2026 年 08 月 05 日

互联网行业研究

买入（维持评级）

行业专题研究报告

证券研究报告

传媒与互联网组
分析师：朱珺（执业 S1130526050001） zhujun1@gjzq.com.cn
分析师：柴梦婷（执业 S1130525030013） chaimengting@gjzq.com.cn

视频生成模型专题深度系列（一）

报告导读：

- 目前视频生成模型相比于 LLM 仍处于较早期阶段，技术范式尚未完全收敛，近期部分观点认为多模态能力更接近模型能力组件，而非推动智能能力上限提升的核心主线，但这均不影响多模态模型离商业化相对比较近。我们认为 AI 视频生成赛道仍有不错的商业模式，产业趋势持续向好，正处于商业化加速验证期。从技术端和产业端来看，AI 视频生成赛道都迎来积极的边际变化：
- 1、技术端：7 月 31 日，Seedance2.5 正式发布，相较于 2.0 重点突破：单次生成时长提升至 30 秒；多模态参考素材提升至 50 个；精准编辑。模型进一步释放生产力，且应用场景扩充至具身智能、工业制造、汽车等行业。同日，Minimax 发布全模态生成模型 H3，支持对文本、图像、视频、声音组成的多模态上下文的统一理解能力、能够输出具备原生双声道的音视频。
- 2、产业端：1）短剧持续爆发：据中国网络视听协会，26Q1，新上线 AI 微短剧数量占比超 95%。据 DataEye，今年 1 月至 5 月 AI 剧/漫剧市场规模已突破 220 亿元，全年有望冲击 400 亿元。2）长剧/电影领域拓展：《太平年》中可灵 AI 深度参与部分虚拟场景及视觉特效镜头的创作。国内首部获得《网络剧片发行许可证》的全流程 AIGC 网络故事片《奇谭：纸刃渡荒墟》登陆爱奇艺，标志着中国 AIGC 影视从 AI 短片/概念片到 AI 商业长片的突破。
- 在技术不断迭代推动之下视频生成产业趋势持续向好，我们看好赛道商业化潜力爆发。后续核心关注及潜在催化包括：Seedance 2.5 实际效果、ARR 转化；H3 实际效果及商业化贡献；其他模型 Kling 等更新及 ARR。
- 本系列将就以下问题做答：1）视频生成模型发展进程与技术迭代路线？技术瓶颈？2）以海外为鉴，Sora 关停启示？3）核心玩家对比及自身迭代复盘？4）商业化变现途径？市场空间？5）竞争壁垒和格局？6）产业链玩家？

投资逻辑：

本篇报告将聚焦于前三个问题进行分析：

- **DiT 为目前主流基础架构，但技术范式尚未完全收敛。**视频生成模型经历了从以 GAN/VAE 为代表到 Token 化 Transformer 与 Diffusion 并行发展再到 DiT 规模化，已呈现出 Scaling 效应，但尚未迎来类似 LLM 的技术范式收敛时刻。下一阶段技术突破仅靠继续扩大模型规模、训练数据或许还不够，核心目标或将是推动模型从拟合表面规律走向学习可预测的动态关系，提升复杂场景下的物理合理性。
- **Sora 是 AI 视频生成行业重要转折点之一，其关停表明模型能力突破并不必然转化成商业成功。**Sora 实现了生成时长突破；复杂提示词理解能力增强；时序一致性与画面稳定性明显提升；部分初步模拟能力。Sora 今年全面关停，主要系，1）战略聚焦：OpenAI 全力备战 IPO，叠加应对 Anthropic 的竞争，将资源聚焦到 Coding、企业服务等核心领域；2）商业模式未跑通：用户热度不持久，粘性低，收入和成本倒挂。3）版权增加约束。
- **从核心玩家对比来看，视频模型各有千秋和侧重。**Seedance 2.5 更偏向“全模态长叙事与创作控制模型”，拥有最大规模的多模态参考数量、最长的单次生成时长和专业级精细编辑；Kling 3.0 系列更偏向“角色一致性与可控分镜模型”，优势在于分镜级精细控制和跨镜头主体一致性；H3 更偏向“全模态理解与参考编辑模型”，核心优势在于跨模态素材理解，对生成、参考和编辑任务的统一支持；Veo 3.1 更偏向“高画质镜头生成与延展模型”，核心优势在于高质量短镜头、首尾帧控制和连续视频延展。

风险提示

技术发展不及预期；竞争加剧；商业化进程不及预期等。

内容目录

图表目录

一、视频生成模型历史发展进程与技术迭代路线如何?

1.1 技术迭代：从 Diffusion/自回归 Transformer 过渡至 DiT 为主流

- 我们复盘视频生成模型历史发展及技术迭代路线，大致可以归为四大阶段，分别是早期学术探索阶段(2021 年之前)、Token 化 Transformer 与视频扩散并行发展期(2021-2023 年)、DiT 规模化期（2024-25Q1）和原生音视频与多模态创作产品化期（25Q2-至今)。

图表1：*视频生成模型历史发展及技术迭代路线*

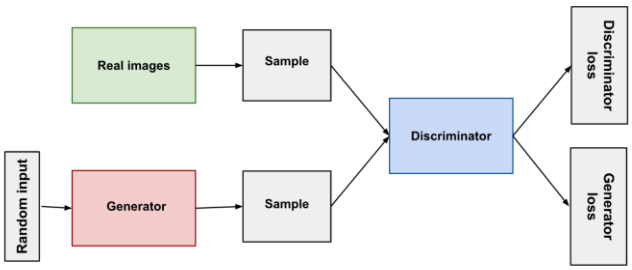
发展阶段	核心技术路线	主要能力与局限	代表模型
早期学术探索期（2016—2020 年）	以 GAN 和 VAE 式变分潜变量模型为代表。GAN 通过生成器与判别器的对抗训练生成视频；VAE 类方法通过学习潜变量分布和采样，建模视频未来状态的不确定性	主要用于无条件视频生成、根据前序帧预测未来及特定数据集实验；生成视频通常较短、分辨率较低，场景和运动类型有限，尚未形成成熟的开放域文生视频能力	GAN 路线：VGAN、MoCoGAN；VAE 类路线：SV2P、SVG-LP
Token 化 Transformer 与视频扩散并行发展期（2021—2023 年）	一条路线先将视频转化为离散视觉 Token，再通过自回归或掩码 Transformer 生成；另一条路线将图像扩散模型扩展至时间维度，并进一步发展出潜空间视频扩散	文生视频能力逐步形成，文本控制、画质及时序连贯性有所提升；但复杂运动、主体一致性、生成效率和较长时间范围内的稳定性仍然有限	Transformer 路线：VideoGPT、CogVideo、Phenaki、VideoPoet；扩散路线：Video Diffusion Models、Imagen Video、Stable Video Diffusion
DiT 规模化期（2024—2025Q1）	视频先通过压缩网络或视频 VAE 映射至低维潜空间，再被表示为时空 Patch 或 Token，由 Transformer 承担核心生成任务	Transformer 更便于扩大模型、数据和计算规模，时空建模、提示词遵循、运动生成和跨帧一致性进一步提升；复杂场景和长时稳定性仍有改进空间	Sora、CogVideoX、HunyuanVideo、Wan 2.1
原生音视频与多模态创作产品化期（2025Q2 至今）	两类能力加速进入商业产品：一是联合生成画面、对白、音效和环境声；二是利用文本、图片、视频和音频等素材进行参考生成、视频修改及上下文编辑	视频生成由无声画面扩展至同步音视频，并进一步支持多素材参考、续写和视频编辑；不同模型支持的输入、输出和编辑能力并不完全相同	原生音视频：Veo 3/3.1、Seedance 1.5 Pro；多模态生成与编辑：Seedance 2.0、Runway Aleph/Aleph 2.0

