国海证券
SEALAND SECURITIES

证券研究报告
海外
2023年11月2日

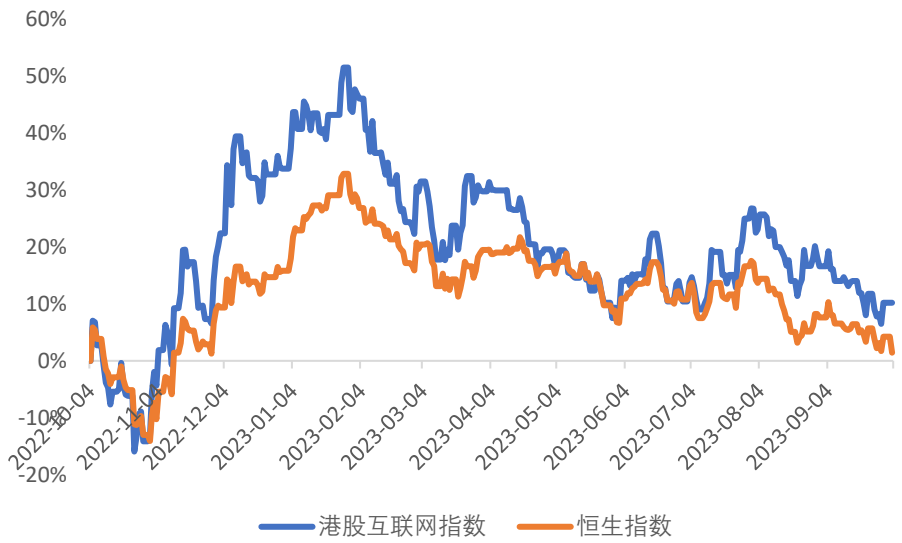
AIGC专题报告：从文生图到文生视频——技术框架与商业化

评级：推荐(维持)

陈梦竹(证券分析师)
S0350521090003
chenmz@ghzq.com.cn

尹芮(证券分析师)
S0350522110001
yinr@ghzq.com.cn

最近一年走势



相对恒生指数表现（%）

表现	1M	3M	12M
港股互联网	-8.60	14.79	50.09
恒生指数	-5.72	-10.23	1.47

相关报告

《从Roblox进化看AIGC游戏未来—Roblox(RBLX.N)深度复盘：二十年沉淀，四阶段演绎（增持）*海外*杨仁文，马川琪，陈梦竹，姚蕾》——2023-09-24

《创新奇智（02121）动态研究报告：业绩维持高速增长，“AI+制造”赋能长期发展（买入）*IT服务Ⅱ*陈梦竹》——2023-09-14

《商汤-W（00020）2023H1业绩点评：生成式AI有望成为业务新驱动，整体亏损收窄（增持）*IT服务Ⅱ*陈梦竹》——2023-09-08

《网易-S（9999.HK）公司动态研究：利润超预期，新游表现强劲，期待后续业绩释放（买入）*游戏Ⅱ*陈梦竹，尹芮》——2023-09-03

《焦点科技（002315）2023H1财报点评：业务调整拖累营收增速，利润端实现稳健增长（买入）*互联网电商*陈梦竹》——2023-09-03

核心提要

1、底层模型技术框架梳理

文生图和文生视频的底层技术框架较为相似，主要包括GAN、自回归和扩散模型三大路径，其中扩散模型（Diffusion model）为当前主流生成模型，多个指标对比下综合占优，能在较为可控的算力成本和较快的速度下生成具备多样性、高质量的图像：①图像质量：扩散模型>自回归模型>GAN模型。FID值（Fréchet Inception Distance score）是用于评估模型生成的图像质量的指标，是用来计算真实图像与生成图像的特征向量间距离的一种度量。FID值越小，可以认为图像质量在一定程度上越优。从不同模型的FID得分来看，扩散模型平均数较小，反应图像质量较高。②参数量：自回归模型>扩散模型>GAN模型。GAN的参数量一般在千万级别，整体较为轻巧，扩散模型的参数量在十亿级别，自回归模型在十亿到百亿级不等。③生成速度（由快到慢）：GAN模型>扩散模型>自回归模型。生成速度与参数量级为负相关关系。④训练成本：自回归>扩散模型>GAN模型。由于参数量级较小，GAN模型训练成本小且开源模型多，仍具备一定优势。而自回归模型参数量级较大，整体训练成本更高。在单张A100GPU下，120亿参数的DALL-E需要18万小时，200亿参数的 Parti更是需要超过100万小时，扩散模型参数量在十亿级别，整体训练成本较为适中。

2、商业化模式及成本拆分

- 文生图商业化模型：当前主要的商业化方式包括基于GPU时间/生成次数/API接口调用/个性化定价等方式。根据我们调研，以Midjourney为例，单张图片生成成本约0.03~0.04美金，单张收入约0.05美金，毛利率约30%~40%，净利率约20%。
- 文生图领域整体创业门槛低于大语言模型：①模型层看：图像生成领域已有生成质量较高的开源预训练模型Stable Diffusion，且SD具有较为丰富的开发者生态，有许多插件供选择。创业公司可基于Stable Diffusion基础版本进行进一步调优和个性化数据训练。②成本端看：从主流模型参数规模看，文生图参数量级多在1-10B之间，而通用大模型入门级门槛达到了70B，文生图整体参数量级较小，成本远低于通用大模型。通过调研文生图初创公司，实际小团队利用开源模型，初期在用户不到1万情况下甚至无需购买A100，通过购买RTX30\40系列、IBS3060（5000~1w/张）也可以启动。我们对文生图推理算力需求也进行了测算，以10亿级参数量的模型、在100万DAU的用户量级为例，若想控制单次推理延迟时间，需部署约143张A100，整体芯片算力需求低于大语言通用模型。
- 文生图商业模式仍存疑问，长期竞争需要技术+产品+场景能力结合突破：①对于垂类AI应用：短期看头部应用通过技术/产品/成本/数据等优势突破，在C端率先开启变现，长期看针对垂类场景C端天花板相对明确，搭建工程化能力可技术输出到B端场景，探索更多变现可能。②对于现有应用叠加AI功能：短期通过AI功能引入提升产品体验和用户粘性；长期看基于现有高频场景，用户壁垒更强、不易流失，用户ARPU和付费率有望提升。

3、文生图代表模型及应用

从模型和应用看，海外OpenAI、谷歌、微软、Meta、Midjourney、Stability AI都推出了各自的文生图模型，国内百度、美图、万兴科技、新国都等均推出各自AI应用。从生成效果看Midjourney、Adobe和Stable Diffusion综合较优，OpenAI最新升级DALL-E3模型将与ChatGPT集成，多模态交互能力持续提升，有望带来新的场景突破。

4、行业评级及理由：文生图和文生视频底层技术不断演进、模型持续迭代，涌现出一批优质原生AI应用，在C端开创了全新的应用体验，同时在B端游戏、营销、影视制作、文旅、电商等多个行业均开启应用，实现降本增效，长期有望进一步打开商业化空间。我们看好AI多模态行业投资机会，维持行业“推荐”评级，建议关注微软、Meta、Adobe、谷歌、百度、阿里巴巴、美图、万兴科技、新国都等相关标的。

5、风险提示：竞争加剧风险、内容质量不佳风险、用户流失风险、政策监管风险、变现不及预期风险、估值调整风险等。

报告目录

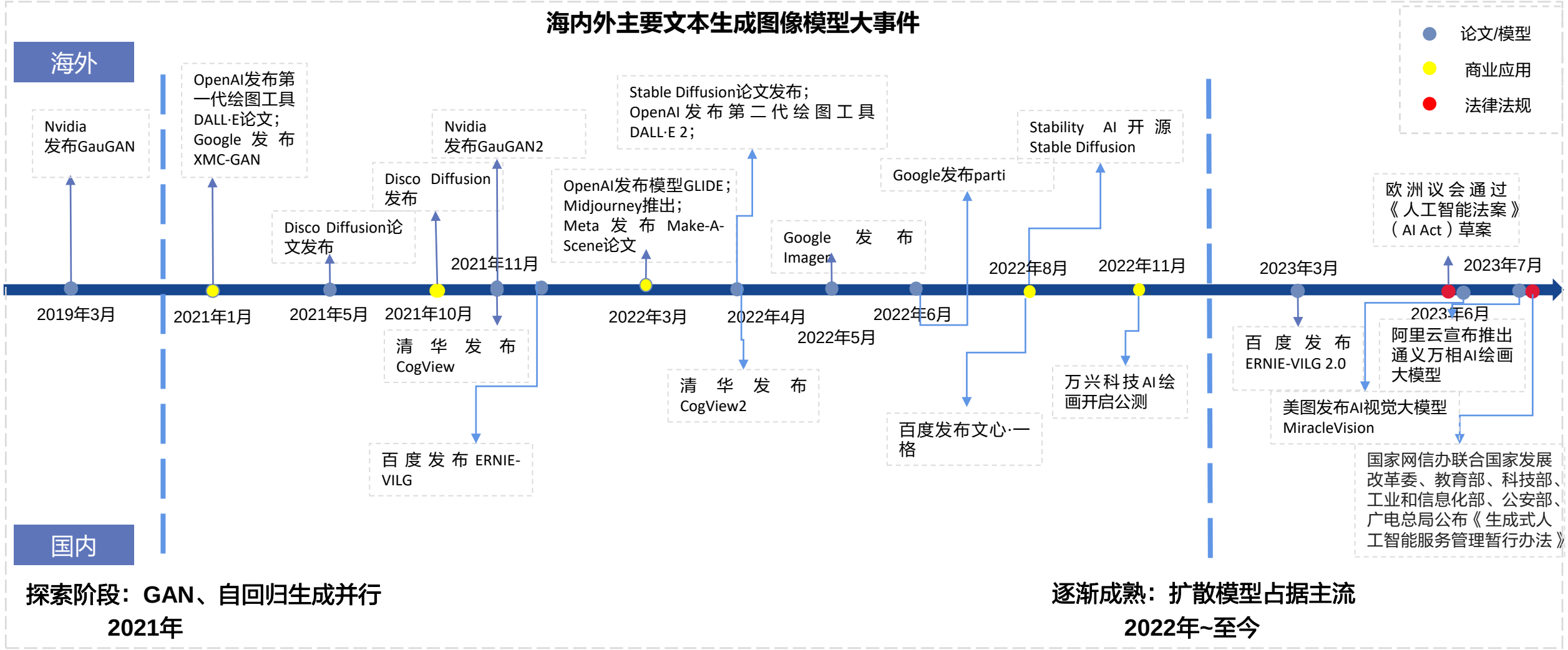
报告目录

底层模型技术框架梳理

文生图：基于文本生成图像，Stable Diffusion开源后迎来快速发展

文生图（Text-to-Image）是基于文本通过生成式AI生成图像的模式。近3年时间，文生图的技术已实现大幅的进步，海外的Stable Diffusion、Midjourney已经能够提供较高质量的图像，国内的万兴科技的万兴爱画、百度的文心·一格也投入商用。文本生成图像的底层模型可以分为GAN、扩散模型、自回归模型三类。目前行业内的明星模型主要基于扩散模型。

海内外主要文本生成图像模型大事件



探索阶段：GAN、自回归生成并行
2021年

逐渐成熟：扩散模型占据主流
2022年~至今

资料来源：论文见附录1，网信中国，央视网，36氪，新智元，智元社区，百度官网，澎湃新闻，证券时报，潮新闻客户端，界面新闻，百度AI微信公众号，百度智能云，国海证券研究所

文生视频：与文生图底层技术一致，自回归和扩散模型为主流

文生视频（Text-to-Video）是基于文本通过生成式AI生成视频的模式。随着文生图技术的精进与成熟，对于文生视频的技术的发展和关注逐渐演变及增加，近3年时间，以Runway为代表的文生视频公司在不断涌现，互联网行业的巨头，比如谷歌、Meta、微软，同样投入人员和精力参与其中，国内目前文生视频技术还在初期发展阶段，目前魔搭社区（Model Scope）里的开源模型ZeroScope表现亮眼。文本生成视频模型的发展经历三个阶段：图像拼接生成阶段、GAN/VAE/Flow-Based生成阶段、自回归和扩散模型阶段。

文本生成视频三大发展阶段

图像拼接生成阶段
2016年以前

GAN/VAE/Flow-Based生成阶段
2016年~2019年

自回归模型及扩散模型生成阶段
2020年~至今

2016.10 VGAN
2017.04 VideoGPT
2017.08 TGAN
2017.12 MoCoGAN
2019.09 DVD-GAN
2021.04 微软亚洲研究院发布GODIVA
2021.11 微软亚洲研究院发布NUWA（女娲）
2022.02 DIGAN
2022.05 清华发布CogVideo
2022.07 微软亚洲研究院发布NUWA-Infinity
2022.09 Meta发布Make-A-Video
2022.10 Google发布Phenaki
2022.10 Google发布Imagen Video
2023.11 字节跳动发布MagicVideo
2023.02 Runway发布Gen-1
2023.03 Runway发布Gen-2
2023.03 微软亚洲研究院发布NUWA-XL
2023.04 Nvidia发布Video LDM
2023.05 Nvidia发布PYoCo

