AI视频生成研究报告

量子位智库 insights

分析师：Xuanhao
xuanhao@qbitai.com

2024. 7

1. 技术侧

2. 应用侧

3. 玩家格局

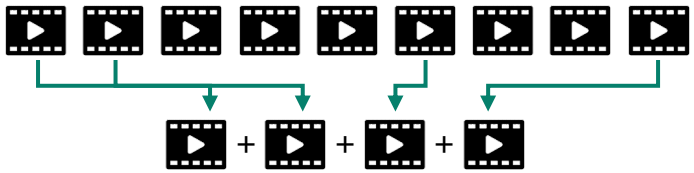
大模型各模态总览：多模态发展趋势清晰，文本、图像商业化规模和成熟度较高，AI视频生成正在迅速发展

	概况	关键节点	代表应用	成熟度
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，例如ChatGPT日活用户已经突破1亿，OpenAI在2024年6月ARR的达到34亿美元**	• 2018年6月，由Alec Radford主导在OpenAI推出GPT-1 • 2020年6月，OpenAI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	✓✓
图像	• 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，OpenAI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（重点讨论！）	• 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年下半年，创业公司推出Runway-Gen2，Stable Video Diffusion、Pika等产品 • 2024年2月，OpenAI发布Sora引发全球关注	• Sora • Runway • 快手可灵 • Pixverse	✗
音频	• 目前主要是音乐生成（语音识别、克隆暂不纳入讨论），市场不如图片生成、视频生成等领域热门，比视频更加早期 • **明星创业公司较少，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 Open	• Suno • Stable Audio	✗
3D	• 技术路线目前尚不清晰，垂直明星创业公司较少，产品大多处于早期阶段，但正在加速发展	• 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Meta 3D Gen	• Luma.AI • Meshy	✗

信息来源：量子位智库

技术趋势：视频生成正在由检索生成、局部生成走向依靠自然语言提示词的全量生成，生成内容更加灵活丰富，应用空间广阔

检索生成



- 检索生成主要是对现有的视频素材根据关键词和标签进行检索匹配，再进行相应的拼接和排列组合

特点

- 采用传统的跨模态视频检索技术，通过视频标签的或者视频语义理解的方式从数据库中的检索，再将这些素材进行剪辑、组合拼接在一起，**本质上还是键值对匹配的逻辑**
- 例如短视频平台的知识类视频、解说类视频，通过文本关键字在数据库中进行素材检索，然后在进行拼接组合生成
- **创意空间有限，没有贡献增量素材，但成本极低，生成速度极快**

局部生成

- 仅针对视频的一部分进行生成，例如视频中人物角色、动作、背景、风格化、特殊效果等

特点

- 采用传统的计算机视觉（CV）、计算机图形学（CG）技术，但生成功能有限，主要是一些局部的垂点功能
- 例如效果生成，在现有视频上添加多种效果，如滤镜、光影、风格化、美颜特效等。也可以做局部动态化，如人物的面部表情生成、搞笑表情、爆款特效、舞蹈动作生成等
- **有一定创意空间，生成部分新元素，成本低但应用的场景有限**

提示词生成

重点讨论！

- 通过文字、图片、视频作为提示词来进行凭空生成，不依赖外部素材，核心在于大模型的能力

特点

- 采用基于Transformer或者扩散模型的大模型路线，可以通过自然语言或者指导图进行全局生成（但也可以嵌入已有内容），视频的内容、风格、长短、分辨率、宽高比都可以进行灵活调整
- 例如生成天马行空的创意视频、艺术视频、卡通视频等等，非常灵活
- **创意空间无限，所有的元素都是全新生成，现阶段成本高昂，但天花板高，应用场景广泛**

无新增内容 → 部分新增内容 → 全量新增内容

信息来源：量子位智库

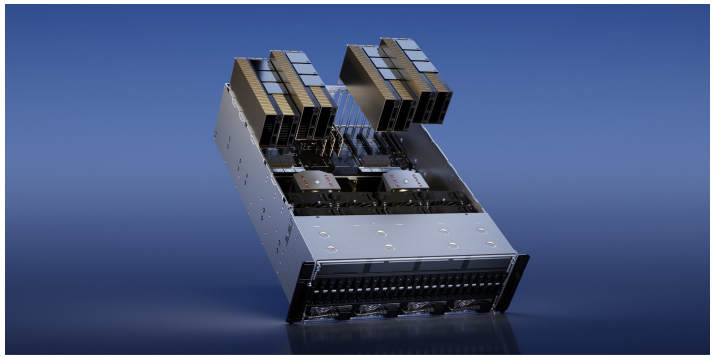
技术趋势：视频生成正由扩散模型主导的格局走向与语言模型结合的路线，Transformer将在视频生成方面发挥主导作用

	探索期	早期	增长期	未来方向
	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Video Vision Transformer	• 2020年 DDPM论文发表 • 2021年 Latent Diffusion Models论文发表 • 2022年 Video Diffusion Model 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	• GODIVA、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	• Stable Video Diffusion、Make-A-Video、Runway Gen1/2、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[1]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）

信息来源：量子位智库，1）在视频生成的语境下，主要指基于Transformer这样的自回归模型

技术挑战：算力需求大，数据要求高，算法复杂是目前制约视频生成模型能力的三大挑战

1 算力需求大



AI计算卡示例

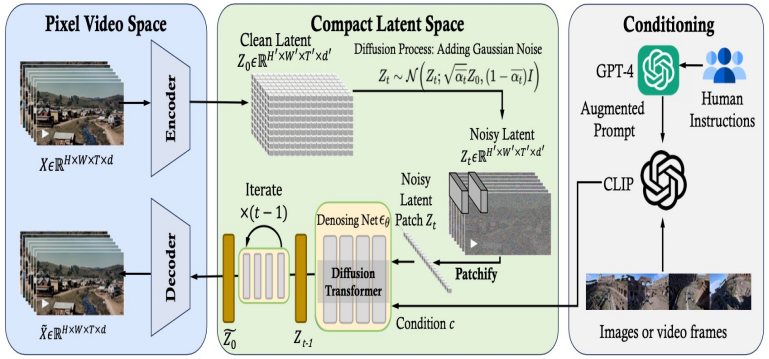
- 训练视频生成大模型所需要的计算量远高于一般的文本和图像模型，这导致开源社区和学术界等相对业界算力不足的玩家难以参与，学界在视频基础模型上工作较少，**相关的模型和科研成果多出自互联网公司和主打视频生成的商业技术公司**
- 以Sora为例，从训练侧看，训练成本大约为数十万英伟达H100 GPU hours（据估算），**需要千卡GPU的计算集群**，以H100的使用价格约为3$/h估算，Sora的训练成本可能达数千万至上亿美元
- 从推理侧看[1]，价格方面目前Sora每分钟的推理成本约数十美元，成本高昂；生成时长方面，单个视频生成时长超过10min，推理速度很慢

2 数据要求高

视频训练数据示例

- **高质量数据少**：最佳的训练数据是高质量的视频-文本对，即针对一段视频，有与之对应详细准确的文字描述，互联网上大部分的视频数据都难以满足需求（如数据不准确甚至是错的），此外视频数据的宽高比、分辨率、时长各异，需要进一步处理。数据量方面，Sora 的训练数据可能超过500万小时的精良视频
- **公开数据质量低**：公开数据集例如WebVid（1070万个文本视频对，仅5.2万小时）、HowTo100M总时长超10万，但都是4s的短视频）、CelebV-Text（超7万个人脸-文本片段描述），数据量小且质量低
- **版权数据获取难**：例如电影、记录片、动漫、MV等影视作品，内容平台版权库，以及YouTube、抖音等UGC内容，成本高且有版权限制

3 算法复杂



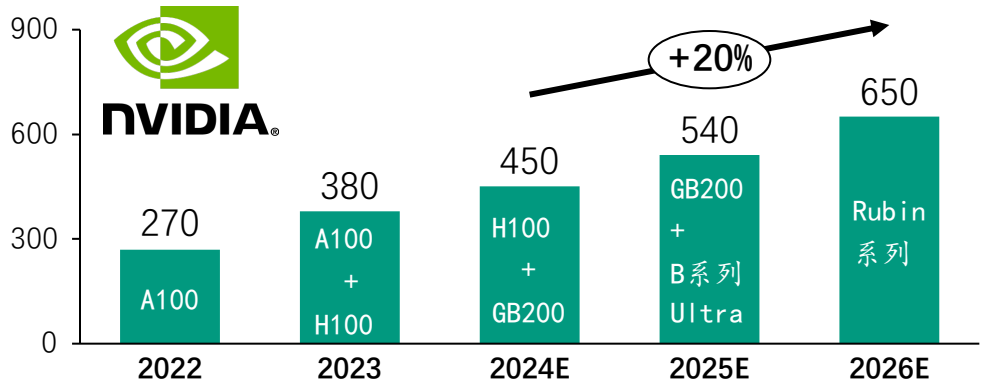
Sora架构（推测）

- **时间维度增加复杂性**：视频生成在图像的基础上增加了时间维度，例如针对时间维度和空间维度结合做数据表示，这对可扩展性、视频生成的时长和生成效果一致性方面有重大影响
- **视频生成更难规模化（scale）**：对于语言模型而言，可以进行大规模的自监督学习，而图像和视频生成模型需要进行图像-文本或视频-文本对标注做监督学习，规模化的难度更大，这是视频模型和LLM的本质差异
- **Tokenizer设计更复杂**：文本模态的tokenizer更成熟，语言已经过人类智能的一次压缩，但图像是现实世界的原始信息，信息密度较低，需要重新设计更好的tokenizer

信息来源：量子位智库，OpenAI，英伟达，《Sora: A Review on Background, Technology, Limitations, and Opportunities of Large Vision Models》，1）推理速度除模型外也看算力供给

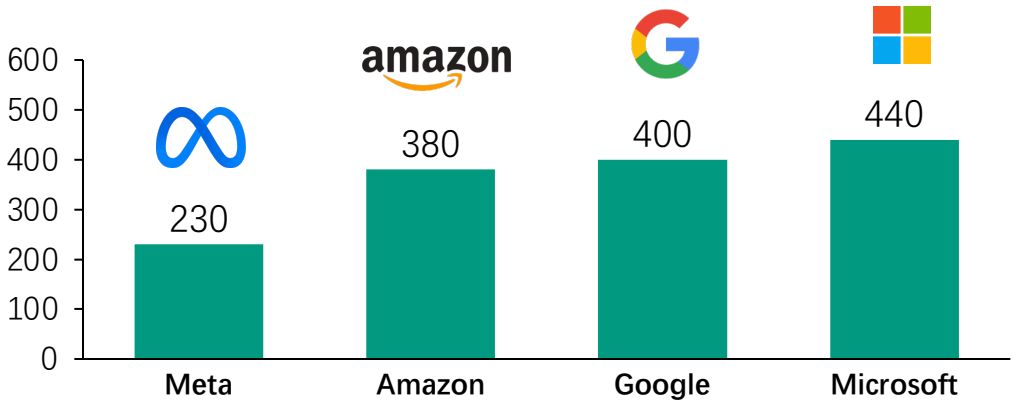
技术驱动：视频生成模型的能力将在算力驱动下快速进步，稳定性、可控性、丰富度将持续提升，解锁更多应用空间

1 英伟达人工智能GPU出货量持续增加（万张）



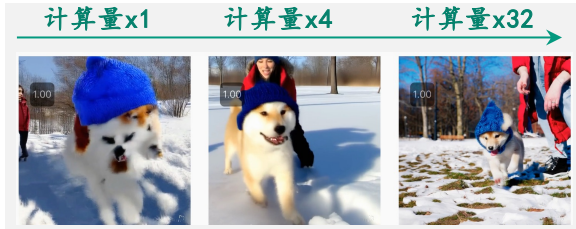
- **英伟达目前占据全球高端GPU市场超95%的市场份额**，是事实上生成式AI全行业算力市场供给量的决定者
- 英伟达的人工智能GPU在2022年出货量约270万，主要以A100为主；2023年出货量约380万块，主要以A100和H100为主；预计2024年出货量可能达到450万块，以H100和最近发布的Blackwell系列为主
- 预计英伟达GPU交付量将保持20%的增速，并在每年进行芯片架构的升级，稳步提升芯片和系统的计算能力

2 北美巨头数据中心2024年Capex投入预期（亿/美元）



- 目前北美最头部的科技公司都在重金押注AI数据中心建设，以规模最大的4家公司为例，预计2024年在数据中心上的投入将达到1500亿美元左右，在一定程度代表了领军玩家对于大模型的信心和预期
- 云厂商旗下的数据中心既支持内部业务需求，也对外提供GPU算力，**将逐步满足市场对于算力的需求，支持各类模型的训练和推理**
- 数据中心成本主要包括AI服务器的采买、土建成本、电力系统、制冷系统、监控系统等

关键分析



OpenAI Sora 示例

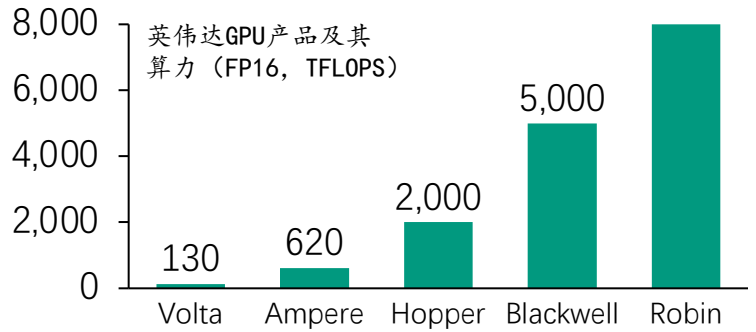
- 从OpenAI Sora的实践成果来看，继续增加模型的数据量和相应的参数规模（Scale up）依然是AI发展的核心路线，强大算力支持是模型进步的必要支撑
- **模型能力**：scale之后可以涌现出更多高级特性，例如：1）随着镜头的旋转和移动，人物、对象、场景在三维世界中保持稳定真实，2）模型可以模拟距离关系和空间关系，生成针对一个角色的多个镜头，3）模拟生成内容中的物理交互关系
- **应用成本**：视频模型的推理成本较高，需要大规模的推理算力来支持市场的大规模应用，充沛的算力供给将推动视频生成从实验阶段推向商业化普及

信息来源：量子位智库，Morgan Stanley

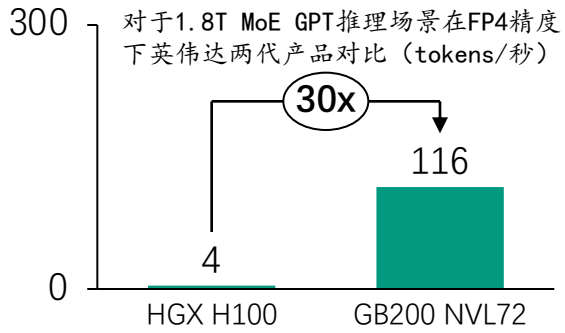
技术驱动：视频生成的推理成本将持续下降，生成速度进一步提高，加速应用层技术扩散和商业化规模增长

