV 方案（2021）、BEVFormer（2022）	架构维度
Transformer 查询解码机制	端到端一体化框架	UniAD（CVPR 2023）	架构维度
多模态大模型 Gemini（Google，2024）	端到端多模态驾驶模型	Waymo EMMA（2024）	模型维度
VLA 范式（源自机器人领域）	车端 VLA 驾驶模型	理想 MindVLA、小鹏世界基座模型（2025）	模型维度
生成式/扩散模型	驾驶世界模型	Wayve GAIA-1（2023）、英伟达 Cosmos（2025）	模型维度

数据来源：《Attention Is All You Need》，《Planning-oriented Autonomous Driving》等其他文献，东吴证券研究所

3.2. 表征、数据与学习范式存在系统差异

在共享底层技术的同时，物理 AI 与数字 AI 存在的系统性差异导致其存在一定的技术隔阂，具体表现在输入表征、数据获取与训练范式三个方面。

其一是表征差异：数字 AI 处理的文本天然是离散的、带语义的符号 Token，输入端已具备较强结构化基础；物理世界的原始输入则是连续的光子、点云等信号，并不存在天然、通用、可被模型直接处理的符号化方案，智能驾驶因此多出一段数字 AI 无需经历的历程——将三维物理空间构建为可学习的模型底座。

其二是数据差异：数字 AI 的训练语料已大量沉淀在互联网中，规模化的核心动作是抓取、清洗与压缩已有数据；物理 AI 最具价值的数据恰恰是长尾、极端与事故场景，天然稀疏、采集成本高且可能涉及安全风险。兰德公司（RAND，2016）的情景测算说明了路测里程的量级问题：若要在 95%置信度、±20%精度下证明自动驾驶致死率达标，约需 88 亿英里，以百辆车规模的车队计算需要数百年；为缓解这一问题，业界发展出以世界模型生成合成数据、并将其定位为数据工厂的路径。

其三是学习范式差异：数字 AI 的预训练以下一词元预测为代表，其模仿对象是整个互联网上的人类知识，模仿可覆盖的边界较宽，强化学习与 RLHF 更多承担对齐与推理增强角色；物理 AI 以 L4 级自动驾驶为代表，所模仿的对象通常是平均水平的人类司机，而安全性目标却要显著超越人类。我们认为，当模仿对象为平均水平人类司机、而目标是大幅超越人类安全性时，单纯模仿学习的上限相对更低，强化学习更接近突破能力边界的关键路径。

表9：数字 AI 与物理 AI 在表征、数据与学习范式上的三重差异

维度	数字 AI（以文本/大模型为主）	物理 AI（以智能驾驶为例）
表征	文本天然离散、带语义，输入端已结构化，Token 化轻量	缺乏天然通用 Token，需先把连续三维空间构建为可学习底座
数据	高质量语料已沉淀于互联网，核心是抓取、清洗与压缩存量	高价值长尾、极端与事故场景天然稀疏、昂贵且具风险，需制造增量
学习范式	预训练模仿互联网级人类知识，强化学习承担对齐与推理增强	模仿对象为平均人类司机、目标远超人类，强化学习成为关键环节

数据来源：《DeepSeek-R1 Incentivizes Reasoning in LLMs through Reinforcement Learning》（Nature 2025），《Will We Run Out of Data? Limits of LLM Scaling Based on Human-Generated Data》等其他文献，东吴证券研究所

4. 物理 AI：从能力外溢到真实世界闭环

物理 AI（Physical AI）通常指能够感知物理世界、并在三维空间中完成决策与行动的人工智能系统，自动驾驶是其最具代表性的落地场景之一。英伟达创始人黄仁勋在

2025 年国际消费电子展（CES 2025）主题演讲中，将人工智能发展划分为感知 AI、生成式 AI、智能体 AI 与物理 AI 四个阶段，并提出通用机器人的 ChatGPT 时刻临近。数字 AI 以大语言模型为代表，主要处理文本、代码、图像等数字信号；物理 AI 则需要在真实三维空间中感知、决策和行动。

物理 AI 的研究以智能驾驶为切入点，沿一条主线依次展开：数字 AI 能力向物理 AI 的迁移及其边界；VLA 的动作生成范式；物理世界表征的演进；世界模型作为数据工厂的角色；从模仿学习到强化学习的范式变化；云端训练与车端部署之间的时延分工。六个议题看似分散，实则共同回答同一个问题，即物理 AI 如何把感知、状态、动作、反馈与部署放入同一真实世界闭环。

我们认为，物理 AI 的核心并非单点提升视觉或语言模型的能力，而是围绕“可执行、可反馈、可验证”重建一整套技术体系。数字 AI 已经证明，统一的模型架构能够在语言、视觉与多模态理解任务中形成广泛迁移能力；物理 AI 沿用这一能力底座，却不止于把已有模型部署到车辆或机器人终端，而需要让模型在连续变化的环境中完成状态理解、行动生成、结果反馈与策略更新，并在安全、实时与算力约束下稳定运行。各环节并非相互独立，而是共同定义了物理 AI 从能力展示走向可执行系统的路径。

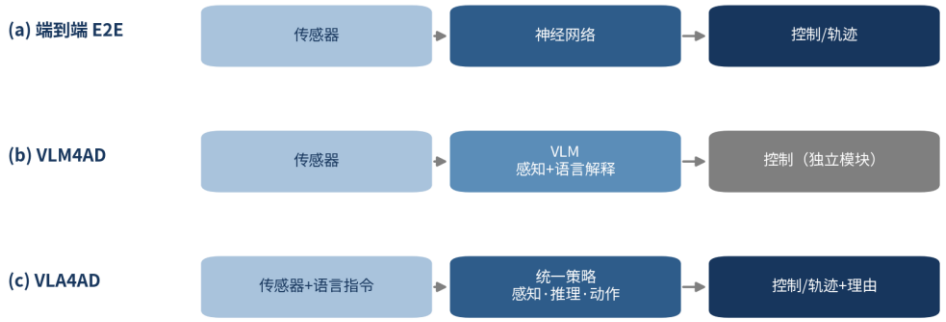
4.1. VLA：从端到端到动作生成的范式升级

4.1.1. 自动驾驶由模块化流水线走向统一策略

自动驾驶的整体技术演进可划分为四个范式。早期主流方案为经典模块化流水线，将任务拆分为感知、预测、规划、控制四个模块，其优势在于各组件可独立开发、测试与优化，缺陷是模块间误差会级联传播、信息碎片化，难以处理长尾极端场景。为解决模块化缺陷，研究转向端到端自动驾驶，将原始传感器输入直接映射为控制指令，消除了模块边界、可端到端优化，但在语义稳定性、可解释性与语言交互能力上仍存在不足。