2016.01 2017.01 2018.01 2019.01 2020.01 2021.01 2022.01 2023.01

生成技术路径：从GAN到Diffusion，模型持续优化迭代

	生成式对抗网络（GAN）	自回归模型	扩散模型
结构	•**生成器（Generator）：**一个神经网络或者其他方式拟合出的函数，给定输入，负责生成整个GAN所需的输出 •**判别器（Discriminator）：**一个判断输入真假的二分类器函数	•**Transformer：**整体主要分为Encoder和Decoder两大部分，能够模拟像素和高级属性（纹理、语义和比例）之间的空间关系，利用多头自注意力机制进行编码和解码	•通过对纯高斯噪声反向降噪推断来生成图像
运行原理	•生成器将抓取数据、产生新的生成数据，并将其混入原始数据中送交判别器区分。这一过程将反复进行，直到判别器无法以超过50%的准确度分辨出真实样本	•通过编码器将文本转化成token或序列，应用自回归预测，经过训练好的模型解码输出图像	•定义一个扩散步骤的马尔可夫链，逐渐向数据添加随机噪声，然后学习逆扩散过程，从噪声中构建所需的数据样本
存在问题	•**训练不稳定：**GAN 的相互博弈过程容易造成训练不稳定，使得训练难以收敛。近期突破思路有Relativistic GAN。有别于传统 GAN 测量生成样本是否为真的概率这一做法，Relativistic GAN 将关注点放在测量生成样本比真实样本“更真”的概率，使得 GAN 获得了更好的收敛性 •**生成样本大量重复相似：**模式坍塌被认为是应用 GAN 进行图像生成时最难解决的问题之一，它会造成训练结果冗余、生成图像质量差、样本单一等问题。近期突破性思路有包含两个判别网络的D2GAN	•**计算成本消耗大：**模型受制于计算效率与训练数据的规模，自回归模型的参数通常是扩散模型参数量的10倍以上 •**大量的训练数据：**自回归模型需要大规模的、高质量的数据进行训练，尤其在文本生成视频的训练中，目前缺少高质量的文本-视频对是文生视频自回归模型的一大难题	•**采样速度慢：**连续模型使用高斯噪声，很难处理离散数据 •**计算消耗过大：**蕴含多个（原始模型可能要上千个）前向传播过程，对显卡硬件需求大，计算较慢
改进方向	•**结构改善：**将 GAN 与机器学习中最新的理论成果进行交叉训练，引入迁移学习、强化学习等，使 GAN 与计算机图形学等技术更好地融合，推动结构的改善 •**模型压缩：**目前图像生成技术想要落地，势必要根据需求调整模型的大小，结合基于“知识蒸馏”的模型压缩方法，进行匹配需求的优化和压缩，从而内嵌入小型软件中，拓宽应用领域	•**创新生成模式：**改进图像生成的方式，比如在视频生成过程中，从“逐像素”改进至逐帧生成，减少计算量 •**提升数据质量：**在文本生成视频中，联合文本-图像对进行训练，规避因为文本-视频对不足的劣势	•**训练方式改进：**知识蒸馏促进模型压缩和加速，改进扩散过程以减少采样时间，调整噪声尺度优化采样程序，数据分布替换降低预测误差 •**无训练采样：**以更少的步骤和更高的精度从预训练的模型中获取知识 •**混合模型改进：**在扩散模型的中加入额外生成模型，以利用其他模型的高采样速度 •**分数与扩散统一：**确定扩散模型和去噪分数匹配之间的联系，有助于统一广义扩散的加速方法
图像	StackGAN++、DF-GAN	DALL-E、CogView、 CogView2、Parti、CM3leon	Stable Diffusion、GLIDE、DALL-E 2
视频	VGAN、TGAN、VideoGPT、MoCoGAN、DVD-GAN、DIGAN	CogVideo、GODIVA、NUWA、Phenaki	Video Diffusion Model、Make-A-Video、Imagen Video、Tune-A-Video、Dreamix、NUWA-XL、Text2Video-Zero、VideoLDM、PYoCo
商用			图像：Midjourney；Stable Diffusion；文心一格 视频：Runway

文生图模型竞争格局

国海证券 SEALAND SECURITIES

	国内	海外
应用	文心一格 AI艺术和创意辅助平台；万兴爱画；WHEE AI视觉创作的灵感激发器；无界AI 创作无限，以致涌现；通义万相；meitu；TIAMAT	DALL·E 2；Midjourney；Clipdrop by stability.ai；Adobe Firefly；runway；dreamstudio.io
代表模型	StackGAN++、DF-GAN	DALL-E、CogView、CogView2、Parti、CM3leon	Stable Diffusion、GLIDE、DALL-E 2
底层架构	GAN（生成式对抗网络）	自回归模型	扩散模型

人工智能监管：中欧美均发布相关条例，引导生成式AI规范发展

国海证券 SEALAND SECURITIES

人工智能主要发展地区的监管发展

时间	地区	法律、法规及监管条例发布	主要内容
2022年10月	美国	美国白宫发布《人工智能权利法案蓝图》	提出了建立安全和有效的系统、避免算法歧视，以公平方式使用和设计系统、保护数据隐私等五项基本原则，且将公平和隐私保护视为法案的核心宗旨，后续拟围绕这两点制定完善细则。
2023年1月	美国	美国商务部下属机构美国国家标准与技术研究院（NIST）发布《人工智能风险管理框架》	鼓励用户全面规划人工智能系统，包括预期的商业目的和使用人工智能可能造成的潜在危害。要求有道德的人工智能从业者确定如何以定量和定性的方式衡量人工智能系统所产生的影响。组织将使用测量的结果来帮助其持续管理人工智能系统：RMF框架为用户提供了管理已部署人工智能系统风险的工具，并根据评估的风险和风险优先级分配风险管理资源。
2023年6月	欧洲	欧洲议会通过《人工智能法案》（AI Act）草案	全球范围内首部系统化规制人工智能的法律，草案提出对人工智能采取分级管理的思路，基于人工智能的四个风险等级（从低风险或无风险、有限风险、高风险、不可接受风险）进行区别管理，要求生成式人工智能的设计和开发符合欧盟法律和基本权利。
2023年7月	中国	国家网信办联合国家发展改革委、教育部、科技部、工业和信息化部、公安部、广电总局公布《生成式人工智能服务管理暂行办法》	提出国家坚持发展和安全并重、促进创新和依法治理相结合的原则，采取有效措施鼓励生成式人工智能创新发展，对生成式人工智能服务实行包容审慎和分类分级监管，明确了提供和使用生成式人工智能服务总体要求。

资料来源：中国网信网，全球技术地图，AI时代前沿，国海证券研究所

GAN：通过生成器和判别器对抗训练提升图像生成能力

GANs（GAN, Generative Adversarial Networks），生成对抗网络是扩散模型前的主流图像生成模型，**通过生成器和判别器进行对抗训练来提升模型的图像生成能力和图像鉴别能力**，使得生成式网络的数据趋近真实数据，从而图像趋近真实图像。

GAN常见的模型结构

- **单级生成网络：代表有DF-GAN等。**只使用一个生成器、一个鉴别器、一个预训练过的文本编码器，使用一系列包含仿射变换的UPBlock块学习文本与图像之间的映射关系，由文本生成图像特征。
- **堆叠结构：多阶段生成网络，代表有stackGAN++、GoGAN等。**GAN 对于高分辨率图像生成一直存在许多问题，层级结构的 GAN 通过逐层次，分阶段生成，一步步提生图像的分辨率。在每个分支上，生成器捕获该尺度的图像分布，鉴别器分辨来自该尺度样本的真假，生成器G1接收上一阶段的生成图像不断对图像进行细化并提升分辨率，并且以交替方式对生成器和鉴别器进行训练。多阶段GAN相比二阶段表现出更稳定的训练行为。（**一般来说，GAN的训练是不稳定的，会发生模式倒塌的现象mode collapse，即生成器结果为真但多样性不足**）

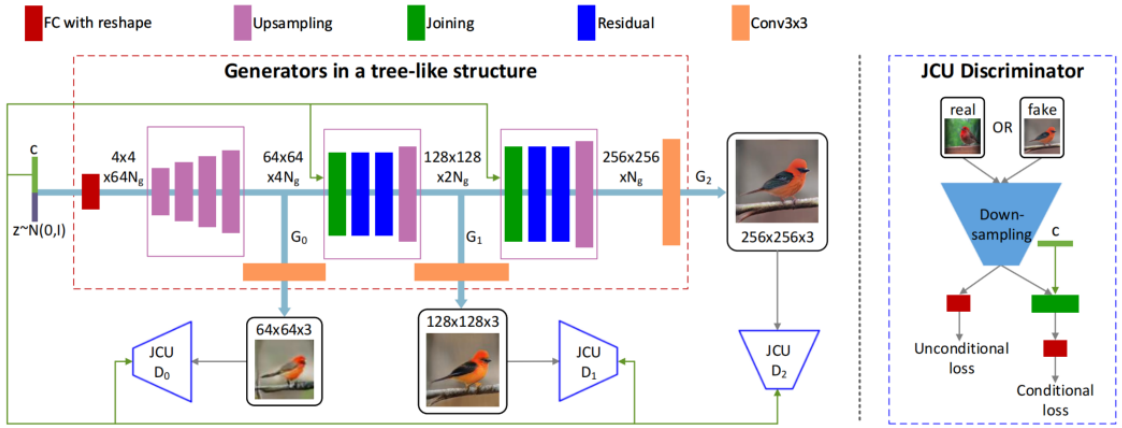


Fig. 2: The overall framework of our proposed StackGAN-v2 for the conditional image synthesis task. c is the vector of conditioning variables which can be computed from the class label, the text description, *etc*.. N_g and N_d are the numbers of channels of a tensor.

stackGAN++的文字生成图片架构原理

生成对抗网络实现文本生成图像主要分为三大部分：文本编码器、生成器和鉴别器。文本编码器由RNN或者Bi-LSTM组成，生成器可以做成堆叠结构或者单阶段生成结构，生成模型捕捉样本数据的分布，不断生成图像，判别模型判别输入是来自是真实数据还是来自生成模型，鉴别器用于鉴别生成器生成的图像是否为真和是否符合文本语义。**两者在对抗中，不断提升各自的能力**，生成器逐渐提升生成图像的能力，生成图像的分布接近真实图像分布，从而提高判别器的判别能力，判别器对真实图像和生成图像进行判别，来提高生成器的生成能力。

- **GAN的特点：**相比于其他模型，GAN的模型参数量较少，比较轻便，因此GAN擅长对单个或多个对象类进行建模。但由于训练过程的不稳定性，扩展 GAN 需要仔细调整网络架构和训练因素，扩展到复杂数据集则极具挑战性，稳定性较差、生成图像缺乏多样性。

GAN：在早期文本生成视频领域也有所应用

在图像拼接阶段后，生成对抗网络（GAN）开始应用在文本生成视频领域，因为它们可以在没有第一帧的情况下执行无条件或类条件视频合成，但由于其稳定性不足，逐渐被自回归模型和扩散模型替代。

经典GAN模型在视频领域应用梳理

模型名称	发布时间	生成方式	IS*（↑）
VGAN	2016.10	第一个将GAN用于视频生成的模型，它将视频分解为静态背景和移动前景，通过分别生成背景和移动对象来生成视频。生成器由两个卷积网络组成：第一个是3D时空卷积网络，用于捕获前景中的移动对象，而第二个是静态背景的2D空间卷积模型。从双流生成器生成的帧被组合在一起，然后馈送到鉴别器以区分真实视频和虚假视频。	8.31 ± .09
VideoGPT	2017.04	模型采用了变分自动编码器（VAE）和生成对抗网络（GAN）从文本中提取静态和动态信息，静态特征用于草绘文本条件背景颜色和对象布局结构。通过将输入文本转换为图像过滤器来考虑动态特征，模型从公开可用的在线视频中自动创建匹配文本视频语料库。	24.69 ± .30
TGAN	2017.08	模型学习未标记视频的语义表示，使用由3D反卷积层组成的单个生成器生成视频，模型利用了两种不同类型的生成器：时间生成器和图像生成器。时间生成器将单个潜在变量作为输入并输出一组潜在变量，每个潜在变量对应于视频中的一个图像帧，图像生成器将一组此类潜在变量转换为视频。	11.85 ± .07
MoCoGAN	2017.12	MoCoGAN将视频分解成内容和运动两个部分，通过将一系列随机向量映射到一系列视频帧来生成视频。每个随机向量由一个内容部分和一个运动部分组成。当内容部分保持固定时，运动部分通过随机过程实现。	12.42 ± .07
DVD-GAN	2019.09	双视频鉴别器GAN（DVD-GAN）基于复杂的数据，从噪声矢量生成视频，生成48帧高达256 * 256的高质量图像。DVD-GAN是在Kinetics-600 数据集上训练的，以前的工作仅使用子集和预处理的样本。与MoCoGAN 类似，有两个鉴别器来处理视频的时间和空间方面。	32.97 ± 1.7
DIGAN	2022.02	模型将隐式神经表示应用于视频编码，包含（a）基于隐式神经表征（INR）的视频生成器，它通过以不同的方式操纵空间和时间坐标来改善运动动态，以及（b）运动鉴别器，无需观察整个长帧序列即可有效识别不自然运动。可以在 128 ×128分辨率的 128 帧视频上进行训练，比之前最先进的方法的 48 帧长 80 帧。	29.71 ± .53

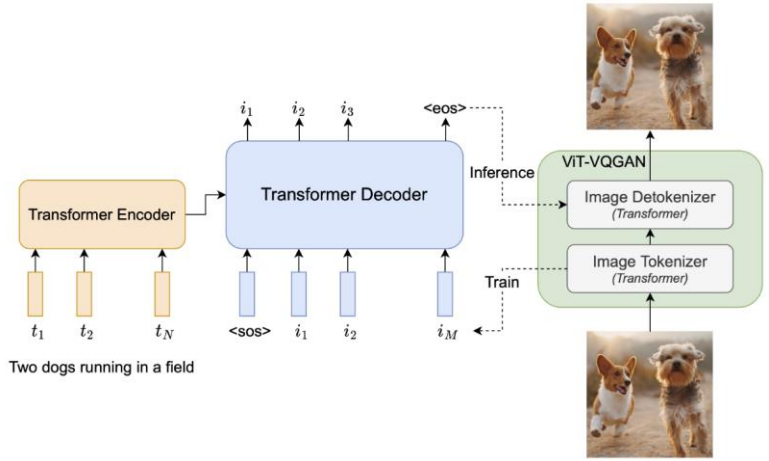
*模型在 UCF-101 数据集上的 IS 得分（分值越高越好）

自回归模型：采用Transformer结构中的自注意力机制

自回归模型（Auto-regressive Model）采用Transformer进行自回归图像生成。Transformer整体主要分为Encoder和Decoder两大部分，能够模拟像素和高级属性（纹理、语义和比例）之间的空间关系，利用多头自注意力机制进行编码和解码。采用该架构模型的文生图**通常将文本和图像分别转化成tokens序列，然后利用生成式的 Transformer 架构从文本序列（和可选图像序列）中预测图像序列，最后使用图像生成技术（VAE、GAN等）对图像序列进行解码，得到最终生成图像。**

经典自回归模型

- **结合ViT-VQGAN：**谷歌Parti将ViT-VQGAN作为图像标记器将图像编码为离散标记序列（使用Transformer GPT-2作为编码生成工具），它将文本到图像的生成视为序列到序列的建模问题，图像标记序列是目标输出，并利用其将此类图像标记序列重建为高质量、视觉多样化图像。首先将一组图像转换为一系列代码条目，类似于拼图，然后将给定的文本提示转换为这些代码条目并「拼成」一个新图像。



- **结合VQ-VAE（矢量量化变分自动编码器）：**清华的 CogView 与百度的 ERNIE-ViLG 均使用 VQ-VAE + Transformer 的架构设计。先将文本部分转换成token，然后将图像部分通过一个离散化的AE(Auto-Encoder)转换为token，将文本token和图像token拼接到一起，之后输入到GPT模型中学习生成图像。

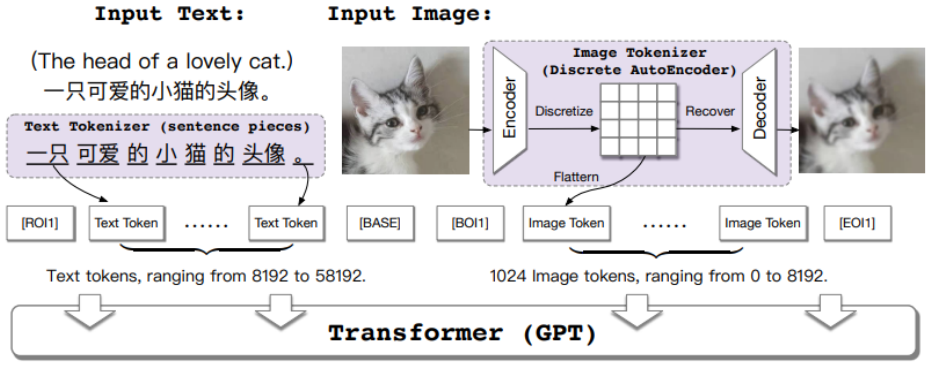


Figure 3: The framework of CogView. [ROI1], [BASE1], etc., are seperator tokens.

