中邮证券
CHINA POST SECURITIES

证券研究报告

AI视频行业深度报告

技术跃迁驱动内容革命，把握产业变革新机遇

传 媒　行业评级：强大于市|维持

证券分析师　王晓萱

证书编号　S1340522080005

中邮证券

发布时间：2026-02-14

投资要点

- **视频生成：从GAN走向DiT，通往AGI的重要路径。**视频作为同时融合文本、图像、音频等多模态信息，并引入时间维度因果结构的内容形态，天然具备更高的复杂性与表达力，代表着AIGC产业能力上限。当前文本、图片、音乐等模态生成技术已相对成熟，视频仍是行业技术短板，其突破将对AIGC的产业应用前景起到重要作用。从技术演进看，AI视频生成技术自2010年代中后期逐步起步，经历了GAN、Transformer等多个架构的尝试，行业技术路线一度出现分歧。直至2022年，Diffusion与Transformer的融合思路逐步成型，叠加2024年OpenAI发布的Sora验证了DiT架构在视频生成中的可行性与效果，行业迎来关键转折点，主流厂商全面向DiT路径演进，视频生成自此进入快速发展阶段。
- **技术进展：短视频生成已近专业水准，世界模型或为长视频生成带来新变量。**当前AI视频已可根据文本提示直接生成包含多人物、动态动作与复杂背景的完整画面，Sora2、Veo3等音画一体化模型的出现进一步推动其从“画面生成工具”演进为“视听内容引擎”，短片段生成质量已接近专业制作水准。但现有架构在生成时长、物理合理性等维度仍存在结构限制，难以支撑更复杂的长视频构建，业界目前重点关注的世界模型可能是解决上述瓶颈的关键路径。世界模型最初研讨主要用于弥补语言模型在物理与因果建模方面的能力缺口，2025年前行业以“表征派”为主，主要聚焦环境感知与状态预测；2025年后，产业研究重心开始向“生成派”倾斜，Genie 3、Marble等代表性成果的推出标志着世界模型具备生成持续存在物体、模拟因果逻辑与动态环境的潜力，直接对应当前视频生成的技术短板。世界模型与现有视频模型技术路径存在差异，不受后者架构限制，且在空间一致性与物理逻辑等关键性能上展现出更快的迭代效率。行业亦已普遍认为视频生成是世界模型的雏形，后续在能力与技术演进上可能进一步重合。目前世界模型已被业内普遍视为与大语言模型同级的重要人工智能发展路径，相关参与者数量仍在持续增加，后续研发节奏预计或将进一步加快，2026年或为实现跃迁的关键节点。
- **商业化进展：C+B端双路并进，影视级项目有望迎来商业元年。**全球AI视频生成市场正加速扩张，预计2026年市场规模将达2.96亿美元，同比增长35.16%。行业商业路径主要分为C端平台与B端工具两类：1）C端：以订阅模式为主要收入来源，用户量是现阶段主要评判标准，Sora体量仍断档领先。目前行业亦在积极探索新商业模式， 例如OpenAI推出了社交化视频创作平台Sora app，未来C端有望进一步向广告、电商等新增路径拓展，并同时为B端内容传播带来新渠道；2）B端：API是当前主流业务模式，核心电商展示、广告等领域应用已基本成熟。“质量+效率+成本”是API核心评价维度，可灵、海螺、Vidu等部分国产模型已实现行业领先。但目前API模式主要应用于素材级生成，仍不具备提供完整影视级项目制作流能力。部分海外厂商已开始试水影视级AI解决方案，并初步在商业层面实现验证。以Utopai为例，其通过《Cortés》《Project Space》等项目累计实现收入约1.1亿美元。主流厂商亦在加快布局：OpenAI参与制作的AI影片《Critterz》预计于2026年上映；Runway设立旗下制作部门Runway Studios；灵AI亦于2025年亮相东京TIFFCOM内容交易市场。随着模型能力演进与工具链完善，2026年有望成为AI影视制作商业化的关键起点。

投资要点

- **传媒：AI视频核心应用场景，广告、影视、游戏均有望受益。**
  - **广告：**伴随用户侧信息获取方式向抖音、快手等短视频平台转移，推动广告形态由图文持续向视频迁移。2025年上半年全网移动广告中，视频类素材投放占比已超过65%，其中竖屏视频占比高达54.8%。竖屏广告主要为短视频广告，其多集中在6～15秒之间，契合现阶段视频生成模型的能力边界。目前AI在广告内容生成中的应用仍集中于内容草拟（70%）、文案创作（59%）等环节，视频创作渗透率（19%）明显偏低，仍然具备后发增长潜力。此外，AI视频工具的普及亦有望打破原有营销服务商的能力边界，使其从单一媒介投放职能，转型为能协助品牌进行内容策划、生成、测试与投放优化的全链路合作伙伴，提升其在营销生态中的战略价值。从资本市场反馈来看，2025全年，海外营销龙头Applovin股价累计涨幅108.08%，充分反映海外市场对AI+营销的价值认可，后续国内厂商有望持续跟进；
  - **影视：**AI漫剧与视频生成契合度最高，已率先实现商业闭环。拟真人短剧方面，据新华网统计，2025年1月抖音TOP5000短剧中仅4部为全AI生成，10月、11月分别增长至69部与217部，内容接受度在快速提升，后续或逐步进入量产阶段。长剧与电影层面，CG特效等高价值环节有望成为首批替代场景。但鉴于CG特效仍为影视工业中技术门槛最高模块之一，具备专业团队与预算的头部厂商短期内仍可能优先采用成熟CG方案。相较而言，中小型影视团队或更可能依托AI实现降本增效，率先受益于技术平权红利；
  - **游戏：**视频生成与3D生成的底层技术路径一致，均依赖扩散模型与Transformer等生成架构。当前3D生成在游戏建模领域已初步实现静态资产的自动化生产，整体进程正由“能力验证”向“实用落地”迈进。例如腾讯内部数十款游戏（如《元梦之星》）已接入混元3D能力，《蛋仔派对》亦与影眸科技合作，支持玩家通过AI生成游戏内物品，推动创作工具升级。后续世界模型等新技术落地或将进一步带动场景/动态资产的生成与应用，有望持续拓展AI在游戏中的应用深度。长期看，视频生成与交互融合亦有望为游戏内容演化带来新方向，当前AI原生交互已在多款文字类游戏中率先落地，未来若3D视频生成与行为驱动结合成熟，或催生具备实时互动能力的新游戏品类，重塑产业增长边界。
- **核心受益上市公司：**1）具备自研算法与模型能力，且具有多场景业务嵌合能力的技术型公司：昆仑万维；2）拥有海量内容资产与版权资源的影视内容提供商：中文在线、捷成股份、华策影视；3）积极布局AI营销、具备内容分发的整合型平台公司：易点天下；4）推动AI生成能力嵌入游戏资产生产流程的大型游戏公司：完美世界、巨人网络。
- **风险提示：**AI视频生成技术发展不及预期、产业应用不及预期、版权保护风险。

目录

1 视频生成的前世今生：从GAN走向DiT，通往AGI的重要路径

1.1 视频生成：融合多模态信息能力，决定AIGC技术上限

1.2 发展历程：从早期分化逐步走向共识，产业进入高速发展期

1.1 视频生成：融合多模态信息能力，决定AIGC技术上限

- **视频同时融合文本、图像、音频等多模态信息，天然具备更高的复杂性与表达力，代表着AIGC产业能力上限。**视频需处理空间、时间、因果与交互等高维结构，并要求将文字、图像、音频等模态映射到同一表征空间，其复杂性要求模型必须具备对真实世界的综合理解与推演能力：**1）空间：**视频需理解物体形状、位置关系、遮挡与深度等三维结构；**2）时间：**视频要求模型在连续帧中保持状态演化一致性，学习动力学规律与行为轨迹；**3）因果与交互：**视频呈现对象间的作用、反应与事件链条，迫使模型掌握因果机制和多实体交互规则。当前文本、图片、音乐等模态生成技术已相对成熟，视频仍是行业技术短板，其突破将对AIGC的产业应用前景起到决定性作用。

图表1：主流AI模态技术进展

模型能力	技术进展	发展历程	代表模型
文本	• 大语言模型在文字处理上面的卓越表现开启了生成式AI的浪潮，基础模型能够基于语言进行推理是智能的重要表现，在各个领域应用最为成熟。	• 2018年6月，由Alec Radford主导在0penAI推出GPT-12020年6月，0penAI推出GPT-3，引发业界关注，验证scaling路线 • 2022年11月，ChatGPT掀起技术浪潮	• ChatGPT • Character.AI • Gemini • Anthropic
图片	• 文生图领域产生了仅次于基础模型的杀手级应用，获得了大量创作者和用户关注，成熟度仅次于文本模态 • Midjhouney已有超过2000万用户，在无投资的情况自我造血，在2023年的营收超过2亿美元	• 2021年1月，0penAI发布初代文生图模型DALL-E • 2022年8月，Stable Diffusion在Stability.ai的支持下开源，推动社区在图像领域快速发展 • 2023年3月，Midjourney V5发布，迅速成为现象级应用	• Stable Diffusion • Midjourney • Dall-E 3
视频	• 视频是图像模态的进一步扩展，但由于技术复杂，对于算力、数据等资源要求较高，成熟相对文本、图像较慢 • 领军企业已经做出标杆，显著加速领域发展，已出现多家视频生成领域创业公司，**但商业化、产品化进展较慢**	• 2022年10月，Google、Meta发布Phenaki、Make-A-Video • 2023年，创业公司推出Runway-Gen2，Stable Video Diffusion、Pika等产品 • 2025年，openAI、Google先后发布sora 2、Veo3.1等 • 2026年2月，抖音发布Seedance 2.0	• Runway • Sora2 • 可灵 • Pixverse • Seedance
音频	• 目前主要是音乐生成，不如图片生成、视频生成等领域热门 • 创业公司较少，但**有加速的发展的态势**	• 2024年2月，Suno.ai发布Suno V3 • 2024年6月，Stability.AI推出文生音频模型Stable Audio 0pen	• Suno • Stable Audio
3D	• 当前技术路径可分为2D图像生成+3D重建、原生3D生成两类：前者前端图像生成沿用了目前图像/视频生成模型的扩散模型体系；后者思路是将原本DiT架构中的2D训练数据替换为3D数据。因此其技术路径均与视频模型同源，发展水平亦受到视频模型底层技术的影响	• 2020年8月，NeRF论文发表 • 2022年9月，谷歌发布DreamFusion • 2023年5月，OpenAI开源Shape-E模型 • 2024年7月，Meta发布Meta3DGen	• Luma.AI • Meshy

资料来源：量子位， MIT科技评论，极客公园，证券时报网，中邮证券研究所整理

1.2 发展历程：从早期分化逐步走向共识，产业进入高速发展期

- **AI视频生成技术自2010年代中后期逐步起步，经历了多个关键架构的迭代升级。**视频生成技术最早可追溯至20世纪90年代的图像序列拼接方法，其开启了将静态帧合成为动态视频的早期尝试，但真正的AI模型化探索始于2014年GAN的提出。2017年，Transformer架构的引入为模型带来了更强的时序建模与语义表达能力，但仍存在计算资源受限、生成质量不稳定等问题。因而在2020年后，部分开源社区尝试将扩散模型应用于视频生成，试图跳出Transformer架构限制，行业技术路线一度呈现分歧。直至2022年，Diffusion与Transformer的融合思路逐步成型，叠加2024年OpenAI发布的Sora验证了DiT架构在视频生成中的可行性与效果，行业迎来关键转折点，主流厂商全面向DiT路径演进，视频生成自此进入快速发展阶段。

图表2：视频生成技术历程概览

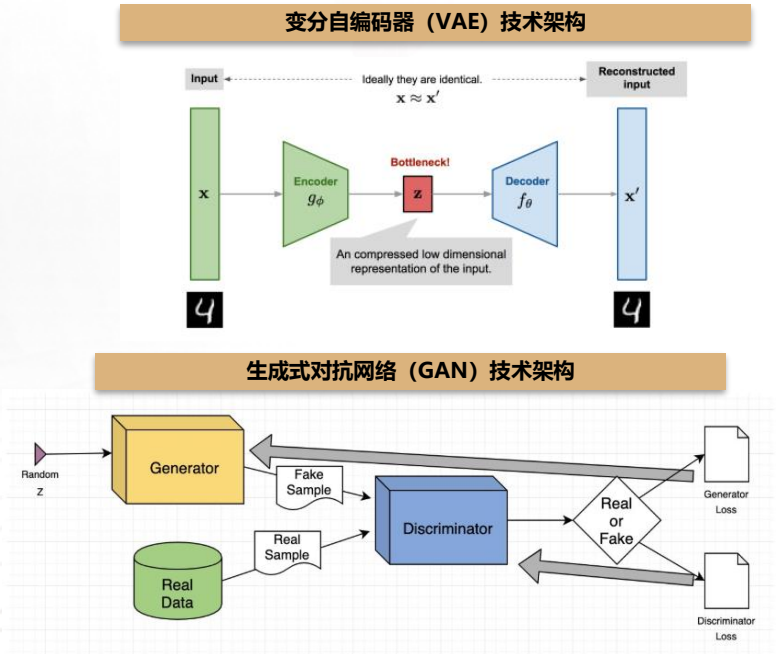
	GAN	Transformer	Diffusion Model	Diffusion + Transformer
技术节点	• 2014年GAN发表，2016年左右开始用于视频生成 • 2013年VAE出现，可以通过编码器和解码器的机制学习数据的潜在表示	• 2017年Transformer发表，逐步向各个领域渗透，早期在视频生成领域也有尝试 • 2021年Google发布VideoVisionn Transformer	• 2020年DDPM论文发表 • 2021年 Latent DiffusionModels论文发表 • 2022年 Video DiffusionModel论文发表	• 2022年12月Diffusion Transformer论文发表 • 2024年2月OpenAI发布Sora，取得惊艳的生成效果，验证了Diffusion和Transformer结合的有效性
案例	• Text2Filter、 TGANs-C, VGAN、TGAN、VideoGPT、MoCoGAN、DVD-GAN、DIGAN • 应用范围有限(只能总特定生成，例如数字)，生成分辨率低、生成时长极短	• G0DIVA、VideoGPT、Phenaki、Cog Video、 NUWA • 早期基于Transformer的视频生成效果有限，在当时的成本比较高增加平方级增加，复杂度高	• Stable Video Diffusion、Make-A-Video、RunwayGen1/2、Video LDM • **由于SD的开源激发社区创新，推动扩散模型在视觉生成领域占据主导地位**	• Sora、 Videopoet、 W. A.L. T. • 由于领军企业的示范效应，Diffusion Transformer已经成为视频生成领域的核心路线，领域内其他玩家都开始聚焦DiT并取得进展目前业界对于Sora的架构大多是猜想，从架构大方向上接近，但具体细节上不同模型、产品各异
特点	• **生成多样性差:**生成器-判别器对抗的模型架构决定生成效果逼真但多样性、创造性差，很难泛化生成速度快:GAN可以一次出图，无需做多步推理	• **可扩展性强:**Scale up路线上目前最好的架构 • **生成速度慢、开销大:**需要把图片进行分割再转换成超长序列，计算量随分辨率增加平方级增加，复杂度高	• **生成质量高:**生成效果好，细节丰富清晰，训练要远比GAN稳定 • **显存开销大:**视频生成有大量相关依赖关系的图片同时生成，对计算显存有极高要求	• **可扩展性强:**基于扩散模型的视频模型，核心的构件是卷积神经网络U-Net，随着模型规模scale up增益会放缓甚至消失。DiT(Diffusion Transformer)将传统扩散模型的U-Net替换为Transformer，从而能够实现更强的可拓展性 • **长期来看语言模型路线可成为主要路线，**可以将多种模态融合到一起(由于架构的成熟和优良的扩展性)

资料来源：量子位，中邮证券研究所

1.2 发展历程：从早期分化逐步走向共识，产业进入高速发展期

- **GAN–VAE阶段（2014-2016）：确立“视频可被端到端生成”的技术方向，是后续技术跃迁的理论起点。**视频生成技术最早可追溯至2016年UC Berkeley提出的VGAN，该模型首次将生成式对抗网络（GAN）引入视频生成任务，并通过空间–时间卷积结构实现低分辨率短时动态序列的合成。同年，京都大学与东京大学提出的TGAN将视频生成分解为时间潜变量序列与图像生成器协同工作的方式，实现捕捉跨帧运动信息。在此基础上，2018年NVIDIA团队提出MoCoGAN，将视频内容与运动显式解耦，分别建模并通过对抗学习生成一致动作序列，从而实现了更具可控性的基础视频生成框架。但该阶段的模型多基于GAN的对抗式重建能力+VAE的连续潜空间表达，受限于模型架构限制，应用范围仅限于简单场景（如数字、基础动作），生成分辨率与时长均较低。

图表3：GAN/VAE技术架构



资料来源：ALU，AWS，中邮证券研究所

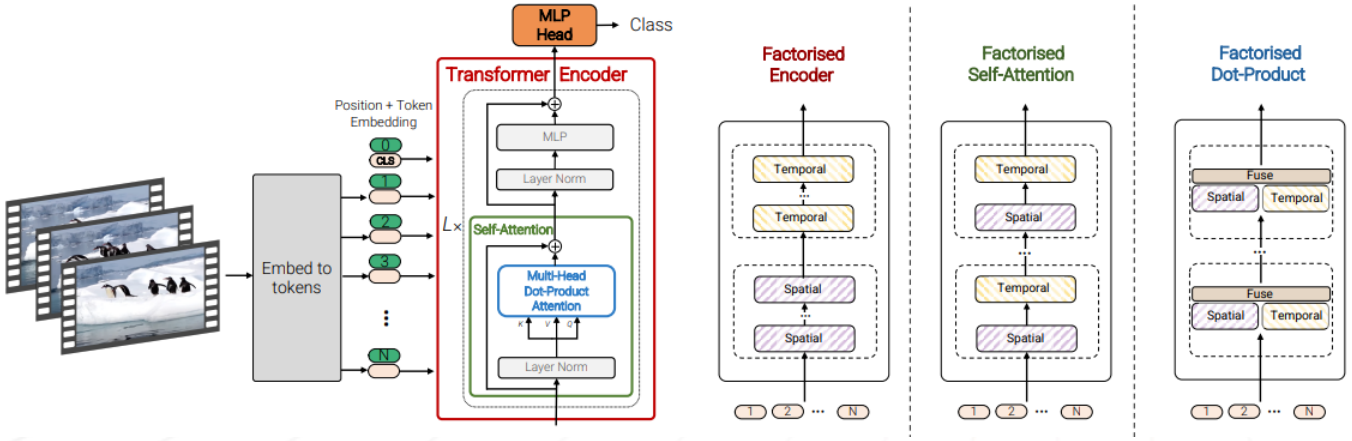
图表4： GAN-VAE阶段生成视频仅能完成基础动作展示，且分辨率与时长均较低

资料来源：github，中邮证券研究所（注：展示效果图为MoCoGAN 生成）

1.2 发展历程：从早期分化逐步走向共识，产业进入高速发展期

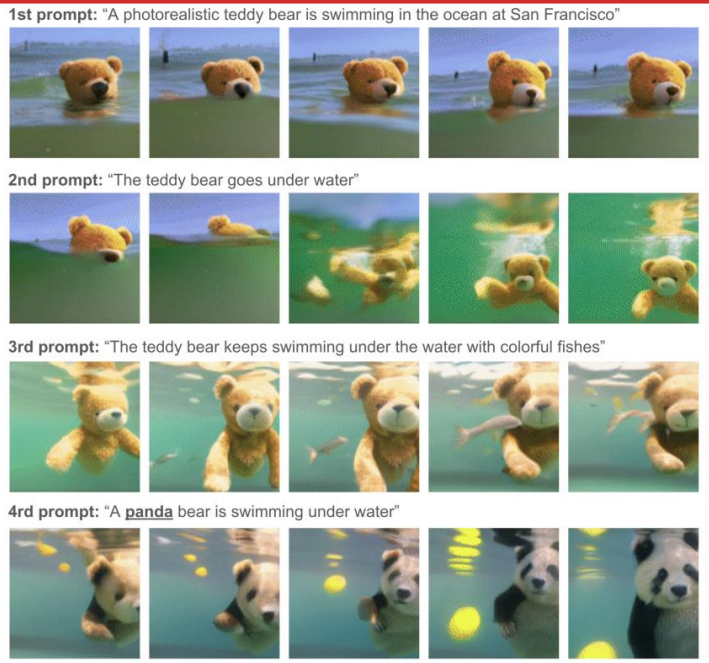
- **Transformer表征阶段（2017–2021）：时空表征能力显著提升，为视频生成真正可用奠定底层基础，但生成质量、成本化能力均属过渡期。**2017年Transformer论文发表后，该架构快速渗透至各类序列建模场景，并在视频生成任务中开启探索。自2021年Google推出Video Vision Transformer（ViViT）起，GODIVA、VideoGPT、Phenaki、CogVideo、NUWA 等视频模型相继出现。相较于GAN系列，Transformer具备明确的概率密度建模能力、收敛过程更稳定，并能够有效捕捉跨帧长程依赖，在生成时序一致、衔接自然的动态内容上更具优势。但由于其计算复杂度随空间与时间token数呈平方级增长，分辨率与时长提升将带来指数级的算力压力，导致该阶段模型在生成效果上仍受限制，其产业价值主要体现在从“能生成”迈向“能理解再生成”。

