东吴证券 SOOCHOW SECURITIES

汽车行业深度报告

AI 系列深度 19 期：多模态如何从模型拼接走向统一？

2026 年 08 月 07 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

- **多模态智能的本质，是把文本、图像、音频、视频等映射进统一表示，并在此基础上完成理解、推理、生成与交互。**我们认为，多模态是 AI2.0“架构、任务、模态三重收敛”在模态维度的集中兑现。
- **我们依据“模态对齐发生在流程的哪个阶段”，将多模态的发展过程划分三阶段：**2021—2023 年为外挂式/组合式阶段，CLIP、Flamingo、BLIP-2、LLaVA 等通过视觉编码器、投影层、Q-Former 或指令微调把图像接入语言模型；2023—2024 年进入原生多模态阶段，Gemini 代表多模态能力从接口层前移到预训练阶段；2024 年至今进入全模态 Omni 阶段，GPT-4o 将文本、视觉和音频输入输出纳入同一实时交互闭环。
- **沿三阶段看，主线是对齐位置不断前移、转换损耗逐级下降。**外挂式阶段工程效率高，但图像先被视觉编码器压缩再送入语言模型，细节、空间关系和时序信息存在损耗；原生阶段让文本、图像、音频、视频从训练早期共同参与表示学习，提升复杂真实任务中的泛化能力，同时显著抬高数据、训练系统和跨模态对齐门槛；Omni 阶段进一步将多模态从“统一预训练”推进到“实时统一交互”，补齐听、看、说的互动性。
- **我们认为，Omni 并非多模态发展的终点，理解与生成统一、多模态深度推理有望成为下一阶段两条核心主线。**其一，理解偏重高层语义，生成需保留空间结构与纹理细节，二者存在表征冲突，业界正沿纯自回归、扩散及二者融合等范式，探索共享编码或“编码解耦、主干统一”。其二，以 o1 为代表的深度推理已由语言扩展至图像、视频和具身场景，但当前仍以语言思维链为主；视觉工具增强、视觉—语言交错推理、多模态潜空间推理等方向正在加速发展，以视觉表示或多模态潜变量为原生载体的推理仍处于早期阶段。
- **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. 复盘多模态：基础模型如何统一语言、视觉与生成

1.1. 多模态的核心问题：统一表示下的理解、推理与交互

多模态大模型关注的核心问题，是模型如何统一处理文本、图像、音频、视频等不同模态，并在统一表示之上完成理解、推理与交互。这个界定包含两层含义：其一是“统一”，即不同模态需要被映射到同一套可供模型处理的表示中，而不是各自为政；其二是“理解、推理、交互”，即统一表示不只是为了识别或匹配，更要支撑跨模态语义理解、组合推理和实时交互。

人类智能天然是多模态的。人在认识世界时，视觉、听觉、语言、动作彼此交织、相互印证：语言中的“红色”、视觉中的红色光谱、动作中对红色信号灯的反应，指向的是同一个概念。机器智能若要逼近通用，同样需要跨越单一模态边界。多模态正是需要把这些分散能力统一到一起。

因此，理解多模态不能只看“模型能接收几种输入”。能接收图像输入的语言模型早已存在，但模态之间是松散拼接还是深度统一、对齐发生在流程的哪个位置、信息在跨模态转换中损失多少，才是区分不同技术路线的关键。

1.2. 多模态是“统一基础模型”在模态维度的兑现

在 AI1.0 时代，处理不同模态需要不同专用架构：图像用卷积网络，语音用隐马尔可夫模型或循环网络，文本用循环网络。模态之间几乎不共享表示，跨模态任务要靠人工设计流程，把多个专用系统串起来。Transformer 出现后，“把任意模态切成 token、喂进同一个注意力骨干”成为可能，这正是架构收敛与模态收敛能够同时发生的技术前提。

