增持（维持）

多模态技术加速，AI商业宏图正启

证券分析师：张良卫
执业证书编号：S0600516070001
联系邮箱：zhanglw@dwzq.com.cn

证券分析师：周良玖
执业证书：S0600517110002
联系邮箱：zhoulj@dwzq.com.cn

研究助理：郭若娜
执业证书：S0600122080017
guorn@dwzq.com.cn

2023年12月18日

核心观点

- **多模态：AGI必经之路与商业宏图起点。**（1）多模态是实现通用人工智能的必经之路。模态数据输入可帮助模型能力和用户体验提高，允许多模态数据输出也更符合真实世界需要。在数据、算法及算力上的要求都要高于单模态，这一波自然语言大模型发展为其他模态提供技术参考，行业有望加速发展。（2）多模态是AI商业宏图起点。多模态大模型有望真正为企业降本增效，且企业可将节省的成本用于提高产品/服务质量或者技术创新，推动生产力进一步提升；C端技术平权下内容创作有望达到一个成本与质量更优的均衡点，或出现新的、空间更大的UGC平台。
- **多模态大模型的技术脉络与前进方向：（1）视觉模型：**数据与算法往往同步发展，大型高质量数据集是模型突破重要基础，而近年视觉算法在泛化性、可提示性、生成质量和稳定性等方面突破将推动技术拐点到来以及爆款应用出现。**其中2D图像生成引领视觉模型前进方向，**由于2D图像生成是视觉模型中要求相对较低的领域，因此更容易实现技术突破，也出现了midjourney这类爆款应用，其兼顾使用门槛及生成效果，数据飞轮效应开始体现。文生图成本仍有优化空间，其中通用类应用由于需求相对刚性且有较强的付费意愿，盈利领先。3D资产生成、视频生成等领域受益于扩散算法成熟，但数据与算法难点多于图像生成，其中视频生成当前可类比2D图像生成的2021年（已有上亿规模数据集、扩散模型取得突破），且考虑到LLM对AI各领域的加速作用以及已出现较好的开源模型，2024年行业或取得更大的发展。3D资产生成则相对更加早期。**（2）听觉模型**：数据仍有缺口，23年以来技术有所突破。未来技术成熟后可为企业/内容制造商/娱乐应用提供高性价比的音乐作品，或基于娱乐属性向C端收费。**（3）具身智能：**相对远期，AI+机器人实现与现实世界交互。
- **海外技术领先，国内技术与应用同步发展。（1）海外：**OPENAI和谷歌在多模态领域布局的广度和技术先进程度上都处于领先地位，且都推出了表现较好的通用多模态大模型。而Stability.ai、 midjourney、runway等垂类独角兽也对技术突破和产品创新发挥重要作用。**（2）国内：**国内数据、算法、算力均有劣势，但海外算法开源有利于国内技术追赶；考虑到中国科技公司在产品运营和迭代方面实力更强，技术与应用有望同步发展。国内大厂及大模型公司均积极布局多模态，有望结合生态优势进行变现；万兴科技、美图等AI视觉应用公司亦有望受益于底层技术进步。
- **投资建议：**我们推荐在多模态方向已有布局或具备布局能力的标的：昆仑万维、万兴科技、美图，建议关注新国都；多模态技术进步利好电商、游戏、教育、营销等领域AI应用发展，推荐焦点科技、中文在线、盛天网络、蓝色光标、凤凰传媒、世纪天鸿等，建议关注掌趣科技等；建议关注受益于AI视频应用发展的多模态技术公司，如虹软科技、当虹科技等；算力方向建议把握板块龙头投资机会，推荐中际旭创等龙头。
- **风险提示：**多模态技术发展不及预期，伦理与隐私问题，商业化拓展不及预期，算力基础设施发展不及预期。

目录

1.1 多模态是实现通用人工智能的必经之路

■ **按照处理的数据类型数量划分，AI模型可以划分为两类：（1）单模态：**只处理1种类型数据，如文本等；**（2）多模态：**处理2种及以上数据，可类比人脑同时对文本、声音、图像等不同类型信息进行处理。

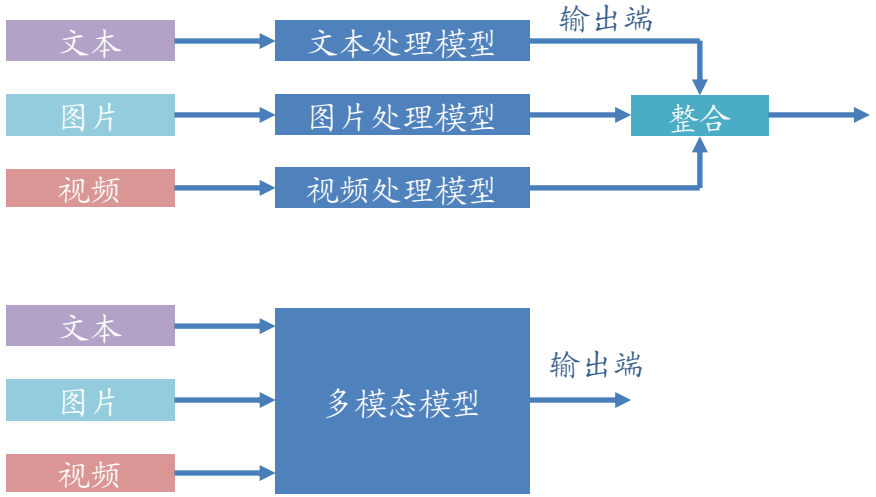
■ **多模态是实现通用人工智能的必经之路。**相比单模态，多模态大模型在输入输出端的优势明显：

- **输入端：1）提升模型能力：**高质量语言数据存量有限，且不同模态包含的信息具有互补性，多元的训练数据类型有助于提升通用大模型能力；**2）提高用户体验：**推理侧更低的使用门槛和更少的信息损耗。
- **输出端：更实用。1）**可直接生成综合结果，省去多个模型的使用和后期整合；2）更符合真实世界生产生活需要，从而实现更大商业价值。

表：主要的多模态大模型类型

类型	功能实现	主流模型代表
视觉模型	视觉理解、文生图、文生视频、文生3D等	DALL·E 3、Stable Diffusion、gen-2、PIKA1.0
听觉模型	音乐生成、音色转换、语音生成和音效生成等	Whisper、MusicLM、MusicGen
具身模型	感知推理、机器操作等	PaLM-E

图：相比单模态，支持多模态的输出更方便实用

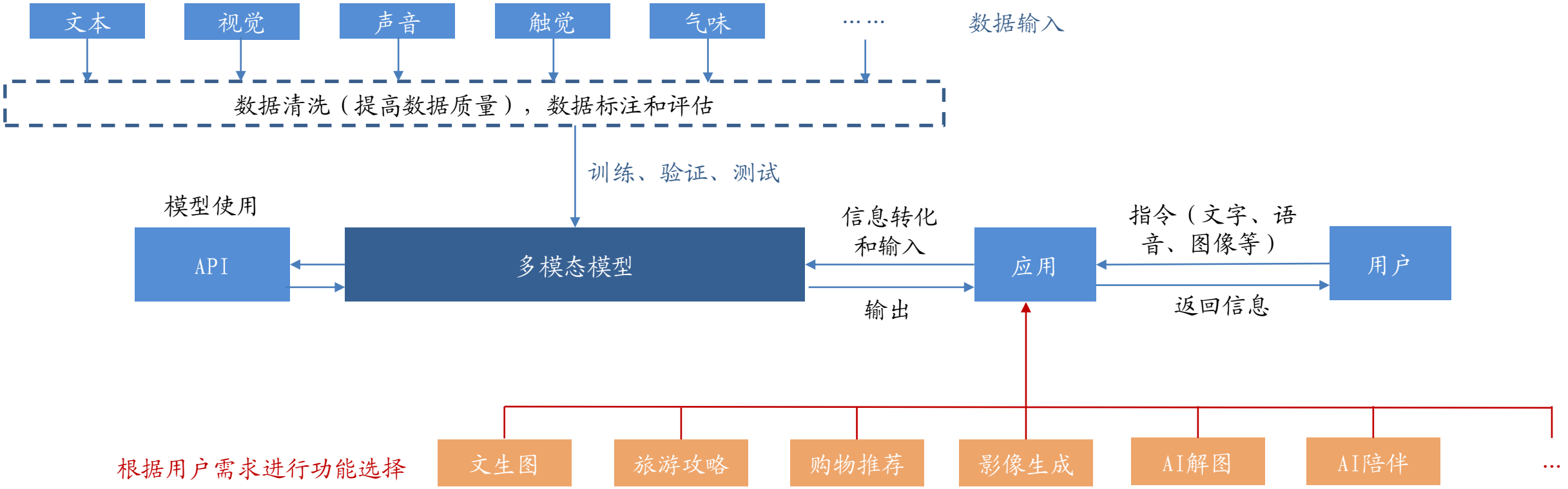


资料来源：东吴证券研究所整理

1.2 多模态大模型框架概览

- **数据：**文本、视觉、声音、触觉、气味等。
- **算法：**通过多模态统一建模，增强模型的跨模态语义对齐能力，打通各模态之间的关系，执行标准化的任务。
- **应用：**办公、电商、娱乐、教育等领域。

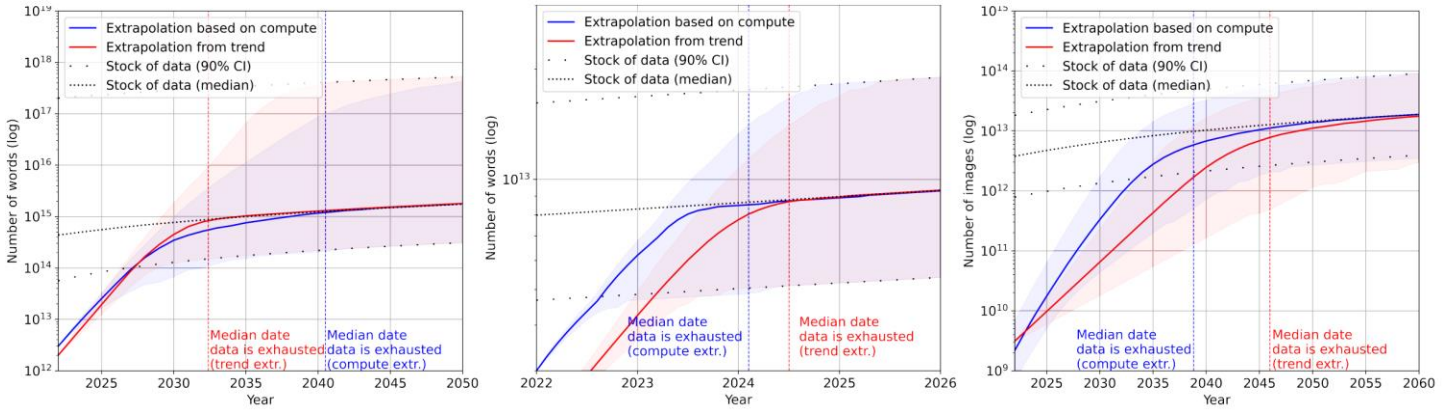
图：多模态大模型框架



1.3 数据：高质量多模态数据有限，合成数据发展或能改善

- **数据存量有限：**1）根据Epochai，在当前大模型高速发展趋势下，高质量语言数据可能在2026 年之前耗尽，而低质量语言/视觉数据存量也可能将在2030~2050/2030~2060年耗尽。2）**高质量多模态数据集有限：**由于不同类型的标注成本差异大，视觉等模态数据的收集成本比文本数据高，导致多模态数据集，尤其是高质量数据集通常比文本数据集少得多。
- **AI合成数据或有望改善数据枯竭问题。**1）与实际数据具有相同的预测特性。2)合成数据获取速度更快，为垂直模型的训练更快定制数据集。3）适应多模态模型的数据模态组合，能够扩大所有数据模态存量的组合，有效增加数据存量。

图：按照 ChatGPT大模型对数据消耗速率的不同数据使用量预测



低质量语言数据使用量预测　　高质量语言数据使用量预测　　视觉数据使用量预测

注：虚线为数据存量

表：合成数据产业分类与代表企业

合成数据产业	代表企业	企业服务
结构化数据	Mostly AI	提供合成数据生成器MostlyGenerate，能够生成与真实数据相当的匿名数据集
非结构化数据	DataGen	为计算机视觉团队提供设计合成数据集的自助服务平台。
测试数据	Informatica	提供具有合成数据功能的测试数据解决方案
开源服务	Twinify	用于为给定的敏感数据集生成1个保护隐私的合成数据集

资料来源：论文“Will we run out of data? An analysis of the limits of scaling datasets in Machine Learning”，天翼智库，东吴证券研究所

