2024 年 03 月 21 日

从世界模型看算力需求变化

行业深度研究

证券研究报告

国金证券研究所
分析师：刘道明(执业 S1130520020004)　　联系人：黄晓军(执业 S1130122050092)
liudaoming@gjzq.com.cn　　huangxiaojun@gjzq.com.cn

AI 模型系列报告：从世界模型看算力需求变化

核心观点

Sora 是第一个表现出"涌现"能力的视频生成模型：随着模型规模增大而出现“理解世界”的能力。虽然许多 LLM，如 ChatGPT 和 GPT-4，表现出涌现能力，但在 Sora 出现之前，展示类似能力的视觉模型一直很少。根据 Sora 的技术报告，它是第一个表现出确认的涌现能力的视觉模型，标志着计算机视觉领域的一个重要里程碑。

Sora 的成功源于 Diffusion Transformer 架构的引入，和过去多年高质量数据的积累。

从架构上看，视频生成模型的技术路线开始收敛，Sora 的 Diffusion Transformer 架构证实了有效 scale-up 也即是增加算力能够对提升视频生成的精细度和效果，是视频生成领域的"GPT-3 时刻"。类似于 GPT-3 证明了更大的训练量、模型参数量、Token 数量，训练效果越好。引入了 Transformer 的 Sora 也表现出了同样的趋势，OpenAI 进行了 32x 训练量和 1x、4x 训练量的对比，从结果上看，32x 训练量的生成效果远远强于 1x 和 4x 的生成效果。在 Sora 发布后 Google、Snap 也发布了采用类似技术的视频生成模型，确定了 Diffusion Transformer 的视频生成路线，并且算力的需求会大大提升。

从数据上看，高质量的数据对视频生成模型的训练至关重要，Sora 利用自有工具增强训练数据集与提示工程。OpenAI 训练了专用的视频字幕模型来为视频生成详细描述，生成高质量的视频-字幕对，用于微调 Sora 以提高其指令跟随能力。同时为了确保用户提示与训练数据中这些描述性标题的格式保持一致，Sora 执行了一个额外的提示扩展步骤，即调用 GPT-4V 模型将用户输入扩展到详细的描述性提示。

我们认为，随着 Diffusion Transformer 类模型大量应用于图像及视频生成，推理需求将大幅增加，与 LLM 推理更需要内存带宽的资源需求不同，视觉模型推理将对芯片本身算力和内存容量提出更高要求。Sora 的 DiT 和大语言模型在推理时的逻辑不同，Diffusion 需要约 20 Steps 优化过程，每次均是计算的完整的 patch，访存需求也会大大下降，从 LLM 推理的访存密集型场景转变成算力密集型场景。

Sora 高质量的视频生成对影视和游戏行业的影响是最直接而深远的，降低制作门槛并且很有可能重塑影视和游戏制作的流程与格局。高质量的视频生成对于影视行业的工作流会有深远的影响，前期可以替代掉分镜以及概念片制作，后期可以取代部分特效制作。对于游戏行业，游戏开发人员可能会使用它来生成自定义的视觉效果，甚至是从玩家叙述中生成角色动作。

风险提示

模型架构的大幅改变影响算力需求分布
算力速度发展不及预期
中美科技领域政策恶化

内容目录

图表目录

一、Sora 模型的特点

1.1 Sora 在生成视频的质量、灵活性和时长上与之前的模型有代际差距

较长的视频生成时长：Sora 可以生成长达 60 秒的高保真度视频。对比之前的视频生成模型，Pika1.0 可以生成 3s-7s 的视频，Runway 的 Gen-2 可以生成 4s-18s 的视频。

灵活的分辨率：得益于其训练数据的灵活性，Sora 可以生成 1080P 的任何比例视频，而不是像之前的模型在生成非原生训练比例时会出现画幅的消失。

高保真渲染：在模拟数字世界时，如 Minecraft 游戏，Sora 能够实现高保真的渲染效果，使得生成的视频内容看起来就像真实游戏画面一样。

存在 Scaling Law：更高的算力、更大的模型规模、patch 数量的增加能对生成视频的效果有明显的正向提升。

三维空间连贯性：Sora 模型能够生成具有正确空间关系和动态相机运动的视频内容，确保视频中的物体在三维空间中保持连贯的运动。

图表1：Sora 在镜头和人物变化下的连贯性和一致性是 Scaling Law 下涌现出的能力

来源：Sora：Technical Report》、国金证券研究所

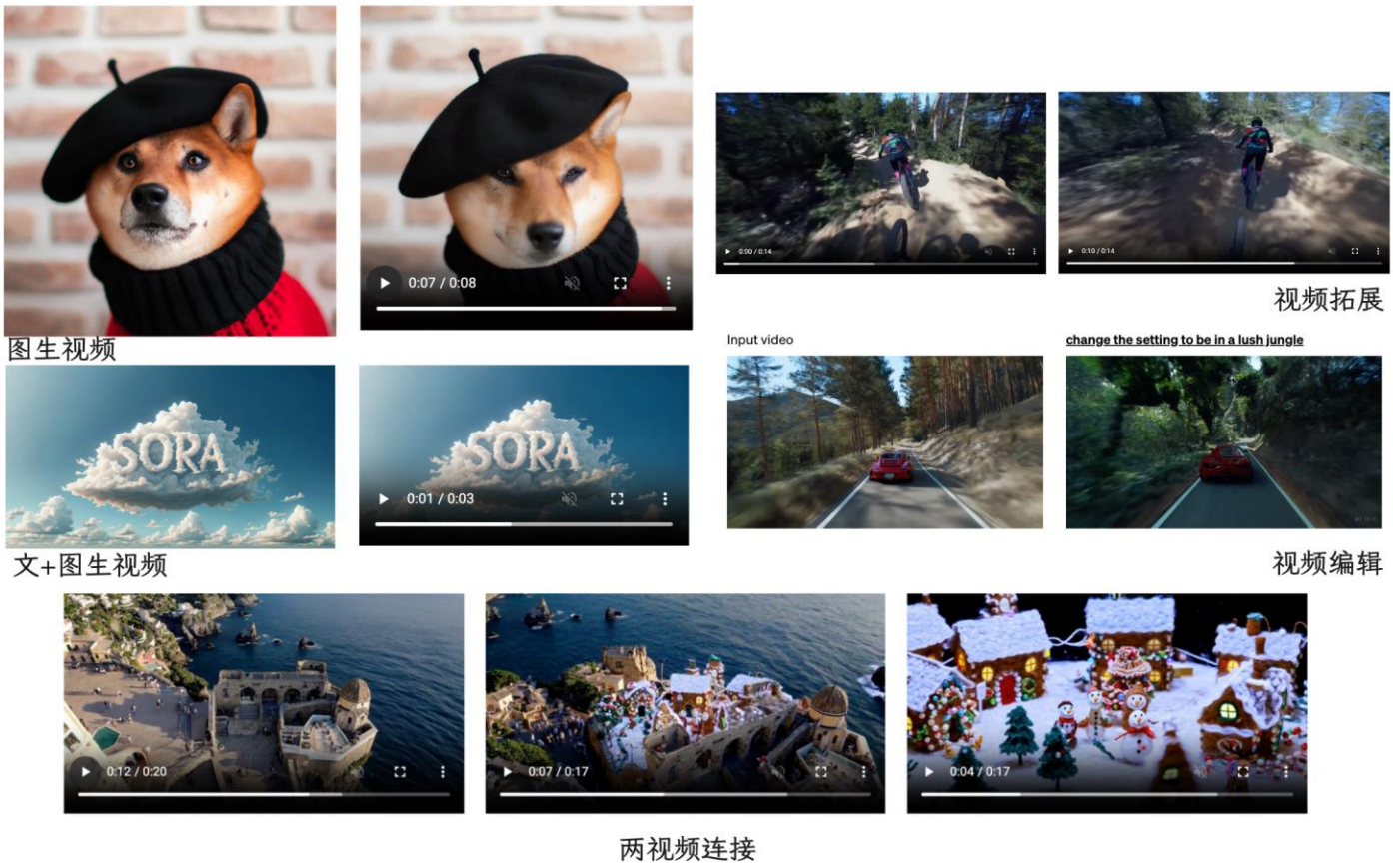
动态相机运动：模型能够模拟包含动态相机运动的视频，使得视频中的人物和场景元素能够随着相机的移动或旋转而相应地改变位置。

空间一致性：Sora 确保视频中的物体在空间上保持一致性，即使在复杂的场景变换中也能保持正确的相对位置和运动轨迹。

长期连续性和物体持久性：Sora 能够在视频中保持角色和物体的长期一致性，即使在视频中出现遮挡或离开画面的情况，也能保持其存在和外观。同时，它能够生成具有连贯故事线的视频，确保视频中的事件和动作在时间上是连续的。