因此，多模态的发展史，可以视作“模态对齐由人工设计逐步升级为模型自行统一”的历史。外挂式阶段，对齐靠人工设计的连接模块完成；原生阶段，对齐前移到预训练，由统一模型在多模态数据上联合学习；Omni 阶段，对齐进一步前移到单一模型实时处理，所有模态进入同一个神经网络。

图1：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：东吴证券研究所

2. 多模态发展阶段：对齐前移与转换损耗下降

2.1. 三阶段框架：对齐位置前移、转换损耗下降

我们对 2021 年以来多模态大模型代表性学术成果进行了梳理，按“模态对齐发生在流程的哪个位置”将其归纳为三个阶段。需要说明的是，三个阶段在时间上有所重叠，且后一阶段并不意味着前一阶段方法被淘汰。例如外挂式结构至今仍是大量视觉语言应用的主流实现，三阶段刻画的是技术前沿的重心迁移，而非简单的新旧替换。

表1：多模态大模型发展三阶段框架

阶段	时间	核心问题	代表模型	阶段特征
一、外挂式/组合式	2021—2023	如何让图像进入语言模型体系？	CLIP、ALIGN、CoCa、Flamingo、BLIP-2、LLaVA	先分别训练视觉编码器、语言模型或图文对齐模型，再通过投影层、Q-Former、交叉注意力、指令微调等方式把视觉接入大语言模型
二、原生多模态	2023—2024	多模态能否从预训练开始统一建模?	Gemini	多模态不再只是后期外挂，而是从模型设计和训练阶段就面向文本、图像、音频、视频等统一输入
三、全模态 Omni	2024 至今	文本、视觉、语音能否由同一模型实时处理?	GPT-4o、Qwen2.5-Omni	从“图像输入+文本回答”走向文本、视觉、语音、视频的实时统一交互

数据来源：各模型技术报告与论文，东吴证券研究所

我们认为，原生多模态阶段始于 2023 年，可视为多模态大模型的元年。这一年起，多模态从“语言模型的外接功能”，开始转变为“基础模型的内生能力”。此时多模态才第一次成为头部基础模型从预训练阶段就共同面向的目标，而非事后接入的扩展。

2.2. 阶段一 · 外挂式/组合式：先把图像“接进”语言模型

2021—2023 年，多模态模型处于外挂式/组合式阶段，核心问题是如何让图像进入语言模型体系。主流方法是先训练好视觉模型、语言模型或图文对齐模型，再通过连接模块把它们拼接起来。该阶段的根本局限在于：多模态能力主要是“接上去”的，而不是模型从一开始就具备的。

表2：阶段一·外挂式多模态代表论文

时间	代表论文/模型	核心贡献	在多模态大模型中的位置
2021	CLIP	用大规模图文对进行对比学习，把图像和文本映射到同一语义空间，奠定开放词表视觉理解基础	图文语义对齐的起点
2021	ALIGN	用十亿级带噪图文对训练双塔对齐模型，证明大规模弱监督图文数据可行	CLIP 路线的规模化验证
2022	CoCa	结合对比损失与图像描述损失，使模型同时具备图文对齐和文本生成能力	从纯对齐走向“理解+生成”
2022	Flamingo	用感知重采样器与门控交叉注意力，把图像/视频接入冻结的语言模型，支持交错图文与少样本	“视觉模块+大模型”路线成型
2023	BLIP-2	用 Q-Former 连接冻结视觉编码器与冻结大模型，大幅降低视觉语言预训练成本	高效“视觉—语言桥接器”范式
2023	LLaVA	将指令微调引入视觉语言模型，用 GPT-4 生成视觉指令数据，形成开源视觉对话模型	多模态从“识别/问答”走向“助手式交互”