在此基础上，研究进一步将视觉语言模型引入自动驾驶（VLM4AD），把感知与自然语言推理统一到同一嵌入空间，利用大模型常识先验提升对罕见场景的泛化能力；但该路线仍以感知解释为核心，语言输出与控制耦合较弱，存在“能解释场景、却难以直接生成可靠控制决策”的动作生成缺口。VLA4AD 范式的关键变化，是把视觉感知、语言理解和控制决策整合到同一策略中，通过显式动作头（action head）弥合动作生成缺口，其目标能力包括遵循自然语言指令、输出决策理由以支持事后验证、以及利用常识先验提升罕见场景泛化。

图15：自动驾驶范式比较



数据来源：《A Survey on VLA Models for Autonomous Driving》，东吴证券研究所绘制

4.1.2. VLA 的三大模块与阶段演进

VLA4AD 在输入侧整合多模态信号：视觉输入常将图像投影为 BEV 表征以强化空间推理，激光雷达提供 3D 结构、毫米波雷达提供速度、IMU 用于运动跟踪、GPS 用于定位，语言输入包括环境查询、任务规则解析与语音交互。其核心模块可拆分为视觉编码器、语言处理器与动作解码器：视觉编码器将异构传感器数据转换为与语言空间对齐的隐层特征（常用 DINOv2、CLIP 等骨干并加入 3D 空间增强）；语言处理器采用预训练大语言模型并以 LoRA 微调或检索增强注入领域知识；动作解码器将融合特征转换为驾驶动作，主要有自回归分词器、扩散头与分层控制器三种技术路线。

VLA4AD 的内部演进可划分为四个阶段。预 VLA 阶段（约 2023 年）多模态大模型更多充当被动解释器，代表如 DriveGPT-4，语言不参与决策；模块化 VLA 阶段（2024—2025 年初）LLM 升级为主动路径规划媒介，采用多阶段流水线、语言作为中间表示，代表如 OpenDriveVLA；统一端到端 VLA 阶段（2025 年）在一次前向传播中直接将多模态传感器流与语言指令映射为控制信号，代表如 Waymo EMMA；推理增强 VLA 阶段将长时域推理、记忆与交互整合进控制循环，代表如引入 QT-Former 记忆模块的 ORION。

图16：VLA4AD 四阶段演进时间轴

数据来源：《A Survey on Vision-Language-Action Models for Autonomous Driving》，《EMMA: End-to-End Multimodal Model for Autonomous Driving》等其他文献，东吴证券研究所

4.1.3. 厂商路径分化与工程边界

头部厂商的 VLA 路线开始分化。Waymo EMMA 构建于 Gemini 之上，以统一语言空间联合处理多任务，引入思维链（CoT）推理后端到端规划性能进一步提升 6.7%，但论文披露其仅能处理少量图像帧、不含激光雷达等精确 3D 感知模态。理想 MindVLA 将三维空间理解、交通语义推理与驾驶动作生成统一到同一框架，采用 3D 高斯作为中间表征并自监督训练，语言模型 MindGPT 采用 MoE 与稀疏注意力，动作侧由 Diffusion 扩散解码器解码轨迹，整个推理在车端实时运行。元戎启行 40B VLA 基座使同一模型同时承担 Driver、Analyst、Critic 三类角色，并通过评价结果反向驱动数据闭环。千里科技 CoWorld-VLA 在 VLM 隐空间构造交通交互、道路几何、动态演化与自车轨迹四类专家 Token，构成面向规划的隐空间世界表征，在 NAVSIM v1 上取得 PDMS 89.8。

小鹏第二代 VLA 去掉语言转译环节，设计原生多模态 Tokenizer 并采用 32 倍超密视觉思维链，实现从视觉信号到动作指令的端到端直接输出，在架构形态上趋近视觉—动作（VA）；作为对照，理想 MindVLA-o1 按公司表述保留语言层、是原生多模态 MoE Transformer。由此，在是否保留显式语言推理层这一维度上，头部厂商路线已出现分化。

表10：代表性厂商 VLA 模型的架构与公开指标对比

模型	视觉/输入	语言/推理	动作解码	公开指标（公司/论文口径）
Waymo EMMA	环视相机，纯视觉，不含 LiDAR/雷达	Gemini 统一语言空间，CoT 推理	轨迹文本 token	nuScenes 规划 SOTA；CoT 使端到端规划提升 6.7%
理想 MindVLA	摄像头+LiDAR，3D 高斯中间表征	MindGPT，MoE+稀疏注意力	Diffusion 扩散轨迹	公司口径：车端实时运行
元戎启行 40B VLA	视觉输入为主	Analyst 因果推理（三角色）	实时驾驶动作	数据闭环周期由 5 天以上压缩至约 12 小时
千里 CoWorld-VLA	图像+导航+车辆状态	Qwen3-VL 隐空间四类专家 Token	扩散式 HMEF 规划器	NAVSIM v1 PDMS 89.8；FVD 32.7
小鹏 第二代 VLA	原生多模态 Tokenizer，视觉/语音/文本	32 倍超密视觉思维链，弱化文本语言	端到端直接输出动作	公司口径：较传统 CoT 预测误差降低 33%

数据来源：《EMMA: End-to-End Multimodal Model for Autonomous Driving》，《A Survey on Vision-Language-Action Models for Autonomous Driving》等其他文献（含各公司公开披露），东吴证券研究所

4.2. 表征：物理世界如何进入模型

4.2.1. 物理世界缺乏天然通用 Token

表征决定物理世界以何种形式进入模型。连续光子、点云、运动与空间关系需要先被转换为可学习的数值表示，模型才能进一步完成识别、预测、规划与控制。嵌入（embedding）是其中最常见的实现形式，通过将对象映射至连续向量空间，使语义、几何与关系信息由向量间的距离和方向承载。数字 AI 与物理 AI 的根本差异，首先在于表征获取方式不同。