来源：arxiv.org，国金证券研究所

- 2021 年之前：早期学术探索期，以 GAN（生成式对抗网络）和 VAE（变分自编码器）为代表。
  - ✓ GAN 通过“生成器—判别器”对抗训练生成视频：生成器根据随机噪声或条件信息合成视频，判别器判断视频来自真实数据还是生成器，生成器据此不断提升结果的真实性。VAE 在训练时将视频压缩成带有随机性的潜变量，并训练解码器根据这些潜变量还原视频，训练完成后，模型可以随机采样一组潜变量，再由解码器生成一段新视频。当前视频模型仍使用 Video VAE 等进行时空压缩与解码，VAE 更多承担“视频压缩器和解码器”的角色，而不是直接决定生成视频的内容。
  - ✓ GAN 需要让生成器和判别器不断博弈，一旦双方训练失衡，模型就可能难以收敛，或反复生成少数几类相似视频，从而影响内容多样性。与图像生成相比，视频生成还需要学习物体运动和帧间一致性，而随着分辨率和视频时长提高，时空数据规模、计算及显存需求增加，对抗训练更加困难。受当时模型架构、算力和训练稳定性等限制，早期视频 GAN 集中于低分辨率、短时长视频生成。VAE 通常要求生成视频在每个像素上尽量接近真实视频。当同一场景存在多种合理的运动或细节变化，而潜变量又未能充分区分这些可能性时，模型可能在不同结果之间取

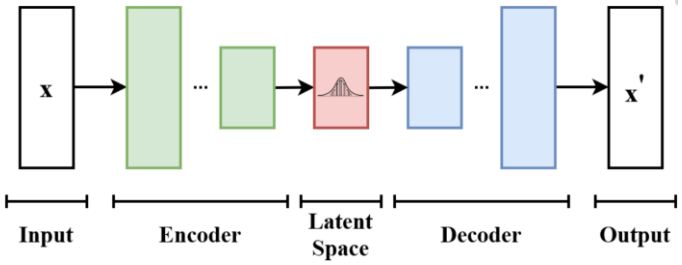
折中，导致画面偏平滑、边缘模糊，细节和纹理不足。

图表2：GAN 结构图



来源：Google for developers，国金证券研究所

图表3：VAE 结构图



来源：Wikimedia Commons，国金证券研究所

- 2021-2023 年：Token 化 Transformer 与视频扩散并行发展。当时主流路线：

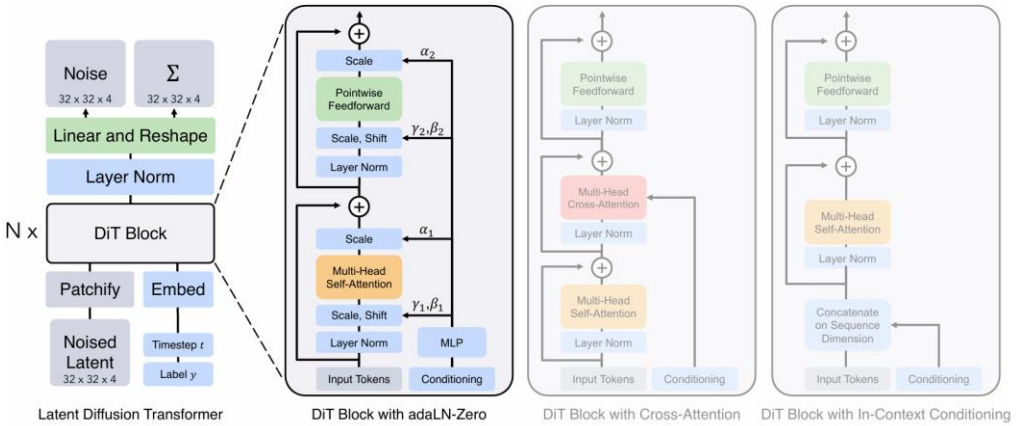
  1）基于扩散模型 Diffusion 生成视频：训练时不断向真实视频加入噪声，让模型学习反向去噪，生成时从随机噪声出发，在文字提示的引导下经过多轮去噪，逐步形成清晰视频。传统扩散模型通常采用 U-Net 作为去噪网络，通过逐级压缩和恢复视频特征、融合不同尺度的信息，预测每一步需要去除的噪声。该路线优势是通常能够生成质量较高的内容、训练相对稳定，缺点是需要多次迭代，生成速度较慢。代表模型为 Runway Gen-1 等。

  2）基于 Transformer 架构大语言模型生成视频：像语言模型一样预测视频 Token。模型先将文字、图像、视频音频压缩为离散化 Token，再根据已有 Token 依次预测后续 Token，注意力机制可以寻找不同画面、时间和输入条件之间的关联。该路线优势是能以统一的 Token 序列处理多种模态，在同一模型中灵活组合多种输入和生成任务，缺点是生成速度较慢，长序列计算开销较大。代表模型为 Google VideoPoet 等。

- 2024-2025 年：Diffusion 与 Transformer 融合，DiT 成为主流架构。

  ✓ DiT 架构：模型通常先通过视频压缩网络（VAE 或类似的自编码器），将原始视频在空间和时间维度压缩到计算量更低的潜空间，再将视频潜变量切分为一系列时空 Patch，并转换为 Transformer 能够处理的 Token。训练时，模型向视频潜表示加入噪声，在文字、图像等条件的引导下，由 Transformer 建模不同画面区域及前后帧之间的关联，并学习如何将带噪表示逐步更新为干净的视频表示。生成时从随机噪声出发，经过多轮迭代得到视频潜变量，最后由解码器将其还原为像素空间中的完整视频。Sora、可灵等均明确采用了 Diffusion Transformer 或相近的 Transformer 扩散架构，DiT 已成为视频基础模型的主流技术路线之一。

图表4：DiT 架构



来源：《Scalable Diffusion Models with Transformer》，国金证券研究所

- ✓ 与此前的自回归 Transformer 路线不同，DiT 一般不是按照固定顺序逐个预测“下一个视频 Token”，而是在扩散框架下，对整组时空 Token 进行多轮更新。其核心变化是以 Transformer 替代扩散模型中传统的 U-Net，作为处理带噪潜变量的主干网络，从而将扩散模型的迭代生成机制与 Transformer 较强的规模扩展能力、长距离依赖建模能力结合起来。
- ✓ 去噪环节从 U-Net 到 Transformer，核心区别在于：U-Net 是从局部出发，通过多层网络逐渐理解整体，再恢复细节；Transformer 则是把视频拆成一组时空 Token，让不同画面区域和前后帧的信息相互参考，在整体关系中不断更新并还原视频。核心优势包括：1）时空关系建模能力更强。更容易联系相隔较远的画面区域和前后帧，提升视频连续性。2）模型规模扩展能力更好。随着模型规模和训练算力增加，性能通常能够持续提升。