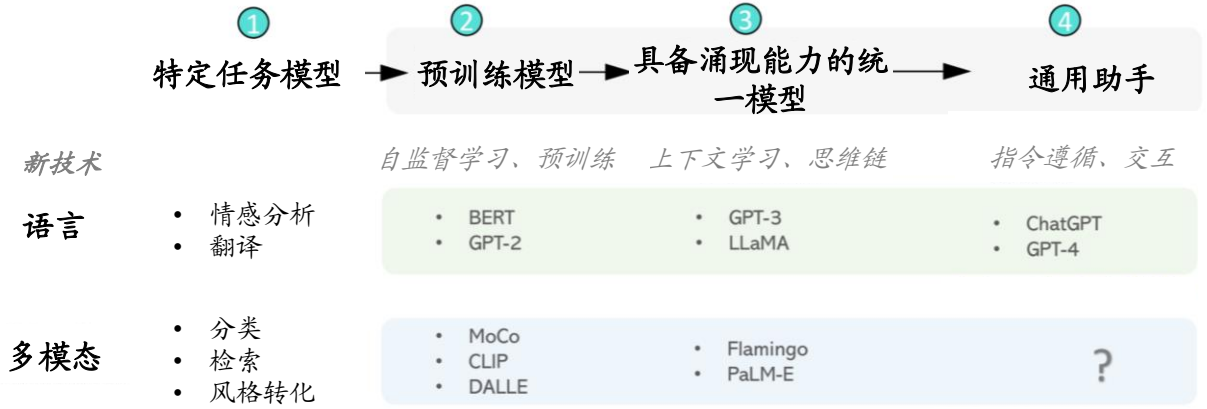
1.4 算法：技术要求更高，LLM发展提供突破口

- **相比单模态，多模态大模型算法和工程难度更大，**在表征、对齐、推理、生成、迁移、量化等环节均面临更多难点。
- **预训练为多模态主流训练方式。**由于高质量的多模态标注数据较少，基于Transformer结构的多模态预训练模型逐渐成为主流，通过海量无标注数据进行预训练，再使用少量有标注数据进行微调。原生多模态大模型是未来发展趋势，即设计时原生支持多模态，具有处理不同形式数据的能力，但各环节难度会更高。23年12月谷歌GEMINI即为原生多模态，一开始就在不同模态上进行预训练，利用额外的多模态数据进行微调以提升有效性，行业技术取得进一步突破。
- **这一波大语言模型发展给多模态带来新突破：**1）大语言模型LLM可充当大脑，处理各种模态信息，将其它模态信息对齐到LLM的语义空间。2）大语言模型在训练方式上给多模态模型提供前进方向参考，如自监督、预训练、上下文学习、指令遵循等。

表：多模态大模型算法面临的技术挑战

环节	难点
表征	对不同模态数据提取特征学习跨模态数据的信息交互，以及多模态数据的共同表示。不同模态间的异构性为表征带来挑战，如语言通常是符号表示，而语音通常是信号表示。
对齐	识别跨模态数据的联系。多模态之间数据异构，可能存在一对一、多对多、甚至不存在对齐关系
推理	结合知识，利用多模态对产与问题结构多步推理。如利用音频信号与视觉描述的嘴唇运动预测声音信号，容易产生不同的预测结果与噪声
生成	学习生成式过程，产生能够反映交叉模态相互关系、结构、连贯性，包括总结、翻译、创造等任务。
迁移	在不同模态的模型间进行知识迁移，尤其当主模态缺乏标注数据、输入噪声大时具有更大的困难，因为次模态信息的迁移会产生主模态从未见过的新行为
量化	多模态模型需要通过更深入的实证与理论研究，提高其在实际应用中的稳健性、可解释性和可靠性

图：自然语言基础模型开发轨迹为多模态的基础模型提供启发

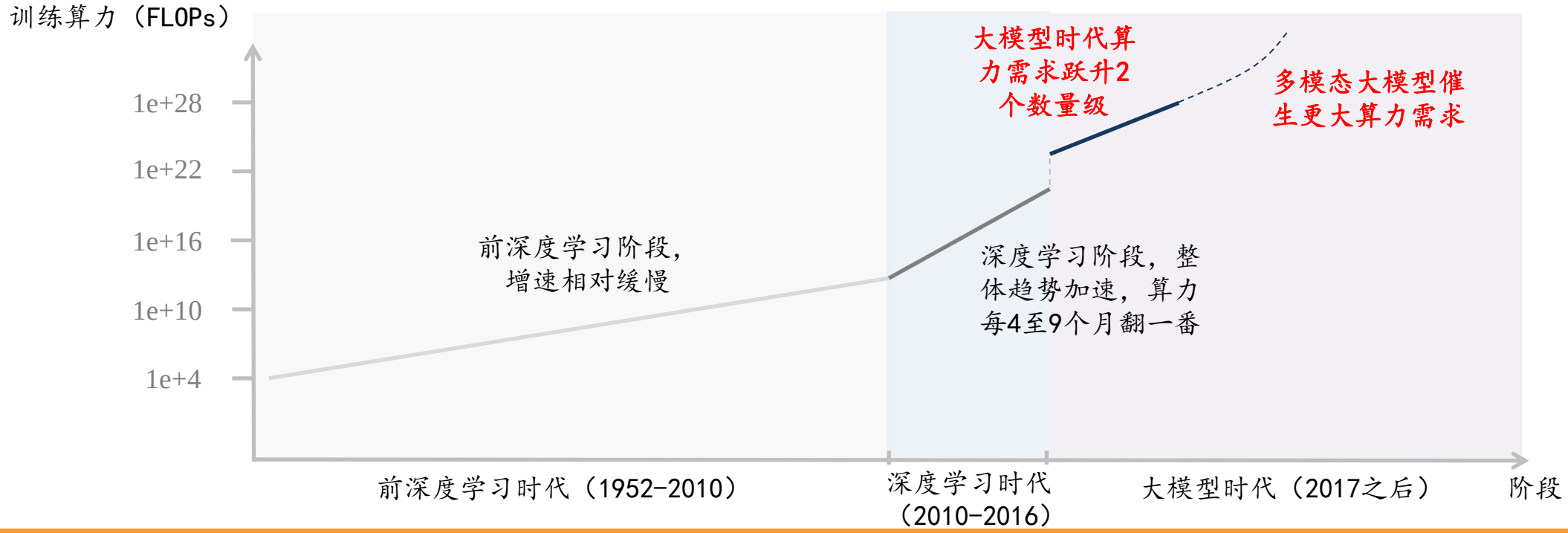


资料来源：CMU，论文“Multimodal Foundation Models: From Specialists to General-Purpose Assistants”，东吴证券研究所

1.5 算力：需求更大，催化产业新机遇

- **多模态大模型对算力的需求高于单模态。**一般在同样信息量情况下，文字数据量<图片数据量<视频数据量，多模态大模型需处理的数据量更大，再加上训练工程上难点更多，对应算力需求更高。参考前深度学习时代向深度学习时代过渡，以及从“大炼模型”进入“炼大模型”切换之后，算力需求均有明显提升。根据机器之心，谷歌Gemini有万亿参数，训练动用的算力是 GPT-4 的五倍。
- 未来随着算力需求的进一步提升，芯片制造、提供云服务以及模型微调的企业有望迎来更多发展机会。

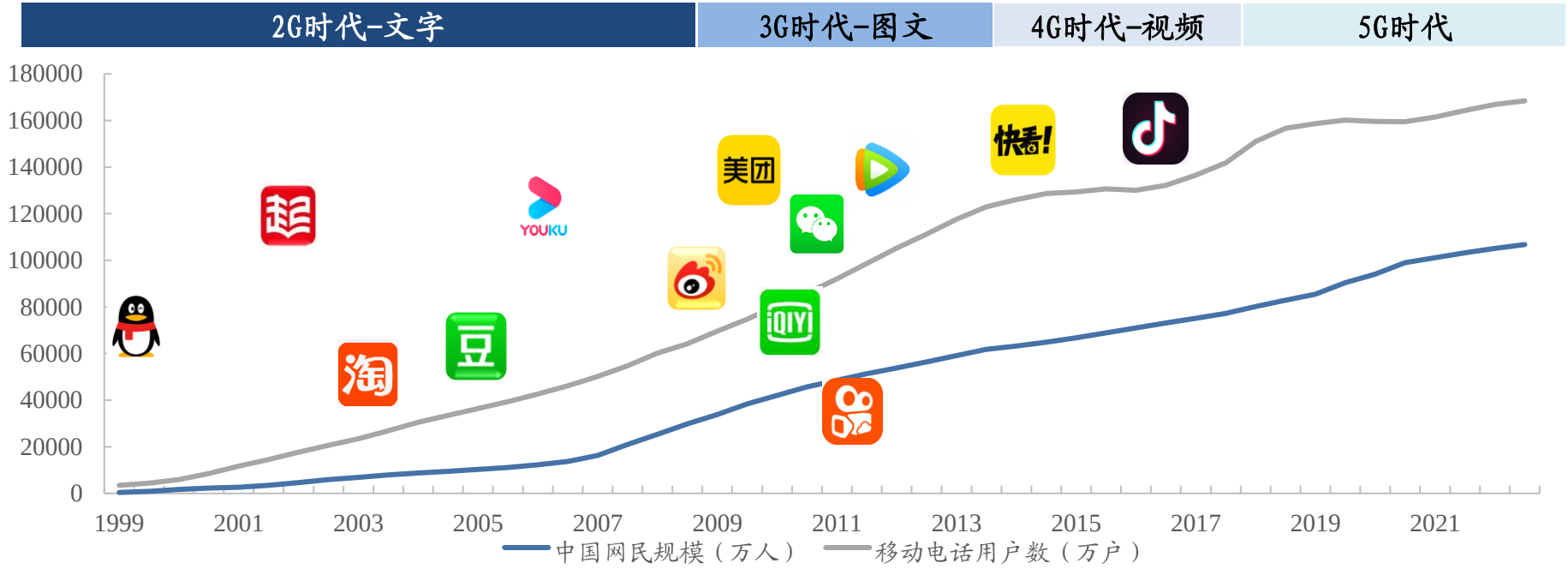
图：机器学习史上算力需求里程碑

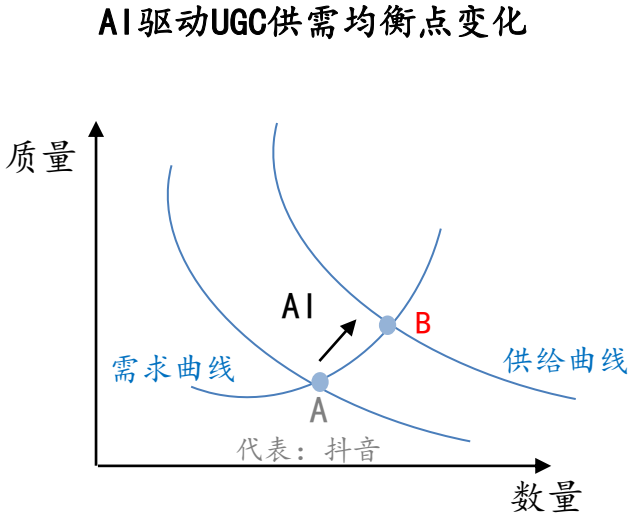


资料来源：论文“COMPUTE TRENDS ACROSS THREE ERAS OF MACHINE LEARNING”，量子位，CSDN，Google APIs，东吴证券研究所

1.6 多模态是AI时代真正的商业宏图起点

- **2B：**更符合真实世界生产需要，有望提高电商、营销、金融、教育等行业的生产力，真正为企业降本增效（我们在此前AI应用系列深度报告中已有较详细分析，此处不多赘述）；企业可将节省的成本用于提高产品/服务质量或者技术创新，推动生产力进一步提升。
- **2C：多模态大模型发展带来技术平权，C端内容创作达到一个成本与质量更优的均衡点，或出现新的空间更大的UGC平台。**过去UGC平台如小红书、知乎、抖音、快手等，用户创造内容的门槛每降低一倍，用户创造内容的数量会增加十倍，对应平台用户规模也会大幅增加。图像、视频、音频、3D资产等多模态技术进一步发展有望驱动AIGC时代真正到来。





资料来源：中国工业和信息化部，中国互联网联络中心，量子位，东吴证券研究所绘制

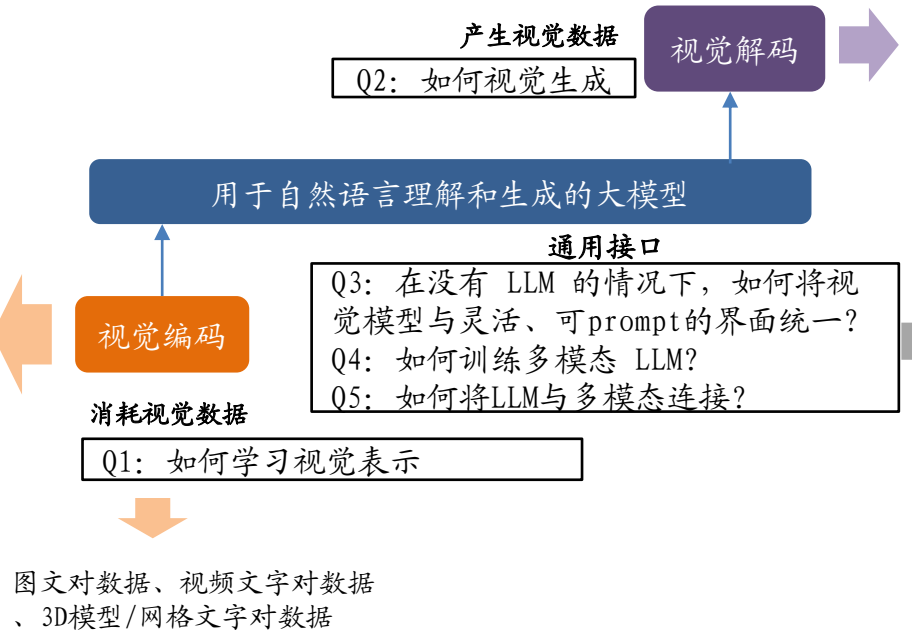
2 多模态模型的技术脉络与前进方向

2.1 视觉模型：数据与算法同步发展，图像生成引领方向

- **数据与算法往往同步发展，大型高质量数据集是模型突破重要基础，算法突破推动爆款应用出现。**
  - 现阶段多模态数据大多需要先用文本标注而非直接用于训练，相比文本数据集，图文对、视频文字对等数据集获取和标注工作量更大，大型高质量数据集的出现将为领域内算法突破奠定基础。
  - 算法在泛化性（21年CLIP，似GPT3时刻）、可提示性（22年Flamingo，似chatgpt时刻）、生成质量和稳定性（2021年扩散算法）等突破将推动技术拐点到来及爆款应用出现。
- **2D图像生成引领视觉模型前进方向。**由于2D图像生成在数据、算法、算力等方面是视觉模型中要求相对较低的领域，因此更容易实现技术突破，更早出现爆款应用（如Midjourney、dalle3），其也为3D资产生成、视频生成等领域提供技术参考。但考虑到后两个方向算法未完全收敛，尚未进入“大炼模型”阶段，距离真正的技术和应用爆发拐点还需要一定时间。

