汽车行业深度报告

AI 系列深度 23 期：智能驾驶的表征演进

增持（维持）

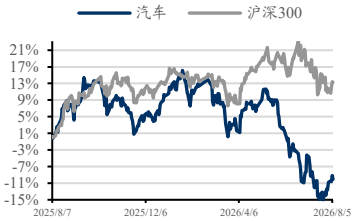
2026 年 08 月 07 日

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

- **物理 AI 与数字 AI 的根本差异，首先体现在表征获取方式上。**文本天然离散且带语义，数字 AI 输入具备较强可 token 化基础；物理世界的输入则由连续光子、点云、运动和空间关系构成，缺乏天然通用的物理世界 token 化方案。因此，物理 AI 需要先构建三维物理空间的可学习表征，再进入规划、控制和行动。
- **以智能驾驶为例，表征演进可概括为原始像素、BEV/VectorNet、Occupancy、多模态 Token 和潜在状态五个阶段。**原始像素端到端简洁但缺少结构约束；BEV+Transformer 与矢量化 VectorNet 是同一层级的两条并行路线，分别以特斯拉和 Waymo 为代表，前者把多摄像头信息统一到鸟瞰三维空间，后者把道路和参与者抽象成数学对象；Occupancy 从识别对象类别转向判断空间占据；多模态 Token 连接视觉、语言和动作；潜在状态则把世界压缩到可预测空间。
- **物理 AI 中的 token 化器本质上是重型编码网络。**文本分词通常只是把离散符号映射为 token，而车端 BEV 编码器、占用网络、VAE、3D 编码器等承担了从连续物理输入到可学习表征的核心转换。离散化和嵌入不是预先给定的，而是由网络学习形成；这也是物理 AI 相较数字 AI 多出的一道关键工程门槛。
- **从公司证据看，特斯拉、小鹏、华为、理想、Waymo、Wayve 等路线虽不同，但共同主线是从目标级识别走向空间级理解，再走向潜在状态预测。**多种表征长期并存，并非简单替换：BEV 与矢量、占用、3D 高斯、世界模型潜空间，分别服务不同层次的感知、预测、仿真和规划需求。
- **世界模型可能成为表征训练的关键手段。**标签稀缺使物理世界嵌入难以完全依赖人工监督，GAIA、Cosmos、JEPA 系等世界模型通过自监督预测未来状态、生成场景或学习潜在动力学，帮助模型在无标注或弱标注条件下形成更统一的物理表征。竞争壁垒正从单点感知精度，前移到如何将物理世界嵌入可预测潜在空间。
- **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 17 期：视觉智能为何引入 Transformer？》

2026-08-06

《AI 系列深度 16 期：语言智能为何从 RNN 转向 Transformer？》

2026-08-06

内容目录

图表目录

1. 表征演进：物理世界如何进入模型

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；相似对象由此形成聚合，可区分属性沿不同方向展开，为下游判别、回归与规划提供统一输入。

张量是承载表征的统一数据形式。标量、向量、矩阵与更高阶数组统称张量，神经网络各层之间传递的数据本质上都是张量：文本经分词后查表映射为嵌入向量序列，形成序列长度 × 维度的张量；图像则天然对应高 × 宽 × 通道的像素张量。模型理解过程，可以视为不断将原始张量转换为信息更紧致、任务更可分的表征张量。

衡量表征质量，通常关注三点：1）是否保留任务相关信息，并使其尽量线性可分；2）是否压缩冗余与噪声；3）是否让相似对象在空间中聚合。上述标准将贯穿后续对各类物理 AI 表征的比较。下表先汇总相关基础概念。

表1：张量与嵌入 Embedding 基础概念

概念	含义	在 AI 中的典型形态
张量	标量/向量/矩阵的统一推广，多维数组	网络层间传递的数据、参数与中间特征
嵌入 Embedding	把离散或连续对象映射为低维稠密向量	词向量、图像/视频特征、潜在状态向量
分词/Token 化	把输入切分为可索引的基本单元	文本子词 BPE、图像 Patch、视觉离散码
潜在空间 Latent space	压缩后的低维表征空间	自编码器/VAE 隐变量、世界模型潜在状态
表征学习	从数据中自动学习有用表征	自监督预训练、对比学习、掩码重建

数据来源：《Efficient Estimation of Word Representations in Vector Space》，《Auto-Encoding Variational Bayes》等其他文献，东吴证券研究所

2. 物理 AI 表征：缺乏天然通用 token 化方案

数字 AI 与物理 AI 的根本差异，首先在于表征获取方式不同。文本天然是离散且带语义的符号序列，语言本身已由人类对世界完成一轮编码；物理世界输入则是连续的原始光子与点云，缺乏天然、通用、可直接处理的物理世界 token 化方案。这意味着物理 AI 迭代中多出一段数字 AI 无需经历的流程：先将三维物理空间构建为可学习底座。