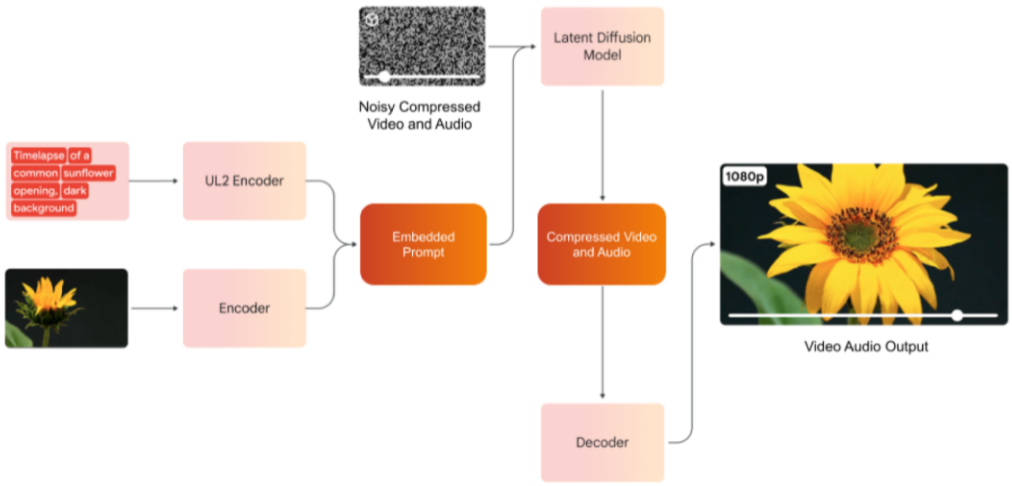
图表5：U-Net 对比 Transformer

对比维度	U-Net 去噪主干	Transformer 去噪主干（DiT）
处理方式	采用 U 形多尺度网络，结合卷积及时间模块逐层处理带噪视频特征	将视频表示为时空 Token，利用注意力计算不同区域及帧之间的关系
主要优势	局部细节和多尺度结构处理成熟，计算相对高效	长程时空关系、跨帧一致性和模型规模扩展能力更强
主要局限	远距离关系主要通过多层网络和额外时间模块传递	视频 Token 数量庞大，注意力计算和显存开销较高
代表模型	Video Diffusion Models、Imagen Video、Video LDM、Stable Video Diffusion	Sora、可灵 1.0、Seedance 1.0、Seedance 1.5 Pro、HunyuanVideo

来源：arxiv.org，国金证券研究所

- 2025-至今：原生音视频与多模态创作成为主流的进化方向，DiT 架构仍为核心底座。
  - ✓ 1）原生音视频：此前视频模型通常先生成画面，再通过其他模型后期添加音效，而原生音视频模型则在同一生成过程中共同生成画面和音效，使人物口型、动作和声音内容更自然地对应。核心优势是减少后期拼接，提升音画同步和整体叙事效果。代表模型包括 Google Veo 3、Seedance 1.5 Pro 及可灵 VIDEO 3.0 等。

图表6：Google Veo 3 架构



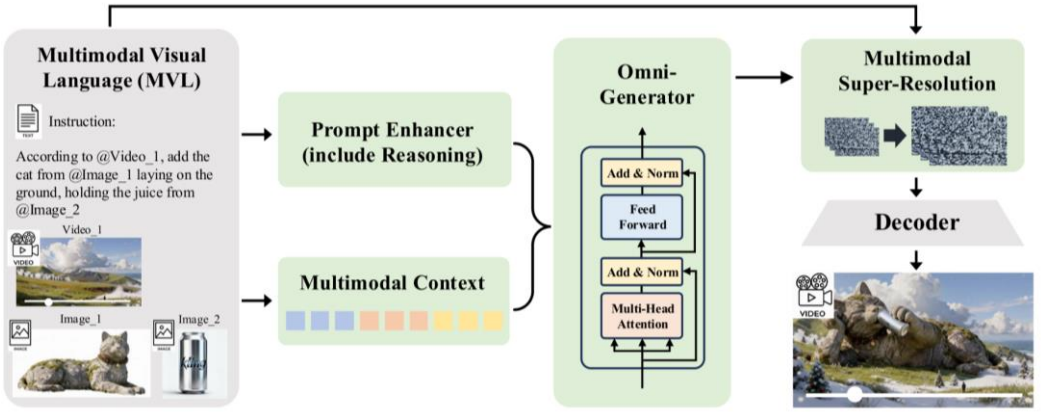
来源：《Veo 3 Technical Report》，国金证券研究所

- ✓ 2）多模态创作：模型不再只接受文字或单张图片，而是可以同时理解文字、图片、视频和音频等多模态素材。其核心变化是将过去分散在多个模型和工具中的生成、参考与编辑能力逐步整合到统一模型中，使视频创作从“单次生成”逐步走向连续、可控制的完整工作流。代表模型包括可灵 O1/VIDEO 3.0 Omni、Gemini

Omni、Seedance 2.0 等。

✓ 以可灵 O1 为例，核心流程如下：1）模型同时接收文字指令、参考图片和参考视频，并以多模态视觉语言（MVL）的形式，将文字与视觉素材组织为统一输入表示。2）Prompt Enhancer 基于多模态大语言模型理解不同输入之间的关系，推断创作者的具体意图，并重新组织提示词，使优化后的提示词更符合模型熟悉的指令表达方式，进而改善主体身份一致性、空间关系连贯性、色彩还原度和物理合理性。3）优化后的指令特征与图片、视频形成的多模态上下文共同进入 Omni-Generator。生成器在共享嵌入空间中联合处理视觉 Token 和文本 Token，使不同模态进行交互，以提高视觉一致性。4）基础生成器首先输出视频的低分辨率潜变量，随后多模态超分辨率模块同时以该潜变量和原始 MVL 信号为条件，进一步补充纹理、边缘等高频细节，最后经解码输出视频。

图表7：可灵 Omni 架构



来源：《Kling-Omni Technical Report》，国金证券研究所

1.2 当下瓶颈：从拟合表面规律走向学习可预测的动态关系

- 目前，视频生成模型的技术评价重点已不再局限于能否生成视觉质量较高、表面合理的单个视频，而是进一步转向长时一致性、复杂交互和物理合理性等能力。头部模型实践及相关研究表明，增加模型规模和训练计算通常能够改善生成质量，扩充并优化视频和文本训练数据也有助于提升指令遵循，视频生成领域已呈现出 Scaling 效应。但现有研究也表明，仅靠 Scaling 或不足以使模型自动掌握能够稳定外推至训练分布外场景的物理规律。在速度、尺寸等条件超出训练分布的物体运动与碰撞场景中，模型仍可能更接近于模仿训练数据中的相似案例，而非学习出能够稳定外推的底层物理规则。
- 下一阶段的突破仅靠继续扩大模型规模、训练数据和算力或许还不够，核心目标之一或将是推动模型从学习表面规律，进一步转向建模物体状态、材料属性、运动轨迹、相互作用及其结果之间的动态关系，让模型学到更具预测性的动态结构。从而使模型在场景条件发生变化时，也能较准确地判断后续会发生什么，提升复杂场景下的物理合理性。
- 在新发布的 Seedance 2.5 描述中，也指出了新模型在复杂运动的物理合理性、极多主体交互场景的稳定性上，模型仍有提升空间。

图表8：视频生成模型在物理推理基准上的表现

所有得分均在 [0, 1] 区间内，分数越高越好。结果按推理环节、物理定律维度和数据来源维度三个维度进行分组展示。
P-T/P-G 表示感知-文本/感知-图形；F-T/F-G 表示定律表述-文本/定律表述-图形；Ded. 表示推演。
Grav.、Mom.、N1 和 Multi 分别表示万有引力定律、动量守恒定律、牛顿第一定律和多定律组合案例。

模型	平均分 (Avg.)	推理环节得分 (Track-wise Score)					物理定律维度得分 (Pillar-wise Score)				数据来源维度得分 (Source-wise Score)	
		感知-文本 (P-T)	感知-图形 (P-G)	定律表述-文本 (F-T)	定律表述-图形 (F-G)	推演 (Ded.)	万有引力 (Grav.)	动量守恒 (Mom.)	牛顿第一定律 (N1)	多定律组合 (Multi)	仿真数据 (Sim.)	真实数据 (Real)
视频模型												
Wan2.2 [56]	0.267	0.645	0.257	0.009	0.275	0.149	0.245	0.274	0.344	0.200	0.310	0.224
HunyuanVideo-1.5 [64]	0.177	0.230	0.292	0.027	0.180	0.155	0.128	0.205	0.272	0.183	0.208	0.146
VBVR-Wan2.2 [58]	0.373	**0.923**	0.502	0.001	0.239	0.201	0.347	0.375	0.459	0.387	0.394	0.352
Seedance 2.0 [49]	**0.473**	**0.597**	**0.490**	**0.478**	**0.485**	**0.315**	**0.450**	**0.510**	**0.495**	**0.389**	**0.487**	**0.459**
Veo 3.1 [63]	0.313	0.408	0.357	0.325	0.316	0.160	0.288	0.314	0.420	0.235	0.356	0.270