任务场景丰富：除了视频生成以外，Sora 还可以用于其他任务，如图生视频、文生图片、文+图生视频、视频拓展、视频编辑、连接两个不同视频等。

图表2：Sora 适配任务场景非常丰富，覆盖了图像生成/编辑领域大多数任务



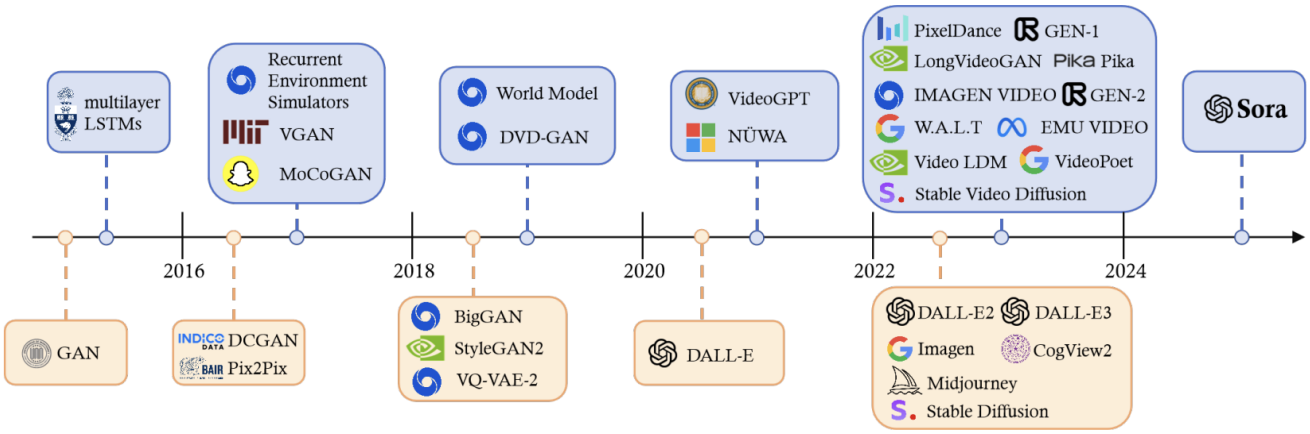
来源：《Sora: Technical Report》、数字未来实验室、国金证券研究所

二、视频生成模型的历史与现状

2.1 文生视频是个年轻的方向，最早能追溯到 15 年的基于 GAN 生成模型

文生视频是个年轻的方向，面临着多方面的独特挑战。主要有 1）计算成本高昂：确保帧间空间和时间一致性需要大量的计算资源，导致训练成本高昂；视频信息的复杂性进一步加剧了计算成本，需要更强大的计算能力来处理海量数据。2）视频信息复杂：视频数据形式多样，分辨率和比例各异，包含空间、时间、内容等多维信息；如何找到一种统一的表示形式，以有效地进行大规模训练，是文生视频技术需要解决的关键问题。3）缺乏高质量数据集：现有的文生视频多模态数据集数量有限，且标注程度不够，难以满足模型训练的需求。4）视频描述的模糊性：如何用文本准确描述视频内容，是文生视频技术面临的另一个难题，简短的文本提示难以完整描述视频，而复杂的描述又会增加模型的训练难度。

图表3：视频生成模型发展历史



来源：《Sora: A Review on Background》、国金证券研究所

2.2 GAN 和 VAE 时代

文生视频模型最早能追溯到 2015 年。早期研究主要使用基于 GAN（生成对抗网络）和 VAE（变分自编码器）的方法在给定文本描述的情况下自回归地生成视频帧（如 Text2Filter 及 TGANs-C）。虽然这些工作为文生视频这一新计算机视觉任务奠定了基础，但它们的应用范围有限，仅限于低分辨率、短距以及视频中目标的运动比较单一、孤立的情况。

图表4：最初的 GAN 文生视频模型在分辨率、上下文和长度方面极为有限

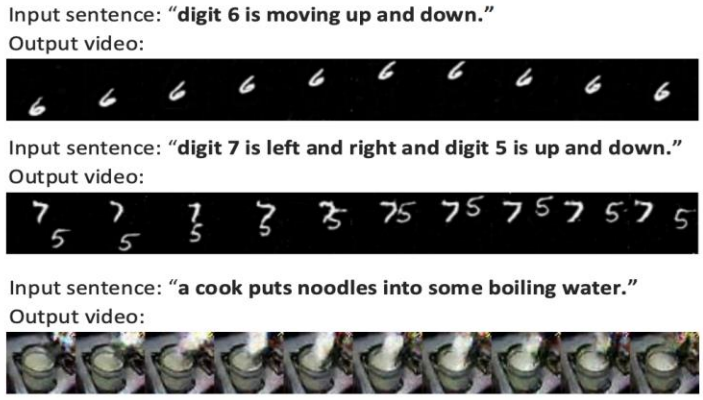


Figure 1: Examples of video generation from captions on Single-Digit Bouncing MNIST GIFs, Two-Digit Bouncing MNIST GIFs and Microsoft Research Video Description Corpus, respectively.

来源：《Text-to-Video: The Task, Challenges and the Current State》、国金证券研究所

2.3 Transformer Based

受文本（GPT-3）和图像（DALL-E）中大规模预训练 Transformer 模型的成功启发，文生视频研究的第二波浪潮采用了 Transformer 架构。Phenaki、Make-A-Vide、NUWA、VideoGPT 和 CogVideo 都提出了基于 Transformer 的框架，而 TATS 提出了一种混合方法，从而将用于生成图像的 VQGAN 和用于顺序地生成帧的时间敏感 Transformer 模块结合起来。在第二波浪潮的诸多框架中，Phenaki 尤其有意思，它能够根据一系列提示（即一个故事情节）生成任意长视频。同样，NUWA-Infinity 提出了一种双重自回归（autoregressive over autoregressive）生成机制，可以基于文本输入合成无限长度的图像和视频，从而使得生成高清的长视频成为可能。

2.4 Diffusion Based

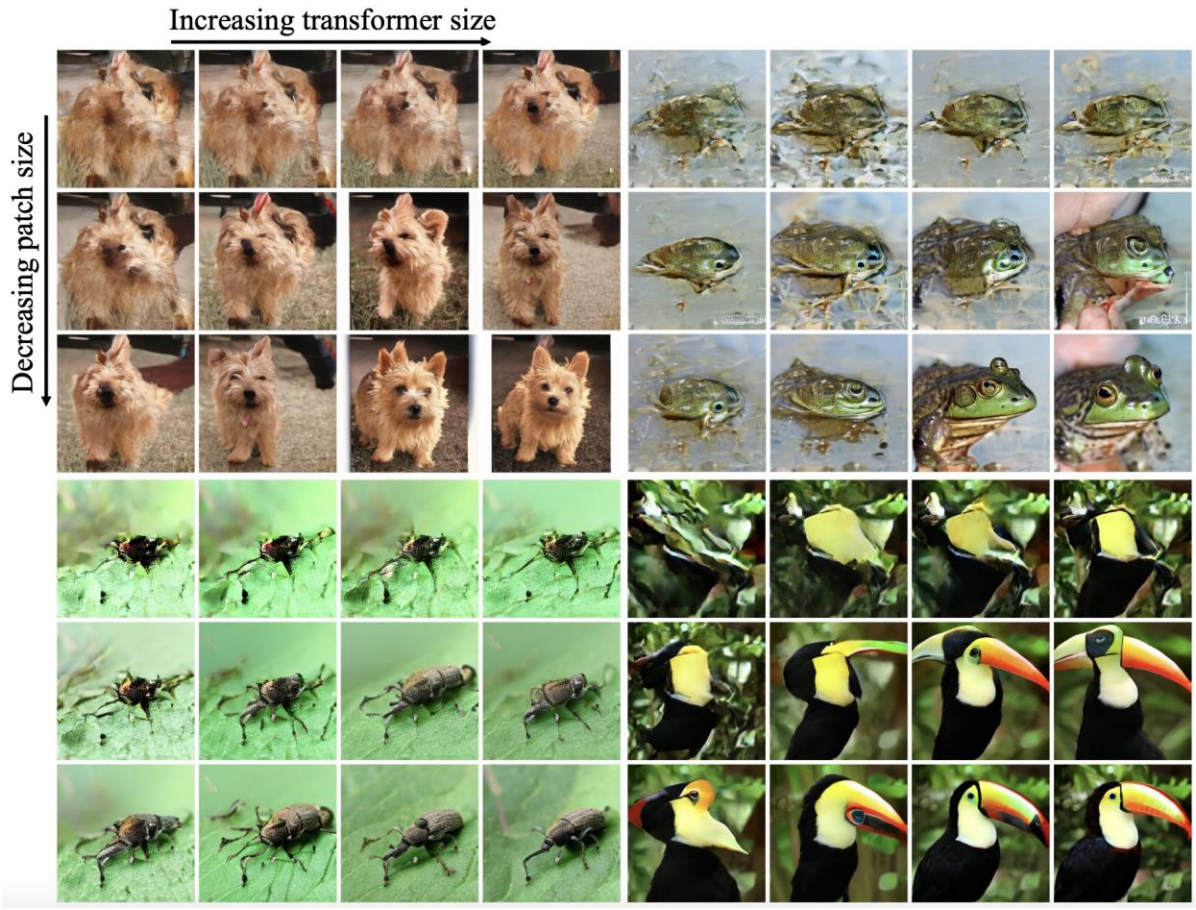
第三波文生视频模型浪潮主要以基于扩散的架构为特征。扩散模型在生成多样化、超现实和上下文丰富的图像方面取得了显著成功，这引起了人们对将扩散模型推广到其他领域

