2024 年数字经济专题：人工智能行业应用如火如荼_数字经济算力基建再接再砺

一、OpenAI 推出 Sora 文生视频模型，AI 全球应用发展更进一步

（一）Sora 文生视频模型推出超预期，有效驱动 AI 应用发展

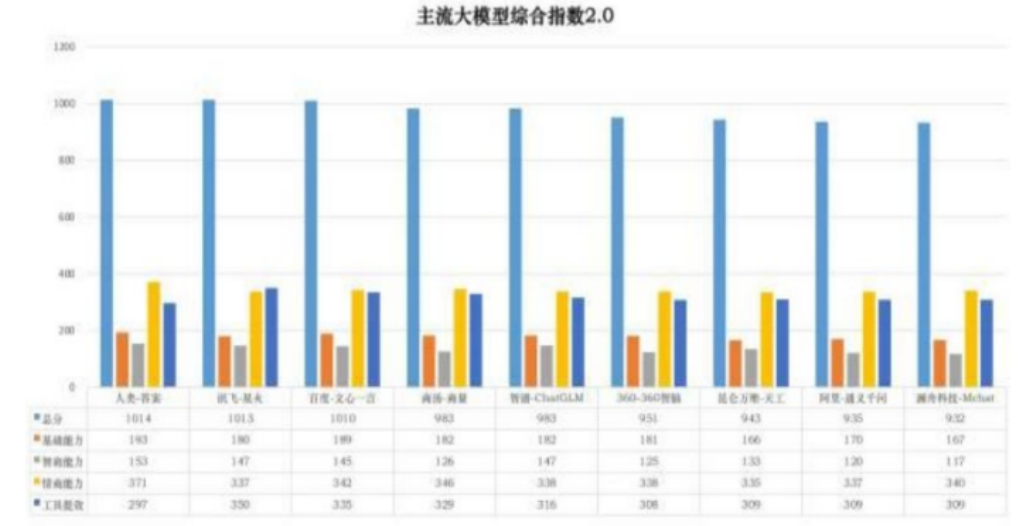
从全球看，OpenAI 推出文生视频模型 Sora，人工智能赋能短视频领域。2024 年 2 月美国 OpenAI 继 2022 年底 ChatGPT 发布后，推出全球首款文生视频模型 Sora，该款模型可以通过输入文字及提示词（最长 135 个）后，生成细节连贯的相关视频。Sora 的发布，使得 ChatGPT 从文字、图片层面正式向 成熟短视频层面进行演进，可生成最长 60 秒的全动态视频，通过深入理解物体在现实世界中的存在方 式，具备创建复杂场景和多人物角色的能力。它能够描绘道具、生成表现出丰富情感的角色，充分展 示了对物体存在的出色理解，确保了生成视频过程中人物、环境等一致性，一经推出备受关注。 Sora 具备“世界模拟器”的潜力，视频长度提升和效果超预期。Sora 发布前，友商 Pika、Runway 等生成模型大多处于生成 4 秒左右的“动图”范畴，60 秒连贯视频叠加 Sora

更强的语义理解能力、对 不同宽高比和分辨率的适应能力、优秀的视频扩展能力等优势，使得 Sora 发布后便同其它模型产生较 大代差，对 AI 制作视频领域带来新一轮突破。

算法原理方面，Sora 本质上基于“Transformer+Diffusion”。Sora 是一个在不同时长、分辨率和宽高 比的视频及图像上训练而成的扩散模型，同时采用了 Transformer 架构，也就是一种“扩散型 Transformer”。 Sora 主要的算法基础原理在于 Transformer+Diffusion，从文字生成视频主要经过三步，分别为语义理解、 生成图像以及图像排序生成视频，语义理解主要基于 ChatGPT，生成图像基于 Diffusion，图像排序生成 视频则基于 Diffusion 及 Transformer。 首先，Sora 需要巨量数据进行学习分析。由于 Sora 属于文生视频模型，故而需要互联网规模的海 量视频数据库进行分析学习，进而通过数据库进行联想，从而对输入的语义有加深的了解；其次，通 过文字生成图片。在文字输入后，Sora 会将文字先利用 ChatGPT 生成（Transform）图片，即 Transformer， 给出的文字越多，生成的图片细节愈发丰富；而 Diffusion 则会根据关键词特征值对应的可能性概率， 在使用视频库中数据进行多次拟合后，将碎片化信息粘合进行完整的图片输出；生成图片后，再多次重复该过程，生成完整视频。将完整的图片进行时间序列排序，利用时空补片技术（Spacetime latent

patches）生成具有语义代表性的视频成品。给定一个压缩的输入视频，模型会提取一系列时空补片， 充当 Transformer 的 token。正是这个基于补片的表示，让 Sora 能够对不同分辨率、持续时间和长宽比 的视频和图像进行训练，在推理时，模型则通过在适当大小的网格中排列随机初始化的补片来控制生 成视频的大小。