预训练方法		技术路径
图像骨干训练方法	标签监督	基于ImageNet、 ImageNet21K等数据集的监督
	语言监督	如CLIP（2021）、ALIGN，使用对比损失对从 Web 抓取的数百万甚至数十亿个嘈杂图像-文本对进行预训练，可实现零样本图像分类，并使传统CV模型能够执行开放词汇 CV 任务
	仅图像自监督	从图像本身挖掘监督信号中学习图像表示，用于掩蔽图像建模
多模态融合		CoCa、Flamingo
区域级和像素级图像理解		GLIP、SAM

图：视觉基础模型框架



领域	应用案例
2D图像生成	DALL-E系列、Stable Diffusion、Midjourney、Imagen、Parti
3D资产生成	DreamFields、DreamFusion、Magic3D、Point-E、Shap-E
文生视频	Imagen Video 、 Make-A-Video、gen-2、PIKA1.0

训练方式		应用案例
设计时原生支持多模态，具有处理不同形式数据的能力		Gemini
统一视觉模型	从特定视觉任务泛化到多个视觉任务	CLIP 、GLIP 、OpenSeg
	不同粒度级别上不同视觉理解任务的统一	UniTAB、Pix2Seq-v2、GLIP-v2、X-Decoder
使用LLM进行训练	将 LLM 的能力扩展到多模态并端到端训练模型	Flamingo、GPT-4
使用LLM链接工具	将ChatGPT等LLM与各种多模态基础模型相结合	Visual ChatGPT、MM-ReAct

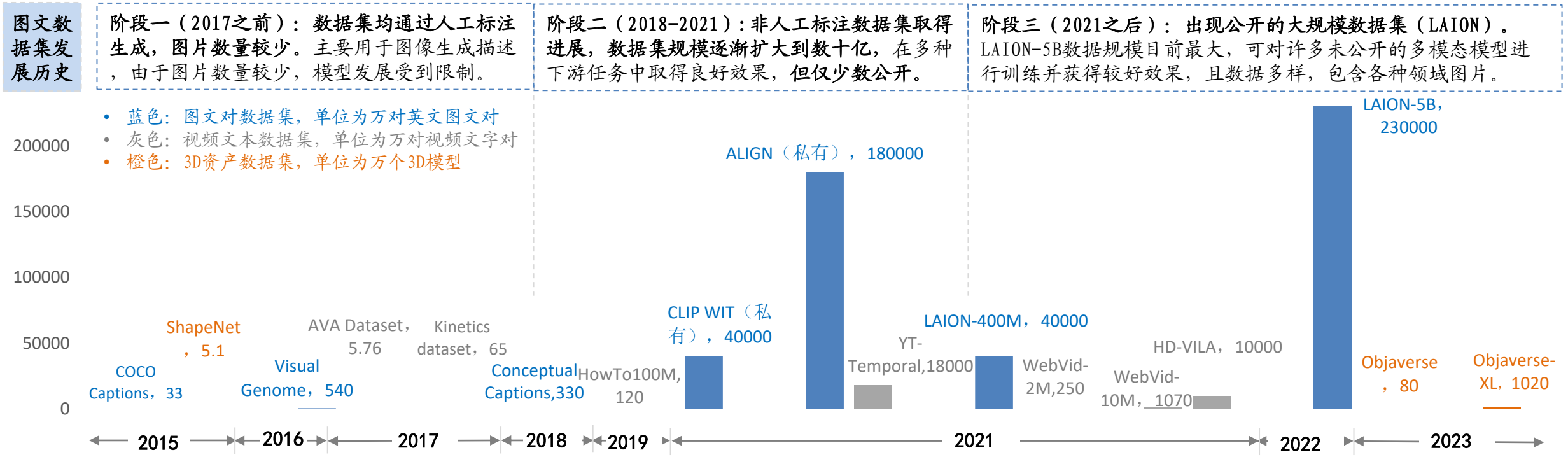
注：图像骨干（backbone）：对图像进行特征提取的网络

资料来源：论文“Multimodal Foundation Models: From Specialists to General-Purpose Assistants”，东吴证券研究所

2.1 视觉模型：数据与算法同步发展，图像生成引领方向

- **图像模型领域已具备大规模高质量的公开数据集，驱动文生图技术加速发展，也为其他视觉模型提供帮助。**2021年LAION-400M数据集发布，大小接近此前已有的私有图文对数据集CLIP，2022年5B版本发布，是目前已知且开源的最大规模多模态数据集，已用于训练当前最先进的文本-图像模型，包括Stable Diffusion等。其他类型视觉模型中也会采用图文对数据集进行训练。
- **视频领域已有上亿规模的高质量数据集，期待加速行业发展；3D领域则仍有待突破。**
  - 视频领域：2021年HD-VILA数据集实现规模、质量、多元性突破，数据集包含来自300万个视频中1亿个视频文本对，视频时长合计37万个小时，所有视频分辨率720p VS 主流视频文本数据集分辨率240p/360p；涵盖YouTube15个最流行的视频类别，如体育、音乐、汽车等。
  - 3D资产领域：数据集规模仍较小，尚未出现上亿规模的数据集，仍待突破。

图：不同视觉数据集规模发展对比

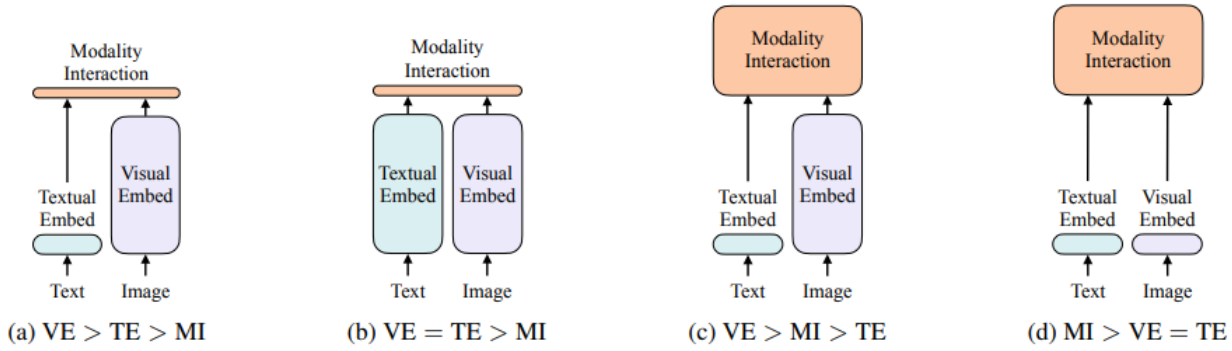


资料来源：OPENDATALAB，新智元，马达智数，Meriot论文，Objaverse论文，东吴证券研究所整理

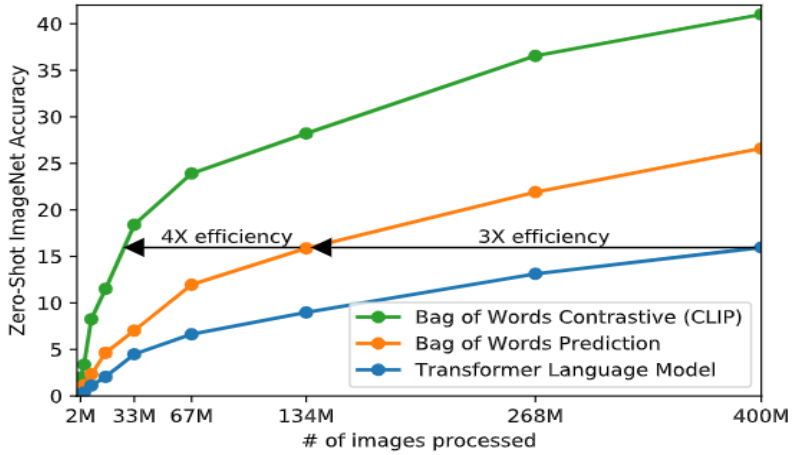
2.1.1 视觉理解：CLIP模型提供重要泛化能力

- **视觉理解模型可分为：（1）双塔模型：浅层语义交互。**对文本和图像分别编码后再输入多模态编码器，如CLIP。优点是适合检索任务，可检索大量文本与图像，缺点是不足以处理复杂分类任务。**（2）单塔模型：深层语义交互。**先将文本特征和图像特征连接后统一输入到多模态编码器，如ViLT，优点是可以充分将多模态信息融合，更擅长分类任务以及需要强交互的多模态理解任务，但不适合检索。**（3）混合模型。**
- **重要模型1——CLIP（2021年发布），第一个可通过零样本和少样本学习推广到多个图像分类任务的模型。**传统的视觉数据集创建成本很高，且任务泛化性差。OPEN AI创建了一个包含4亿图文对数据集，并借助大规模自然语言监督训练CLIP 模型，将不同模式、文本和图像的数据映射到共享向量空间，实现了可以用自然语言指示进行大量的分类基准，即 “Zero-Shot” 能力（将ImageNet上的zero-shot分类精度从11.5%提升到76.2%）。CLIP可识别图像、生成图像、回答与图像相关的问题，搜索与文字描述相符的图像，且其结构松散耦合，在保证学到多模态表征的基础上可随意拆分，从而将Encoder模块很好用到其他模型或者任务上，如Flamingo和LLaVa使用CLIP作为图像编码器， DALL-E用CLIP筛选生成的图像。

图：视觉编码模型的四个类别（b为CLIP，d为ViLT）



图：不同方法的 Zero-Shot ImageNet-1K 精度

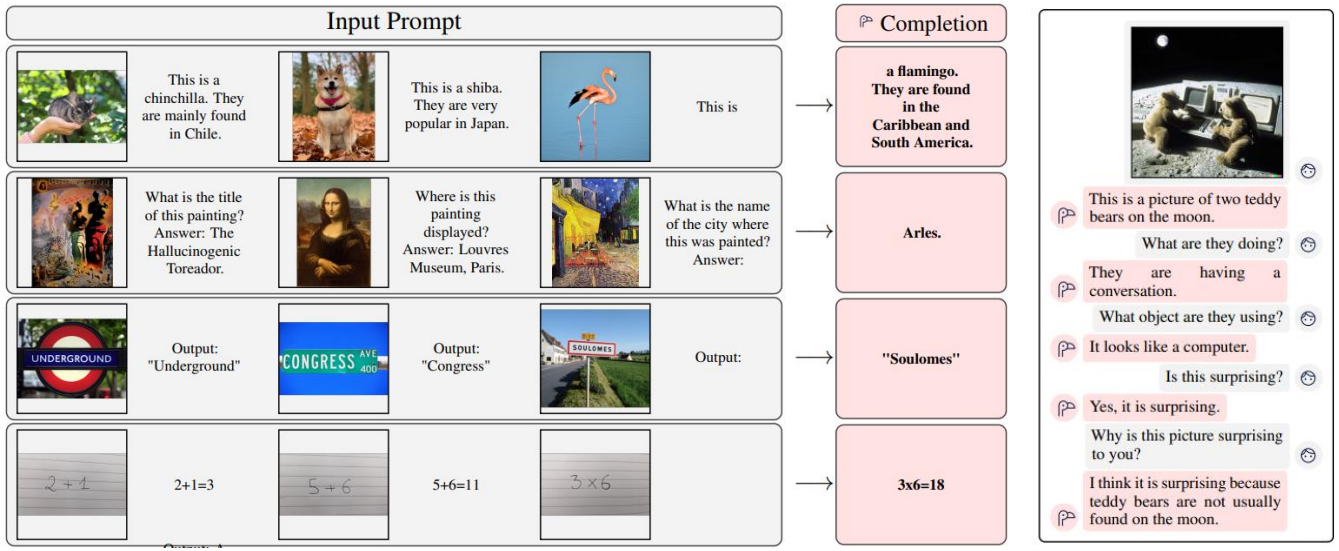


资料来源：ViLT论文，Clip论文，东吴证券研究所

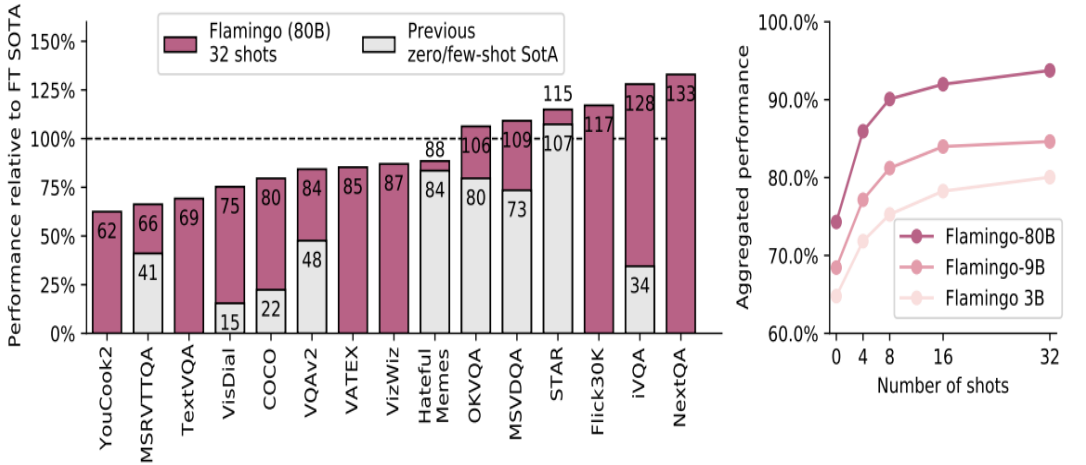
2.1.1 视觉理解：Flamingo推动预训练+微调 转向 预训练+prompt