（如音频、3D，最近又拓展到了视频）的兴趣。这一波模型是由 Video Diffusion Models (VDM) 开创的，它首次将扩散模型推广至视频领域。然后是 MagicVideo 提出了一个在低维隐空间中生成视频剪辑的框架，据其报告，新框架与 VDM 相比在效率上有巨大的提升。另一个值得一提的是 Tune-a-Video，它使用 单文本 - 视频对微调预训练的文生图模型，并允许在保留运动的同时改变视频内容。随后涌现出了越来越多的文生视频扩散模型，包括 Video LDM、Text2Video-Zero、Runway Gen1、Runway Gen2、Stable Video Diffusion 以及 NUWA-XL。

2.5 视频生成模型的前沿：把卷积网络卷出了 Diffusion Model

这些模型缺点比较明显，比如支持视觉数据的类别少、视频时间短、视频尺寸固定等。当时还在 Meta 实习、现任 Sora 项目的负责人之一的 William Peebles 于 23 年 3 月发表的《Scalable Diffusion Models with Transformers》中的 Diffusion Transformers (DiTs) 对新的视频生成路线起到了关键的作用。DiT 的主要工作是替换了 Stable Diffusion 中的 UNet 为 Transformer，证明了在图像生成领域的 Scaling Law，也即是减少 patch size 增加参数量对生成图像有较大的积极影响。

图表5：DiT 证明了 Scaling Law 在图像领域的生效

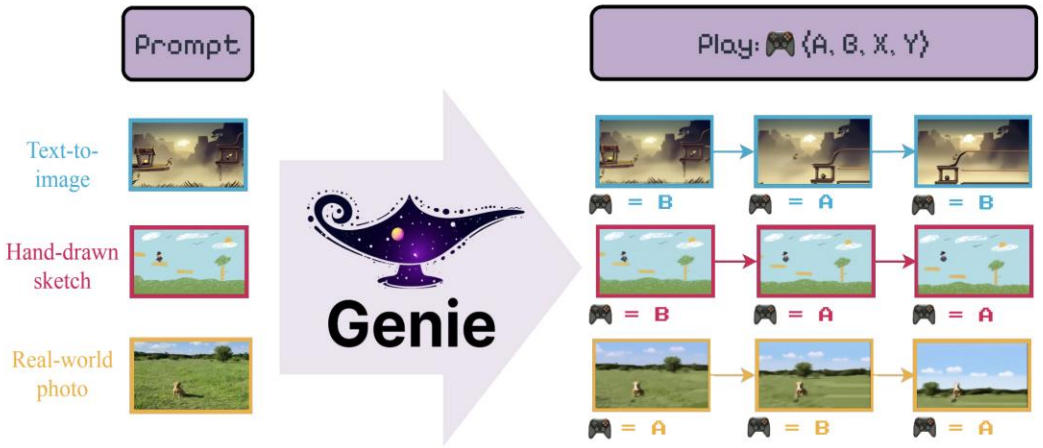


来源：《Scalable Diffusion Models with Transformers》、国金证券研究所

Sora 在 DiT 图像生成的基础上拓展到了视频生成，能够生成多样化的视频和图像，解决了先前方法在视频长度、尺寸和固定大小方面的限制，能够生成任意比例的 1 分钟 1080P 的高质量视频。Sora 没有公布详细的模型架构，后文中我们会对其架构进行逆向工程分析。

Sora 发布大约一周后，Google 也公布了其采用了类似模型架构的 Genie 视频生成模型，论文中明确指出是采用了 Spatiotemporal (ST) Transformers 代替了 Stable Diffusion 中的 UNet。Genie 使用户能够逐帧地在生成的环境中操作，而且是在无监督（数据没有标注）的情况下进行训练，更接近去世界模型的定义。虽然 Genie 目前只能生成 160X90 大小的视频，但是随着数据质量提升、模型规模的扩大，视频生成的尺寸和质量也会有所提升。

图表6：Genie 在生成视频中对主体动作的识别更为优秀



来源：《Genie: Generative Interactive Environments》、国金证券研究所

在差不多的时间，Snap 也发布了其使用了 Spatiotemporal(ST) Transformers 的视频生成模型，主要区别是其采用了 FIT（Far-reaching Interleaved Transformers）技术，该技术能降低在 Token/Patch 扩大的情况下的计算复杂度。一般来说，n 倍长度的 Token/Patch 在经过 Multi-head Self-Attention 时会有 n^2 倍的计算复杂度，经过 FIT 优化后可以实现 n^4/3 的计算复杂度，降低了生成长视频或者高分辨率视频的算力需求。

2.6 国内的绝大多数文生视频模型还处于 Diffusion 阶段，研发机构也在快速跟进

国内已有超 15 家企业推出了视频生成工具，既包括字节、百度、阿里、腾讯等 6 家巨头，也包括爱诗科技、生数科技、智象未来等 9 家创企。 智东西观察发现，文生视频领域大厂与创企各有领头羊，字节和 Morph Studio 在稳定性和成像质量方面表现出色。然而，大部分产品仍处于测试阶段，存在临时下线、排队时间长、无独立站点等问题。此外，生成视频效率低，2-4 秒视频的等待时间通常需要 3-5 分钟甚至更久。同时，现阶段文生视频的运动程度普遍较低，多为平移式运动或镜头运动，且对于人手、动物等非现实场景，大模型仍难以理解和生成。

图表7：国产视频生成模型比较

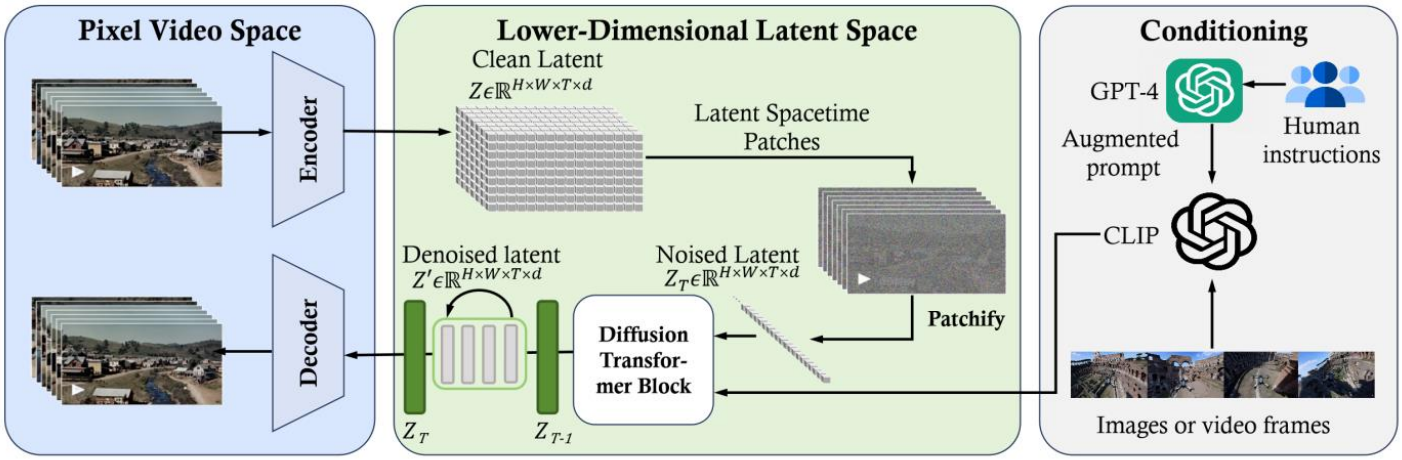
产品/模型	语义理解	运动程度	成像质量	总分
CapCut	4	4.4	3.5	16.1
Morph Studio	4	4	3.5	15.6
NeverEnds	3.8	3.8	2.9	14.4
艺映 AI	3.4	3	3.7	13.1
VideoCrafter 2	2.3	4	3.7	13
PixVerse	3.1	3	3.3	12.5
Vega AI	2.6	3	2.8	11.8
Pixeling	3.3	2.9	2.3	11.3

来源：智东西、国金证券研究所

目前已公开的国内视频生成模型还多数处于 Video Diffusion Models 阶段，还没有使用 Diffusion Transformer 架构的。国内公司和机构也在快速跟进，北大的 OpenSora 项目已经立项，计划复现 Sora 的模型架构与生成效果；字节在 3 月也将对自研的视频生成工具开启内测，鉴于字节已经拥有上万张计算卡的集群，并且原抖音 CEO 转向剪映业务，字节的新的视频生成模型也值得期待。

三、Sora 模型逆向工程

图表8：Sora 模型概览



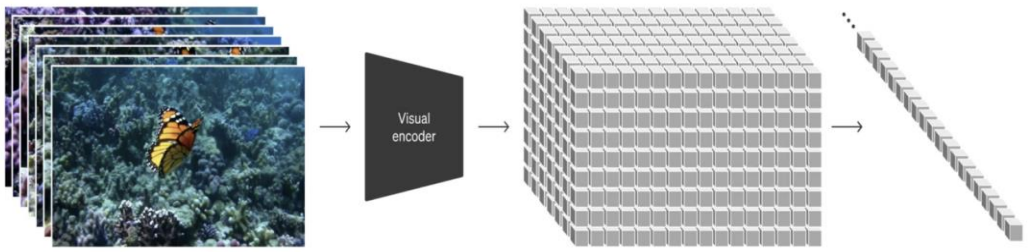
来源：《Sora: A Review on Background》、国金证券研究所