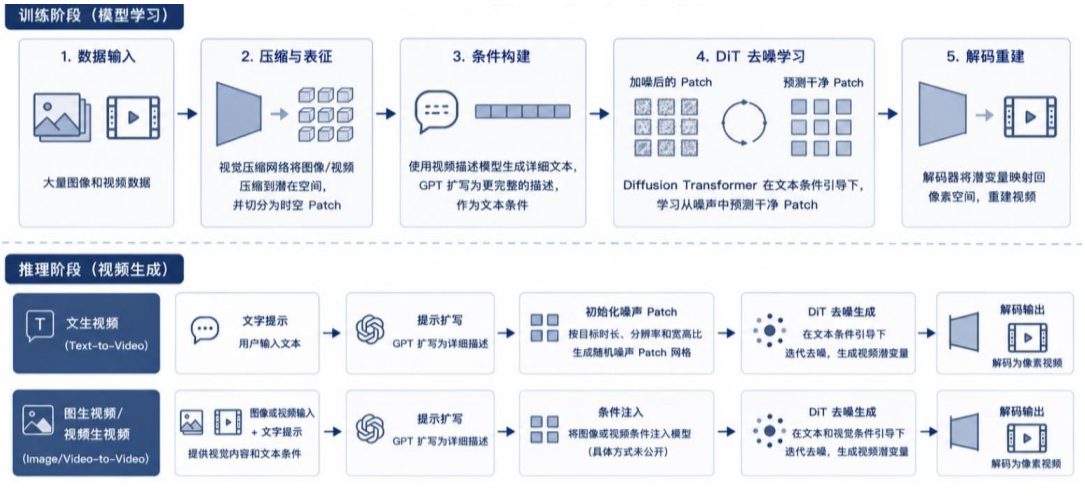
来源：《Apple-π:Benchmarking Thinking with Video Towards Law-Grounded Physical Intelligence》，国金证券研究所

二、以海外为鉴，Sora 关停的启示？

2.1 Sora 的突破：时长+文本理解+时序一致性+初现世界模拟

- 核心流程：1）构建统一视觉表征。训练阶段，Sora 通过视觉压缩网络将图像和视频从像素空间映射到经过空间和时间压缩的潜在空间，再将潜变量切分为一系列时空 Patch，作为 Transformer 处理的视觉 Token。2）构建文本与视觉条件。训练阶段，OpenAI 利用视频描述模型为训练视频生成更加详细的文字标签，以提高文本与视频内容的对应关系。生成阶段，GPT 会将用户的简短提示扩展为更完整的描述。3）Diffusion Transformer 迭代去噪。训练阶段，模型接收带噪声的时空 Patch，并在文本等条件引导下学习预测原始的干净 Patch。文生视频时，模型根据目标视频的时长、分辨率和宽高比，排列相应规模的随机初始化 Patch 网格，再通过扩散去噪逐步生成视频潜变量；图生视频和视频生视频则在输入图像或视频的条件下完成动画化、延展或编辑等任务。4）解码输出视频。生成完成后，视频潜变量通过对应的解码器映射回像素空间，形成最终视频。

图表9：Sora 视频生成核心流程



来源：OpenAI 官网，国金证券研究所

- Sora 的突破体现在以下几个方面：

1）生成时长与规格实现突破。Sora 能够生成最长约 1 分钟的高清视频，并支持不同的时长、分辨率和宽高比。Sora 通过将不同规格的图像和视频统一转换为“时空 Patch”，扩大了可利用的训练数据范围，也增强了模型处理长视频和多种画幅的能力。

2）复杂提示词理解能力增强。Sora 可以更准确地呈现提示词中的主体、动作、场景细节和镜头要求。得益于用重描述技术为训练视频生成更详细的文字标签，并利用 GPT 将用户的简短提示扩展为完整描述，从而提高文字条件与视频内容之间的匹配度。

3）时序一致性与画面稳定性明显提升。Sora 在部分长视频中能够维持人物、动物和

物体的外观，在主体被遮挡、短暂离开画面或切换镜头后，仍保留一定的连续性。其基于 Transformer 的扩散模型直接在时空 Patch 上建模，有助于学习相隔较远的画面区域和时间片段之间的关系，从而改善长时序一致性，但这种能力并非始终稳定。

4）初步展现“世界模拟”能力。 Sora 能够在动态镜头下保持一定的三维空间关系，并表现物体持续存在、人物动作改变环境状态等现象，说明大规模视频训练使模型开始学习现实世界中的部分空间与运动规律。但仍属于统计意义上的视觉模拟，Sora 尚不能可靠还原玻璃破碎、物体形变等基础物理过程。

- Sora 2 进化到视频生成 GPT-3.5 时刻：第一代 Sora 被 OpenAI 称为视频生成的“GPT-1 时刻”，强调其意义在于证明扩大预训练规模可以带来物体持久性等此前较难稳定出现的能力，而 Sora 2 进一步将行业推进至近似“GPT-3.5 时刻”——视频模型已由展示性生成迈向具备原生音频、复杂指令控制和初步物理模拟能力的实用阶段。

图表10：Sora 2核心升级

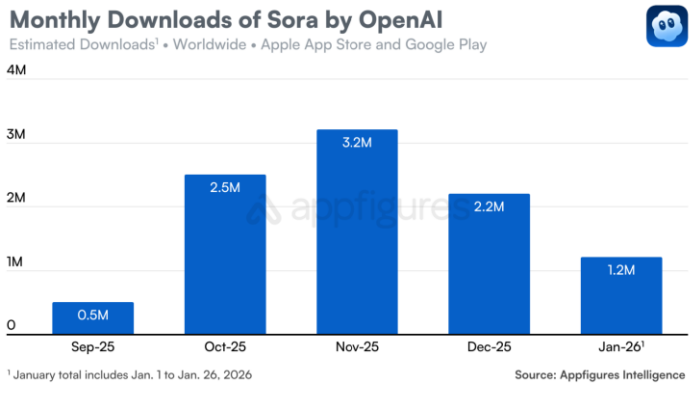
Sora 2 核心升级	相较第一代 Sora 的变化
物理模拟更准确	更能遵循物理规律，合理表现浮力、刚体运动以及动作失败后的结果
原生同步音频	从视频生成扩展为视频与音频生成，可同步生成对白、环境声和音效
可控性明显增强	能够遵循跨多个镜头的复杂指令，并更好地保持人物、物体和环境状态
真实感与风格范围提升	画面真实感提高，同时增强写实、电影和动漫等风格的生成表现