数据来源：各论文原文，东吴证券研究所

CLIP 是这一阶段的起点。它的核心贡献不是生成图像，而是用大规模图文对进行对比学习，把图像和文本映射到同一个语义空间。CLIP 使用约 4 亿组图文对训练，证明自然语言可以作为视觉概念的开放式标签，使模型具备零样本迁移能力。换言之，CLIP 解决的是“视觉世界如何接入语言语义空间”的问题，其图文对齐思想对 Flamingo、BLIP-2、LLaVA 乃至 Stable Diffusion 等大量后续模型有深远影响。其局限在于复杂推理、多轮对话和细粒度视觉解释能力不足。

Flamingo 解决的是视觉信息如何接入大语言模型，并让模型基于图像进行生成式回答的问题。它用预训练视觉模型与语言模型进行桥接，使模型能够处理任意交错的图像、视频和文本序列，并通过少样本提示适应视觉问答、图像描述等任务。其技术意义在于，多模态模型开始从“图文相似度匹配”，进入“视觉条件下的语言生成”。Flamingo 确立了外挂式多模态的基本形态：冻结或复用强大的单模态模型，再通过桥接模块组合成多模态模型。

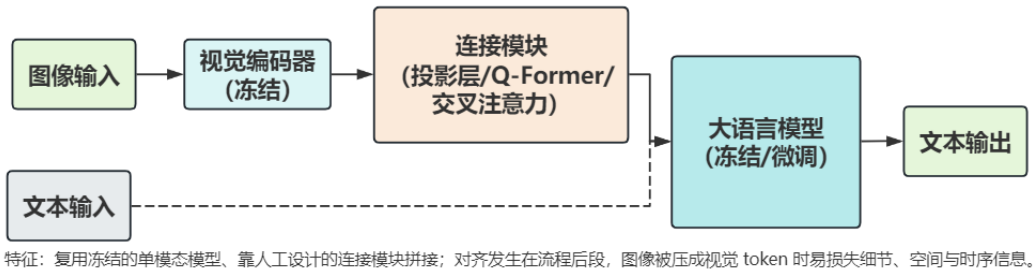
BLIP-2 解决的是降低视觉编码器与大语言模型连接成本的问题。论文认为，大规模端到端训练视觉语言模型成本过高，因此冻结已有图像编码器和大语言模型，只训练一个轻量级 Q-Former 来桥接视觉与语言，从而大幅降低视觉语言预训练门槛，成为后续大量开源多模态模型的基础结构。

LLaVA 将指令微调引入多模态模型，用 GPT-4 生成多模态指令跟随数据，并把视

觉编码器与大语言模型连接起来，训练出面向通用视觉语言理解的开源视觉助手。它标志多模态从“识别/问答”走向“助手式交互”，同时确立了“视觉编码器+投影层+大语言模型+指令微调”这一至今仍被广泛沿用的开源多模态范式。

外挂式阶段的共同特征，是把强大的单模态模型当作“零件”，靠人工设计的连接模块组装成多模态系统。这种做法工程上高效，也便于复用现成模型；但代价是模态对齐发生在流程后段：图像要先被视觉编码器压缩成视觉 token，再经连接模块送入语言模型，这一过程可能损失部分视觉细节、空间关系与时序信息。如何把对齐前移、减少这种损耗，正是下一阶段要解决的问题。

图2：外挂式/组合式多模态的典型结构



数据来源：东吴证券研究所

2.3. 阶段二·原生多模态：把对齐前移到预训练

2023—2024 年，Gemini 标志原生多模态阶段开启，推动多模态能力在预训练阶段统一形成。

在原生多模态路线出现前，图像往往要先被视觉编码器压缩成视觉 token，再通过连接模块送入大语言模型，这一过程会丢失一部分视觉细节、空间关系和时序信息。原生多模态路线则让文本、图像、音频、视频等模态在模型训练早期就共同参与表示学习，使模型更自然地形成跨模态对应与推理能力。两种路线的对比如下。