从本质上讲，Sora 是一个 Diffusion Transformer，具有灵活的采样尺寸，如图所示。它分为三个部分：1）Time-space compressor（时空压缩器）首先将原始视频映射到潜在空间中。2） 然后，ViT 处理标记化的潜在表示并输出去噪的潜在表示。3）类似 CLIP 的调节机制接收 LLM 增强的用户指令（使用 GPT-4 增强）和潜在的视觉提示，以指导扩散模型生成样式或主题的视频。经过多次降噪，得到生成的视频的潜在表示，然后用相应的解码器映射回像素空间。

3.1 Video Encoding：将视频信息有效的转化为机器理解的方法是至关重要的

视频生成模型的核心问题之一是视频数据的形式多种多样，包括分辨率、宽高比等。同时，视频包含的信息纬度是高于文本和图片的，其中包含着空间位置、时间、内容信息。因此 Sora 的重要工作之一就是找到一种方式，可以将多种类型的多维视觉数据转化为统一的表示方法，方便进行大规模的训练。

图表9：Sora 技术报告中的 Encoding 模式



来源：《Sora: Technical Report》、国金证券研究所

其中第一步是将视频原始内容提炼成一个潜空间特征（Latent representation），这一步与大语言模型的 tokenization 类似，将人类可以理解的内容转化成机器可以理解的内容，区别是视频内容需要保留时间、2D 空间位置和内容信息，而文字模型只需要保留内容和 1D 位置信息。由于视频单帧的像素量过大，所以这一步也承担了压缩的功能，当前模型会把单帧压缩成 16x 16 或者 32x32 的数量。

图表10：视频生成模型 Patch 方法对比

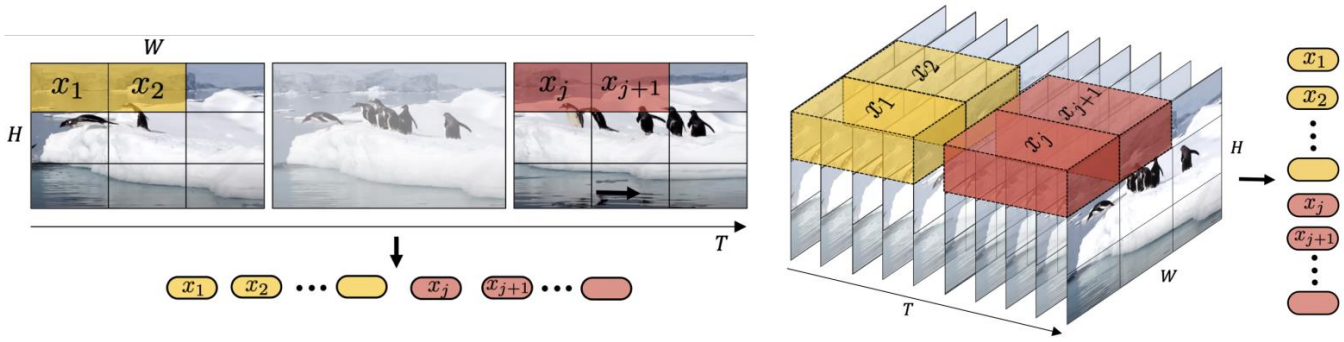
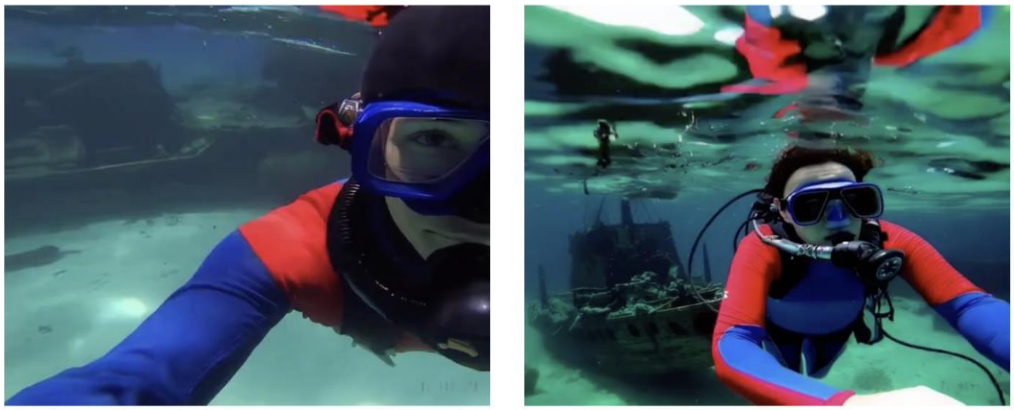


Figure 9: Comparison between different patchification for video compression. Source: ViViT [38]. (**Left**) Spatial patchification simply samples n_t frames and embeds each 2D frame independently following ViT. (**Right**) Spatial-temporal patchification extracts and linearly embeds non-overlapping or overlapping tubelets that span the spatiotemporal input volume.

来源：《Sora: A Review on Background》、国金证券研究所

由于 Transformer 无法直接接受高维的数据进行训练，下一步是将视频特征拆分成时空图像块（Spacetime Patches），根据 Sora 技术报告中引用的文献，目前主要有两种方法，其一（左图）是将每一帧分成 H * W 个 patch，然后根据时间线性排列成一维，patch 总量为 HxWxT。其二（右图）是从一段视频片段中提取一系列 patch，也就是所谓的 tube patch。这种方式是将 VIT 的 embedding 扩展到 3D 形式。一个 patch 的大小是 t * H * W，即时间窗口乘以图片分块的宽和高，这种方法 patch 总量在相同 patch size 下与第一种方法相同，也是 HxWxT，对单帧图像来说 H 和 W 均除以 t，能够提供更精细的图像信息。但是这种方法 patch 内部包含了时间信息，导致 patch 的大小变大，且增加了生成 patch 的计算量。Sora 技术报告没有披露具体的生成 patch 方法，鉴于其 patch 名称为 spacetime patch，并且生成的视频有更好的时空连续性，我们倾向于其使用第二种 patch 方法。

图表11：Sora 生成不同比例的视频内容保存度更好



来源：《Sora: Technical Report》、国金证券研究所

Sora 另一个不同于以往的视频生成模型的特点是可以生成自由宽高比的视频，并且视频的关键元素能很好的保留下来。主要原因是，过去的模型比如使用的 ViT（Vision Transformer）的每个图像块（patch）都必须是同一个固定尺寸，且原图必须是正方形。

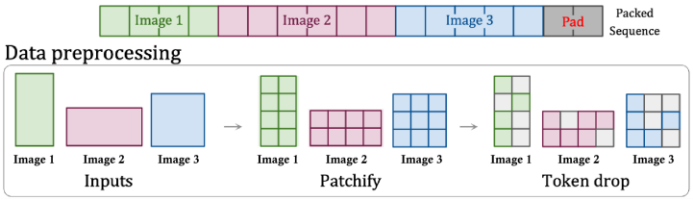
根据参考文献，Sora 大概率参考了 Navit 的实现方式，在组成时空块的时候，通过一种称为“Patch n’ Pack”的技术，允许在训练过程中处理不同分辨率和宽高比的输入。在这种技术下，不同宽高比和分辨率的内容都可以拆成图像块。但拆图像块的逻辑可以灵活调整，可大可小，从而适应不同分辨率。而来自于不同图像的图像块内容，可以被打包在同一序列里。这样的话，不同分辨率、宽高比的内容都可以灵活组合成图像块。并且这篇

论文中还有一个技术可以根据图像相似度，丢掉雷同的图像块，实现更快的训练。

3.2 模型的核心部分：Diffusion Transformer

Sora 是一个基于 Transformer 的 Diffusion Model。模型结构最初由 Scalable Diffusion Models with Transformers 这篇论文提出，也就是 DiTs。

图表12：Navit 的数据处理方法



来源：《Sora: A Review on Background》、国金证券研究所

图表13：DiT 的核心架构

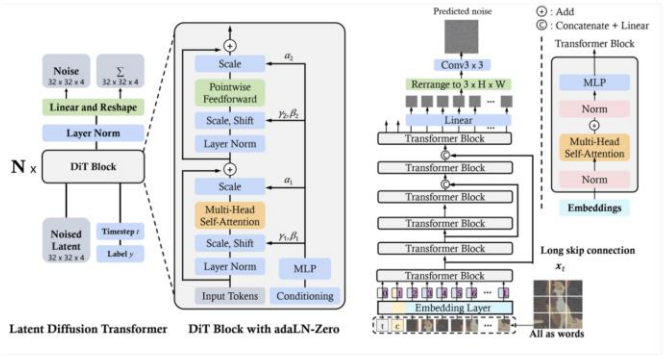


Figure 11: The overall framework of DiT (left) and U-ViT (right)

来源：《Sora: A Review on Background》、国金证券研究所

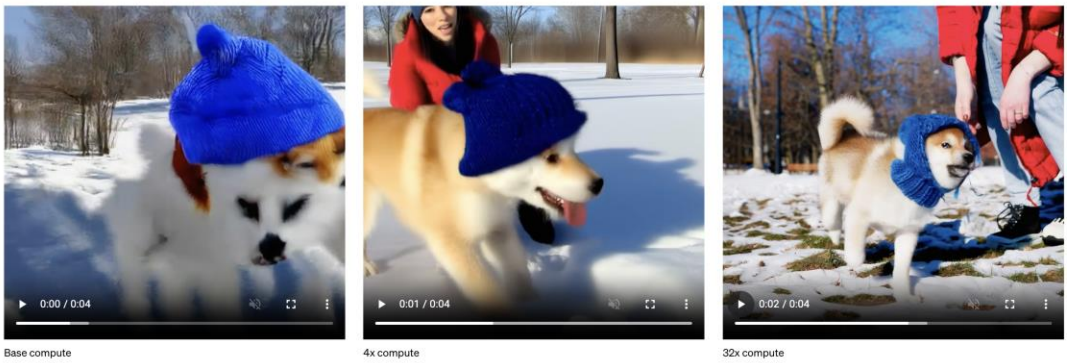
Stable Diffusion 由三个主要模块组成，每个模块都由独立的神经网络实现：