1 硬件的计算能力、推理速度不断提升

A 芯片层性能提升

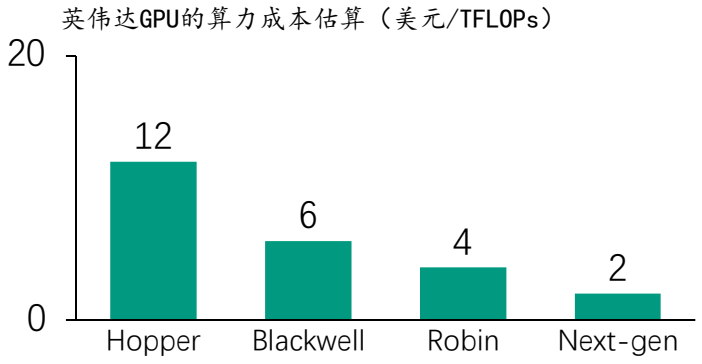


B 系统层性能提升

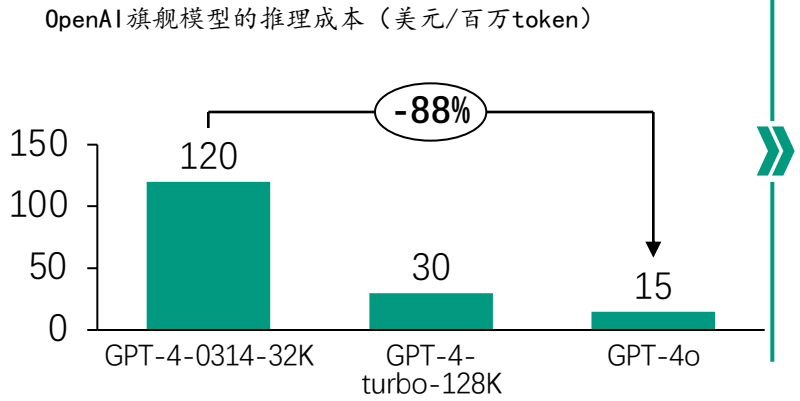


2 模型应用的成本将不断降低

A 芯片层成本优化



B 系统层成本优化



关键分析

- **当前痛点**：目前制约视频生成应用普及的重要因素之一是生成速度，生成5s左右的视频需要等待数分钟，且需要尝试多次才能获得理想结果，对用户体验造成的影响较大
- **加速生成**：单卡芯片算力提升和系统、集群上面的优化可以大幅增加模型推理速度（tokens/秒），缩短视频生成的等待时间

- **降本趋势**：视觉模型的价格快速下降尚未开始，但随着市场需求驱动和产品化的成熟，类似LLM的降价趋势也将出现在视频模型上
- **FLOPs成本下降**：单位计算量的成本将持续降低，主要来源于芯片架构的提升和服务器、数据中心系统优化
- **软件层优化**：从LLM来看，推理成本正在迅速降低，头部模型在过去一年降幅约90%，降本趋势将持续

信息来源：量子位智库，英伟达，OpenAI，Morgan Stanley

技术展望：视频生成模型不仅限于生成视频内容，长期将统一多模态的生成和理解，成为通向AGI的重要路径

Scaling Laws

GPT-3.5
语言模型

DALL-E
视觉模型

Whisper
音频模型

数据
算法
算力

GPT-4V
多模态理解

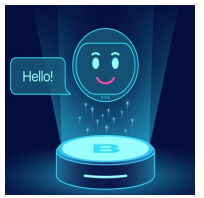
多模态生成
Sora

多模态大模型

“Everything to Everything”
统一生成和理解

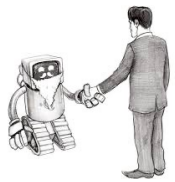
物理世界交互
- 具身智能
- 自动驾驶
- …

数字世界交互
- 智能管家
- 虚拟Agent
- …



System 2
- 复杂推理
- 长线规划
- …

超级对齐



早期阶段
- 内容生成和理解独立
- 不同模态相互独立

目前各个模态正在相互融合：
- Text-to-Video：OpenAI Sora打通文字与视觉
- Audio-to-Audio：GPT-4o实现音频交流
- Text-to-Audio：Suno打通文字与音乐
- Video-to-Audio：Deepmind发布匹配Veo的视频生音频技术

关键分析

- **生成和理解在本质上是统一的**，语言模型的next token prediction越准确，意味着模型对于语言和世界的理解越准确。对于视频模型，对下一帧或下一个patch[1]的预测的越准确，上代表了模型对物理世界的理解越准确
- **视频模态包含大量信息：**从仿生的角度看，人脑有80%的信息来自视觉，因此视觉信息的理解与生成对于多模态大模型至关重要
- **视觉模型可以压缩一切：**“当多模态训练达到一定规模时，语言智能就会融入到视觉智能中，这是一条获得世界模拟器的路径，可以通过这样的模拟器获得任何东西。”---Aditya Ramesh，OpenAI Sora及DALL-E项目负责人

信息来源：量子位智库，1）patch是OpenAI Sora中类似token的一种数据表示

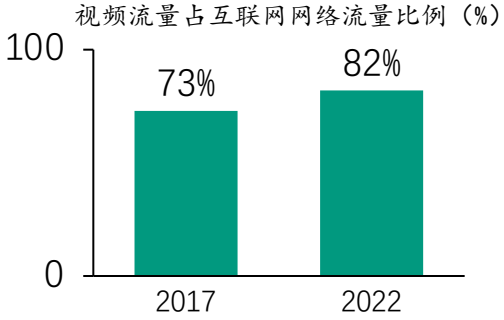
1. 技术侧

2. 应用侧

3. 玩家格局

场景广阔：互联网内容正在全面视频化，视频内容的消费场景丰富，AI视频内容生成的潜在市场规模巨大

1 视频流量是主要的互联网信息流量



- 2017到2022年，**全球互联网视频流量占消费互联网流量的比例从73%增长到82%**，成为流量最大的内容形式
- 2022年，每月有500万年的视频内容通过互联网传输。相当于每秒钟有110万分钟的视频被流式传输或下载

2 视频是移动互联网最大的内容消费形式

11亿人

64小时

- **内容视频化是大势所趋**，移动互联网的用户使用总时长占比中，短视频稳居第一达到28%
- 移动视频行业用户规模达10.76亿，月人均时长为64.2小时，视频平台成为流量核心，可以将用户引向电商、音乐、影视、本地生活、旅游服务等等垂直赛道

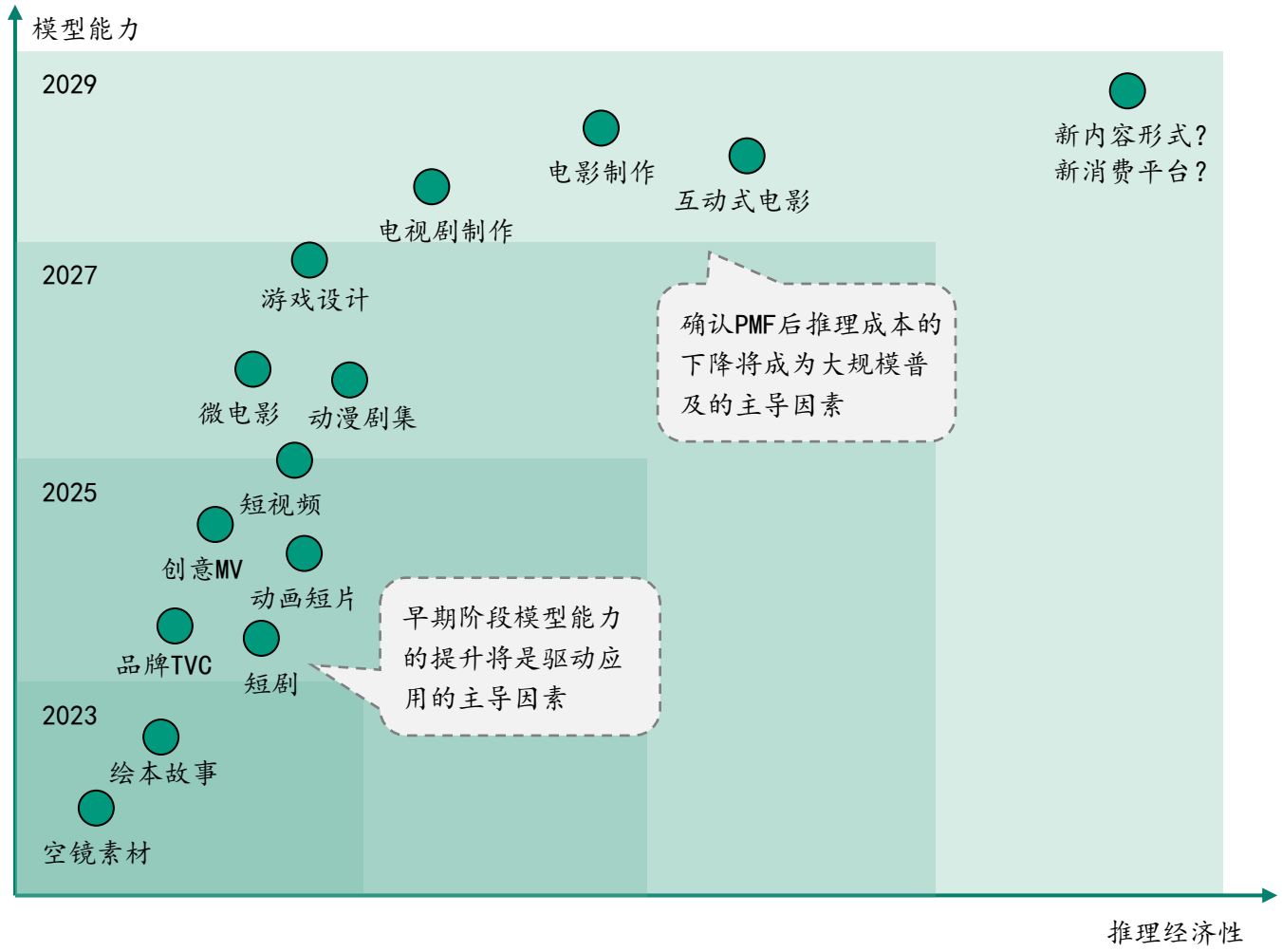
		长视频平台			短视频平台		电影市场（年度票房）
海外视频应用	平台	YouTube	Netflix	Disney+	Instagram	TikTok	
	用户	25亿MAU	2.7亿	1.5亿MAU	20亿MAU	16亿MAU	N/A
	年营收	315亿美元	340亿美元	84亿美元	100亿美元+	160亿美元	330亿美元（全球）
本土视频应用	平台	iQIYI爱奇艺	腾讯视频	bilibili	抖音	快手	
	用户	5亿MAU	4亿MAU	3.4亿MAU	8亿+MAU	7亿+MAU	N/A
	年营收	320亿	100亿+	230亿	1500亿	1135亿	550亿（本土）

关键分析

- **从消费端来看，视频是用户消费时间最长的内容形态**，有丰富的应用的场景和大型内容分发平台，长期或有诞生超级应用的机会
- 随着AI视频生成的能力不断提升，**AI生成视频占视频消费内容的比例将不断提升**，推动内容供给端变革，逐步渗透视频消费市场

信息来源：量子位智库，Questmobile，Cisco

应用趋势：2024年将成为AI视频的应用元年，未来3-5年更多应用场景将随着模型能力提升和推理成本下降逐步解锁



成熟阶段

- **模型能力**：通过自然语言及其他方式可以实现对内容的精准控制，深度理解物理世界规律，稳定性、丰富度达到在各个领全面达到商用水准。 1分钟的视频片段生成时间达到缩短到数秒，接近实时生成
- **经济性**：视频生成的成本继续降低1个数量级
- **产品**：新一代视频交互界面开始普及，**视频生成内容融入大部分视频制作场景，重塑内容生态**

发展阶段

- **模型能力**：实现复杂语义理解，同时满足多个生成条件，视频的活动度、丰富度、稳定性可以媲美影视级内容，有效时长超过一分钟，在部分场景可以充分满足需求。1分钟的视频片段生成时间缩短到到分钟级
- **经济性**：推理成本下降1个数量级
- **产品**：**视频模型与传统工作流进行深度集成**，同时萌生AI原生工作流，商业化规模达到Midjourney的水平

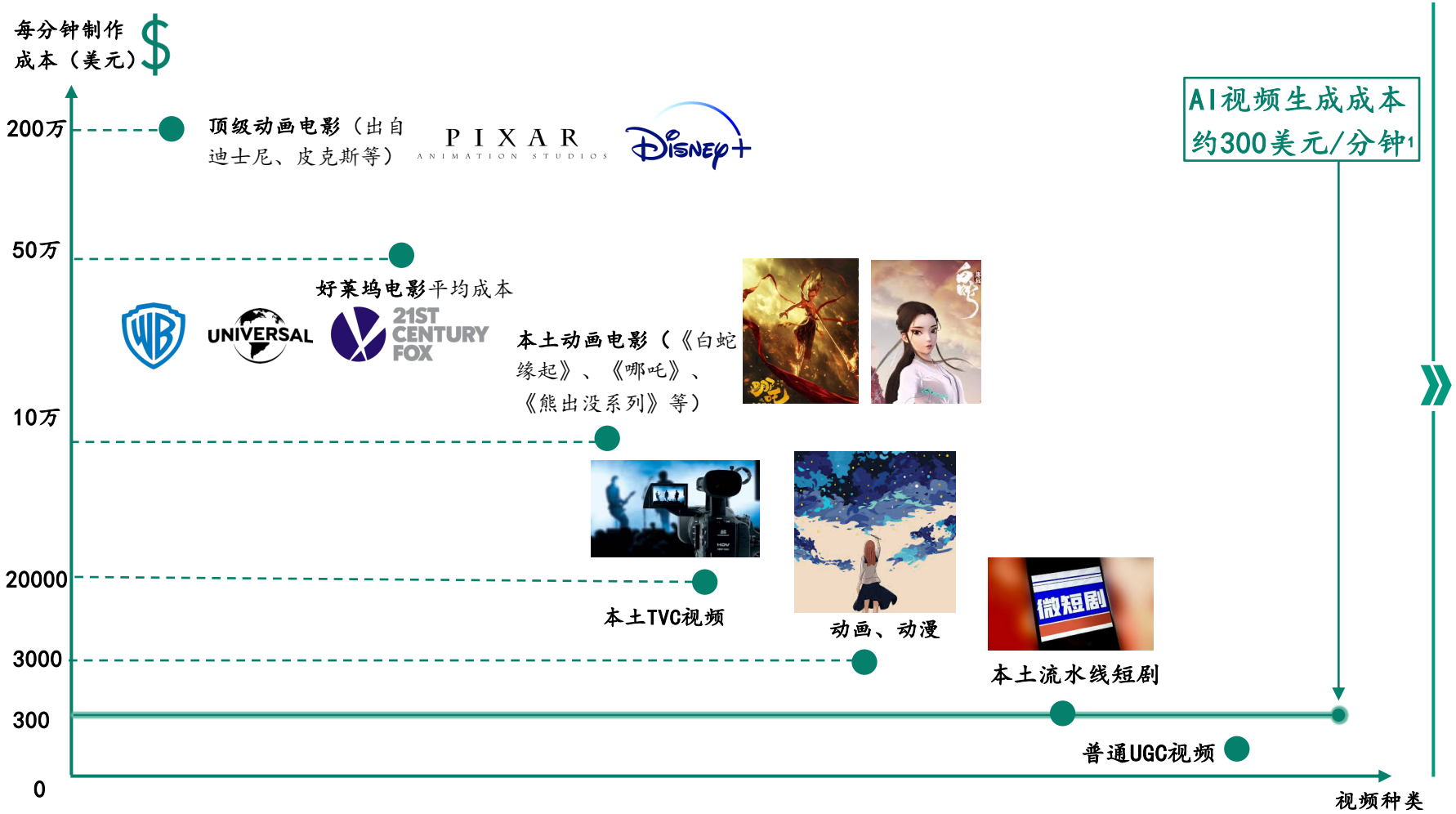
早期阶段

- **模型能力**：生成视频时长度短、活动度低、稳定差，人物对象、背景经常畸变。语言理解能力有限，只能理解简单指令，难以同时满足多个生成条件，指令遵从能力差，10秒左右视频片段需要3-5分钟生成
- **经济性**：成本高昂，每分钟视频约3美元[1]
- **产品**：**简单的文成视频、图生视频功能，以网页端和移动的简单应用为主，功能较为单薄**