图4：国内互联网大厂不断加快大模型迭代节奏，重视 AI 商业化落地



资料来源：《人工智能大模型体验报告 2.0》，新华社研究院，中国银河证券研究院

（二）Sora 将 AI 潜力具象化，全球未来 AI 发展潜力无限

Sora 是对已有信息的整合，未来发展仍可持续演进。根据 Sora 算法原理，我们可以发现其核心是 基于互联网上已有的视频信息，根据文字输入要求进行碎片化拼接整理，从而具

备基于现有数据库的基础联想能力，虽然 Sora 目前突破了文生视频模型的时长限制及连贯性的问题，但尚未完全理解现实

世界中的物理法则和随机应变，未来 AI 发展潜力仍有较大提升空间。 Sora 是 ChatGPT 的延伸，商用前景大有可为。鉴于 Sora 的算法及底层核心逻辑机制，我们认为当 前 Sora 更多的意义在于将 AI 潜力具象化，当前处于该具象化进程早期阶段，其本质仍然是以 ChatGPT 为底座的文生视频模型，与其它文生视频模型相比，拥有时长更久、长期一致性、多样化视频格式输 出等特点，其内核仍以 ChatGPT 及自身视频训练量关联度较大。我们认为 Sora 作为在 ChatGPT 上衍生 的文生视频模型，未来主要发展方向也正如其所说，或将以“世界模拟器”为前景，逐步提升其创作 能力和推理能力。长期来看，Sora 将远远不只是内容生产工具，其构建的基于三维物理世界来创造数 字原生世界的强大引擎，将给一些产业从底层工具层面带来变化，形成深远影响。

二、我国数字基建“适度超前”，有效助力 AI+快速发展

（一）政策持续支持，大力转型数字经济发展

我国大力推进现代化产业体系建设，“人工智能+战略”明确提出。2024 年《政府工作报告》中提 出“制定支持数字经济高质量发展政策，积极推进数字产业化、产业数字化，促进数字技术和实体经 济深度融合。深化大数据、人工智能等研发应用，开展人工智能+行动，打造具有国际竞争力的数字产 业集群。” 2024

年或将成为我国数字经济发展的重要一年，在人工智能迅速发展的大背景下，我国政 府工作报告中提出开展“人工智能+”行动，有望在 2024 年实现从基础设施建设，到产业链逐步自主 可控，再到行业应用的稳步推进。在基础设施建设方面，2023 年 10 月工信部等六部门联合印发《算力 基础设施高质量发展行动计划》，提出到 2025 年，智算中心计算力、运载力、存储力应用赋能等方面 具体指标要求。“东数西算”工程八大算力枢纽及国家超算中心陆续建设中，目前全国智算中心已投运 25 个、在建超 20 个，建设总量与节奏或超预期。

适度超前建设数字基础设施，推动经济社会数字化转型。十四届全国人大二次会议发表政府工作 报告，该报告中强调，积极推进数字产业化、产业数字化，促进数字技术和实体经济深度融合；适度 超前建设数字基础设施等。作为建设网络强国的基石和重要内容，数字信息基础设施正在成为衡量国家核心竞争力的重要标志。我国在 5G、云计算、大数据等领域具有先发优势，新一代信息技术与各行各业深度融合应用，已成为经济社会发展的战略性公共基础设施。适度超前建设数字信息基础设施， 有助于为数字中国建设和数字经济发展提供高质量的产品和服务，高效满足千行百业、千家万户对美好数字生活的新需求，塑造发展新动能新优势，进一步推动经济社会数字化转型。