1）文本编码器（Text Encoder）：采用 Transformer 模型，将文本中的每个词/Token 编码为向量特征。2） 图像信息生成器（Image Information Creator）：Stable Diffusion 的核心部分，负责将文本编码后的向量特征与初始化噪声结合，生成包含图像信息的数组。3）图像解码器（Image Decoder）： 将图像信息数组还原为清晰的图像。

DiTs 主要工作也就是 Sora 主要应用的部分，就是将第二部分，由 U-Net 替换成了 Transformer。换成 Transformer 的原因是，使用 Transformers 可以很好地保持原有的优秀特性，比如可伸缩性、鲁棒性、高效性等，并且使用新的标准化架构可能在跨领域研究上展现出更多的可能。Sora 的技术报告并未披露其 Transformer 的架构，紧跟着 Sora 推出的 Google 的 Genie 和 Snap 的 Snap Videos 均采用了 ST-Transformer(Spatio-temporal Transformer)，在模型的架构层也针对视频的时空性进行了优化。

Sora 证明了视频生成模型的 Scaling Law 正是因为采用了 Transformer，类似于 GPT3.0 发布的时候证明了大模型的能力可以随着算力的提升、模型规模的扩大而提升生成效果。OpenAI 进行了 1x、4x、32x 算力情况下的生成效果对比，32x 算力生成的视频明显好于更低算力的结果。

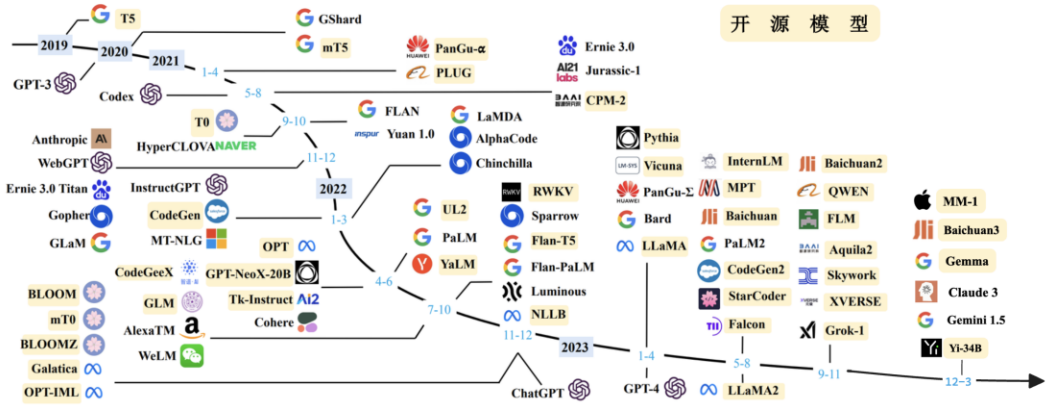
图表14：不同算力下 Sora 生成视频的对比



来源：《Sora: Technical Report》、国金证券研究所

3.3 大语言模型训练和推理对计算资源的需求分布不同

图表15：大语言模型最新发展追踪



来源：《A Survey of Large Language Models》、数字未来实验室、国金证券研究所

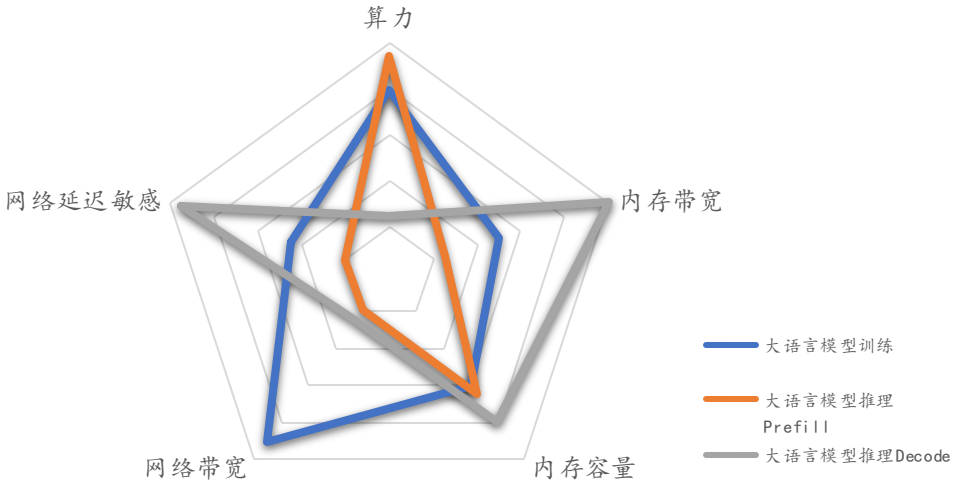
截止到 2024 年 3 月初，我们跟踪了国内外推出的大模型，可以发现模型机构和公司的竞争在加剧，推出新的大模型的速度在加快。我们总结出以下几点趋势：

1）长上下文(Long-Context)：最新的模型如 Gemini1.5 和 Kimi 支持到百万级别的 Token，对训练和推理时的内存容量和算力提出更高要求。

2）多模态(Multi-Modal)：理解图片、视频、音频信息是大模型的确定趋势，这些信息同样有这更大的 Token 数量，也会增大内存容量的需求。

3）MOE(Mixture-of-Experts)：越来越多模型包括 Mixtral、Gemini1.5 和 Grok 在内的模型在应用 GPT 的 MOE 提升效果。除了直接扩大参数规模，MOE 的多个子模型能够处理不同问题，虽然也会增加参数数量，但是在推理时只调用部分子模型，增加计算效率。

图表16：大语言模型训练和推理过程的计算需求分布

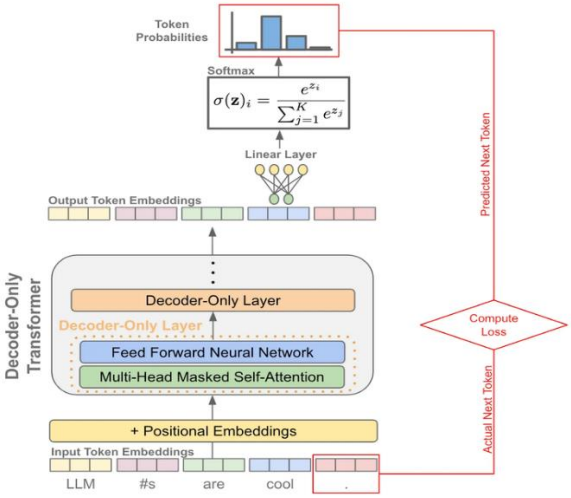


来源：Opening AI Infrastructure by Meta、数字未来实验室、国金证券研究所

对于大模型，其训练和推理过程中对计算资源的需求也大相径庭，其中训练时算力和网络带宽的资源比较紧缺，推理分为两个过程，prefill 对算力和内存容量的需求比较紧缺，decode 过程更需要内存带宽、内存容量和较低的网络延迟。

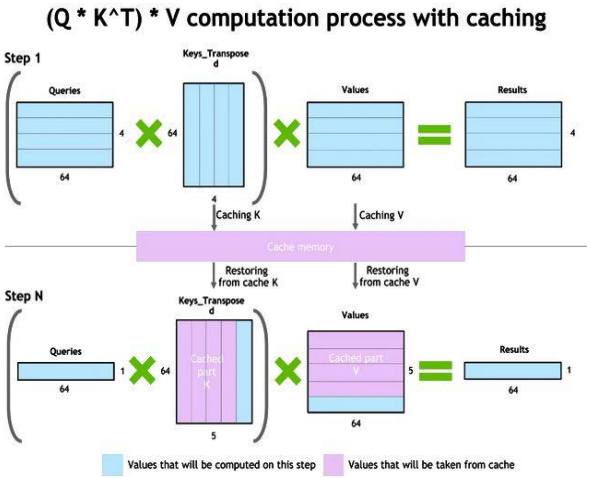
大语言模型训练时一次性对一整个句子每个 Token 进行下一个 Token 的预测，并计算所有位置 Token 的 Loss 并逐步优化，可以并行计算，需要大规模的算力和集群，所以训练对机器之间的网络带宽要求较高。