2.1. 文本侧：离散 Token 降低输入表征成本

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，属于相对低成本的输入处理环节。

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。（1）DeepSeek-V3 采用字节级 BPE，词表约 12.8 万；（2）阿里 Qwen 系列在 tiktoken 的 cl100k_base 基础上扩充中文，词表约 15.1 万；（3）智谱 GLM 系列采用字节级 BPE，词表约 15 万；（4）MiniMax-01 词表为 200,064。（5）月之暗面 Kimi、字节豆包 Seed、百度文心、腾讯混元等也主要沿文本或多模态文本接口展开。

整体来看，文本侧的世界编码主要由人类语言预先完成，模型接收的输入已经具有离散性、语义性和较强的结构化基础。文本分词环节集中解决符号颗粒度和词表效率问题，模型能力学习主要发生在分词之后；物理 AI 则需要在进入主体模型前，先完成连续信号到可学习表征的转换。

2.2. 物理世界侧：连续信号需要先构建可学习底座

物理 AI 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。其输入包括多相机连续像素、激光雷达点云等原始信号，无法像文本那样通过查表直接获得 Token。

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。这一视觉 Token 化过程正是物理世界表征问题的缩影。需要区分的是，图像 Patch 或视觉 Token 主要解决 2D 视觉输入的模型化问题，智能驾驶还需要三维空间、动态交互与动作后果表征。

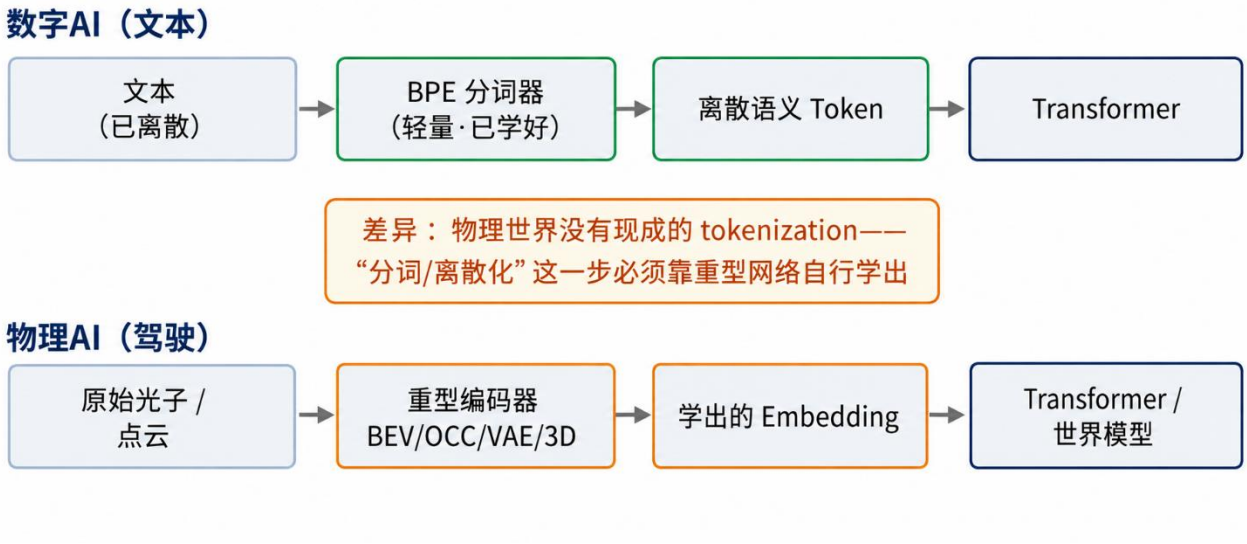
智能驾驶领域进一步形成了面向三维空间的可学习底座。BEV+Transformer 与矢量化 VectorNet、Occupancy 占用网络、3D 高斯，可归纳为一条表征演进主线，并将在第三部分结合公司样本展开。下表与下图先对照数字 AI 与物理 AI 两类表征问题。

表2：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：《BEVFormer: Learning Bird's-Eye-View Representation from Multi-Camera Images via Spatiotemporal Transformers》，《Occupancy Networks: Learning 3D Reconstruction in Function Space》等其他文献，东吴证券研究所