来源：OpenAI 官网，国金证券研究所

2.2 Sora 的关停：战略+商业模式未跑通+合规等问题共同导致

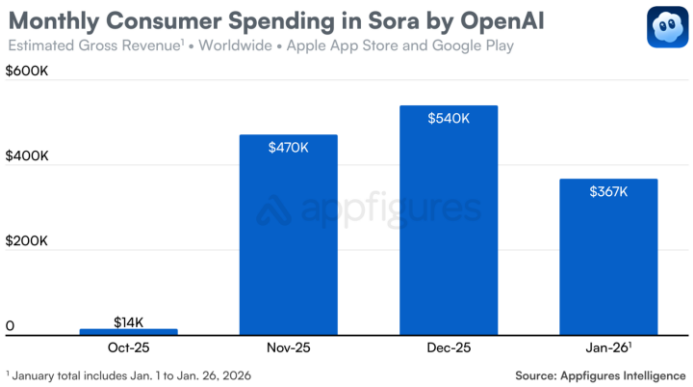
- 战略：OpenAI 全力备战 IPO，叠加应对 Anthropic 的竞争，将算力、人力、资金等资源聚焦到 Coding、企业服务等核心领域。Sora 的关停也是公司重新分配稀缺算力资源、减少非核心产品投入的举措。
- 商业模式：
  - ✓ 用户数据在上线初期表现亮眼，热度并不持久：Sora 上线 5 天内下载量突破 100 万，并迅速登顶 App Store 榜首。但一个月后，下载量就开始明显下滑，2025 年 12 月，Sora 下载量环比下降 32%；2026 年 1 月，下载量继续暴跌 45%。
  - ✓ 用户粘性低：据 a16z 合伙人 Olivia Moore 转引的 Sensor Tower 早期数据（2025 年 11 月），Sora App 的 7 日用户留存率约为 2%，30 日留存率约为 1%，早期 60 日留存曲线接近 0%。用户体验一次新鲜感后，由于缺乏持续使用的刚需场景，难以找到长期使用的价值，从而导致用户粘性极低。
  - ✓ 商业模式未跑通，收入和成本严重倒挂：收入端，Appfigures 估算 Sora 整个生命周期的应用内购收入仅约 210 万美元。成本端，根据《福布斯》在较高用户使用量假设下的测算，其日算力成本可能达到 1,500 万美元。《华尔街日报》援引知情人士称，Sora 实际每天亏损约 100 万美元。整体来看，Sora 虽然实现了较大用户触达，但用户付费和持续度不足以覆盖高昂的视频生成成本，UE 尚未跑通。

图表11：Sora APP月度下载量



来源：Appfigures，国金证券研究所

图表12：Sora APP月度消费者支出



来源：Appfigures，国金证券研究所

- 合规：Sora 2 仍未完全解决版权问题。产品上线初期，OpenAI 采取“主动退出”（opt-out）机制，即版权方需主动申请，才能限制相关角色被生成，引发影视公司和版权组织反对。2025 年 10 月，OpenAI 宣布向更接近 opt-in 的权利人控制机制调整。此后，美国电影协会和日本 CODA 均要求 OpenAI 加强生成端的拦截，并停止未经许可使用版权作品。2025 年 12 月，OpenAI 与迪士尼达成授权合作，允许 Sora 在授权范围内使用超过 200 个旗下角色，显示其开始转向明确授权模式。但该合作仅覆盖特定 IP，训练数据授权、未授权角色识别及规模化版权合作机制等问题仍未完全解决。

三、核心玩家对比以及自身迭代版本复盘?

3.1 对比：Seedance 重全模态长叙事与控制、Kling 重角色与分镜、H3 重全模态理解与参考编辑、Veo 重镜头质量与延展

我们选取目前头部视频生成最新模型进行对比，核心结论：

- Seedance 2.5 更偏向“全模态长叙事与创作控制模型”：1）输入与参考方面，模型延续 2.0 统一的多模态音视频联合生成架构，单次最多可输入 30 张图片、10 段视频和 10 段音频，并强化白模、运动和创意参考。2）生成与控制方面，模型可在 30 秒内组织多个有逻辑关联的镜头，使故事从铺垫、推进、转折到收尾逐步展开，并通过优化镜头衔接和场景过渡提升连贯性；在延展中也可保持主体特征、场景环境和叙事节奏的连贯。同时，模型支持时间戳级控制与编辑，可定向修改特定时段的角色、动作和剧情，并增强绿幕、视角、运镜及参考编辑能力。3）输出与延展方面，单次可生成最长 30 秒的音视频，并支持多轮延展。4）整体来看，Seedance 2.5 将大规模多模态参考、30 秒长叙事和精细音视频编辑相结合，核心优势在于更长的单次生成时长、长叙事连贯性和专业级精细控制。

- Kling 3.0 系列更偏向“角色一致性与可控分镜模型”：1）输入与参考方面，3.0 Omni 支持文字、图片、视频和 Element（将角色外观和声音沉淀为可重复调用的主体资产）组合输入，无视频输入时最多可使用 7 张图片或 Element，有视频输入时可使用 1 段 3—10 秒视频，并搭配最多 4 张图片或 Element。2）生成与控制方面，角色外观和声音可沉淀为重复调用的主体资产，在不同场景和镜头中保持一致；支持自动多镜头和自定义多镜头生成，用户可逐镜头设定时长、景别、视角、叙事内容和运镜。3）输出与延展方面，单次可生成 3—15 秒、720p 或 1080p 音视频。4）整体来看，可灵 3.0 系列将角色资产复用、逐镜头编排和原生音视频生成相结合，核心优势在于分镜级精细控制和跨镜头主体一致性。

- Minimax H3 更偏向“全模态理解与参考编辑模型”：1）输入与参考方面，H3 可统一理解文字、图片、视频和音频组成的多模态上下文，单次最多支持 9 张图片、3 段视频和 3 段音频，混合素材总数不超过 12 个，并可通过自然语言描述不同素材之间及其与目标视频之间的关系。2）生成与控制方面，模型将文生视频、首尾帧生成、多模态参考、视频动作迁移和音视频编辑等任务统一建模，可分别参考或修改人物、动作、运镜、风格、声音和剪辑节奏。3）输出与延展方面，单次可生成 4—15 秒、最高 2K、24 FPS 的视频。4）整体来看，H3 将全模态理解、多模态参考与生成编辑统一到同一套系统中，核心优势在于跨模态素材关系理解，对生成、参考和编辑任务的统一支持。

- Veo 3.1 更偏向“高画质镜头生成与延展模型”：1）输入与参考方面，模型支持文字、首帧、首尾帧和 Ingredients 参考生成，最多可输入 3 张人物、角色或产品参考图。2）生成与控制方面，Veo 3.1 强调电影化画面表现、提示词遵循、首尾帧之间的自然过渡和原生音频生成，但公开接口不支持多段视频和独立音频混合参考。3）输出与延展方面，单次可生成 4 秒、6 秒或 8 秒视频，支持 720p、1080p 和 4K 及原生音频；视频延展每次增加 7 秒、最多延展 20 次，合并后最长约 148 秒，但仅支持 720p，且输入必须是此前由 Veo 生成的视频。4）整体来看，Veo 3.1 的核心优势在于高质量短镜头、首尾帧控制和连续视频延展，但其素材参考规模相对中国的模型更有限。