数字基础设施“适度超前”+AI 发展，政策支持助力算力全产业链快速发展。在当前全球人工智 能发展迅猛的背景下，我们认为数字基础设施超前建设，配合推进“人工智能+”行动或将成为未来 AI 赋能千行百业的重要一步。由于我国下游场景所需计算数量规模较大，叠加近年来政策的持续推动，我们认为算力产业链上游智算中心建设将保持较高景气度，同时国内外算力需求高增也将拉动光模块 产业链市场规模的进一步提升。下游应用场景受政策及需求拉动，从 0 到 1 的不断突破将带动市场空 间的逐步扩容。2021 年工信部发布《新型数据中心发展三年行动规划》，同时 2021 年国务院发布的《“十 四五”数字经济发展规划》，再到 2023 年国务院发布的《数字中国建设整体布局规划》，我国数字经济 基础设施发展方向及路径逐步明晰。虽然人工智能产业链上游方面，目前海外厂商占据较优势地位， 我国产业链相关公司追赶态势较猛，算力产业链上游国产化程度有望进一步加速，规模化、国产化有 望带来更大的发展空间。

（二）AI 发展如火如荼，市场规模复合增长率或超预期

人工智能不断取得进展，GPT 结论精准度持续提升。ChatGPT 的发展带动行业从专注于特定任务 的人工智能（AlphaGO 等）向更广泛的强人工智能发展，ChatGPT 的基础逻辑为利用大量数据进行模型

训练，实现深度学习，从而能够根据输入信息给出针对性的输出和预测，故而 ChatGPT 给出结论精确 与否主要取决于参数量的多寡。从 GPT 的发展阶段来看，随着代际的提升，其功能不断完善，参数量 及预训练量均呈现出指数级别的增长，从而显著提升了 ChatGPT 结论的准确性。

中国 AI 市场规模 2021-2026 五年复合增长率（CAGR）将超 20%。根据 IDC 的《2023 年 V1 全球 人工智能支出指南》（IDC Worldwide Artificial Intelligence Spending Guide）预测数据，中国人工智能 （AI）市场支出规模 2023 年增至 147.5 亿美元，约占全球总规模十分之一。受、地缘政治及宏观 经济等因素的影响，IDC 小幅下调了 2022 年中国 AI 市场规模预测值，相比 2021 年增长约为 17.9%。长远来看，AI 技术的创新迭代驱动了应用场景的进一步落地，以 AIGC、数字人、多模态、AI 大模型、 智能决策为代表的热点为市场带来了更多想象力和可能性。同时，企业对自身“数字化”、“数智化” 转型的积极推动催生出对 AI 技术的多元化需求，为中国 AI 市场规模的长期增长奠定了基础。IDC 预 计，2026 年中国 AI 市场将实现 264.4 亿美元市场规模，2021-2026 五年复合增长率（CAGR）将超 20%。 在五年预测期内，AI 领域的主要支出仍将来自于专业服务领域的行业用户，紧随其后的是政府和金融

行业，三者合计约占市场总量的一半以上。增长最快的行业分别为银行和地方政府，五年 CAGR 均超 23%。具体来看，AI 在专业服务领域，可以广泛应用于搜索和推荐、广告营销等，目前 ChatGPT 已经 嵌入必应搜索，带来了更多的智能化和人性化的特性，提升用户的搜索体验和搜索效率，为国内市场 提供了新思路。在政府行业，主要应用在公共安全、城市管理和社会服务方面，通过人脸识别和大数 据相关技术，识别出潜在的安全风险，此外还可以在办理业务时进行人员核对，提高办事效率。在金 融行业主要的应用包括风险管理、欺诈检测、投资分析等，随着数字人的不断进步，金融行业的服务 模式也将重塑。

（三）AIGC 驱动千行百业加速裂变，行业应用或快速渗透

工业 AI 大模型市场空间广阔。根据工信部数据，我国工业互联网 2023 年核心产业规模达 1.35 万 亿元，工业互联网融入 49 个国民经济大类，有一定影响力的工业互联网平台超过 340 个，覆盖全部工 业大类，工业设备连接数超过 9600 万台套，带动投资超 1700 亿元，工业经济整体呈现稳中向上、回升向好的态势。从软件业经济运行来看，软件产品收入 29030 亿元/+11.1%，其中工业软件产品实现收 入 2824 亿元/+12.3%

，工业软件收入占比逐步提升，工业 AI 大模型不断赋能，盈利能力保持稳定。

工业垂直大模型提质增效，赋能工业生产各个环节。从产业技术变革发展来看，数据成为新生产 要素、算力网络基建成为生产资源、人工智能算法成为新生产工具，共同构成新质生产力重要驱动因 素。数据成为新型工业化关键性生产要素，构成提升工业垂直大模型专用场景的关键。工业各环节围绕语言、专用、多模态和视觉四类大模型开展探索，4 类模型应用占比：75%、15%、8%和 2%，当前 以大语言模型为主。2024 年 3 月，政府工作报告指出全面部署推进新型工业化，提高全要素生产率， 推动重点产业链高质量发展，工业企业利润由降转升。总体来说，伴随政策发力加快推进新型工业化， AI 大模型算法升级赋能传统制造业数字化转型升级，行业盈利能力有望增强，带来正向增益。