信息来源：量子位智库，1）参考Runway的公开价格

成本驱动：AI生成视频的成本远低于各类现有视频内容的制作成本，将逐渐驱动视频生成内容渗透到各内容种类

不同类型视频内容制作成本



关键分析

- **AI视频生成的成本远远低于影视行业的制作成本**，有若干个数量级的降本效果，**但目前阻碍应用的主要因素是模型能力不足**，生成效果尚无法与传统制作方式竞争，**但预期模型能力将持续迭代**，未来3-5年达到可以与传统制作方式的媲美的水平
- **动画类电影的制作成本尤其高**，需要渲染大量的2D和3D内容，传统制作方式包括角色建模、场景贴膜、纹理贴图、渲染合成等环节，**需要数百人耗时数月进行制作，工程量非常大**，视频生成可以大量削减制作成本的
- **局部应用已经开始：**在对于制作质量要求较低、制作方式和内容较为模板化的短剧行业，已经出现AI短剧生成的应用，例如Reel.AI

信息来源：量子位智库，专家访谈，1）Runway目前的生成价格约每分钟3美元，按100:1的生成可用率计算，实际成本约每分钟300美元

应用案例-MV、品牌广告：Sora作为目前头部模型，在创意视频和品牌广告领域已具备应用价值，但仍存在诸多局限

1 创意短片《Air Head》--- 2024年3月

- 时长1分20秒，由Shy Kids团队3人花费2周时间完成制作，**总体呈现效果精良**

可控性差

- **抽卡率高，生成素材可用率约300:1**
- **片段间一致性差**：难以保证人物在不同视频片段之间的一致性，目前只能通过详细的提示词描述来弥补，但效果欠佳
- **镜头难以控制**：对于专业摄影术语理解有限，类似镜头平移的功能需要通过后期裁剪实现
- **生成稳定性低**：同样的提示词会产生不同的生成内容，例如要求生成黄色气球但实际生成式红色
- **生成能力局限**：生成的气球上总会有面部表情，需要后期抹除，不同片段画面风格难以保持一致，需要后期统一调色

版权限制

- 为了避免版权问题，OpenAI对提示词进行了限制，例如拒绝生成“35mm胶卷，未来宇宙飞船中，一名男子拿着光剑靠近”类似星球大战的提示词

生成速度慢

- 虽然Sora原生支持1080P视频生成，但由于生成速度太慢团队选择生成480P的视频，再用其他工具再后期进行超分处理，生成3-20秒的视频需要10-20分钟的生成时间（和云算力供给也有关）

2 品牌广告《玩具反斗城的起源》--- 2024年6月

- 时长1分06秒，由玩具反斗城团队和导演Nik Kleverov共同构思制作，并在戛纳国际创意节亮相
- Sora生成的第一个商业广告，**效果接近可以和传统品牌短片的水准**，可以传达品牌方的关键视觉元素和风格

不足之处

- 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）
- 背景元素存在畸变，例如背景中的自行车的有畸变特征

信息来源：量子位智库，量子位，Shy kids，Toys"R"Us

应用案例-短剧、动画：井英科技发布AI短剧APP Reel.AI，自研短剧视频生成模型Reel Diffusion，生成效果接近可消费水平

Reel.AI

（井英科技生成的AI短剧）

30分钟

用户日均使用时长

15%

付费用户长期留存

- **市场空间大**：2024年短剧在国内的市场规模为400-500亿元，已经接近国内电影市场规模，海外市场发展情况和渗透率要低于本土，市场空间更加广阔
- **制作效果接近成熟**：目前AI短剧的制作水平还难以与传统实拍模式媲美，但在的动画短剧领域已基本达到可用水平。随着模型能力逐步迭代，未来一年内普通的短剧生成也将达到用户可消费的水平
- **制作流程介绍**：目前采用与外部导演合作的模式，1）由导演进行剧本创作，并将其分解为分镜剧本，2）井英科技将分镜剧本转化为提示词并输入视频生成模型中（该环节替代了短剧演员）获得结果，3）导演从生成结果中中选择满意的分镜视频，或再通过提示词进行二次生成调整，4）选定视频后在传统视频工作流中进行后期的剪辑和处理
- **互动功能**：用户在App内可与短剧主角聊天，类似Character.AI，可增加用户粘性

Reel Diffusion

- AI短片《M.A.D》（俱皆毁灭），在全球AI电影马拉松大赛上荣获B站观众选择奖。该短片利用CreativeFitting 自研的视频大模型 Reel Diffusion 生成，**效果媲美传统动画短片**

（Reel Diffusion生成的动画短剧）

- Reel Diffusion 视频大模型在叙事型视频的生成方面能力领先
- 模型支持人物角色的细腻情感表达及复杂场景的生成，从算法到训练数据及工程实现，都进行了专门设计，帮助创作者讲述引人入胜的故事
- **动画短剧生成要比普通短剧生成更加成熟**

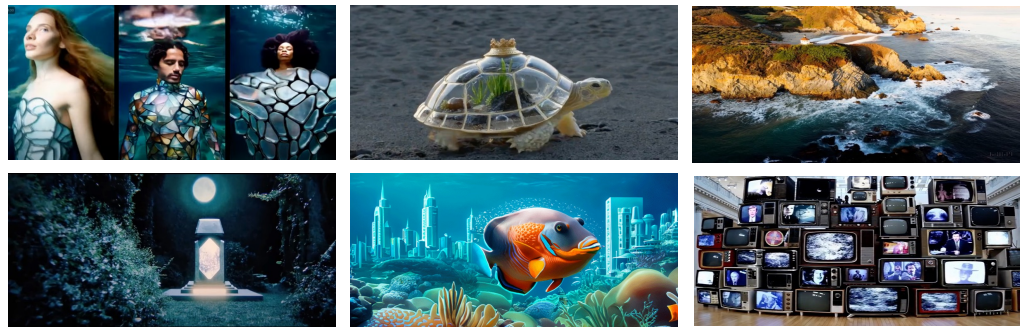
信息来源：量子位智库，井英科技

应用趋势：视频生成模型正在赋能传统视频制作工作流，目前主要价值在于素材生成环节，其他环节有少量渗透

	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
工具	Office 365 + 各种生产力软件			Adobe shutterstock	Pr Au Ae	
视频模型赋能点	• 视频生成模型缩短视频生产周期，快速进行创意验证，生成粗略的原型视频，加快创意、情节构思规划过程 • 通过大量新素材和创意素材可以进一步激发创造力，提升内容创作质量					• 传统工作流在后期阶段对于质量的要求较为精细，目前视频生成模型的后期编辑功能还不能充分满足需求

相互激发

通用素材生成



信息来源：量子位智库，OpenAI

应用趋势：新一代AI视频工作流正在萌生，将整合音视频创作全流程提高创作效率，降低AI视频内容的制作摩擦

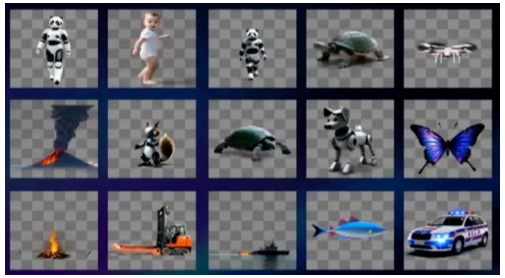
	前期制作		中期制作		后期制作	
	大纲脚本	分镜规划	素材拍摄	素材选取	剪辑	渲染/音效
内容	• 故事创意、主题风格构思，明确故事大纲，如主要情节和角色发展 • 将大纲扩展成完整的剧本，包括对白、场景描述和情节细节	• 文字剧本视觉化，确定镜头类型、镜头运动、角色位置和动作 • 编制镜头列表，包括每个镜头的编号、描述、拍摄角度、镜头运动等	• 根据分镜规划制定拍摄计划，完成器材准备和技术测试、准备演员、服化道及布景设置	• 根据情节和主题需要进行素材整理、分类、剔除无效素材 • 对于无须拍摄类内容，需要从素材库进行素材选取	• 初步剪辑、逻辑剪辑，拼接出粗略的视频版本 • 精剪微调，对影片进行精细剪辑，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	• 快速进行创意验证，生成粗略的原型视频，加快创意、情节构思规划过程。通过大量新素材和创意素材可以进一步激发创造力，提升内容创作质量		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。**目前头部文生图产品的图片生成质量最好，一般先用文生图应用生成指导帧**，然后再作为视频生成的输入进行动态化		runway • 不会涉及专业级影视剪辑，视频大部分关键元素在生成阶段完成，用户主要的做一些简单的逻辑剪辑或者音效编辑，可以快速出片	
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 → 正在萌生

信息来源：量子位智库

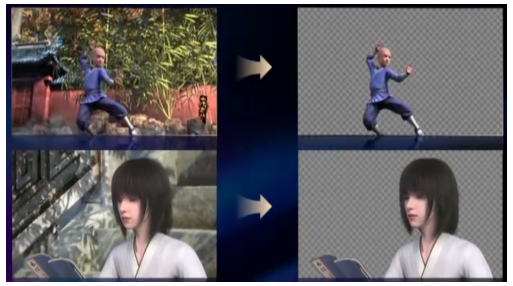
应用案例-工作流（精细化生成）：阿里达摩院发布寻光视频制作平台，通过图层编辑方式和工作流整合提升创作全流程效率

1 基于图层组合的编辑方式



图层生成

- 用户可以单独生成视频中的角色、物体和环境对象，生成的视频为透明背景，可以整合覆盖到其他视频内容中，实现对于视频内容的细颗粒度操作和局部编辑



图层拆解

- 用户也可以上传自己视频，寻光平台可以对视频进行图层拆解，分解出创作者需要的视频内容，例如人物角色，方便用于其他视频内容的组合、编辑

图层融合

- 可以把用户自己生成的视频内容或者上传内容进行图层组合，可以实现视频背景、人物的任意切换，以解决目前AI视频生成的场景一致性问题，增加可用性。支持场景和人物的批量替换，功能强大

2 易用的工作流界面

类PPT的图形化操作界面

- 将视频分解为多个场景的组合，再将每个场景分解为多个分镜视频，方便用户预览整个视频，对每个镜头进行精细编辑，可以直接通过拖拽完成顺序调整，在任意位置进行添加、删除
- 针对每个分镜视频提供一揽子的的编辑功能

3 整合大量AI编辑功能

生成素材

上传素材

全局型元素

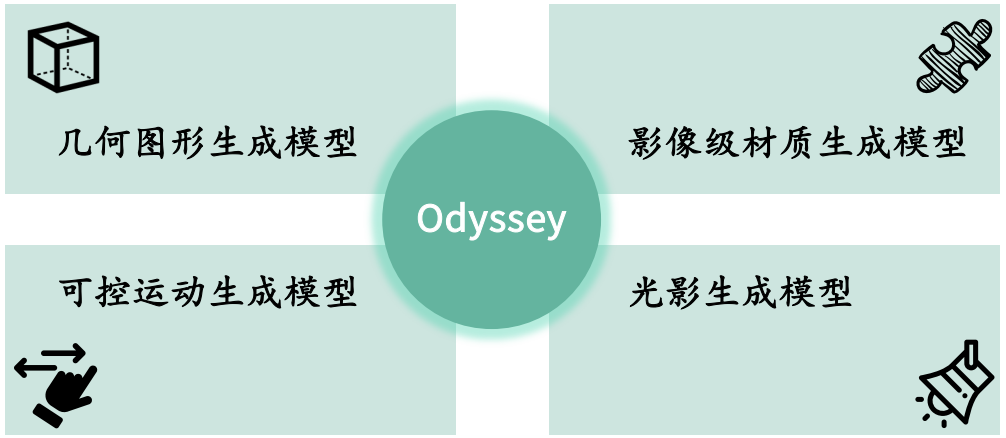
- 视频风格化：莫奈、浮世绘、水彩、水墨、卡通等20种风格
- 镜头运镜控制：左右平移、上下平移、推进拉远、左右环绕等
- 帧率控制：修改不同镜头的帧率修改使得视频更加一致丝滑
- 清晰度控制：提供不同清晰度的生成选择
- 画质增强：提供视频超分工具

局部型元素

- 目标编辑：可以消除、替换、新增视频中的各类目标
- 移动目标：通过拖拽可以实现目标的运动效果，人体控制：控制视频中角色的肢体动作
- 人脸控制：批量替换、编辑人脸
- 前景、背景控制

信息来源：量子位智库，达摩院

应用案例-工作流（精细化生成）：Odyssey结合4种生成模型，可以实现对视频内容的精确控制和生成，主打好莱坞级的视频内容生成



- **主打高端影视场景**：能够生成好莱坞级的山脉、平原、植被、海洋、河流、火焰、烟雾、建筑、人物以及任何其他东西创作者可以完全控制场景中生成的每个元素和位置方向，无论是几何形状、材质、灯光、动作还是其他方面。场景由可提示和可操作的对象组成，这些对象可以独立运行，同时还能保持上下文感知
- **多元化团队背景**：主要是技术人员+创作者的组合，例如来自 Cruise、Waymo、Tesla、Microsoft、Meta 和 NVIDIA 等公司的技术人员，首席工程师来自《孢子》、《模拟城市》、《模拟人生》、《异形：隔离》等视频游戏，**艺术家**则曾参与制作《沙丘 2》、《哥斯拉》、《造物主》、《复仇者联盟：奥创纪元》、《艾丽塔：战斗天使》和《侏罗纪世界：失落王国》等电影
- **投资方**：包括谷歌风投、Elad Gil、Garry Tan、Jeff Dean等以及来自OpenAI、Deepmin、Meta、Midjourney、Pixar的研究人员