- **自回归模型的特点：**1）相比于其他模型，自回归模型的稳定性及生成图像的逻辑相对合理。2）但计算效率总体较低，生成速度较慢，训练成本相对较高，其实际应用受限于计算效率和训练成本相对不足，目前Meta发布的CM3leon在计算效率有较大的提高，优化了模型的计算速度。

资料来源：《Hierarchical Text-Conditional Image Generation with CLIP Latents》Aditya Ramesh等，《CogView: Mastering Text-to-Image Generation via Transformers》Ming Ding等，Google Research，易观，电脑学习网，机器之心，国海证券研究所

自回归模型：生成视频相比GAN更加连贯和自然

与GANs相比，自回归模型具有明确的密度建模和稳定的训练优势，自回归模型可以通过帧与帧之间的联系，生成更为连贯且自然视频。但是自回归模型受制于计算资源、训练所需的数据、时间，模型本身参数数量通常比扩散模型大，对于计算资源要求及数据集的要求往往高于其他模型，随着扩散模型的火热，自回归模型的热潮逐渐降低，基于文本生成图像的文本生成视频的热潮渐起。

早期自回归模型：生成质量差，成本高

早期的视觉自回归模型，如，以“逐像素”的方式进行视觉合成。然而，其只能应用于低分辨率图像或视频，并且难以放大，是因为模型高维视觉数据上的计算成本很高。

- 模型代表： PixelCNN，PixelRNN，Image Transformer，iGPT 和 Video Transformer

中期自回归模型：生成质量提升，生成规模增大

随后，VQ-VAE 作为一种离散视觉标记化方法的出现，高效和大规模的预训练可以应用于图像和视频的视觉合成任务。

- 模型代表：GODIVA、VideoGPT

自回归模型现状：高质量生成，训练成本降低

后续，出现利用预先训练的文本到图像模型来生成高质量的视频的方法，视频被视为图像的时间序列，这减少了视频token的数量，并降低了训练成本。但这可能导致动力学建模不佳并产生运动伪影。

- 模型代表：NUWA、CogVideo、Phenaki

资料来源：《NUWA: Visual Synthesis Pre-training for Neural visUal World creAtion》Chenfei Wu等，《CogVideo: Large-scale Pretraining for Text-to-Video Generation via Transformers》Wenyi Hong等，国海证券研究所

扩散模型：当前主流路径，通过添加噪声和反向降噪推断生成图像

扩散模型（Diffusion Model）是通过定义一个扩散步骤的马尔可夫链，通过连续向数据添加随机噪声，直到得到一个纯高斯噪声数据，然后再学习逆扩散的过程，经过反向降噪推断来生成图像，通过系统地扰动数据中的分布，再恢复数据分布，逐步优化过程。

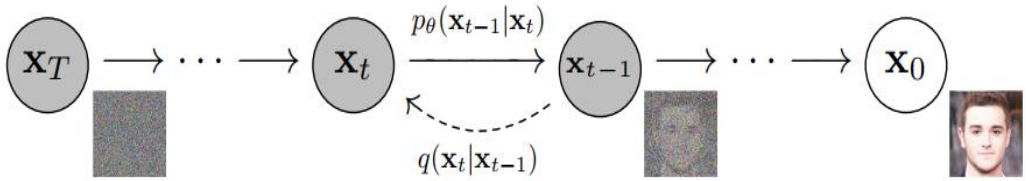
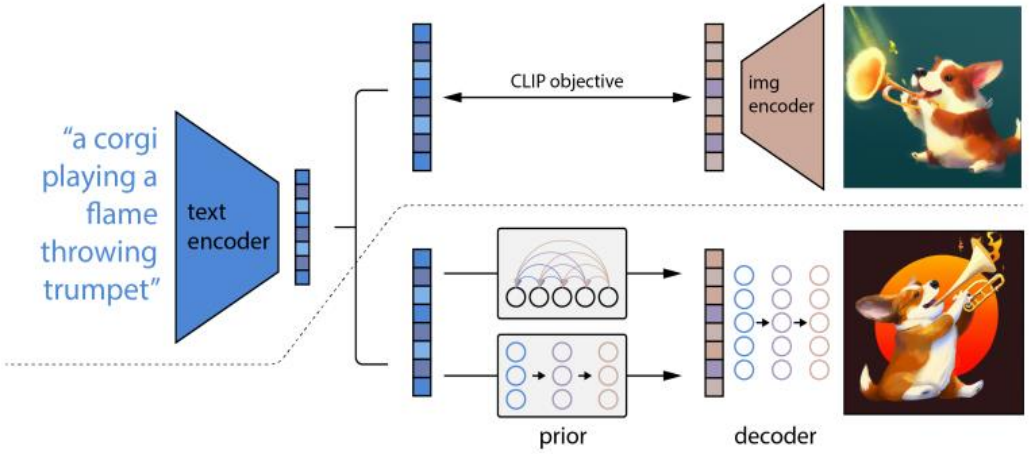
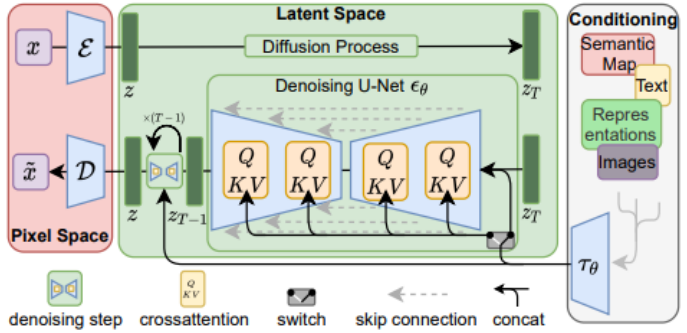


Figure 2: The directed graphical model considered in this work.



扩散模型在图像生成应用

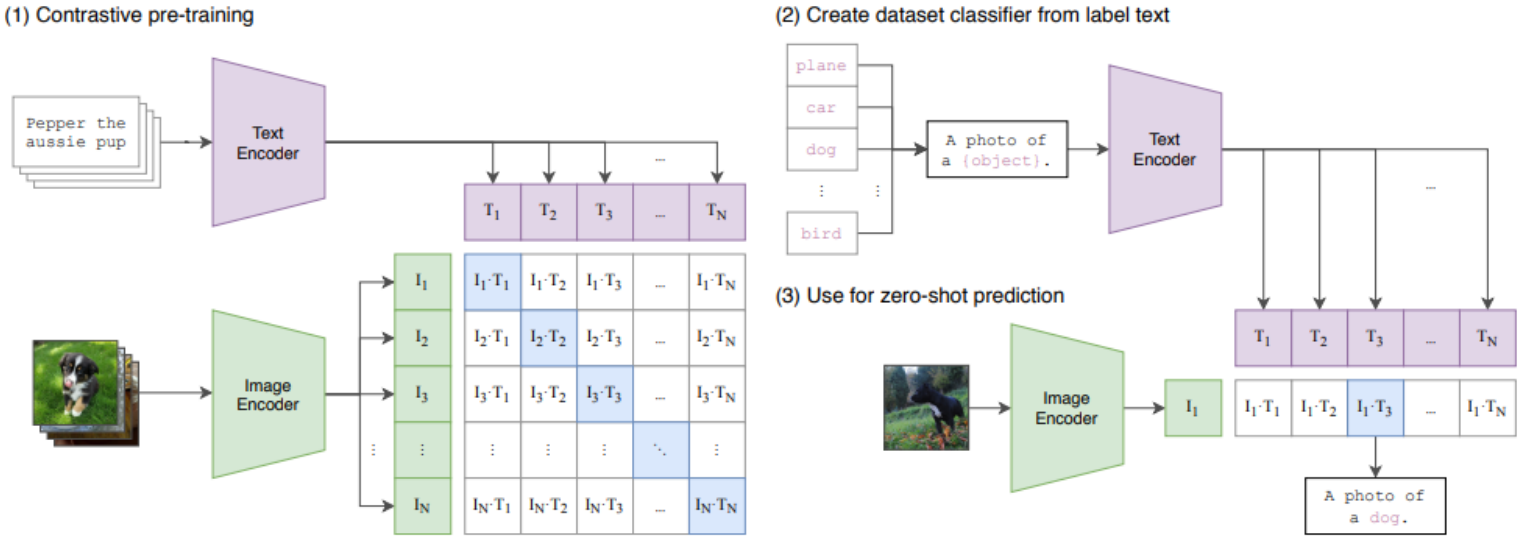
- **结合CLIP：**比如DALLE-2，采用Diffusion Model结合CLIP， CLIP文本嵌入首先被馈送到自回归或扩散先验以产生图像嵌入，然后该嵌入用于调节扩散解码器，后由扩散解码器产生最终图像。
- **结合潜在空间（Latent Space）：**Stable Diffusion将模型应用于预训练自动编码器的潜在空间（Latent Space），这使得扩散模型的训练能够在一个有限的计算资源的环境下进行，并且能够保持图像的质量和灵活性。Latent Diffusion Models通过在一个潜在表示空间中迭代“去噪”数据来生成图像，然后将表示结果解码为完整的图像，**让文图生成任务能够在消费级GPU上，在10秒级别时间生成图片，大大降低落地门槛。**



- **扩散模型的特点：**
- 1）相比先前的模型，扩散模型在训练稳定性和结果准确性能力提升明显，**替代了GAN成为目前主流模型**。2）当应对大量跨模态图像生成的需求，通过结合CLIP，能够实现图像生成速度和质量的显著提升，生成的图片具有较好的多样性和写实性。3）相比于其他模型，扩散模型有较强的表现及相对中等的计算成本。

CLIP：实现文本和图像特征提取和映射，训练效果依赖大规模数据集

CLIP（Contrastive Language-image Pre-training）是**基于对比学习的文本-图像跨模态预训练模型**，由文本编码器（Text Encoder）和图像编码器（Image Encoder）组成，编码器分别对文本和图像进行特征提取，将文本和图像映射到同一表示空间，通过文本-图像对的相似度和差异度计算来训练模型，从标签文本创建数据集分类器，从而能够根据给定的文本生成符合描述的图像。



预训练模型：

预先在数据量庞大的代表性数据集上训练模型，当迁移到自定义的数据中，通过权重和偏差调优后，使模型达到需要的性能 。

预训练模型能够节省从零开始的高昂时间成本和计算成本，降低模型对标注数据数量的要求，能够处理一些难以获得大量标注数据的场景。

CLIP的特点

- **优点：**由于CLIP完成了基于多模态的对比学习和预训练，在过程中已经将文本特征和图像特征进行对齐，该模型无需事先标注数据，减少了标注数据的工作量及对应成本，能够在零样本图像文本分类任务中高质量运行。
- **缺点：**1）CLIP在包含时间序列数据和需要推理计算的任务中由于模型本身的局限性，生成图像的效果不佳。2）CLIP的训练效果依赖大规模的文本-图像对数据集，对训练资源的消耗比较大，**CLIP是由OpenAI团队通过4亿对图像-文本对训练后提出的。**

资料来源：巴比特资讯，Nvidia Developer，机器之心，《Improving Language Understanding by Generative Pre-Training》，《 Learning Transferable Visual Models From Natural Language Supervision 》，国海证券研究所

扩散模型：当前也为文生视频主流技术路径

当前主要的文本到视频模式主要采用基于扩散的架构，由于扩散模型在图像生成方面的成功，其启发了基于扩散模型的视频生成的模型。Video Diffusion Model的提出标志着扩散模型在视频生成领域的应用，该模型将扩散模型拓展到视频领域。

经典扩散模型在视频生成领域应用梳理

模型名称	组织	发布时间	生成方式
Video Diffusion Model	Google	2022.04	视频扩散模型（Video Diffusion Model）是标准图像扩散架构的自然延伸，是首个将扩散模型延展到视频生成领域的模型，模型支持图像和视频数据的联合训练，这能够减少小批量梯度（Variance of minibatch）的方差并加快优化，生成长和更高分辨率的视频。
Make-A-Video	Meta	2022.09	Make-A-Video通过时空分解扩散模型将基于扩散的T2I模型扩展到T2V，利用联合文本-图像先验来绕过对配对文本－视频数据的需求，这使得潜在地扩展到更多的视频数据。
Imagen Video	Google	2022.10	Imagen Video基于Imagen图像生成模型，采用级联扩散视频模型，并验证了在高清视频生成中的简单性和有效性，文本生成图像设置中的冻结编码器文本调节和无分类器指导转移到视频生成仍具有有效性。
Tune-A-Video	新加坡国立大学、腾讯	2022.12	Tune-A-Video是第一个使用预训练T2I模型生成T2V的框架，引入了用于 T2V 生成的一次性视频调谐的新设置，消除了大规模视频数据集训练的负担，提出了有效的注意力调整和结构反转，可以显著提高时间一致性。
Gen-1	Runway	2023.02	Gen-1将潜在扩散模型扩展到视频生成，通过将时间层引入到预训练的图像模型中并对图像和视频进行联合训练，无需额外训练和预处理。
Dreamix	Google	2023.02	Dreamix提出了第一个基于文本的真实视频外观和运动编辑的方法，通过一种新颖的混合微调模型，可显著提高运动编辑的质量。通过在简单的图像预处理操作之上应用视频编辑器方法，为文本引导的图像动画提供新的框架。
NUWA-XL	微软亚洲研究院	2023.03	NUWA-XL是一种“扩散超过扩散”（Diffusion over Diffusion)的架构，“从粗到细”生成长视频，NUWA-XL支持并行推理，这大大加快了长视频的生成速度。
Text2Video-Zero	Picsart AI Resarch (PAIR)，UT Austin，U of Oregon，UIUC	2023.03	Text2Video-Zero提出零样本的文本生成视频的方法，仅使用预先训练的文本到图像扩散模型，而无需任何进一步的微调或优化，通过在潜在代码中编码运动动力学，并使用新的跨帧注意力重新编程每个帧的自我注意力，强制执行时间一致的生成。
VideoLDM	英伟达	2023.04	VideoLDM提出了一种有效的方法用于训练基于LDM的高分辨率、长期一致的视频生成模型，主要是利用预先训练的图像DM并将其转换为视频生成器通过插入学习以时间一致的方式对齐图像的时间层。
PYoCo	英伟达	2023.05	PYoCo提出一种视频扩散噪声，用于微调文本到视频的文本到图像扩散模型，通过用噪声先验微调预训练的eDiff-I模型来构建大规模的文本到视频扩散模型，并实现最先进的结果。