在文本侧，将输入转为 Token 通常是轻量且可学习的步骤。文本经子词分词（如字节级 BPE）切分为 Token 后查表得到嵌入，分词器训练成本低、词表规模可控。主流大模型的分词方案虽在算法与词表规模上存在差异，但整体仍建立在成熟的子词切分框架之上：DeepSeek-V3 采用字节级 BPE、词表约 12.8 万，阿里 Qwen 系列在 tiktoken 的 cl100k_base 基础上扩充中文、词表约 15.1 万，智谱 GLM 采用字节级 BPE、词表约 15 万，MiniMax-01 词表为 20 万余。文本侧的世界编码主要由人类语言预先完成，模型能力学习主要发生在分词之后。

物理世界输入没有天然离散语义单元，必须由网络学习如何切分与编码。即便是文本大模型，一旦转向多模态，也必须额外引入视觉编码器，将连续像素压缩为模型可处理的视觉 Token；以 Qwen3-VL 为例，其通过 SigLIP2 视觉编码器提取图像特征。图像 Patch 与视觉 Token 主要覆盖二维视觉输入，智能驾驶还需进一步表示三维空间结构、连续时间变化、交互与动作后果，因此车端表征网络的工程复杂度显著高于一般图文多模态模型，BEV、Occupancy、VAE 与 3D 编码器由此成为物理 AI 输入链路中的核心组

成部分。

表11：数字 AI 与物理 AI 表征问题对照

维度	数字 AI（以文本为主）	物理 AI（以智能驾驶为例）
原始输入	离散符号序列	连续光子/点云等原始信号
是否有天然通用 Token	有，语言已离散、带语义	无天然通用方案，需网络学出离散化/编码
Token 化代价	轻量（子词 BPE，可学但廉价）	重（BEV/占用/VAE/3D 编码器等重网络）
编码器与离散化关系	分词与嵌入分步、边界清晰	编码器直接产出嵌入，二者无清晰边界
世界编码来源	人类语言事先完成	需模型自行构建物理世界表征

数据来源：《DeepSeek-V3 Technical Report》，《VectorNet: Encoding HD Maps and Agent Dynamics from Vectorized Representation》，东吴证券研究所

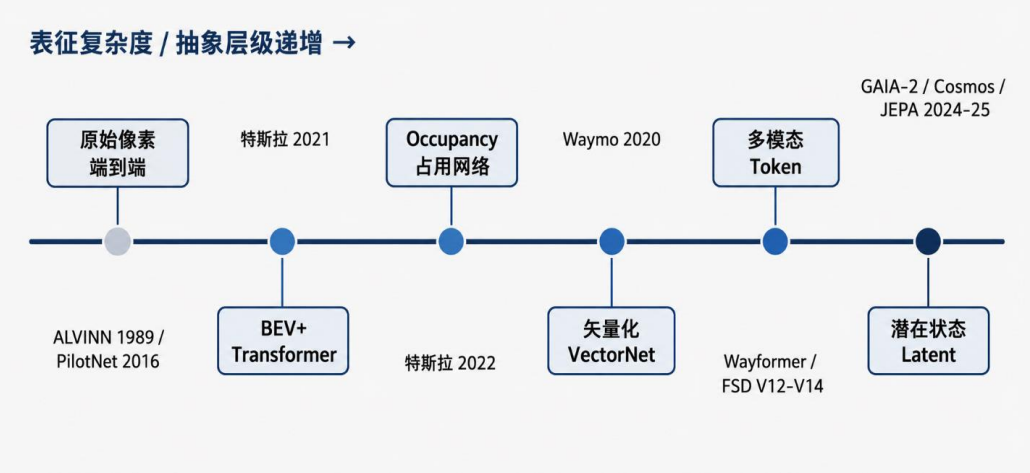
4.2.2. 智能驾驶表征由像素走向潜在状态

我们将智能驾驶表征的演进归纳为一条主线：原始像素→BEV→Occupancy 占用→矢量化→多模态 Token→潜在状态，各阶段在产业中长期并存、相互叠加。贯穿演进过程的主线，是上一阶段的表征约束推动下一阶段表征形成。

第一阶段以原始像素直接回归控制量，感知与决策被耦合在隐式表征中，ALVINN 与英伟达 PilotNet 等可在缺少车道线显式标签的情况下自主形成敏感特征，但内部表征难以检视，闭环运行中易受协变量偏移与误差累积影响。**第二阶段 BEV 结合 Transformer，将多相机图像统一至三维鸟瞰空间**：特斯拉 2021 AI Day 以鸟瞰空间位置编码为查询、各相机特征为键值，由自注意力学习向量空间取信息，缓解跨相机目标错位与外参波动；小鹏 XNet 1.0 为国内首个量产的动态 BEV 感知架构。

第三阶段 Occupancy 以体素占据描述空间，把感知对象由预定义实体扩展至完整空间占据，并预测占据状态与速度信息，特斯拉 2022 AI Day 将其落到车端，华为 ADS 2.0、理想 AD Max 3.0 亦相继纳入量产感知栈。第四阶段矢量化将地图、轨迹与道路参与者抽象为矢量对象，并通过图网络与注意力建模交互，VectorNet 相对栅格化基线将计算量降低约一个数量级、参数量节省 70%以上。第五阶段多模态 Token 尝试将视觉、语言与动作纳入统一序列建模，特斯拉 FSD、Wayve LINGO 等体现视觉—语言—动作逐步融合。第六阶段潜在状态进一步走向可预测空间，GAIA、Cosmos 等世界模型将视频与场景压缩至离散 Token 或连续潜在空间，并在历史图像、文本与动作条件下预测未来状态。