材质生成（如荒漠、草地、土地）

对象生成（如树木、石块、森林）

光影生成（不用强度、方位、风格）

对象纹理生成

信息来源：量子位智库，Odyssey

应用案例-工作流（流程整合化）：美图发布AI短剧制作平台MOKI，整合包括创意生成、后期编辑、音效制作等视频创作全流程

前期

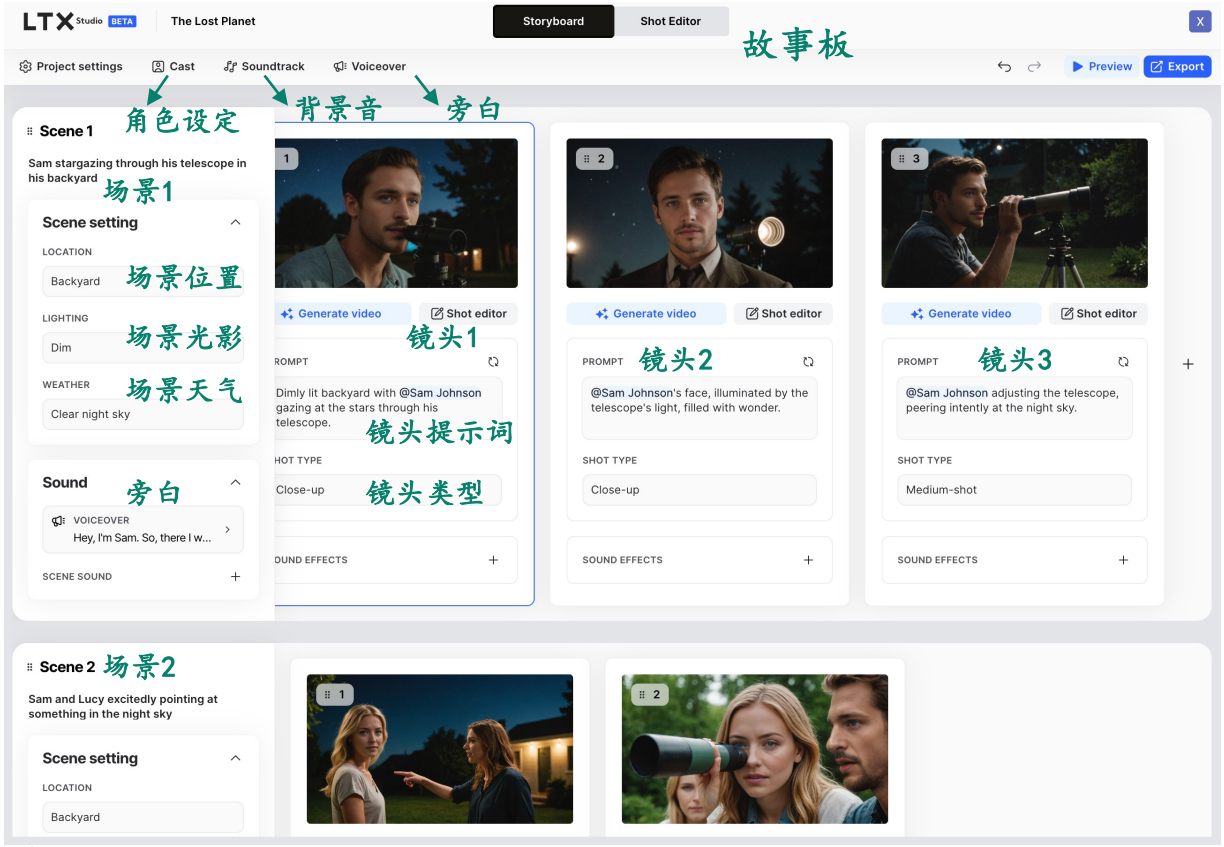
中期

后期

可制作各类短片

信息来源：量子位智库，美图

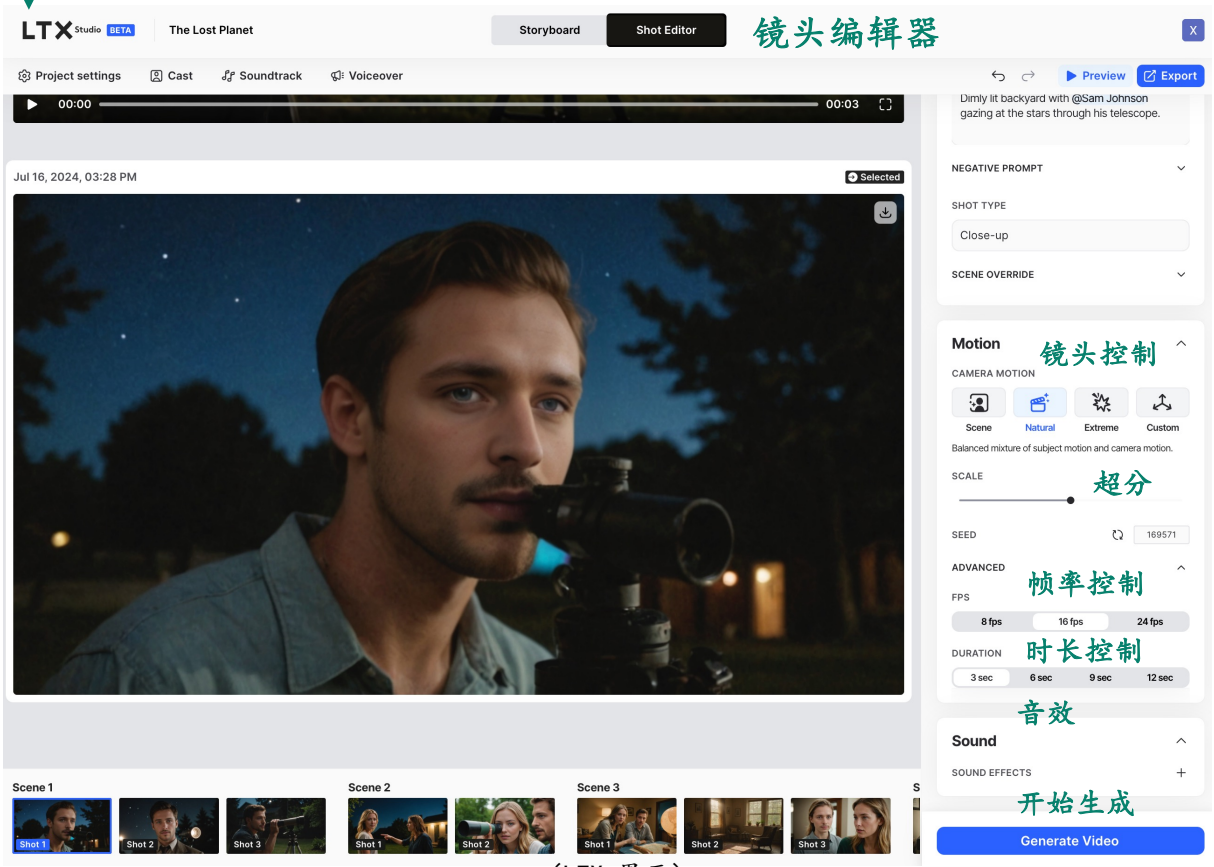
应用案例-工作流（流程整合化）：LTX Studio采用基于故事版和分镜的生成编辑方式，同时整合音效、旁白等功能



（LTX 界面）

- 故事板界面：用户需要先进行角色设定，包括人物的肖像、风格、名字等，然后故事版可以帮助用户构思视频的整体内容，包括从场景和分镜头两个层次，可以设定每个场景的基本情况，如位置、光影、天气等，也可以添加该场景的音效和旁白。

- 镜头编辑界面：可以对每个镜头进行精细化编辑，例如镜头的控制（LTX提供了超过10种镜头运动方式）、分辨率调整、视频帧率控制、音效旁白等，确认好基本参数后可以开始生成视频



（LTX 界面）

信息来源：量子位智库，LTX Studio

产品路线：视频生成目前仍处于早期阶段，从应用路线上看主要分为通用类生成和垂直类生成两类产品

通用类

- **场景广泛**：不针对某一类风格、行业、角色或其他方面进行垂直优化，旨在用视觉信息建模物理世界，通过自然语言作为提示词生成视频
- **天花板高**：通用生成的想象空间更大大，创意性强，未来将有更多应用形态涌现，预计未来视频的生成和理解将会逐步统一，强大的视频生成能力也代表视觉理解的进步

"As great as Sora is generating things that appear real - what excites us is its ability to make things that are totally surreal."

---Shy Kids

特点

- **研发难度大、算力、数据资源要求高**：模型本身是对数据集的拟合，要求模型能够生成任意内容的视频，本质上是要求训练数据集的场景丰富程度极高、内容质量好，标注质量详尽、准确，以及经过大规模scale来学习视频中包含的各类知识和物理规律，目前大多数视频生成技术公司都属于此类
- **内容合规和本土化问题难以避免**：视频输出内容可以包含更多维度的信息，其中可能涉及内容安全、意识形态及不同文化背景的偏好差异，例如本土模型对本土文化理解力更好，海外模型的输出会凸显欧美审美偏好和价值观

案例

OpenAI SORA　Pika　runway　PixVerse　KLING 可灵大模型　生数 ShengShu　Asphere 爱诗

垂直类

- **场景细分**：垂直类视频生成主要指围绕细分需求进行视频生成，主要是针对细分场景，用垂类数据或者私有数据做适配训练和可控性、稳定性优化
- 商业化路径清晰，有稳定的商业模式和营收

特点

- 需要的算力资源和数据资源少，主要是用少量垂类场景数据和算法对模型进行加强，模型不追求“大”，且在模型层选择灵活，可以把文生视频、图生视频作为外部能力接入传统模型作为辅助增强，**核心要素还是行业知识**
- 目前垂直类产品主要是在营销场景下，针对人物、或者某一类风格进行微调，几千条数据就可以显著增强模型在垂直领域的表现

案例

FancyTech　奥创光年　BOOLV　CreativeFitting

信息来源：量子位智库，OpenAI

商业模式：通用视频生成在海外市场以SaaS产品为主，国内市场项目制为主，服务内容多样化，但订阅制有待成熟

付费点

SaaS产品

生成点数 | 生成时长 | 生成速度 | 团队协作

增值功能（视频超分、音效功能、编辑功能、各类动效）

定制化

模型训练 | 客户专员支持 | 业务沟通 | API定制化

私有化部署 | 生成数量

本土市场

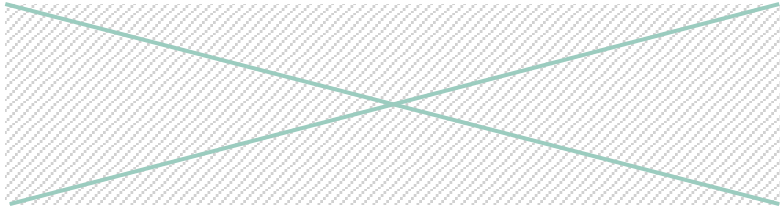
- 目前本土的SaaS市场成熟度相比海外仍有欠缺，用户的主要画像是自媒体创作者、创意工作者，覆盖人群比较垂直，商业化规模有限
- 随着新一代用户的年轻化、专业化，为内容工具的付费的习惯正在逐渐形成，但仍需时间培育

- 目前主要客户以各行业头部公司为主，预算比较充足且愿意拥抱AI新技术，一般大客户都会要求部分定制化服务
- 视频生成领域的定制化一般不涉及技术上的二次开发，工作量主要在具体的需求沟通、微调模型，帮助客户熟悉产品，以及提供技术支持服务等
- 一些场景需要客户和公司结合行业知识进行共创，例如营销视频在内容结构、风格、审美等方面的选择

海外市场

- 海外市场是文生视频类应用的主打市场，**生成式AI的创意市场已有类似产品完成市场验证**，例如Midjourney年收入已经超过2亿美元、超2000万用户
- 海外市场的用户主要是一些C端用户或者中小B端用户，主要通过在社交媒体和创意工作者人群中构建社群并以PLG的方式进行增长
- 目前主流的通用视频生成应用大都采用SaaS应用服务模式，向用户收取每月订阅费用或者生成视频的消耗量分不同付费版本灵活计费

- 海外市场的定制化目前较少，主要是SaaS服务，但头部公司也会提供模型训练服务和API定制化，例如Adobe和Runway
- 创业公司也会做一些大客户或者知名案例，但主要目的是**打造知名度，同时探索用户需求、打磨产品**，例如Runway曾参与《瞬息全宇宙》的制作、Sora完成《Air Head》制作



关键分析

- **本土市场**：如果不同视频生成模型之间能力没有显著差异化，很可能出现类似大语言模型领域的价格战，产品盈利将面临较大压力。C端市场、专业消费者是最理想的市场，但如果场营收不好，可能要持续做B端服务
- **本土市场目前的商业化路径一般是“从大到小”**，例如从大B的影视公司，到中B的影视工作室，游戏工作室、广告工作室、短剧团队，再到自媒体创作者、专业创作者等Prosumer、小B用户
- **海外市场：C端会是长期主线**，目前竞争并不激烈，处在逐步拓展市场的阶段

信息来源：量子位智库

对比分析：视频生成相比图片生成的应用复杂度更高，大规模普及或需要从模型到工作流的全面进步才能打开市场

	内容生成	编辑修改	后期效果	应用规模
图片生成	耗时：• 1分钟左右 产出数量：• 输入一次提示词可以生成4张图片供用户选择（以Midjourney为例） 产出质量：• 生成的大部分图片可以满足普通用户的需求标准，技术已经比较成熟 • 抽卡次数较低，一般生成5次左右可以得到理想结果 成本：• 约0.03美元/张（以Midjourney为例）	• 图片生成的二次编辑比较简单，一般用Photoshop或者其他编辑软件就可以满足需求，主要是在细节和特殊需求上做精调		Midjourney 2亿美元年营收[1]
视频生成	耗时：• 2分钟左右（以最新的Runway Gen-3为例） 产出数量：• 10秒左右的视频片段（以最新的Runway Gen-3为例） 产出质量：• 生成视频存在畸变、分辨率低、指令遵从差、稳定性差等问题 • 目前抽卡率非常高，可能到几百，例如Shykids团队用Sora制作的MV作品抽卡率约300，需要进行大量尝试 • 视频生成对于创作者的要求更高，功能和提示词的复杂度更高 成本：• 不同模型各异，取决于生成视频的质量和长度，Runway的价格约为3美元每分钟，Sora可能高达数十美元每分钟 • 生成片段并不足以构成可用作品，实际场景需要大量的视频生成片段组合为完整的视频（时长数分钟或更长），进一步推高了应用成本	• **修改难度大**：在视频生成内容上做编辑很难，例如人物对象、环境等内容发生畸变的修改难度和工作量非常大 • **一致性难题**：视频制作要保证多个视频片段中人物角色、物体和周围环境的一致性，如何精准衔接多个片段，目前还没有很好的解决，对二次编辑的依赖度较高	• **音频需求**：视频和图像的另一个的不同是视频需要匹配音频内容，需要考虑背景音乐、旁白等内容，增加了额外的复杂度 • 视频需要考虑时间维度和视频逻辑，在后期需要逻辑剪辑等环节	runway 约500万美元年营收[2]