模型对比：扩散模型图像质量最优，自回归模型相对训练成本最高

国海证券 SEALAND SECURITIES

①图像质量：扩散模型>自回归模型>GAN模型

FID值（Fréchet Inception Distance score）是用于评估模型生成的图像质量的指标，是用来计算真实图像与生成图像的特征向量间距离的一种度量。如果FID值越小，则相似程度越高，可以认为图像质量在一定程度上越优。从不同模型的FID得分来看，扩散模型平均数较小，反应图像质量较高。

②参数量：自回归模型>扩散模型>GAN模型

GAN的参数量一般在千万级别，整体较为轻巧，扩散模型的参数量在十亿级别，自回归模型在十亿到百亿级不等。

③生成速度（由快到慢）：GAN模型>扩散模型>自回归模型

生成速度与参数量级为负相关关系。

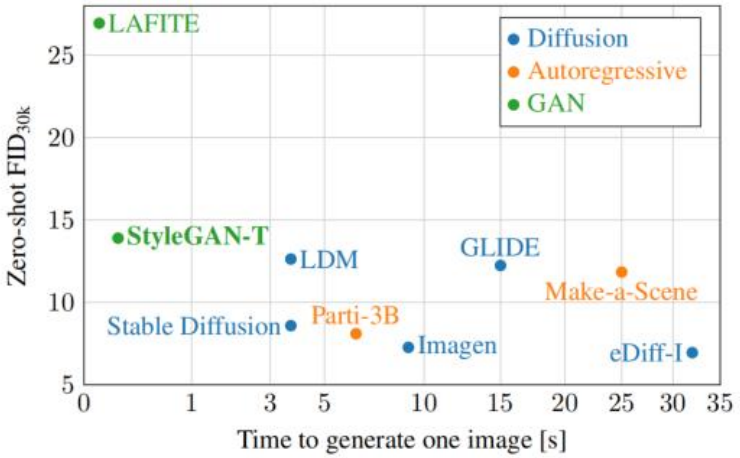
④训练成本：自回归>扩散模型>GAN模型

由于参数量级较小，GAN模型训练成本小且开源模型多，仍具备一定优势。而自回归模型参数量级较大，整体训练成本更高。在单张A100GPU下，120亿参数的DALL-E需要18万小时，200亿参数的 Parti 更是需要超过100万小时。扩散模型则较为适中。

模型名	模型类型	参数量级
GLIDE	扩散模型	35亿
DALLE-2	扩散模型	35亿
Imagen	扩散模型	34亿
Re-Imagen	扩散模型	36亿
DALLE	自回归模型	120亿
Cogview	自回归模型	40亿
Cogview2	自回归模型	60亿
Parti	自回归模型	200亿
DFGAN	生成对抗网络	0.19亿

主要图像生成模型比较

	扩散模型	自回归模型	GAN模型
图像质量	优	良+	良
参数量	中	差	优
生成速度	中	差	优
易扩展性	中	中	优
优势原因	基于马尔可夫链的正向及反向扩散过程，未对图片进行降维压缩，能够更加准确地还原真实数据，对图像细节的保持能力更强，具备多样性和真实感	Transformer的多头自注意力机制提升了生成的稳定性，可以实现更大规模参数的模型训练	生成器和判别器动态对抗的特点，避免了马尔科夫链式的学习机制，无需在学习过程中进行推断
优点	生成的质量高	相比GAN生成质量较高，采样速度较快，灵活的生成分布更加均匀	采样速度较快，灵活的设计框架
缺点	大量扩散步骤导致采样速度慢、模型成本较高	需要将图像转为token进行自回归预测，采样速度慢、模型成本高	可解释性差，容易出现模式崩溃



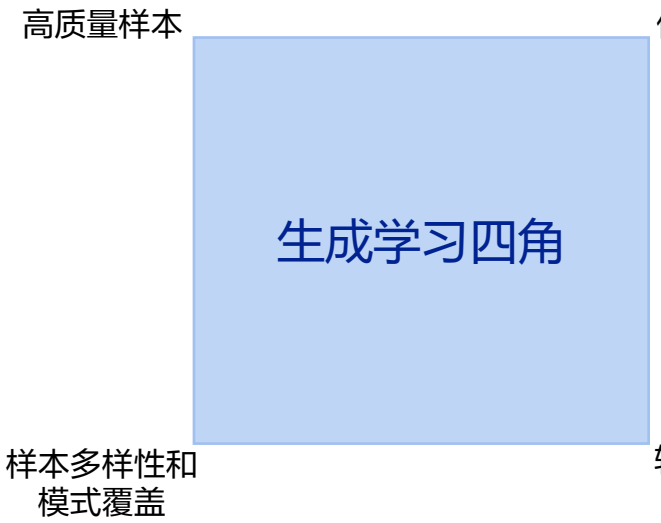
Model	Model type	Zero-shot FID_{30k}	Speed [s]
LDM	Diffusion	12.63	3.7
GLIDE	Diffusion	12.24	15.0
DALL·E 2	Diffusion	10.39	–
Stable Diffusion *	Diffusion	8.59	3.7
Imagen	Diffusion	7.27	9.1
eDiff-I	Diffusion	**6.95**	32.0
DALL·E	Autoregressive	27.50	–
Ernie-ViLG	Autoregressive	14.70	–
Make-A-Scene *	Autoregressive	11.84	25.0
Parti-3B	Autoregressive	8.10	6.4
Parti-20B	Autoregressive	7.23	–
LAFITE	GAN	26.94	**0.02**
StyleGAN-T *	GAN	13.90	0.10

* downsampled to 256×256 pixels using Lanczos　　– not available

图像生成模型的困境：多个指标中求取平衡，目前Diffusion综合占优

生成式学习框架有四个关键要素：样本高质量、样本多样和模式覆盖、低廉的计算成本和快速的计算能力，目前没有一个模型能够充分满足四个要素。

	高质量样本	样本多样行和模式覆盖	低廉的计算成本	较快生成速度	现状
扩散模型	√	√	×	√（部分）	图片生成质量较高且速度尚可，具有较强的多样性，目前是主流模型，但模型成本相较GAN仍然偏高。
自回归模型	√	难以同时满足	难以同时满足	×	样本多样性和低廉计算成本难以同时满足，自回归模型先验的学习使用的是文本到中间离散表征的映射，依赖于大规模数据集，导致其很难在低廉的计算成本下产生较为多样的样本。
生成对抗网络（GAN）	√	×	√	√	能够快速生成高质量样本且成本低，但模式覆盖率较差，容易出现模式崩塌。



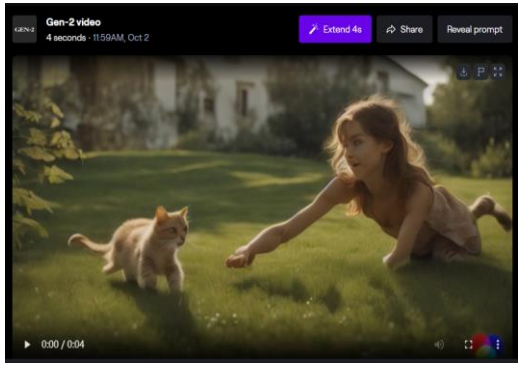
技术改进探索

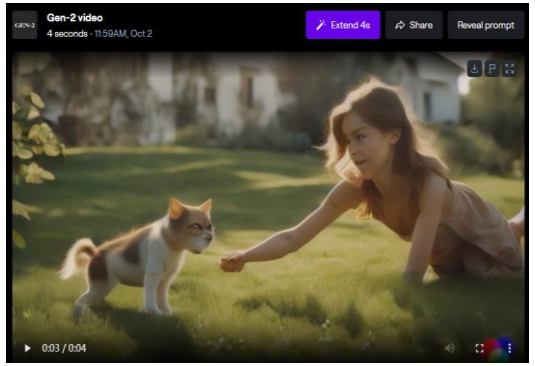
- OpenAI提出的全新图像生成模型Consistency Models，不仅能够解决**扩散模型**迭代步骤多、采样速度慢的问题，并且无需对抗训练可以直接生成高质量样本，可以快速完成图像修复、图像超分辨率等多种图像任务，表现出了更强的应用潜力。
- Meta的CM3Leon采用了基于 token 的**自回归模型**方法，但计算量仅相当于以往基于Transformer 方法的五分之一，因此既具备自回归模型的功能多样性和有效性，也保持着较低的训练成本和良好的推理效率，并获得了 4.88 的 FID。
- **GAN的潜力仍然存在：**来自浦项科技大学（韩国）、卡内基梅隆大学和Adobe研究院的研究人员提出了一种全新的生成对抗网络架构GigaGAN，打破了模型的规模限制，在推理速度和图像生成效果方面展现了更好的性能，对应解决传统的GAN在增加架构容量导致的不稳定问题，可以看到GAN在图像编辑、图像转换等场景的应用潜力仍然存在。

文本生成视频模型仍存在许多技术难点，生成效果有待提升

缺少大规模、高质量的文本-视频对

在文本生成图像的情景中，收集高质量的文本-图像对是可能的，但是高质量的文本-视频对是稀缺的，许多数据库中的视频很少和文本强相关，并且许多只描述了场景而缺少时间信息。文本生成视频模型需要大量数据来学习字幕相关性、帧照片写实感和时间动态，但与丰富的图像数据资源相比，视频数据在样式、数量和质量方面受到更多限制。除此以外，视频片段的长度是不等的，为了训练将视频切成固定帧数的片段，会破坏文本和时间之间的“对齐”情况（文本和时间信息不匹配），进而影响模型的训练。

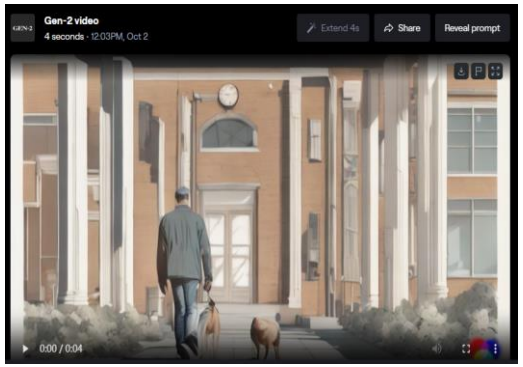


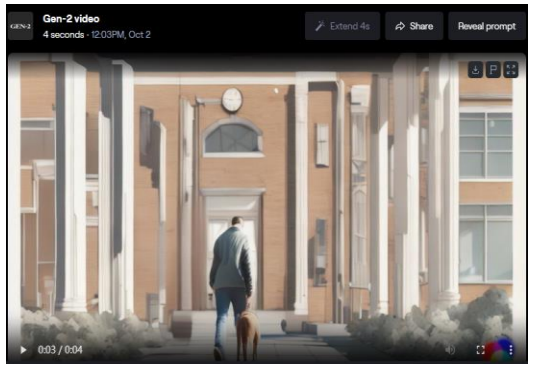


Prompt：a girl is chasing a cat on the grass,full shot,classic

高维度视频数据建模的复杂性

视频生成除了考虑空间信息，还需要考虑时间信息，高质量的视频的生成需要更高强度的计算及复杂的推理能力，在考量视频质量时，视频长度、逼真度、连贯性目前还无法完全达到。被人眼识别为连贯的视频需要帧率为每秒24帧以上，目前在帧率上虽然技术有达到，但是图像质量和前后帧的逻辑联系等仍有待进一步改进。





Prompt：a man is walking the dog in the school,wide angle,cinematic

用户Prompt表达不确定性

用户在进行文字描述（prompt）时，通常有个性化的用语和表达方式，模型对于文字描述的理解（prompt）会较大的影响生成，同时在此中，模型可能无法详细理解多主体交互关系、动作在时间轴上的演进、一词多义等。用户在表达时，可能会出现要素的缺失、描述模糊等情况，致使模型没有获得足够的信息进行生成，而产生用户预期与模型生成的差异。同样，AIGC平台是否能够“突破”用户表达的瓶颈，生成更具创新性的内容，提高优质内容的含量，也是目前的困境。

商业化模式及成本拆分

文生图商业化模型

主要成本费用拆解

算力成本

主要包括芯片采购和电费
前期训练+后期推理调用

运维成本

训练数据集获取成本

人力成本

模型开发训练等人员

运维成本

产品运营和维护

销售费用

产品买量、营销等

主要商业化方式

基于GPU时间

基于生成次数

个性化定价

API接口调用

主要应用场景

图像分类

图像分割

可以在工业、工程设计等领域辅助进行目标识别、图像相似度检索，辅助CAD设计等；在医学领域可以帮助进行医学影像标注、解剖、病理结构变化分析等。

图像生成

图像风格转换

在艺术设计、产品设计、动画与游戏制作等方面均有商业化潜力，可以根据设计者的草稿图、创意概念来生成图像，以及图像合成、图像编辑、增强图像艺术性等，从而能够帮助设计师、建模师进行动漫人物、游戏场景的制作，帮助完成海报、产品LOGO和产品包装设计等工作。在电商的应用方面，图像生成可以在虚拟试衣间、模拟商品展示等场景提升用户的在线购物体验。