图1：数字 AI 与物理 AI 输入处理链路对照

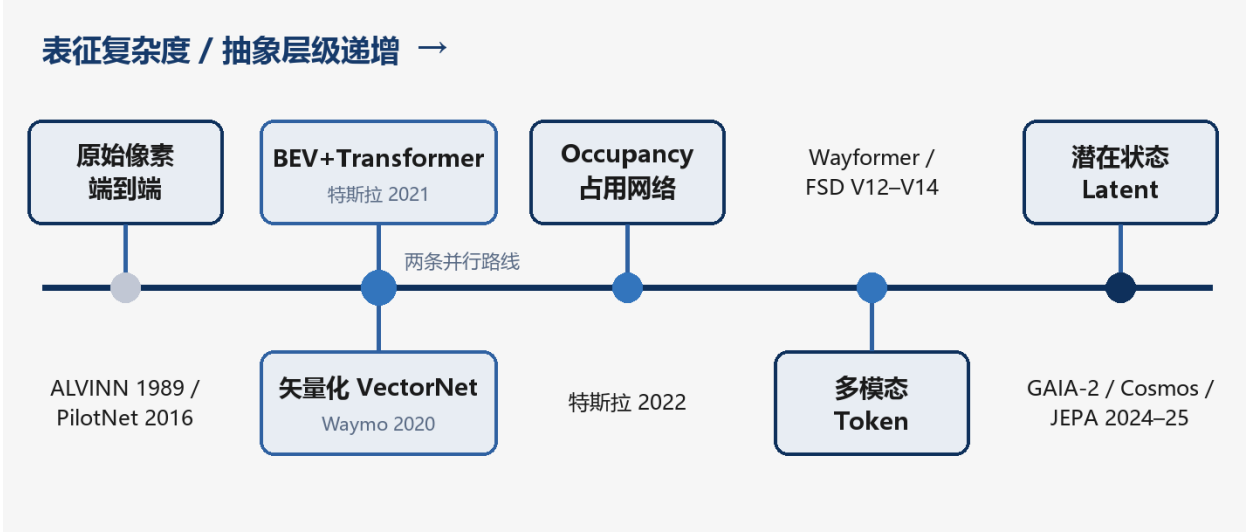


数据来源：《Qwen3-VL Technical Report》，《Efficient Estimation of Word Representations in Vector Space》，东吴证券研究所

3. 智能驾驶表征演进：从像素、BEV 到潜在状态

我们将智能驾驶表征的演进归纳为一条主线：原始像素→BEV/VectorNet→Occupancy 占用→多模态 Token→潜在状态，各阶段在产业中长期并存、相互叠加。其中 BEV 与矢量化属于同一层级的两条并行路线，前者以特斯拉为代表，把多相机像素统一到鸟瞰三维空间；后者以 Waymo 为代表，把地图与轨迹抽象为矢量对象。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。

图2：智能驾驶表征演进主线示意



数据来源：《End to End Learning for Self-Driving Cars》，《VectorNet: Encoding HD Maps and Agent Dynamics from Vectorized Representation》等其他文献，东吴证券研究所

贯穿智能驾驶表征演进的主线，是上一阶段表征约束推动下一阶段表征形成：原始像素将感知与决策耦合在隐式表征中，诊断和泛化难度较高，因此 BEV 将多相机图像统一到三维鸟瞰空间，同期矢量化将地图与轨迹表示为矢量并以图和注意力建模交互，二者分别从空间侧与关系侧摆脱原始像素；BEV 与矢量表征仍以目标框、车道等结构化实体为中心，对不规则或未知障碍以及自由空间表达不足，因此 Occupancy 以体素占据描述空间；占用表征稠密但计算负担较高，且与矢量化一样仍偏任务专用，多模态 Token 尝试将视觉、语言、动作统一到序列建模框架；而 Token 更多是对已观测信息的编码，面向长尾和安全约束，进一步走向可预测潜在状态。

3.1. 原始像素端到端：起点与局限

以原始像素直接回归控制量，是端到端驾驶最初的表征形态。ALVINN，Pomerleau，**1989**，用一个浅层全连接网络，把低分辨率前视图像直接映射到转向；英伟达 PilotNet/DAVE-2 则用约 9 层卷积网络，仅以人类驾驶的前视像素与转向角做端到端模仿训练。网络在没有车道线显式标签的情况下，内部形成了对车道与道路边界敏感的特征，作者以显著性可视化加以佐证。这说明表征可以被隐式学习，但其形态不可控、难以检视。

纯像素模仿的脆弱性，推动业界转向显式、可检视的三维表征。纯模仿在开环评估中可能表现良好，却会在闭环中因协变量偏移与误差逐步累积而失效；改进路径包括显式合成扰动轨迹，加入碰撞、偏离车道等几何损失，并将输入从原始像素改为中层鸟瞰 BEV 表示，以降低感知噪声对规划的传导，构成 BEV 阶段的直接动机。