图17：智能驾驶表征演进主线



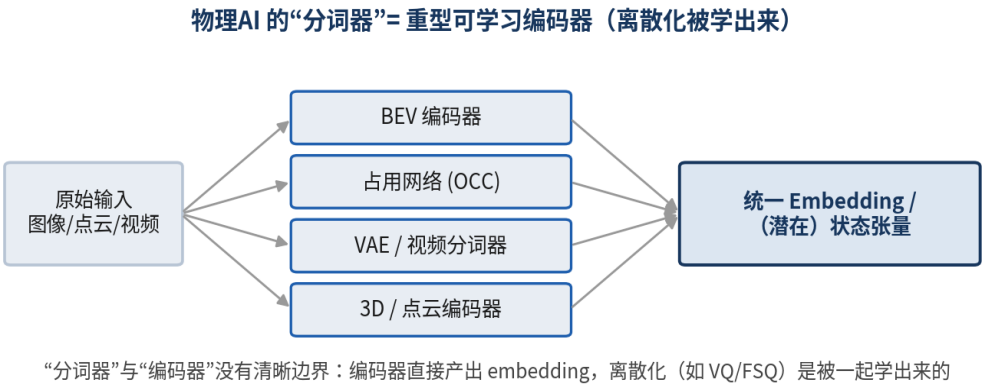
数据来源：《VectorNet: Encoding HD Maps and Agent Dynamics from Vectorized Representation》，《BEVFormer: Learning Bird's-Eye-View Representation from Multi-Camera Images via Spatiotemporal Transformers》等其他文献，东吴证券研究所

4.2.3. 世界模型承担自监督表征训练工具

物理世界标注昂贵且难以覆盖长尾，面向物理世界的嵌入更多需要依靠自监督学习形成。世界模型通过预测未来，为表征学习提供海量自监督信号：无论预测对象是像素还是潜在状态，模型都可在无需人工标注的情况下学习动态内部表征。Yann LeCun 自 2022 年起倡导的 JEPA 路线（Meta 的 I-JEPA、V-JEPA、V-JEPA 2，以及面向点云的 Point-JEPA）主张在潜在/嵌入空间而非像素空间做预测，从而学习更抽象、更稳健的表征。

因此，物理 AI 中的“分词器”并非文本侧那样轻量的查表步骤，而是 BEV、占用网络、VAE 与 3D 编码器等重型可学习网络。竞争壁垒正从单个感知模块的精度，前移到如何把物理世界嵌入统一、可预测的潜在空间；在该方向上，已出现把推理过程放到潜在空间而非文本空间的尝试。

图18：物理 AI 的分词器——BEV、占用、VAE 与 3D 编码器等重型网络



数据来源：《VectorNet: Encoding HD Maps and Agent Dynamics from Vectorized Representation》，《Self-Supervised Learning from Images with a Joint-Embedding Predictive Architecture》，东吴证券研究所

4.3. 数据：从复现场景到制造场景

4.3.1. 物理 AI 的数据范式由抓取已有转向制造所需

物理 AI 与数字 AI 训练处于两种不同的数据供给环境。数字 AI 的训练红利建立在文本、代码与图像等语料早已存在于互联网这一前提上，获取方式是对公开来源进行抓取、过滤与压缩；Epoch AI 关于数据规模的研究估计公开人类生成文本的有效存量约为 300 万亿 Token，并预计可公开获取的高质量文本数据可能在 2026 至 2032 年间基本用尽。由此看，数字 AI 数据问题的性质更接近对有限存量的高效利用，即存量管理。

物理 AI 最具训练价值的恰是长尾、极端与事故场景，而它们在现实中天然稀疏。自动驾驶真正需要学习的是突发避让、恶劣天气、临界碰撞等低频但关键的边界场景，这些场景在日常行驶记录中出现频率极低。RAND 以美国道路安全统计为基准测算，2013 年人类驾驶死亡率约为每亿英里 1.09 人；由于致死与受伤是相对于行驶里程的稀有事件，仅凭实车路测在统计上清晰证明自动驾驶安全性需要数亿乃至数千亿英里，其给出的一个示例情形即需约 88 亿英里，而当时单一车队积累的最多实测里程约为 130 万英里。真实采集这类场景还面临高成本与安全风险。由此，物理 AI 数据问题的性质与数字 AI 相反，不是存量管理而是增量制造，即如何安全、低成本地获得现实中难以遇到的关键场景。

表12：数字 AI 与物理 AI 在数据获取性上的对比

维度	数字 AI	物理 AI
数据是否现成	文本、代码、图像已存在于互联网	长尾、极端、事故场景需主动制造
主要获取方式	抓取、清洗、压缩已有数据	实车采集，叠加仿真与生成
边际成本与风险	边际成本低，安全风险相对较低	采集成本高，部分场景具危险性
价值数据的分布	高质量语料相对可得	价值高度集中于稀疏长尾
存量约束	公开高质量文本预计 2026—2032 年趋于用尽	真实事故场景天然稀疏，路测难以穷尽
Scaling 核心动作	抓取与压缩存量	制造与覆盖未观测状态

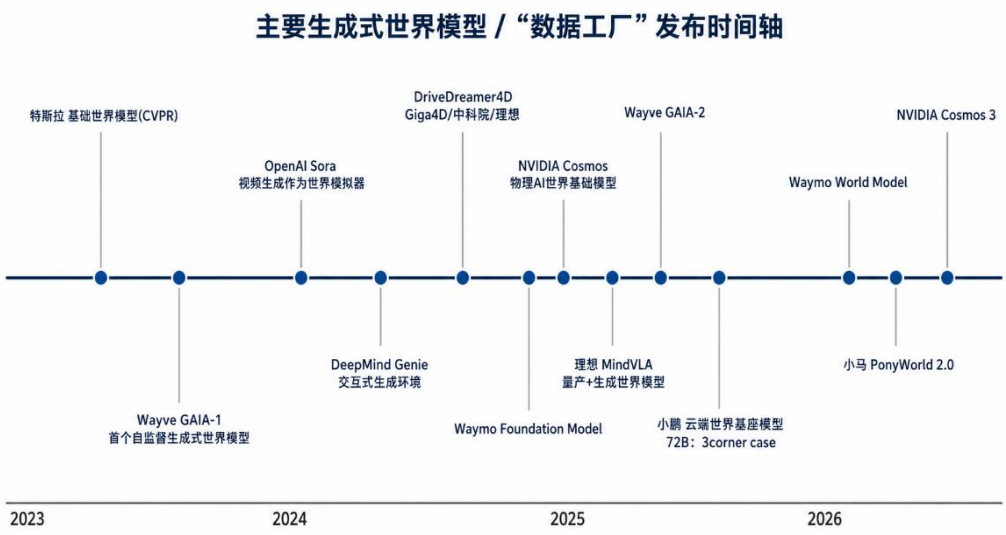
数据来源：《Will We Run Out of Data? Limits of LLM Scaling Based on Human-Generated Data》，《Driving to Safety: How Many Miles of Driving Would It Take to Demonstrate Autonomous Vehicle Reliability?》，东吴证券研究所