三、AI 发展催化算力网络产业链迎来新机遇，相辅相成共赢

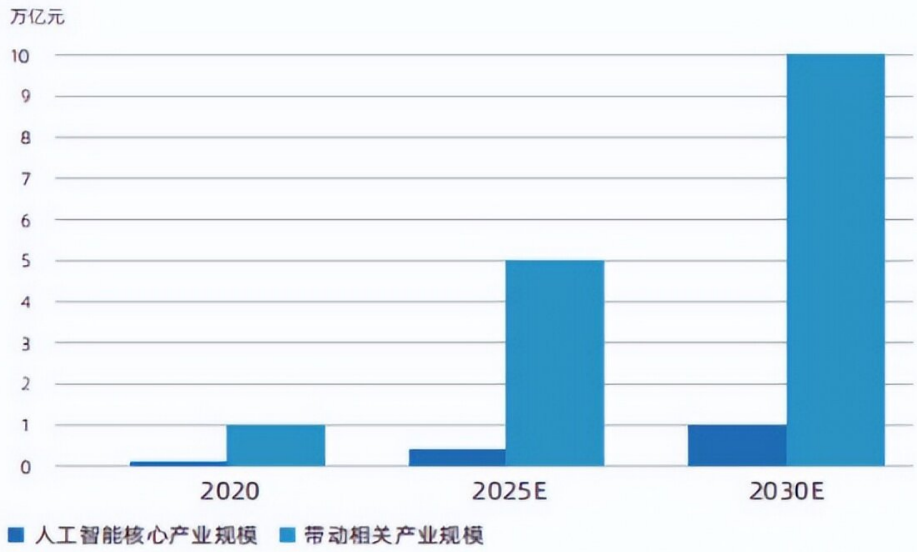
（一）AI 智算中心建设有望加快，产业链环节中光模块持续受益

人工智能产业链中，软件及硬件互相促进创新。我们认为大模型的问世，不仅是因为人类创造力 的提升，也是因为硬件的更新迭代推动了应用端的创新，工欲善其事必先利其器，随着英伟达 V/A/H 系列 GPU 的面世，以及数通侧光模块速率从 100-200-400-800G 的演变，叠加存储技术的不断突破，

数据传输速率，并行处理量及存储量实现较快增长，使得数据处理数量级实现快速提升，从而推动应 用革新。展望未来，我们认为在大模型开发如火如荼的背景下，在其上开发的应用或将遵循该规律， 在软件及硬件的不断进步下，逐步完成从量变到质变的飞跃。

人工智能发展催化全球算力规模保持高速增长。《国务院关于印发新一代人工智能发展规划的通 知》(国发(2017)35号)提出“要推进人工智能理论、技术与应用；到 2025 年，人工智能核心产业规模 超过 4,000 亿元，带动相关产业规模超过 5 万亿元；到 2030 年，人工智能核心产业规模超过 1 万亿元， 带动相关产业规模超过 10 万亿元”。信通院预测，“十四五”期间，在智算中心实现 80% 应用水平的 情况下，城市/地区对智算中心的投资可带动人工智能核心产业增长约 2.9-3.4 倍、带动相关产业增长 约 30 倍，我国算力规模高增可期。

图10：人工智能带动相关产业链规模不断提升



资料来源：《智能计算中心创新发展指南》，国家信息中心，中国银河证券研究院

智算中心的服务器/交换机及光模块占 BOM 成本比例有望增长，冷却方式有望从风冷转向液冷。 相较于此前传统数据中心，智算中心因人工智能的发展，整体折旧时间缩短（预计从传统数据中心普 遍 10 年期折旧降至 3-5 年），同时 AI 服务器/交换机技术因不断迭代，成本比例或将逐步提升。由于 数据中心本身即为高能耗行业，所需电力较多，故而 PUE 控制一直以来便相对较严，智算中心能耗更 为严重，预计未来液冷代替风冷，降低数据中心 PUE 势在必行。 数据中心市场份额方面，2022 年全球数据中心市场中，我国占比约 25%-30%，2023 年我国算力 总规模占比全球居次席，年增速约为