表3：外挂式多模态与原生多模态对比

维度	阶段一：外挂式/组合式	阶段二：原生多模态
构建方式	先有视觉模型和大模型，再连接	从模型设计与训练开始就面向多模态
能力来源	多模态是后期接入的能力	多模态是基础模型的原生能力
典型结构	视觉编码器+连接模块+大语言模型	强调统一预训练与跨模态联合建模
主要输出	以文本为主	朝多模态输入、多模态推理、多模态输出演进

数据来源：Gemini 技术报告及相关论文，东吴证券研究所

Gemini 的核心思路，是以 Transformer 解码器为主体架构，将文本、图像、音频、视频等不同模态转化为可被模型统一处理的序列表示，并支持多种模态在同一上下文窗口中交错输入。换言之，模型不再只是“先由视觉编码器理解图像、再把结果翻译成语言交给大模型”，而是从预训练阶段开始，就在多模态数据上学习不同模态之间的对应、组合与推理关系。

这一路线的关键变化，是多模态能力从“接口层能力”前移到“基础模型能力”。Flamingo、BLIP-2、LLaVA 等早期模型，本质上仍是“视觉接入语言中枢”，即在已有语言模型基础上，通过视觉编码器、连接器或指令微调把图像信息接进来。Gemini 则强调从底层训练范式上实现文本、图像、音频、视频的联合建模，使模型能在统一上下文中处理图文问答、图表理解、视频理解、语音理解以及跨模态推理。这意味着，多模态不再是语言模型的外部扩展，而逐步成为基础模型本身的内生能力。

2.4. 阶段三·全模态 Omni：把统一处理前移到实时交互

2024 年至今，GPT-4o 标志多模态大模型进入全模态 Omni 实时交互阶段，推动多模态能力从“统一预训练”进一步走向“实时统一交互”。

原生多模态阶段已经开始从训练早期统一处理多种模态，核心问题是多模态能力能否在基础模型内部共同形成。进入 Omni 阶段后，行业关注点进一步前移：文本、视觉、语音、视频能否由同一个模型实时处理，并在同一个交互闭环中完成理解、推理和生成。

GPT-4o 的核心思路，是以同一个全模态模型统一处理文本、视觉和音频的输入输出，使所有模态都进入同一个神经网络建模。过去的语音多模态系统通常采用“语音识别（ASR）—语言理解（大模型）—语音合成（TTS）”三段式流程：用户语音先被识别为文本，再交给语言模型理解推理，最后由语音合成模块生成回答。这一流程虽能实现语音交互，但本质仍是多个单模态或弱多模态模块串联，存在信息丢失与高延迟问题，难以达到单一模型原生多模态交互水平。GPT-4o 代表的 Omni 路线则把流程简化为：文本、图像、音频、视频进入同一个全模态模型，并由同一个模型生成文本、音频、图像等输出。两种范式的对比如下。