3.2. BEV 加 Transformer：从 2D 像素到 3D 向量空间

特斯拉 2021 AI Day 使用 Transformer，将多相机图像统一到三维向量空间。感知

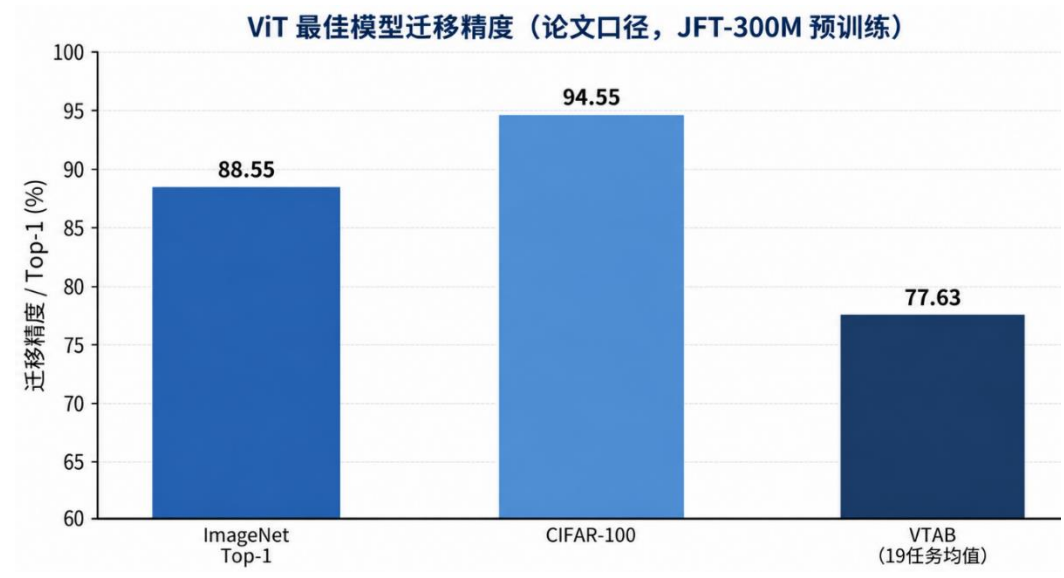
先以共享骨干 RegNet 逐相机提取多尺度图像特征，经 BiFPN 做跨尺度融合，再以鸟瞰空间的位置编码作为查询 Query、各相机图像特征作为键值 Key/Value，由自注意力学习向量空间中每个位置应从各图像的哪些区域取信息，从而降低对逐相机标定固定投影的依赖。

直接在向量空间预测的原因在于，先在图像平面预测再投影会引入难以消除的误差。当一个目标横跨多个相机时，逐相机预测拼接到向量空间会出现错位与不一致，且每辆车的相机外参存在装配抖动；让网络直接在向量空间输出，可同时吸收这两类误差。时序上，特斯拉用基于时间和基于空间的双特征队列缓存历史特征，再由空间 RNN 视频模块融合，以处理遮挡与运动估计。

在国内量产侧，小鹏率先把动态 BEV 做进了量产车。小鹏 XNet 1.0，2022 年 10 月 24 日发布，为国内首个量产的动态 BEV 感知架构，以 View Transformer 把多相机 2D 特征投影到统一 3D 坐标系，直接输出动态目标的 3D 边界框与速度。

BEV 所用的 Transformer 内核，与自然语言、视觉领域同出一脉。Transformer，Attention Is All You Need，2017，以自注意力替代 RNN 与卷积，在 WMT 2014 英德/英法翻译上分别取得 28.4/41.8 BLEU；ViT 把图像切成 16 × 16 的 Patch、线性投影后加位置编码当作 Token 序列输入标准 Transformer，但需在 JFT-300M 等超大规模数据上预训练才显著超越卷积网络；BEV 的学术谱系亦可参考 LSS 与 BEVFormer，以 BEV 查询做空间交叉注意、并辅以时序自注意。

图3：ViT 在大规模预训练后的迁移分类表现



数据来源：《An Image is Worth 16 × 16 Words: Transformers for Image Recognition at Scale》，东吴证券研究所