■ **重要模型2——Flamingo（2022年发布），在广泛的开放式多模态任务中建立了少样本学习新 SOTA。**DeepMind发布的Flamingo架构中包括一个预训练语言模型（DeepMind的Chinchilla）+预训练Vision Encoder（DeepMind NFNet-F6，采用CLIP对比损失在图像文本对数据集上预训练的）+Perceiver Resampler模块实现最终输出固定长度特征。Flamingo可在多种开放式视觉和语言任务中实现快速学习（文本描述补全、VQA / Text-VQA、OCR、数学计算、文本描述、物体计数、语言文本混合理解、人物常识等等），不需要微调，同时在大部分多模态任务上能实现和 GPT-3 一样的 In-context few-shot推理能力。在众多基准测试中，Flamingo 的表现优于在数千倍于特定任务数据的基础上进行微调的模型。根据机器之心，行业内推测OPEN AI最新发布的多模态模型GPT4-V是一个类似 Flamingo的架构。

图： Flamingo 可通过少量prompt快速适应各种图像/视频理解任务（左），还能进行多图像视觉对话（右）



图： Flamingo 在16 项未微调任务中的6项表现优于最先进的微调模型。在9项few-shot任务中，Flamingo 表现最优



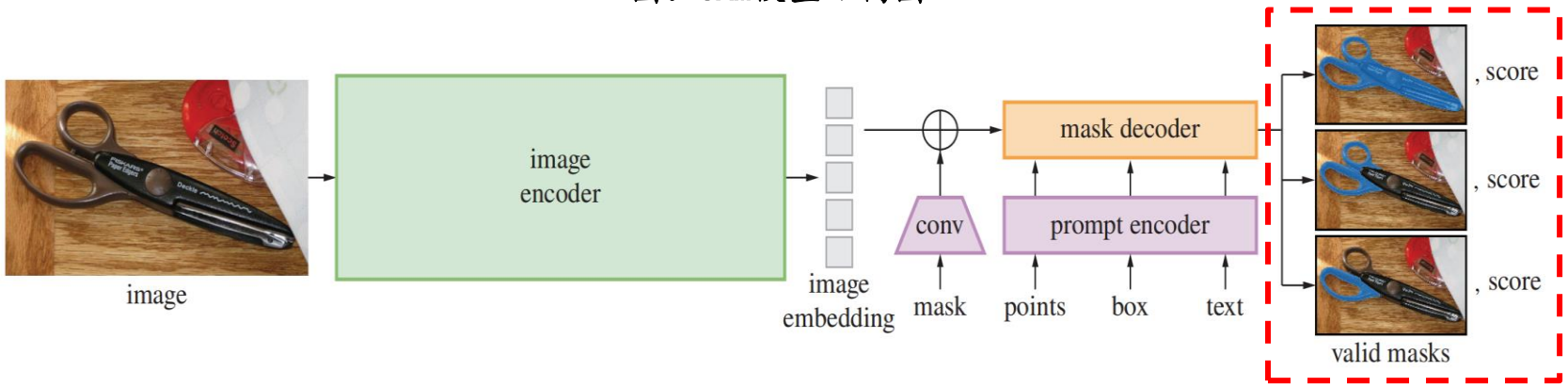
2.1.1 视觉理解：SAM，用prompt分割一切

- **重要模型3——SAM（2023年发布），第一个致力于图像分割的基础模型，零样本分割一切。**图像分割是计算机视觉的核心任务之一，是指识别目标、并沿目标边缘进行区域分割的技术，此前方法大致分为：1）交互式分割，允许分割任何类别的对象，但需要专家通过迭代细化掩码指导该方法。2）自动分割，允许分割提前定义的特定对象类别，但需要大量手动注释对象训练。SAM 模型的可提示界面允许用户以灵活方式使用，只需为模型设计正确的提示prompt就可完成范围广泛的分割任务。
- META受到语言模型中prompt的启发，训练了基于prompt的视觉 Transformer（ViT）模型，视觉模型是在一个包含来自1100 万张图像的超过 10 亿个掩码的视觉数据集 SA-1B 上训练的。SAM 可以为任何 prompt（点击、boxes、文本等）返回有效的分割掩码，完成范围广泛的分割任务。
- Meta 预计，与专门为一组固定任务训练的系统相比，基于 prompt 工程等技术的可组合系统设计将支持更广泛的应用，SAM 可成为 AR、VR、内容创建、科学领域和更通用 AI 系统的强大组件，如 SAM 可以通过 AR 眼镜识别日常物品，为用户提供提示。但由于 SAM 中的 ViT-H 图像编码器参数较大，实际使用的计算和内存成本还较高。

图：SAM图像分割效果展示

SAM 可通过单击或交互地选择点来分割对象以包括或排除对象，可通过使用多边形工具绘制边界框或分割区域创建分割，它会捕捉到对象。
当在识别要分割的对象时遇到不确定性时，SAM 能生成多个有效掩码。
SAM 能为图像中存在的所有对象自动识别和生成蒙版。
在预计算图像嵌入后，SAM 可立即为任何提示提供分割掩码，从而实现与模型的实时交互。

图：SAM模型结构图



2.1.2 视觉生成：文生图技术开始收敛至扩散算法，应用有望加速

- 主流生成算法包括VAE、GAN、Diffusion等，生成算法与视觉理解算法可实现多种组合关系。如OPENAI 经典文生图模型DALLE包括三个独立训练的模型：dVAE（decoder），Transformer（encoder）和CLIP（筛选）。
- 目前2D图像生成是相对成熟的应用方向，一方面受益于大规模公开数据集和表征模型等基础环节的进步，另一方面也受益于生成算法中扩散模型的突破。目前基本大部分文生图模型/应用的decoder环节都用Diffusion扩散算法，如DALLE 2 & 3（OPENAI）、Stable Diffusion（stability.ai）、Midjourney等。

随着技术收敛，**提高模型参数正成为趋势**，如：stability.ai最新模型SDXL 0.9 参数量较 Beta 版本显著增加，是目前所有开源图像模型中参数量最大模型之一，基础模型35 亿参数+模型集成管线66 亿参数；而beta 版权只有31 亿参数并使用单一模型。

图：主要生成模型发展情况

- 主流技术：GAN、VAE等
- 常见应用：人脸生成、风格迁移、超分辨率、图像补全和可控图像编辑。
- 图像生成/编辑网络与文本的多模态交互非常有限。

- 20年DDPM在更加庞大的数据集上展现出与GAN模型相媲美的性能，扩散模型理论逐渐成熟
- 21年扩散生成技术进一步突破、出现公开大规模多模态数据集（LAION）和多模态表征模型（CLIP）
- 22年Midjourney、开源模型Stable Diffusion走红
- 23年大模型微调技术（ControlNet和LoRA）推动自定义调整、扩展AI模型，更好适应不同的具体应用（如二次元风格化、logo生成等）

导入期 | 爆发期

	计算量	效果	应用	2014	2015	2016	2017	2019	2020	2021	2022	2023
VAE 变分自编码器	较大	图像模糊	较少作为独立主干网络结构设计，可组合使用	VAE			VQ-VAE	D-VAE		DALLE		
GAN 生成对抗网络	低，但较难训练	图像连续性强，但常遇到模式坍塌和不稳定等问题，生成数据多样性较差	3D生成科研中热门候选技术之一	GAN	CGAN 增加生成类别条件	StackGAN 文生图模型	CycleGAN 图生图模型 StackGAN++	styleGAN		VQ-GAN 文生图模型	Parti	
Diffusion 扩散模型	大	质量高，但连续性较差	2D文生图主流技术，3D生成和视频生成领域有所应用		Diffusion 扩散模型				DDPM	Latent Diffusion	midjourney DALLE2 Imagen Stable Diffusion	DALLE3 Imagen 2

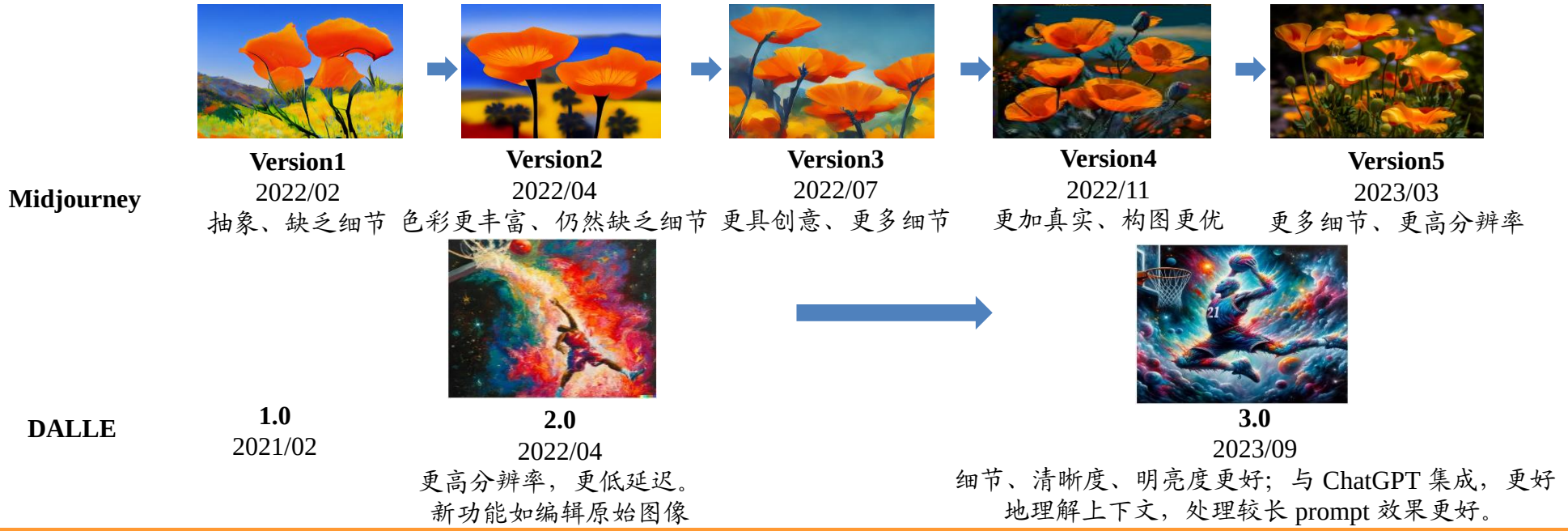
资料来源：半轻人，智源社区，硅兔赛跑，别有一洞天，新智元，半亩塘，东吴证券研究所

2.1.2 2D图像生成：开源模型引领生态，闭源应用飞轮加速

■ **海外文生图应用已初显生态，可分为开源模型和闭源应用两类。**

- **基于开源模型stable diffusion的开发者生态百花齐放。**Stability.ai的开源模型Stable Diffusion加上LoRA等插件即可实现用少量图片训练文生图模型。根据智源社区， Stable Diffusion已有超过 20 万开发者下载和获得授权，各渠道累计日活用户超过 1000 万。团队开发的付费在线平台DreamStudio获得超过 150 万用户，生成超过 1.7 亿图片。很多垂类文生图应用，如专注二次元形象生成的 NovelAI、专注头像生成的 Lensa、AI写真的妙鸭都是在Stable Diffusion 微调得到。
- **闭源应用兼顾使用门槛及生成效果，数据飞轮效应开始体现。**Midjourney、Dalle、adobe Firefly等应用只需输入prompt即可输出精美图画，无需微调，做到使用门槛和生成效果的平衡。其中Midjourney是先于Stable Diffusion推出的基于扩散模型的应用，有一定先发优势，通过不断收集用户反馈数据（每次生成4张图片，用户可以让模型再次修改任意图片），反向推动技术迭代，提供更好的产品体验，实现数据飞轮效应。截至23/12/13，服务器成员数量超过1700万人，是Discord上最大服务器；根据海外独角兽，其年收入已超过1亿美金。

图：闭源应用Midjourney和dalle产品迭代情况



资料来源：智源社区，midjourney，PaperWeekly，海外独角兽，东吴证券研究所

2.1.2 DALLE3 vs Midjourney vs Firefly vs Imagine测试对比

■ AI文生图应用测试对比：

- **DALLE 3：** OPEN AI表示“比以往系统更能理解细微差别和细节，让用户更加轻松地将自己的想法转化为非常准确的图像”。我们实测中语义理解能力优秀（OPENAI强项）；但图片细节、构图、精美程度相对较弱。
- **Midjourney：** 公司参考 CLIP及Diffusion，抓取公开数据训练的模型。我们实测中画面精美程度和细节表现佳，语义理解能力不如DALLE3。
- **Adobe Firefly：** 集成到Photoshop中，支持多人协作、在线评论等。我们实测中语义理解能力最弱；但写实表现优，构图、色彩能力突出。
- **META Imagine：** 基于Meta的Emu学习模型， Emu是使用11亿图像-文本对数据集对28亿参数的UNet进行预训练得到，再使用几千张高质量图像进行质量微调提高模型效果。我们实测中生成效果中等，但产品功能较少难以对同一张图片进行后续调整。