表4：传统串联式语音助手与全模态 Omni 路线对比

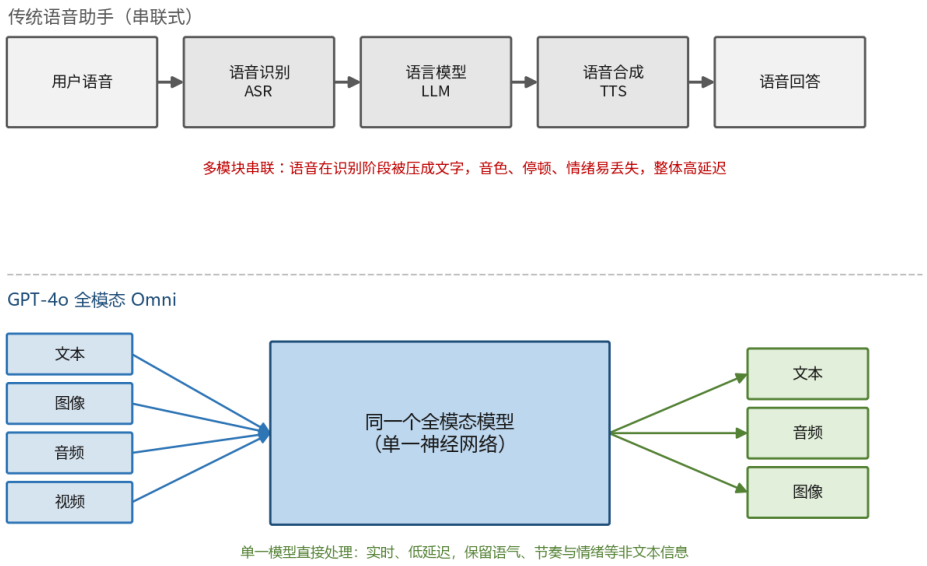
对比维度	传统语音助手：串联式流程	GPT-4o：全模态 Omni 路线
技术范式	多个模块串联组合	单一模型一体化处理
典型流程	语音→识别→语言理解→语音合成→回答	文本/图像/音频/视频→同一全模态模型→文本/音频/图像
模型结构	识别、理解、合成分属不同模型或模块	各模态输入输出由同一个神经网络处理
语音理解	先将语音转写为文本再理解	直接处理音频输入，保留语气、节奏、情绪等非文本信息
信息保留	语音在识别阶段被压成文字，音色、停顿、情绪易丢失	语音信息直接进入模型，保留细粒度交互信号
交互延迟	多模块串联，响应链条较长	单一模型直接处理，响应更接近实时
视觉参与	视觉作为额外输入模块，难与语音实时协同	可结合摄像头、屏幕、图像或视频进行连续对话
核心意义	实现了语音入口，但仍是模块组合	从“看图说话”走向“听、看、说、实时交互”

数据来源：GPT-4o 相关公开资料，东吴证券研究所

我们将 Omni 阶段突破归纳为四点：第一，语音直接理解，模型不再只依赖识别后的文本，而能利用音频中的非文本信息；**第二**，视觉实时参与，模型可结合图像、屏幕、摄像头或视频流内容进行对话；**第三**，低延迟对话，语音交互从“问答式等待”走向更接近人类交流的即时反馈；**第四**，多模态共同推理，文本、视觉、音频可同时作为推理条件，共同影响模型判断和输出。**由此，多模态大模型竞争焦点，从“能否理解多模态内容”，升级为“能否以自然、实时、低延迟的方式完成多模态交互”。**

2025 年，Qwen2.5-Omni 以单一模型统一处理文本、图像、音频、视频，并以流式方式生成文本和自然语音；其提出的 TMRoPE 用于对齐视频与音频时间戳，Thinker-Talker 架构则将“理解与文本生成”和“流式语音生成”分工协同。这使 Omni 路线从闭源产品走向有公开技术细节的可复现方案，也为国内厂商在该路线上的跟进提供参照。

图3：全模态 Omni 路线与传统串联式流程对比



数据来源：东吴证券研究所

3. 理解生成统一与多模态推理仍待验证

3.1. 全模态 Omni 路线的瓶颈

我们认为，Omni 路线在补齐“实时互动性”之后，主要面临以下瓶颈。

其一，理解与生成的表征冲突。“理解”任务，例如看懂一张图并回答问题，更需要高层语义表示，希望模型抓住“图里有什么、意味着什么”；“生成”任务，例如画出一张图，则需要保留空间结构与纹理细节，关心“每个像素长什么样”。在单一模型里同时做好二者，存在一定冲突。这正是 Janus 等模型要“解耦视觉编码”的根本原因：统一模型在理解侧和生成侧对视觉表示的需求并不一致。