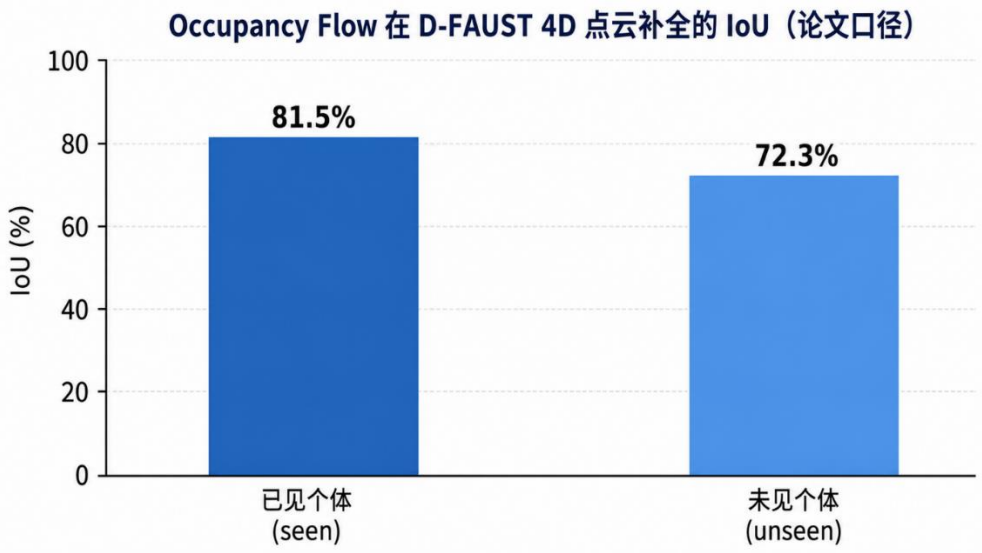
BEV 给出了度量一致的三维空间，却仍以预定义实体为中心。其输出多为目标框、车道线等结构化实体，对形状不规则、类别未知的障碍物，以及可通行自由空间本身的

表达力不足，这推动了对空间占据本身建模的 Occupancy 路线。

3.3. Occupancy 占用：从“是什么”到“是否被占据”

占用网络把三维表面表示为连续空间中的占据概率边界，从而摆脱固定体素分辨率与内存约束。其机制是对空间中任意一点在给定条件下输出占据概率，概率等值面可被提取为物体表面；该表示可在推断时通过多分辨率等值面提取 MISE 和八叉树式细分生成网格。Occupancy Flow 进一步为空间叠加随时间演化的运动向量场，把静态三维重建扩展为四维。

图4：Occupancy Flow 在 D-FAUST 上的 4D 点云补全表现



数据来源：《Occupancy Flow: 4D Reconstruction by Learning Particle Dynamics》，东吴证券研究所

表3：占用网络与占用流学术指标

论文/任务	指标	数值
Occupancy Networks / ShapeNet 单图 3D 重建	mean IoU	0.571
Occupancy Networks / ShapeNet 单图 3D 重建	Chamfer-L1	0.175
Occupancy Networks / ShapeNet 单图 3D 重建	Normal Consistency	0.834
Occupancy Flow / D-FAUST 4D 补全 seen	IoU	0.815
Occupancy Flow / D-FAUST 4D 补全 unseen	IoU	0.723

数据来源：《Occupancy Networks: Learning 3D Reconstruction in Function Space》，《Occupancy Flow: 4D Reconstruction by Learning Particle Dynamics》，东吴证券研究所

特斯拉 2022 AI Day 把占用网络落到车端，按体素直接预测空间占据及其速度。占用网络把车身周围空间划分为体素，重点预测每个体素的占据状态与占用流，即速度信息，可在任意点查询；该方法对未知形状的一般障碍物同样有效，并在车端以较高分辨率近实时运行，从而在无激光雷达的纯视觉条件下输出三维占据。

国内头部厂商在 2023 年前后相继把占用纳入量产感知栈。1）小鹏 XNet 2.0，2023 年 10 月 24 日，在动、静态 BEV 之外引入 Occupancy 实现三网合一。**2）**华为 ADS 2.0，2023 年 4 月，在 BEV 底座上引入 GOD 通用障碍物检测，对未在白名单内的异形障碍亦可识别，支撑无图道路理解。**3）**理想 AD Max 3.0，2023 年，OTA 5.0 于 2023 年 12 月落地，采用静态 BEV、动态 BEV 与 Occupancy 三类网络，并以 NeRF 增强静态几何还原。