注：受测试次数限制，对比结果仅供参考，可能与实际情况存在差异

图：主要文生图应用对比（左上DALLE3 ，右上Midjourney，左下adobe Firefly ，右下META Imagine）

prompt1（超现实）：黑暗美学，带有齿轮和电线的未来主义机械机器人，在高峰时段拥挤不堪的地铁中上下班，令人惊叹的照片

Prompt2（复杂场景）： 微缩丰收场景、四川火锅店厨师制作食物、工作中的人、创意摄影海报、微缩摄影、美丽照明、重点照明、全球照明超现实、超精细、8K、简单构图

Prompt3（写实）：电影般，令人惊叹，简约，薄的轻量级的光,可爱的红色蓬松的狗在大雪纷飞的纽约街头坐在梅西百货公司门口。8K。壁纸。

资料来源：adobe Firefly，chatgpt，discord，META imagine，DataLearner,东吴证券研究所。注：为保证测试结果有效性，我们测试prompt为英文，选用第一次生成的图片中表现最好的一张。

2.1.2 商业模式：成本仍有优化空间，通用类应用盈利领先

■ **2D图像生成模型训练成本可控，推理成本有下降空间：**

- **训练成本低于通用大模型：**Stable Diffusion训练成本60万美元，对比GPT3单次训练成本460万美元；
- **推理成本：**Midjourney生成一张图片成本约0.5美分，未来随着模型成熟及底层硬件迭代，推理成本有望进一步下降。
- 其他费用：以Midjourney Discord服务器为例，Discord额外收取约10%的手续费。

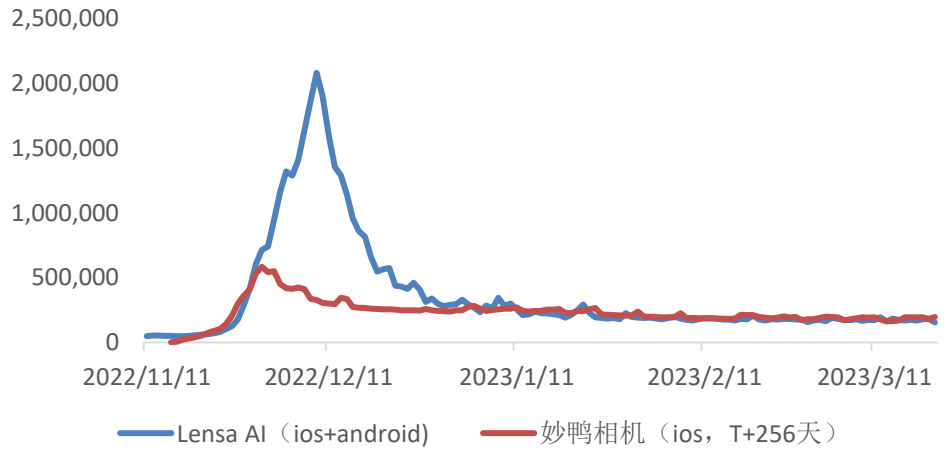
■ **文生图产品基本采用订阅制变现，但受众差异带来不同的盈利难度。**

- 初步验证模式跑通的是Midjourney，定价模式高于开源模型（完全能覆盖成本），22年8月已实现盈利，主要面向有文生图刚性需求、付费意愿较强的用户，如专业设计师、电商从业者等。即使是人力成本相对较低的国内，商家找淘宝模特拍一张商品图片成本可能在200-500元左右，相比之下Midjourney更有性价比。
  - Midjourney技术快速迭代，不断逼近开源模型调优后的表现，同时通过功能扩展给予创作者更大的发挥空间。
  - Adobe Firefly、DALLE 3等在生态上更胜一筹，未来技术进一步改进后亦有可观前景。
- 开源模型生态价值>变现价值：Stable Diffusion面向中小创作者或者创业团队，每月价格5-15美金，定价低但客户付费持续性较差，部分客户调优出自己的模型之后会流失，目前仍处在亏损状态。
- 众多垂类文生图应用，短期盈利或不是难点，核心问题在于用户留存，如Lensa、妙鸭等垂类文生图应用都在短时间内出现用户数和流水爆发，但较窄的应用范围加上需求刚性不足使其只能“一波流”，短期数据难以维持。

表：Midjourney收费模式

	basic	Standard	Pro	Mega
月度	$10	$30	$60	$120
年度	$96 ($8/月)	$288 ($24/月)	$576 ($48/月)	$1152 ($96 /月)
GPU加速	3.3 小时/月	15 小时/月	30 小时/月	60小时/月
多任务处理	同时进行3项任务 等待区10项任务	同时进行3项任务 等待区10项任务	同时进行12项加速任务和3项常速任务 等待区10项任务	同时进行12项加速任务和3项常速任务
版权归属	拥有商用版权	拥有商用版权	拥有商用版权	拥有商用版权

图：LENSA及妙鸭日活曲线有一定相似之处（单位：人）



资料来源：AI前线，海外独角兽，机器之心，SENSOR TOWER，Lambda Labs首席科学官Chuan Li,AI科技评论，稀土掘金，midjourney，东吴证券研究所

2.1.3 视频生成：可类比图像生成的2021年，期待24年发展

- 视频生成包括文生视频、图生视频等。相比图像生成，视频生成模型训练面临更多难点：**1）算力和存储需求高：**视频比图像更大，训练时需更大GPU内存，推理时生成大量帧，确保帧间空间和时间一致性会产生长期依赖性，计算成本更高。**2）大规模高质量数据集仍较少；3）技术复杂，控制难度高。**需考虑流畅性、动作、逻辑问题；涉及到更多空间维度，当视频很长时，确保每一帧都协调一致相当复杂；prompt过于简单难以为每一帧都提供详细的描述。
- **2023年以来技术加速突破，2024年或有望取得更大发展。**1）2021年受GPT3和DALLE启发，行业开始采用Transformer 架构，出现了Make-a-video等只需prompt即可生成视频的模型；2）2022年扩散模型从图像扩展到视频领域，相关研究论文数量从22年的14篇显著增加至23年前10个月的103篇，23年GEN-2、PIKA1.0等视频模型效果取得明显突破，开源玩家入场，共同推动视频生成行业加速发展。虽然由于数据、算法等难点，行业技术仍未收敛，生成效果仍有提升空间（仍有生成痕迹，流畅度/清晰度/时长/动作复杂度不够），但我们认为视频生成的2022-2023年可类比2D图像生成的2021年（已有上亿规模的数据集、扩散模型取得突破），且考虑到LLM对AI各领域的加速作用以及已出现较好的开源模型，2024年行业或取得更大的发展。

图：视频生成技术复盘

	第一波浪潮	第二波浪潮（2021至今）	第三波浪潮（2022至今）
优缺点	生成视频速度快，但质量和分辨率低，长度短，控制能力弱，应用范围有限。	可捕捉上下文，对长视频建模更好，可实现细粒度语义控制，**但计算量大**。	在生成多样性和质量方面表现更佳，**相比transformer计算量更小**，可与transformer架构结合
主要模型	Text2Filter TGANs-C	2021：VideoGPT 2022：Make-a-video(META，文生图进化)、Phenaki（谷歌，交互生成视频,2min以上长镜头）、NUWA、CogVideo	2022： 谷歌Video Diffusion Models (VDM) + Imagen Video（更高清） 2023： Runway Gen1+Gen2、微软NUWA-XL、字节MagicVideo等
底层技术	GAN、VAE	受GPT3和DALLE启发，采用Transformer 架构	受stable diffusion等文生图启发，采用扩散架构

阶段

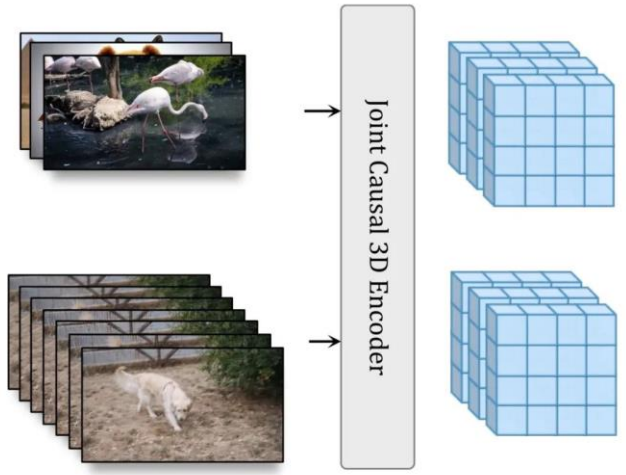
资料来源：HUGGING FACE，Bay的设计奥德赛，复旦大学，东吴证券研究所

2.1.3 视频生成：W.A.L.T新框架或带来技术突破口

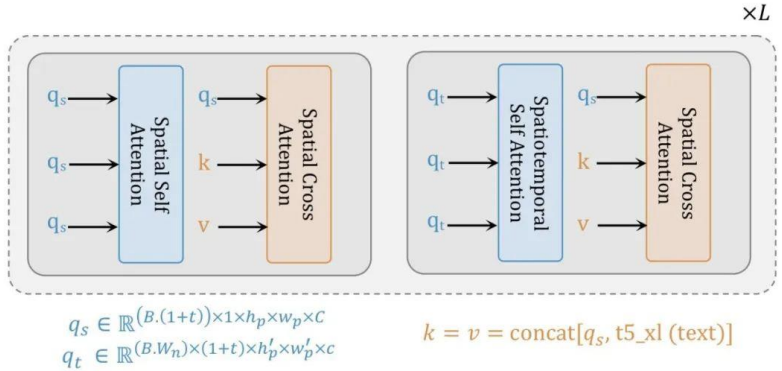
- 12月12日，李飞飞及其学生团队与谷歌合作，推出AI视频生成模型W.A.L.T（窗口注意力潜在Transformer ，Window Attention Latent Transformer），可以以每秒8帧的速度生成512 x 896分辨率的视频，支持文生视频、图生视频、和3D相机拍摄视频等方式。
- **W.A.L.T将Transformer 与扩散算法结合，同时改善计算成本和数据集问题。**Transformer在处理视频等高维数据时成本过高，W.A.L.T将Transformer架构与潜在扩散模型（Latent Diffusion Models，LDM）结合，在一个共享潜在空间中压缩图像和视频，一方面降低Transformer的计算要求，提高训练效率；另一方面能同时在图像和视频数据集上进行训练（W.A.L.T使用来自公共互联网和内部来源约970M文本-图像对，和约89M文本-视频对的数据集），有望为视频生成模型训练增加可用数据集。
- 团队基于W.A.L.T训练了三个模型的级联（Cascade），用于文本到视频的生成任务，包括一个基本的潜在视频扩散模型、两个视频超分辨率扩散模型，实现了无需使用无分类器指导的情况下，在视频生成基准UCF-101和Kinetics-600、图像生成基准ImageNet测试上SOTA 。

图：W.A.L.T的文生视频示例

图：W.A.L.T使用因果编码器在共享潜在空间中压缩图像和视频



图：W.A.L.T使用基于窗口注意力的Transformer架构进行潜在空间中的联合空间和时间生成建模



资料来源：W.A.L.T论文，东吴证券研究所

2.1.3 视频生成：格局尚早，期待各类玩家共同推动行业前进

■ 目前在视频生成领域表现较好的是runway、pika等独角兽，在语义理解能力、生成的视频画质、精美程度、画面一致性等方面有各自的优劣势。

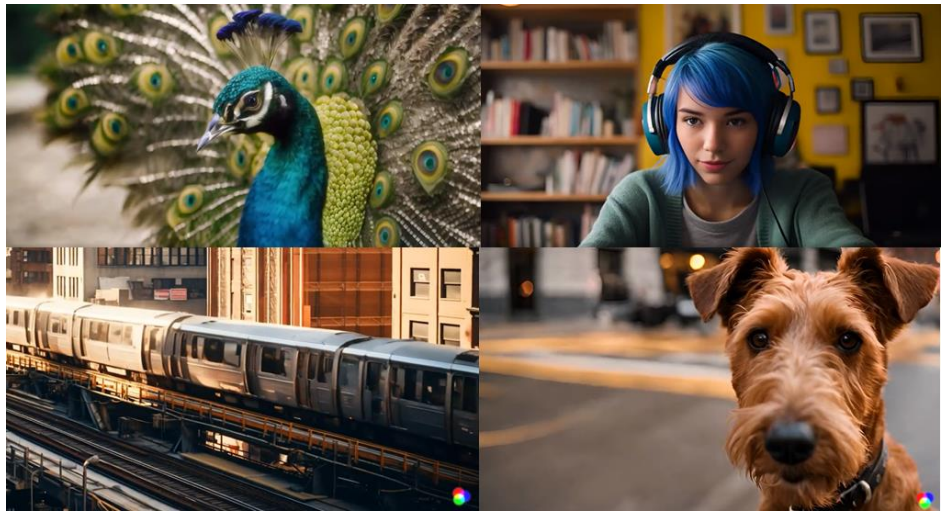
■ 开源派玩家Stablility ai也推出了开源视频生成模型Stable Video Diffusion，有望推动视频生成领域的开发者生态繁荣。

■ 科技大厂亦加速布局，如META推出Emu Video、李飞飞团队联合谷歌推出的W.A.L.T等；OPENAI虽暂时未有直接布局，但我们认为只是时间问题，未来其若入局将进一步推动行业加速发展。