图表17：大语言模型训练过程



来源：《Sequential Modeling for Reinforcement Learning》、国金证券研究所

图表18：大语言模型推理过程



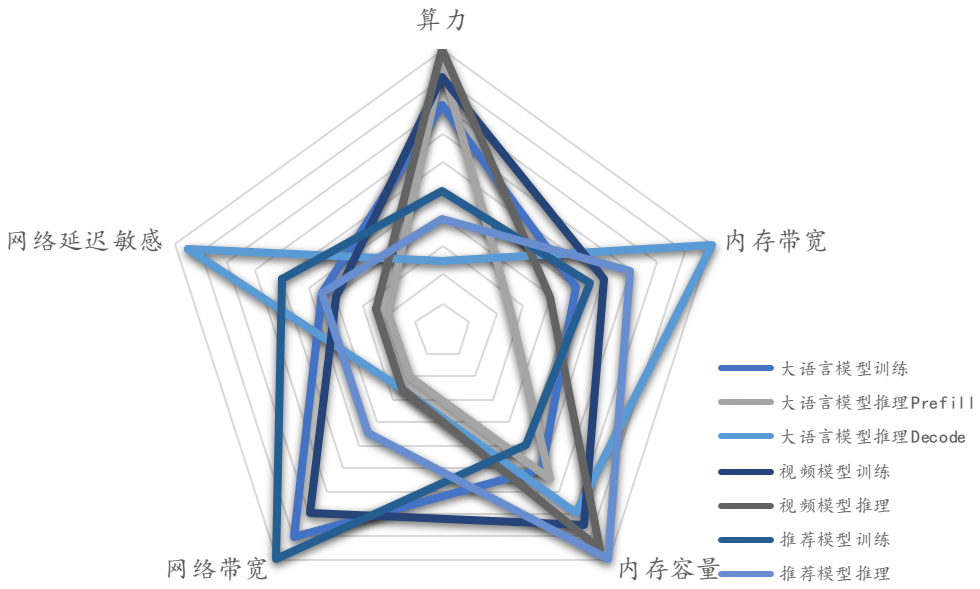
来源：Nvidia 开发者文档、国金证券研究所

在大语言模型的推理过程中，首先进入的是 Prefill Phase，也就是预处理阶段。在这个阶段，模型会进行一次计算密集型的操作，即计算并缓存每一层的 key 和 value。这个过程对于每一个请求的 prompt 来说都是必要的，但它只需要进行一次。因为模型内部主要是矩阵乘法运算，所以这个计算过程是并行执行的，而生成的缓存被称为 KV Cache，是大语言模型的核心。

在生成回答时推理会进入 Decoding Phase，这是一个串行的过程，主要任务是生成新的 Token。这一阶段采用了自回归的方式，即利用上一步生成的 Token 以及之前的所有 Token 作为输入，来预测并生成下一个 Token。这个过程包含两个关键步骤：首先，使用前一阶段创建的 KV Cache 来计算并输出下一个 Token 的 embedding；其次，在计算过程中，会得到当前 Token 在每一层的 key 和 value，这些信息会被缓存起来，并更新到 Prefill Phase 阶段的 KV Cache 中。通过这样的方式，模型能够持续优化其预测，确保生成的序列既连贯又符合逻辑。

3.4 对算力需求的影响：Patch/Token 数量的大幅提高对内存容量需求有积极影响

图表19：视频生成模型与大语言模型对计算资源的不同需求



来源：Opening AI Infrastructure by Meta、数字未来实验室、国金证券研究所

视频生成模型和大语言模型在对算力的要求上最大的区别 Patch/Token 的数量区别，视频的 Patch 与视频的时长（T）、宽度（W）、高度（H）、Patch 密度（单帧 patch 的数量 Ps）都有关系。由于在 Transformer 模型中，每个 Patch 都会经过 Multi-head Self-Attention 层，该层的计算复杂度与 Token 数量的平方成正比，并且在训练过程中，模型需要存储 Self-Attention 层的中间结果，这些中间结果的存储空间与 Patch/Token 数量的平方也成正比。最后与视频信息的复杂度关系可以总结为：

1）视频时长：O(n^2)

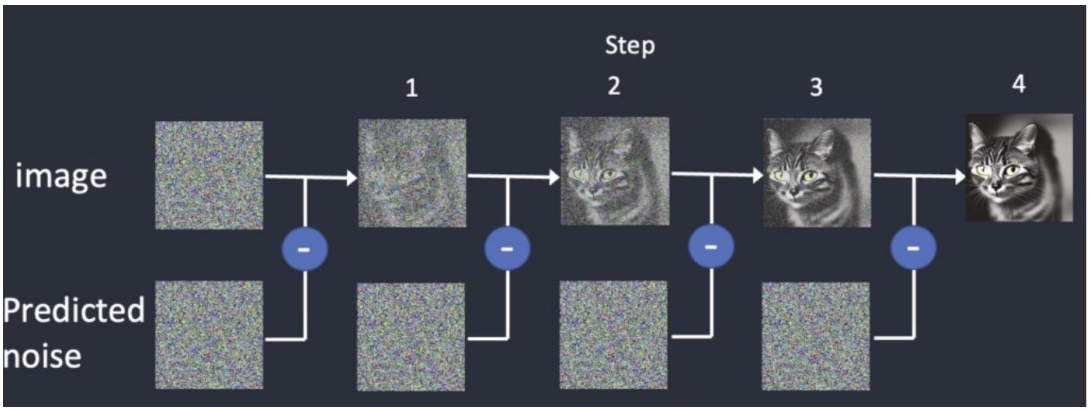
2）宽度或者高度(以正方形视频为例)：O(n^4)

3）Patch 密度：O(n^2)

以 Sora 目前能够生成的最长、分辨率最高的视频为例（1080P、30 帧/s、60 秒），在 Patch 密度为 32x32 的情况下，Patch 数量超过了 180 万，目前支持最长 Token 的大语言模型 Gemini 1.5 Pro 也只支持到了 100 万。考虑到 Sora 的参数规模据估计在 20-50B，与 GPT-4 相比也有一个数量级的减少，但是平均 Patch 数量至少增加了两个数量级，因此推理过程中所需内存的大小也会有数量级的提升。所以说在同一模型下，更大的算力和内存可以生成分辨率更高、时长更长的模型。同样的，对于同一提示词，模型参数量更大，生成的视频效果也更好，对算力和内存的要求也更高。

3.5 对算力需求的影响：推理时算力需求的增长大于内存速率需求的增长

图表20：Diffusion 模型推理生成图片的过程



来源：《How does Stable Diffusion work?》、国金证券研究所

Sora 和 GPT 的核心虽然都属于 Transformer，但是 Sora 的 DiT 和大语言模型在推理时的逻辑不同，Diffusion 需要基于一个随机的 noise latent 矩阵按照多个时间步迭代生成，每一步都在迭代细化 latent（图像/视频），使其更接近输入的提示词，这个步数在优化之后能减少到约 20 Steps 即可产生算力与效果均衡的结果。包括 GPT 在内的大语言模型是 Decoder-Only Transformer 架构，通过 Auto Regression 的方式预测下一个 Token，是一个完全的访存密集型场景，推理时性能瓶颈在内存带宽。 而 Sora 的 DiT 是一个 Encoder-Only Transformer 架构，推理的每一个 Step 时会输出全部长度的 Patch，一次性生成全部长度的 Patch，对计算卡内存的访存次数要远小于 GPT，是一个计算密集型场景。

图表21：目前用于训练和推理计算卡的算力/内存对比

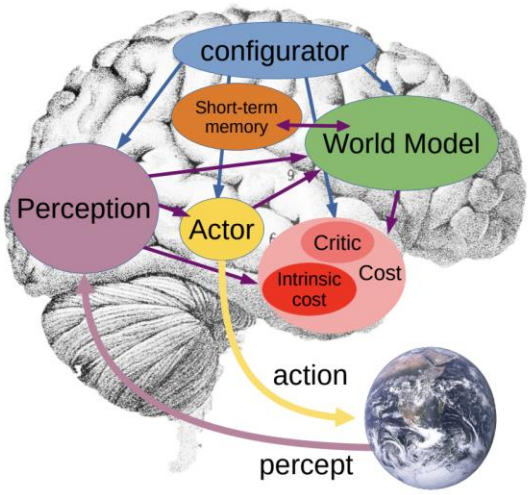
计算卡	算力（FP16）	内存大小	内存带宽
A100	624 TFLOPS	40/80GB HBM2e	2 TB/s
H100	1979 TFLOPS	96GB HBM3	3.35 TB/s
L40S	733 TFLOPS	48GB GDDR6	0.8 TB/s
H20	148 TFLOPS	96GB HBM3	4 TB/s
L20	119.5 TFLOPS	48GB GDDR6	0.8 TB/s
TPU v5p	459 TFLOPS	95GB HBM3	2.76 TB/s
昇腾 910B	376 TFLOPS	64GB HBM2e	\
Groq	188 TFLOPS	230MB SRAM	80 TB/s

来源：SemiAnalysis、Google TPU 产品网站、Arthurchiao、数字未来实验室、国金证券研究所

以国内特供的 H20 为例，由于美国商务部的禁令，算力做了较大的阉割，内存带宽相比 H100 反而有提升，所以在 LLM 的推理过程中反而能实现对 H100 有 10%的领先，但是对于视频生成模型来说，H20 相比 H100 会有较大的劣势。此外，Groq 的基于 SRAM 的芯片，内存带宽达到了 80TB/s，推理 Token 生成速度相比 GPT 和 Gemini 有超过十倍的提升，这种低算力、高带宽的芯片在视频生成模型推理中也是毫无优势。所以，我们认为高质量视频生成模型的普及和推理次数的增加会增加算力需求和内存大小而不是内存带宽，高算力、大显存的芯片是更适合视频生成模型的推理的，显存带宽相比 LLM 推理重要性降低，GDDR 亦可满足。