图表13：Seedance 2.5 VS Kling 3.0 Omni VS H3 VS Veo 3.1

模型	Seedance 2.5	Kling 3.0 Omni	MiniMax H3	Veo 3.1
特征	更偏“全模态长叙事与创作控制模型”：围绕 30 秒多镜头叙事，强化镜头衔接、场景过渡、多轮延展、大规模多模态参考和时间戳级编辑，提升复杂内容创作的可控性	更偏“角色一致性与可控分镜模型”：以 Element 为核心复用角色外观和声音，并将主体一致性、原生音频与逐镜头编排结合	更偏“全模态理解与参考编辑模型”：统一处理文字、图片、视频和音频，将生成、参考、动作迁移及音视频编辑等任务纳入同一框架	更偏“高质量镜头生成与延展模型”：突出画面质量、电影化表现、提示词遵循、原生音频、参考图控制和视频延展
总体架构	延续 Seedance 2.0 的统一多模态音视频联合生成架构；未进一步披露具体 Transformer、VAE 及参数量等底层结构	Kling 3.0 系列采用深度融合的统一模型训练框架，将多模态输入、原生音频、Element 和多镜头生成整合至统一流程；具体网络结构及参数量未披露	完整系统由 H3-Context-IR、H3-Base 和 H3-Regenerate-2K 组成。H3-Base 采用 33B 参数的稠密单流 H3-Omni-Transformer，将不同模态编码为统一序列并联合预测音频与视频潜变量	Veo 3 采用音视频联合潜空间扩散架构：音频和视频分别经自编码器压缩为潜变量，再由基于 Transformer 的去噪网络联合生成；Veo 3.1 未披露单独的底层架构变化
输入模态	支持文字指令，以及图片、视频和音频等参考素材混合输入	支持文字、图片、视频和 Element 混合输入；Element 可通过多角度图片或角色视频建立，并可上传声音样本绑定角色音色	支持文字、图片、视频和音频输入；覆盖文生视频、首帧、尾帧、首尾帧及多模态参考生成	Gemini API 支持文字和图片输入；此前由 Veo 生成的视频可作为延展输入
参考素材规模	单次最多输入 30 张图片、10 段视频和 10 段音频，可混合使用	无视频输入时最多输入 7 张图片或 Element；有视频输入时可输入 1 段 3—10 秒视频，并合计输入最多 4 张图片或 Element	最多输入 9 张图片、3 段视频和 3 段音频，混合素材总数不超过 12 个；视频和音频单段不超过 15 秒；音频不能单独输入	最多输入 3 张参考图片，用于保持人物、角色或产品的外观；首帧和尾帧属于独立的帧控制模式
多模态参考能力	可综合理解参考素材中的构图、场景、风格、人物、道具和声音；强化白模、运动和创意参考，可控制空间结构、人物姿态、运动轨迹、镜头角度和运镜	图片、视频、Element 和文字均可作为提示信息，可参考人物、商品、场景、动作及运镜，并将人物外观和声音沉淀为可复用 Element	可通过自然语言指定不同素材与目标视频之间的关系，例如分别参考视频运镜、图片人物和音频音色；支持人物、动作、镜头、风格、声音及剪辑节奏参考	通过 Ingredients 参考图、首帧和首尾帧控制人物、产品、物体、场景及视觉风格；公开 API 不支持多段视频和独立音频混合参考
复杂动作与指令	可结合文字、白模、运动参考和时间戳控制复杂动作、运动轨迹及镜头；复杂运动的物理合理性和多主体交互稳定性仍有提升空间	提升文本指令响应和动作表现，在多人群像及互动场景中可分别锁定人物或物体特征，并将动作、运镜和主体一致性结合	强调复杂多模态指令遵循、跨模态关系理解和视频到视频动作迁移；H3-Context-IR 负责理解、整理和细化复杂多模态指令	强调提示词遵循、复杂动作描述、电影化镜头运动和视觉物理真实感；可通过详细提示词控制主体动作、环境变化和镜头运动
多镜头与分镜	30 秒内可组织多个具有逻辑关系的镜头，使故事形成铺垫、发展、转折和收尾；可利用时间戳控制特定时段的叙事、视角、运	支持自动 Multi-Shot 和 Custom Multi-Shot，可逐镜头指定时长、景别、角度、叙事内容和运镜，并优化镜头之间的衔接	预训练任务包含原生多镜头建模，可通过自然语言描述不同镜头及视听关系；未披露逐镜头参数化分镜界面	可通过提示词描述景别、视角、运镜、动作和镜头变化；公开 API 未提供逐镜头时长和参数化分镜设置

	镜和节奏			
角色一致性	在多人同框和群像叙事中，可保持多个人物的外观、声音和主体特征；视频延展时可保持主要人物、环境和叙事节奏	可将角色外观和独特声音绑定为 Element，并在不同视频、场景和镜头中反复调用；复杂群像中可分别维持多个主体特征	参考生成可在单段生成视频中保持参考人物、产品或其他资产的主要特征	可利用最多 3 张参考图保持同一人物、角色或产品的外观，并将其置于不同场景中
原生音频	基于统一音视频联合生成架构，可同步生成音视频并接受音频参考；2.5 进一步提升音质、人物声音保持和音画同步	支持原生音视频同步生成，可生成对白、环境声和音效；Element 可绑定人物音色并跨视频复用；有参考视频输入时暂不支持同时开启原生音频	H3-Omni-Transformer 联合预测音频和视频潜变量，输出 32kHz 双声道音频；语音、音效和音乐统一建模，稳定支持 11 种对白语言	原生生成对白、音效和环境声，并提升音视频同步；自然且连贯的短对白仍属于持续改进方向
视频编辑	支持时间戳级音视频编辑，可修改指定时段的角色、动作、声音或剧情；同时强化绿幕换背景、视角、运镜及参考编辑，并保持修改前后的连续性	视频编辑能力延续 O1，支持主体增加、移除或替换，背景修改，局部及整体调整，风格、色彩、天气、构图和视角修改，以及根据参考视频生成前后镜头	将参考与编辑统一为通用任务，可通过自然语言表达素材的保留、修改、迁移和复用关系，覆盖图片到视频、视频到视频及音视频联合编辑	官方模型体系提供扩画、增加或移除物体、角色控制、运动控制和场景延展等能力；Gemini API 主要开放参考图、首尾帧和视频延展
单次生成时长	最高 30 秒	3—15 秒，支持整数时长	4—15 秒，支持整数时长	4/6/8 秒；参考图、1080p 或 4K 仅支持 8 秒；延展每轮 7 秒，最多 20 轮，最终≤148 秒
输出分辨率	当前公开服务支持 480p、720p 输出	模型指南支持 720p 和 1080p；Kling 3.0 系列已上线原生 4K 视频生成能力，可直接生成 3840×2160 视频	支持 768P 和 2K，输出帧率为 24 FPS	支持 720p、1080p 和 4K；4K 不适用于 Veo 3.1 Lite，视频延展仅支持 720p
视频延展	支持多轮延展，并在延展过程中保持主要人物、环境、声音和叙事节奏	可根据参考视频生成前一镜头或后一镜头；Kling 平台另提供自动延展和自定义延展功能，最长约 3 分钟	未披露独立的视频续写或时长延展功能；H3-Regenerate-2K 用于高分辨率再生成	可将此前由 Veo 生成的视频每次延长 7 秒，最多延展 20 次；输入视频最长 141 秒，合并后最长约 148 秒，延展仅支持 720p
开放状态	未公开模型权重	未公开模型权重	部分开放权重：H3-Base 的任务检查点、Transformer、Visual VAE、Audio VAE 及推理代码已开放；H3-Context-IR、H3-Regenerate-2K 及稀疏注意力实现尚未开放	未公开模型权重

来源：各公司官网，国金证券研究所

其他主流视频生成模型如阿里巴巴的 HappyHorse、爱诗科技的 PixVerse、生数科技的 Vidu，技术相关信息披露相对有限，我们适当减少对比维度，核心结论如下：

- HappyHorse 1.1 重点突出多参考图理解与融合能力，可提高参考图片与生成结果之间的一致性，更准确地保留商品、人物和场景等视觉信息；同时通过优化动作建模与时序一致性，改善复杂动作的流畅性和连贯性，并提升指令遵循、画面细节及音画同步能力。模型支持文生视频、首帧图生视频和参考图生视频，其中参考图生视频可输入 1—9 张图片，最高输出 1080P，单次最长 15 秒。
- PixVerse V6 重点强化多片段生成、音频生成及图像 / 视频混合参考能力；其参考生视频模式可理解输入视频中的主体、动作、场景、运镜和视觉风格，并根据提示词进行主体替换、视频重构和动作模仿，最多可同时输入 10 张参考图和 2 段参考视频，参考视频总时长不超过 15 秒。V6 支持文生视频、图生视频、首尾帧生视频、视频延展和参考生视频，常规模式支持 1—15 秒、360P 至 1080P 输出，并可选择是否生成音频。
- Vidu Q3 系列以音视频同步生成、镜头控制和叙事节奏为核心能力，可直接输出包含对白、旁白、音效和音乐的视频。Q3 单次最长生成 16 秒，并提供帧级镜头运动与叙事节奏控制。参考生视频最多可输入 1—7 张参考图；在主体参考模式下，可通过图片或文字定义主体，图像或文字素材合计不超过 7 项，每个主体最多配置 3 张图片。部分子型号支持智能切镜，Q3 还支持多人对话以及中文、英文和日文音视频生成。不同子型号分别覆盖文生视频、图生视频、首尾帧生视频和参考生视频，最高支持 1080P 输出。