图表5： 基于Transformer 的ViViT四类技术架构



资料来源：《ViViT: A Video Vision Transformer》，AWS，中邮证券研究所

图表6： Transformer 类模型已能根据提示词生成对应视频

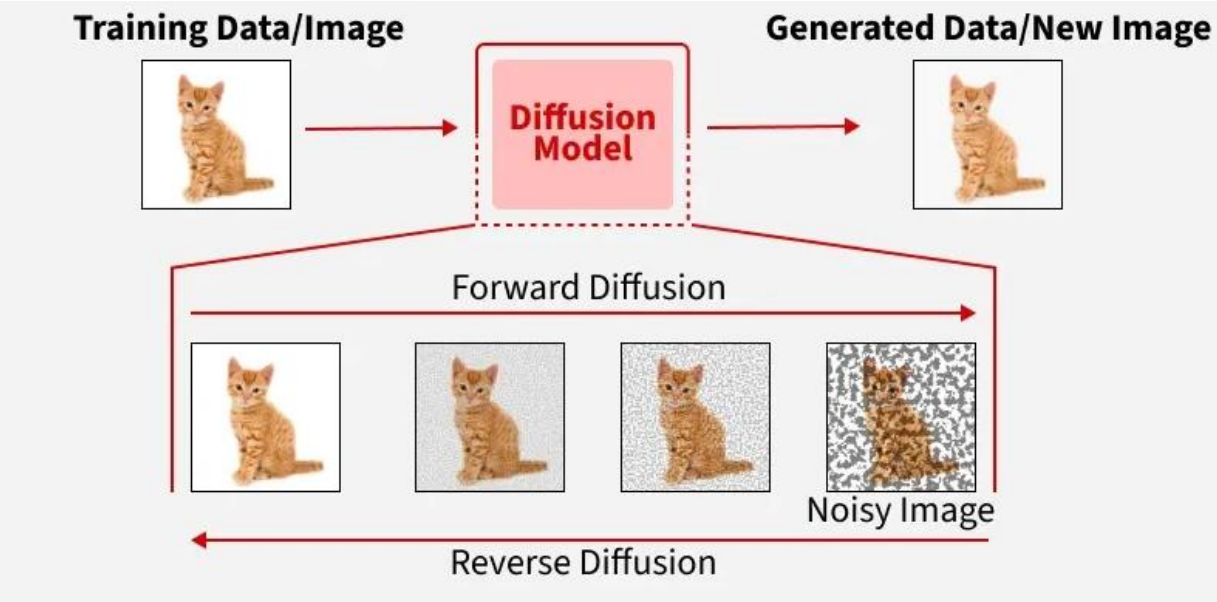


资料来源：《PHENAKI: VARIABLE LENGTH VIDEO GENERATION FROM OPEN DOMAIN TEXTUAL DESCRIPTIONS》，中邮证券研究所（注：展示效果图为Phenaki 生成）

1.2 发展历程：从早期分化逐步走向共识，产业进入高速发展期

- **Diffusion扩散模型阶段（2020–2023）：实现高质量短视频生成，但受限于时长与物理一致性，存在技术上限。**扩散模型（Diffusion）通过“逐步加噪—逆向去噪”的显式概率建模范式，解决了GAN在训练稳定性和可控性上的核心缺陷，为高质量视觉生成奠定了基础。2022年，Meta发布Make-A-Video，其可根据自然语言生成约5秒短视频，是推动视频生成技术进入商业化探索阶段的早期代表之一。但传统扩散模型的去噪网络基于U-Net，其本质是一种以局部卷积为主的二维图像编码器，只能在空间维度内进行局部感受野建模，缺乏对时间维度的统一表征，也无法捕捉跨帧的长程依赖、物体状态延续与物理一致性。基于此结构的视频扩散模型，误差会沿时间轴不断累积，导致跨帧漂移、运动不连续，使视频生成在时长与整体一致性上存在上限。

图表7： 扩散模型的前向扩噪/逆向去噪过程展示



资料来源：GeeksforGeeks，中邮证券研究所

图表8： Make-A-Video 生成视频效果展示

资料来源：Meta，中邮证券研究所

1.2 发展历程：从早期分化逐步走向共识，产业进入高速发展期

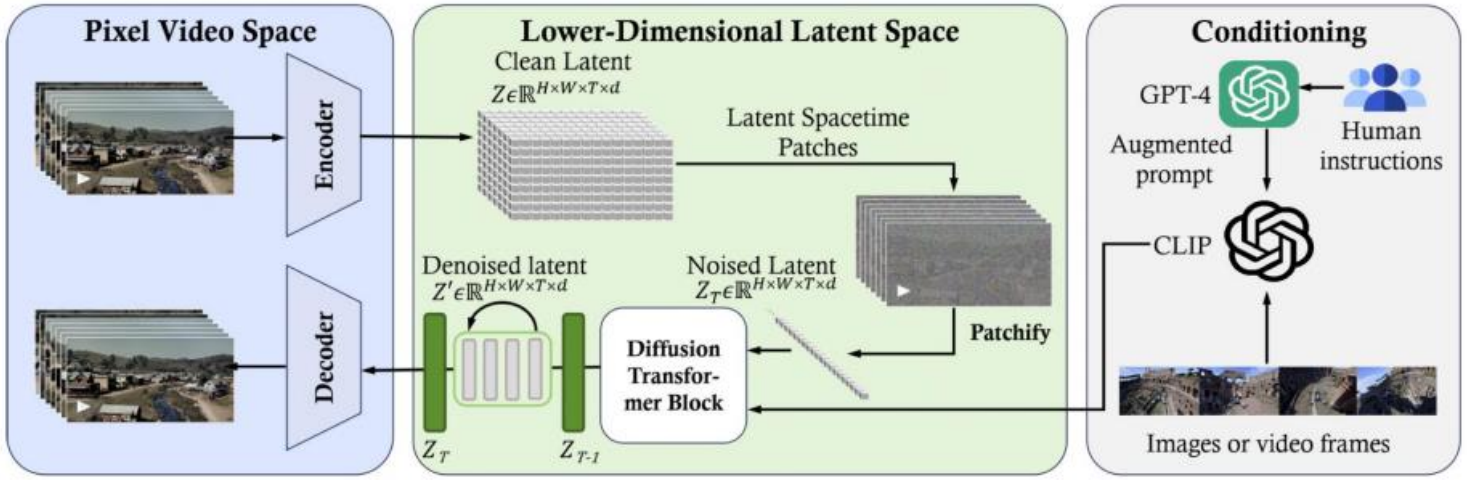
- **DiT扩散模型阶段（2024至今）：在Sora推动下进入商业化周期，形成视频生成的主导技术路线。**DiT的核心思想是以Transformer结构取代传统扩散模型中的U-Net作为去噪网络。2024年2月，OpenAI发布Sora，首次在工业级规模上验证了Diffusion+Transformer结合的有效性：在更长时长、更高分辨率、更复杂场景物理一致性以及更强的帧间连贯性上实现突破。

图表9： Sora技术架构

负责将高维度的视频像素序列转换为可建模的低维潜在表示，并在生成结束后将潜在表示解码回复像素视频。

通过将视频表示切分为时空 Patch，并利用 **Diffusion Transformer Block** 执行逐步去噪，实现对视频在空间与时间维度上的全局建模

为生成过程提供语义指导，通过 CLIP、GPT-4 以及人工指令构成多模态的条件信号



输入视频首先通过Encoder映射到紧凑的潜在空间，为后续的时空建模提供结构化表征；模型完成生成与去噪后，再通过 Decoder将干净的潜在表示还原为真实可播放的视频帧

模型从加入噪声的潜在表示出发，通过 Transformer 捕捉长程依赖与物理一致性，逐步生成干净的潜在视频。相比传统使用 U-Net 的扩散模型，基于 Transformer 的去噪网络具有更强的可扩展性和时空关系建模能力，使 Sora 能够稳定生成更长时长、更高分辨率、逻辑一致性更强的视频

1）CLIP 将用户输入的图像或视频帧编码为视觉特征；2）GPT-4 对文本提示词进行语义扩展和增强，使模型能够理解更细致的场景设定、动作逻辑与风格要求；3）人工反馈则用于训练与强化模型对复杂指令的遵循能力

资料来源：《Sora: A Review on Background, Technology, Limitations, and Opportunities of Large Vision Models》，中邮证券研究所

图表10： Sora生成图片展示及技术细节

能力维度	解读
分辨率与时长	支持 1080p 高清长时长视频，此前模型多为1080p以下水平；单段可生成最长60秒的视频。
构图与取景	相比以往需要裁成方形的视频训练方式，Sora 出现主体被截断或局部错位的问题得到优化，整体画面更自然、主体更完整。
语言理解	利用 GPT 生成更详细的 prompt，提升模型对指令的理解度；从而使生成内容更准确贴合用户意图，文本到视频的语义对齐显著增强。
多镜头叙事及场景复杂度	能连续生成多镜头、多角度的影片段落，同时能够生成包含多个角色、特定动作类型以及精确的主体和背景细节的复杂场景。

资料来源：OpenAI，中邮证券研究所

1.2 发展历程：从早期分化逐步走向共识，产业进入高速发展期

- **DiT架构融合了diffusion与Transformer的双重能力，推动视频技术进入高速迭代期。**DiT 在继承扩散模型生成稳定性、训练可控性等基础优势的同时，引入了Transformer的推理能力、长程依赖建模能力与多模态统一表示能力，具体来看：

① **DiT遵循Scaling Law，画面表现等能力提升形成外推性。**传统扩散视频模型受限于卷积结构的局部建模特性，技术进展呈现不连续的断点式突破。而DiT架构通过引入Transformer，使视频模型能够遵循Scaling Law，生成能力可随参数规模、数据体量与训练算力提升而持续增强。近两年主流厂商在此基础上持续扩大模型规模并优化训练策略，使视频生成在分辨率、细节刻画及光影一致性等方面较早期模型显著改善；

② **融合Transformer推理能力，复杂性与叙事表达持续提升。**Transformer架构强化了对长程依赖与因果关系的建模能力，使视频模型能够更高效地融合大语言模型的推理能力，在画面复杂性与叙事结构表达层面获得增强；

③ **多模态融合能力增强，实现音画一体化发展。**基于Transformer的统一token表达与自注意力机制，文本、图像、视频与音频等多模态信息可在同一语义空间内对齐与协同生成，推动视频生成技术由早期“无声视频生成”逐步向“音画一体化生成”演进。

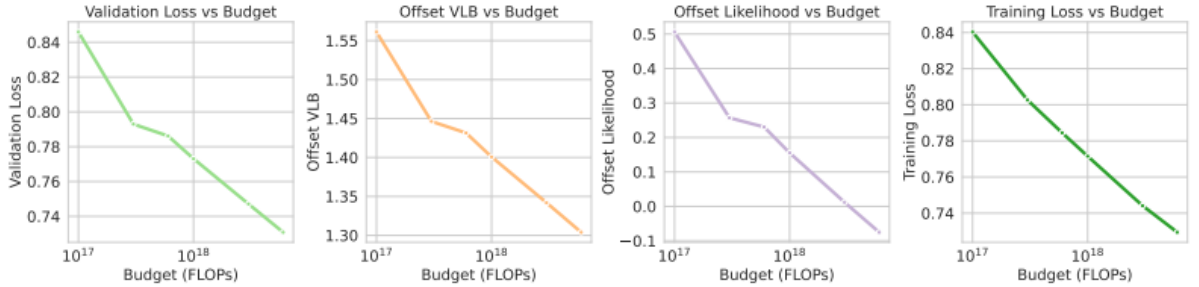
图表11： DiT架构亦遵循大模型scaling law规则

随着训练算力预算增加，验证集损失呈稳定下降趋势，模型泛化能力随规模扩展持续提升

算力规模提升使扩散模型的变分下界持续改善，模型对真实数据分布的拟合能力不断增强

模型对真实样本的对数似然随算力增长而提升，表明生成质量在概率意义上持续优化。

训练损失随算力预算呈幂律下降，验证了DiT模型的的Scaling Law特征



资料来源：《SCALING LAWS FOR DIFFUSION TRANSFORMERS》，中邮证券研究所

图表12： 视频生成技术的核心评价维度

评测维度	核心关注点
美学质量	单帧与空间层面的视觉表现
物理规律遵循	人物、物体及环境的运动是否符合物理规律。
时空一致性	背景、人物、物体在不同帧间的连续性，是否出现突然的消失或跳变。
指令遵循	生成内容与文本指令的匹配程度。
创造性/ 多样性	元素、风格、运镜手法等的丰富性。

资料来源：久谦咨询，中邮证券研究所

1.2 发展历程：从早期分化逐步走向共识，产业进入高速发展期

- **目前业内主流视频厂商模型均已向DiT架构收敛。**Sora 发布之后，字节、Google、腾讯等主流厂商以及各类开源项目亦在向DiT框架迁移。尽管各家主干架构技术仍有差异，但路线本质上均是在DiT架构内的技术演进。

图表13： 主流视频模型技术架构概览（部分）

模型	主干架构	文本塔	视频塔	位置编码	分辨率
Seedance（字节）	MM-DiT	Qwen2.5-14B	VAE	3D RoPE + MM-RoPE	720p, 1080p
混元-Avatar（腾讯）	MM-DiT	LLaVA	Two Hunyuan 3D VAE	3D RoPE	704p, 1216p
MAGI-1	DiT	T5	Transformer-based VAE	3D RoPE	720p
混元-Custom（腾讯）	Hunyuan-MM-DiT	LLaVA	Two Hunyuan 3D VAE	3D RoPE	512p, 720p
Veo3（google）	DiT	–	–	–	1080p
SkyReels-v2	Wan-DiT	umT5	Wan VAE	Learnable Frequency Embeddings	256p, 360p, 540p, 720p
Open-Sora 2.0	Flux (MM-DiT)	T5-XXL, CLIP-Large	HunyuanVideo 3D VAE、Deep Compression Autoencoder	3D RoPE	256p, 768p
WAN2.1	DiT + Cross-attn	umT5, Qwen2-VL	Wan-VAE	Standard Sinusoidal Spatialpositional Encodings	480p, 720p
VACE	Wan-T2V-14B, LTX-Video-2B	Inherited	Inherited	Inherited	480p, 720p
Phantom	MMDiT	T5 Dinov2(Ref.Img)	(CLIP VAE), (Qwen2.5, 3D VAE)	3DRoPE	480p, 720p
StepVideo	DiT	Hunyuan-CLIP, Step-LLM	Video-VAE	3DRoPE	544p
ConceptMaster	Transformer-based latent diffusion	T5；CLIP	3D VAE	3D Self-attention	–
VideoAlchemist	DiT	DiT Text Encoder, CLIP, Arcface	CogVideoX-5B VAE、DiT Tokenizer 、CLIP VIT-L/14、DINOv2 VIT-L/14	RoPE	256p
混元（腾讯）	Flux (MM-DiT)	Hunyuan MLLM Decoder + CLIP	3D VAE	3D RoPE	720p
LTX-video	DiT + Cross-attn	DiT Text Encoder	Video-VAE	RoPE	512p
MovieGen（Meta）	LLaMa3 Design	U2I, ByT5, Long-prompt MetaCLIP	TAE + VAE	Factored	256p, 1080p
PyramidFlow	MM-DiT	–	PyramidStages Autoregressive TemporalPyramid	–	768p
Sora（OpenAI）	DiT	–	–	–	480p, 1080p

资料来源：《A Survey on Long-Video Storytelling Generation: Architectures, Consistency, and Cinematic Quality》，中邮证券研究所

2 技术进展：短视频生成已近专业水准，长视频或迎重要变革节点

2.1 技术进展：美学、多模态化能力已近专业水准，物理性、生成时长是主要瓶颈

2.2 技术趋势：世界模型或将带来新技术变革，视频生成有望再迎发展拐点

2.1 技术进展：美学、多模态化能力表现优异，物理性、生成时长是主要瓶颈

1）美学质量方面：当前AI视频生成模型已能够根据提示直接生成包含多人物主体、动作、背景与光影的完整动态画面，短片段生成能力已接近专业影视制作水准

- 主流视频生成模型普遍已支持1080p及以上分辨率，部分模型可生成4K及以上画面；帧率方面，多数模型可稳定支持24fps，部分已提升至30fps。以Sora 2为例，其在拟真性、风格表达以及复杂场景生成方面已取得显著进展，整体水平以基本满足影视与商业内容制作需求：**1）真实性方面**，人物表情与动作连续性提升，同时在光影关系、纹理细节与景深层次等环境维度表现愈发成熟，整体画面真实感明显，已接近工业级CG制作能力；**2）风格层面**，模型可覆盖写实、动漫等多种视觉风格，能够适配悬疑、科幻等不同题材与叙事氛围的创作需求；**3）复杂人物主体及多镜头连续叙事能力方面，**模型已能够在同一场景中生成多人物、多动作的协同表现，并支持多镜头角度切换下的连续叙事。

图表14： sora 2在人物刻画方面已能以假乱真

Sora 2生成的奥特曼：手机屏幕与机身在光源作用下形成清晰且方向一致的反光效果，与环境光照逻辑相符；人物面部表情细微变化自然，眼神、眉部与口部状态协调，情绪表达与肢体动作保持一致，整体画面在真实感方面已十分逼近真实拍摄影像。

Sora 2生成的女性角色：眼周细纹、面部肤质纹理及轻微色斑等非理想化特征刻画自然，妆容层次与皮肤反射关系清晰，呈现出不同区域在光影作用下的不同表现，具备真实感。

资料来源：量子位公众号，中邮证券研究所

图表15： 主流视频生成模型分辨率及帧率

模型名称	分辨率	帧率（fps）
Google Veo 3	1080P	24
MiniMax Hailuo 02	1080P	24
快手 可灵 2.1	1080P	24
字节 Seedance 1.0	1080P	24
阿里 Wan 2.2	1080P	30
Runway Gen-4	4K	24
爱诗科技 PixVerse V5	1080P	24
生数科技 Vidu Q1	1080P	24
OpenAI Sora Turbo	1080P	30

资料来源：久谦咨询，中邮证券研究所

图表16： Sora 2生成能力概览

支持风格种类多样化

提示词：以吉卜力工作室动漫风格呈现，一个男孩和他的狗奔跑在绿草如茵的山丘上，天空云卷云舒，远景中是宁静的村庄，风景如画。

细致还原实际水纹、光影及场景

提示词：一名男子从跳水板上跳下，完成团身入水

多人物主体的复杂画面生成

提示词：一群人正在打排球

多镜头连贯叙事

提示词：维京人出征作战——北海启航（10.0秒，冬日清冷的日光 / 早期中世纪）...

资料来源：CG世界公众号，中邮证券研究所

2.1 技术进展：美学、多模态化能力表现优异，物理性、生成时长是主要瓶颈

2）多模态方面：从“无声”向“视听”阶段全面演进，路径收敛或将推动技术加速迭代

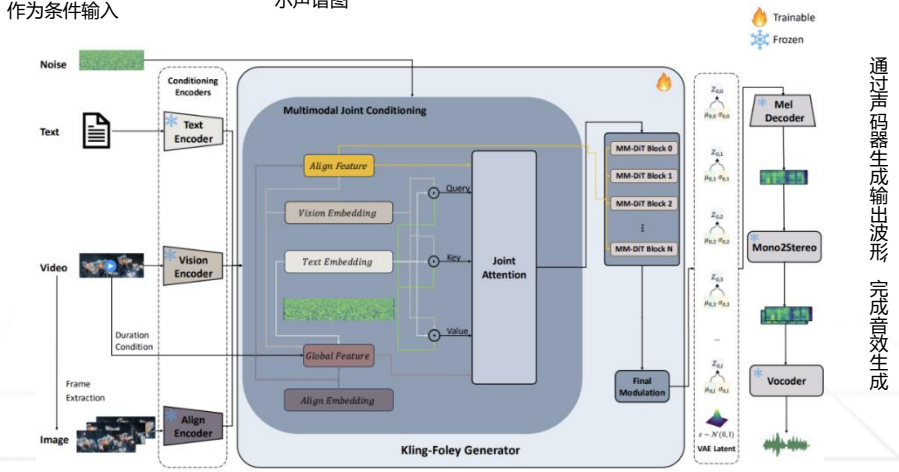
- **AI视频音效生成技术主要分为一体化生成和后期分离生成两类技术路径。**1）原生音视频一体生成：采用多模态联合训练架构，在视频合成过程中同步生成高保真音频流，一步到位生成带有音效的视频；2）后期分离式生成：采用解耦式的跨模态推理框架，将音频生成剥离为一道独立工序。该类模型通过对视频帧序列进行时序特征提取与事件识别，驱动合成物理属性匹配、情感语义一致的音效轨迹。
- **从技术路径看，一体化音画生成在技术原理上具备天然优势，但实现门槛较高；分离式方案则因更强的可行性，长期占据行业主流。**一体化路径将音效直接嵌入视频生成的底层流程，在统一时间轴与语义空间内完成联合建模，因而能够实现物理事件与声音的高精度对齐，相比分离式具备先天技术优势。但由于一体化生成壁垒较高，行业早期较多侧重分离式研究。典型产品包括2024年Pika、Google推出的Sound Effects与V2A系统，以及2025年国内厂商可灵、腾讯发布的Kling-Foley、HunyuanVideo-Foley 等。但**严格意义上讲，由于分离式音频生成并未纳入视频生成的统一建模过程，本质仍是独立音频模块，并不代表视频模型本身已具备了多模态生成能力。**

图表17：两类AI视频音效生成技术对比

	一体化生成	后期分离生成
表现能力	真实感和声音丰富度等关键指标上，**整体表现更稳定**	**容易出现音画错位或声音表达不规范等问题**
适用场景	极致的影视级沉浸式视听，尤其是对音画高度协同有较高要求的项目	预算有限、批量内容生产或对音效要求不高的使用场景
技术难度	高	低
算力成本	高	低
同步能力	精准同步	灵活同步

资料来源：302.AI，中邮证券研究所整理（注：对比基准模型为Veo 3 Pro及Kling音频模型）

图表18： 分离式音频生成架构演示（以Kling-Foley为例）



资料来源：机器之心公众号，中邮证券研究所

图表19：Pika-Sound Effects功能展示

资料来源：量子位，中邮证券研究所

2）多模态方面：从“无声”向“视听”阶段全面演进，路径收敛或将推动技术加速迭代

- **Google Veo3是首个实现商业化落地的音视同步生成模型，真正意义代表了视频生成技术从“画面生成工具”向“视听内容生成引擎“的形态跃迁。**2025年5月，Google 发布第三代视频生成模型 Veo 3。该模型可根据文本提示生成高质量视频内容，并在生成过程中同步输出与画面高度一致的对白、唇动对齐语音、拟真环境音效及情绪氛围音轨。
  - **从能力维度看，Veo3音频生成能力生成音效种类多样，已能够覆盖较为广泛的视频场景需求。**根据302.AI的对比测评，其在赛车音效、宴会场景、ASMR及史诗级战争等多维视频场景中，对Veo 3 Pro与分离式架构产品Kling Video-to-Audio进行了系统测试。结果显示，尽管Veo 3在个别场景下仍存在跳音或局部音频缺失等问题，但相较Kling Video-to-Audio，其在大多数场景中已能够实现更为完整的音画覆盖，整体适配能力与生成稳定性表现更优；
  - **从商业化维度看，一体生成在效率与使用门槛方面具备天然优势，推动Veo 3用户渗透率实现快速提升。**一体化生成将原本相互独立的视听生成工作流在统一模型体系内完成融合，从而简化了内容生产流程，对缺乏专业后期能力的C端用户尤为友好。根据Google介绍，Veo 3在发布后两个月内累计生成视频超过7,000万条，其B端版本亦在上线一个月内生成约600万条视频。