1 视频生成环节复杂、昂贵、缓慢 ✗

2 视频编辑修改难度大 ✗

3 视频后期环节多 ✗

信息来源：量子位智库，1、2均代表2023年数据

1. 技术侧

2. 应用侧

3. 玩家格局

竞争要素：基础模型、产品、场景是AI视频生成发展的三个主要方向，自底向上由模型进步逐步驱动上层发展

	描述	举例	发展方向
场景层	• **外接场景**：通过的视频生成模型API赋能其他场景，给其他产品输出模型能力，为其它产品集成API • **应用场景**：可以直接接入现有应用，作为增值功能，获得用户流量资源，支撑应用、功能启动	• **外接场景**：Sora、Runway给Adobe、Canvas提供模型API • **应用场景**：OpenAI把DALL-E集成到ChatGPT中，营销视频模型有直接的视频消费场景，AI视频营销、AI影视制作、AI游戏设计、自媒体创作、泛娱乐创作等	• **存量场景**：主要针对各行业已有场景的赋能，随着视频生成能力提升将逐渐渗透 • **增量场景**：还处在早期阶段，**需要用户和产品进行双向探索，需要时间等待涌现**
产品层	• **交互界面**：目前视频生成产品的交互界比较早期，就是输入提示词输出视频结果，主要是网页和本地两种方式 • **应用类模型**：在基础模型之上进行局部优化、调整、编辑的算法和模型 • **工作流**：视频的制作不仅视频素材的生成，也包括创意生成、后期的剪辑精调、音频素材添加、分享协作管理 • 社区：	• **交互界面**：目前主要是三种形式，Discord对话框，WebUI或者移动应用，以及节点式ComfyUI • **应用类模型**：已经产品化的包括动态笔刷，镜头控制，表情控制等。相关的科研论文也比较多，致力于增加更多维度控制性、精确性 • **工作流**：Runway提供了全面的视频工作流，提供从制作到剪辑再到后期的丰富工具包	• **交互界面**：尚不清晰，但基础模型的推理速度提升和推理成本下降可能是本质因素 • **应用类模型**：视频生成下一阶段的核心核心是可控性的提升，例如如何保持一个角色在多个生成片段中的一致性，预计短期会有明显进步 • **工作流**：但生成式内容正在渗透传统的工作流，但是比较碎片化，需要试用多个工具各取所需
基础模型层	• 模型层主要指底层视频生成的基础模型，研发难度大，成本高昂，资源投入密集 • 下一代头部视频生成大模型的训练的成本可能超过1亿美元	• Sora，Veo，Runway、Pika、Pixverse、Vidu、可灵大模型、Sable Video Diffusion	• **优化方向**：模型架构优化、训练数据优化、Scale up（增加模型规模）、推理成本、速度优化 • **格局**：长期来看，头部玩家的视频基础模型将持续Scale up成为世界模拟器，其他玩家可能深耕场景，做差异化产品

产品市场匹配

模型产品匹配

信息来源：量子位智库

玩家格局概览：目前AI视频生成领域主要有OpenAI、互联网公司、技术创业公司、内容工具软件、垂类创业公司5类玩家

	OpenAI	互联网公司	技术创业公司	内容工具软件	垂类创业公司
场景层	✓ • 中，内部来说ChatGPT有坚实的用户量基础，DAU已经过亿，外接方面可以高举高打找行业头部公司合作打造标杆效应，BD能力优秀	✓ • 中，公司旗下的内容平台、工具平台都可以进行集成视频生成产品，有视频内容的应用场景和流量基础	✓ • 弱，目前只能为用户提供单点的视频生成能力，缺少流量支持、应用场景、打开市场较难	✓ • 中，离视频创意类用户比较近，内容工具是视频制作工作流中的重要关节，用户有生成视频素材的需求	✓ • 中，生成视频以直接商业化为导向，在场景层最为成熟，玩家一般有行业背景和资源，可以更好匹配客户需求快速商业化
产品层	✓ • 弱，目前来看仅为基础模型提供简单的使用界面和API服务，不会做更加精细化的产品，**产品层面的创新不是OpenAI的核心优势**，业务上优先级不高	✓ • 互联网公司积累了较多的应用类科研成果和丰富的产品化经验，但大公司和组织和监管等因素会拖累产品研发，**动作空间和迭代效率有限**	✓✓ • 强，会为用户提供多样化的功能，产品迭代速度和产品功能灵活度最好，主流应用目前主要来自技术创业公司，优势在于联动产品和模型	✓ • 中，**深耕视频制作领域，更能理解视频制作方面的用户需求**，优势在于为用户提供全面的创作体验，为用户提供功能丰富便捷的工作流	✓ • 中，重点在于垂直行业知识，这是其他玩家无法具备的，本质上和其他玩家的直接竞争并不大，主打的客户也是垂类的领域的客户的
模型层	✓✓ • 强，目前OpenAI在**基础模型能力上相比其他玩家有代际领先**，且有充足的资源条件继续持续创新，开拓边界	✓ • 中，相对其他玩家有充足的数据资源和算力资源，但在整体业务上重要性和优先级各异	✓ • 中，**模型层是公司的核心竞争力和价值点**，团队一般有较强的技术背景，集中发力模型层	✗ • 弱，在算法和算力资源上没有优势的，模型层不是业务重心，**可以直接选择外接其他基础模型，灵活度高，基础模型并不是业务核心**	✗ • 弱，不追求通用生成能力，主要从业务场景出发驱动底层模型的开发，或者**可以直接选择外接其他基础模型，模型层并不是业务核心**

信息来源：量子位智库

OpenAI：聚焦AGI，认为视频模型是世界模拟器、通往AGI的重要路径，有充足的资源和决心重注scaling路线，主要发力模型层

OpenAI

"Video generation will lead to AGI by simulating everything." ---OpenAI

- OpenAI在视频生成领域目前处于大幅领先的状态，**目前主要发力模型层，追求基础模型上的能力跃升，产品层面动作较少**，未来视频模型的商业化模式可能主要是API服务，更多的产品化、场景化更可能交给合作伙伴或者客户来做
- **OpenAI认为Sora可以类比为视频生成领域的"GPT-1时刻"，Sora只是视频生成领域Scaling curve上的第一个data point，且相信视频生成模型是世界模拟器、通往AGI的关键路径**
- 应用方面，目前OpenAI主要寻求和大型影视公司和行业头部客户、创作者进行合作，逐渐摸索市场需求，完善模型的安全性、对齐程度、并打出Sora在AI视频应用方面的标杆案例

场景层

- 自有场景方面可以和自家旗舰应用集成，ChatGPT日活近亿，可以直接触达大量用户，但没有直接的视频应用、消费场景。外部场景主要是把视频生成API接入第三方应用，例如Sora接入Adobe、Canva等内容制作工具，或者类似Dalle-3接入Shutterstock等；此外也会提供专业客户服务做标杆，例如与Shy Kids、玩具反斗城合作，但是PMF不清晰、大规模应用尚未成熟

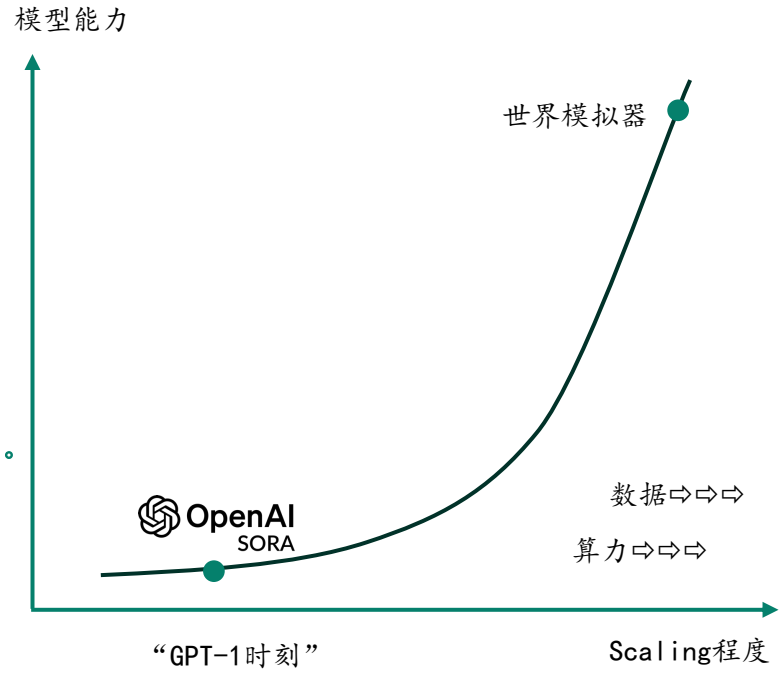
产品层

- OpenAI声称Sora将于24年年内投放市场使用，具体的产品使用形式未知，目前也没有清晰的产品路线图，可能类似Dall-E系列集成到ChatGPT中供C端用户使用，猜测不会涉及专业的视频工作流

模型层

- OpenAI在基础模型层各个方面都有领先优势：1）算力方面资源充足，例如微软Azure在算力和工程上的支持，2）有极高的人才密度，例如吸引关键论文DiT的作者加入团队，3）内部技术的正反馈，例如利用自有多模态模型（GPT-4V）进行数据标注和提示词增强等
- **从核心团队将Sora定位为"GPT-1"来看，预计视频模型也将继续保持迭代（类似DALL-E系列），空间依然非常大**，OpenAI在视频基础模型方面将继续保持代差式领先，持续开拓技术边界

"Sora已经开始显现出对于人类交互的细致理解，随着我们在这一范式上继续scale，最终我们可以建模人类如何思考。模型能生成真正的逼真视频、动作序列的唯一方式就是得到一个关于人类、环境如何运作的内部模型，这项技术将很快会变得更好。" ---Sora 团队



信息来源：量子位智库，OpenAI

OpenAI：Sora的成功源自其在数据、算法、算力等层面的优势累加，在未来将对模型持续进行迭代部署，引领行业

1 数据

A 数据规模大

- **数据量大**：Sora的视频训练数据据推测在500万小时左右（供参考），不仅包含视频数据，也包含图片数据，OpenAI在技术报告中暗示其数据量可能是类似大语言模型的“互联网级”
- **合成数据**：OpenAI可能使用了一些物理引擎、游戏引擎渲染的合成数据，以帮助模型更好地学习物理规律，例如OpenAI收购了Global Illumination（一家3D引擎渲染的公司）
- **版权数据**：建立多方数据合作，例如与全球最大的图库Shutterstock合作获得图片、视频、音乐及其他元数据

B 数据质量好

- **Recaption**：使用Recaption技术可以提高数据质量，主要是对视频和图片进行精细化打标。OpenAI复用了之前在DallE 3中Recaption技术，可以利用GPT-4V对视频文本对数据进行标注
- **训练侧**：用合成数据进行Recaption可以丰富、校准视频数据的文字标注。DALL-E 3中的Recaption显示使用合成数据进行训练的模型生成效果要比不使用更好，且合成数据占比更高的模型生成效果更好
- **推理侧**：可以利用GPT进行文本拓展，把用户输入的prompt进一步扩展，更丰富细节更多，使得生成的内容更加生动细密逼真

2 算力

A 算力投入大

- **训练方面**：预计Sora的训练使用万卡英伟达H100集群，训练成本数千万到上亿美元，远超其他模型

B 系统运营强

- 基础设施能力强大：例如万卡计算集群的运营、监控、维护、管理等，在训练和推力方面的优化能力，可以增加算力的利用效率
- **OpenAI和微软Azure团队是少数有数万张先进GPU集群运营实践的团队**。OpenAI最近一年增建了大规模的模型推理优化团队，专注于在Infra层面进行基础优化降本增效，最近也在从Google的TPU团队招揽人才，进一步下探技术栈

3 算法

A 发明通用的数据表示

- 时空潜在图块（Spacetime latent patches） 的表示形式解决了不同数据来源中分辨率、宽高比、时长各异的问题，patch在模型中扮演类似token的角色，统一的数据表示使得利用transformer进行scale成为可能

B 押注Transformer的可扩展性

- 引入Diffusion Transformer代替传统的U-Net，解决了卷积神经网络在参数量到达一定规模后，进一步scale增益减小甚至程消失的问题
- 原始的DiT架构并不能直接用于视频，需要针对视频进行改进，Sora之前业界已有将Transformer引入视频生成领域的工作（例如李飞飞与Google 研发的的W.A.L.T项目和谷歌的Videopoet项目），OpenAI并不是首创，**但在工程上把scaling推到了极致**

C 单独训练视频压缩、解压网络

- **减少训练算力要求**：OpenAI训练了一个能降低视觉数据维度的网络，可接受原始视频作为输入并输出一个在时间和空间上都被压缩的潜在表示，Sora在这个压缩的潜在空间内进行训练
- 压缩网络网络可以看做包含时空信息的压缩视频编码器，压缩后可以大量节省算力资源，同时尽可能保留视频原始信息，并为后续训练过程奠定基础。相应地，OpenAI也训练了解码模型，在完成生成过程后新的潜在表示映射回像素空间（即输出视频）

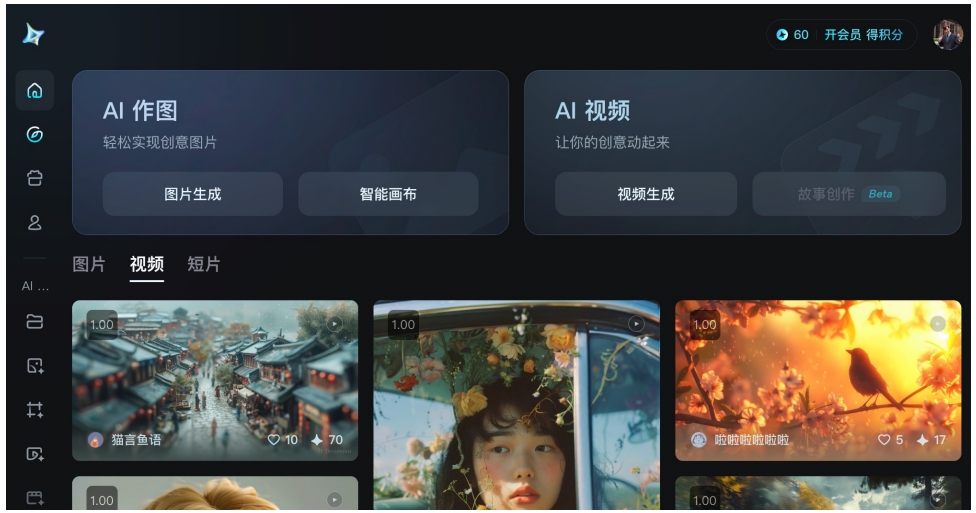
信息来源：量子位智库，OpenAI

互联网公司-字节跳动：AI视频生成和核心业务强相关，在战略层面高优先，资源充足保持自研策略，视频生成已经开始初步产品化

ByteDance 字节跳动

- **字节在AI方面布局以内部开发为主，内部产品多，外部投资少**，在视频生成领域也延续了这一策略，尚未对投资视频生成创业公司
- **AI视频生成和字节核心业务强相关**，长期看可能对整个视频内容的生产、消费业态造成冲击，**视频生成相关业务在集团优先级很高**（例如抖音集团CEO张楠已转至负责剪映业务），对于抢占新技术、新产品高地的需求强烈
- **公司有强大的视频产品基因**：作为头部内容平台积累了大量视频数据，在视频处理方面有丰富的产品、工程经验，算力资源相对充足