图表14：HappyHorse 1.1 VS PixVerse V6 VS Vidu Q3 系列

模型	官方重点能力	主要生成方式	参考素材能力	镜头 / 分镜控制	音频能力	最高输出规格
HappyHorse 1.1	提升文本语义理解、镜头调度和动态生成表现；强化多参考图下主体、场景风格及画面一致性，并改善图生视频的画面质感、跨片段一致性、动作流畅度、文字稳定性和音画同步	文生视频、首帧图生视频、参考图生视频	首帧图生输入 1 张图片；参考图生输入 1—9 张参考图；不支持参考视频或自定义音频输入	官方强调镜头调度，可通过提示词描述景别、镜头切换和运镜；未披露时间戳或逐镜头独立设置参数	文生、首帧图生和参考图生均支持生成有声视频；未提供自定义音频上传或主体音色设置参数	最高 1080P；3—15 秒
PixVerse V6	支持多片段生成、内联音频、视频延展及图像 / 视频混合参考；可理解参考视频中的主体、动作、场景、镜头运动和视觉风格，并按提示词进行内容调整或动作模仿	文生视频、图生视频、首尾帧生视频、视频延展、参考生视频	图生输入 1 张图片；首尾帧输入 2 张图片；最多输入 10 张参考图和 2 段参考视频，参考视频总时长不超过 15 秒	文生和图生视频支持多片段生成参数；首尾帧、延展和参考生视频未提供相同参数，未披露时间戳级逐镜头设置	文生、图生、首尾帧、视频延展和参考生视频均支持内联音频生成；另提供音效生成和口型同步能力	最高 1080P；1—15 秒

Vidu Q3 系列	系列重点覆盖原生音视频输出和智能切镜；Pro 侧重综合表现，Mix 侧重效果平衡，Drama 侧重对白、角色站位、动作及电影化叙事，Ad 侧重短广告节奏和细节一致性，Turbo 侧重速度与性价比	Pro 支持文生、图生和首尾帧；Mix、Drama、Ad 支持参考生视频；Turbo 支持文生、图生、首尾帧及参考生视频	图生输入 1 张首帧图；首尾帧输入 2 张图；参考生视频支持 1—7 张参考图。部分模式可用图像或文字定义主体，素材总数不超过 7 个，每个主体最多使用 3 张图片；Q3 系列不支持参考视频输入	Pro、Mix、Ad 和 Turbo 明确支持智能切镜；Drama 强调多角色对白、角色空间站位和电影化叙事；未披露时间戳级逐镜头时长设置	Pro、Mix、Drama 和 Turbo 明确支持同步音视频生成，可输出语音和音效；Ad 官方模型表未明确标注同步音视频	最高 1080P；24fps；Pro / Mix / Turbo 为 1—16 秒，Drama / Ad 为 3—15 秒

来源：各公司官网，国金证券研究所

3.2 Seedance 历史迭代：从高质量视频生成迈向统一多模态与长叙事创作

- Seedance 经历了几轮迭代（自 2025 年 6 月以来已完成四个主要版本迭代），从“高质量视频生成”向“原生音视频、多模态创作与长视频生产”持续升级。其中，Seedance 2.0 的推出是最关键的里程碑，从模型技术上做到全球 SOTA，也是第一个跨过生产质变点的视频生成模型。Seedance 2.0 采用统一的多模态音视频联合生成架构，支持文字、图片、视频和音频四种模态输入，同时具备稀疏架构的效能优势，并将多模态参考、视频生成、视频延长和编辑能力纳入统一创作体系。
- Seedance 2.5 版本于 7 月 31 日正式发布，延续了 Seedance 2.0 统一的多模态音视频联合生成架构，重点突破长叙事能力、多模态参考能力和编辑能力。主要包括以下升级：
  - ✓ 单次生成时长达 30 秒，支持多轮延长：Seedance 2.5 可单次生成 30 秒高质量视频片段，并支持多轮延长。同时，模型优化了镜头衔接和场景过渡，让长视频保持更好的连贯性。模型还大幅优化了画质、音质以及运动质量，并有效弱化了视频生成中常见的“油腻感”。用户可以生产数分钟具有统一视听语言、高质量的连贯内容，精彩故事一次呈现。
  - ✓ 多模态参考能力全面升级：Seedance 2.5 支持用户单次输入最多 30 张图片、10 段视频、10 段音频作为参考素材。模型还提升了白模参考、运动参考、创意参考等各类参考能力，使模型能更好地理解创作意图，实现多主体、多场景、多镜头变化的复杂创意。
  - ✓ 更精准、稳定的编辑能力：Seedance 2.5 支持精准的时间戳控制，可定向编辑视频内容，提升创作的效率和可控性。同时，模型强化了绿幕编辑、视角编辑、参考编辑等多种编辑能力，满足影视、广告等更多专业、复杂场景的创作需求。
- 此外，围绕 Seedance 的版权治理与商业化体系进一步完善。2026 年 6 月 10 日，火山方舟版权商业化平台正式上线，并与周星驰旗下比高集团达成 AI 视频创作版权合作，《喜剧之王》《长江七号》《食神》三项影视 IP 首批入驻。平台通过“授权—保护—审核—分发—变现”的全链路机制，为 IP 在 AI 场景下的合规使用和商业价值开发提供支持。

图表 15：Seedance 历史版本

模型	Seedance 1.0	Seedance 1.5 Pro	Seedance 2.0（升级后）	Seedance 2.5
发布 / 状态	2025 年 6 月 11 日正式发布	2025 年 12 月 16 日正式发布	2026 年 2 月 12 日正式发布；后升级原生 4K 能力	2026 年 7 月 31 日正式发布

核心架构	空间层与时间层解耦的 Diffusion Transformer；统一支持文生视频与图生视频;原生多镜头	双分支 Diffusion Transformer，由视频分支、音频分支及跨模态联合模块构成	统一多模态音视频联合生成架构，并采用稀疏架构	延续 Seedance 2.0 统一的多模态音视频联合生成架构；具体网络结构未披露
输入模态	文字、图片	文字、图片；支持文生音视频、图生音视频及无声视频生成	文字、图片、视频、音频；最多同时输入 9 张图片、3 段视频和 3 段音频	文字、图片、视频、音频；单次最多输入 30 张图片、10 段视频和 10 段音频
输出分辨率	API 支持 480p、720p、1080p	API 支持 480P、720p、1080p	API 支持 480p、720p、1080p 及 4K，其中 4K 为 10-bit 位深输出	当前公开服务支持 480p、720p 输出
原生音频生成	无	有；可生成人声、环境音、音乐及空间音效，支持多语种和方言	有；支持双声道音频，可并行输出对白、背景音乐和环境音效并与画面同步	有；支持音视频联合生成、音频参考和多语言内容创作，可在多主体场景中保持人物形象与声音
单次生成时长	API 支持 2-12 秒	API 支持 4-12 秒	API 支持 4-15 秒；可直接输出 15 秒多镜头音视频	最长 30 秒单段直出
多镜头及叙事	原生支持多镜头;镜头切换时保持主体、视觉风格和整体氛围一致	增强叙事连贯性，支持复杂镜头运动、角色表演和情绪表达	支持 15 秒多镜头音视频，可响应长脚本并自主规划镜头语言和叙事节奏	支持 30 秒连续叙事，可承载连续镜头运动、情绪推进和完整故事节拍
视频编辑	未公开通用视频编辑能力	未公开通用视频编辑能力	支持自然语言视频编辑，可定向修改指定片段、角色、动作和剧情	支持时间戳控制，可定向修改指定片段中的角色、动作、声音或剧情；强化绿幕编辑、视角与运镜编辑及参考编辑，并保持修改前后连贯