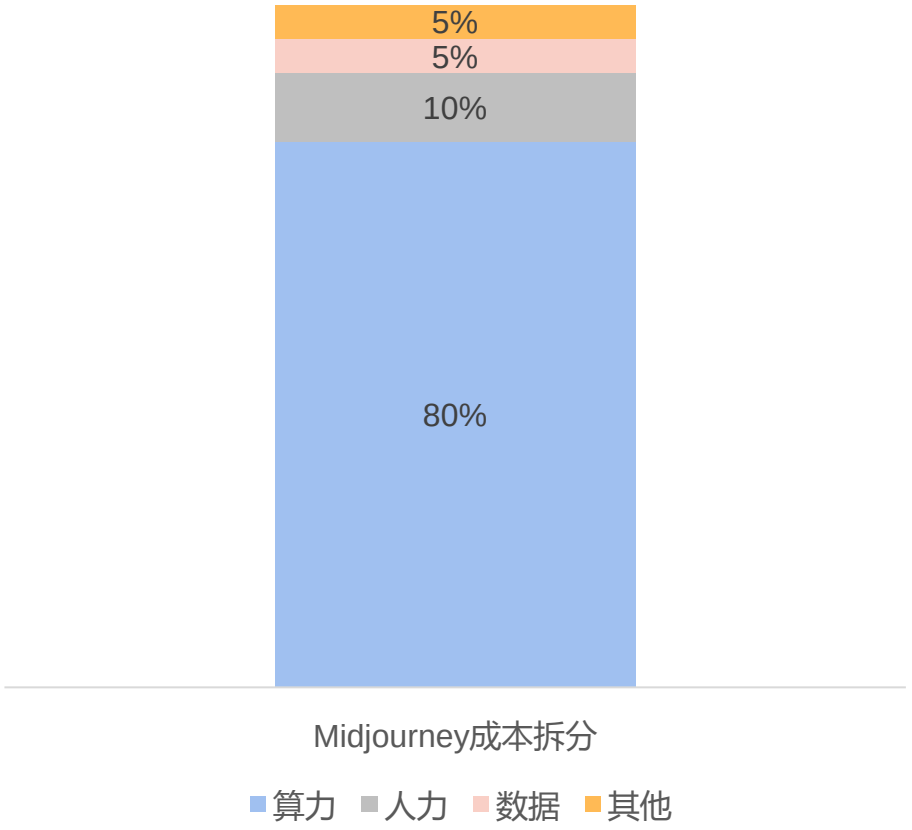
图像修复

能够根据已有图像的上下文信息修复缺失部分，对图像对比度、锐度或色彩鲜艳度等图像要素的增强等，可以应用于数字化历史文献的修复、图像修补等。在摄影与影视制作方面应用于老照片、老电影的修复和画质提升。

图像超分辨率

能够从低分辨率图像恢复和重建高分辨率图像，在医学影像处理场景中能够帮助医生创建逼真的病例和解剖结构，生成CT扫描图像，辅助进行病情的分析诊断。另外，在天文观测和卫星遥感观测等方面，利用图像超分辨率能够提升成像设备的性能，并克服时间差异、气象变化等因素引起的图像场景变化，为天文探索发现增加了更多可能性。

图片生成模型成本拆分：以Midjourney为例



数据成本

数据购买相对一次性，约1000万~2000万美金，假设每年摊销500万美金

人力成本

硅谷一线公司比如OPENAI或者Midjourney人均人力成本大概是80-90万美金/年，目前Midjourney总共11个员工，人力成本约1000万美金/年。

芯片投入

考虑Midjourney庞大的用户规模，按照使用1万张英伟达A100卡计算总成本约1.8-1.9亿美金左右，按照3年折旧摊销一年平均约花费6000万美金。

电力消耗

按照A100算力的每一张卡的功率是250瓦，一年大概需要400万美金的电费。

总成本约7500~8000万美金/年
约0.03~0.04美金/张图片

年收入约1亿美金
约0.05美金/张图片

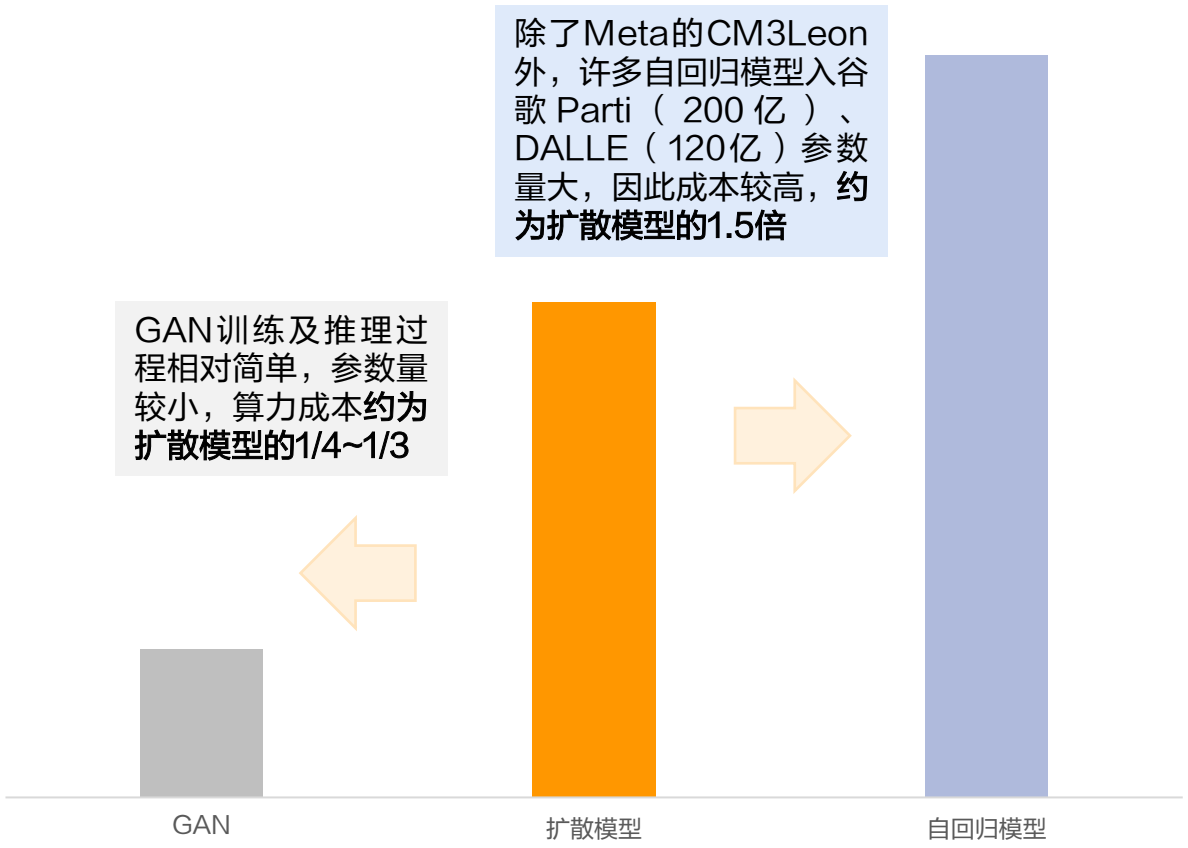
毛利率~30%-40%

净利率~20%

平均来看自回归模型成本最高，生成视频成本远高于生成图片

国海证券 SEALAND SECURITIES

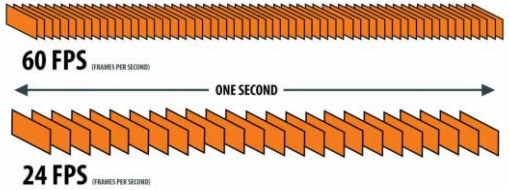
不同文生图模型的算力成本消耗对比

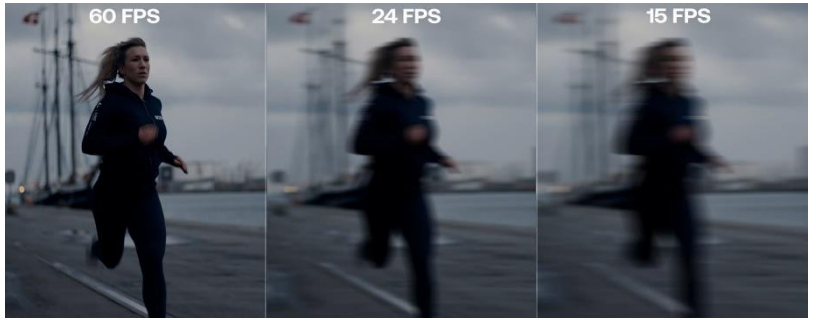


但在实际模型应用中，成本不仅取决于参数量大小，也取决于训练时间和用户规模。
前期训练阶段，若对模型训练时间没有要求，可以通过延长训练时间降低GPU成本；若对训练时间要求较短，则需要布局更多芯片提高训练速度。
上线阶段，如果用户体量很大，比如OpenAI和Midjourney规模用户体量，线上运营推理的成本可能占到整体成本80-90%，训练阶段成本只占10-20%。

文生视频的成本可能为文生图24倍以上

- 人眼看到的视频是透过一连串的静态影像连续快速播放的结果，由于每一张静态画面的差异很小，因此连续快速播放时，一张张快速闪过的静态画面在人眼视网膜上产生“视觉暂留”现象，原本静态的图像仿佛连贯运动了起来。
- 通常来说，人看到视频是连贯的需要帧率为每秒24帧以上，电影放映的标准也是每秒24帧以上。如果文生图一次性消耗的算力是一个单元，文生视频一次产生消耗约24个单元。实际应用可能是小于24，但不会小特别多，并且很有可能大于24，因为文生视频不仅仅是简单的把图片快速播放起来，还需要内容具备多维性和多元性。目前主流文生视频模型生成视频长度仅支持2秒~4秒。

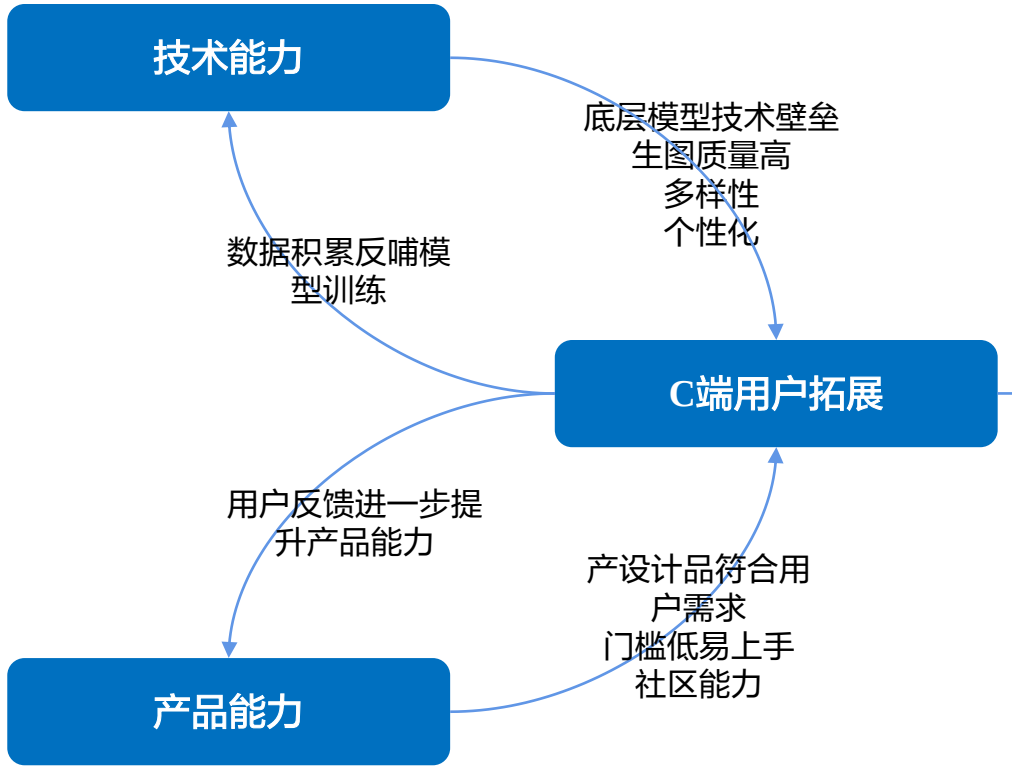




资料来源：产业调研，GITHUB，the gamer station，中国电子报，国海证券研究所

图像生成应用的竞争壁垒依赖技术和产品能力双驱动下的飞轮效应

为什么Midjourney脱颖而出?



工程化能力

AI工程化的过程包括需求分析、模型设计、模型训练、模型部署等环节。AI在产业里成功落地，所需要的不仅是基础的算法能力、数据获取等，还需要较为完备的产品化能力和解决方案的搭建能力，形成一整套软件、硬件、数据、算法、营销等体系的建设

B端用户拓展

进一步打开天花板

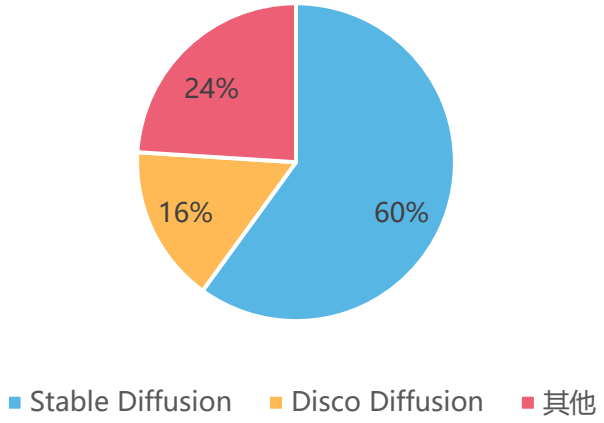
多场景针对性输出：
游戏开发
影视制作
工业设计
营销
电商
……

文生图领域整体创业门槛低于大语言模型，商业模式仍存疑问

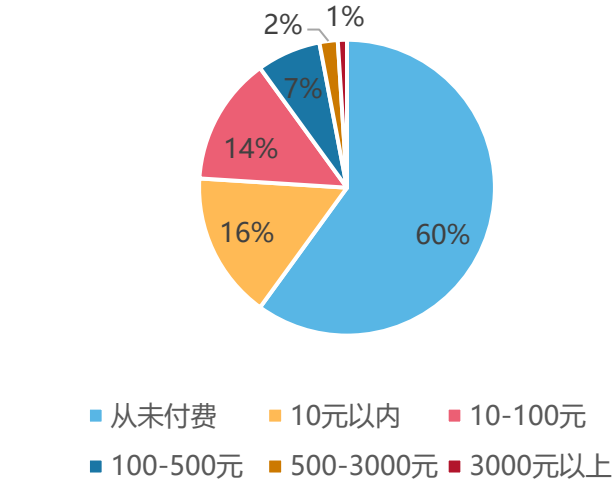
NLP预训练模型											
模型名称	GPT	BERT (Large)	GPT-2	T5	GPT-3	PanGu	LaMDA	ERNIE 3.0	ERNIE 3.0 Zeus	GLM-130B	GPT-4
参数量	110M	340M	1.5B	11B	175B	200B	137B	10B	千亿参数	130B	万亿以上
训练数据集大小	5GB		40GB	750G	45TB	1.1TB	1.56TB	4TB	4TB		
文生图预训练模型											
模型名称	DFGAN	DALL-E	Cogview	Cogview2	Imagen	GLIDE	DALL-E2	StableDiffusion	Parti	CM3Leon	SDXL1.0
参数量	19M	12B	4B	6B	3.4B	3.5B	3.5B	800M	20B	7B	3.3B
训练数据集大小								400M	5B	340M	