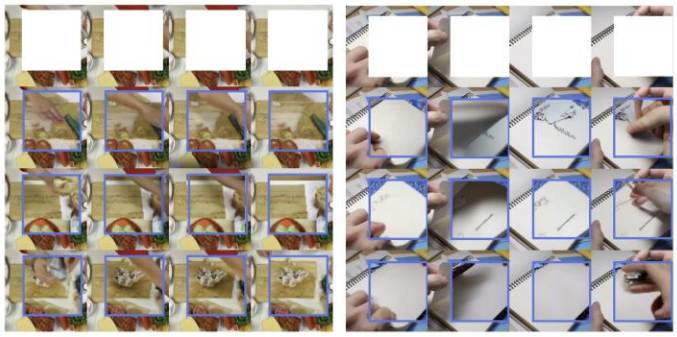
四、世界模型之争：三种 AI 路线的争论

图表22：LeCun 提出的世界模型



来源：《A Path Towards Autonomous Machine Intelligence》、国金证券研究所

图表23：V-JEPA 实现的视频预测



来源：《V-JEPA》、国金证券研究所

OpenAI 称 Sora 表现出来的涌现能力使其像是一个世界模拟器（World Simulator），但是世界模型（World Model）的提出者也就是 Meta 的首席科学家 Yann LeCun 并不同意，Meta 同日发表论文《Revisiting Feature Prediction for Learning Visual Representations from Video》并推出 V-JEPA 模型，通过学习图像和视频的表示，主要用于预测视频缺失的部分或者被遮住的部分，目标是希望从内在学习并理解物理世界的概念。他认为，大多数根据提示生成的逼真视频并不意味着模型能够理解物理世界，生成模型与基于世界模型的因果预测是两种截然不同的任务。生成模型的目标是生成看起来真实的视频，而世界模型的目标是理解物理世界并预测其未来状态。对于生成模型来说，可信视频的数量空间非常庞大，因此只需生成一个符合逻辑的样本即可算作成功。而对于世界模型来说，真实视频的合理延续数量空间要小得多，生成一个有代表性的片段是一个更难的任务，特别是在需要满足特定条件的情况下。

图表24：通往 AGI 的不同流派

Encoder特征理解派	Autoregressive生成派	语言中轴派
• 以Yann LeCun为代表的meta、google、stanford等科学家认为生成模型没有理解内容，并不是AGI的正确道路。 • 训练Encoder提取目标的特征才是理解图片/视频，通过预测修复不完整的图片/视频可以学习到图片/视频的重点与特征，同时使用自监督学习，更类似于人类通过观察来学习。 • 语言不是AGI的必要条件。	• Sora涌现的能力表示它是能够理解和模拟现实世界的基础模型，并且认为是实现AGI的重要里程碑。 • Scaling Law在当前算力水平下仍然有效，随着算力提升，可能涌现出更多接近AGI的能力。 • 语言是高效的数据，通过语言描述图片/视频/推理过程，生成模型也可理解这些内容。	• 以王小川为代表的认为AGI理想一定要以语言为中轴做模型。 • 认为Sora是需要把语言加进去，或者需要视频把语言加进去，才能变成往AGI走的引擎。 • 希望构建构建的是模型大楼，包括虚拟世界模型、生命模型和真实世界模型。 • 对不同场景和领域有不同模型，AGI中没有单一通用模型的存在。

来源：Meta、OpenAI、腾讯科技采访、数字未来实验室、国金证券研究所

生成派则相信 Scaling Law 会一直存在，随着算力、训练数据、参数规模不断扩大，会有更多接近于 AGI 的能力出现，同时与特征理解派想要提取图像或者视频的特征不同，生成派认为通过语言来描述图片视频甚至人类推理事物的过程，然后学习与语言的相似程度也可以理解这些内容。

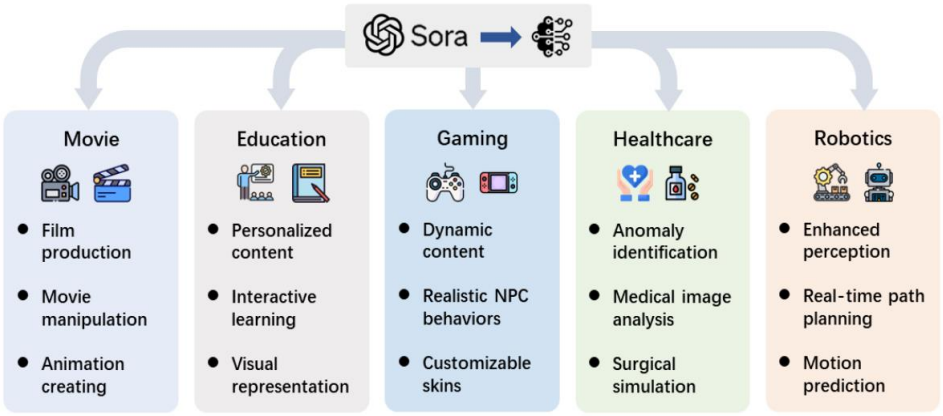
语言中轴派的王小川表示 AGI 需要以语言作为中轴、并且认为 Sora 需要加入语言才能成为通往 AGI 的引擎，并且他认为 AGI 中没有单一通用的模型的存在，需要模型大楼，包括虚拟世界模型、生命模型和真实世界模型。

总体来看，科学家们更支持重点在 Encoder 的特征理解派，而工程师和用户则更倾向于生

成派，Yann LeCun 成今年会发布的 Llama3 会更多的使用特征理解，与同样今年会发布的 GPT-5 对比后，通往 AGI 的路线会更清晰。

五、高质量视觉模型的出现的应用和对行业的影响

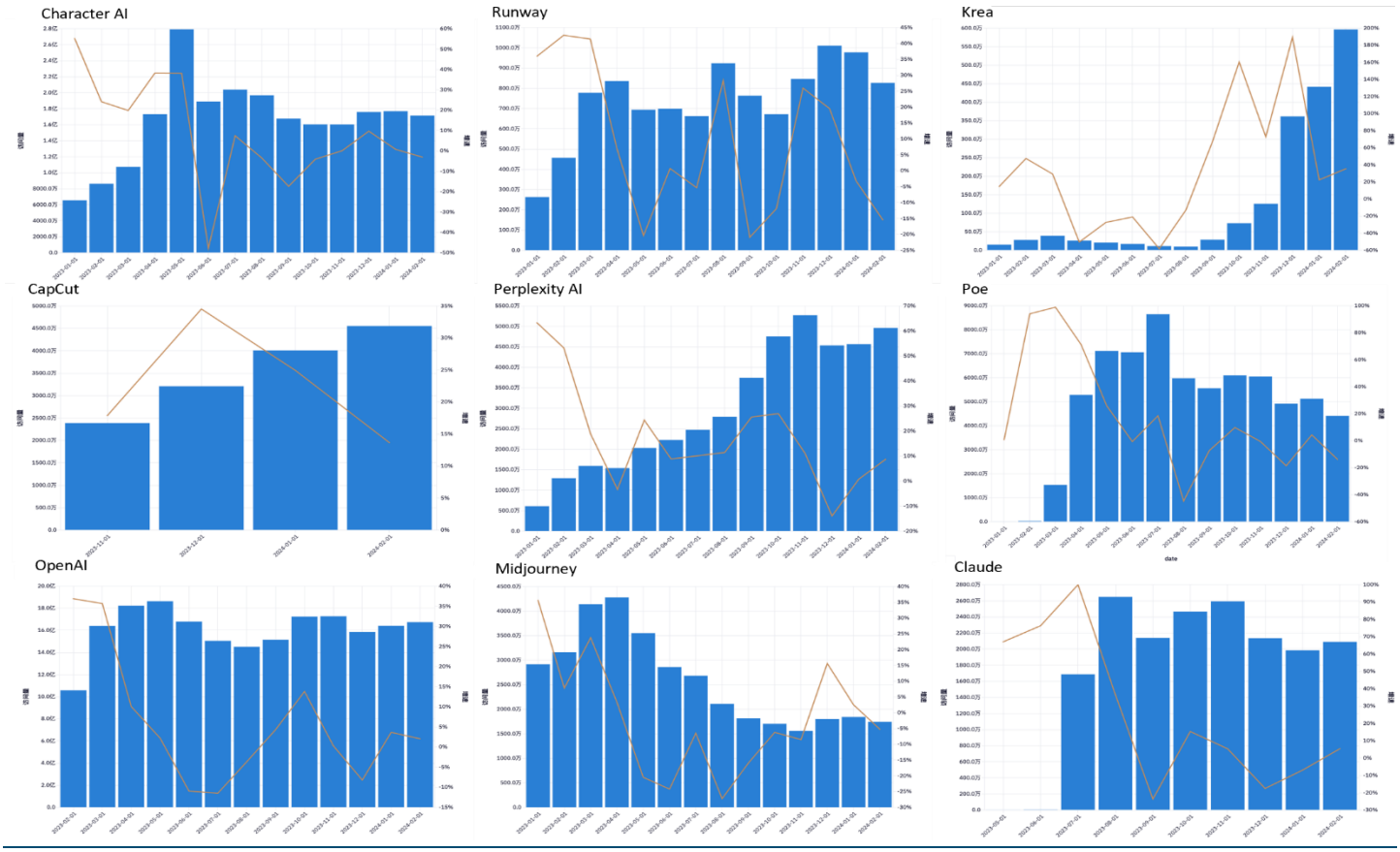
图表25：视频生成模型的应用行业



来源：《Sora: A Review on Background》、国金证券研究所

视频模型的应用仍处于高速发展阶段，发布时间比较晚的应用，比如 CapCut 和 Krea 的访问热度仍处在快速增长，较为成熟的应用比如 Runway 的热度也较为稳定。我们认为随着以 Sora 为代表的视频扩散模型成为一项前沿技术，视频生成的质量会不断提升，它们在不同研究领域和行业的应用也会继续加速，为从自动化内容生成到复杂决策过程等任务提供了变革性的潜力。