来源：公司官网，国金证券研究所

3.3 Kling 历史迭代：从视频生成可用化迈向统一多模态与专业叙事创作

Kling 迭代速度较快，逐步从基础视频生成扩展至多模态交互、生成编辑统一、原生音视频生成和专业级叙事编排。其中，1.0 和 O1 分别代表模型“真正面向用户可用”和“生成与编辑统一”两个关键里程碑；2.0、2.6 和 3.0 则分别补齐多模态交互、音频输出和专业创作能力。

- 1.0 达成的核心成就：全球第一个发布的、用户真正可用的 DiT 架构视频生成模型（当时 Sora 只有 demo，12 月才真正发布产品）。
- 2.0 达成的核心成就：推出 MVL（多模态视觉语言），本质是解决输入侧的问题：将交互方式从单一文字提示扩展至文字、图片和视频等多模态输入，更准确表达创作意图。
- O1 达成的核心成就:侧重多模态输入，允许用户在文本指令中插入各类非文本文件，来表达文字难以描述的意图，比如具体的人物形象、细微的动作指令等。同时，将多种视频生成与编辑任务统一到同一模型中。
- 2.6 达成的核心成就：侧重多模态输出，在生成高质量视频的同时，同步输出匹配的音频，实现音画同出，是 Kling 首个支持原生音频视频的模型。
- 3.0 达成的核心成就：迈向 All-in-One（AIO）模型，进一步统一多模态输入、音视频生成与编辑能力。其中 3.0 侧重通用生成和多镜头叙事，3.0 Omni 侧重多参考一

致性与专业分镜控制。

图表16：Kling历史版本

版本	核心架构/定位	输入模态	输出分辨率与单次时长	原生音频	参考/编辑/叙事能力
可灵 1.0 2024 年 6 月	采用 Diffusion Transformer 架构；自研 3D VAE 进行时空压缩，并采用时空联合注意力机制	文字、图片	最高 1080P、30fps；单次最长 2 分钟	无	强调复杂运动、物理规律和长视频生成；支持文生视频、图生视频及视频续写
可灵 1.5 2024 年 9 月	延续 1.x 基础架构，主要提升基础生成质量和运动表现；底层结构变化未披露	文字、图片	支持 1080P 高品质输出；5 秒或 10 秒	无	提升动态表现和语义响应；新增运动笔刷，可控制局部主体的运动区域和轨迹
可灵 1.6 2024 年 12 月多图参考于 2025 年 1 月上线	延续 1.x 技术路线，重点提升生成质量、文字响应和主体一致性；底层结构变化未披露	文字、图片，支持多图参考	支持 5 秒或 10 秒生成；分辨率沿用标准及高品质模式	无	新增多图参考，可结合多张人物、物体或场景图片，提升多主体及跨画面一致性
可灵 2.0 2025 年 4 月	推出 MVL（多模态视觉语言），以文字作为语义骨架，通过多模态信息补充人物、动作、场景和运镜等细节	文字、图片、视频片段；声音和运动轨迹可作为进一步扩展模态	支持 720P 和 1080P；5 秒或 10 秒	无	支持图片及视频参考、元素增删替换和多模态编辑，提升主体、动作及运镜控制能力
可灵 2.1 系列 2025 年 5 月	延续 2.0 技术路线，分为标准模式、高品质模式和大师版；底层架构变化未披露	文字、图片	标准模式 720P，高品质及大师模式 1080P；5 秒或 10 秒	无	强化复杂动作、物理表现、提示词响应和画面质量；支持首尾帧等控制能力
可灵 2.5 Turbo 2025 年 9 月	侧重基础模型质量、可控性和生成效率升级；未披露新的 DiT、VAE 或蒸馏结构	文字、图片	最高 1080P；支持 5 秒或 10 秒生成	无	强化复杂指令、人物互动、大幅运动、运镜和微表情，提升物理表现、风格一致性及画面美感
可灵 01 2025 年 12 月	基于 MVL 的统一多模态视频模型，核心系统包括 Prompt Enhancer、Omni-Generator 和 Multimodal Super-Resolution，将指令理解、视频生成和画面增强整合到统一系统中	文字、图片、视频、主体元素	支持 720P 和 1080P；可自由生成 3—10 秒；输入参考视频最高 2K	无	将参考生成和视频编辑统一到同一模型，支持主体替换、内容增删、风格修改和镜头延展，并强化人物与场景一致性
可灵 2.6 2025 年 12 月	可灵首个支持原生音频的视频生成模型；具体音视频联合建模结构未披露	文字、图片	单次最长 10 秒；原生音频模式支持 1080P，无原生音频模式支持 720P 和 1080P	支持原生音频，可同步生成语音、对白、旁白、歌唱、动作音效和环境音	强调音画同步、多人对白和口型同步；后续 2.6 Motion Control 进一步加入动作参考控制

可灵视频 3.0 2026 年 2 月	基于 All-in-One 产品框架和统一多模态训练框架，由 2.6 路线升级而来，进一步整合视频生成、原生音频、参考和叙事控制	文字、图片、多图或视频元素、声音特征	首发支持 720P 和 1080P，可自由生成 3—15 秒；后续升级支持原生 4K	支持原生音频、多语言、多口音及多角色不同语言对白	强化多镜头叙事、主体与元素一致性、画面文字保留和复杂动作表现
可灵视频 3.0 Omni 2026 年 2 月	01 路线的升级版本，基于统一多模态训练框架，进一步统一多模态参考、原生音频、视频编辑和分镜控制	文字、图片、视频、主体元素、声音特征	首发支持 720P 和 1080P，可自由生成 3—15 秒；后续升级支持原生 4K	支持原生音频；可将声音特征绑定至指定人物并跨镜头复用	支持全模态参考、自然语言编辑和自定义多镜头分镜；可逐镜头设置时长、景别、视角、叙事内容和运镜
可灵 3.0 Turbo 2026 年 6 月	3.0 系列效率版本，面向高并发和规模化生产，强调更低延迟、更高性价比和稳定输出；具体蒸馏及压缩方案未披露	文字、图片	支持 720P 和 1080P，可自由生成 3—15 秒	支持原生音频和音画同步	支持多镜头生成，侧重批量生产和规模化稳定性；官方未表明其具备 3.0 Omni 的完整统一编辑与全模态参考能力

来源：公司官网，国金证券研究所

四、风险提示

1）技术发展不及预期。视频生成模型仍处于技术尚未完全收敛时期，如若技术发展不及预期，或将影响整个行业发展进程，影响相关公司多模态模型相关业绩。

2）竞争加剧。视频生成模型行业参与玩家较多，且各家都在不断推出新模型，竞争加剧，或将影响行业格局和相关公司业绩。

3）商业化进程不及预期。视频生成模型在下游应用如若渗透率持续较低，AI 视频商业化进程不及预期，将影响模型收入。

行业投资评级的说明：

买入：预期未来 3—6 个月内该行业上涨幅度超过大盘在 15%以上；

增持：预期未来 3—6 个月内该行业上涨幅度超过大盘在 5%—15%；

中性：预期未来 3—6 个月内该行业变动幅度相对大盘在 -5%—5%；

减持：预期未来 3—6 个月内该行业下跌幅度超过大盘在 5%以上。

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