4.3.2. 世界模型成为数据工厂

世界模型概念源自强化学习，指智能体对环境状态及其演化的内部预测模型。Ha 与 Schmidhuber 用变分自编码器压缩高维感知、用循环神经网络建模时间动态，使智能体可在学习到的模型内部生成状态并训练策略。随大规模生成式建模发展，世界模型由辅助智能体的预测器逐步演变为可生成完整交互世界的系统，并存在“理解世界”与“预测未来”两条路径，分别以 LeCun 的 JEPA 与生成式视频模型为代表。

特斯拉在 CVPR 2023 提出构建自动驾驶领域的基础世界模型，其神经网络可根据过去数秒视频预测并生成未来全部 8 个摄像头画面，标志系统重心由判别式感知向生成式预测转移。Waymo 云端仿真则由 2016—2020 年以真实数据回放为主的 Carcraft，经 SimAgents 等生成模型，演进至 2026 年 2 月发布的 Waymo World Model，直接建模并生成未来世界状态。国际前沿机构中，英伟达于 2025 年 1 月发布 Cosmos 世界基础模型平台以生成物理一致的合成数据，Wayve 的 GAIA-2 采用潜在扩散框架、可按需生成临界碰撞与分布外危险场景。生成式世界模型在技术上以视频生成模型为底座，阿里通义万相、字节 Seedance、腾讯混元 HunyuanVideo、智谱 CogVideoX 等中国大厂与头部创业公司在这一基座层快速追赶，但通用视频生成能力尚不能直接等同于可被安全闭环依赖的数据工厂。

图19：生成式世界模型发布时间轴



数据来源：《GAIA-1: A Generative World Model for Autonomous Driving》，《Cosmos World Foundation Model Platform for Physical AI》等其他文献，东吴证券研究所

4.3.3. Corner Case 制造的价值与边界

世界模型被寄望于规模化生成 Corner Case，但其自身在长尾与长时域处的可靠性同样构成约束。长时程一致性是生成式世界模型反复强调的难点，Wayve 在 GAIA-2 中专门强调要维持长时域时空一致性; DriveDreamer4D 亦指出现有重建方法在跨原始采集轨迹渲染时容易出现目标错位、车道模糊与鬼影。这意味着越稀缺、越偏离训练分布、时间序列越长的场景，生成越不可靠，而这些区域恰是最需要 Corner Case 的区域。我们认为，世界模型既是 Corner Case 生产者、本身也面临 Corner Case，正视这一张力是评估数据工厂价值的前提。

业界围绕 Corner Case 形成两类思路。第一类为外生受控生成，生成内容由外部预设扰动参数或生成条件指定，并以真实轨迹与可重建结构为锚、以可执行性约束筛选：DriveDreamer4D 将原始自车轨迹经横向偏移与纵向变速构造新轨迹并按可行驶区域筛除不可执行路径，理想以“重建+生成”统一世界模型、以 3D 高斯泼溅把场景重建提速约 7 倍，小鹏 720 亿参数云端世界基座面向撞车、事故等稀缺极端场景规模化补全。第二类为内生问题驱动生成，把 Corner Case 发现权交给系统自身：小马智行 PonyWorld 2.0 以 Intention Layer 做自诊断，将模型表现偏差拆解为可训练问题并据此生成定向采集与 hard cases，把模型迭代由数据回流驱动推进至问题诊断驱动。两类做法路径不同，但都指向让生成更可控、更有针对性、更可校验。

表13：Corner Case 生成的两类思路对比

思路	代表	Corner Case 的发现/触发来源	典型做法	主要约束与筛选
外生：受控扰动或生成	DriveDreamer4D、理想 MindVLA、小鹏云端世界基座	由人工预设或外部条件指定（扰动参数、场景类型、生成目标）	真实轨迹横向偏移与纵向变速等受控扰动，重建加生成补全	可行驶区域、最小碰撞距离、结构条件投影、真实锚定
内生：问题驱动地发现	小马 PonyWorld 2.0	由系统对自身表现的诊断指定	生成定向采集与 hard cases	意图-动作-结果一致性、车端策略误差与世界模型误差归因

数据来源：《DriveDreamer4D: World Models Are Effective Data Machines for 4D Driving Scene Representation》等公开资料，东吴证券研究所整理

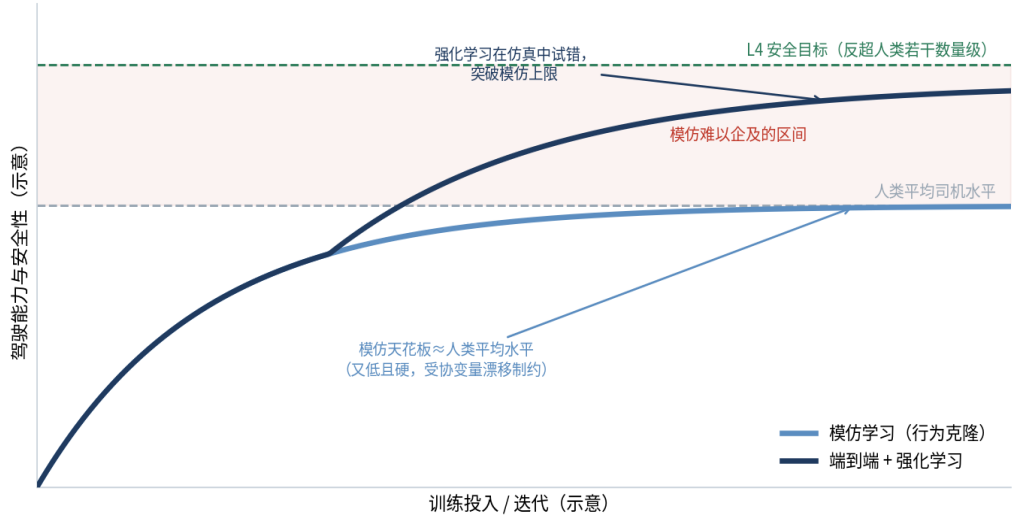
4.4. 强化学习：从可选增强走向高物理闭环关键环节

4.4.1. 模仿学习的能力上限在数字 AI 与物理 AI 之间分化

数字 AI 与物理 AI 均以模仿作为重要训练范式，但模仿对象差异决定了强化学习在两类系统中的角色分化。大语言模型的预训练在机制上可视为一种模仿，其学习对象是经过筛选的互联网级人类文本语料，模仿可支撑较高能力底座；强化学习更多承担事后对齐与推理增强角色，DeepSeek 团队 2025 年发表于《自然》的研究显示，在不使用监督微调的前提下纯强化学习即可诱发推理行为，但模仿本身已提供足够高的起点。