字节已经推出包含视频生成功能的产品即梦Dreamnia



视频相关成果	类型	简介	时间
Story difffusion	应用模型	Story Diffusion可以增强视频生成中角色一致性，包括风格服装等等，有助于生成长视频和连贯的图像	2024.5
Dreamina	产品、基础模型	AI作图和AI视频生成功能已全量上线，是一款文生图、文生视频的商业化产品，之前集成在剪映中，现在单独分拆面向市场	2024.5
AnimateDiff-Lightning	基础模型	文本到视频的生成模型，生成效率能够比原版AnimateDiff快十多倍，采用跨模型扩散蒸馏方法实现，已开源研究模型发布	2024.1
MagicDance	应用模型	一个基于人物图片生成逼真生动舞蹈视频的模型	2024.1
MagicEdit	应用模型	文生视频的编辑工具，通过自然语言提示改变视频风格、场景甚至替换视频里的对象或添加元素，同时保持原视频一致性	2023.9
Boximator	应用模型	通过文本控制生成视频中人物或物体的动作，基于 MagicVideo-V2 模型	2024.2
Magic-Me	应用模型	一个虚拟人物视频生成框架，特点包括：稳定人脸、保持视频中人脸的稳定性；提取人物信息、高清重绘：在合成后，对视频进行高清重绘，提升画面质量	2024.2
MagicVideo-V2	基础模型	文生视频模型，可生成具有出色保真度和平滑度的美观、高分辨率视频。用户评估表现优于 Runway、Pika 1.0、Morph、SVD等	2024.1
MagicAnimate	应用模型	一种基于扩散模型的人体图像动画框架，旨在增强时间一致性、忠实地保留参考图像并提高动画保真度	2023.12
PixelDance	应用模型	结合了图像指令和文本指令来生成视频，针对复杂场景和大幅运动进行了优化	2023.11
MagicAvatar	应用模型	通过简单的文本提示就能创建虚拟人物，也可以根据源视频生成跟随给定动作生产，还能对特定主题的虚拟人物进行动画化。	2023.9
MagicVideo-V1	基础模型	早期视频生成模型，能在单个GPU卡上生成256x256分辨率的视频	2022.10

信息来源：量子位智库，Dreamnia

互联网公司-阿里巴巴：阿里延续了AI方面的总体策略，偏好通过内部自研加对外投资的方式加强在视频生成领域的布局

阿里巴巴

- **阿里总体的AI战略是以阿里云为核心从B端突破，同时注重外部投资布局。**阿里内部视频相关的内容场景比较少，视频生成和核心业务相关性较低，且阿里比较缺乏内容方向的基因（大文娱集团一向表现不佳），内部孵化相关产品难度比较大，且在集团层面视频生成大模型的业务优先级有限，不是主推方向
- **内部研发方面**，视频生成方向的研究主要依靠下属阿里云集团的达摩院，整体风格更偏向科研，但也有视频生成产品的尝试
- **对外投资方面**，阿里在视频生成赛道延续在AI方向的整体策略，重视对外投资，例如参与了视频生成创业公司生数科技的亿元Pre-AI轮融资、领投爱诗科技的A2轮亿元融资
- 阿里内部业务场景主要在B端，除阿里云外，小部分技术成果也是从B端业务团队诞生，例如由淘天集团阿里妈妈团队发布的AtomoVideo，在商品营销视频方面有潜在应用场景

阿里巴巴 投资布局 爱诗AIsphere 生数ShengShu

视频相关成果	类型	简介	时间
寻光视频创作平台	产品	一站式AI视频创作平台，提供AI视频的工作流和丰富AI编辑工具	2024.7
UniAnimate	应用模型	支持合成一分钟的人类动作高清视频	2024.6
EasyAnimate	应用模型	DiT-based视频生成框架，它提供了完整的高清长视频生成解决方案，包括视频数据预处理、VAE训练、DiT训练、模型推理和模型评测等	2024.6
AtomoVideo	应用模型	高保真图像视频生成框架，该框架利用高质量的数据集和训练策略，保持了时间性、运动强度、一致性和稳定性，并具有高灵活性	2024.3
EMO	应用模型	仅需一张人物肖像照片和音频，就可以让照片中的人物按照音频内容唱歌、说话，且口型基本一致，面部表情和头部姿态非常自然	2024.2
Animate Anyone	应用模型	只需一张图片即可生成平滑稳定的视频。这项技术对短视频、电商和动漫行业都有一定的影响，用户只需提供一个静态的角色图像和一些预设的动作（或姿势序列）然后会生成该角色的动画视频	2024.1
Livephoto	应用模型	让真实图片根据用户的指令动起来，LivePhoto能很好的保持参考图片的细节，精准跟随文本指令，生成大幅度运动的视频	2023.12
DreaMoving	应用模型	一个基于扩散模型的可控视频生成框架，用来生产高质量的定制人物视频,提出了一个用于动作控制的“视频控制网（Video ControlNet）”和一个用于保持身份一致性的“内容引导器（Content Guider	2023.12
VGen	基础模型	一个视频生成的全面框架生态，包括了阿里在此领域的多项研究，如：VideoComposer：具有高度灵活可控性的视频合成框架；InstructionVideo：通过人类反馈，优化视频扩散模型；开源图生视频大模型 I2VGen-XL、文生视频模型ModelScopeT2V	2023.9
ModelscopeSythesis	基础模型	阿里早期的视频生成的模型	2023.3

信息来源：量子位智库

互联网公司-腾讯：主打混元DiT多模态大模型，将支持高质量图片和视频生成，产品化进展较慢

Tencent 腾讯

- **腾讯在视频生成方面布局相比字节阿里更加中和**，内部自研产品方面节奏没有字节激进，对外也没有投资视频生成相关的创业公司，**但视频生成模型是必选项，主要策略是内部开发**
- 混元大模型项目主要由腾讯TEG负责，团队C端产品能力不是很强，主要还是做底层技术，过去一年大模型的应用战略主要是支持腾讯内部大量场景，腾讯内部有600+场景，各个场景对大模型需求很大，但视频生成产品化程度还不清晰
- 腾讯的自身的C端流量的非常稳固，核心业务不在视频生成的潜在冲击范围内，视频生成业务的优先级相比抖音、快手等视频内容平台更低，产品化进度较慢
- **目前腾讯主打多模态模型是混元DiT**，据称支持文生视频、图生视频、图文生视频、视频生视频等多种视频生成能力，已经支持 16s 视频生成，预计在24年第三季度可以实现30s视频生成，但尚未明确何时向C端用户开放

视频相关成果	类型	简介	时间
MOFA-Video	应用模型	通过箭头控制视频内容的运动方向，控制手动轨迹、迁移面部表情等	2024.6
Follow Your Emoji	应用模型	可以通过人脸骨架信息生成任意风格的脸部动画，一键创建“表情包”	2024.6
Follow-Your-Posev2	应用模型	输入人物图片和动作视频，让图片上的人跟随视频上的动作动起来	2024.6
ToonCrafter	应用模型	一个帮助动画师生成和优化卡通动画过渡效果的工具	2024.5
ID-Animator	应用模型	文本驱动的人物视频生成框架。由参考图片生成角色定制化视频	2024.5
Revideo	应用模型	通过指定内容和动作，在特定区域进行精确的视频编辑	2024.5
混元DiT	基础模型	中文原生的 DiT 架构文生图、视频开源模型，支持中英文双语输入	2024.5
MuseV	应用模型	虚拟人物视频生成框架，可以将各类人物角色图片动态化	2024.3
DynamiCrafter	应用模型	根据文本提示将静态图像转换成动态内容，强调更真实视频动态	2024.3
MovieLLM	应用模型	生成高质量、多样化的视频数据，减少了人力的投入	2024.3
AniPortrait	应用模型	腾讯游戏团队开发的由音频驱动的人像动画合成工具	2024.3
Follow-Your-Click	应用模型	点击对应区域加提示词，让图片中静态的区域动起来一键转换成视频	2024.3
FreeNoise	应用模型	一种视频长度增长推理方法，可以适用于多种视频生成模型	2024.1
VideoCrafter2	基础模型	该模型框架是在之前的 VideoCrafter1 基础上进行了大幅改进	2024.1
AnimateZero	应用模型	一种基于视频扩散模型的零样本图像动画生成器	2023.12
VideoCrafter	基础模型	腾讯的首个文生视频、图生视频模型	2023.10

信息来源：量子位智库，腾讯

互联网公司-快手：推出可灵视频生成大模型，同时匹配生成控制和文生图功能，视频生成效果领先，在本土和海外社区受到广泛关注

快手

- **快手处于和抖音类似的生态位，视频内容是快手的核心业务，内部优先级较高**，希望把视频生成模型和生产者工具结合起来，不断帮助创作者降低创作门槛，提升短视频制作质量和效率。除助力内容生态外，图生视频等功能也可以赋能来自快手电商业务的商品视频生成需求。
- **快手在6月份发布了可灵大模型**，可生成长达2分钟的高质量视频，海内外关注度较高，生成效果可优于目前大部分创业公司的产品，在快影App和网页端处于内测阶段，**已经开放给30万用户使用**

（快手可灵应用网页端）

视频相关成果	类型	简介	时间
可灵大模型	基础模型+产品	可灵支持生成长达2分钟的30fps的超长视频，分辨率高达1080p，且支持多种宽高比，质量优异，已经集成在快影app中供少量用户使用	2024.6
VideoTetris	应用模型	能够直接增强现有模型的组合生成能力，还能支持涵盖多复杂指令、多场景变更等更高难度的长视频生成	2024.6
Direct-a-Video	应用模型	用户像导演一样指定一个或多个对象的运动和/或相机运动，可以分离控制对象运动和相机运动	2024.3
Video-LaVIT	应用模型	图文视频生成模型，将视频表示为关键帧和时间运动，并设计分词器适配LLM，实现视频、图像和文本的统一生成预训练	2024.4
I2V-Adapter	应用模型	该研究引入了一个创新的图像到视频转换方法，能够在不需要改变现有文本到视频生成（T2V）模型原始结构和预训练参数的情况下，将静态图像转换成动态视频。	2024.1

可灵大模型技术方案

- **模型设计**：用Transformer代替U-net，使用3D注意力机制和3D VAE压缩
- **数据建设**：建设视频数据平台，全流程自动化支持模型的训练和评估，采用多维度的视频标签体系精细筛选数据，自研captioner提升标注的完整度和准确度
- **功能方案**：可变宽高比保证输出尺寸灵活，支持多种应用模式，包括时序延展、图生视频、插帧等，为用户提供多种控制手段

可灵大模型的能力亮点

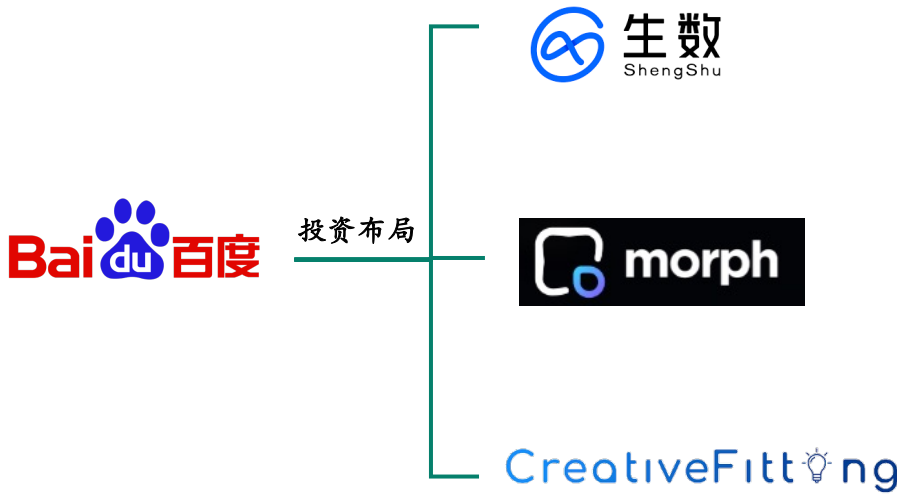
- 大幅度的合理运动的
- 模拟物理世界特性
- 电影级的画质生成
- 分钟级的长视频生成
- 强大的概念组合能力和想象力
- 支持自由的输出视频宽高比

信息来源：量子位智库，快手

互联网公司-百度：尚未推出视频生成相关产品，主要通过对外投资方式布局视频领域

- **百度在视频生成领域也发表了一定成果，但不如腾讯、阿里、字节等头部公司活跃，也没有产品化尝试**
- **目前偏好通过对外投资补足视频能力**，百度在大模型方面投资较少，在大模型领域主打全套自研，但却直接投资了生数科技，垂类应用方面投资了AI视频短剧公司井英科技，百度风投（Baidu Ventures）则投资了Morph Studio，对视频领域的关注度较高

视频相关成果	类型	简介	时间
Hallo	应用模型	首个基于扩散技术实现端到端生成高度逼真人脸视频的开源项目。用户只需提供一段音频和所选人像，即可轻松制作出具有极高真实感的人脸视频	2024.6
UniVG	基础模型	一个视频生成系统，根据生成自由度的不同，将视频生成任务分为高自由度和低自由度。高自由度的视频生成任务输入条件较弱，如文本和图像，低自由度的视频生成任务通常涉及强约束条件，如图像动画和视频超分辨率，解决方案空间较小，自由度较低	2024.1
VideoGen	基础模型	VideoGen可以生成高质量图片，引入了一个以参考图像和文本提示为条件的高效级联潜在扩散模块用于生成潜在视频，之后通过增强型视频解码器将潜在视频表示映射为高清视频	2023.9