30%，新增算力设施中智算中心占比过半，相较于美国数据中心 CAGR>10%的增速，我国算力总规模市场份额有望进一步增长。

随着数据中心的不断建设，我们认为光模块具备长期高成长性。智算中心拥有更短时间的折旧，以及更高的速率及更低的时延需求，故而对于服务器的要求相对较高，衍生出服务器的“1 VS 多”的 配套产品光模块需求大幅提升，我们认为光模块具备量价齐升的逻辑： 量升：800G 光模块陆续放量，海外互联网巨头不断加单。网络集群中的拓扑结构决定光模块用量， IDC 逐渐扁平、计算资源池化、SDN/NFV 兴起，叶脊架构成为主流。理论上当交换芯片速率达 51.2Tbps 时，400G 光模块成为主流、800G 光模块需求初步产生，当交换芯片速率达 102.4Tbps 时，800G 光模 块成为主流、1.6T 光模块需求初步产生。每代高速光模块新品进入客户供应商名单基本需要经历 0.5- 1 年的认证周期，产品推出时间靠前的供应商被采纳为主流方案的可能性更大，并且当前竞争格局较 好，2022 年国内典型光通信企业海外营收占比均在 75%以上。目前全球光模块需求 400G 主要集中于 Amazon（约 45%）和 Google（约 25%）、800G 主要集中于 Nvidia（约 50%）、Google（约 30%）和 Meta （约 20%）等，前期已优先得到客户验证的公司将优先受益。随着海外互联网公司对相关速率要求的

不断提升，800G 光模块有望快速放量中。 价升：光模块单位速率成本已接近 1 美元/Gbps，新产品 800G 价格有望超出 400G 两倍。过去光模 块的价格，一方面受制于光芯片的摩尔定律，光芯片在光模块中的价值量占比 30%-50%，随速率的增 加而增加，另一方面新产品随着时间的推移良率逐步抬升。光模块市场规模的边际变化主要体现在高 速光模块份额与价格的相对增长，以及低速光模块份额与价格的相对下降。对于特定接口速率的光模 块，其出货量曲线预计是一个长尾的正态的倒 U 型曲线，价格曲线是一个类反比例函数，两者相乘得 到的市场规模曲线是一个厚尾的右偏的倒 U 型曲线，市场规模会先于出货量触顶。

（二）人工智能对算力投入要求高，算力底座运营商数据中心有望不断扩张

随着人工智能的不断发展，其基础设施当前仍处于快速扩张中。由于 GPT 对算力需求以推理卡和 训练卡为主，其中训练卡主要支持模型的深度学习训练，推理卡则在训练完成的模型上做快速的响应 及执行。故而训练卡的计算能力相对较强，需求量更大，使得市场空间及单片价值量处于相对较高位置。我们认为随着人工智能的发展，以及大模型训练量的快速增长，推理卡和训练卡预计在未来一段

时间仍将保持高速增长，人工智能高成长推动硬件逐步升级换代，对算力投入提出更高要求。

当前硬件价值量较高，英伟达出货量全球居首。在当前全球算力紧缺的背景下，GPU 仍是投资的 主要方向。GPU 性能方面，英伟达较为领先，且领先幅度相对较大，其下 H/A 系列 GPU 较 AMD 的 MI300L 系列具备较大存储及计算速率优势，未来随着技术的升级换代，其它具备规模出货能力以及技 术积累的厂商有望弯道超车。

我国算力设施建设正迎头追赶。2023 年 10 月，工信部等六部门联合印发《算力基础设施高质量 发展行动计划》，提出到 2025 年：（1）计算力方面，算力规模超 300EFLOPS，其中智算占比达 35%， 智算中心 2023 年完成 30 个，2025 年完成 50 个，东西部算力平衡协调发展；（2）运载力方面，国家枢 纽节点数据中心集群间基本实现不高于理论时延 1.5 倍的直连网络传输，重点应用场所光传送网（OTN） 覆盖率达到 80%，骨干网、城域网全面支持 IPv6、SRv6 等创新技术使用占比达到 40%；（3）存储力方 面，存储总量超 1800EB，先进存储容量占比达到 30%以上，重点行业核心数据、重要数据灾备覆盖率 达到 100%；（4）应用赋能方面，打造一批算力新业务、新模式、新业态，工业、金融等领域算力渗透