图表20： Veo 3 Pro VS Kling Video-to-Audio音频生成效果评测

场景	赛车音效		宴会		ASMR触发音		史诗级战争场面		音乐表演	
对应视频截图										
	Veo 3-Pro	Kling-Video-to-Audio	Veo 3-Pro	Kling-Video-to-Audio	Veo 3-Pro	Kling-Video-to-Audio	Veo 3-Pro	Kling-Video-to-Audio	Veo 3-Pro	Kling-Video-to-Audio
简评	除了引擎声，还有一脚刹车制动的音效，但缺乏风噪声，实现音画同步	音频不通透，听感平淡，没有体现出引擎的高频轰鸣。除引擎声外，缺乏其他音效，不够丰富	在兼顾画面主体人声、碰杯声的处理时，也伴有嘈杂的背景环境音	生成了接近口型但却不符合规范表达的人声	Kling的表现优于Veo 3，生成了蜂蜜倒在面包上的声音。虽然Kling生成的音效不如Veo 3饱满有力，但也一定程度上避免了失真		Veo3虽然生成了兵器对抗、人物嘶吼、战马奔腾、沙石滚动的音效，但是有几帧打斗画面也明显出现了跳音和空白音效的问题	Kling的音效漏帧更为严重，打斗场面几乎只出现了一次兵器碰撞的声音，层次单调，缺乏真实感	Veo 3生成的曲目音效相对来说层次更丰富，听感更接近真实。但两个模型都出现了同样的问题，扫弦的手看似看似对应上了节拍，而按弦的手和传出的曲调其实并不匹配，细节上的音画同步还是较难准确实现	
音效真实度	4	3	4	1	4	4	3	2	5	4
音效丰富度	4	3	4	2	2	5	4	2	4	3
音画匹配度	4	3	4	4	5	5	3	1	3	2

资料来源：302.AI公众号，中邮证券研究所

2）多模态方面：从“无声”向“视听”阶段全面演进，路径收敛或将推动技术加速迭代

- **在Veo3示范效应带动下，主流厂商均在加快其音视同步生成模型的落地，发布节奏明显趋于密集；同时伴随路径共识形成，或将推动后续音视生成模型技术的迭代进程进一步提速。**Veo3 商业层面的成功提升了行业对音画一体生成路径的重视度，主流模型厂商均在先后推出相关产品，可灵等部分此前以分离式方案为主的厂商亦开始转向一体化生成技术。尽管当前一体化生成在高复杂度场景下仍存在不足，但在技术路线形成共识背景下，后续各厂商对音画一体能力的研发与工程化投入有望持续提升，并推动该方向进入更快的能力迭代阶段。

图表21：主要音画同步生成模型概览（部分）

模型名	所属公司	发布时间	主要音频功能
Seedance 1.5 pro	字节跳动	2025/12	支持环境音、背景音乐、人声等多种元素，实现了毫秒级的音画同步输出。在对白处理上，模型支持多人多语言对话，口型对齐精准，覆盖中文方言（如四川话、粤语等）、英文及小语种。
可灵2.6	快手	2025/12	适用于单人独白（商品展示 / 生活 Vlog / 新闻播报 / 演讲表达）、旁白解说（商品讲解 / 赛事解说 / 纪录片 / 故事叙述）、多人对白（访谈节目 / 短剧等）、音乐表演（唱歌 / 说唱表演 / 多人合唱 / 乐器演奏）等场景。
PixVerse V5.5	爱诗科技	2025/12	人物对白、环境声和背景音乐的生成等
Runway Gen 4.5	Runway	2025/12	原生对话、背景音乐等，用户还可以编辑现有音频并添加对话。
万相2.6	阿里巴巴	2025/12（首个支持音视频版本为同年9月万相2.5）	支持音画同步、声音驱动等功能
VEO 3.1	Google	2025/10（首个支持音视频版本为同年5月发布的VEO 3）	全链路音频生成：支持“Ingredients to Video” “Frames to Video” “Extend”等功能的同步音效生成
Sora2	OpenAI	2025/9	背景音景、人声及音效等。

资料来源：新华网，凤凰网，新浪财经，新浪科技，IT之家，OpenAI，TechCrush，中邮证券研究所

2.1 技术进展：美学、多模态化能力表现优异，物理性、生成时长是主要瓶颈

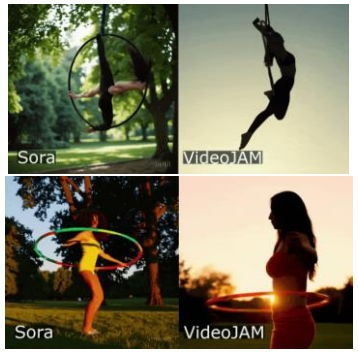
3）物理能力方面：“效率-能力”间的权衡，双路径协同并进

目前在提升物理能力方面主要可以概括为两种路径：1）隐式物理学习；2）显式物理约束：

- **隐式物理学习：以复杂任务能力迭代快为主要优势，但稳定性存在局限，更适用于强调效率与表现力的C端应用场景。**其基本机制是基于大规模视频数据，学习不同物体状态在时间维度上的联合概率分布，从而在统计意义上复现真实世界中的运动模式。在Scaling Law推动下，随着数据规模、模型容量与算力增长，高复杂度动作的生成能力能够快速提升。但由于路径本质仍是对概率分布的外推，模型并不具备对物理因果关系的显式建模能力，生成偏差仍难以避免。因此，该路径更适合以创作效率、多样性和视觉表现力为核心诉求的C端内容生成场景。
- **目前Sora 2尚未对其技术路径进行公开，但根据业内判断，其大概率仍沿用“生成模型+时空建模”的隐式物理学习范式。**根据OpenAI官网信息显示，Sora2模型物理表现能力的提升主要依赖于大规模视频数据的预训练与后训练体系优化。生成效果亦可予以侧面佐证：相较一代模型，其在高复杂度动作还原能力（如复现奥赛季体操运动）上实现显著提升，但在诸如灭火器喷口位置等个别基础物理或事实逻辑层面仍会出现偏差。

图表22： sora 2在复杂运动的物理还原能力提升明显

Sora1（左）在生成简单体操运动时仍不足以满足基本的物理逻辑



Sora2能够完成奥赛季的体操完整动作还原

资料来源：新智元公众号，CG世界公众号，中邮证券研究所

图表23： sora 2在基础物理或事实逻辑层面仍会出现偏差

灭火器出口位置错误

凭空生造文字

资料来源：第一财经公众号，中邮证券研究所

2.1 技术进展：美学、多模态化能力表现优异，物理性、生成时长是主要瓶颈

- **显式物理约束：物理正确性上限更高，但工程成本与生成自由度受限，更偏向B端专业与垂直场景。**其核心思想是在生成过程中引入物理先验、规则引导等结构性约束，从而降低违背因果规律的生成结果。但由于现实物理规则高度复杂且层级繁多，使得该路径在算力成本、系统复杂度等方面面临挑战。此外，显式约束在提升稳定性的同时可能降低结果的发散性与创造性，因此行业当前更多采用的是弱显式物理的工程化折中方案：即在不破坏生成模型主体灵活性的前提下，局部引入物理相关数据或约束信号。以Runway为例，根据AITurbo介绍，其产品Gen-4通过集成物理感知模拟层，使其能对走路姿态、物体重量等多种物理现象实现深刻理解。但由于折中方案仍牺牲了部分物理正确性，模型在因果推理与物体恒存性方面表现仍有不足。尽管如此，从综合性能来看，得益于更为出色的物理模拟能力等优势， Gen-4.5仍能以1243的Elo评分在Artificial Analysis文生视频基准测试中位居首位。

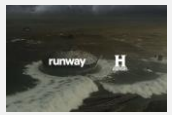
图表24： Runway Gen 4.5生成图像展示

真实还原汽车疾驰下的尘土飞扬，并展现出真实颠簸感

能够还原粉刷墙壁后，涂刷区域随时间推进逐渐干燥的真实物理过程。

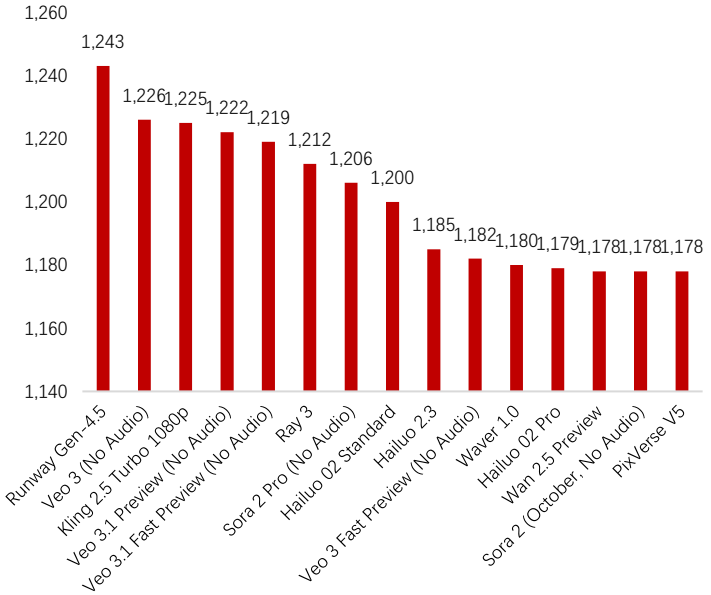
资料来源：量子位公众号，中邮证券研究所

图表25： Runway已参与众多工程级影音项目

领域	项目名	图示	时长	简介及Runway贡献
电视剧	House of David		432分钟/共八集	一部以传奇人物大卫为核心的历史史诗，基于真实事件改编。影片上线前17天观众人数突破2200万，并登顶亚马逊 Prime 播放榜。借助 Runway 的工具支持，项目后期制作周期缩短约5个月。
纪录片	Life After People		360分钟/共八集	该剧主要探讨当人们消失后世界会发生什么，实地拍摄并使用传统的视觉特效和剪辑方法将耗资巨大且耗时，曾因预算不足而被搁置。借助 Runway 的技术支持，项目后期制作周期缩短约1个月，单个场景的合成工作量减少约4小时。
广告	安德玛首只智利本土广告		-	该广告片以智利国家冰球队为主角，需要呈现更衣室氛围、赛场对抗及摔倒等动作画面，并对火焰效果提出视觉要求。Runway 在后期制作阶段参与了约530个资产的创作，帮助项目整体成本降低约80%。

资料来源：Runway，中邮证券研究所

图表26： 文生视频模型排名（ELO）

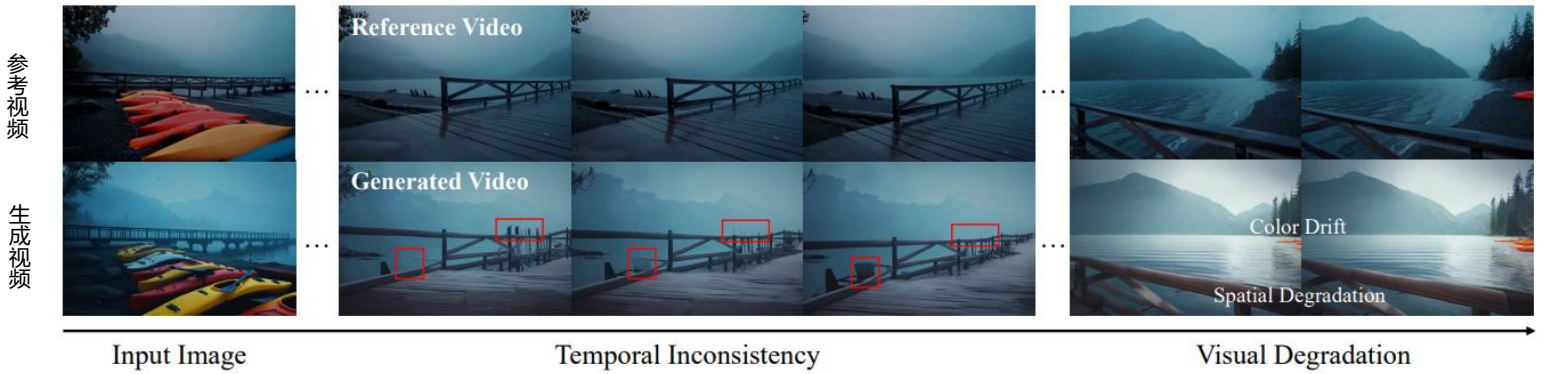


资料来源：Artificial Analysis，中邮证券研究所

4）生成时长方面：能力主要短板，短期以拼接过渡方案为主，原生长视频仍待技术突破

- **目前主流视频生成模型的单次原生生成时长仍处于“数秒级”。**根据我们统计海内外主要产品能力，单次生成普遍集中在5~10s区间，即梦、Sora 2等有所提升但上限目前也仅约20s，长时段连续生成仍未形成规模化突破。
- **长视频生成要求模型具备持续记忆与一致性维护能力，目前仍然存在两大技术限制。1）模型本身上下文记忆力有限，时序依赖捕捉仍然不足：**长视频需持续跟踪物体位置、人物身份、环境变化等状态。但DIT架构中的Transformer对“长时间依赖”的捕捉能力有限，易导致前后帧信息断裂，例如人物服装突然变化、场景逻辑矛盾（如室内场景无过渡切到户外）；**2）生成时长增加下，误差累积仍不可控：**短视频可容忍局部质量瑕疵，但在长序列生成中误差会逐帧累积，导致视觉质量随时间逐步下降，表现为内容漂移、模糊化或动作停滞等现象。

图表27： 目前长时视频生成技术仍然存在颜色退化、空间退化等问题



对比给定参考视频，目前长时生成视频技术会存在颜色退化（色度/亮度不一致）、空间退化（细节丢失、清晰度下降）等问题

资料来源：《LongVie: Multimodal-Guided Controllable Ultra-Long Video Generation》，中邮证券研究所

图表28： 目前主流视频模型生成时长仍普遍维持在数秒级

模型名	模型类型	单次生成时长	清晰度	视频尺寸	支持模式
可灵2.6	音画同步模型	5S、10S	不可选	文生音画支持16:9、1:1、9:16；图生暂不提供选择	文生/图生视频
即梦3.5 Pro	音画同步模型	5S、10S、12s	不可选	21:9、16:9、4:3、1:1、9:16	文生/图生视频
Hailuo 2.3	单视频模型	6s、10s	720P、1080P	不可选	文生/图生视频
Vidu Q2	单视频模型	文生视频普通版5s，会员可延展至8s；图生不可选	1080P	文生音画支持16:9、1:1、9:16；图生暂不提供选择	文生/图生/参考生视频
百度蒸汽机2.0（有声版）	音画同步模型	5s、10s	720P	不可选	图生视频
OpenAI Sora 2	音画同步模型	5s、10s、15s、20s	480P、720P、1080P	16:9、4:3、1:1、9:16、3:2、2:3	文生/图生（含视频）视频
Runway Gen 4.5	音画同步模型	5s、8s、10s	720P，可升级至4K	基础支持16:9	文生/图生（含视频）视频
Google VEO 3.1	音画同步模型	4s、6s、8s	720P、1080P	16:9、9:16	文生/图生视频

资料来源：各公司官网，AI工具集，Google Cloud，中邮证券研究所

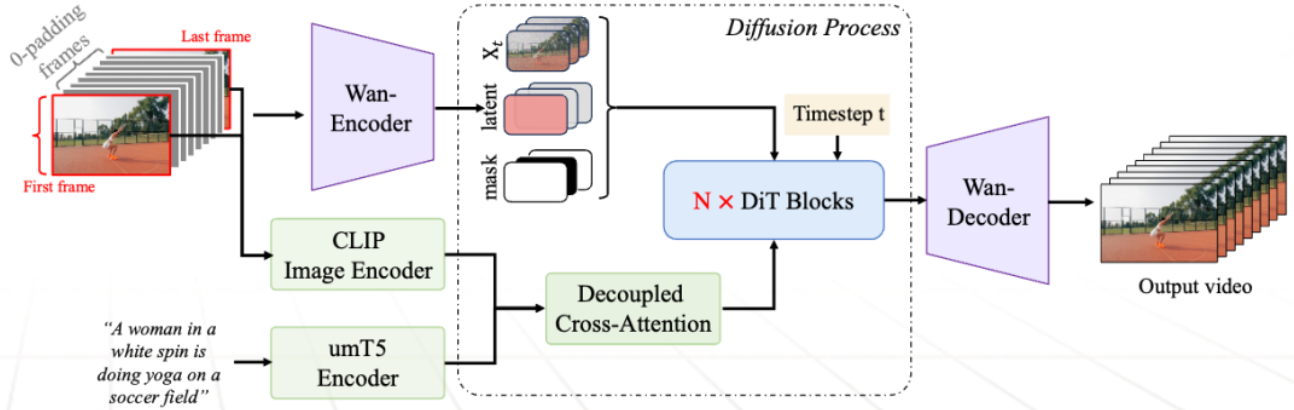
4）生成时长方面：能力主要短板，短期以拼接过渡方案为主，原生长视频仍待技术突破

- **目前行业主要通过“关键帧/首尾帧生成+分段式生成+拼接”的方式来实现长视频生成能力。**常见流程为先生成一组关键帧或锚点画面，用于锁定场景布局、人物状态与关键动作节点，再以锚点为条件逐段扩展中间帧或镜头片段，并通过拼接与平滑处理完成跨段衔接，以此实现长视频的续写能力。
- **分段式方法对错误累积有所优化，但仍受架构记忆上限约束，且使用端门槛仍明显偏高。**分段式方法通过关键帧锚点约束生成方向，可在一定程度上缓解长时生成中的错误放大问题，但底层DiT架构仍未建立真正覆盖整段视频的长时记忆，因此在生成时长拉长后仍可能出现人物状态漂移、物体消失、叙事不连贯等现象。同时，分段式方法需要先构思并生成分镜，再以提示词逐段续写，本质上类似“先定分镜脚本再补拍镜头”的专业级生产逻辑，对C端大众的快速内容表达场景仍存在明显门槛。

图表29：分段式生成长视频技术逻辑

资料来源：机器之心公众号，中邮证券研究所

图表30：首尾帧模型架构图（以万相模型为例）



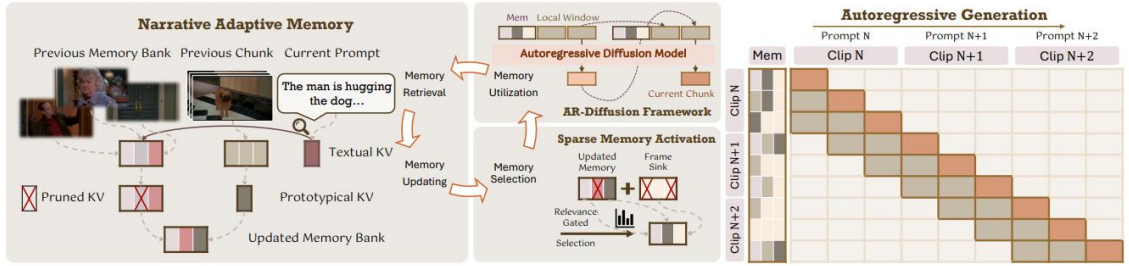
资料来源：通义大模型公众号，中邮证券研究所

2.1 技术进展：美学、多模态化能力表现优异，物理性、生成时长是主要瓶颈

4）生成时长方面：能力主要短板，短期以拼接过渡方案为主，原生长视频仍待技术突破

- **生成时长上的技术迭代仍在持续推进，短期分钟级视频生成可能迎来较快进展。**当前各大厂商虽技术路径各异，但仍可按延长记忆跨度与降低错误累积两大方向分类，比较有代表性的主要是来自可灵的记忆增强方案与字节的错误累积抑制方案：**1）记忆改善层面：**香港大学与可灵团队联合推出MemFlow方案，其采用叙事自适应记忆+稀疏记忆激活架构，即首先创建记忆库，在新片段生成时根据提示词检索库中相关视觉记忆，从而实现角色形象与叙事状态的跨段一致性；**2）错误累积优化层面：**UCLA与字节Seed团队联合提出Self-Forcing++方案，其采用“教师-学生双模型”纠错机制，即学生模型首先生成长视频（允许崩坏），通过教师模型对画面进行修正并更新，使模型在长时间范围内形成稳态生成能力。从效果看，两类技术均已将能力扩展至分钟级以上，其中 Self-Forcing++最长生成时长可达4分15秒，短期分钟级视频的技术化落地已具备较高可及性。
- **电影、剧集级别的超长叙事生成目前仍存明显距离，部分早期探索提供了潜在新思路，但仍需等待进一步论证。**包括清华团队提出的 Riflex 以及Google推出的 CoF（帧链）等技术，通过算法层面优化与搭建新型推理框架侧为未来视频生成扩展时长提供了可能的新路径，但相关研究仍停留在理论与早期实验阶段，实际价值仍有待验证。

图表31：MemFlow方案提升视频记忆力技术逻辑



构建并更新“记忆库”

让记忆库参与生成 & 控制显存与效率

生成链式结构：每一段（Clip）生成时，都沿用了前面的一部分信息（斜向延展的重叠块）

资料来源：《MemFlow: Flowing Adaptive Memory for Consistent and Efficient Long Video Narratives》，中邮证券研究所

图表32：Self-Forcing++方案优化错误累积技术逻辑

在传统短视频蒸馏中，模型每次都从随机噪声生成。Self-Forcing++ 改为在长视频 roll-out 后，把噪声重新注入到已生成的序列中，使后续帧与前文保持时间连续性。这一步相当于让模型「重启但不失忆」，避免时间割裂。