图表26：代表 AI 应用访问量热度变化



来源：SimilarWeb、国金证券研究所

5.1 影视制作

Sora 轻松生成吸引人电影内容的能力，预示着电影制作大众化的新时代。这展现了一个未来景象，任何人都有机会成为电影制作人，极大降低了进入电影界的门槛，并引入了一种新的电影制作维度，将传统叙述与 AI 驱动的创意完美融合。这些技术不仅简化了电影制作过程，还有望彻底改变电影制作领域的面貌，使其更加开放、多样化，更好地适应观众不断变化的偏好和分发渠道的发展。24 年 2 月 23 日央视频联合上海 AI 实验室等在 AI 的辅助下创作了《千秋诗颂》。导演表示，AI 使得团队创作从一个月 1 集加速到一个月 3 集，速度提升两倍。并在 29 日发布了联合清华大学元宇宙文化实验室制作的国内首部 AI 全流程微短剧《中国神话-补天》的片花，美术、分镜、视频、配音、配乐全部由 AI 完成。目前运用的模型还处于上一代 Stable Diffusion 技术路线，Sora 这种高质量的视频生成对于影视行业的工作流会有深远的影响，前期可以替代掉分镜以及概念片制作，后期可以取代部分特效制作。

图表27：AI 辅助制作的《千秋诗颂》

来源：CCTV 央视网、国金证券研究所

图表28：AI 全流程制作的《中国神话-补天》片花

来源：央视频、国金证券研究所

5.2 游戏

高质量的视频生成模型对游戏行业的影响是深远且革命性的，特别是在提升真实感和沉浸体验方面。这一技术的发展和应用，为游戏设计与开发打开了全新的视野，以下是几个关键方面：

1）动态环境与实时反馈：利用高质量的视频生成模型，游戏开发者可以创造出随玩家行为和游戏事件自然变化的环境。这不仅限于天气和景观的变化，更包括城市的发展、植被的生长或是季节的更替等。这种技术能够让游戏世界实时响应玩家的决策和行为，提供更加丰富和多样的游戏体验。

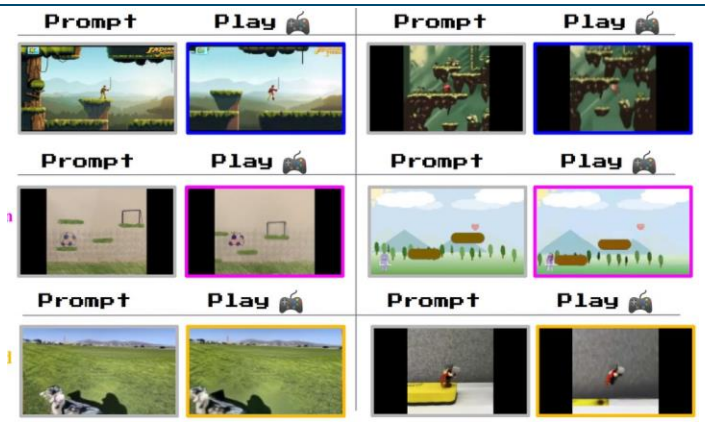
2）增强的故事叙述能力：通过高质量视频生成模型，游戏中的故事叙述能力将大大增强。开发者可以根据游戏的进展和玩家的选择实时生成对应的场景和剧情，使得每个玩家的体验都是独一无二的。这种个性化的故事叙述方式，能够极大提升玩家的沉浸感和情感投入。

3）更高效的资源利用：传统游戏开发中，创建高质量环境和角色模型需要大量的时间和资源。而利用视频生成模型，开发者可以快速生成高质量的游戏内容，减少了对专业艺术家和模型师的依赖。这不仅能够加速游戏的开发周期，也能够降低开发成本。比如使用 Sora 生成的高质量的场景视频，利用其涌现的 3D 一致性获取不同视角下场景的照片，通过 3D 重建工具可以生成对场景的建模，后期优化 prompt 实现理想化的运镜之后可以完成效果更佳的建模。

图表29：根据 Sora 生成的视频制作的 3D 模型

来源：《Sora: Technical Report》、数字未来实验室、国金证券研究所

图表30：Genie 实现操作输入图片中的主体



来源：《Genie: Generative Interactive Environments》、国金证券研究所

4）创新的游戏机制：高质量视频生成模型的应用，也促使开发者探索新的游戏机制。例如，利用这项技术模拟真实世界的物理反应和声音效果，可以创造出更为真实的战斗和交互体验。此外，运用类似 Google 的 Genie 这种动作模型，游戏中的角色和敌人也可以根据玩家的行为和游戏的进程进行实时的适应和变化，提供更具挑战性和可玩性的游戏内容。

六、风险提示

1. 模型架构的大幅改变影响算力需求分布：目前的大语言模型和最新的视频生成模型均基于 Transformer 架构，其训练过程对算力要求高，推理过程对内存芯片带宽要求较高，未来模型架构可能会发生变化，对算力需求的分布亦会有影响。

2. 算力速度发展不及预期：目前算力主要受限于芯片制程和互联技术的发展，随着摩尔定律的逐渐失效，未来算力速度的发展可能会放缓，可能影响模型的训练和推理。

3. 中美科技领域政策恶化：中美在 AI 领域竞争激烈，美国限制先进芯片和半导体对中国的出口，随着竞争的加剧，未来可能会推出更严格的限制政策，限制国内 AI 模型的发展。

特别声明：

国金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。

本报告的产生基于国金证券及其研究人员认为可信的公开资料或实地调研资料，但国金证券及其研究人员对这些信息的准确性和完整性不作任何保证。本报告反映撰写研究人员的不同设想、见解及分析方法，故本报告所载观点可能与其他类似研究报告的观点及市场实际情况不一致，国金证券不对使用本报告所包含的材料产生的任何直接或间接损失或与此有关的其他任何损失承担任何责任。且本报告中的资料、意见、预测均反映报告初次公开发布时的判断，在不作事先通知的情况下，可能会随时调整，亦可因使用不同假设和标准、采用不同观点和分析方法而与国金证券其它业务部门、单位或附属机构在制作类似的其他材料时所给出的意见不同或者相反。

本报告仅为参考之用，在任何地区均不应被视为买卖任何证券、金融工具的要约或要约邀请。本报告提及的任何证券或金融工具均可能含有重大的风险，可能不易变卖以及不适合所有投资者。本报告所提及的证券或金融工具的价格、价值及收益可能会受汇率影响而波动。过往的业绩并不能代表未来的表现。

客户应当考虑到国金证券存在可能影响本报告客观性的利益冲突，而不应视本报告为作出投资决策的唯一因素。证券研究报告是用于服务具备专业知识的投资者和投资顾问的专业产品，使用时必须经专业人士进行解读。国金证券建议获取报告人员应考虑本报告的任何意见或建议是否符合其特定状况，以及（若有必要）咨询独立投资顾问。报告本身、报告中的信息或所表达意见也不构成投资、法律、会计或税务的最终操作建议，国金证券不就报告中的内容对最终操作建议做出任何担保，在任何时候均不构成对任何人的个人推荐。

在法律允许的情况下，国金证券的关联机构可能会持有报告中涉及的公司所发行的证券并进行交易，并可能为这些公司正在提供或争取提供多种金融服务。

本报告并非意图发送、发布给在当地法律或监管规则下不允许向其发送、发布该研究报告的人员。国金证券并不因收件人收到本报告而视其为国金证券的客户。本报告对于收件人而言属高度机密，只有符合条件的收件人才能使用。根据《证券期货投资者适当性管理办法》，本报告仅供国金证券股份有限公司客户中风险评级高于 C3 级（含 C3 级）的投资者使用；本报告所包含的观点及建议并未考虑个别客户的特殊状况、目标或需要，不应被视为对特定客户关于特定证券或金融工具的建议或策略。对于本报告中提及的任何证券或金融工具，本报告的收件人须保持自身的独立判断。使用国金证券研究报告进行投资，遭受任何损失，国金证券不承担相关法律责任。

若国金证券以外的任何机构或个人发送本报告，则由该机构或个人为此发送行为承担全部责任。本报告不构成国金证券向发送本报告机构或个人的收件人提供投资建议，国金证券不为此承担任何责任。

上海
电话：021-80234211
邮箱：researchsh@gjzq.com.cn
邮编：201204
地址：上海浦东新区芳甸路 1088 号紫竹国际大厦 5 楼

北京
电话：010-85950438
邮箱：researchbj@gjzq.com.cn
邮编：100005
地址：北京市东城区建内大街 26 号新闻大厦 8 层南侧

深圳
电话：0755-86695353
邮箱：researchsz@gjzq.com.cn
邮编：518000
地址：深圳市福田区金田路 2028 号皇岗商务中心 18 楼 1806

【小程序】
国金证券研究服务

【公众号】
国金证券研究