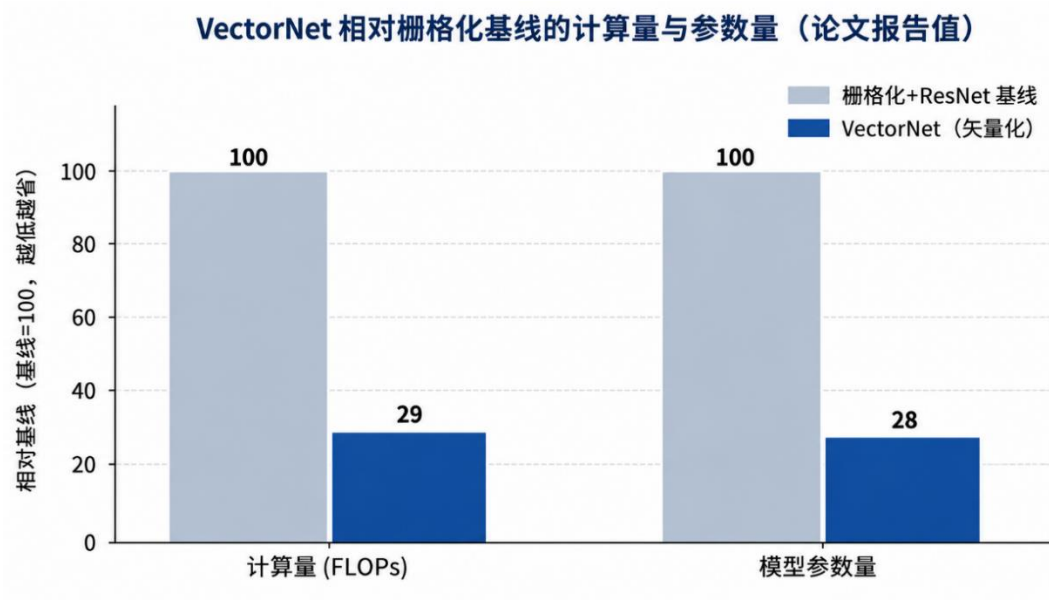
占用提供了稠密的自由空间，却偏重且不显式编码实体之间的关系。而预测与规划恰恰需要结构化、关系化的场景表征。在特斯拉沿 BEV 与占用推进的同时，Waymo 从关系侧给出了另一条并行答案：把地图与轨迹直接矢量化，并以图与注意力建模交互。

3.4. 矢量化：从栅格图像到数学矢量

VectorNet 不再把场景渲染成图像，而是把地图与轨迹表示成矢量并以分层图网络建模。它把高精地图与运动物体表示为带方向与属性的矢量折线序列，先在每条折线内部用类 PointNet 的局部子图聚合得到折线级节点特征，再在所有折线节点之间用自注意力构成全局交互图，建模高阶关系；训练时还设有自监督的节点补全辅助任务，随机遮蔽部分地图或轨迹节点并根据上下文重建。

矢量化在大幅降本的同时，把模型重心从生成轨迹转向学习场景表征。相比渲染成栅格再用卷积网络处理的基线，VectorNet 将计算量降低约一个数量级、参数量节省 70%以上，并在 Argoverse 预测基准上取得更优结果，K=1 时 DE@3s 约 4.01；Wayformer 进一步把地图、历史轨迹、交互等多模态场景统一为 Token，用注意力完成跨模态与时空融合，并研究早、中、晚不同融合方式，以隐查询注意力降本，更接近 Transformer 范式。

图5：VectorNet 相对栅格化基线的计算量与参数对比



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

占用与矢量化分别兼顾空间覆盖与关系建模，但仍是面向感知与预测的任务专用表征。数字 AI 的大序列模型范式提示，驾驶模型也可以将多相机视频、地图、动作乃至语言统一为一条 Token 流。

3.5. 多模态 Token：走向视觉—语言—动作统一

特斯拉 FSD v12 用单一神经网络实现像素输入到控制输出，把大量规则代码纳入端到端模型。FSD v12 以单一神经网络替代此前超过 30 万行的 C++规则代码；FSD V14 进一步扩大模型规模并强化多模态端到端闭环，输入覆盖视频、导航、自车状态与音频，并把语言与三维重建用作中间监督，路径已接近视觉—语言—动作 VLA 的统一建模。

把语言纳入驾驶、把视觉编码为 Token，正成为中外厂商共同方向。1）Wayve 的 LINGO/LINGO-2 把自然语言纳入驾驶模型，可解释、可闭环，属 VLA 方向。**2）**中国厂商的多模态模型，如 Qwen3-VL、MiniMax-VL-01，则在文本 Token 之外，额外引入学习得到的视觉 Token，这正是物理世界表征问题在大模型侧的映射。

表4：代表性中国大模型分词器与潜在表征要点

机构/模型	分词/词表口径	表征相关要点
DeepSeek-V3	字节级 BPE，词表约 12.8 万	MLA 低秩潜在 KV 压缩
阿里 Qwen / Qwen3-VL	tiktoken cl100k_base 扩中文，约 15.1 万	多模态用 SigLIP2 视觉编码器产视觉 Token
智谱 GLM	字节级 BPE，约 15 万	以文本为主，多模态版本另加视觉编码
MiniMax-01 / VL-01	词表 200,064	线性与 softmax 混合注意力
腾讯 混元-T1	—	引入 Mamba 状态空间结构
月之暗面/字节/百度	各自 BPE 体系	Kimi、豆包 Seed、文心，文本为主路线