Self-Forcing++ **在训练阶段也采用滚动缓存**，实现训练-推理对齐，消除了「曝光漂移」和「帧重复」的问题。

将原本只在 5 秒窗口内进行的教师-学生分布对齐，扩展为滑动窗口蒸馏。教师不必生成长视频，也能「局部监督」学生的长序列表现，从而实现长期一致性学习。

资料来源：机器之心公众号，中邮证券研究所

2.2 技术趋势：世界模型或将带来新技术变革，视频生成有望再迎发展拐点

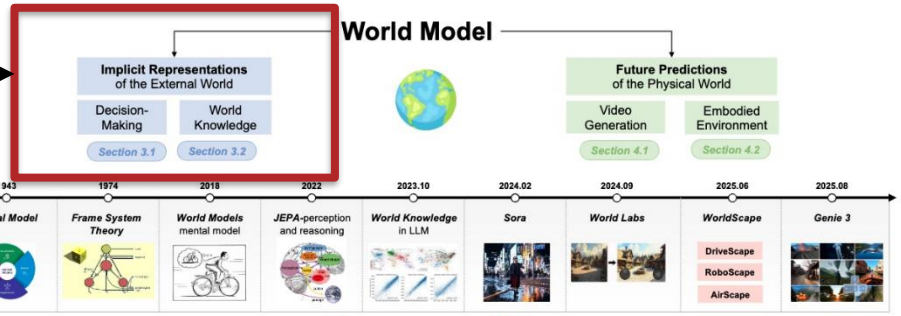
- **当前市场尚未形成对世界模型的明确共识，其讨论仍停留在科研突破与业界案例层面，未与视频生成商业路径产生直接连接，因而暂未获得估值与产业关注度的同步反映。我们认为，要理解世界模型的产业意义，需要从其技术演化、结构分类、落地场景与对现有视频生成技术的赋能四个维度展开，从而才能明确其产业价值。**

1）过往的世界模型指什么？

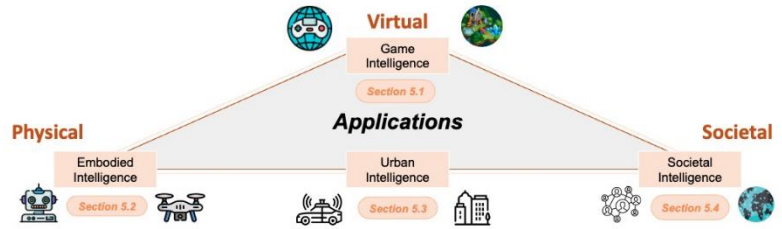
- **源于弥补现有大语言模型的认知缺陷，早期预期应用于自动驾驶、具身智能与人机协作机器人领域。**世界模型概念最早可追溯至2018年的《Recurrent World Models Facilitate Policy Evolution》论文，其借鉴了认知科学中心智模型理论，认为人类之所以能够推理、规划与决策，是因为大脑内部存在对世界的抽象建构，并能进行反事实推理，即不依赖真实试错，而在脑内“假想”未来后果。彼时世界模型主要旨在解决大语言模型的先天认知局限：即其学习机制本质仍停留在符号序列之间的线性关联与概率外推，缺乏对现实世界三维结构、物体属性、动态因果关系的理解。因此，早期世界模型核心目标并非单纯生成视觉或预测轨迹，本质仍是期望创建一个能从当前状态和动作预测下一状态的函数。

图表33：早期世界模型思想主要是为了解决大模型认知缺陷的问题

早期世界模型架构主要是为了打破现有模型概率建模的认知短板，让AI能够真正实现因果推理、理解世界

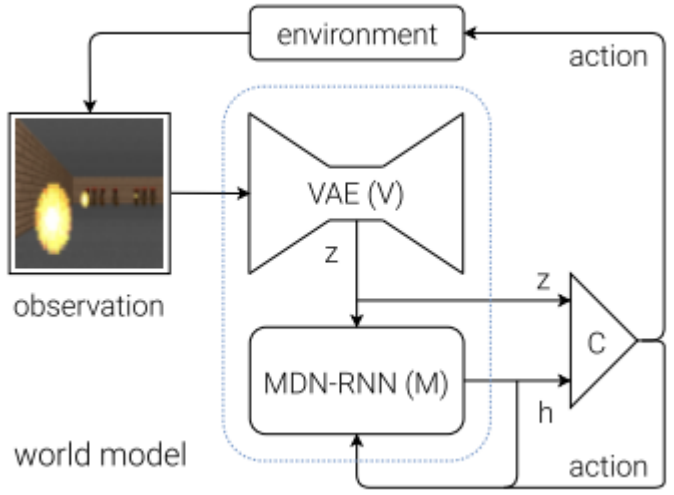


由于核心并不是创造显性像素世界，因此其产业落地更偏机器人、自动驾驶领域，视觉生成类的世界模型则是在2024年后才被提及



资料来源：《Understanding World or Predicting Future? A Comprehensive Survey of World Models》，中邮证券研究所

图表34：初代世界模型架构主要用于潜空间模型推理



早期世界模型架构的核心思想是：让AI在压缩的潜在空间中进行世界建模，而**不是直接体现在原始像素空间**

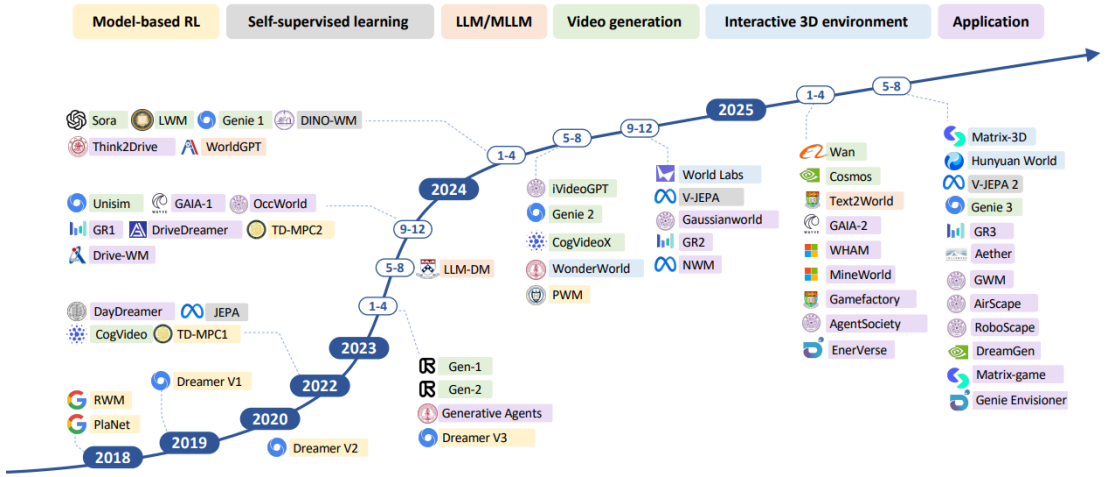
资料来源：《Recurrent World Models Facilitate Policy Evolution》，中邮证券研究所

2.2 技术趋势：世界模型或将带来新技术变革，视频生成有望再迎发展拐点

2)现今的世界模型指什么?

- **概念已发生演化，并逐渐分化出两条“同名但内核不同”的技术路线。**2024年，业内对于世界模型的理解开始延伸，并出现一种新的趋势判断：世界模型能力存在“语言生成→图像生成→3D 生成→世界生成（同时具备时序与空间序建模）”的趋势链条。同年2月，OpenAI将Sora定义为“world simulators”，强调其在像素空间直接学习现实世界的三维结构、物理规律与动态演化过程；但Meta同期推出的V-JEPA仍延续了传统世界模型“内部状态预测与抽象建模”路径，因此产生了路径分歧。目前世界模型最终形态仍无定论，但方向上已逐渐形成两大流派：
  - **1）表征派：**以Yann LeCun的JEPA路线为代表，认为世界模型是一个深藏在系统后端的“大脑”，只在表征处理后的潜在空间里运作，预测的是“抽象状态”；
  - **2）生成派：**以Google Hassabis的Genie 3路线为代表，认为世界模型应能够直接生成可交互的“可玩世界”、环境与动态演化。
- **在生成派内部目前已出现了更细分的路径划分。Entropy Town将生成派拆解为两类：界面型（Interface）与模拟器型（Simulator），对李飞飞团队的Marble和Google的Genie系列做了进一步区分。**但我们认为，从长期视角看，两条路径的边界可能会逐步收敛，因为二者的共同目标都是构建一个显性建模的3D虚拟世界。

图表35：世界模型理解的衍生推动了不同技术方向的产品落地



资料来源：《Understanding World or Predicting Future? A Comprehensive Survey of World Models》，中邮证券研究所

图表36：目前三类世界模型流派的核心原理

世界模型派系	方向	代表产品	定义
生成派	世界模型即界面	Marble	让人们能够从文字或二维素材，直接生成可编辑、可分享的三维环境。在这种模式下，「世界」是呈现在VR头显、显示器或电脑屏幕上的那片可供人观看与游走的空间。
	世界模型即模拟器	Genie 3	这类模型能生成连续、可控制的视频式世界，让智能体在其中反复尝试、失败、再尝试。
表征派	世界模型即认知框架	JEPA	一种高度抽象的形式，没有像前两种一样可供人欣赏的画面。关注点不在于渲染，「世界」以潜在变量和状态转移函数的形式呈现，是机器人完美的训练基地。

资料来源：Entropy Town，中邮证券研究所整理

2.2 技术趋势：世界模型或将带来新技术变革，视频生成有望再迎发展拐点

3)当下世界模型技术重心出现了什么边际变化?

- **2025年前：世界模型研究以“表征派”为主，且成果多停留在学术论文与实验室验证层面。**早期代表工作包括World Models、Dreamer系列以及Meta提出的JEPA等，其本质均依赖隐性潜变量学习与未来状态预测，旨在提升智能体学习效率，均属于表征派技术路线。Sora虽宣称自己是物理世界模拟器，但由于无法回应LeCun对于动作与世界状态的因果律问题（即模型无法回答“若施加动作，环境将如何改变”这一类具身推理问题），因此在学术界和产业界对于其是否属于世界模型仍有争议，彼时也未促成产业对生成派路线的系统性关注或资源投入。
- **2025年后：部分主流厂商入局生成类赛道，叠加Genie3交互能力首次验证了生成模型具备掌握因果律的潜力，重心开始向生成类倾斜。**自2025年起，海内外多家企业开始将世界模型作为明确布局方向，其中，NVIDIA推出Cosmos，李飞飞创办World Labs，腾讯混元及昆仑万维等本土厂商亦在积极跟进。2025年8月，DeepMind发布可交互世界模型Genie3，首次在工程层面为“生成类模型能否回应因果律”提供了可验证证据，证明生成式路线在长期演化下具备成为世界模型标准范式的可行性。在此信号作用下，市场对生成类技术重视度提升，并进一步推动Runway等视频模型厂商入局。

图表37：目前已发布生成世界模型厂商及产品概览（部分）

模型名	厂商	发布时间	模型能力
Cosmos	英伟达	2025 年1月发布首个版本， 8 月发布更新版本	可用于世界生成和推理，或进行后训练以开发专用的物理 AI 模型，分为三个方向：1）Cosmos Predict：根据单个图像和文本提示生成 30 秒的预测世界状态视频；2）Cosmos Transfer：可跨各种环境和光照条件快速扩展单个模拟或空间视频；3）Cosmos Reason：使用视频和图像的结构化推理。
Genie3	Google	发布时间为2025年8月（此前版本为23、24年发布的Genie 1及2代）	Google首个支持实时交互的世界模型，用户只需输入文本提示，Genie 3就能以每秒24帧的速度实时生成可供自由探索的动态世界，并在720p分辨率下保持数分钟的画面一致性。应用场景主要包括虚拟训练场、游戏与内容创作、AI 智能体研究等。
Marbles	World Labs	2025年11月	核心能力可以概括为三点：1）多模态生成：可以根据一张图片、一段视频，甚至一句文字提示，重建出结构完整、细节丰富的 3D 世界；2）AI 原生的世界编辑能力：允许用户像调整真实场景一样对世界进行局部替换、材质变化、光照调整或布局重构；3）可落地的制作流程：支持将生成的世界导出为高斯溅射、三角网格或视频格式，可直接进入Unreal、Unity、Blender 等常见创作工具，融入游戏、影视等行业的工作流。
GWM-1	Runway	2025年12月	目前包含三个变体：1）GWM Worlds：用于实时环境的模拟与探索；2）GWM Avatars：能够模拟人类对话；3）GWM Robotics：用于机器人操作。
Matrix-Zero	昆仑万维	2025年2月	Matrix-Zero世界模型包含两款子模型。其中：1）自研3D场景生成大模型，支持将用户输入的图片转化为可自由探索的真实合理的3D场景；2）自研可交互视频生成大模型，提供以用户输入为核心驱动的可交互空间智能视频生成方案，支持根据用户实时输入生成互动视频效果。
混元3D	腾讯	2025年7月	可基于一句文本描述或一张图像输入，生成一个 360 度沉浸式的三维场景。用户可在其中进行视角切换、自由环视、行走，视觉体验接近 VR 世界，且支持物理仿真与二次编辑，亦可导出为全景贴图用于虚拟展示。

资料来源：各公司官网，36Kr，腾讯科技公众号，量子位公众号，证券时报网，APPSO，中邮证券研究所整理

2.2 技术趋势：世界模型或将带来新技术变革，视频生成有望再迎发展拐点

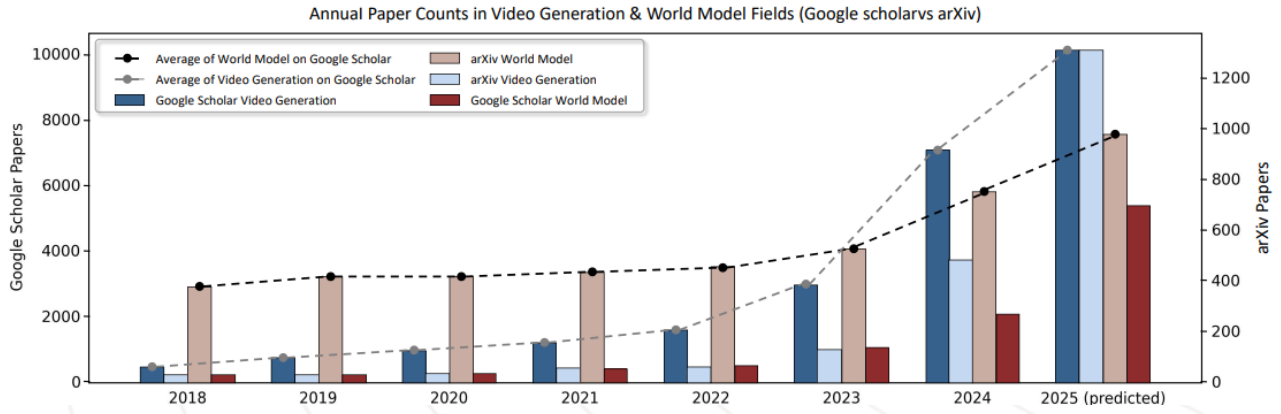
4) 生成世界模型和视频生成模型有什么关联?

- **世界模型的终极愿景在具身智能，但其目前最具确定性的价值是提供可长期维持状态并遵循物理规律的虚拟世界生成能力，正对应视频生成在长时一致性与物理可信度上的核心短板。**如前文述，目前视频生成技术在长时一致性、物理准确性等层面仍存技术瓶颈。相比之下，世界模型自底层设计要求模型能够持续维护环境状态、跟踪实体关系并在内部模拟动力学与因果变化，使生成行为从“预测下一帧”转向“维持一个可运行的世界”，因此其技术演进对于视频生成具有直接的工程补短意义。
- **当前行业亦普遍将视频生成视作世界模型的雏形，对世界模型的能力变化预期，亦与视频模型的技术演化路径高度耦合。**世界模型的发展已形成相对清晰的代际规划：第一阶段：侧重画面与文本的一致性与短秒级运动生成，即为当下的视频生成类模型技术；第二阶段：要求模型支持更长甚至无限时长的连续生成，同时保持语义级与导航级的交互能力，实现跨镜头、跨场景的一致叙事；第三阶段：要求内部具备物理规律与动态反馈机制，使生成逻辑从像素预测转向“世界内部运行”；最终阶段：指向开放式随机推理与多时空尺度的自主演化，是具身智能的终极形态。可以看到，前三阶段的每一步能力跃迁，都与视频生成的技术需求高度耦合。

图表38：世界模型的四代演进路径

资料来源：《Simulating the Visual World with Artificial Intelligence: A Roadmap》，中邮证券研究所

图表39：视频生成模型与世界模型论文的发表数量同步变动，呈现相互依赖关系



资料来源：《Simulating the Visual World with Artificial Intelligence: A Roadmap》，中邮证券研究所

2.2 技术趋势：世界模型或将带来新技术变革，视频生成有望再迎发展拐点

- **从架构来看，世界模型与现有视频生成模型（DiT）是不同技术路线，其发展不受DiT范式的进展约束。**后者侧重对下一帧像素进行条件式外推，本质仍属于“表层渲染逻辑”，缺乏对长期状态、空间关系与物理反馈的内部表达，因此在长时序生成与一致性维持方面迭代速度仍相对缓慢。而世界模型通常由状态表征模型、动态模型与决策模型三部分构成，通过显式维护环境状态并模拟动态变化，使生成过程转向“世界内部演化”，从机制上更具备长时空一致性和物理遵循能力。
- **从推进速度来看，世界模型在“时间维度、空间一致性与物理逻辑”等关键性能上的迭代也要明显更快，未来可能成为视频生成的另一条主流演进路径。**以Google Genie系列为例：2024年底发布的Genie2仅能支持浅层三维环境搭建与基础交互，画面维持约10~20秒即出现信息崩溃；不到一年时间，Genie3已能够以24fps实时生成可供自由探索的动态虚拟世界，并在720p分辨率下维持数分钟级画面一致性。同时，其新增“可提示的世界事件”与“视觉记忆”机制，使得房间内物品布局、涂鸦等实体特征在用户多次往返时仍保持稳定，空间一致性呈现量级式提升。

图表40：目前已发布生成世界模型厂商及产品概览（部分）

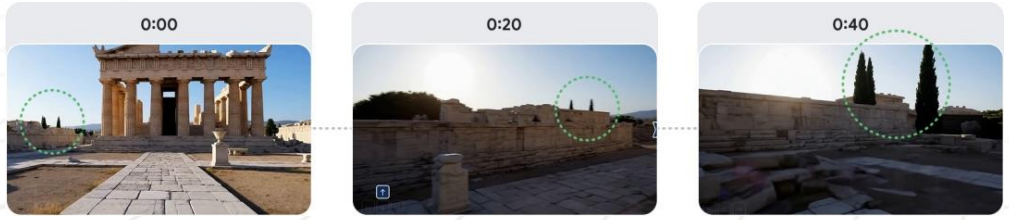
长时画面稳定及物理还原

对世界物理属性的建模：还原波浪拍打路面的自然物理现象，并能实现复杂的环境交互。

多风格虚拟世界创造力

动画与幻想场景建模：具备与视频模型同等的场景想象力，能够创造奇幻场景与富有表现力的动画角色。

空间一致性



Genie 3的一致性非常高：建筑物左侧的树木在整个交互过程中始终保持一致，即使时而进入视野时而消失

资料来源：腾讯科技公众号，新智元公众号，中邮证券研究所

图表41：Genie世界模型系列技术迭代进展与其他模型对比概览

	GameNGen	Genie 2	Veo	Genie 3
分辨率	320p	**360p**	720p 至 4K	**720p**
领域	针对特定游戏	**3D 环境**	通用	**通用**
控制方式	针对特定游戏	**限制的键盘/鼠标操作**	视频级描述	**导航、可指令触发实时反馈的世界事件**
交互时长	几秒	**10–20 秒**	8 秒	**数分钟**
交互延迟	实时	**非实时**	不适用	**实时**

资料来源：Google DeepMind，中邮证券研究所

图表42：世界模型技术架构

模块	功能定位	常见技术/方法	核心作用
状态表征模型	将原始观测数据压缩为低维潜在状态	VAE（变分自编码器）、自编码器、降维/滤波技术	保留关键特征、过滤噪声，使模型能够高效处理复杂输入
动态模型	根据当前潜在状态和动作预测下一状态的变化	RNN（循环神经网络）、LSTM（长短期记忆）、随机状态空间模型（SSM）等	模拟状态转移规律，提供“虚拟沙盘”供模型进行内部试错与物理逻辑推演
决策模型	基于预测的未来状态规划最优动作序列	MPC（模型预测控制）、深度强化学习（Actor-Critic）等	选择动作策略并根据价值或奖励信号进行调整，使智能体执行合适行为

资料来源：智见AGI公众号，中邮证券研究所整理

2.2 技术趋势：世界模型或将带来新技术变革，视频生成有望再迎发展拐点

- **目前世界模型已被业内普遍视为与大语言模型（LLM）同级的重要人工智能发展路径，相关参与者数量仍在持续增加。**除前文提及厂商外，字节跳动、华为盘古、xAI、快手可灵等企业亦在积极布局世界模型相关技术路线，后续可能逐步实现产品化落地。此外，尽管OpenAI的Sora 2尚难严格界定为世界模型，但其明确提出“world simulation”叙事，也表明其在战略层面已经认可生成派并可能在后续推出真正的世界模型产品。
- **行业共识逐步形成背景下，研发节奏预计或将进一步加快，2026年有望成为世界模型跃迁的关键节点。**目前Genie3等首批世界模型仍处于研究预览阶段，仅向特定研究人员与创作者开放，核心价值仍在于验证生成式路径的可行性，我们认为其行业位置更类似于2022年GPT-2向GPT-3的跃迁前夜。当前随着入局者持续增多，叠加3D资产与物理仿真资源逐步丰富，技术框架有望加速成熟，2026年或将迎来世界模型的GPT-3时刻，其能力亦有望从技术展示逐步迈向基础场景的商业化应用，进入真正的产品验证周期。

图表43：目前主要布局世界模型技术路线厂商进展

厂商	世界模型计划及进展
xAI	根据量子位介绍，**xAI已公开宣布进入世界模型方向**，并从英伟达相关团队引入多名资深研究人员；其内部正在尝试构建可由AI自动生成自适应、逼真3D场景的模型体系，**首批试点方向可能面向电子游戏应用场景**。
luma AI	一家专注于影片生成的初创企业，估值已超过40亿美元。2025年9月推出其首个推理式视频生成模型 Ray3。根据华尔街见闻报道，**2025年12月，公司宣布进一步扩充模型研发与团队规模，研发方向将延伸至“世界模型”体系**。
华为盘古	2025年6月20日，在华为开发者大会2025（HDC 2025）上，**华为发布了基于盘古多模态大模型的世界模型，可以为智能驾驶、具身智能机器人的训练，构建所需要的数字物理空间**。
字节跳动Seed	已成立Seed-多模态交互与世界模型团队，致力于研发具备人类水平的多模态理解与交互能力的模型。目前公司已于25年10月推出3D生成大模型——Seed3D 1.0，实现从单张图像到高质量仿真级3D模型的端到端生成。根据镁客网介绍，**字节世界模型项目已进入攻坚阶段**。
可灵	2025年6月，在CVPR 2025大会上，快手可灵AI事业部万鹏飞博士围绕模型架构与生成算法、互动与可控能力、效果评估与对齐机制、多模态理解与推理四方面公开介绍团队在视频生成与世界模型方向的最新研究进展，显示**其技术规划已开始向世界模型体系延伸**。
阿里巴巴	阿里巴巴目前未公开提出独立的世界模型产品或路线，但**2025年6月达摩院联合湖畔实验室与浙江大学发布了将视觉-语言-动作模型（VLA）与世界模型相融合的统一框架 WorldVLA**，体现其在世界模型方向的研究探索。但由于VLA主要应用于自动驾驶与机器人场景，未来是否延展至生成式世界模型仍存在不确定性。