自动驾驶端到端模型的主流训练方式同样具有模仿属性，但模仿对象主要是人类驾驶演示数据，其能力上限更容易受到示范数据质量、覆盖范围与人类驾驶分布约束。行为克隆式模仿学习在闭环控制中还存在协变量漂移问题: Ross 等指出策略执行早期的微小动作偏差会使车辆进入训练分布之外的状态，误差随时间累积放大。多家机构公开表述指向同一判断——L4 自动驾驶目标是在安全性上超越人类而非趋近人类平均水平，卓驭科技将预训练模仿概括为可将系统能力推升至约 80 分、而从 80 分提升至 95 分需要强化学习。我们认为，当模仿对象为平均水平人类司机、而安全目标远超人类时，模仿天花板更低、约束更强，强化学习由可选的事后增强转向突破上限的关键环节。

图20：模仿学习与强化学习的能力天花板（概念示意）



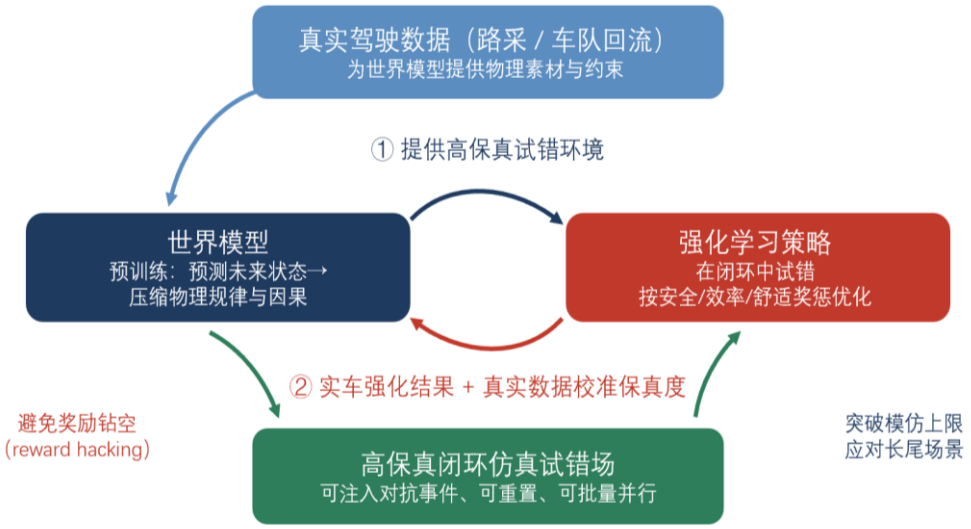
数据来源：《A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning》，《DeepSeek-R1 Incentivizes Reasoning in LLMs through Reinforcement Learning》，东吴证券研究所

4.4.2. 世界模型为强化学习提供高保真试错环境

强化学习要在驾驶中发挥作用，离不开可反复试错的训练环境，而真实道路难以承担这一角色——碰撞类数据极难且不应主动获取，危险场景也无法随意复现。世界模型的意义在于提供可控、可重置、可注入危险事件的闭环环境。理想在阐述 MindVLA 时表示，强化学习若要避免“奖励钻空”，必须与足够逼真的世界模型交互，其“重建+生成”统一世界模型将三维场景重建与新视角生成结合，在其上开展大规模闭环强化学习；华为乾崑 ADS4 以云端世界引擎的难例扩散模型生成场景，难例密度可达真实世界的约 1000 倍。

世界模型的保真度又依赖真实数据持续校准，由此形成双向共生。理想指出重建式方法符合物理约束但新视角泛化受限、生成式方法泛化较强但物理约束不足，两者结合方能得到既符合物理规律又具泛化能力的世界模型，这一结合依赖真实路采数据为重建提供素材、为生成提供约束。Momenta 首席执行官曹旭东在发布 R7 时表示，预测是智能进化的核心基石，世界模型通过预测物理世界未来状态来理解物体属性与因果关系。我们认为，世界模型先通过对真实数据的预测学习物理规律，再以此充当强化学习环境，真实数据质量直接决定环境保真度、进而影响强化学习有效性。

图21：强化学习与世界模型的双向共生闭环



数据来源：东吴证券研究所绘制

4.4.3. 厂商实践由端到端训练延伸至后训练

代表性厂商的强化学习实践主要围绕量产辅助驾驶与 L4 Robotaxi 两类场景展开。量产乘用车侧，“端到端之上叠加强化学习后训练”正收敛为共识：特斯拉 FSD 自 v12 起强化端到端路线，Elluswamy 在 2025 年 ICCV 介绍了可合成 8 路相机画面、注入对抗事件的学习型世界模拟器；小鹏在云端训练约 720 亿参数世界基座并经后训练、强化学习与蒸馏生产车端模型；理想以“重建+生成”世界模型作为强化学习训练场；Momenta 在端到端基础上引入强化学习推出 R6 飞轮大模型，地平线 HSD 采用“一段式端到端+强化学习”。

L4 Robotaxi 侧更强调以世界模型驱动的闭环仿真与安全冗余：小马智行第七代 Robotaxi 以 PonyWorld 世界模型基于强化学习范式承载生成场景、高保真仿真与行为评估，楼天城提出“安全不是模仿而是超越”；Waymo 公开口径更多以基础模型与大规模仿真支撑 L4 运营，文远知行侧重端到端、博弈论与飞轮式大规模仿真。我们归纳出若干共性与分歧：共性在于代表性厂商均在不同程度上围绕“端到端模型、世界模型仿真、强化学习”构建提升长尾安全性的技术组合；分歧主要体现在是否以语言为中介、强化学习披露的明确程度，以及面向量产辅助驾驶还是 L4 运营。相关分类系我们整理，非业界统一标准。