表：2023年主要视频生成模型/技术对比

模型	开发团队	推出时间	特点	是否开源	生成视频表现		
					长度	每秒帧数	分辨率
Gen-2	Runway	6月	影视级构图运镜，画面清晰度精美度最强，最新版本可生成4K画质视频	否	4~16秒	24	768 x 448（免费） 1536 x 896（付费）
Pika 1.0	PIKA Labs	11月	语义理解能力强，画面一致性较佳	否	3秒	8~24	-（低）
Stable Video Diffusion	Stablility ai	11月	第一个基于图像模型 Stable Diffusion 的生成式视频基础模型	是	2~4秒	3~30	576 × 1024
Emu Video	Meta	11月	在生成质量和文本忠实度上表现较好	否	4秒	16	512 × 512
W.A.L.T	李飞飞及其学生团队、谷歌	12月	Transformer 与扩散算法结合，同时改善计算成本和数据集问题	否	3秒	8	512x896

图：runway最新gen-2可生成高清高质量的视频画面

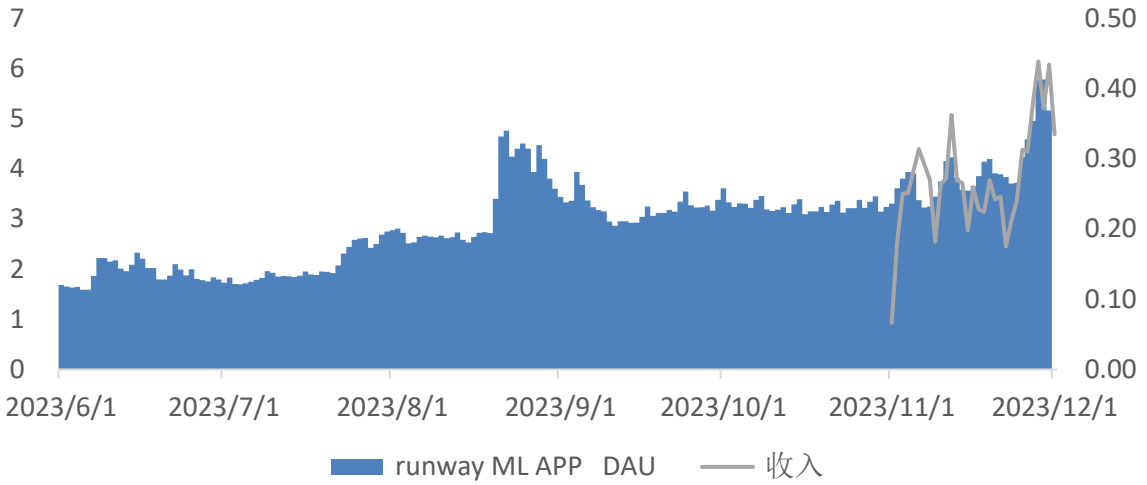


资料来源：RUNWAY官网，AI创新视界，AI变革指南，智东西，汽后观察员，东吴证券研究所

2.1.3 视频生成：Runway技术领先，具备商业化潜力

- **多年深耕AI视觉领域，实现技术领先。**Runway成立于2018 年，合伙人是三位来自纽约大学 Tisch艺术学院ITP项目的研究生同学。Runway从创立之初的ML模型平台到转型成AI Tools工厂，过程中基于视频抠图类的AI Tool打造一套云原生的视频编辑工具，并在寻找图像生成方法时发现Diffusion模型并构建Stable Diffusion早期版本，后又于23年先后推出视频生成模型Gen-1和Gen-2。 Runway既保持对行业前沿技术的敏锐度，又能坚定自研，从而保持在视觉生成模型领域的相对领先。
- **随着模型更加成熟，商业化潜力有望进一步释放。**Runway目前提供免费试用版和付费版（15~95美金/月），对于专业视频制作方而言，传统视频后期团队一个资深编辑后期制作费用250 美元/小时，一个只负责抠图的编辑费用60-80 美元/小时， 而Runway可为其节省几个小时甚至几天的工作；而随着技术更加成熟及算力成本下降， Runway有望吸引更多非专业用户。自11月对Gen-2进行重大更新以来， Runway APP的日活和收入数据均有明显增长，我们认为其商业化潜力未来有望进一步释放。

图： 11月更新GEN2以来，runway APP日活明显增长（DAU单位为万人；收入单位为万美金，右轴）



表：Runway 的订阅模式（1个积分=0.01美元）

	免费试用版	标准版	高级版	无限版
月度会员价格	-	$15	$35	$95
年度会员价格	-	$144 ($12 / 月)	$336 ($28 / 月)	$912 ($76 / 月)
积分/月	125（不可额外购买）	625（可额外购买）	2250（可额外购买）	无限
分辨率	720p	4K和绿屏alpha遮罩	4K和绿屏alpha遮罩	4K和绿屏alpha遮罩
去水印	×	✓	✓	✓
视频时长	Gen-1长达4s Gen-2长达16s	Gen-1长达15s Gen-2长达16s	Gen-1长达15s Gen-2长达16s	Gen-1长达15s Gen-2长达16s
视频项目	3	无限	无限	无限
存储空间	5GB	100GB	500GB	500GB

资料来源：Runway，海外独角兽，SENSOR TOWER，东吴证券研究所

2.1.4 3D资产生成：在视觉生成模型中相对早期

- 3D资产生成具有广阔的应用空间，如智能3D打印生成、虚拟现实设备、元宇宙生成等。目前3D生成技术包括NeRF（神经辐射场） 、GAN、DIFFUSION等，主要难点包括：**1）缺少大量高质量的3D数据集。2）算力要求更高。**每次优化所需迭代次数更多，耗时耗力，如22年11月英伟达推出的Magic3D模型生成单个3D网格模型可能需长达40分钟，而midjourney一般10秒钟左右即可生成4张图片。**3）技术难点更多。**3D模型远比2D图像复杂，且必须具备从不同角度看物体形状的一致性，更容易出现常识性问题，如AI生成的3D对象有多个头或者多个面。
- 受益于扩散模型等生成算法发展及文生图成功应用的出现，**2022年以来的Magic3D、Point-E等3D生成模型技术上有所突破，但在生成效率及精度上仍未找到平衡点，在视觉生成模型中属于相对早期的领域。**如Magic3D分辨率比DreamFusion提升8倍，但完成一次渲染仍需40分钟；OPENAI的Point-E通过使用点云模型极大提升生成效率，只需单个GPU用1~2分钟即可完成，但精度相对有所降低。

表：主要文生3D模型对比

模型	公司	推出时间	底层技术	精度	效率
DreamFields	谷歌	2021年底	NeRF 3D场景技术+DALLE+CLIP	-	-
DreamFusion	谷歌	2022/10	Imagen+Mip-NeRF，预训练2D文本到图像扩散模型实现文本到三维合成	一般	较差
Magic3D	英伟达	2022/11	分两个阶段，由粗到细渲染生成3D模型	较好	一般
Point-E	OPENAI	2022/12	文生图+图像到3D的扩散模型，实现文本到3D点云快速生成	较差	较好
Shap-E	OPENAI	2023/5	编码器+条件扩散模型，直接生成3D资产	较差	好

图:文字生成3D效果图，从左至右分别为Magic3D、DreamFusion、Point-E

2.2 听觉模型：数据仍有缺口，23年以来技术有所突破

- 听觉模型可分为音频识别与音频生成两大方向，其中识别技术已相对先进，生成可分为语音生成、音色转换、音乐生成和音效生成等。
- **现有训练数据集范围较窄，规模有限。**由于声音信号有自由度高、动态化特点，生成连贯、高质量音频需依靠大量文本-音频数据进行训练。
  - 语音：主要来自开源数据集、企业自有数据等，但方言、小语种等低资源语音数据，用于语音翻译全流程对齐的标注数据仍然较少。
  - 音乐：考虑到版权问题，大多与音乐版权公司合作获取，如Stable Audio与AudioSparx合作，MusicGen与ShutterStock及其子公司Pond5合作。但也因此高质量数据较少，如23年推出的Stable Audio训练数据包括超过80万条音频文字对，谷歌的MusicLM为5500个音乐-文本对，相比视觉数据集明显较少。
- **受大模型及扩散模型等工作的启发，音频生成算法2023年以来取得进一步突破，其中音乐生成技术相对更加复杂，开源模型的出现有望推动行业前进。**

表：音频识别发展历史

时期	技术方案	效果
1950~1970s	基于模板匹配	1952年“奥黛丽”系统能识别数字语音。1970年代，IBM的Shoebox系统能识别约16个英文单词。
1980s	统计模型	隐马尔可夫模型（HMM）被引入到语音识别中，大大提高识别准确性。
2000s - 现今	深度神经网络	系统能够自动从大量数据中学习特征，极大地改善识别效果。

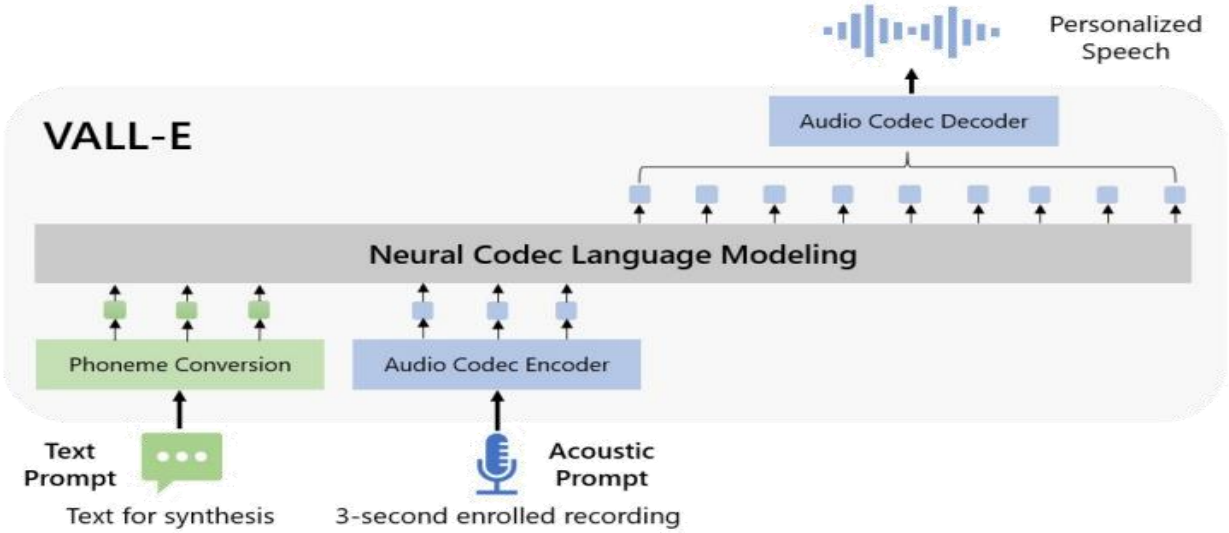
资料来源：techlead，易观分析，Midifan，MusicGen论文，华尔街见闻，东吴证券研究所

2.2.1 语音合成：23年在泛化性、生成质量上取得突破

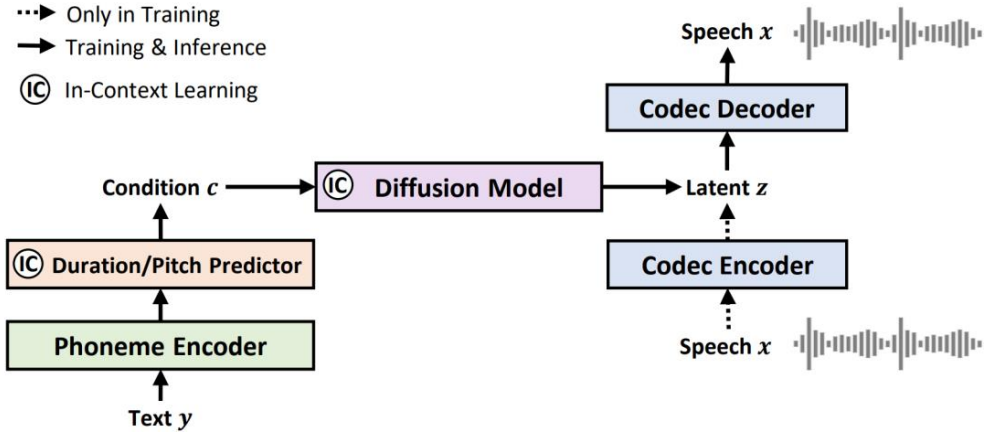
■ **23年以来，语音生成TTS领域算法亦开始受益于GPT和扩散模型等技术，在泛化性、生成质量上取得突破。**

- TTS主流方法包括基于统计参数的语音合成、波形拼接语音合成、混合方法及端到端神经网络语音合成，其中基于参数的语音合成包含隐马尔可夫模型（HMM）以及深度学习网络（DNN）。
- 过去TTS系统存在泛化能力差、生成音频机器感过重等问题，进入23年以来，行业开始出现类似自然语言领域的GPT和视觉领域的扩散模型等技术的尝试，并取得了较好的表现。1） 1月微软发布的VALL-E是第一个基于语言模型的TTS框架，利用海量的、多样化的、多speaker的语音数据（训练数据数百倍于以往TTS系统），实现zero-shot最优表现。只需提供三秒的音频样本即可模拟输入人声，并根据输入文本合成出对应的音频，而且还可以保持说话者的情感基调；2）5月微软发布NaturalSpeech2，利用扩散模型实现了 zero-shot 的语音合成，并改善了VALL-E丢音多音等问题。