其二，离散与连续模态生成范式的统一。文本通常以离散 token 进行自回归建模，而图像、视频及原始音频属于高维连续信号，当前高保真生成更多采用扩散或流匹配。尽管连续信号也可经 token 化后纳入自回归框架，但仍面临信息损失、序列过长与生成效率等权衡。因此，对于进一步走向理解与生成统一的多模态模型，如何在共享骨干中协调自回归与扩散/流匹配目标，或形成统一生成范式，仍是核心技术与工程难题。

其三，推理过程仍未多模态化。三阶段推进的是输入输出层面的统一，深度推理仍以语言思维链为载体。现有模型虽可将图像、视频等编码为视觉 token 并与文本联合输入，但多步推理的中间状态仍主要由语言思维链承载；随着推理链延长，视觉证据容易被语义压缩、注意力稀释或逐步遗忘，细粒度纹理、空间结构及时序关系因而难以持续参与推演。因此，多模态统一已由输入输出推进至交互层，并开始向推理层深入，但整

体仍处于由“多模态输入—语言推理”向“多模态证据持续参与—潜空间推演”过渡的早期阶段。

3.2. 方向一：理解与生成的统一——三条技术路线

针对“理解与生成表征冲突”这一核心瓶颈，业界提出多条代表性技术路线：

表5：理解与生成统一的代表性探索

时间	代表模型	技术路线	核心思路
2024	Chameleon	纯离散 token/早融合	把图像和文本都表示为离散 token，用统一 Transformer 从头训练，支持交错图文的理解与生成
2024	AnyGPT	纯离散 token	将语音、文本、图像、音乐统一转为离散 token，使多模态像“多语言”一样接入大模型
2024	Transfusion	自回归+扩散混合	在单一 Transformer 中结合“预测下一个 token”与扩散损失，同时处理离散文本与连续图像
2024	Show-o	自回归+离散扩散混合	用单一 Transformer 统一理解与生成，融合自回归与离散扩散，支持文生图、图像补全等
2024	Emu3	纯 next-token	将图像、文本、视频 token 化，仅靠“预测下一个 token”完成理解与生成，把多模态拉回大语言模型式自回归范式
2024	Janus	解耦视觉编码	解耦“理解用”与“生成用”的视觉编码，缓解统一模型中理解与生成对视觉表示需求不同的冲突
2025	Janus-Pro	解耦+规模化	在 Janus 基础上通过训练策略、数据扩展和模型规模提升，增强理解与文生图能力

数据来源：各论文原文，东吴证券研究所整理

归纳来看，主流路线有三条：纯自回归、自回归加扩散混合、解耦再统一。

纯自回归路线（Chameleon、AnyGPT、Emu3）把所有模态都离散化为 token，用统一的“预测下一个 token”来建模一切。其优点是与大语言模型的架构和训练范式高度一致，简洁、可规模化；代价是图像生成质量目前仍逊于扩散模型。Emu3 “Next-Token Prediction is All You Need”的命名，鲜明表达了这一路线“把多模态彻底拉回语言模型范式”的主张。

自回归加扩散混合路线（Transfusion、Show-o）在同一个 Transformer 中，文本走自回归、图像走扩散，把两种生成范式并轨。其思路是：离散模态用自回归，连续模态用扩散，在统一骨干内各司其职。2025 年出现的 BAGEL 即采用混合专家式 Transformer，理解侧用视觉编码器，生成侧用变分自编码器加修正流，在统一预训练中

让理解与生成相互促进，其理解能力可比肩同期开源视觉语言模型，图像生成质量可对标专用扩散模型。

解耦再统一路线（Janus、Janus-Pro）不强求理解与生成共用一套视觉表示，而是解耦二者的视觉编码，再共享同一个语言骨干，从而正面化解“表征冲突”。Janus 由国内团队提出，Janus-Pro 进一步通过数据与规模放大效果，是这条路线的代表。