资料来源：品玩，字节跳动，观察者网，量子位，镁客网，Cmoney，新浪科技，华尔街见闻，36Kr，中邮证券研究所

3 商业化进展：C+B端双路并进，影视级项目有望迎来商业元年

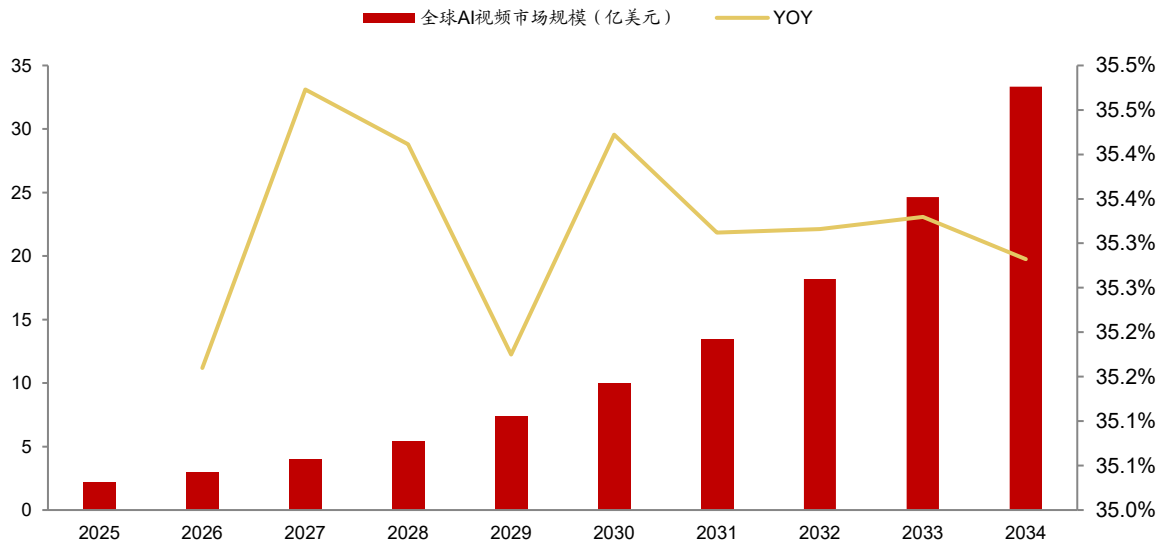
3.1 C端：订阅模式为主，社交化等新方向探索有望拓宽商业路径

3.2 B端：API模式持续催生素材级生成需求，影视级项目有望迎来落地元年

3 商业化进展：C端双路并进，影视级项目有望迎来商业元年

- **市场规模持续扩张，2034年有望突破30亿美元，C/B端双轮驱动商业化路径逐步清晰。**全球AI视频生成市场正处于高速成长阶段，根据Precedence Research预测，2025年市场规模约为2.19亿美元，预计2026年将增长至2.96亿美元，同比增长35.16%；至2034年有望达到33.32亿美元，2025–2034年间的复合增长率（CAGR）达35.32%。随着模型能力提升与应用场景拓展，行业商业化路径已逐步分化为C端订阅与B端API/解决方案两大模式：
  - **To C端：**主要通过订阅制收费，用户可按需选择免费版、标准版、高级版、尊享版等不同等级，月度订阅价格从几元至数百元不等，典型客户包括内容创作者、短视频用户及泛娱乐消费群体；
  - **To B端：**主要通过API调用与定制化解决方案变现，客户覆盖影视制作、互联网平台、电商营销、广告代理等行业，费用按调用次数、生成时长或项目定制程度计价，月度支出从几十元至数万元不等。

图表44：全球AI视频市场规模及增速



资料来源：Precedence，中邮证券研究所

图表45：AI视频商业模式概览

客群方向	收费模式	计价模式	客户群体
C端	订阅制	用户可按需选择免费版、标准版、高级版、尊享版等不同等级，月度订阅价格从几元至数百元不等	内容创作者、短视频用户及泛娱乐消费群体
B端	API调用与定制化解决方案	生成时长或项目定制程度计价，月度支出从几十元至数万元不等	影视制作、互联网平台、电商营销、广告代理等行业

资料来源：第一财经公众号，中邮证券研究所整理

3.1 C端：订阅模式为主，社交化等新方向探索有望拓宽商业路径

- **目前C端仍然以订阅模式为主要收入来源，业内多采用“免费试用+多档订阅+积分充值”服务架构。**目前海内外主流视频模型面向C端均普遍采用“免费试用+多档订阅+积分计量”的三段式结构，即先通过免费额度引导用户试用，再以订阅机制分层解锁核心能力（如生成次数、清晰度、时长限制），并辅以积分体系满足超额使用需求，形式上延续了传统SaaS的基本框架。从定价来看，海外主流厂商的基础订阅价格多集中在20~30美元/月，国内则普遍为60~80元/月，国内具备性价比优势。但从产品丰富度看，Sora 2、Veo 3等海外模型并未单独计费，而是作为ChatGPT、Gemini等大模型生态会员中的一部分，用户在订阅后可同步使用其他生成类能力，整体使用维度更为丰富。

图表46：主流AI视频模型C端会员

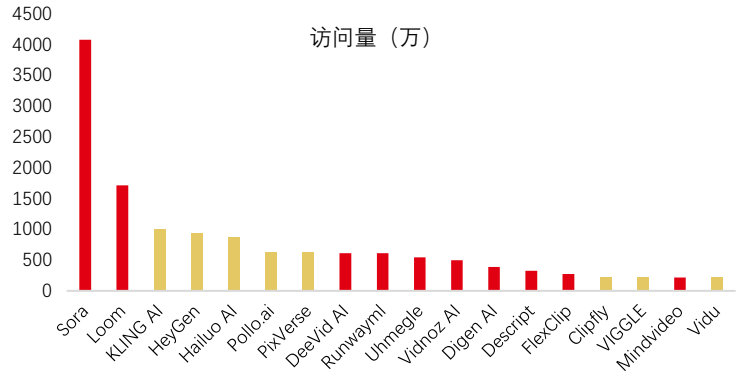
模型	厂商	模式	订阅价格（月价格）	订阅使用权限	增值服务	说明
Sora 2	OpenAI	免费+多档订阅	Plus：20美元/月；Pro：200美元/月	免费：约 12 次/天(邀请制)；Plus：50-100 次/天，1080p；Pro：无限生成	-	Sora2暂未独立提供会员，其服务内容内嵌于ChatGPT会员体系
Veo 3.1	Google	免费+多档订阅+积分充值	Pro：19.99美元/月；Ultra：249.99美元/月（标准定价）	Pro：每月可获得1,000点 AI 点数，Ultra每月可获得 25,000 点；1000点对应Veo 3 .1 Fast模型生成约20部，Veo 3.1 Quality模型生成约10部。未订阅用户每月可获得 100 点 AI 点数	提供点数充值服务选择，折合100点/美元	Veo3目前内嵌于Gemini会员体系，未单独提供服务
Gen 4.5	Runway	免费+多档订阅（积分兑换）	分为免费、Standard、Pro、Unlimited四档会员，订阅版本收费15~95美元/月不等（连续包年提供折扣）。此外，亦提供定制级企业订阅服务	Standard：每月包含 625 个积分，对应生成约25s Gen-4.5视频；Pro：每月包含 2250 个积分，对应生成约90s Gen-4.5视频；Unlimited：无限生成。免费版一次性提供125积分，但仅可使用前代产品	-	会员除Gen4.5外，亦可使用其他前代产品
HeyGen	HeyGen	免费+多档订阅（分辨率及用户数）	Creator：29美元/月；Team：39美元/月	Creator、Team均无生成视频数量上限要求，单视频任务均支持30分钟级扩展，区别主要在Team可支持4K级分辨率及用户数量上限	-	-
Hailuo2.3	MiniMax	免费+多档订阅+积分充值	分为免费、基础、标准、大师、至臻、尊享六级会员，收费版本68~1399元/月不等（连续包年提供折扣）	免费版本提供一次性积分使用，付费版本提供1000~20000贝壳（积分）/月不等，基于Hailuo 2.3/2.0生成6s、1080p单视频耗费80积分，同规格下Fast版本50积分	提供点数充值服务选择	会员除Hailuo 2.3外，亦可使用其他前代产品
Kling 2.6	可灵AI	免费+多档订阅+积分充值	分为免费、黄金、铂金、钻石、黑金五级会员，收费版本66~1314元/月不等（连续包年提供折扣）	黄金：每月660灵感值（积分），约生成33个标准视频；铂金：每月3000灵感值，约生成150个标准视频；钻石：每月8000灵感值，约生成400个标准视频；黑金：每月26000灵感值，约生成1300个标准视频	提供点数充值服务选择，基础充值折合10点/元，充值金额增多享有不同程度折扣	会员除Kling 2.6外，亦可使用其他前代产品，同时支持使用图片生成模型
即梦视频3.5	即梦AI	免费+多档订阅+积分充值	分为免费、基础、标准、高级四级会员，收费版本79-649元/月不等（连续包年提供折扣）	基础：每月1080积分，约生成216个视频；标准：每月4000积分，约生成800个视频；高级：每月15000积分，约生成3000个视频	提供点数充值服务选择，基础充值折合10点/元，充值金额增多享有不同程度折扣	会员除即梦视频3.5外，亦可使用其他前代产品，同时支持使用图片生成模型
PixVerse V5.5	拍我AI	免费+多档订阅+积分充值	分为免费、标准、专业、尊享四级会员，收费版本79-459元/月不等（连续包年提供折扣）。此外，亦提供定制级企业订阅服务	标准：每月1200积分，约生成60个视频；专业：每月6000积分，约生成300个视频；尊享：每月15000积分，约生成750个视频	提供点数充值服务选择，基础充值折合每500点需35元，充值金额增多享有不同程度折扣	除V5.5标准版外，亦可使用其他同代产品，同时支持使用图片生成模型

资料来源：各公司官网，API易，中邮证券研究所

3.1 C端：订阅模式为主，社交化等新方向探索有望拓宽商业路径

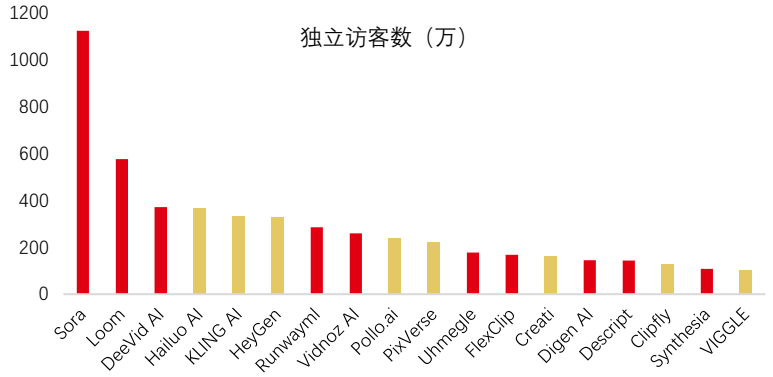
- **用户量是现阶段C端商业模式的主要评判参考，Sora体量仍断档领先。**由于目前视频模型主要采用SaaS模式变现，因此可以直接沿用其评判标准。但由于行业目前披露ARR、留存率等指标厂商数较少，对于复杂性指标的测算仍然困难，因此用户量是现阶段评价模型商业化能力的主要参考：
  - **从访问量与用户数来看，**Sora在整体访问量与独立访客数量上仍处断层领先地位，但可灵、海螺、HeyGen等本土头部企业发展亦较快，整体访问量亦基本达到千万级别；
  - **从使用时长来看，**Sora凭借用户数体量，总使用时长仍居首位。但在单次访问时长维度，Sora与其他产品差别不大，在一定程度上反映出各平台产品技术并未存在显著差距。因此，短期Sora依靠用户基数与品牌优势，仍能维持体量领先，但长期竞争仍取决于各平台的技术迭代与创作链条延伸能力，不排除部分后发产品在技术或产品表现上实现反超。

图表47：25年11月全球视频模型访问量排名



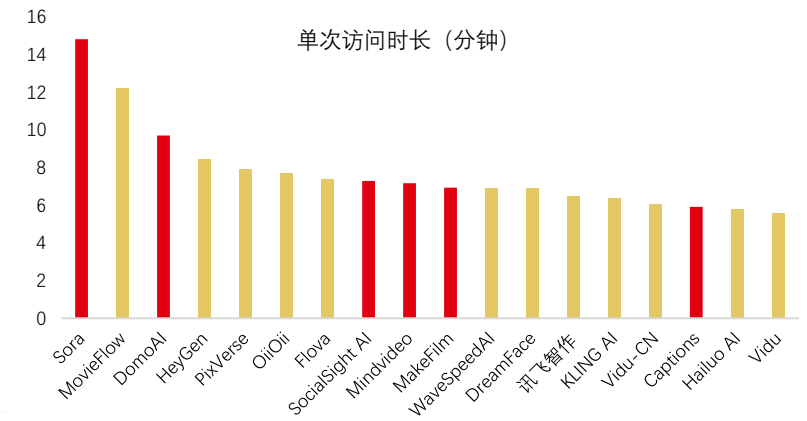
资料来源：非凡产研，中邮证券研究所（注：红色为海外企业）

图表48： 25年11月全球视频模型访客数排名



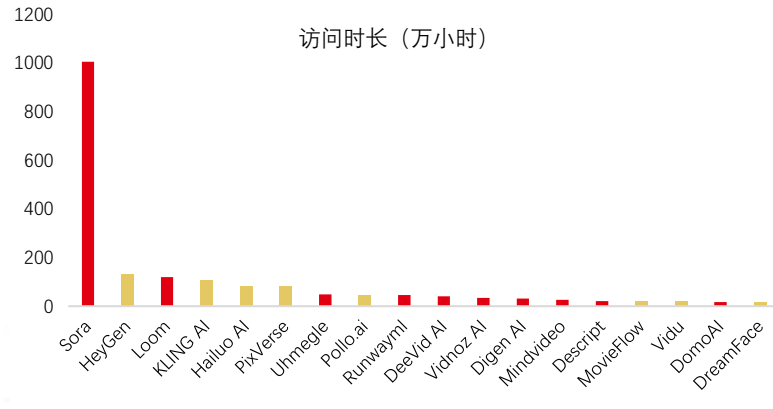
资料来源：非凡产研，中邮证券研究所（注：红色为海外企业）

图表49： 25年11月全球视频模型单次访问时长



资料来源：非凡产研，中邮证券研究所（注：红色为海外企业）

图表50： 25年11月全球视频模型访问总时长量排名

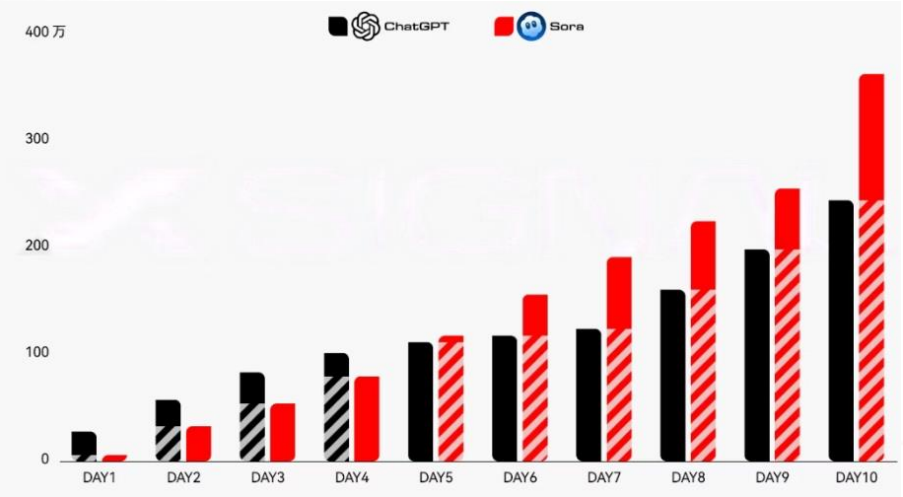


资料来源：非凡产研，中邮证券研究所（注：红色为海外企业）

3.1 C端：订阅模式为主，社交化等新方向探索有望拓宽商业路径

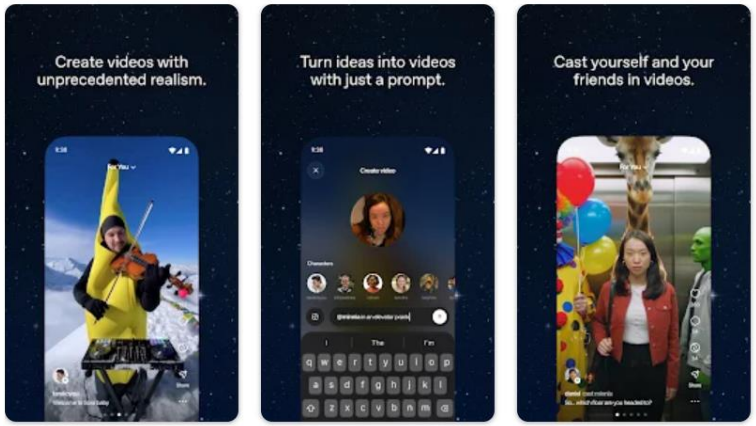
- **Sora App上线为AI视频走向“社交化”带来一定启示，有望同时为C端变现与B端分发带来新流量路径。**2025年9月30日，OpenAI发布Sora 2时，同步推出iOS端应用“Sora”。Sora App被定位为面向普通用户的社交化视频创作平台，其核心功能主要聚焦互动与再创作。根据奇异AI公众号报道，Sora App上线初期DAU虽仅5.72万，不及同期ChatGPT（27.68万）表现，但发布5天即反超ChatGPT，第10天DAU已达365.26万，超过ChatGPT同期水平47.2%。
- 尽管Sora APP仍处于早期，留存与粘性尚待验证，但其初期流量的大幅攀升验证了“生成+社交”融合逻辑的可行性。当前可灵等主流厂商亦已在自身应用中嵌入内容分享与展示机制，AI视频产品的社交化趋势正逐步形成。后续伴随厂商应用的社交生态逐步形成，或在传统订阅模式上，为AI视频C端产品打开广告、电商等新增收入路径。此外，相比传统的“下载-发布”链路，厂商自带社交生态将使视频内容在生成完成后即具备被观看、被互动的天然渠道，亦有望为B端用户生成的视频内容提供新的分发与触达场景。

图表51：Sora app对比chatGPT前十日DAU



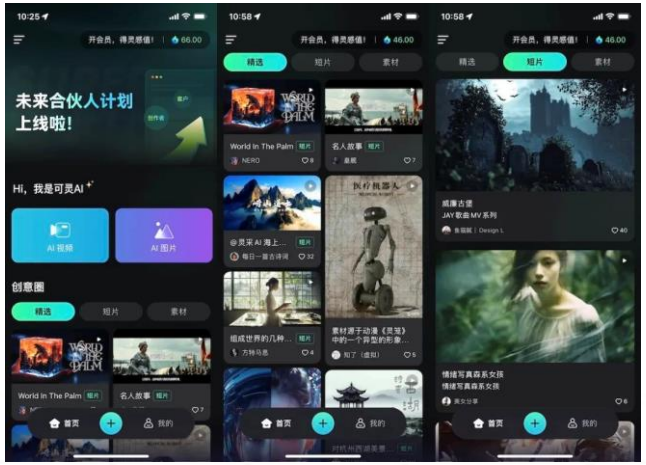
资料来源：奇异AI公众号，中邮证券研究所

图表52：Sora APP社交功能展示



资料来源：Google Play，中邮证券研究所

图表53：可灵APP社交功能展示



资料来源：腾讯网，中邮证券研究所

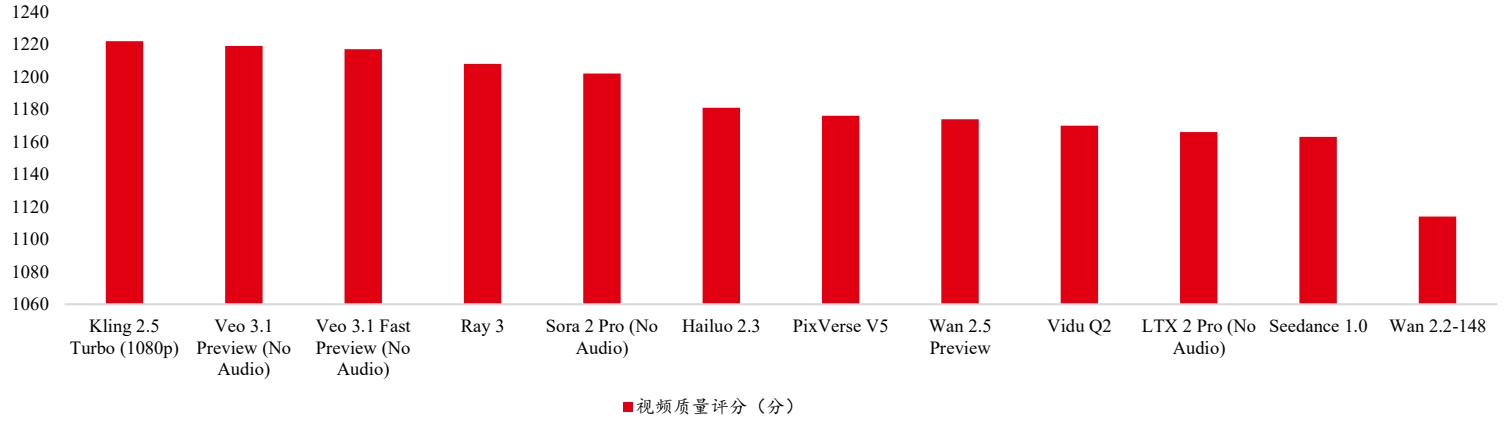
3.2 B端：API模式逐步跑通，影视级项目有望迎来商业化元年

1）素材级生成

- **API是主流业务模式，核心电商展示、广告等领域应用已基本成熟。**如前所述，由于现阶段视频生成模型仍难以支撑长叙事、镜头连贯与角色一致性等高要求的影视级内容创作，B端应用重点聚焦于短时长、可结构化的内容场景，典型如电商商品展示、广告创意视频等。