表14：代表性厂商“端到端+世界模型+强化学习”技术布局梳理

厂商	定位	主要技术标签	强化学习的角色	世界模型/仿真环境
特斯拉	量产辅助/Robotaxi	端到端神经网络，FSD V12→V14	V14 在模仿外引入更激进 RL	学习型世界模拟器，可合成 8 路相机、注入对抗事件
小鹏	量产辅助驾驶	世界基座模型 72B→车端 XVLA	云端后训练含 RL，再蒸馏上车	在研世界模型作为基座训练场
理想	量产辅助驾驶	MindVLA/MindVLA-o1	重建+生成世界模型中做大规模闭环 RL	重建+生成统一世界模型，MindSim
Momenta	量产辅助驾驶供应商	一段式端到端→R6 飞轮→R7	R6 量产端到端+RL，R7 在世界模型中做 RL	R7 三层世界模型
地平线	芯片+量产方案	HSD 一段式端到端+强化学习	交互式博弈强化学习，仿真平台训练	端到端世界模型+交互博弈
华为乾崑	量产方案 L2-L4	WEWA：世界引擎+世界行为模型	以仿真环境中 RL 探索为核心	云端世界引擎生成高密度难例
Waymo	L4 Robotaxi	Waymo 基础模型；EMMA 研究	公开口径侧重基础模型与仿真	大规模仿真，EMMA 基于 Gemini
小马智行	L4 Robotaxi	第七代 Robotaxi；PonyWorld	基于 RL 范式的 PonyWorld	生成场景、高保真仿真、行为评估
文远知行	L4 Robotaxi/通用	WeRide One；一段式端到端	公开侧重端到端、博弈论、飞轮	大规模云端仿真

数据来源：各公司官网、公告，东吴证券研究所整理

4.5. 时延：云端训练最优与车端部署最优的分离

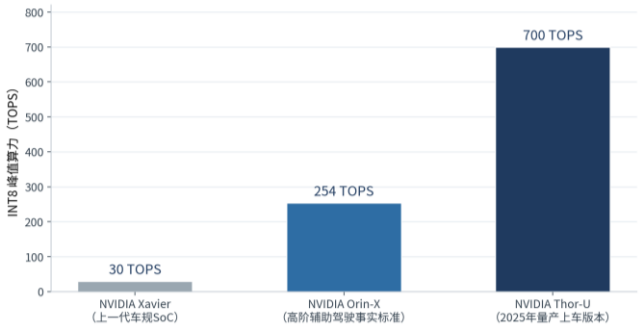
4.5.1. 训练最优不等于部署最优

经典缩放定律以给定训练算力下的损失最小化为主要目标，“最优”指向训练算力最优前沿而非部署时延。Kaplan 等（2020）的结论更偏向扩大参数规模（约 1.7 个训练 Token/参数），Hoffmann 等（2022）提出 Chinchilla 式算力最优配比、给出参数与数据近似等比扩张（约 20 个 Token/参数），但两项研究均未将推理成本与实时时延纳入约束。Sardana 等（2024）在缩放定律中显式加入推理成本后指出，在高频部署场景下算力最优点会移向参数更小、训练数据更多的模型；Llama2、Llama3 分别以 2 万亿、15 万亿 Token 远超 Chinchilla“最优”数据量训练，正是这一部署导向的体现。

智能驾驶把部署约束推向更高强度：在固定车规算力上以不低于 30Hz 运行大型视觉 Transformer 叠加大语言模型，在现有车载算力下仍具较高难度。与云端算力可弹性扩展不同，车端算力平台相对固定：NVIDIA Orin-X 约 254 TOPS 长期被广泛用于高阶辅助驾驶平台，其继任者 Thor 自 2025 年起量产上车、主流落地版本 Thor-U 单颗约 700 TOPS。当前量产大模型的实际推理帧率仍低于 30Hz 参考水平，语言模型与链式推理带来的串行计算是重要时延来源：理想披露其 VLA 推理帧率约 10Hz（较上一代 VLM 的约 3Hz 提升三倍多），NVIDIA 在 Alpamayo-R1 论文中报告端到端推理时延 99ms（约合