- **模型层看：**图像生成领域已有生成质量较高的开源预训练模型Stable Diffusion，且SD具有较为丰富的开发者生态，有许多插件供选择。创业公司可基于Stable Diffusion基础版本进行进一步调优和个性化数据训练， Stable Diffusion最新发布的开源模型SDXL1.0采用更大参数量级进一步提升了生成图像质量。例如初创公司无界 AI 便是国内最早基于 SD 模型推出 AI 绘画工具的平台之一。
- **成本端看：**从主流模型参数规模看，文生图参数量级多在1-10B之间，而通用大模型入门级门槛达到了70B，文生图整体参数量级较小，成本远低于通用大模型。通过调研文生图初创公司，实际小团队利用开源模型，初期在用户不到1万情况下甚至无需购买A100，通过购买RTX30\40系列、IBS3060（5000~1w/张）也可以启动。用户1万左右的文生图公司，生成单张图片的成本在0.1元左右。
- 文生图领域虽然创业门槛低，但商业模式仍存疑问。但国内C端用户付费意愿偏低，B端则需要和场景强相关，会有较多定制化的场景，要针对不同客户的产品需求去打造相应的图片生成的引擎，对工程化能力有很高的要求，长期看大公司可能具备更强的场景和工程化能力。以无界AI为例，其用户量接近300万，C端付费率约20%，营收主要来源于B端客户。

中国AI绘画行业算法模型使用占比情况



中国AI绘画用户为AI绘画产品或服务付费比例



资料来源：产业调研，51TCO，华为云，澎湃新闻，阿里云，电科技，智谱官网，新智元，华经产业研究院，论文见附录，国海证券研究所

部分文生图&视频应用商业化情况

国内			国外		
应用	应用类型	收费模式	应用	应用类型	收费模式
文心一格	AI作画助手	白银/黄金/铂金会员分别定价69/139/339元/月	DALLE1&2	AI绘画平台	15 美金换115 个点数
万兴爱画	AI绘画平台	10次/30次/100次分别为5元/12元/20元	Stable Diffusion	文本到图像扩散生成	开源免费
美图WHEE	AI视觉创作工具	收费模式尚未明确	Midjourney	AI绘画平台	Basic/Standard/Pro分别定价10/30/60美元/月
无界版图	AI绘画工具	青铜/白银/黄金/铂金会员分别定价99/199/1299/4699元/月	Designs.ai	人工智能图片视频创作工具	Basic/Pro分别定价19/69美元/月
妙鸭相机	AI写真生成软件	限时特惠9.9元，附赠10颗钻石，钻石可用来高清化（2颗/张）和下载照片（2颗/张）	Lumen5	人工智能在线视频制作平台	Basic/Starter/Professional分别定价19/59/149美元/月，企业可定制化
PicSo	AI绘画生成器	用户可每天免费生成一张图，会员：9.99美元/月或49.99美元/年	Runway	人工智能在线视频制作平台	通过销售月度“点数”（credits）供用户使用Gen-1、Gen-2成等产品及增值服务，分别有标准版（$12/月-625点）和高级版（$28/月-2250点）
6Open art	AI绘画工具	20/200/800/5000点分别为5/30/100/500元	Synthesia	人工智能在线视频制作平台	个人版本收取固定订阅费用，价格为29美元/月，全年订阅享受25%折扣，264美元/年；企业版本根据座位数的不同费用不同

文生图推理算力需求测算

文生图	情景1	情景2	情景3	情景4	情景5	Midjourney	核心假设
模型参数量（亿）	10	20	30	40	50	30	MJ参数量预估在9-40亿区间，SD在10亿上下，大部分文生图模型参数量在几十亿级别
所需显存容量（GB）	3.7	7.5	11.2	14.9	18.6	11.2	根据经验公式（参考右方推算思路）等比例换算得到
A100显存容量（GB）	40	40	40	40	40	40	
（1）单次推理所需显卡数量	**0.14**	**0.29**	**0.43**	**0.57**	**0.71**	**0.43**	A100可拓展7个GPU，所以1/7张A100已可满足单次推理需求
DAU（万）	100	300	500	800	1000	700	MJ用户数量2023年5月在1500万左右，考虑文生图付费用户较为活跃，DAU/用户数设定为50%
并发推理需求最大设计容量（次）	1000	3000	5000	8000	10000	7000	Google搜索引擎10亿日活对应10万并发推理需求，大模型推理约为5000次，考虑到文生图耗时更长，假设并发次数设计比例为谷歌搜索设计比例的10倍
单次推理时合并的推理需求数量	1	1	1	1	1	1	通常来说文生图并未对推理次数进行合并计算
（2）并发推理所需要推理次数	1000	3000	5000	8000	10000	7000	并发推理最大设计容量/单次推理合并的推理需求容量
推理所需要显卡数=（1）x（2）	143	857	2143	4571	7143	3000	

推理思路介绍：

显存容量经验公式：10亿参数量对应3.7GB显存容量需求。假设每个参数为FP32格式，假设每个参数为FP32格式即4个字节（文生图一般不需要做精度缩减），则原始理论需求为10*4*10^8/1024/1024/1024=3.7GB。

计算单次推理所需显卡数量：A100显存容量为40GB/80GB，以40GB为例，A100可拓展的GPU是7个，40/7=5.7>所需显存需求3.7，因此单次推理所需A100显卡数量为1/7。

根据日活数量推算并发推理需求最大设计容量：以Google的日活与单秒所需要处理的并发需求作为基础，考虑到文生图所需要的耗时较长，要让用户具备一定用户体验，并发容量的设计次数应该是10倍于谷歌搜索。

计算并发推理所需要推理次数：假设同一时间可承受的最高推理请求次数，以及单次推理时模型合并的推理需求数量，得到在并发推理时所需要的推理次数。

计算并发推理所需要显卡数量：单次推理所需显卡数量与并发推理时最高所需要的推理次数相乘即为所需显卡的数量。

文生视频推理算力需求测算

文生视频	情景1	情景2	情景3	情景4	情景5	核心假设
模型参数量（亿）	100	150	200	250	300	Runway GEN2参数未公布，预估在100亿左右，整体参数量级高于文生图
所需显存容量（GB）	37.3	55.9	74.5	93.1	111.8	根据经验公式（参考右方推算思路）等比例换算得到
A100显存容量（GB）	40	40	40	40	40	
（1）单次推理所需显卡数量	1	2	2	3	3	
DAU（万）	10	50	100	300	500	整体用户体量目前低于文生图，DAU/用户数设定为50%
并发推理需求最大设计容量（次）	200	1000	2000	6000	10000	考虑文生视频速度更慢，假设文生视频假设为文生图的2倍
单次推理时合并的推理需求数量	1	1	1	1	1	通常来说文生视频并未对推理次数进行合并计算
（2）并发推理所需要推理次数	200	1000	2000	6000	10000	并发推理最大设计容量/单次推理合并的推理需求容量
推理所需要显卡数=（1）x（2）	200	2000	4000	18000	30000	

推理思路介绍：

显存容量经验公式：100亿参数量对应37GB显存容量需求。假设每个参数为FP32格式，假设每个参数为FP32格式即4个字节（文生图一般不需要做精度缩减），则原始理论需求为100*4*10^8/1024/1024/1024=37GB。

计算单次推理所需显卡数量：A100显存容量为40GB/80GB，以40GB为例。

根据日活数量推算并发推理需求最大设计容量：以Google的日活与单秒所需要处理的并发需求作为基础，考虑到文生图所需要的耗时较长，要让用户具备一定用户体验，并发容量的设计次数应该是10倍于谷歌搜索。考虑文生视频速度更慢，假设文生视频假设为文生图的2倍，即5倍于谷歌搜索。

计算并发推理所需要推理次数：假设同一时间可承受的最高推理请求次数，以及单次推理时模型合并的推理需求数量，得到在并发推理时所需要的推理次数。

计算并发推理所需要显卡数量：单次推理所需显卡数量与并发推理时最高所需要的推理次数相乘即为所需显卡的数量。

如何看待文生图竞争格局？与高频场景结合更容易突围

	代表应用	短期	长期
垂类AI应用	Midjourney Stable Diffusion Runway 文心一格 万兴爱画 ……	头部应用通过技术/产品/成本/数据等优势突破，在C端率先开启变现； 创业门槛不高导致出现许多中长尾应用，缺乏竞争优势将逐渐被淘汰，用户留存率低	针对垂类场景头部应用C端天花板相对明确； 搭建工程化能力可技术输出到B端场景，探索更多变现可能
现有应用叠加AI功能	Adobe Firefly 美图 ……	短期收入端贡献不明显，主要盘活现有用户，通过AI功能引入提升产品体验和用户粘性	基于现有高频场景，长期用户壁垒更强，用户不易流失，用户ARPU和付费率有望提升

文生图代表模型及应用

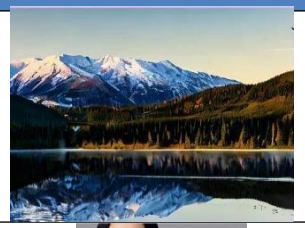
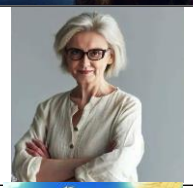
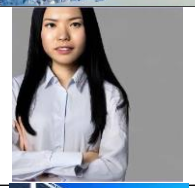
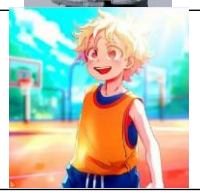
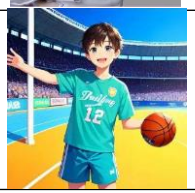
图像生成模型一览：国内外厂商积极布局探索

图像生成模型															
模型名称	DALL-E	Disco Diffusion	CogView	GLIDE	ERNIE-VILG	Make-A-Scene	Midjourney	DALL-E-2	Stable Diffusion	Imagen	CogView2	Parti	ERNIE-VILG 2.0	CM3leon	DALL-E-3	发展趋势
发布时间	2021.02	2021.05	2021.11	2021.11	2021.12	2022.03	2022.03	2022.04	2022.04	2022.05	2022.04	2022.06	2023.03	2023.07	2023.10	
研发机构	OpenAI	OpenAI	清华	OpenAI	百度	Meta	Midjourney	OpenAI	慕尼黑大学、海德堡大学、Runway	Google	清华	Google	百度	Meta	OpenAI	整体厂商数量较少；最新研究以国外厂商为主
支持语言	英文	英文	中文、英文	英文	中文、英文	英文	英文	英文	英文	英文	中文、英文	英文	中文，英文	英文	中文，英文	以英语为主
底层算法	自回归模型	扩散模型	自回归模型	扩散模型	自回归模型	自回归模型	扩散模型	扩散模型	扩散模型	扩散模型	自回归模型	自回归模型	扩散模型	自回归模型	扩散模型	Diffusion Model占据主流
参数量	12B		4B	3.5B	10B	4B		3.5B	1B	1.5B	6B	20B	24B	350M/760M/7B		参数量↑
训练数据集大小	250M文本-图像对		30M文本-图像对		145M文本-图像对	35M文本-图像对		900M图像	LAION-5B	460M图像-文本对；400MLaion图像-文本对	30M文本-图像对	LAION-400M；FIT400M；JFT-4B dataset	170M图像-文本对	Shutterstock datasets		训练集大小、种类↑
是否开源	否	是	否	否	否	否	否	否	是	否	否	否	否	否	否	开源模型较少；国外厂商新模型不开源
Zero-shot FID-30k（↓）	27.5		27.1	12.24	14.7	11.84		10.39	12.6	7.27	24.00	7.23	6.75	14.20/6.61/4.88		图像质量↑
Speed [s]	-		-	15	-	25	-	-	3.7	9.1	-	-	-	-	-	生成速度↑

资料来源：论文见附录5，国海证券研究所

主流商用文生图模型效果对比：综合看Midjourney和Adobe相对领先

国海证券 SEALAND SECURITIES

风格	prompt	DALL-E	DreamStudio（Stable Diffusion v2.1）	Midjourney V5	Adobe Firefly	百度文心一格	Tiamat	点评
写实风景	一张美丽的风景照，夕阳的余晖照射在平静的湖面上，茂盛的森林，远处的雪山，专业摄影							生成质量都比较高，但在画面风格上有一定区别，且对倒影处理也很逼真
人物肖像	一位经验丰富的女设计师肖像，半身照，浅灰色背景，丰富的面部细节							Midjourney V5 和 Adobe Firefly效果逼真、完成度高；文心一格和Tiamat完成度相对落后
二次元	一个开心的男孩，穿着短袖运动服，阳光明媚，背景是篮球场，动漫，明亮的色彩							DALL-E和DreamStudio相对较差，其余效果均不错；文心一格和Tiamat有专门“二次元”模型
概念场景	宏伟的古代中国建筑群，漂浮在云端，广角全景，成群的仙鹤，瀑布，盛开的桃花，游戏概念设计							Adobe、MJ生成效果领先，文心一格具备中国风特色，Tiamat有专门“概念场景”模型，部分应用缺乏仙鹤元素
3D人偶	一个可爱小公主，黑色长波浪形头发，穿着华丽的蓝色衣服，快乐地微笑着，迪士尼风格，被向日葵包围，3D渲染，超高分辨率							Midjourney、Tiamat 和 Adobe Firefly 的生成质量都不错，Dreamstudio 的细节太粗糙，DALL-E生成的人物质感略差，文心一格忽略“向日葵”描述

Open AI：先后推出自回归和扩散图像模型，最新发布DALL-E3

国海证券 SEALAND SECURITIES

DALL-E 2021年2月

DALL-E（120亿参数），文生图鼻祖模型，**主要基于自回归模型。**

- 生成策略：在第一阶段，首先训练一个编码图像的编码-解码结构，并将中间表示作为图像特征。在第二阶段，提取文本的特征，并将其与图像特征拼接起来，从而得到图文对的特征。接着，通过Transformer 模型生成图像。

(a) a tapir made of accordion. a tapir with the texture of an accordion.