数据来源：《DeepSeek-V3 Technical Report》，《Speculating LLMs' Chinese Training Data Pollution from Their Tokens》等其他文献，东吴证券研究所

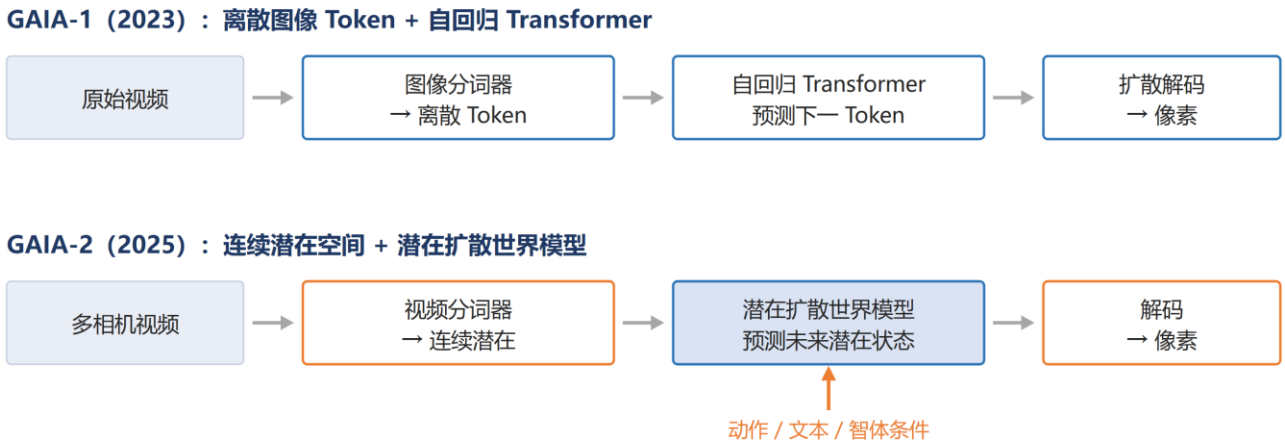
多模态 Token 让驾驶模型可以沿大语言模型的序列建模范式扩展，但 Token 仍主要是对已观测信息的编码。面向长尾场景与安全验证，模型需要进一步预测尚未发生的未来状态，这推动了可预测潜在空间的世界模型路线。

3.6. 潜在状态：从观测编码走向可预测空间

GAIA-1 把视频压缩为离散 Token，再用自回归 Transformer 以预测下一个 Token 的方式建模世界。其用矢量量化 VQ 图像 Token 化器把视频帧压缩为离散 Token，码本约 8192，世界模型是约 65 亿参数的自回归 Transformer，在历史图像、文本与动作 Token 条件下预测下一 Token 以建模场景动态，并用视频扩散解码器把 Token 渲染回像素，约 4 倍时序上采样；其训练还借 DINO 特征蒸馏向 Token 注入语义，并表现出与大语言模型类似的规模化规律。

随后的 GAIA-2 与英伟达 Cosmos，把表征从离散 Token 进一步推进到连续潜在。GAIA-2 改用连续潜在表示，约 32 倍空间压缩、64 通道，并以潜在扩散世界模型在动作、自车与文本条件下直接预测未来潜在状态；Cosmos 的 Token 化器同样是学习得到的编码器，可输出连续或离散 FSQ 两类 Token，并以潜在扩散构建世界基础模型，公开信息显示小鹏为首批采用方之一。这说明，物理 AI 中的 Token 化器本身就是一类重型可学习网络。

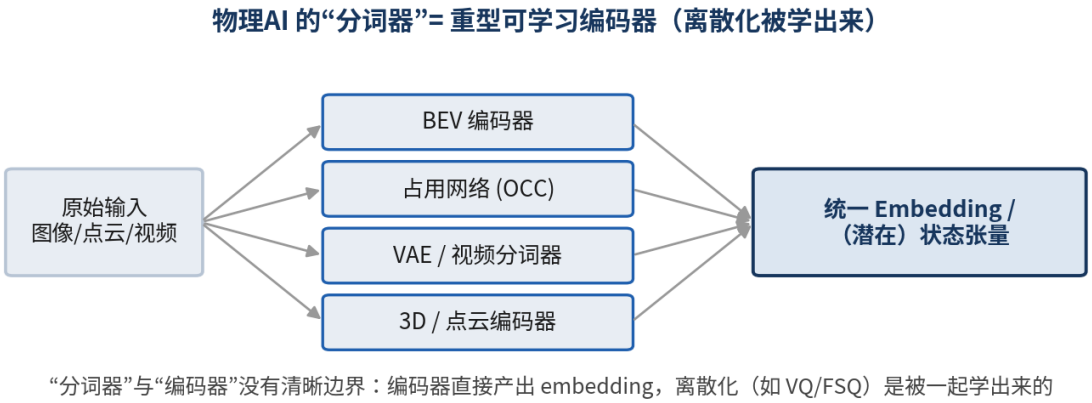
图6：GAIA-1 到 GAIA-2：从离散 Token 加自回归到连续潜在加潜在扩散