信息来源：量子位智库

互联网公司-谷歌、Meta：发表了多项视频生成领域的前沿研究和基础模型，但整体风格偏研究向，产品化进展速度缓慢

Google

- **谷歌拥有和OpenAI同级别的资源**，发表了大量在视频生成领域的前沿论文和模型，例如Sora在其技术报告里也提及数篇来自谷歌的技术论文，是视频基础模型的有力竞争者
- **谷歌的产品化节奏较慢**：内部有多个视频团队，不够聚焦，大公司的组织结构较为低效，产品迭代速度不如OpenAI，其在2024 Google I/O上发布了旗舰视频生成大模型Veo，但距离产品化较远。此外大公司对于模型的安全可控性要求很高，视频生成模型的对齐工作也很复杂，需要考虑内容安全和数据安全，例如此前谷歌Gemini的文生图功能和SGE都因输出结果不安全和带有歧视性内容导致公关灾难
- 谷歌自身有大量的视频场景，例如Youtube是世界上最大的长视频平台，拥有高质量的创作者用户
- 对外投资方面，谷歌在23年6月投资了视频生成明星公司Runway，并成为其首选云服务商

Meta

- Meta在视频生成领域进展不如Google和OpenAI，但也发表了多项研究和基础模型，风格上存在和谷歌类似的产品化进展缓慢的问题，**但整体上AI的战略优先级和实力次于谷歌**
- 场景方面，Meta旗下的Facebook和Instagram Reels都是十亿用户级别的视频场景，但与谷歌类似，视频模型的内容可控性、对齐及组织问题可能会拖慢产品化速度

视频相关成果	类型	简介	时间
Veo	基础模型	谷歌旗舰模型，可生成长达60s的高质量视频，对标OpenAI的Sora	2024.5
vlogger	应用模型	主要基于扩散模型，只需要一张用户的头像、一段讲话录音，就能得到一个本人的演讲视频，视频时长可变	2024.3
Lumiere	基础模型	扩散模型，可以实现一件换装，视频的局部动态修改，图片转视频，视频风格化等功能，模型能够一次性生成视频中的所有帧，连贯性好	2024.1
W.A.L.T.	基础模型	李飞飞团队与谷歌合作的模型，结合了Transformer和扩散模型，可以生成3s视频，文生视频，图片转视频，3D效果	2023.12
VideoPoet	基础模型	采用多模态的Transformer 架构，可以执行各类生成任务，包括文本生成视频、图像生成视频、视频补全及视频风格转换	2023.12
Imagen Video	基础模型	基于扩散模型，能生成1280*768分辨率、每秒24帧的视频片段，Imagen Video继承自22年5月份的当时的图像生成SOTA模型Imagen	2022.10
Phenaki	基础模型	基于Transformer架构，图像加文本输入，它能够生成2分钟以上的长视频，通过输入长达200多个字符的系列提示来得到	2022.10
Flowvid	应用模型	视频生视频合成框架，主打视频风格的变换，可与现有的模型无缝协作，完成各种修改，包括风格化、对象交换和局部编辑	2024.1
Fairy	应用模型	视频生成、编辑的加速框架，大大增强了AI在视频编辑上的表现	2023.12
Emu Video+Emu Edit	基础模型	Emu Video是一种基于扩散模型的文本到视频生成方法，可以分解步骤生成高质量的视频。Emu Edit可以基于文本指令就对图像进行编辑	2023.9
Make-a-video	基础模型	根据输入的自然语言文本生成一段5秒钟左右的短视频。并且在此基础上，拓展到从图像生成视频，和从视频生成视频	2022.10

信息来源：量子位智库

内容工具软件-Adobe：自研Firefly模型，同时外接第三方视频生成模型，旨在为用户提供多样选择和最佳体验

- **总体策略**：Adobe在AI视频方面发力整个视频制作的全栈技术，不仅是单点的视频生成能力。在模型层面，Adobe采用自研+外接的方式为用户提供视频生成能力。在工作流方面，Adobe会把视频生成能力的接入现有专业级工作流（例如Adobe Premiere Pro），并为生成式AI进行产品适配
- **商业模式**：参考目前的Firefly图片生成，视频生成服务可能包括，1）提供API让企业能将模型嵌入内部工作流；2）为企业提供强化或训练模型的能力，使视频模型能够生成具有品牌特色的视频；3）按生成点数计费，普通的订阅制服务（**既有内部App集成，也有单独的网页应用**）
- 现阶段Adobe面向企业级用户、专业用户群体的产品地位难以撼动，类似视频生成的AI能力将全面搭载到Adobe产品体系中，提供便捷程度高、兼容性强的云化服务将有利于Adobe沉淀行业用户数据构建生态闭环，在未来继续保持其竞争优势

自研模型

Adobe Firefly

优势

定制化	• 企业可以使用自有版权的品牌图像、徽标、角色和风格来训练定制版 Firefly，仅供内部使用，可加强品牌内容生产，发挥知识产权价值
精细控制	• 基于Adobe数年服务客户的经验可以帮助客户实现生成的精准控制，例如结构参考（上传具有所需构图的图像，生成的每张图像都将与该结构相匹配）
商业安全	• Adobe的专有模型可以控制模型输出内容。例如Firefly的图像生成模型商业安全性极高，是基于 Adobe Stock 等授权内容和版权已过期的公共内容进行训练，对于客户来说内容安全是构建商业内容关键考量

相互补充

外接模型

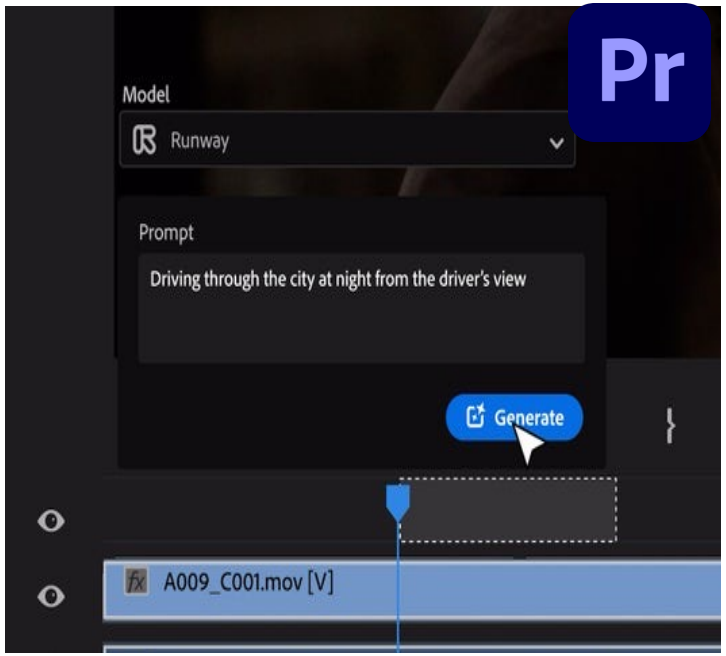
OpenAI SORA

runway

Pika

接入

- Adobe对于外接模型非常开放，认为客户将针对不同的用途使用不同的模型，将为用户提供能满足其需求的灵活选择，会把各类模型无缝接入Adobe的工作流，为用户提供最佳体验



信息来源：量子位智库，Adobe

内容工具软件-万兴科技：自研万兴天幕音视频多媒体大模型，可实现文生视频、视频生视频、文生音效、文生音乐等多媒体能力

wondershare 万兴科技

- 万兴“天幕”聚焦数字创意垂类创作场景，基于15亿创作者及100亿本土化高质量音视频数据沉淀，以音视频生成式AI技术为基础，具有多媒体、垂直解决方案以及本土化数据三大特点，全链条赋能全球创作者，让大模型应用落地更有针对性、更具实效
- 万兴“天幕”已迭代近百项音视频原子能力，其文生视频能力已实现不同风格、丰富场景及主体的连贯性；万兴“天幕”已与银行、电视台、电商和影视制作机构等多个行业领导者展开合作，助力其降本增效

万兴天幕音视频多媒体大模型

	视频模型	图像模型	音频模型
原子能力	• 文生视频 • 文生3D视频 • 视频内容编辑 • 数字人播报 • 视频换脸 • 视频调色 • 视频插帧 • 视频高光卡点 • 视频去水印 • 无限缩放	• 文/图生图 • 腹肌生成 • AI写真 • AI焕新家 • AI文字特效 • AI海报生成 • 风格迁移 • 图像分割 • 图片换脸 • 图像生成	• 文生声音 • 语言增强 • 人生分离 • 视频配乐 • 组合音乐 • 语言去噪 • 人声克隆 • 音效生成 • 节拍检测
	专业提升视频制作效率	精细化提升垂类质量	特色化补齐创作闭环
基础设施	自研推理框架	数据管理生产平台	高性能算力平台	云原生容器平台

“我们希望万兴的每一款产品，都是「技术+应用」的结合，能够解决某个细分领域的具体问题，让产品用户真正获得价值”。——万兴科技市场商务负责人意达

算力投入

- 万兴在模型层面有较大投入，2024年计划在算力上投入1亿，已经构建千卡集群及自研推理框架
- 万兴将持续加码NPU+GPU强力算力底座，积极拥抱国产算力，未来将构建万卡集群

数据建设

- 万兴有多年的数据内容积累，包括超过100万小时的视频训练数据以及创意版权库万兴喵库，本土数据+海外数据双向加持
- 万兴已组建上百人的数据标注团队和数据管理生产平台，优化模型训练数据，建立竞争壁垒

产品整合

- 万兴喵影（海外版Wondershare Filmora）是万兴面向大众用户的拳头产品，主打音视频剪辑，帮助用户对视频、照片、音频等内容进行个性化编辑、制作和转换，在视频领域已有超10年积累。万兴天幕大模型将作为技术底座为产品提供视频、图像、音频等AI生成能力

信息来源：量子位智库，万兴科技

内容工具软件-美图：模型层自研奇想大模型MiracleVision，应用方面推出设计类WHEE和AI短片制作应用MOKI输出视频能力

- 奇想大模型提供文生视频、图生视频、视频生视频等能力，Sora发布后美图升级视频大模型至MiracleVision V5，技术架构方面转向Diffusion Transformer，在语义理解、画面稳定性、动态连续性、主体一致性、内容可控性以及生成时长等方面获得全方位提升
- 应用方面，美图在设计类应用WHEE中添加了视频生成功能，在AI短片工具MOKI的工作流中集成了视频生成能力，以及在美图设计室集成了图生视频功能，侧重静态商品的视频生成，产品化节奏较快

奇想智能MiracleVision大模型

场景层	电商	广告	游戏	影视	动漫
产品层	MOKI	WHEE	美图设计室		
模型层	MiracleVision 奇想智能	DiT架构			
基座层	大数据平台	美学评估系统	标注平台		
	算法框架	分布式训练平台	算力运维平台		

业务策略

- 在基础模型方面会持续投入，但对大规模投入模型军备竞赛保持克制，**不会与头部大模型进直接竞争，主要以公司擅长的应用侧为核心，驱动订阅业务和利润增长**

数据建设

- 作为内容工具软件有多年的优质数据积累，深耕国内市场多年，与国外顶级视觉大模型相比具有明显本土化、差异化优势，**比如更擅长亚洲人像角色的处理，擅长中国传统文化元素与现代设计相结合的国风国潮视觉效果等**
- 收购站酷，获得大量图像、视频数据，可用于模型训练，为自研AI视觉大模型MiracleVision的生态带来协同效应

产品整合

- MiracleVision主要作为AI模型层底座，通过API、SDK、SaaS、模型训练等形式向行业客户、合作伙伴全面开放模型能力，助力多场景工作流，帮助企业降本增效
- 接入美团旗下各类功能产品，包括MOKI（AI短片工具）、WinkStudio（桌面端 AI视频编辑工具）、美图设计室（主打电商场景的 AI 视频工具）、WHEE（AI视觉创作工具）等

信息来源：量子位智库，美图

技术创业公司-Runway：AI视频生成的先驱产品，目前主要定位电影级视频制作，产品功能最为全面

runway

- Runway的创立时间比较长，2018年建立，A轮融资后想做轻量化的创意工具包，后来转向生成式人工智能，团队技术实力优秀，公司核心团队曾在2022年4月发表Stable Diffusion生态的核心论文High-Resolution Image Synthesis with Latent Diffusion Models》
- Runway在23年6月完成第六轮融资后的估值为15亿美元，一共筹集了2.36亿美元，接受了谷歌、英伟达的投资，**是视频生成领域目前融资规模最大的公司，团队上百人，最新产品Gen-3已经投放市场，效果优异**

2023.2	2023.6	2023.11	2024.6
Gen-1推出，主要是视频风格化和视频动态渲染	Gen-2推出，新增文生视频、图生视频、文图生视频功能	增加导演模式、运动笔刷、镜头控制等功能	Gen-3推出，新增更加细节的控制工具，大幅提升视频生成质量，如人像生成、一致性等

产品定位

- **Runway的产品主打影视、艺术制作，终极目标是可以制作两小时的电影，支持完成整个迭代制作的过程**，主要服务电影人、艺术家等创意工作者，**产品做得比较重，功能非常丰富**
- 商业模式主要是分阶梯的订阅制服务（12/28/76/125美元每月，四种），以及定制化的模型训练服务，已有服务传媒娱乐客户的案例，定制模型可以帮助客户更好地控制艺术风格、角色和叙事

运营举措

- **Runway主要从电影制作的定位出发，举办相关活动组建社区、吸引电音人、创作者**。例如推出“Runway Watch”功能，让用户分享观看通过Runway创作的长篇幅的视频作品，多次举办人工智能电影节AIFF，为获奖电影提供奖金和展映机会
- **Runway提供了非常详尽的AI视频制作和产品使用教程**

产品功能

- **Runway的产品线包括超过30种AI驱动的创作工具，覆盖视频、图像、语音、3D等模态**。除视频生成外，其产品功能包括文本-图像的转换和生成，图像-图像转换，通过文本修改图像，扩展图像，以及3D对象和纹理的生成等
- Runway还提供视频和图像编辑功能，包括背景去除、物体移除、超分，调色、动态编辑等，在视频方面提供基于时间轨道的传统视频编辑工具，并提供团队云端协作等功能

Runway合作案例

Canva	把Runway接入到Canva产品中中供上亿用户使用
media.monks	为客户提供模型和创意工具的高级权限
musixmatch	为Musixmatch社区的100多万艺术家提供Runway的生成模型来基于歌词生成音乐视频
gettyimages	在Runway的标准模型之上，用客户专有创意数据集进行微调服务以实现更符合客户需求的生成能力

信息来源：量子位智库，Runway

技术创业公司-Pika：视频生成新秀，团队背景优异，背靠硅谷，投资阵容强大，主要定位个体普通创作者

Pika

- 创始人郭文景（Demi Guo）和联合创始人兼CTO孟辰霖都是斯坦福大学AI Lab博士生退学创业，**创始人技术背景优秀，投资人团队豪华**，包括Lightspeed Venture Partners、前Github CEO Nat Friedman、Quora创始人Adam D ‘Angelo、OpenAI联创 Andrej Karpathy、Perplexity CEO等
- **团队人才密度高**，目前团队规模已经扩展到20人左右，CTO孟辰霖是DDIM、Img2Img、Model Distillation的作者，创始团队成员共取得10块IOI竞赛金牌，其他成员包括谷歌视频大模型Lumiere项目的第一作者，Hugging Face Diffusers的核心开发者等