(b) an illustration of a baby hedgehog in a christmas sweater walking a dog

GLIDE 2021年11月

GLIDE（35亿参数），**主要基于扩散模型。**

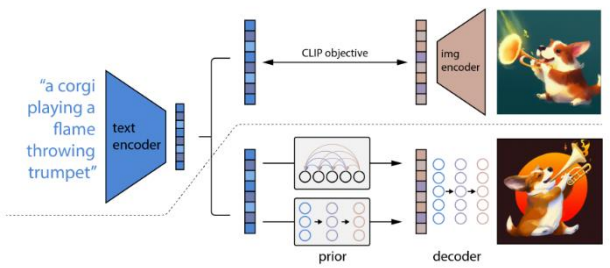
- 生成策略：算法采用了Guided Diffusion方法中相同的Autoencoder结构，但是进一步扩大了通道数量，使得最终的网络参数数量达到了3.5 billion；数据采用了和DALLE相同的大规模文本-图像对数据集。
- 从模型效果（FID值）来看，GLIDE在较小参数量的基础上实现了比DALL-E更好的图片效果。以猫在跳棋为例，GLIDE模型生成具有阴影和反射的逼真图像，并以正确的方式组合多个概念，产生新颖概念的艺术效果图。

"a surrealist dream-like oil painting by salvador dalí of a cat playing checkers"

DALL-E2 2022年4月

DALL-E 2（35亿参数），**主要基于扩散模型。**

- 生成策略：主要包括三个部分：CLIP，先验模块prior和img decoder。其中CLIP又包含text encoder和img encoder。DALL-E 2的工作是训练两个模型。第一个是Prior，接受文本标签并创建CLIP图像嵌入。第二个是Decoder，其接受CLIP图像嵌入并使用扩散模型生成图像。
- **DALL-E2核心在于Diffusion与CLIP模型的结合。**DALL-E 2中的文本语义和与其相对的视觉图片之间的联系，是由CLIP学习的，
- 在图像映射环节，OpenAI使用了GLIDE的修改版本来执行图像生成，实现了逼真的还原效果。



DALL-E3 2023年9月

DALL-E 3，**主要基于扩散模型。**

- 相比DALLE-2，DALLE-3主要改进两个方面：1）DALL-E 3 更加基于现实基础，能够更有效地完善细节内容，使得生成的图片更加真实与吸引人；2）DALL-E 3 可以更好地理解上下文，能够根据prompt更加精准输出图像，即使是简单的输入也能得到较为细节的图像。
- **此外，DALL-E3与ChatGPT集成，允许用户使用 ChatGPT 创建提示并包含更多安全选项。**DALL · E 3 将于 10 月初向 ChatGPT Plus 和 Enterprise 客户提供。

DALLE-2

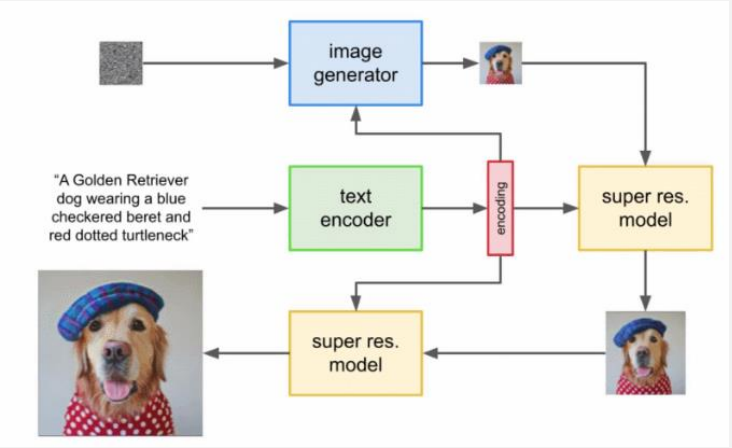
DALLE-3

谷歌：先后推出基于扩散模型的imagen和基于自回归模型的Parti

Imagen

Imagen为谷歌在2022年5月发布的一款图像生成模型（15亿参数），**主要基于扩散模型。**

- 生成策略：Imagen首先将文本输入编码器（使用谷歌基于Transformer的T5编码器，确保文本编码理解标题中的单词如何相互关联），转换成数值表示。此外，Imagen使用扩散模型作为图像生成器，创建能够将输入到 Imagen 的字幕的语义信息封装起来的图像，图像生成器或基础模型先输出一个小的 64x64 图像，随后Imagen使用两个超分辨率模型（同样基于扩散模型）将该图像放大到最终的 1024x1024 分辨率。
- 研究结果：表明大型预训练冻结文本编码器对于文本到图像任务非常有效，且预训练文本编码器的大小比扩散模型的大小更重要。
- 生成效果： Zero-Shot FID 30k值为7.27，优于同Open AI同样基于Diffusion模型的DALL-E2（10.39），主要原因或在于Imagen的文本编码器比 DALL-E 2 的文本编码器大得多，并且接受了更多数据的训练。



Parti

Parti为谷歌在2022年6月发布的另一款图像生成模型，**主要基于自回归模型。**

- 生成策略：Parti将Transformer与ViT-VQGAN结合。将文本到图像的生成视为序列到序列的建模问题，类似于机器翻译——这使其能够受益于大型语言模型的进步，尤其是通过扩展数据和模型大小来解锁的功能。Parti 使用功能强大的图像标记器ViT-VQGAN将图像编码为离散标记序列，并利用其将此类图像标记序列重建为高质量、视觉多样化图像的能力。
- 研究结果：对四种比例的 Parti 模型（350M、750M、3B 和 20B）进行了详细比较，并观察到：1）模型功能和输出图像质量得到持续和实质性的改进，**最大版本的 Parti 模型甚至可以拼写单词**，而 OpenAI的DALL-E 2 只能生成图像。2）在比较 3B 和 20B 模型时，评估者大多数时候更喜欢后者，具体来说：图像真实度/质量为 63.2%、图文匹配率为75.9%；3）20B 模型尤其擅长抽象、需要世界知识、特定视角或书写和符号渲染的提示。
- 生成效果： Zero-Shot FID 30k值Parti-3B 为8.10、Parti-20B为7.23
- 缺陷：高质量图像生成依赖大参数量，训练成本较高；且对于部分情形生成能力有待提升，比如计数、否定的文本描述、多物体空间位置等

A portrait photo of a kangaroo wearing an orange hoodie and blue sunglasses standing on the grass in front of the Sydney Opera House holding a sign on the chest that says Welcome Friends!

Meta：公布基于自回归的模型CM3Leon，生成质量媲美主流扩散模型

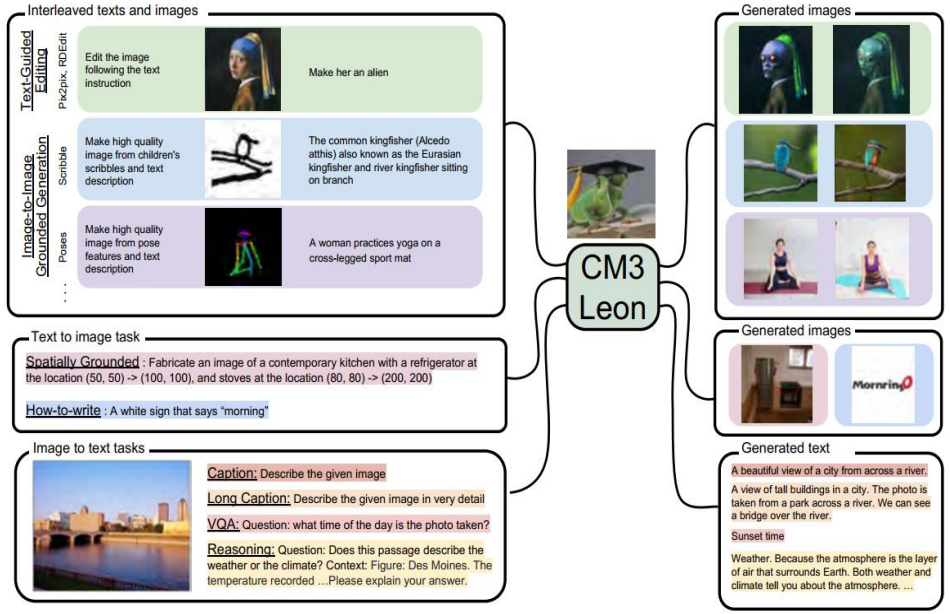
7月16日，Meta在官网公布CM3leon，是首个使用纯文本语言模型配方改编和训练而成的多模态模型，使用了30亿文本token，并经历了大规模检索增强预训练和随后的多任务监督微调（SFT）阶段。

模型架构及数据库

- CM3Leon采用自回归模型。在预训练阶段，Meta 使用了数百万张来自 Shutterstock 的授权图片，有70 亿参数，达到了 OpenAI DALL-E2 模型的两倍。
- 在架构方面，CM3Leon采用了一个和成熟的文本模型相似的仅解码器 Transformer，但不同的是CM3Leon能够输入和生成文本和图像。通过采用论文「Retrieval-Augmented Multimodal Language Modeling」中提出的训练检索增强技术，Meta大大提高了CM3Leon模型的效率和可控性。
- **CM3Leon强大性能的关键在于加入大规模的检索增强预训练阶段和第二个多任务加入监督微调的技术（SFT）阶段。**通过应用跨模态的监督微调技术，Meta显著提高了CM3leon在图像标注、视觉QA和文本编辑方面的性能。

模型特色

- **CM3leon 的强大之处在于将模态组合成单一模型，让它能够在文本、图像和构图任务之间流畅地自由转换。**除了文生图的功能，CM3leon还可以为图像生成标注、回答有关图像内容的问题，甚至可以根据边界框和分割图的文本描述创建图像，此前公开披露的AI系统中是没有的。CM3leon还有一个独特的功能——根据任意格式的文本指令对现有图像进行编辑，比如更改天空颜色，或者在特定位置添加对象。
- **兼顾计算量和成本的高质量图像生成模型，突破多模态模型的新疆界。**据Meta介绍，CM3leon训练时的计算量仅相当于以往基于Transformer方法的五分之一，但CM3leon在文本到图像的生成方面还是获得了同类领先的性能，获得了 4.88 的 FID，超越谷歌的文本到图像模型 Parti。**CM3leon 既具备自回归模型的功能多样性和有效性，也保持着较低的训练成本和良好的推理效率。**



左侧：各种多类型任务的常见输入；右侧：相应的模型输出

	Retrieval in Training	Responsible	# of Retrieved Documents	Dataset Size	Model Size	Zero-shot FID-30K
RA-CM3	✓	✗	2	150M	2.7B	15.70
StableDiffusion	✗	✗	-	400M	800M	12.60
KNN-Diffusion	✓	✗	10	70M	400M	12.50
MUSE	✗	✗	-	500M	3B	7.88
PARTI	✗	✗	-	5B	20B	7.23
RE-IMAGEN	✓	✗	2	450M	3.6B	5.25
CM3Leon-7B	✓	✓	0	340M	7B	10.82
CM3Leon-7B	✓	✓	1	340M	7B	5.78
CM3Leon-350M	✓	✓	2	340M	350M	14.20
CM3Leon-760M	✓	✓	2	340M	760M	6.61
CM3Leon-7B	✓	✓	2	340M	7B	**4.88**

Table 1: Summary of various text-to-image models on the zero-shot MS-COCO task as measured by FID. For all of our models, we generate 8 samples for each input query, and use a CLIP model to select the best generation.

CM3leon的效率显著高于同类Transformer架构模型

资料来源：MetaAI官网，《Scaling Autoregressive Multi-Modal Models:Pretraining and Instruction Tuning》，新智元，AI前线，IT之家，新智元，巴比特，国海证券研究所

Midjourney：基于扩散模型的文生图龙头，用户规模超千万

Midjourney 是AI基于文字生成图像的工具，由David Holz创立于2021年。Midjourney以拥有充沛流量的Discord为载体，实现低成本获客和低成本营销，**在此中拥有超过1000万人的社区，不到一年完成了1亿美元的营收**，但至今未融资。Midjourney的模型是闭源的，参考CLIP及Diffusion开源模型的基础上抓取公开数据进行训练。

应用优势

- **迭代速度快于同行，图像质量提高迅速。**Midjourney从V1推出到V5.2版本仅仅一年半的时间，图像质量显著提高，并且积累了数量可观的用户群体。
- **依托Discord社群，降低用户使用门槛。**Midjourney以拥有充沛流量的Discord为载体，实现低成本获客和低成本营销。
- **拥有较高的图像质量和独特的艺术风格。**Midjourney能够生成不同的风格，用户可以在提示词中选择默认艺术风格的应用强度， Midjourney尤其擅长环境效果，特别是幻想和科幻场景，生成图片具有较强商业价值和艺术特色。
- **基于庞大用户规模和数据反哺模型训练，形成飞轮效应。**通过庞大的用户量及用户数据，Midjourney积累的数据集具有独家性，可以进一步进行针对性训练。

第一代
2022 年 3 月发布

- 图像色彩丰富；
- 包含了提示中的许多细节；
- 穿着略有不同的衣服；
- 每个图像的角度不同；
- 面部不够协调

第二代
2022年4月发布

- 角色脸部更匀称，边缘也更清晰；
- 衣服更逼真；
- 灯光效果越来越好； 更多角度和细节；
- 每一张图片之间差异增大；

第三代
2022年7月25日发布

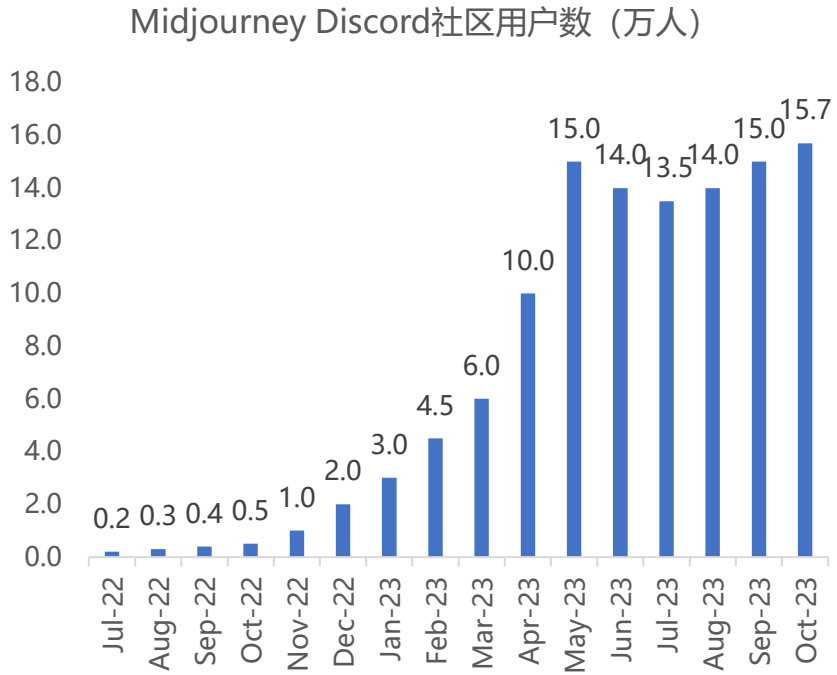
- 对提示有了更深入的理解，包含了更具体的细节；
- 配色没有之前突兀，混合方式更时尚；
- 将角色添加到背包中，而不是让她自己戴上，说明在逐渐改进。