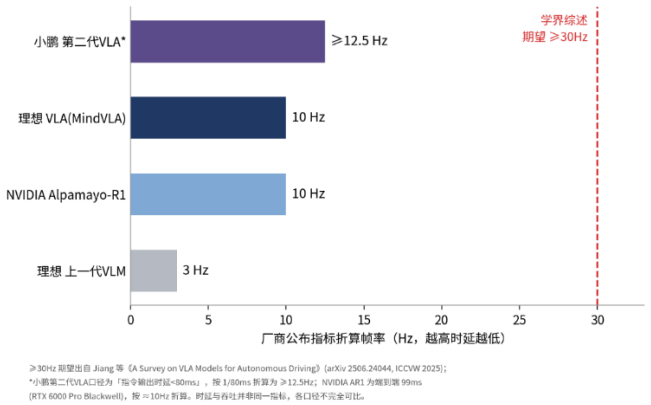
10Hz），小鹏称第二代 VLA 把指令输出时延压缩至 80ms 以内。

图22：主流车端 AI 算力平台峰值算力对比（Xavier→Orin-X→Thor-U）



数据来源：NVIDIA 官网，东吴证券研究所绘制

图23：智驾大模型车端推理帧率对比（各家公布口径）

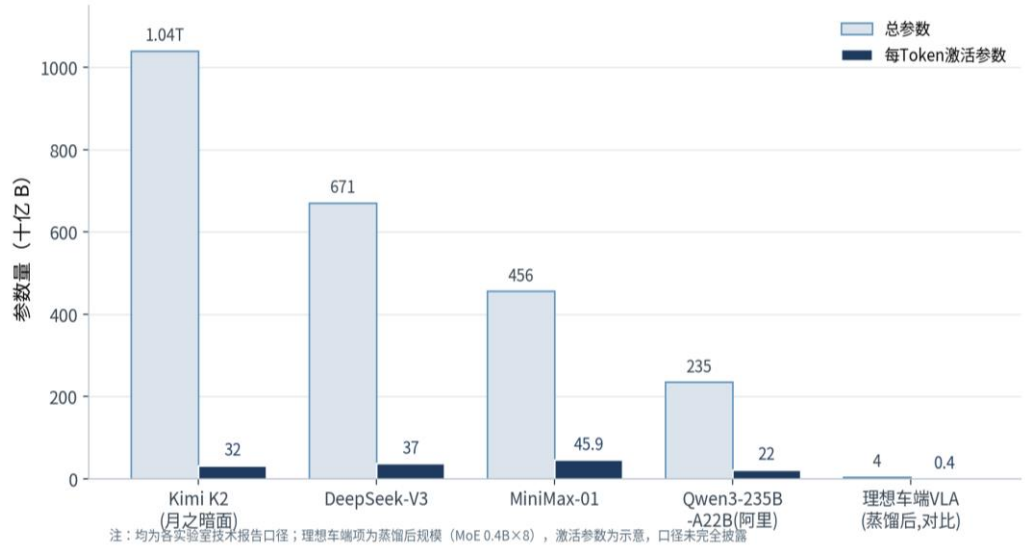


数据来源：《A Survey on VLA Models for Autonomous Driving》，公司公开信息，东吴证券研究所绘制

4.5.2. 云端压缩与车端压缩形成不同效率工具箱

云端压缩与车端压缩的边界在于优化对象不同：云端侧主要降低模型固有算力与复杂度，重点在算法和架构层面；车端侧则在固定车规算力约束下实现低时延部署，重点在系统和工程层面（上述划分为我们归纳，实际路径存在交叉）。云端侧以“稀疏激活 + 注意力/缓存压缩 + 低精度 + 投机解码”为主：混合专家（MoE）通过每个 Token 只激活一小部分专家将模型容量与单次推理算力解耦，DeepSeek-V3 以 671B 总参仅激活 37B；DeepSeek 提出的 MLA 以低秩联合压缩把 KV 缓存压至约 70KB/Token、较 GQA 类降低约 2.7—4.7 倍；MiniMax-01 以闪电线性注意力把序列复杂度由二次降为线性。面向驾驶，云端侧还发展出视觉 Token 剪枝（理想 LightVLA、小鹏 FastDriveVLA）与以隐式/动态 Token 替代冗长思维链（FutureX、华为 DynVLA）两类方法。

图24：云端 MoE 基座模型总参数与激活参数对比（含车端蒸馏规模）



数据来源：《DeepSeek-V3 Technical Report》，《MiniMax-01: Scaling Foundation Models with Lightning Attention》等其他文献，东吴证券研究所

车端侧压缩主要围绕三条路径展开：其一是把云端大模型蒸馏为可上车的小模型，Waymo 在 2025 年 12 月描述以基础模型适配 Driver/Simulator/Critic 三类大教师、再安全蒸馏为车端学生模型，理想在云端以约 32B 多模态基座经强化学习与蒸馏压缩为约 4B 车端 VLA 部署于 Thor 芯片；其二是以快慢双系统在运行时分离快速响应与复杂推理，理想较早采用基于卡尼曼“思考快与慢”的端到端 + VLM 双系统，Waymo 基座采用 Think Fast/Think Slow；其三是量化、投机解码、采样步数压缩与编译优化等推理工程，理想以 INT4/INT8 混合精度运行、并以 ODE 采样器把去噪步数由约 10 步压缩至 2 步。

表15：代表性厂商的“云端基座—车端部署”路径

主体	云端基座	车端部署	蒸馏与快慢架构
理想	约 32B 多模态基座；云端 13EFLOPS	约 4B，MoE 0.4×8；Thor；INT8/FP8；10Hz	基座→RL+蒸馏→车端 VLA；早期 E2E+VLM 双系统统一为 VLA
小鹏	720 亿参数物理世界基座；云端约 10EFLOPS	蒸馏版第二代 VLA 推送 MAX 车型/自研图灵芯片	基座蒸馏上车；CoT 效率约 32×保实时；编译效率约 12×
Waymo	Waymo Foundation Model，统一 Driver/Simulator/Critic	学生模型车端实时运行+独立轨迹校验层	教师→学生蒸馏；Think Fast/Slow 双系统
华为/引望	诺亚前瞻研究+引望量产；DynVLA 以 EMU3 8B 为骨干	面向量产落地	以动态 Token 替代冗长 CoT 降低推理负担

数据来源：公司公开信息，东吴证券研究所

5. 数字 AI 价值在确定性，本轮 AI 增量在物理 AI

数字 AI 已实现在语言、视觉与多模态任务中通用底座模型的跑通，但其难以实现物理世界的建模闭环。我们认为物理 AI 将在一段时间内将与数字 AI 并列成为“任务通解”，成为物理世界场景的 AI 解决方案。物理 AI 将实现在连续、实时、安全约束的真实世界中建立"感知—行动—反馈—优化"的闭环。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,本轮 AI 更大的增量在物理 AI,智能驾驶是其最先跑通闭环的代表场景。

我们认为，物理 AI 正从复用数字 AI 能力转向补齐自身底层能力。早期主要迁移 Transformer 等架构，当前 VLM、VLA 进一步建立在 LLM 及多模态基座之上，但数字 AI 红利或已接近阶段性瓶颈。根本原因在于，物理 AI 面临三类差异：一是缺少天然 Token，需要自主构建物理世界表征；二是长尾、极端场景稀缺且采集成本高，难以直接复制互联网数据 Scaling；三是模仿学习受限于人类驾驶上限，必须依赖强化学习持续试错。世界模型或成为突破上述瓶颈的核心抓手：向前完成潜在空间表征，向中游生成高价值长尾数据，向后为强化学习提供可规模化试验环境，进而形成“表征—数据—学习”的闭环。

物理 AI 从表征到部署的闭环路径与数字 AI 差异明显,使物理 AI 形成区别于数字 AI 的独立而统一的发展方向。我们看好物理 AI 后续发展:随着表征向潜在状态收敛、VLA 持续演进、世界模型与强化学习闭环走向成熟、云车部署逐步兑现,物理 AI 有望由能力展示走向大规模落地,成为 AI 发展当前阶段最具成长弹性的方向。

6. 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