2023.5	2023.11	2024.2	2024.6
公司成立	完成总计5500万美元的三轮融资，发布Pika 1.0	增加视频配音功能Lip Sync和Sound Effects	获得由Spark Capital领投的8000万美元融资，估值接近4.7亿美元，之前轮次的明星投资人也继续加码

产品定位

- 主要服务创意人士，包括艺术家、影视制作者，**目前主要面向个体创作者**，而不是类似好莱坞、传媒行业的中大型客户
- 重视产品的创意成分，团队有超过50%的人从事产品工作，超过1/3的人从事创意相关工作。团队邀请了斯坦福教授、计算机视觉专家Ron Fedkiw加入团队作为顾问，曾参与加勒比海盗、终结者、星球大战等知名电影的特效制作
- **具体的产品定义目前存在空间，偏好提供更好的模型能力、通用化的产品，具体的应用场景交由用户发掘，做产品留白**

运营举措

- Pika 与 ElevenLabs 联合发起了一场名为 FilmFAST 的 AI 电影比赛，为参赛者免费提供Pika和ElevenLabs产品的使用权限，向全球创作者征集优秀AI视频作者

产品功能

- 生成形式：文生视频、图生视频、视频生视频
- 生成风格：动漫卡通、穆迪、3D、水彩、自然、黏土、黑白
- 声音特效：视频场景配音（Sound Effects），人物角色配音（Lip Sync）
- 生成控制：镜头控制、长宽比控制、帧率控制、运动强度控制、自由度控制、视频延长、分辨率调整
- 商业模式目前是分层订阅制，8/28/58美元每月三个阶梯

Pika Labs合作案例

ElevenLabs	与Eleven Labs进行合作把声音克隆、音频生成能力引入产品，例如Lip Sync功能
Adobe	Pika的视频生成模型作为Adobe Premier Pro的外接模型集成到Adobe的工作流
G!lab	Pika与华为、商汤科技、小米等公司共同作为流浪地球制作方旗下电影工业化实验室G!Lab的合作伙伴

信息来源：量子位智库，Pika

技术创业公司-爱诗科技、生数科技：本土视频生成头部公司，已获得数亿元融资，视频生成效果优秀

爱诗科技

产业背景强

- 创始人王长虎曾任字节跳动视觉技术负责人，参与了抖音、TikTok 等视频产品的建设和发展，公司23年4月成立，视频生成产品 PixVerse 于2024年1月正式发布，是国内较早押注视频生成领域的创业公司

2023.4	2023.8	2024.1	2024.3	2024.4	2024.6	2024.7
公司成立	完成千万级天使轮融资	视频生成产品Pixverse发布	完成超亿元级A2轮融资	完成亿级A1轮融资	推出新功能MagicBrush	发布Pixverse V2

产品介绍	• 目前主打产品Pixverse主要面向海外市场，**已经有超过100万用户**，可以在在Web端和Discord平台进行使用 • **产品的长期定位在市场规模广阔的短视生态，创始人之前有很强的短视频技术背景**，在视频处理、工程化方面经验丰富 • 24年5月，智源研究院发布的大模型评测结果显示在提供公开服务的文生视频大模型中，Runway 和爱诗科技的 PixVerse 处于第一梯队，模型生成能力较好
产品功能	• 文生视频：提供写实风格、卡通风格、3D风格、CG风格等生成风格，以及多种宽高比选择 • 图生视频：动态笔刷、镜头控制、动作强度控制、生成质量选择 • 角色生视频：可以生成视频写真，上传人物图片进行动态化，针对人物视频生成进行了增强

生数科技

科研背景强

- 核心成员来自清华大学人工智能研究院和多个海内外顶级学术机构和企业，早在2021年就开始扩散模型研究，拥有针对深度生成式模型的骨干网络、高速采样、可控生成、大模型训练等全栈底层原创研发能力

2023.3	2023.6	2023.8	2024.3	2024.4	2024.6
公司成立	完成近亿元天使轮融资	数千万天使+融资	数亿元天使++融资	发布视频生成模型Vidu	数亿元Pre-A轮融资

产品介绍	• **视频生成大模型Vidu**：该模型采用团队原创的Diffusion与Transformer融合的架构U-ViT，支持一键生成长达32秒、分辨率高达1080P的高清视频内容，能够模拟真实物理世界，还拥有丰富想象力，具备多镜头生成、时空一致性高等特点 • 文生图产品Pixweaver：支持个性化的视觉创作，融合多元风格，出色的语义理解和丰富的细节表现，具备艺术级美学水准 • 3D资产创建工具VoxCraft：通过输入文本或图像，仅需数分钟即可生成高质量的3D模型，支持灵活定制、兼容传统工业管线
科研积累	• 团队长期致力于贝叶斯机器学习的基础理论和高效算法研究，是目前在扩散概率模型领域发表论文成果最多的国内团队 • 生数团队在通用架构、高速采样、高效训练、可控生成、多模态训练、强化学习、基础理论等方面**发表了近20篇论文**

信息来源：量子位智库，爱诗科技，生数科技

技术创业公司-大模型创业公司：各家的研发资源、技术路径、业务侧重各异，在视频生成模型方面选择不同

选择因素

研发资源

- **国内大模型厂商的算力资源相对海外有限**，优秀视频生成模型的训练和推理目前需要大量数据资源和算力成本
- 视频模型方面大模型厂商并没有特别的资源优势，组织上偏好聚焦自身核心业务

技术路径

- 大模型厂商对于视频生成模型在通用人工智能路线上的发挥的作用判断各异，技术路径上存在非共识

业务侧重

- 大模型厂商在业务上的聚焦和侧重不同，例如部分厂商侧重B端的行业应用，或者侧重C端生产力场景
- 目前视频生成的需求和场景尚不清晰，对大模型厂商来说一般不是核心方向

智谱·AI	✓	布局较多，7月发布了视频生成产品清影，可以实现文生视频和图生视频功能，并投资了视频生成创业公司生数科技
Moonshot AI	✓	会探索研发视频生成大模型和多模态路线，但是业务侧重上会长期聚焦Kimi
商汤 sensetime	✓	商汤科技在4月发布了文生视频生成平台，强调对人物、动作、场景的可控性（尚未投放市场）
MINIMAX	✓	8月发布文生视频模型和的文生图模型
零一万物	✗	是否研发视频生成大模型未知
阶跃星辰	✗	是否研发视频生成大模型未知
百川智能 BAICHUAN AI	✗	百川智能明确表示不会做类似Sora的视频生成大模型，将聚焦于语言模型路线，公司业务侧重医疗行业

信息来源：量子位智库

垂类创业公司-FancyTech：借助AI视频生成能力，主打实物还原与控制，结合广告营销行业理解，深度服务头部品牌客户

FancyTech

- 核心团队为阿里系，创始人有资深的阿里从业背景，对于电商营销、广告有深刻的行业认知和积累。FancyTech目前主要为客户提供基于AIGC技术的营销视频和广告视频服务，客户主要是各行业的头部品牌
- **FancyTech的商业化非常成功，在23年已经实现月收入破千万且仍在快速增长，是目前视频生成领域营收规模最大的公司之一**

产品模式

电商场景	AI生成新素材	• 完成电商平台账号授权后，AI系统能够自主学习商品详情、收集基础素材，并根据这些信息生成视频脚本驱动其他图像模型，生成视频中的各种元素，并输出给文字模型生成标题文案，从而完成整个视频的制作 • 适合没有广告素材的商家，可以借助AIGC，实现“从0到1”的高性价比电商广告视频制作
	AI素材组合	• 授权AI访问直播间、小红书或抖音等平台的相关资料，然后AI自主学习资料并自动生成视频脚本、完成配音 • 适合已经拥有大量素材的客户，商家无需上传任何新素材，也无需进行人工的视频剪切操作，AI可以对这些繁杂的素材进行智能重组和放大，“从有到优”地生成高质量的视频内容
	服装行业模特视频生成	• 客户需要提供产品的白底图，由AI技术生成其余的所有内容，包括模特、穿搭示范以及动态效果，每次生成都是全新效果，模特的外貌、服装、背景等都会有所变化
广告场景	广告视频拍摄	• 根据广告主的产品实物，使用激光雷达精确扫描商品，形成全新的3D模型，作为AI系统的重要输入项用于生成高质量的广告视频 • AI能够学习并识别扫描结果中的颜色、形状、花纹等细节信息，并根据这些信息生成逼真的素材和生动到位的文字、配音

应用现状

市场	• **FancyTech目前主要的市场在国内，主要走高举高打的路线，已经在服装、户外、数码家电、美妆个护、视频酒水等行业服务了大量头部品牌**，面向大中型客户，未来将发力出海市场 • 通过服务头部客户可以进一步积累经验和标杆效应，为未来扩大规模服务中小客户打磨产品 • FancyTech是第一家获得由法国奢侈品巨头LVMH 集团创新大奖的中国公司，已经打出标杆性效应
商业模式	• **商业模式方面客单价较高，年期服务费较高在10万左右**，具体价格根据客户实际需求浮动 • 除为客户提供电商营销视频和广告视频生成外，也提供针对客户特有数据集进行调优训练和模型私有化部署的服务（主要还是针对投放量大的客户，成本具有规模效应且倾向于拥抱AI）

FancyTech的优势

- **深度行业知识**：对不同电商平台和营销有深度理解，例如淘宝、抖音、小红书营销风格各异，可以做优秀的跨平台运营
- **服务灵活度**：可以和客户进行深度共创，精细匹配客户需求，把产品和服务内嵌到客户的生产工作流中，为每一个品牌做个性化服务

信息来源：量子位智库，FancyTech

垂类创业公司-布尔向量：为用户提供轻便强大的营销视频生成工具，助力营销场景降本增效，在本土和海外市场迅速增长

BOOLV

- 一家专注于研发AI营销视频的跨境电商2B SaaS软件技术服务公司，成立于2021年，获得多家知名一线美元基金投资，核心团队来自腾讯、字节等头部公司。**公司聚焦基于AI与数据挖掘技术自动生成商用短视频解决方案，服务于跨境电商品牌客户**，愿景是用AI和数据挖掘技术实现商业化视频内容生产，业务覆盖本土和海外市场

产品模式

海外市场	SaaS服务	• 布尔向量为用户提供一键生成营销短视频的功能，具体包括一键URL生成视频，Idea生成视频，脚本生成视频，Blog生成视频，图像生成视频等。产品可以把商品链接中的产品介绍、功能解读、介绍图示、视频等信息进行解析整理，生成可用于投放的营销短视频 • 为用户提供丰富的视频生成模版、素材库和功能灵活的视频编辑器，用户可以使用类似传统时间轨道的轻量化编辑器对视频进精确编辑，选择最佳的风格、背景元素和产品形象 • 市场方面主要依靠产品自然增长，主打用户体验走PLG的路线，通过社交媒体、Product Hub等平台进行传播增长，**主要客户画像是一些各行各业的中小规模商家**
本土市场	定制化服务	• 商业模式方面，定制化服务的客单价可以达到数十万，具体的需要按照客户对视频需求量决定，**主要客户是行业的一些KA，对于AI视频生成的接受度较高，有一定的付费实力和意愿** • 定制化服务不会涉及产品的二次开发，主要是为用户提供产品的使用支持，或者结合客户的数据对模型进行调整 • 本土客户的需求比较复杂，变化很多且频繁，难以用标准化产品满足，例如在具体的设计和素材和视频的内容细节上要求较多，在服务过程中会涉及和客户的深度沟通，需要消耗一定的人力和时间

布尔向量的服务优势

行业知识	• **营销短视频的核心在于唤起观众的购买行为，促成投放ROI的增长**，很多商家并不清楚什么的好的营销视频，商品的视频展示有一些内在逻辑和清晰的结构，例如有不同的镜头语言、信息的传达方式，具体的组合形式需要大量的营销经验来判断用户的偏好，需要来自专业视频营销公司的行业知识 • **在训练模型过程中，需要和客户进行共创**，总结过往的投放经验，确认合适的投放内容，不同的产品、投放渠道和目标人群投放内容各异，布尔向量可以补足客户的营销知识盲区
技术匹配	• 功能方面，不同于类似文生视频模型，组合类的视频生成模型基于多模态机器学习技术，**有很多参数可以进行调节，在生成方面的控制精准且灵活** • 成本方面：相比于生成式大模型（Transformer、Diffusion等）视频生成成本极低，没有推理成本的限制

借力基础模型

- 布尔向量在产品中**也会整合文生视频能力，但主要目的是为用户提供额外选择，在模型层面非常灵活**，可以接入其他视频生成基础模型赋能业务，和大模型没有竞争关系，属于模型能力进步的受益方

信息来源：量子位智库，布尔向量

AI视频生成玩家图谱

本土互联网公司

ByteDance 字节跳动
Tencent 腾讯
阿里巴巴
快手
Baidu 百度
HUAWEI

内容工具软件

wondershare 万兴科技
meitu
Clipfly
新壹科技
LibAI Lab 李白人工智能实验室

本土技术创业公司

爱诗AIsphere
生数 ShengShu
智象未来 HiDream.ai
morph
Vega AI
艺映AI
七火山
极佳 GigaAI
YoYo 鹿影科技 Möbius
深氧科技
NeverEnds
商汤 sensetime
智谱·AI
徐图智能

本土垂类公司

FancyTech
BOOLV
奥创光年
CreativeFitting

开源类项目

北京大学 PEKING UNIVERSITY
上海人工智能实验室 Shanghai Artificial Intelligence Laboratory
潞晨科技
Tencent

海外大公司

OpenAI
Google
Meta
Adobe

海外技术创业公司

runway
Pika
stability.ai
Midjourney
genmo
Haiper
LUMA AI
LTX Studio
Odyssey
KREA
Kaiber
Leonardo.Ai
Moonvalley

海外垂类公司

Fliki
invideo AI
Peech
PICTORY

信息来源：量子位智库

量子位 智库
insights

关于量子位智库：

量子位旗下科技创新产业链接平台。致力于提供前沿科技和技术创新领域产学研体系化研究。

面向前沿AI&计算机，生物计算，量子技术及健康医疗等领域最新技术创新进展，提供系统化报告和认知。

通过媒体、社群和线下活动，基于专题技术报道及报告、专项交流会等形式，帮助决策者更早掌握创新风向。

量子位智库公众号

微信号：Qbitbot020
量子位智库小助手

关于量子位：

量子位（QbitAI），专注人工智能领域及前沿科技领域的产业服务平台。

全网订阅超过500万用户，在今日头条、知乎、百家号及各大科技信息平台量子位排名均为科技领域TOP10，
内容每天可覆盖数百万人工智能、科技领域从业者。

量子位公众号

分析师：Xuanhao（微信：feeltheagi）　智库负责人：李根（微信：ligen603）　商务合作：赵萌（微信：13343397239）