图：VALL-E模型采取了基于语言模型的TTS框架



图：NaturalSpeech 2 系统



资料来源：微软亚洲研究院，仙林智能，VALL-E论文，东吴证券研究所

2.2.2 音乐生成：难度更大，期待开源模型推动行业前进

- **音乐生成模型比语音生成更复杂：**需对长序列进行建模，捕捉音乐全频谱；需更高采样率（音乐44/48 kHz VS 语音16 kHz）；包含来自不同乐器和声和旋律，结构复杂；需避免旋律错误；对创作者而言需要音调、乐器、旋律、音乐风格可控。
- **行业最新模型亦受到Transformer和扩散算法技术影响，期待MusicGen、Stable Audio等开源模型带动行业技术前进。**23年微软建立在Transformer模型上的MusicGen实现较好的生成质量； Stability.ai的Stable Audio实现音质进一步突破、更高的音频采样率、在90秒长度上保持连贯性（VS其他人工智能模型在几秒后演变成随机、不和谐噪音），推理时间也有所减少（可在Nvidia A100显卡上以不到一秒的运算时间渲染生成音频），两个模型都是开源模型，有望推动行业开发者生态繁荣，推动技术进一步突破。
- **商业化：可为企业/内容制造商/娱乐应用提供高性价比的音乐作品，或基于娱乐属性向C端收费。**如利用AI技术生成功能音乐的初创公司Endel，截至2022年8月，已拥有超过8万每月听众，其音乐已经进入了Spotify的一些氛围音乐播放列表；澳大利亚AI音乐生成产品Splash中，用户可通过点击代表节拍、循环和音效的网格来创作歌曲，其登陆游戏平台Roblox的半年后玩家数量便达到2100万人。

表：主要音乐生成模型对比

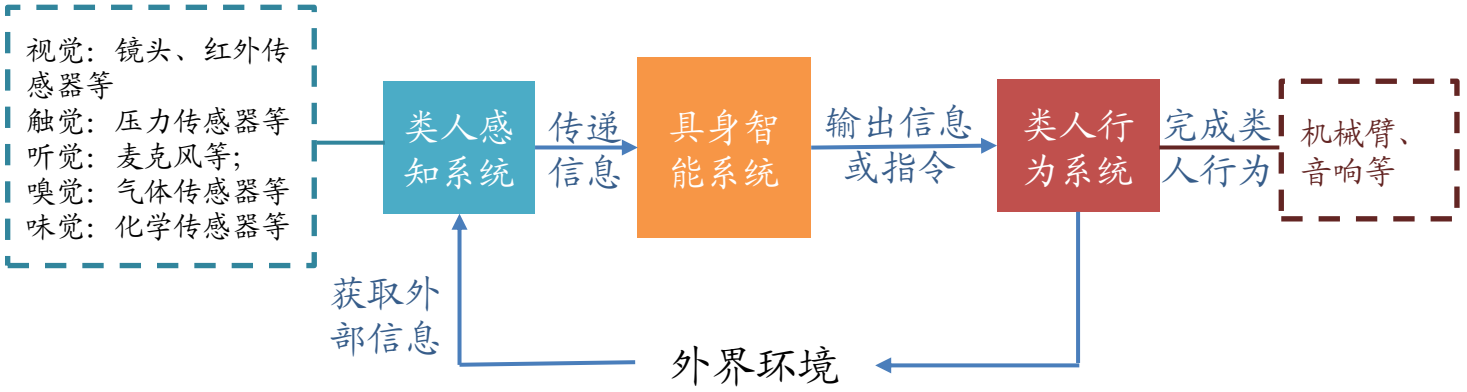
模型	开发公司	推出时间	底层技术	训练数据	音频采样率	功能	是否开源	评价
MusicLM	谷歌	23年1月	分层序列到序列建模任务	5.5k	24kHz	在已有旋律基础上进行创作、通过一组图片和字幕作曲、生成一段以某种乐器按照某种风格演奏的音乐	否	在音频质量和与文本描述的一致性方面优于以前系统
MusicGen	META	23年6月	单级转换器LM+高效标记交错模式	10K内部数据+390K 授权纯乐器音乐曲目	32kHz	根据现有音频进行可控生成和续写功能，可根据已有音乐元素再加工	是	没有严格遵循prompt要求，但准确反映所要求的音乐流派，且生成的音乐展示出对主旋律的不同诠释
Stable Audio	Stability.ai	23年9月	隐藏空间扩散结构	80K授权音乐	44.1kHz	根据prompt生成音乐	是	在90秒长度上保持连贯性、推理速度快

资料来源：MusicLM论文，MusicGen论文，Midifan，华尔街见闻,东吴证券研究所

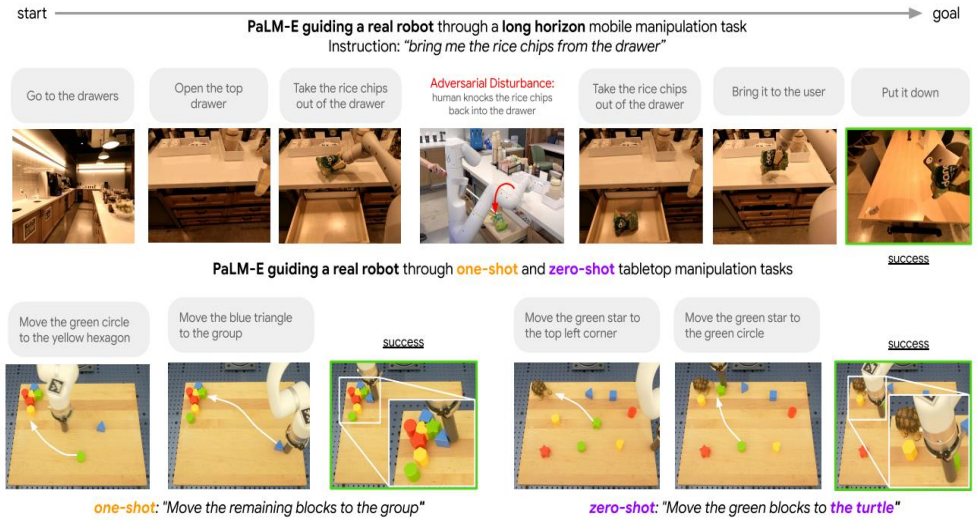
2.3 具身智能：相对远期，AI+机器人实现与现实世界交互

- 具身智能多指有类人身体并支持物理交互的智能体，如家用服务机器人、无人车等。具身智能是最为复杂的多模态能力，不仅要具备理解和推理能力，还要能够接受视觉、触觉等多模态的信息，同时对物理机械技术和工程等也提出要求。目前尽管已有多种机器人与软硬件设备出现，但还只是较为简单的机械控制系统与AI技术的结合。
- **随着多模态模型的视觉感知与推理能力不断增强，可集成至机器人硬件系统，具备与现实世界交互的功能。**GPT4V目前已支持泛化的空间感知与推理能力，如模拟家用机器人并完成居家任务等；PaLM-E能够支持机器人控制规划任务。未来随着多种模态的进一步整合，机器人设备能够实现集视觉、触觉、听觉为一体的完整具身智能。

图：具身智能机器人结构图



图：PaLM-E执行具身智能任务



资料来源：东吴证券研究所整理

3 海外技术领先，国内技术与应用同步发展

3.1 海外：OPENAI和谷歌领先，垂类独角兽加速行业发展

- **OPENAI和谷歌在多模态领域布局广度和技术先进程度上都处于领先地位，且都推出了表现较好的通用多模态大模型。**在垂类领域上，OPENAI在文生图等技术已接近拐点的方向表现较好，而谷歌在3D资产生成等技术还处于较早期的方向表现较好。
- Stability.ai、 midjourney、runway等公司在部分生成领域保持领先，这些独角兽对行业技术突破和产品创新发挥了重要作用，加速孵化爆款应用。

		通用多模态大模型	2D图像生成	视频生成	音频识别/生成	3D资产生成	具身智能
初创公司	openai	GPT4V	DALLE 3.0		whisper	Shap-E	GPT4V
	其他公司		Stable diffusion (Stability.ai) midjourney	Stable Video Diffusion(Stability.ai) Gen-2(runway) Pika1.0(pika)	Stable Audio (Stability.ai)		
科技大厂	谷歌	Gemini	Imagen2、StyleDrop	W.A.L.T、Imagen Video	MusicLM	DreamFusion	Palm-E
	微软	KOSMOS-2	Composable Diffusion（支持文本/图像/音频/视频生成）			RODIN	
	META	AnyMAL	CM3leon	Emu Video	MusicGen	Make-A-Video3D	

资料来源：东吴证券研究所整理。注：均选用玩家最新模型，红色代表该产品在行业内处于领先地位

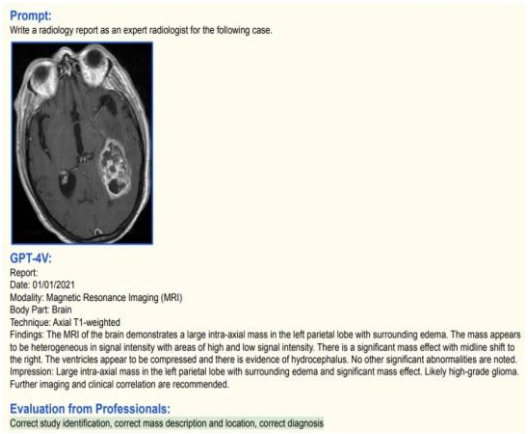
3.1.1 OpenAI：多模态能力不断增强，技术与应用正循环

■ **OpenAI是多模态大模型领头军，在LLM能力基础之上持续增强多模态能力，并注重与chatgpt生态的融合，技术与应用双向驱动。**

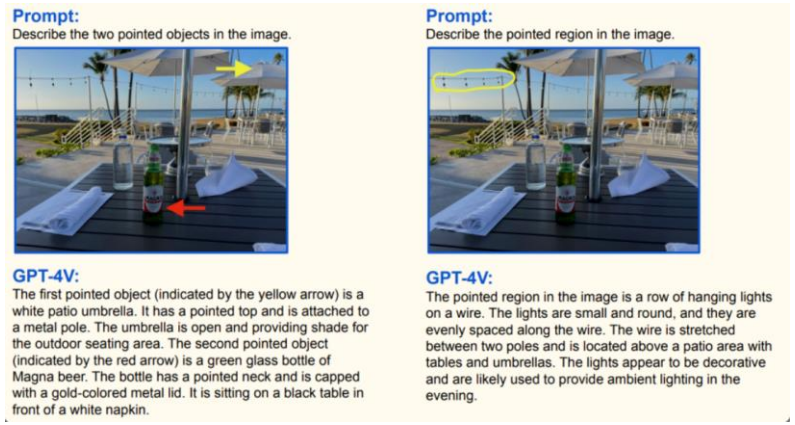
- **2023年10月，GPT-4 新增了视觉功能：1）实现准确且低门槛的识别、判断与推理，与外部工具与插件无缝集成，有望实现更多创新和协作应用。**如根据医学图像生成诊断报告，引用先前医学扫描和诊断历史提升诊断效率。2）**推出视觉参考提示功能，强化C端个人助理职能。**如用户可在图像中用箭头或圈进行标注，指示GPT4V进行聚焦性推理回答。**3）具备情感理解与美学判断能力，展现情感意识人机交互的潜力。**
- **文生图模型DALL·E不断迭代：**23年9月迭代至第三代，简化用户提示词学习过程，在图像表现力方面有明显提升，与MIDJOURNEY差距明显缩小。
- **多模态功能集成于ChatGPT体系中，有望增强技术与应用的正循环：**ChatGPT已基于Whisper、GPT4V、DALL-E·3推出语音和图像多模态功能，支持用户直接与ChatGPT进行语音对话、图像问答和图像生成，在提升用户体验的同时也有望积累更多数据及反馈帮助模型能力提升。

图：OpenAI多模态发展时间线

图：ChatGPT生成诊断报告



图：GPT4V根据视觉参考提示完成作答



图：DALL·E3降低prompt使用门槛，提升作图精细度

3.1.2 Google：2023年底推出多个多模态模型，推动行业技术加速

基础模型

PaLM-E(2023.3)：5620亿参数，可将现实环境连续感知信号嵌入到超大LLM中，从而能够建立语言单词和感知之间的直接联系。可用于连续的机器人控制规划，视觉问答，图像字幕生成等机器人和多模态任务。

GEMINI(2023.12)：原生多模态大模型，可归纳并流畅理解、操作及组合不同类型信息，包括文本、代码、音频、图像和视频，官方测试结果超过GPT4。

- 3种尺寸（Gemini Ultra规模最大且功能最强大，适用于高度复杂任务）、Gemini Pro适用于各种任务的最佳模型、Gemini Nano端侧设备上最高效的模型）
- 已在包括Bard、Pixel等多种产品和平台上推出，未来将应用于更多产品服务，如Search（已开始试验，用户在美国英语搜索延迟降低40%，质量有所提高）、Ads、Chrome 和 Duet AI。

图像生成

Imagen2（2023.12）：使用训练数据自然分布生成更逼真图像

Muse(2023.1)：基于transformer模型

StyleDrop(2023.6)：可捕捉用户提供的风格细微差别和细节，如颜色方案、阴影、设计模式。通过微调极少量可训练参数，并通过人工或自动反馈迭代训练提高质量，即能有效地学习一种新风格。