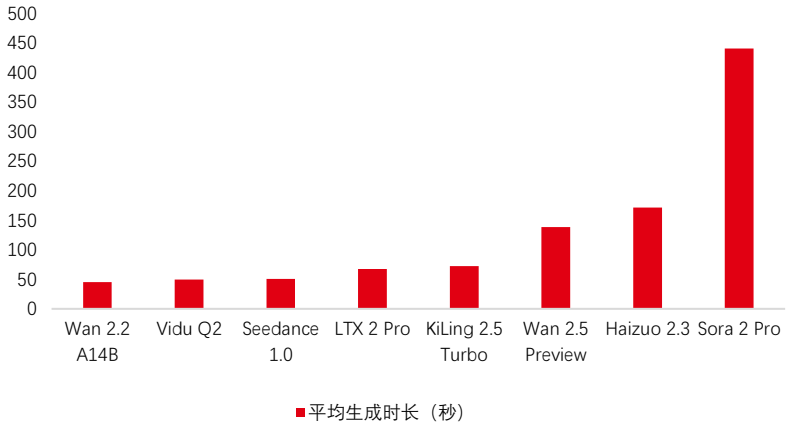
- **“质量+效率+成本”是API模式核心评价维度，部分国产模型已实现行业领先。**API 商用能力主要取决于生成质量、生成速度、调用成本三个关键考察维度：1）生成质量，可灵2.5 Turbo 的表现已超越Veo 3.1和Sora 2等海外代表产品，Hailuo 2.3、PixVerse V5等国产模型亦具备竞争力；2）生成效率，国内主流产品普遍实现分钟级输出，Wan 2.2单视频生成耗时仅45.2秒，而Sora 2 Pro仍需超过7分钟，响应速度相对滞后；3）价格层面，Sora与Veo 3.1系列API单秒调用价格在0.15~0.5美元之间，而部分国产模型如海螺、万相已将成本压缩至美分级，显著降低下游应用的试错门槛与规模化生成成本。

图表54：视频模型生成质量评分



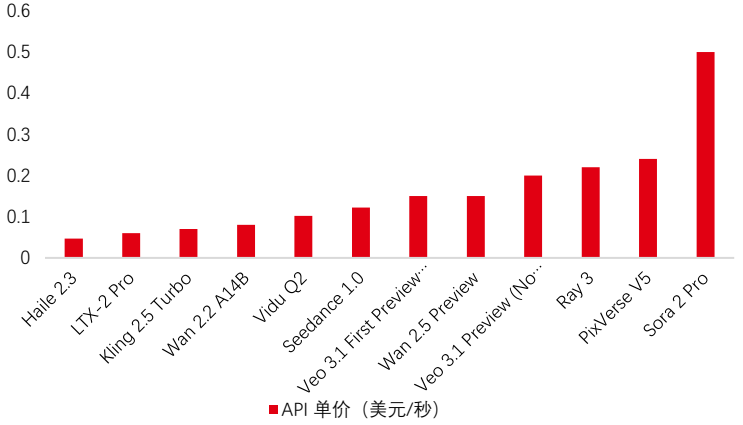
资料来源：Artifical Analysis，中邮证券研究所

图表55：视频模型API单视频生成时间



资料来源：Artifical Analysis，中邮证券研究所

图表56：视频模型API调用价格



资料来源：Artifical Analysis，中邮证券研究所

2）影视级制作

- **高度复杂的系统性工程，传统API模式不具备提供完整制作流能力，需采用多工具串联的组合式路径。**影视生产链条涉及从剧本创作、分镜规划，到素材生成、剪辑合成与后期渲染等多个专业子流程。现阶段AI模型多聚焦于画面生成，尚不具备全流程统筹与跨环节能力整合。目前行业主要是通过多类AI工具组合使用，按需嵌入各影视环节。如文本生成阶段使用大语言模型，分镜设计阶段调用图文转化工具，画面生成阶段依赖视频模型，后期编辑环节则搭配配音、剪辑、调色等其他AI辅助产品。

图表57：影视环节多样复杂，目前单一模型不具备整合工作流能力

	前期制作		中期制作		后期制作	
	大纲脚本	分镜规划	素材拍摄	素材选取	剪辑	渲染音效
内容	故事创意、主题风格构思，明确故事大纲，如主要情节和角色发展将大纲扩展成完整的剧本，包括对白、场景描述和情节细节	文字剧本视觉化，确定镜头类型、镜头运动、角色位置和动作 编制镜头列表，包括每个镜头的编号、描述、拍摄角度、镜头运动等	根据分镜规划制定拍摄计划，完成器材准备和技术测试、准备演员、服化道及布景设置	根据情节和主题需要进行素材整理、分类、剔除无效素材 对于无须拍摄类内容需要从素材库进行素材选取	初步剪辑、逻辑剪辑，拼接出粗略的视频版本 精剪微调，对影片进行精细剪辑，调整镜头长度、顺序、节奏等	添加视觉特效，确保效果自然、完成颜色校正、调色与整体风格一致 同步处理音频素材，确保对白清晰、搭配音效，增强视频的听觉效果
传统	Office 365 + 各种生产力软件			Adobe shutterstock	Pr Au Ae	
混合	ChatGPT Gemini Qwen 利用自然语言处理技术，AI可以帮助创作者生成多种剧本创意，提供不同的故事走向和对话选项。通过分析观众偏好和市场数据，AI能够实时反馈并优化剧本内容。			OpenAI SORA runway stability.ai 目前主要有三种模式:文生视频、图生视频、文生图-图生视频。目前头部文生围产品的图片生成质量最好，一般先用文生围应用生成指导帧，然后再作为视频生成的输入进行动态化	runway 视频大部分关键元素在生成阶段完成，用户再通过搭配其他工具、垂类模型实现出片	
原生	整合音视频生成、编辑等所有功能的一站式应用，可能产生新的交互界面					

已经成熟 → 过渡阶段 → 未来阶段

资料来源：量子位，CSDN，中邮证券研究所整理

3.2 B端：API模式逐步跑通，影视级项目有望迎来商业化元年

- **过往多个项目已在技术层面验证了AI全流程参与影视级制作的可行性。**国内的《带我去飞》《团圆令》，以及海外的《generAIdoscope》《Our T2 Remake》《海上女王郑一嫂》《再见机器人》等，均被明确为AIGC全流程制作电影，其中《再见机器人》甚至早于2024年即已完成制作，表明 AI 在影视级内容生产上的技术路径已趋于成熟。

- **但从产业实践看，过往项目主要以制作方自行整合工具为主，模型厂商直接提供商业化服务的案例较少。**过往项目大多依赖制作方自行拆解制作流程、拼装多种模型与工具完成生产。例如《Our T2 Remake》被拆分为约50个独立片段，由不同创作者分别生成后再进行人工整合。因此，尽管技术可行性已被验证，影视制作层面的商业化交付能力并未真正形成，行业整体仍处于从“工具验证”向“工业化交付”过渡的阶段。

图表58：海内外AI影视级项目概览（部分）

电影名	电影截图/宣传图	国家	简介
带我去飞		中国	全球首部AIGC空战电影，影片以抗日战争重庆大轰炸时期为历史背景，主要讲述空军王牌飞行员赵骧云与重庆少女朱倩茹的战火爱情，电影全长38分钟。该片已于25年10月11日在腾讯视频、极光TV上线。
团圆令		中国	国内首部院线AIGC动画电影，该片实现了剧本策划、场景布景、角色设计、画面生成等关键环节的AI全链条赋能，将传统动画电影2-3年的常规制作时间缩短至5-6个月。目前该片已于25年12月于京点映。
然后呢		中国	全球首部入围柏林、戛纳、威尼斯三大国际电影节的AI长片电影。根据中国电影报报道，该片从最初的创意到制作、剪辑和配乐，全部由导演曹译文一人在AI帮助下完成。
音符中的密码		中国	由长影集团出品，并于25年9月在长影电影院放映。该片通过AI数字人技术复原了长影表演艺术家李亚林、刘世龙的经典形象，通过科技手段实现了两位电影表演艺术家在大银幕上的首次同框合作。
带我去飞		日本	日本首部完全由AI制作的电影，该片由三个独立且风格迥异的故事组成，目前已于25年8月公映。
Our T2 Remake		美国	由50位AIGC创作者历时数月分段合作完成，全片达到了近90分钟的常规商业电影放映长度，目前该片已在 YouTube 和 RAD TV 上映。
海上女王郑一嫂		马来西亚/新加坡	全片总时长70分钟，全球首部政府批准走进院线公映的AIGC大电影。人物原型是传奇女海盗“郑一嫂”，《加勒比海盗》系列的女海盗王“清夫人”亦是以该传奇女性为蓝本创作。
再见机器人		英国	该片早于2024年即完成制作，全长87分钟，制作成本约8000美元/分钟，相比传统好莱坞动画片降低200至300倍。

资料来源：搜狐网，北京市科学技术委员会，中国电影报，中国影视网，新浪财经，腾讯网，搜狐网，红杉汇公众号，36Kr，Facebook，t2remake，中邮证券研究所

3.2 B端：API模式逐步跑通，影视级项目有望迎来商业化元年

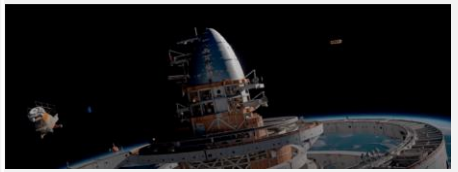
- **目前海外已有初创企业开始尝试提供影视级AI解决方案，并在收入层面验证了商业可行性。**以Utopai为例，其成立于2025年，是一家AI原生影视工作室，前身为3D生成AI公司Cybever。不同于传统模型厂商以API或单点工具授权变现的路径，Utopai选择向下游内容生产与发行端纵向延伸，直接参与影视项目的制作与收益分配，并通过多类AI工具的系统化整合，规避了单一模型能力不足与制作方整合成本高企的问题，使AI成功以解决方案形态嵌入影视工业流程。目前公司已通过《科尔特斯》（Cortés）、《太空计划》（Project Space）等项目实现累计约1.1亿美元收入，成为当前在影视级AI制作领域实现规模化变现的重要案例之一。

图表59：Utopai影视AI解决方案分层协同架构概览

模型工具	定位	介绍
规划层	自回归模型作为「导演大脑」	**序列预测机制：**AR 模型以剧本为输入，通过前帧预测后帧的机制，生成涵盖角色 ID 向量、摄像机轨迹、光影变化等要素的时空计划。该计划本质是一个机器可执行的「拍摄蓝图」，确保长达数十分钟的片长中元素演进逻辑保持一致。 **状态记忆与因果推理：**模型能够维护可回放的长程状态记忆，例如追踪角色从第 1 镜到第 50 镜的动作轨迹，避免传统模型因局部生成导致的逻辑断裂。
渲染层	扩散模型作为「执行引擎」	**条件化生成：**扩散模型不再随机「抽卡」，而是严格依据规划层输出的结构化指令（如深度图、光流信号）生成画面。例如，当规划层指定「摄像机以俯角拍摄雨夜小巷」 时，扩散模型就会据此渲染细节。 **物理规律注入：**通过训练时引入带精确标注的 3D 合成数据，模型学习空间遮挡、材质反射等规则，避免生成内容违反重力或碰撞逻辑。
协同接口	统一状态空间	**规划层与渲染层通过统一状态空间交换信息：**规划器输出未来帧的几何与语义约束，渲染器据此生成像素，并反馈生成结果供规划器优化后续计划。这一闭环解决了扩散模型「生成即遗忘」的缺陷。

资料来源：机器之心公众号，中邮证券研究所整理

图表60：Utopai参与制作电影介绍

项目名	电影截图	介绍
《科尔特斯》		奥斯卡提名编剧Nicholas Kazan操刀，概念设计由《第九区》幕后美术Kirk Petruccelli完成，计划制作两部各100分钟的史诗巨作，讲述殖民时期传奇人物埃尔南·科尔特斯的生平。这部片曾被《名利场》评为“好莱坞最难拍的十部剧本”之一，长年因预算与技术难度未能立项
《太空计划》		由Martin Weisz执导的8集科幻剧，被形容为“《壮志凌云》遇上《世界大战》”。目前已完成欧洲市场的版权预售。

资料来源：硅谷科技评论公众号，中邮证券研究所整理

3.2 B端：API模式逐步跑通，影视级项目有望迎来商业化元年

- **头部厂商亦已开始布局影视级项目领域，其中OpenAI参与制作的《Critterz》预计于26年上映，AI影视级制作有望步入商业化元年。**2025年9月，OpenAI宣布将参与由英国Vertigo Films与洛杉矶Native Foreign联合出品的动画电影《Critterz》。在该项目中，OpenAI将利用其大模型能力及图像、视频生成模型参与剧本构思、视觉概念、美术素材生成及镜头预演等环节，同时由其创意专家Chad Nelson以顾问制片人身份参与协作，该电影计划于2026年在戛纳电影节首映。此外，海内外厂商亦在持续加码影视级制作方向的布局。Runway已成立旗下制作与娱乐部门Runway Studios，通过与电影制作人、工作室、音乐家、作家及独立艺术家合作，推进生成式技术驱动的创意项目实践。本土厂商方面，可灵AI亦于2025年10月亮相东京影视节内容交易市场TIFFCOM，与全球创作者就AI赋能影视创作展开交流。

图表61：OpenAI 参与制作的《Critterz》预计将于2026年上映

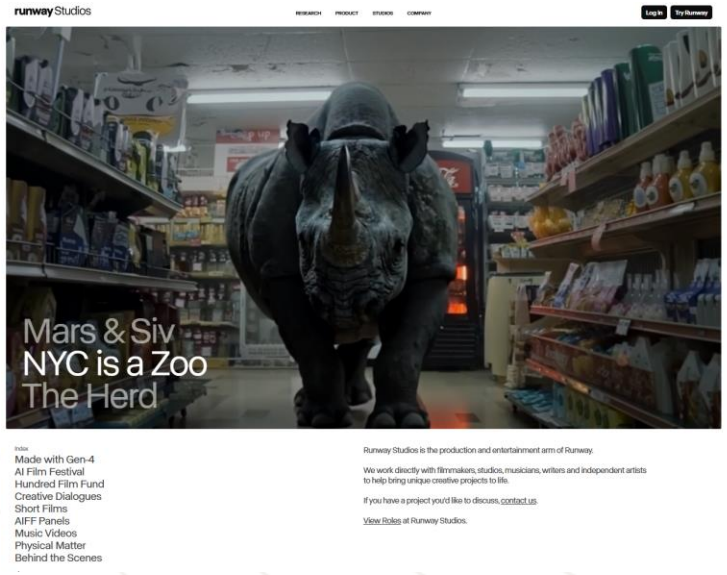
《Critterz》动画形象

《Critterz》动画草图

OpenAI团队将使用GPT-5作为开发与创作“中台”，并辅以OpenAI的图像/视频模型（DALL·E、Sora）完成美术素材、镜头预演与镜头级内容生成

资料来源：腾讯网，中邮证券研究所

图表62： Runway成立工作室探索影视制作领域



资料来源：Runway，中邮证券研究所

4 传媒：AI视频核心应用场景，广告、影视、游戏均有望受益

4.1 营销：内容变革有望带动AI广告需求提升，营销服务商或迎新增量价值

4.2 影视：短剧全场景赋能在即，长剧/影视高价值环节有望优先受益

4.3 游戏：短期创作环节持续降本增效，长期新品类有望拓宽产业边界

4 传媒：AI视频核心应用场景，广告、影视、游戏均有望受益

- **传媒行业是“AI+应用“的主阵地，广告、影视、游戏等细分板块均高度契合。**传统传媒行业内容供给高度依赖创意与人力，生产链条长、成本高、更新频率快，而AI具备大规模内容生成、个性化分发与自动化运营能力，能够有效缓解传统传媒在效率与成本上的瓶颈，提升供给弹性与商业化效率：1）在广告营销环节，AI可基于多模态大模型精准洞察用户兴趣、生成多版本创意素材，并动态优化投放，实现千人千面的内容分发，提升转化效率与ROI；2）在影视制作环节，AI可介入策划、剧本、分镜、剪辑、特效等各个流程，缩短制作周期、降低人力成本并提升内容质量，助力长周期IP资产化运营；3）在游戏环节，AI可赋能角色设计、剧情生成与场景搭建，显著提升用户沉浸感与付费意愿，延长内容生命周期，增强商业变现能力。

图表63：传媒行业是“AI+应用”主阵地

资料来源：易观分析，中邮证券研究所

4.1 营销：内容变革有望带动AI广告需求提升，营销服务商或迎新增量价值

- **广告营销天然契合AIGC，具备全链路赋能能力。**广告营销因其链路结构清晰、目标导向强，天然适配智能化改造。无论是人群洞察、内容生成、渠道投放还是后链路评估，每一环节都具备高度数据化基础和可被建模的决策逻辑。与企业内部流程型场景相比，广告更强调“创意表达+效率释放”的平衡，而AI正好具备自动化、个性化与生成式能力的三重优势。
- **内容创作是广告营销的主要支出环节，具备多维场景AI赋能价值。**广告制作费用通常占广告营销全流程成本的20%以上，在品牌类广告中该比例可进一步提升至30%~40%，是广告营销中投入占比最高、且对传播效果起决定性作用的核心支出环节之一。从AI赋能路径看，AI可协助完成平台素材检索，并根据不同模态需求，生成相应的广告文案、图片及视频内容，从而提升内容生产效率与创意供给能力。

图表64： AI具备赋能营销全链路能力

资料来源：头豹研究院，中邮证券研究所

图表65：AI生成在营销环节中的应用

生成式AI + 内容生产	降低了内容创作的门槛	营销人员可以将精力更多集中于消费者心理与艺术创意上	营销人员可以将精力更多集中于消费者心理与艺术创意上	营销人员可以将精力更多集中于消费者心理与艺术创意上
	使内容批量化生成成为可能	内容素材的需求相比之前将有数量级的提升	内容素材的需求相比之前将有数量级的提升	内容素材的需求相比之前将有数量级的提升

广告营销行业中的内容生产部分中营销素材生产主要包括三种形式:文案生成、图片生成、视频生成。

文案生成

案例：借用Jasper等平台生成符合营销理论的专业文案，相对以前的时间成本的维度，至少提升了70%。出门问问推出AI写作平台"奇妙文"，可根据不同目的生成营销文案。

图片生成

案例：tudouman数字艺术家用Midjourney为麦当劳打造的系列产品图"麦麦文物"，生产效率能达到传统制作的十倍以上。

视频生成

案例：百度营销旗下的AIGC营销创意平台"擎舵"，在短入短几分钟内，就能够为上汽大众朗逸新锐的新品定制生成文案和数字人口播视频并进行上线投放。

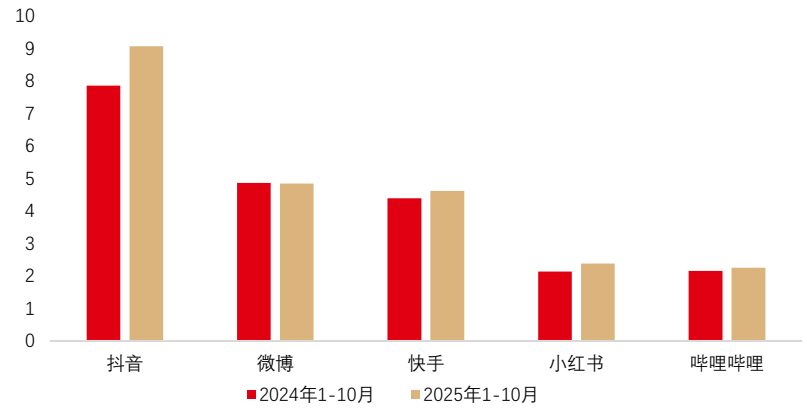
资料来源：量子位，中邮证券研究所

4.1 营销：内容变革有望带动AI广告需求提升，营销服务商或迎新增量价值

- **媒介变化驱动内容形态升级，视频广告成为核心增量方向。**从广告形式划分来看，主要包括图文、图片与视频三类。在早期电商时代，广告内容以图文为主，核心诉求在于“跑量换点击”，内容只需结构清晰、信息完整，渠道获取能力远重于内容表达。但伴随信息获取平台逐步从电商转向抖音、快手等短视频平台，广告媒介结构发生系统性变化，推动内容形态由静态向动态迁移，视频从附属素材逐步跃升为核心传播载体。2024年，信息流视频广告已占整体广告市场的18.03%，若加上贴片式视频广告合计份额将达到23.83%；从内容素材维度看，2025年上半年全网移动广告中，视频类素材投放占比已超过65%，其中竖屏视频占比高达54.8%。