数据来源：《GAIA-1: A Generative World Model for Autonomous Driving》，《GAIA-2: A Controllable Multi-View Generative World Model for Autonomous Driving》，东吴证券研究所

同一种潜在化思路，在文本侧也有对应。DeepSeek-V3 的 MLA 以低秩潜在向量压缩注意力的键值，是这一思路在语言模型中的体现。下两图分别给出 GAIA 潜在表征演进与物理 AI Token 化器的本质。

图7：物理 AI 的 Token 化器：BEV、占用、VAE 与 3D 编码器等重型网络



数据来源：《BEVFormer: Learning Bird's-Eye-View Representation from Multi-Camera Images via Spatiotemporal Transformers》，《3D Gaussian Splatting for Real-Time Radiance Field Rendering》等其他文献，东吴证券研究所

表5：主要厂商车端感知表征路线

厂商	代表阶段/版本	表征路线要点
特斯拉	AI Day 2021 / 2022 / FSD V12-V14	BEV+Transformer→占用→多模态端到端闭环
小鹏	XNet 1.0 在 2022 年 / XNet 2.0 在 2023 年	动态 BEV→引入 Occupancy 三网合一
华为	ADS 2.0 2023.04	BEV 底座+GOD 通用障碍物，无图道路理解
理想	AD Max 3.0，2023，OTA5.0，	静态/动态 BEV+Occupancy+NeRF
Waymo	VectorNet 在 2020 年/Wayformer	矢量化+全模态统一 Token、注意力融合
Wayve	GAIA-1/2、LINGO	视频 Token 化器+世界模型，潜在状态与 VLA

数据来源：《VectorNet: Encoding HD Maps and Agent Dynamics from Vectorized Representation》，《GAIA-2: A Controllable Multi-View Generative World Model for Autonomous Driving》等其他文献，特斯拉、华为等公司公开资料，东吴证券研究所

4. 世界模型：标签稀缺下的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。这使得如何在缺少人工标签的情况下学到高质量表征，成为物理 AI 区别于文本大模型的关键问题。

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。需说明的是，本节讨论世界模型对自监督表征学习的作用；系列四将讨论其作为数据生成、仿真和评估基础设施的作用。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，均主张在潜在/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。

在驾驶场景，世界模型已分化出数据生成、场景重建与仿真预测等不同分工。GAIA-2 与 Cosmos 代表潜在空间世界模型方向；DriveDreamer4D、ReconDreamer 与 GeoDrive 则在 4D 重建与仿真中分别侧重数据生成、在线修复与动作可控。

表6：代表性驾驶世界模型对比

模型	核心定位	技术关键词
GAIA-2 Wayve	潜在空间世界模型	连续潜在、潜在扩散、动作/文本条件
Cosmos，英伟达，	世界基础模型	学习式 Token 化器、连续/离散、潜在扩散
DriveDreamer4D	数据生成器	世界模型先验、4D-GS、扩散合成
ReconDreamer	场景重建器	在线修复、几何先验
GeoDrive	仿真预测器	3D 几何约束、精确动作控制

数据来源：《GAIA-2: A Controllable Multi-View Generative World Model for Autonomous Driving》，《GeoDrive: 3D Geometry-Informed Driving World Model with Precise Action Control》等其他文献，东吴证券研究所

物理 AI 此前的演进更多集中在表征探索，且这一探索仍在继续。竞争壁垒正从单个感知模块精度，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。

表7：自监督潜在表征代表性工作

工作	机构	表征思路
JEPA / V-JEPA / V-JEPA 2	Meta	在潜在空间预测，自监督学抽象表征
Point-JEPA	学术界	面向点云的联合嵌入预测
Cosmos Tokenizer	英伟达	学习式连续/离散 Token 化器
LCDrive / Latent-CoT-Drive	英伟达等	在潜在空间而非文本中推理

数据来源：《Self-Supervised Learning from Images with a Joint-Embedding Predictive Architecture》，《Point-JEPA: A Joint Embedding Predictive Architecture for Self-Supervised Learning on Point Cloud》等其他文献，东吴证券研究所

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