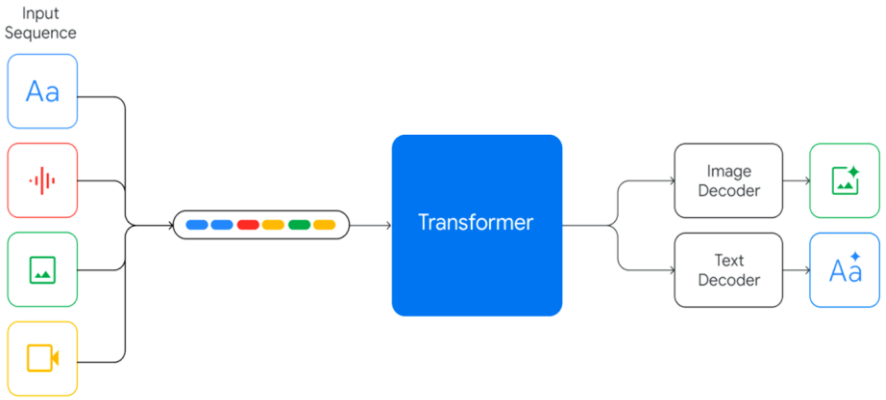
其他

音频生成：MusicLM(2023.1)

视频生成：W.A.L.T（2023.12）、Imagen Video

3D资产生成：DreamFusion

图：Gemini支持文本、图像、音频和视频交错输入，支持图像文本输出



图：谷歌最新文生图模型Imagen2（左）生成效果对比DALLE 3（中）和Midjourney（右）

Prompt：一小幅油画，描绘了摆放在砧板上的橙子。阳光穿过橙子的切片，柔和的橙色光线洒在砧板上。画的背景是一块蓝白相间的布，画面巧妙地捕捉了光的折射、反射效果，同时展示了画家富有感情的笔触

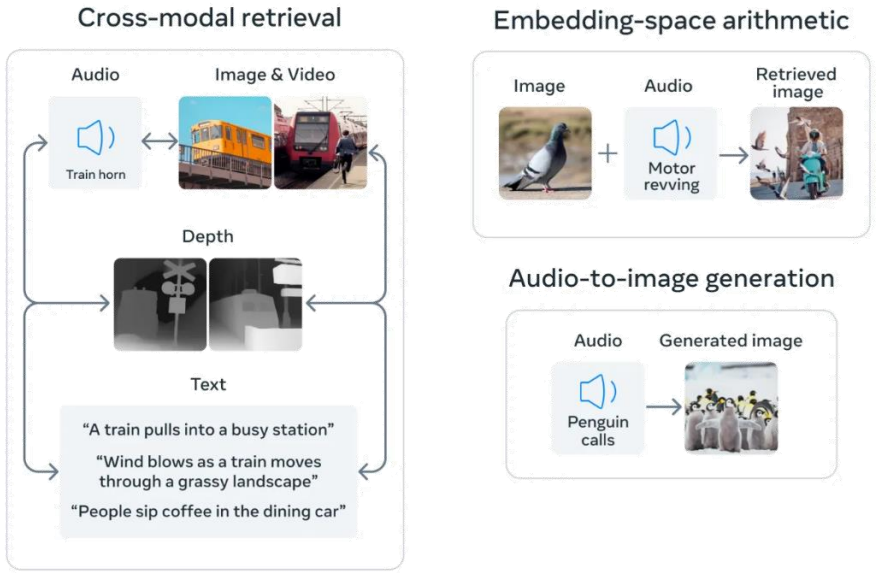
3.1.3 Meta：擅长图像模型，建设开源生态

- **META在图像模型领域的技术积淀相对较多**（数据优势+原有算法优势），推出的视觉大模型DINOv2、图像分割SAM等均有较好表现。
- **META在大模型领域选择开源路线**（如Opt是业内首次开源的大规模预训练模型），**通过生态建设追赶头部玩家。**

表：Meta主要多模态模型

模型名称	类型	发布时间	最大参数量	亮点/优势
OPT	多模态模型	2022/5	1750亿	连接文本、图片以及音频，支持跨模态理解和生成任务
DINOv2	视觉大模型	2023/4	10亿级	自监督训练，不需要微调
SAM	图像分割模型	2023/4	6.32亿	支持通过多语言prompt完成分割，无监督学习
ImageBind	多模态模型	2023/5	-	可跨六种模态进行转换与整合（图像、视频、音频、深度、热量和空间运动）
AnyMAL	多模态模型	2023/9	700亿	为构建多模态LLM 提出一种高效、可扩展解决方案，可将各种模态数据转换到 LLM 的文本嵌入空间。

图：一张鸽子图片+一个摩托音频， ImageBind能生成一张骑摩托车时有众多鸽子飞过的场景



资料来源：Meta，东吴证券研究所

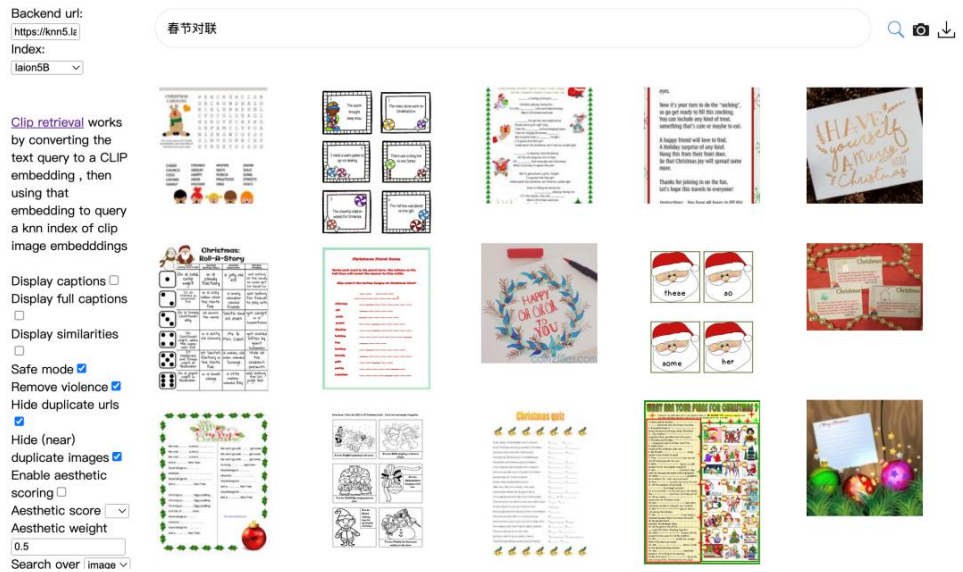
3.2 国内：海外开源有利于国内追赶，技术与应用同步发展

■ **相比英文多模态数据集，中文多模态数据集仍有提升空间。**以图文对数据集为例：

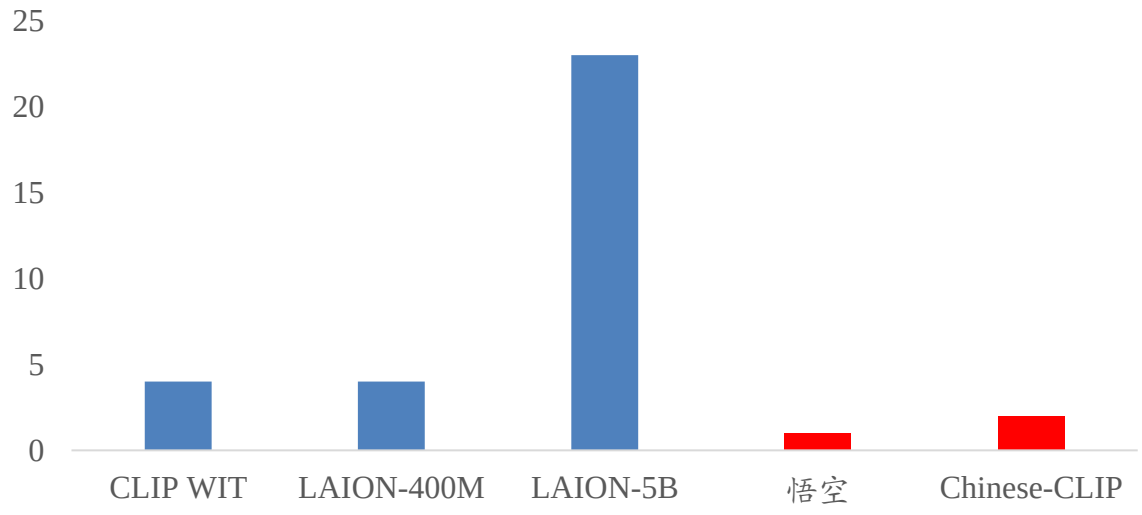
- 英文数据集难以很好满足中文需求，如在蒸馏多语言版本Multilingual-CLIP (mCLIP)中搜索“春节对联”，返回的是圣诞相关的内容。
- 2022年华为诺亚开源国内首个亿级中文多模态数据集悟空，随后阿里达摩院发布2亿规模的Chinese-CLIP，训练图文绝大部分来自公开数据集，大大降低了复现难度，推动中文图像生成模型的发展。但相比英文数据集（CLIP和LAION-400M 4亿图文对、LAION-58B 23亿图文对），中文数据集仍有提升空间。

■ **国内算法相对落后，算力方面亦有劣势，但海外算法开源（如META等）有利于国内技术追赶；考虑到中国科技公司在产品运营和迭代方面实力更强，技术与应用有望同步发展。**

图：在mCLIP检索demo 搜索“春节对联”的返回结果



图：相比英文图文对数据集，国内主流中文数据集规模仍较小（亿对图文对）



资料来源：机器之心，诺亚实验室，AICV与前沿，东吴证券研究所

3.2 国内：海外开源有利于国内追赶，技术与应用同步发展

- **国内科技大厂（百度、阿里巴巴、字节跳动、腾讯等）及大模型公司（昆仑万维、科大讯飞、商汤等）均积极布局多模态，并有望结合自身应用生态优势进行商业变现。**如阿里巴巴应用在电商领域，腾讯应用在营销领域，昆仑万维应用在AI游戏、AI音乐等领域。
- **万兴科技、美图等AI视觉应用公司亦有望受益于底层技术进步，特别是其海外应用。**以美图为例，12月发布自研大模型Miracle Vision 4.0版本，拥有AI设计与AI视频两大新能力，并将于2024年1月陆续应用于美图旗下产品。目前Miracle Vision的视频生成能力已能融入行业工作流，尤其是电商和广告，MV4.0的迭代加速将推动公司向生产力场景应用渗透，助力行业工作流提效。

图：国内科技大厂多模态模型布局

百度	多模态大模型：文心一言（23/3） 文生图：文心一格；视频生成：VideoGen（23/9）
阿里巴巴	多模态大模型：通义千问（23/4），mPLUG-Owl2（23/11） 文生图：通义万相（自研composer架构）；视频生成：I2VGen-XL（23/9）、Animate Anyone（23/12）
字节跳动	多模态大模型：云雀大模型&BuboGPT（23/8）、UniDoc（23/9） 文生图：CLIP-GEN（22/8）；3D生成：MVDream（23/9） 视频生成：MagicAnimate & MagicAvatar（23/12）PixelDance（23/11）
腾讯	多模态大模型：混元大模型（23/9） 3D生成：3D场景自动生成方案（23/3）；视频生成：VideoCrafter（23/7）

表：文心一格收费模式

	基础版	白银会员	黄金会员	铂金会员
定价	免费	69元/月	139元/月	339元/月
电量礼包	-	600~900	1700~2300	4500~6000
同时生图	1组	3组	5组	10组
充电打折卡		9折3张	8.5折3张	8折3张
免费生图模式		赠送200张	无限生图	无限生图

资料来源：文心一格官网，公司官网，东吴证券研究所整理

4. 投资建议

- 随着多模态技术迭代，图像生成、视频生成、3D生成、音频生成等AIGC应用有望加速；后续MR+AI技术共振更有望共同驱动下一代生产力工具及文娱体验升级。

- 我们首推在多模态方向已有布局或具备布局能力的标的：昆仑万维（模型能力国内领先，正进行多模态研发及布局）、万兴科技、美图，建议关注新国都。

- 多模态技术进步利好电商、游戏、教育、营销等领域AI应用发展，推荐焦点科技（AI+电商）、中文在线(AI+游戏、短剧等）、盛天网络（AI+游戏）、蓝色光标（AI+营销）、凤凰传媒（AI+教育）、世纪天鸿（AI+教育）等，建议关注掌趣科技（积极与行业头部厂商合作，23年6月与悠米达成业务合作，共同开发AI游戏创作平台，降低开放世界游戏的开发门槛；23年7月与行者AI达成战略合作，扩大AI游戏创作平台的技术和创新工具储备；23年11月与蓝亚laya达成战略合作，共同打造AI游戏引擎）等。

- 建议关注受益于AI视频应用发展的多模态技术公司，如虹软科技、当虹科技等。

- 算力方向建议把握板块龙头投资机会，推荐中际旭创等龙头。

5. 风险提示

- **多模态技术发展不及预期：**多模态大模型对算法依赖度高，且有赖于技术进步与突破。若后续算法迭代不力，或进一步影响多模态大模型的发展。
- **伦理与隐私问题：**多模态大模型仍存在一定安全风险，包括生成有害内容以及隐私泄露风险，可能会面临监管压力。
- **商业化拓展不及预期：**多模态大模型仍处于开发与初步应用阶段，尚没有成熟的商业模式，存在付费意愿不及预期的风险。
- **算力基础设施发展不及预期：**多模态大模型的算力需求庞大，对于硬件设施的依赖度高，可能存在芯片供应不足等影响模型发展的风险。

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

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）），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

东吴证券 SOOCHOW SECURITIES

东吴证券 财富家园