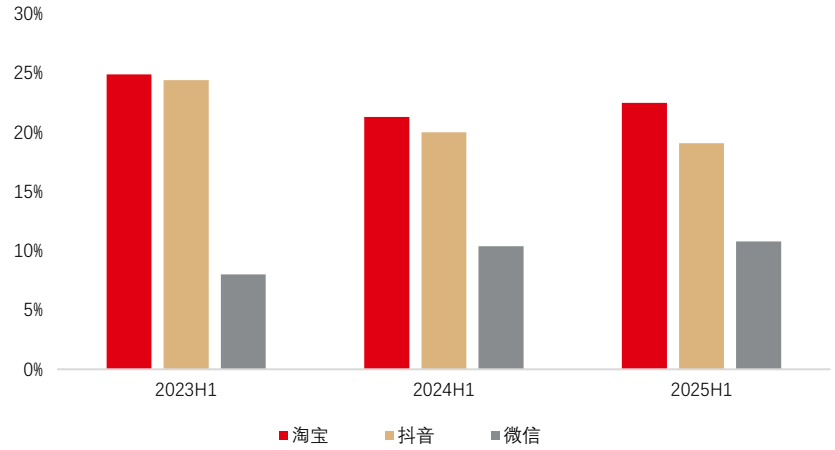
资料来源：头豹研究院，中邮证券研究所

图表66： 新媒体平台月活用户数（亿人）



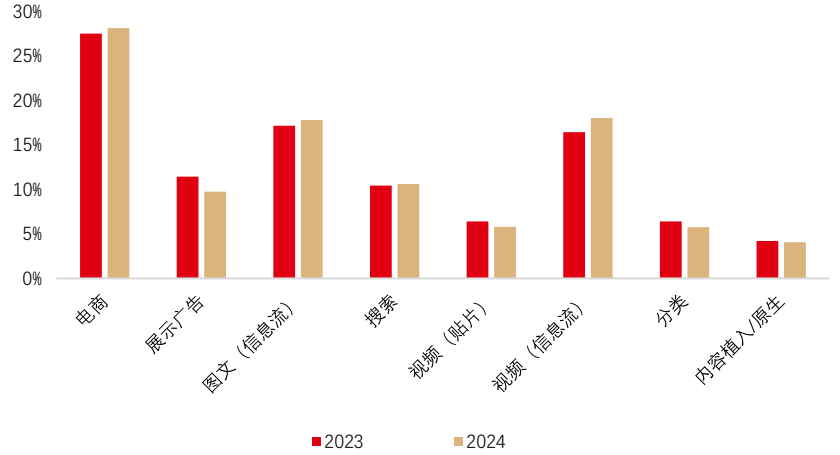
资料来源：QuestMoblie，中邮证券研究所

图表67： 硬广投放费用TOP3平台



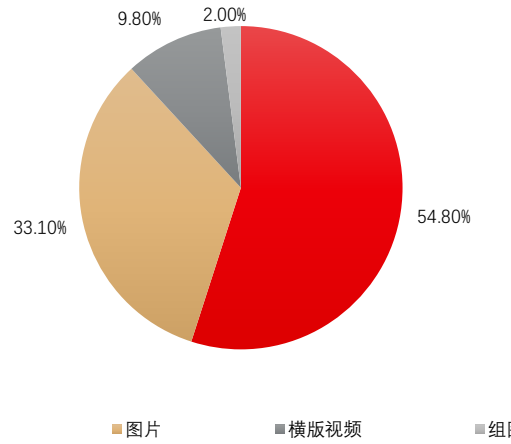
资料来源：QuestMoblie，中邮证券研究所

图表68： 2023-2024年广告形式收入占比



资料来源：中关村互动营销实验室，中邮证券研究所

图表69： 25H1全网在投移动广告素材占比



资料来源：AppGrowing，中邮证券研究所

4.1 营销：内容变革有望带动AI广告需求提升，营销服务商或迎新增量价值

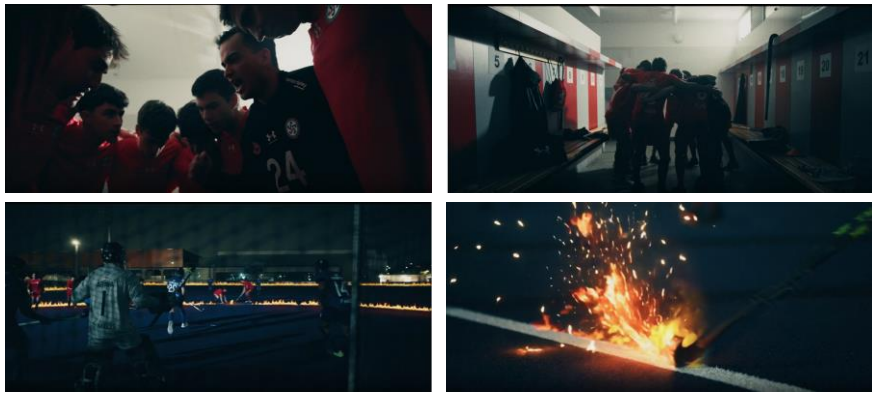
- **短视频广告需求结构与当前视频生成能力高度契合，低渗透率下短期仍具需求释放潜力。**当前视频生成技术与短视频广告场景之间具备高度适配性，短视频广告多集中在6～15秒之间，内容结构简洁、场景单一、对叙事连贯性要求相对有限，恰好契合现阶段视频生成模型在镜头长度、语义驱动与可控性上的能力边界。从效率侧看，相比传统广告制作所需的脚本、拍摄、美术与后期等环节，AI生成视频具备明显的降本增效优势，使广告主得以以更低成本、更高效率批量生成可投放内容。根据头豹研究院统计，AI介入广告内容创作后，文案撰写效率提升约500%，创意图片效率提升200%，混剪视频提升300%，图文助手效率提升600%，创意拓展效率更是高达800%。过去由于视频生成能力相对滞后，AI在广告内容生成中的应用仍集中于文案、图文等环节，视频内容在营销中的渗透率依然偏低，具备显著的后发增长潜力。

图表70： AI在广告领域的视频创作能力持续提升

《玩具反斗城的起源》-Sora（2024）

Sora生成的第一个商业广告：时长1分06秒，由玩具反斗城团队和导演NikKIeverov共同构思制作，并在戛纳国际创意节亮相
表现评价：1）人物角色的细节在不同片段一致性不足(例如衣物细节颜色、纹理、眼镜样式、细节面部特征等细节有轻微畸变)；2）背景元素存在畸变，例如背景中的自行车的有畸变特征

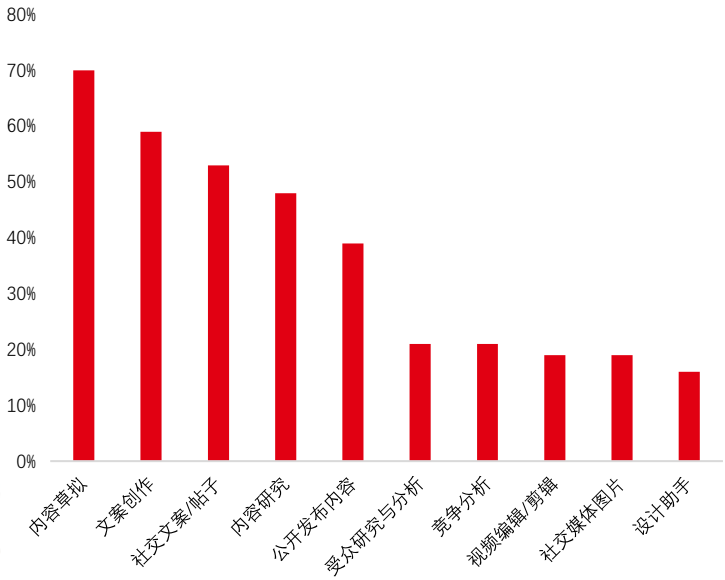
《安德玛冰球题材广告》-Runway（2025）



Runway与安德玛合作广告：时长30秒，由智利Trópico制作公司创。Runway 在后期制作阶段参与了约530个资产的创作，帮助项目整体成本降低约80%.
表现评价：1）具备较强的复杂场景理解与生成能力，能够支持多人、多主体的动态场景创作，镜头切换间画面风格与人物状态保持一致性；2）具备物理建模能力，能够真实还原火花等高动态元素的光影扩散与微粒运动特征

资料来源：量子位，Runway，中邮证券研究所

图表71： AI广告创作各环节渗透率（2024）

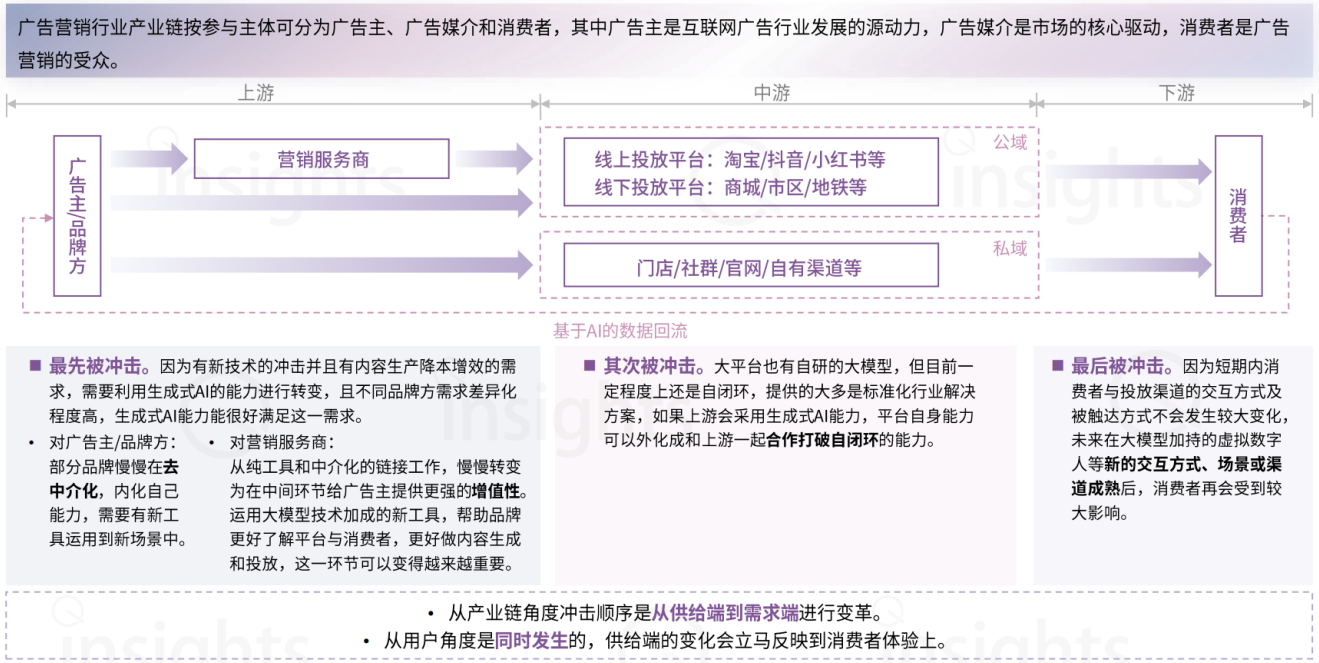


资料来源：艾瑞咨询 ，中邮证券研究所

4.1 营销：内容变革有望带动AI广告需求提升，营销服务商或迎新增量价值

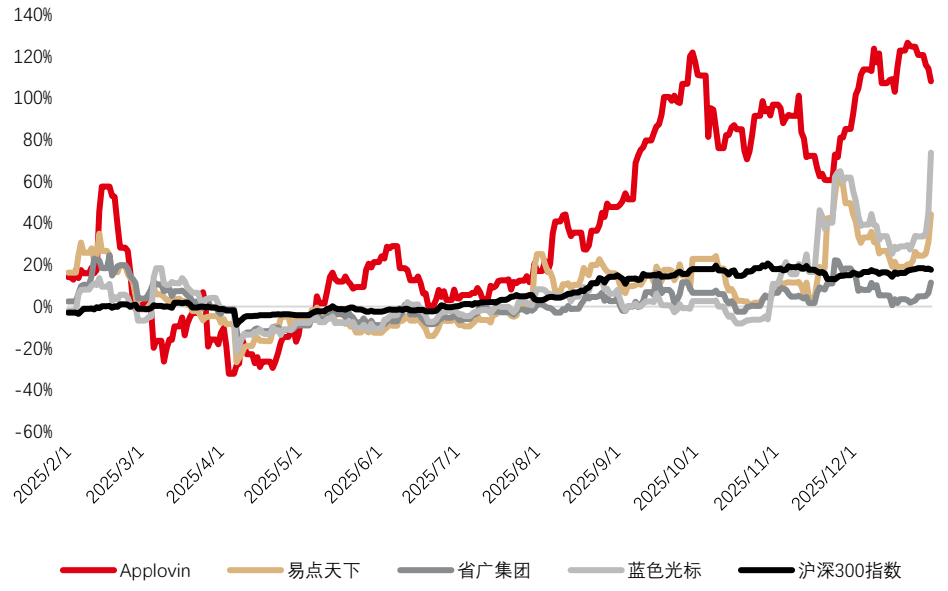
- **内容创作环节重视度提升或将重塑产业链角色分工，营销环节价值权重有望进一步抬升。**当前广告营销链条主要由广告主、营销服务商与媒介平台三类主体构成。以往在“流量链接+投放执行”为核心的传统逻辑下，营销服务商更多扮演工具性和中介型角色，核心价值集中于渠道投放与资源对接。而在生成式AI等新技术推动下，内容生产环节的降本增效需求日益凸显，品牌方对内容定制、场景适配与快速测试的需求提升，使内容创作在整体营销中的战略地位迅速上升。AI工具的普及有望打破原有服务商的能力边界，使其从单一媒介投放职能，转型为能协助品牌进行内容策划、生成、测试与投放优化的全链路合作伙伴。传统营销商或将从“执行型中介”升级为具备技术理解与内容创作能力的“智能增值节点”，在AIGC营销生态中扮演更具战略价值的关键角色。从资本市场反馈来看，回顾2025年，Applovin股价全年累计涨幅108.08%，充分反映海外市场对AI+营销产业的认可，国内头部营销类企业跟进速度较跟进。

图表72：AI冲击从供给端到需求端依次进行，营销服务商价值凸显



资料来源：量子位，中邮证券研究所

图表73： 25年海内外主要营销公司股价涨幅变动情况



资料来源：iFind，中邮证券研究所

4.2 影视：短剧全场景赋能在即，长剧/影视高价值环节有望优先受益

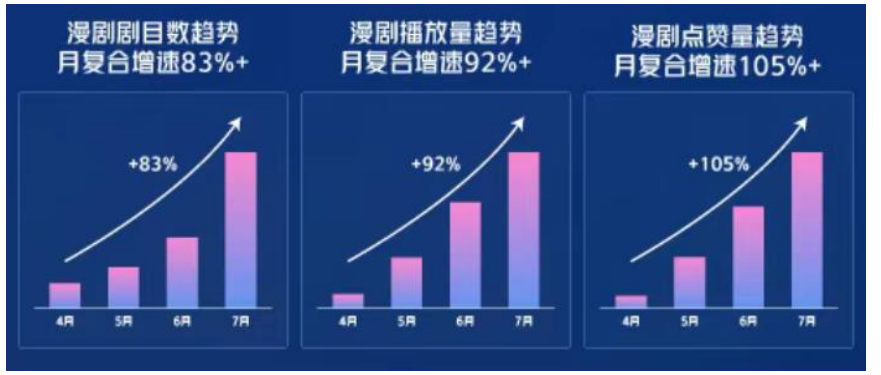
短剧：漫剧率先跑通商业模式，拟真剧有望进入量产阶段

- **AI漫剧：融合短剧+动画两端优势，是当前与视频生成契合度最高的影视应用方向。**从形态上看，短剧单集时长通常为1~3分钟，集与集之间的情节连续性要求较低，天然契合当前视频模型的片段化生成特征，使AI能够实现全流程参与；从内容上看，当前AI漫剧以动态漫与2D动画为主，其画面复杂度、物理精度及光影细腻度均显著低于CG或拟真场景，对模型能力要求更低。
- **得益于效率与成本优势，AI漫剧已率先实现商业闭环。**据短剧自习室统计，目前通过AI赋能，AI漫剧的平均制作周期可较传统动态漫缩短50%以上（从50~60天压缩至30天以内），整体成本可降至传统模式的10%~30%。2025年4~7月期间，AI漫剧剧目供给量以83%的复合增长率持续扩容，播放量与点赞量亦分别实现92%与105%的复合增长。

图表74：目前AI能力能够全流程赋能漫剧制作

资料来源：巨量引擎营销观察公众号，中邮证券研究所

图表75：目前AI漫剧呈现市场井喷态势



资料来源：巨量引擎营销观察公众号，中邮证券研究所

图表76：得益于效率与成本优势，AI漫剧商业化能力凸显

AI应用对于漫剧的商业化能力提升

	传统动态漫	现代漫剧(AI驱动)
更新频率	周更	月更
回款速度	慢	快
成本	高	10~30%
供给	少	多
市场	综上所得会小	综上所得变大

AI应用对于漫剧的效率提升

	纯人工	人工+AI
文本内容	2周	1周
出图上色	1.5~2周	3~6天
动效剪辑	1.5周	1周
配音	2周	3~4天
总时长	50~60+天	30天内

资料来源：短剧自习室公众号，巨量引擎营销观察公众号，中邮证券研究所

4.2 影视：短剧全场景赋能在即，长剧/影视高价值环节有望优先受益

- **AI拟真人短剧：技术侧成熟度提升，从“配文式视频” 向 “叙事型剧集”迈进。**早期受限于画面精度、动作自然性及音视频同步能力，AI拟真内容大多采用“AI旁白+配音+PPT式画面”结构，通过文字叙述兜底剧情推进，人物互动仅作点缀，难以还原真实剧集的沉浸感。而2025年以来，Sora 2、Veo 3等新一代多模态模型相继面世，显著提升了镜头连续性、情绪表达与物理一致性，使得更复杂的剧情构建与角色交互成为可能，AI短剧的内容表现力出现跃升。
- **部分厂商已开始试水落地，成本+收益共振下有望逐步进入量产阶段。**据新华网介绍，2025年1月抖音平台TOP5000短剧中采用全AI生成的仅4部，10、11月即分别上升至69、 217部。目前在“降本增效”与“内容供给扩展”双重催化下，后续入局厂商可能持续增长，AI拟真剧有望进入量产阶段：1）成本端：微短剧制作成本持续增长，2023年单部成本约20~30万元，2024H1已涨至70~80万元，甚至出现单部成本超300万元的头部项目。AI生成具备明显的提效降本优势，以现有案例来看，夫子AI团队一部10集短剧，仅由3人团队在10天内完成，总成本仅为5594元；《白狐》剧组仅用两周即完成短剧全流程制作，整体成本降至传统模式的数十分之一；2）收入端：AI生成具备更强的题材适配能力，能够实现以往因预算或实拍条件受限而难以制作的内容表达，有望进一步拓展平台内容供给边界。

图表77：技术快速迭代，拟真剧画面表现力短期提升已十分明显

《奶团太后宫心计》
发行方：爱微剧场
上映时间：25年8月

以基础生成能力为主，人物呈现相对僵硬，情绪与表情变化有限；画面构图与镜头调度较为单一，主要依赖“镜头缩放+平移”完成叙事，整体仍处于技术验证阶段

《小小神尊穿书反杀白莲花》
发行方：爱微剧场
上映时间：25年12月

人物表情与动作细节明显提升，情绪传达更加自然；场景构建与切换复杂度显著提高，已引入多景别、多机位与节奏化剪辑，叙事流畅度与视觉完成度同步增强

资料来源：影视独舌公众号，独舌短剧公众号，中邮证券研究所

图表78：25年12月播放量超5000万AI真人短剧作品已超12部（抖音平台）

剧集名称	开播日期	累计播放量	剧场名称	类型	集数
重生不做舔狗，我靠千亿遗产横扫末世	2025-11-16	1.1Y	AI漫说剧	重生 / 玄幻仙侠	60
我靠唱歌打脸全团	2025-10-01	1.1Y	爱微剧场	逆袭 / 年代爱情 / 打脸虐渣	61
原创AI短片·致命毒液	2025-04-14	8319.3w	未来派	AI生成	15
末日来袭，我靠缩小系统逆转人生	2025-11-10	7860.6w	小润AI梦工厂	系统 / 玄幻仙侠 / 奇幻脑洞 / 剧情	27
穿越乞丐：开局丞相府大小姐跟我混	2025-10-17	7791.4w	爱微剧场	逆袭 / 穿越 / 古代 / 异能 / 奇幻脑洞	61
749秘档·秦岭邪雨	2025-11-12	6872.3w	749秘档AIGC	民国 / 架空 / 奇幻脑洞 / 悬疑推理 / 年代	24
小小神尊穿书反杀白莲花	2025-12-20	6415.9w	爱微剧场	古代 / 剧情	53
四人一崽闯末世	2025-11-18	5583.9w	九九AI漫解	架空 / 重生 / 玄幻仙侠 / 奇幻脑洞	37
我靠前世记忆改写末世结局	2025-11-11	5525.7w	AI漫说剧	奇幻脑洞 / 重生 / 架空	43
神君鬼夫	2025-12-20	5376.7w	桃桃剧场	剧情 / 古代 / 玄幻仙侠	17
大后传第1部	2025-12-21	5079w	牧神制片	玄幻仙侠 / 古代 / 异能 /萌宝	70
穿书修仙世界，全宗门都在偷听我的心声	2025-12-06	5071.2w	爱微剧场	剧情 / 穿越 / 玄幻仙侠 / 异能 / 萌宝	54

资料来源：独舌短剧公众号，中邮证券研究所

4.2 影视：短剧全场景赋能在即，长剧/影视高价值环节有望优先受益

长剧/电影：CG特效等高价值场景或为率先替代环节，中长尾厂商有望借助技术平权成为核心受益方

- **短期CG特效等技术密集环节有望迎来AI替代，全流程生成仍待发展。**尽管目前已有多部AI原生影视作品落地，但我们认为距离全流程AI生成的规模化商用尚有距离，主要系：1）影视制作并非所有环节都具备AI替代价值，在部分布景简单、实拍效率较高的场景中，传统拍摄依然具备性价比优势；2）目前长剧生成需依赖多工具串联使用，操作门槛仍然偏高，虽然部分厂商已开始提供解决方案，但距离通用化、规模化交付尚需时间。短期来看，AI更有望率先应用于降本增效价值突出的环节，如CG特效等。
- **从使用主体来看，头部厂商短期可能仍将以传统制作模式为主，中长尾影视公司或为主要应用方。**CG特效历来是影视工业中技术壁垒最高的环节之一，也是头部厂商构筑核心竞争力的重要手段。以《阿凡达3》为例，其全片包含超过3,500个视效镜头，涉及大量火焰、爆炸等高复杂度物理效果。尤其在灰烬渲染环节，卡梅隆团队打造了每秒可处理2.3亿个灰烬粒子的渲染系统，每个粒子均携带独立的光谱参数与运动轨迹，短期内AI生成技术仍难以与之匹敌。因而对头部厂商而言，目前AI替代价值仍相对有限。但是对于预算有限、缺乏特效制作能力的中小影视厂商，凭借成本、效率与可获得性上的优势，能够让其具备生成复杂视效画面的能力，有望成为AI特效生成应用的主要群体。