图4：理解与生成统一的三条技术路线

数据来源：东吴证券研究所

3.3. 方向二：多模态推理——从语言 o1 到多模态深度思考

以 o1 为代表的测试时计算范式已由语言扩展至图像、视频和具身场景，多模态模型的发展重点也开始由“接入更多模态”转向“让多模态信息进入推理循环”。Omni 路线率先补齐了模型看、听、说的实时交互能力，新一代旗舰模型又开始引入可调节推理强度和快慢思考机制；但低延迟交互与增加测试时算力之间仍存在结构性权衡，当前多模态深度推理的核心载体仍以语言思维链为主。

2025 年以来，多模态推理沿四个方向加速演进：一是视觉工具增强，模型在推理过程中主动缩放、裁剪、标注或重新检视图像；二是视觉—语言交错推理，通过生成草图、中间图像等方式，将视觉由静态输入转变为动态“思维草稿”；三是多模态潜空间推理，以连续视觉表示或联合潜变量承载中间计算，减少语言转译造成的信息损失；四是面向视频和具身场景的闭环推理，引入时空记忆、世界模型与行动反馈，结合强化学习使模型依据环境结果调整决策。

现阶段，语言思维链与视觉工具结合已进入实用化阶段，而以视觉表示或多模态潜变量为载体的原生推理，以及面向长视频、三维空间和物理世界的因果推理仍处于早期；数据构造、过程监督、能力评测、奖励设计和实时推理效率尚未成熟。因此，下一阶段竞争焦点或将是视觉证据利用、时空记忆、预测—行动闭环与推理效率。

4. 结论：多模态收敛于同一个 Transformer，深度推理仍是关键变量

我们对多模态大模型从外挂式拼接、到原生统一、再到全模态 Omni 的发展历程进行了复盘，核心结论如下。

首先，多模态这轮升级的技术支点是 Transformer。统一的前提不是出现新的专用架构，而是“任意模态切成 token、进入同一个注意力骨干”的序列建模范式：模态之间如何对齐，不再依赖人工设计的接口，而是交给注意力在统一序列内学习。三阶段的实质，是模态进入同一个 Transformer 的时点不断前移——外挂式靠连接模块把压缩后的视觉特征接入语言模型，原生阶段让多模态数据从预训练起进入统一骨干联合建模，全模态 Omni 阶段由同一个模型实时处理全部模态的输入输出。对齐位置每前移一步，人工设计成分就减少一分，转换损耗随之下降。

其次，多模态下一阶段发展或将围绕两个核心问题展开，即理解与生成统一、多模态深度推理。

理解与生成统一仍存在技术瓶颈，其难点不只在于两类任务能否共享表征，也在于两种生成范式能否在同一 Transformer 骨干内协同。前者源于任务属性：理解偏重高层语义，生成需保留空间结构与纹理细节，同一套表征难以两头兼顾。后者源于模态属性：文本以离散 token 建模、天然适配自回归，图像、视频等连续信号的高保真生成更多依赖扩散或流匹配。当前纯自回归、扩散及二者融合路线并行演进，表征架构在共享编码与编码解耦之间持续探索，业界尚未形成单一表征、单一目标统一所有模态的共识。我们认为，短中期技术可能向"模态表征与生成模块适度解耦、语义与推理主干共享"的中间形态发展，从而实现模型理解与生成能力同步提升。

多模态原生推理有望成为下一阶段的核心演进方向。测试时计算范式已由语言延伸至图像、视频与具身场景，技术路线正由视觉工具增强、视觉—语言交错推理，走向多模态潜空间推理、世界模型与行动闭环；视觉信息由一次性输入，转变为推理过程中可反复读取、更新与验证的中间状态。现阶段，深度推理仍以语言思维链为主要载体，以视觉表示或多模态潜变量为原生载体的推理尚处早期，表示方式、训练目标与评测体系均未收敛。我们认为，随着多模态感知与交互能力逐步完善，行业竞争重点或将进一步转向多模态原生推理，关键在于视觉证据利用、时空记忆、预测—行动闭环与推理效率。

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。

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