图表79：《阿凡达》动捕现场

资料来源：世界互联网大会公众号，中邮证券研究所

图表80：《阿凡达》系列视效技术仍处行业顶级，短期AI难以与之匹敌

资料来源：世界互联网大会公众号，中邮证券研究所

4.3 游戏：短期创作环节持续降本增效，长期新品类有望拓宽产业边界

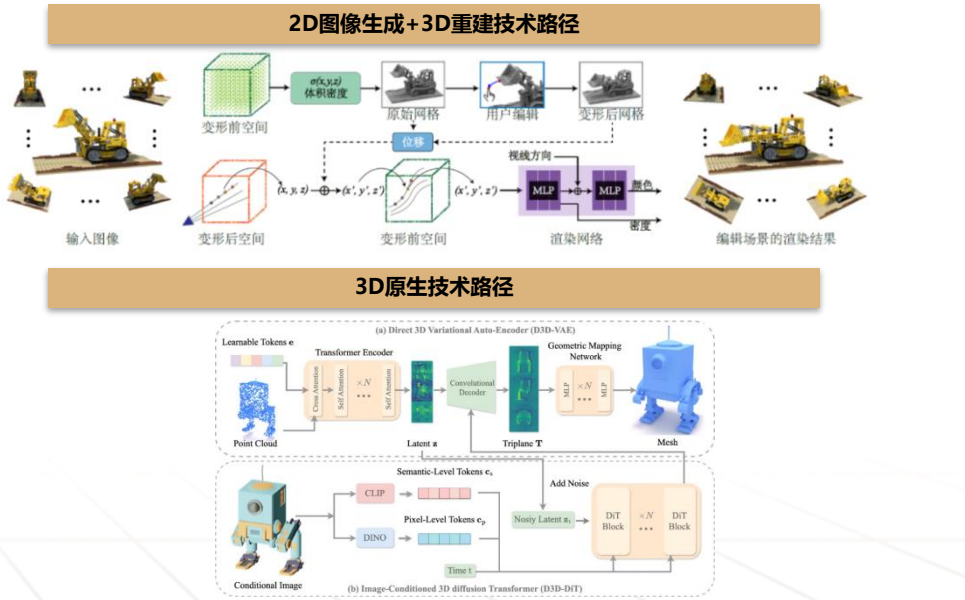
- **视频生成与3D生成在技术路径上具有同源性，其底层均依赖扩散模型与Transformer等生成式架构的发展。**当前3D资产生成主要可分为两类：
  - **2D图像生成+3D重建**：先生成一组多视角图片，再通过3D重建为三维资产。前端图像生成沿用了目前图像/视频生成模型的扩散模型体系，区别在于后端额外引入3D重建模块。当前重建技术主要为NeRF，本质是通过多视角图像学习一个连续的体渲染函数，视觉连续性与真实感较强。但其结果以隐式表示存在，难以直接编辑或用于复杂交互场景。行业内目前有厂商开始尝试以3D Gaussian Splatting替代传统NeRF，即通过在三维空间中显式构建高斯分布点集，使生成结果能够直接转化为可用于后续渲染、动画等场景的通用3D资产。例如前文提及的Marble模型，其输出即可直接导出为3D Gaussian Splatting数据；
  - **原生3D生成：**核心思路是跳过中间形态，将原本DiT架构中的2D训练数据替换为3D数据，使模型直接学习和生成三维结构本身，在几何还原度、多视角一致性以及生成效率方面具备更高的理论上限。但受制于3D训练数据稀缺，当前仍处于早期阶段。

图表81： 3D生成主流技术路径

技术路径	技术原理	优势	劣势	代表产品
2D图像生成+3D重建	先生成一组多视角图片，再通过3D重建为三维资产，**前端图像生成沿用了目前图像/视频生成模型的扩散模型体系，区别主要在于后端需额外引入3D重建模块**	可以利用大量的2D图像数据进行预训练，数据的丰富性使生成的3D模型复杂度提高，富有“想象力”	1、生成速度慢；2、生成质量较低；3、存在兼容性问题	Dreamfield、Dreamfusion（Google）、Point-E（OpenAI）、Magic3D（Nvidia）、ProlificDreamer（生数科技）、One-2–3–45等
原生3D	使用3D数据集进行训练，从训练到推理都基于3D 数据，**通常也是基于diffusion模型和transformer模型的方法进行训练。**	1、生成速度快；2、生成质量高；3、兼容性好	**丰富性不足：**缺乏高质量、大规模的3D数据集，目前比较大的3D数据集基本在百万级别，相比于十亿级别的图像数据集有三个数量级的差距，并且数据质量和一致性较差，制约了模型的“想象力”	Get3D（Nvidia）、Shap-E（OpenAI）、Dreamface（影眸科技）等

资料来源：MIT科技评论，极客公园，中邮证券研究所整理

图表82： 3D生成技术路径架构



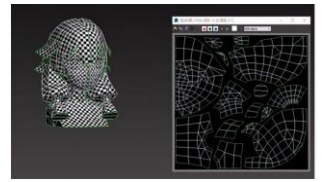
资料来源：华尔街见闻，Jittor，中邮证券研究所

4.3 游戏：短期创作环节持续降本增效，长期新品类有望拓宽产业边界

短期：3D生成逐步替代传统建模，成本端有望持续降本增效

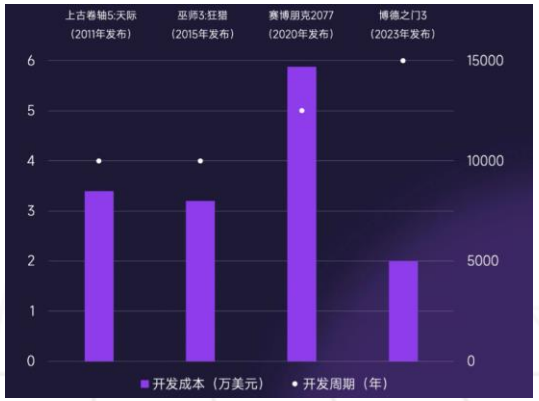
- **传统游戏建模流程高度依赖人工经验，整体链路复杂、周期长、成本高。**在常规流程中，建模师需先进行原画定稿，继而开展模型搭建，再完成UV、贴图与纹理烘焙等环节，流程高度依赖人工与美术经验，同时需要投入大量资金资源。诸如《上古卷轴5》《巫师3》等3A游戏作品，其整体研发投入普遍可达5,000万美元以上，《赛博朋克2077》的研发成本更是接近1.5亿美元。同时，此类项目的开发周期通常长达4~6年，回本周期缓慢。
- **3D生成技术在游戏资产制作环节具备明确降本增效价值。**以《赛博朋克2077》为例，其单一场景区域的建模通常需要约2~3个月完成，AI 3D建模则可在数分钟内生成雏形。此外，3A游戏中单一角色资产的建模通常需要3~5人协同完成，人力成本占整体项目预算的比例能达到25%~35%。AI 3D生成工具能够替代大量重复性工作，使人力投入规模减少约60%~70%。同时，在流程自动化层面，AI 3D建模在自动化率可超过70%，亦能有效减少人工操作与反复修改需求，从而进一步提升整体制作效率与成本可控性。

图表83： 传统游戏建模流程

原画定稿	模型搭建	UV 拆分	贴图烘焙
			将高模细节映射到低模表面的主要方式，烘焙之后的低模贴图将直接影响模型视觉质量与渲染性能
1）对模型进行整体规划，决定UV 贴图数量、公用部分、UV 朝向等问题； **2）对模型的、材质进行分析，**比如毛发，布料，皮革，金属等； **3）对原画结构进行分析，**查看模型穿插是否合理，模型结构是否准确，不同模型之间的比例大小等	进行高、中、低模的搭建，其中：**1）中模：**高模和低模的基础，主要用于确定比例、形状等；**2）高模：**确定纹理细节；**3）低模：**由高模拓扑或者中模减面而来，用于减少计算量、提高渲染速度，确保游戏在各类设备上的运行 **涉及工具**：Zbrush等	对模型进行规划，按照模型的材质(金属/布料)、功能(可动部件/静态部件)、角度(夹角大于 90°)等规范把模型拆分成独立的 UV 壳 **涉及工具**：RizomUV等	**涉及工具**：Marmoset Toolbag等

资料来源：《传统次世代建模流程和AI生成建模流程的对比（陈娜娜著）》，中邮证券研究所整理

图表84：传统游戏研发支出及开发周期



资料来源：量子位，中邮证券研究所

图表85：AI建模对比传统模式

流程类型	次世代建模流程	AI 建模流程
建模周期	单场景约 2–3 个月	单模型生成仅需几分钟
自动化率	小于 30%（UV 展开、贴图烘焙等环节需人工干预）	大于 70%（端到端生成模型 / UV / 材质）
核心技能	美术功底 + 软件精通	提示词设计 + 质量控制

资料来源：《传统次世代建模流程和AI生成建模流程的对比（陈娜娜著）》，中邮证券研究所

4.3 游戏：短期创作环节持续降本增效，长期新品类有望拓宽产业边界

- **当前3D生成技术已能够较好满足游戏中基础静态资产的生成需求，并开始以工具化形态嵌入实际生产流程，应用阶段正由“能力验证”向“实用落地”过渡。**以混元3D为代表，其生成能力已在《元梦之星》等数十款腾讯内部游戏接入与应用，全球首个设计Agent Lovart在3D生成任务中优先选择腾讯混元3D，头部3D打印厂商如拓竹科技、创想三维等亦已接入混元3D模型。此外，其他厂商亦在推动自身技术走向实用落地。例如《蛋仔派对》已与影眸科技达成业务合作，通过引入其生成算法Rodin，使玩家能够在游戏创作界面中直接使用AI生成符合需求的各类物品。
- **动态物体与场景资产的生成亦在加速探索，后续技术落地将持续推动游戏制作环节的降本增效。**1）动态资产：相关探索已开始出现实质性进展，已有模型尝试面向动态人体与动作序列进行生成。例如HY-Motion1.0已能够生成覆盖移动、竞技、健身等多类复杂动作场景的运动人体。后续相关技术有望逐步向游戏中角色动画、可运动道具等动态资产环节渗透；2）场景资产：世界模型的发展有望提供新技术基础，并推动3D生成能力从“单一资产”进一步向“场景级组合与演化”延伸。

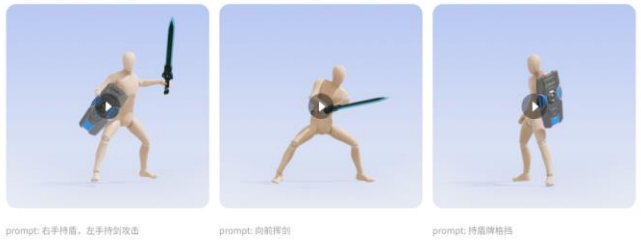
图表86： 3D模型已逐步应用于游戏的静态资产生成

《蛋仔派对》已在游戏中应用3D生成技术

资料来源：游戏研究社公众号，中邮证券研究所

图表87：动态物体与场景资产生成亦在加速探索

动态资产



混元Motion1.0

采用DiT架构，结合流匹配目标函数，探索模型与数据规模扩展在动作生成领域的潜力，能够生成覆盖移动、竞技、健身等多类复杂动作场景的运动人体，并满足游戏场景中人物的动作要求

场景资产

Matrix-game 2.0

基于昆仑万维世界模型架构打造，该模型能够在生成过程中同步构建游戏场景与可操作物体，使生成系统不仅输出静态资产形态，还具备对物体状态与交互关系的统一建模能力，为场景级、交互型资产生成提供了可探索方向

资料来源：《Matrix-game 2.0: An open-source real-time and streaming interactive world model》，腾讯混元，第一财经，中邮证券研究所

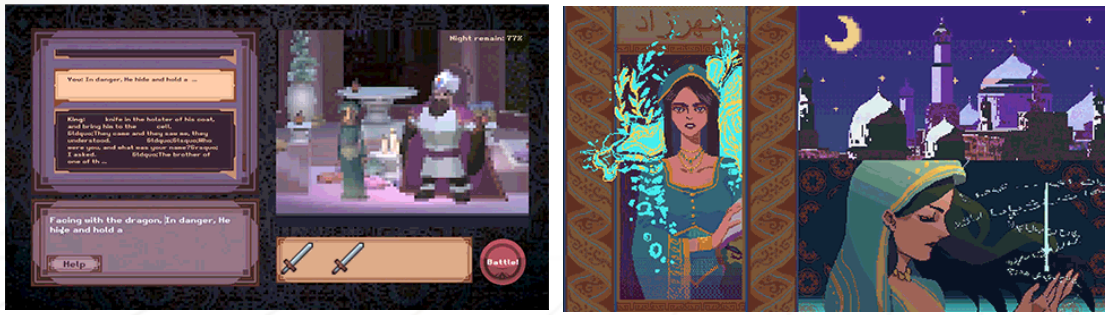
4.3 游戏：短期创作环节持续降本增效，长期新品类有望拓宽产业边界

长期：新游戏品类有望诞生，收入端边界或将拓宽

- **技术持续演进有望推动游戏内容从“线性玩法+静态资产”向“实时交互+生成式世界”演进，并催生3D交互新游戏品类。**当前主流游戏内容依赖开发者预设，玩家在封闭场景中按线性流程完成任务，内容供给权集中于开发团队。而随着3D生成与自然语言交互能力的持续提升，开发者无需预制全部资产和逻辑结构，系统可根据玩家输入或行为按需生成环境、角色与交互规则，使游戏世界具备“实时生成+自主演化”的能力，有望催生空间建造、动态叙事等新兴品类产生。尤其是在头戴显示设备等沉浸式平台上，3D 生成能力或将在空间场景构建、实时物体生成等方面展现应用价值，进一步拓宽内容边界与产业空间。
- **AI原生交互已率先落地于多种品类的文字游戏，验证了动态交互驱动的玩法可行性。**自GPT-2发布以来，已有一批早期产品探索将语言模型能力用于叙事与世界交互类游戏，其中包括《AI Roguelite》《AI Explorer》《Dreamio》《AIdventure》《The Infinite Story》《AI Tales》等典型代表。2025年3月15日，米哈游蔡浩宇旗下AI创业公司Anuttacon的首款游戏《Whispers From The Star》亦正式开启海外封闭测试招募。除文字冒险与叙事类产品外，当前亦有大量基于AI模型驱动的养成/社交、RPG、Roguelite等品类涌现。

图表88：AI交互游戏《1001夜》展示

游戏画面



基于《一千零一夜》改编，游戏中，玩家可以与人工智能共同创作故事。

资料来源：Fakecheese，中邮证券研究所

图表89：AI原生交互已率先落地于多种品类的文字游戏

游戏名称	游戏公司	游戏类型	原生 AI 玩法说明
《Source of Madness》	Carry Castle	AI Roguelite	一款 AI 原生的类 Roguelite 游戏，内嵌子生成式 AI，玩家挖掘了 AI 设定关卡的能力，游戏中的怪物及地图结构会在每局随机产生，玩家甚至无法在游戏中重遇两个完全相同的敌人。
《1001夜》	Ada Eden	AI 叙事 / 文字冒险	灵感来源于文学作品《一千零一夜》，玩家需以上帝的身份，通过讲述故事来影响游戏中角色的言行选择，使角色使用合适的武器词汇（如“刀”“剑”等）一旦成功，玩家能获得得属性武器道具，用于后续的战斗。
《太空杀》	巨人网络	AI 推理	升级版 AI 角色挑战，AI 扮演内鬼，玩家通过语言向 AI 找出真凶。真人玩家与 AI 进行实时博弈、推理与对抗。AI 的发言和决策均由大模型实时生成，具备骗术、协作、伪装等高阶行为表现。
《终外就医》	喵吉拉	AI 解谜	以精神世界探索为核心的 AI 驱动剧情解谜游戏。AI 会根据玩家提问和选择动态生成角色反应、情节推进与交互（如误诊店老板、服务员等）并串联关键线索，解开逻辑复杂的谜题。
《Pick Me Pick Me》	Optillusion	AI 社交	聚焦人设生成与决策博弈的社交派对游戏，结合了 AI 对话互动和卡牌策略玩法。玩家通过对话与决策争夺 AI 的青睐，最后由 AI 判定胜负。
《萌派对》	喵吉拉	AI 社交 + 养成	宠物养成 + 社交玩法。在传统养宠基础上，AI 伙伴不仅可互动、陪玩，还具备自主记忆与成长行为。玩家可与 AI 建立情感关系，形成深度陪伴机制。

资料来源：游戏陀螺公众号，中邮证券研究所

5

核心受益上市公司

5 核心受益上市公司

- **投资建议：**Sora 2等代表性模型在拟真度、时长、叙事一致性等维度的跨越式突破，显著提升视频模型的表达力，重新定义了内容生产的边界，推动视频生成模型的发展迈入关键拐点。当前，AI视频生成技术已在广告创意、短剧生成、游戏建模等多个垂直场景中实现初步落地，并开始显现商业化路径。未来，伴随生成质量的持续提升与端到端视频生成范式的成熟，视频模型有望成为内容产业的新型基础设施，引领下一轮生产力变革。基于此，我们认为以下四类公司有望优先受益：
  1. **具备自研算法与模型能力，且具有多场景业务嵌合能力的技术型公司：**昆仑万维（以“天工”大语言模型为底座，构建覆盖视频SkyReels、音乐Mureka、智能体R1V4等多模态AIGC场景的能力体系，前瞻布局世界模型“Matrix”系列。其中，Matrix-3D在全景视频生成任务中已取得SOTA成绩，技术稀缺性与领先优势显著）；
  2. **拥有海量内容资产与版权资源的影视内容提供商：**中文在线、捷成股份、华策影视；
  3. **积极布局AI营销、具备内容分发的整合型平台公司：**易点天下、蓝色光标；
  4. **推动AI生成能力嵌入游戏资产生产流程的大型游戏公司：**完美世界、巨人网络。

6

风险提示

- AI视频生成技术发展不及预期
- 产业应用不及预期
- 版权保护风险

中邮证券
CHINA POST SECURITIES

感谢您的信任与支持！

THANK YOU

证券分析师 **王晓萱**

E-MAIL **wangxiaoxuan@cnpsec.com**

证书编号 **S1340522080005**

免责声明

分析师声明

撰写此报告的分析师（一人或多人）承诺本机构、本人以及财产利害关系人与所评价或推荐的证券无利害关系。

本报告所采用的数据均来自我们认为可靠的目前已公开的信息，并通过独立判断并得出结论，力求独立、客观、公平，报告结论不受本公司其他部门和人员以及证券发行人、上市公司、基金公司、证券资产管理公司、特定客户等利益相关方的干涉和影响，特此声明。

免责声明

中邮证券有限责任公司（以下简称“中邮证券”）具备经中国证监会批准的开展证券投资咨询业务的资格。

本报告信息均来源于公开资料或者我们认为可靠的资料，我们力求但不保证这些信息的准确性和完整性。报告内容仅供参考，报告中的信息或所表达观点不构成所涉证券买卖的出价或询价，中邮证券不对因使用本报告的内容而导致的损失承担任何责任。客户不应以本报告取代其独立判断或仅根据本报告做出决策。

本报告所载的意见、评估及预测仅为本报告出具日的观点和判断。该等意见、评估及预测无需通知即可随时更改。过往的表现亦不应作为日后表现的预示和担保。在不同时期，中邮证券可能会发出与本报告所载意见、评估及预测不一致的研究报告。

中邮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头寸并进行交易，也可能为这些公司提供或者计划提供投资银行、财务顾问或者其他金融产品等相关服务。

《证券期货投资者适当性管理办法》于2017年7月1日起正式实施，本报告仅供中邮证券签约客户使用，若您非中邮证券签约客户，为控制投资风险，请取消接收、订阅或使用本报告中的任何信息。本公司不会因接收人收到、阅读或关注本报告中的内容而视其为签约客户。

中邮证券对于本申明具有最终解释权。

免责声明

中邮证券 CHINA POST SECURITIES

公司简介

中邮证券有限责任公司，2002年9月经中国证券监督管理委员会批准设立，是中国邮政集团有限公司绝对控股的证券类金融子公司。

公司经营范围包括:证券经纪，证券自营，证券投资咨询，证券资产管理，融资融券，证券投资基金销售，证券承销与保荐，代理销售金融产品，与证券交易、证券投资活动有关的财务顾问等。

公司目前已经在北京、陕西、深圳、山东、江苏、四川、江西、湖北、湖南、福建、辽宁、吉林、黑龙江、广东、浙江、贵州、新疆、河南、山西、上海、云南、内蒙古、重庆、天津、河北等地设有分支机构，全国多家分支机构正在建设中。

中邮证券紧紧依托中国邮政集团有限公司雄厚的实力，坚持诚信经营，践行普惠服务，为社会大众提供全方位专业化的证券投、融资服务，帮助客户实现价值增长，努力成为客户认同、社会尊重、股东满意、员工自豪的优秀企业。

投资评级说明

投资评级标准	类型	评级	说明
报告中投资建议的评级标准：报告发布日后的6个月内的相对市场表现，即报告发布日后的6个月内的公司股价（或行业指数、可转债价格）的涨跌幅相对同期相关证券市场基准指数的涨跌幅。 市场基准指数的选取：A股市场以沪深300指数为基准；新三板市场以三板成指为基准；可转债市场以中信标普可转债指数为基准；香港市场以恒生指数为基准；美国市场以标普500或纳斯达克综合指数为基准。	股票评级	买入	预期个股相对同期基准指数涨幅在20%以上
		增持	预期个股相对同期基准指数涨幅在10%与20%之间
		中性	预期个股相对同期基准指数涨幅在-10%与10%之间
		回避	预期个股相对同期基准指数涨幅在-10%以下
	行业评级	强于大市	预期行业相对同期基准指数涨幅在10%以上
		中性	预期行业相对同期基准指数涨幅在-10%与10%之间
		弱于大市	预期行业相对同期基准指数涨幅在-10%以下
	可转债评级	推荐	预期可转债相对同期基准指数涨幅在10%以上
		谨慎推荐	预期可转债相对同期基准指数涨幅在5%与10%之间
		中性	预期可转债相对同期基准指数涨幅在-5%与5%之间
		回避	预期可转债相对同期基准指数涨幅在-5%以下

中邮证券研究所

北京
邮箱：yanjiusuo@cnpsec.com
地址：北京市东城区前门街道珠市口东大街17号
邮编：100050

上海
邮箱：yanjiusuo@cnpsec.com
地址：上海市虹口区东大名路1080号大厦3楼
邮编：200000

深圳
邮箱：yanjiusuo@cnpsec.com
地址：深圳市福田区滨河大道9023号国通大厦二楼
邮编：518048


中邮证券
CHINA POST SECURITIES