第四代
2022年11月5日发布

- 每幅图像显示出具有略微不同的气候特征；
- 头发、脸、衣服、背景和其他特征都非常详细；
- 艺术色彩丰富，充满活力
- 服装保持动漫外观；

第五代
2023年3月16日发布

- 绝佳分辨率、细节、调色；
- 背景具有城市夜晚的鲜明特征，背景灯光，角色衣服和头发颜色很好地融合在一起；
- 背景没有聚焦，增加了图像的深度；
- 多种多样的角色穿着不同衣服，背包也各不相同。

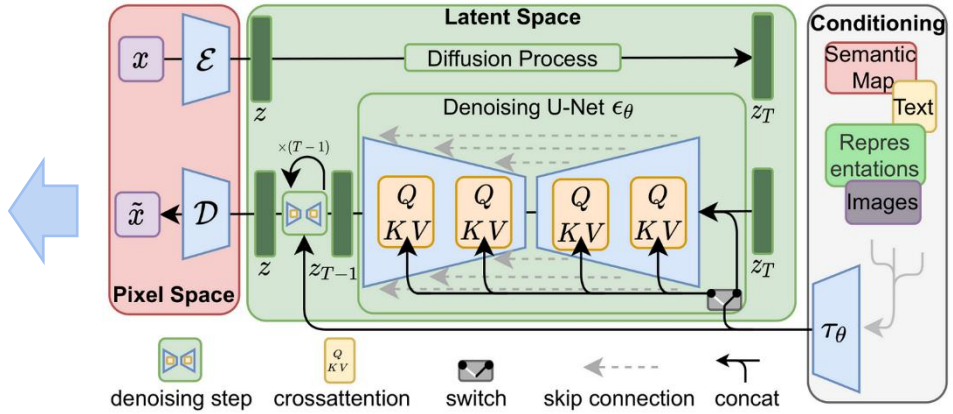


StabilityAI：发布Stable Diffusion开源模型

Stable Diffusion是Stability AI下的一款基于生成式AI的文本生成图像的模型，于2022年8月首次推出。2022年10月Stability AI获得了由全球风险投资公司Lightspeed Venture Partners和Coatue Management领投的1.01亿美元融资，估值突破10亿美元，是AI绘画领域的第一家“独角兽”。

模型架构及原理

- Stable Diffusion采用的底层模型是扩散模型，将扩散模型与Latent Space结合，能够大大减少计算复杂度，同时也能达到不错的图片生成效果。首先需要训练好一个自编码模型，利用编码器对图片进行压缩，然后在潜在表示空间上做diffusion操作，最后再用解码器恢复到原始像素空间，论文将这个方法称之为感知压缩（Perceptual Compression）。引入感知压缩是通过VAE这类自编码模型对原图片进行处理，忽略掉图片中的高频信息，只保留重要、基础的一些特征，能够大幅降低训练和采样阶段的计算复杂度，让文图生成等任务能够在消费级GPU上，在10秒级别时间生成图片，大大降低了落地门槛。



产品版本

- 2022年8月，Stability AI推出Stable Diffusion 1.0版本，11月，Stable Diffusion 2.0版本上线。
- 2023年2月，ControlNet的通用型插件发布，在此基础上，Stable Diffusion可以更精准地呈现人体姿态、画面层次感以及复杂的三维结构，用户可以调整图片细节。
- 2023年4月，Stable Diffusion改进版本——SDXL发布，6月推出 SDXL 0.9 版本更新，对 Stable Diffusion 文本生成图片模型进行了升级。升级之后的 Stable Diffusion 生成的图片效果更加逼真，改进了图像和构图。
- **2023年7月，Stability AI公布了最新的开源绘图模型——SDXL1.0，分别有两个版本：用于文生图的33亿参数模型，和用于66亿参数的图生图模型。Stability AI表示，SDXL1.0能生成更加鲜明准确的色彩，在对比度、光线和阴影方面做了增强，可生成100万像素的图像（1024×1024）。而且还支持在网页上直接对生成图像进行后期编辑。**

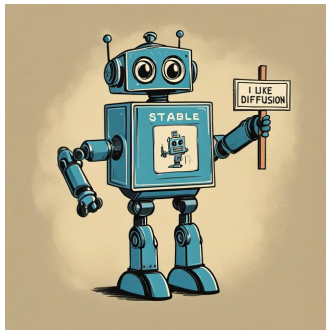
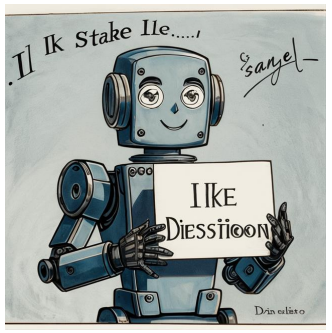
应用特色

- **开源模型吸引开发者，代码速度迭代快，**由于开源免费属性，SD已经收获了大量活跃用户，开发者社群已经为此提供了大量免费高质量的外接预训练模型（fine-tune）和插件，并且在持续维护更新。在第三方插件和模型的加持下，SD 拥有比 Midjourney 更加丰富的个性化功能。
- **商业化不足，产品使用门槛较高。**相比Midjourney，SD对于硬件要求较高，需要本地的独立显卡；部署相对麻烦，需要从Github下载许多部署文件；产品有一定使用难度，若想生成精致个性化的图片，需要一定学习门槛。

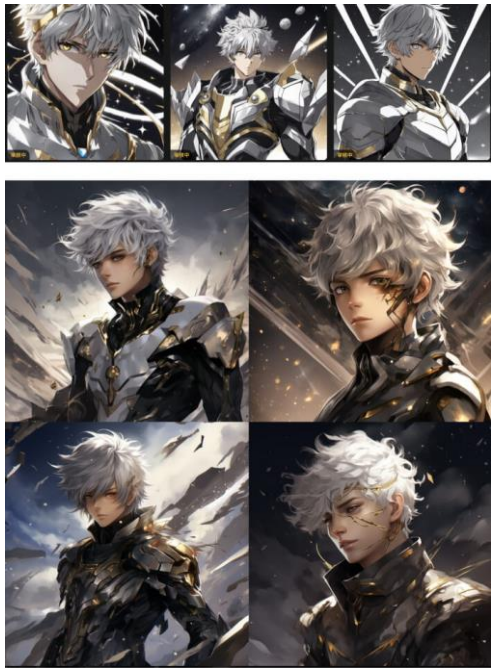
StabilityAI：最新发布SDXL1.0开源版本，图像生成能力进一步提升

国海证券 SEALAND SECURITIES

对比Stable Diffusion 1.5版本

	SDXL1.0	Stable Diffusion 1.5
A robot holding a sign with the text “I like Stable Diffusion” drawn in 1930s Walt Disney style		
Margot Robbie and Keanu Reeves, film grain, action movie, RAW 4K UHD, masterpiece		
Ukiyo-e Sakura Blossom: A blank greeting card in ukiyo-e style depicting a serene landscape with cherry blossom trees in full bloom, colorful flowers carpeting the ground, a meandering river, and birds chirping joyfully in the background		

对比Midjourney（三图为SDX1.0，四图为MJ）



anime guy, short silver hair, black white silver gold clothes, comet, asteroids in background, comet in background, silver eyes, serious/neutral expression, serious/neutral face, forehead armor headband silver and gold, flying, cool, creative, powerful look, 4k

soccer player in white jersey celebrating his goal, in the style of zeiss batis 18mm f/2.8

Clipdrop被Stability AI收购，融入多项AI功能图像处理能力优秀，数据显著增长

国海证券 SEALAND SECURITIES

- **公司简介：Clipdrop是Init ML公司旗下的AI图像编辑和生成应用。**该应用包含超过10种图像处理工具，也加入了AI智能生成图片功能。母公司Init ML于2020年创立于法国，于2023年3月被AI图像生成模型Stable Diffusion的母公司Stability. AI收购。2022年6月，Stability. AI发布SDXL 0.9，表示其是“Stable Diffusion文本-图像模型套件”的最先进开发版本。在收购Clipdrop后， SDXL 0.9功能应用于Clipdrop中。2023年7月26日，Stability. AI发布SDXL 1.0，进一步提升Clipdrop性能。其后数据出现明显增长，2023年7月网站访问量接近1500万。

▼ AIGC功能：建立以AI为动力的图像创作生态系统

文生图：Stable Diffusion XL

- Stable Diffusion功能被集成于Clipdrop中，用户可以使用文本生成图像。
- 提供动漫、3D渲染、写实等风格选择，目前生成的图片不可商用。

图像变体：Reimage XL

- 根据用户上传的图片，识别图片内容和风格，自动生成类似图片。

图像外绘：Uncrop

- 根据用户上传的图片，自动向外拓展图片。

——Clipdrop产品随着生成模型更新而更新，更新速度稳定。功能覆盖广，包含了视觉剪辑的各个方面。

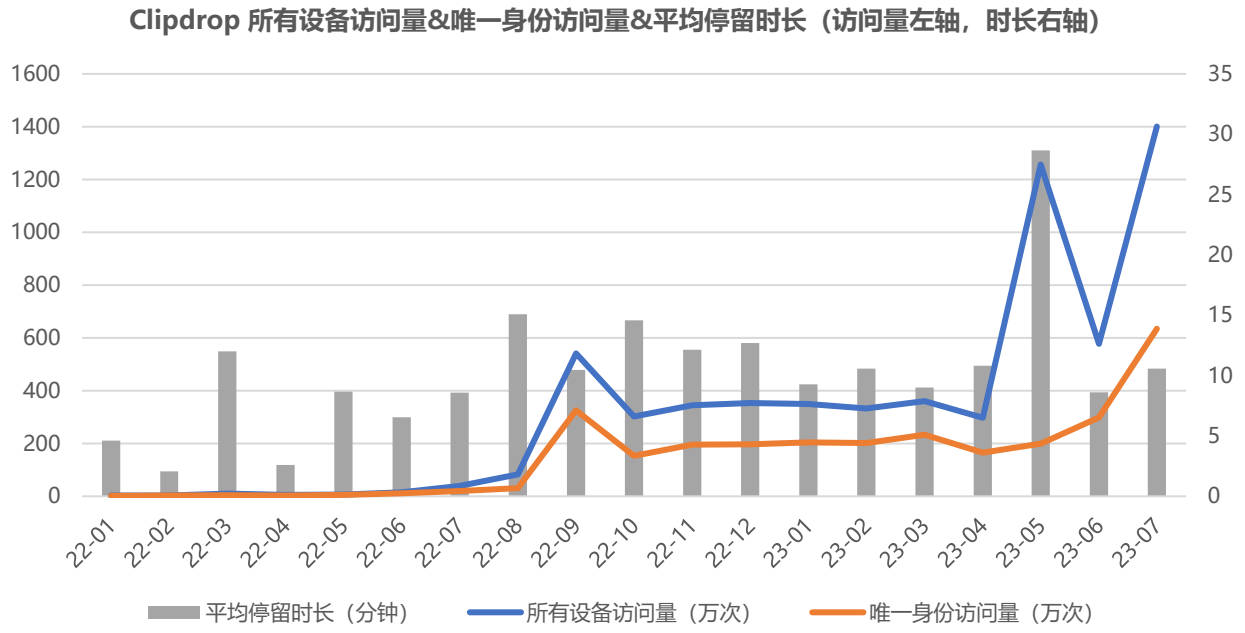
▼ 商业模式：同时进行固定费用和灵活收费，适应更多需求

- Clipdrop的产品有两种付费模式。免费版本可使用几乎所有工具，但是限制图片大小和处理次数。
- **Free升级为Pro收取固定订阅费用。**按月订阅价格为**9美元/月**，全年订阅价格为**7美元/月**，两种订阅方式可使用功能的类型和次数没有区别。订阅Pro版本可以获得Stable Diffusion XL功能的免排队使用，以及解锁各图片编辑功能的高清模式。
- **API基于具体工具和调用工具的次数收费。**API模式下，用户付费充值credit，并用credit去兑换相应的工具及次数。

- **Clipdrop测评：图像生成效果自然，细节处理优秀。**

本测评主要集中在Clipdrop中涉及AIGC的三项功能：文生图、图像变体和图像外绘以及一些基础图像编辑工具。Clipdrop的几项AIGC功能均可以免费使用，但每日可使用次数和图像清晰度有限制。

- **操作简单便捷。**图片的生成和编辑均可以在三步之内完成，图片生成速度较快，无需等待很长时间。但是，未订阅版本会存在排队的情况。
- **AI生成和修改图片质量较好。**生成的图片细节清晰，立意明确；拓展的图片与原图片连接处自然流畅。
- **图像编辑工具细节处理优秀。**在图像物体边缘等细节处处理效果较好，一键编辑速度很快。



资料来源：Clipdrop官网，Crunchbase，eu startups，stability ai官网，Lab 212，Similarweb，国海证券研究所

Adobe Firefly：与Adobe旗下图像编辑软件结合，具备较强可编辑性

Firefly是Adobe的一款基于生成式AI的工具，能够通过100多种语言，使用简单的文字建立影像、生成填色、对文字套用样式和效果、生成式重新上色、3D转换为影像、延展影像等。目前的 Firefly 生成式 AI 模式使用 Adobe Stock 资料集、开放授权作品和著作权已到期的公共内容进行训练。2023年9月，Adobe公布旗下AIGC工具Firefly AI的商业化方案：点数制收费，用户使用AI作图时消耗生成点数，每个点数对应一张图，每月可免费获得25点生成点数，同时可以付费购买额外点数。以单独购买Adobe Firefly的价格计算，生成每幅图像的价格大约为5美分。

产品版本

- 3月，Adobe发布了AI工具Firefly的测试版，测试版使用使用 Adobe Stock 资料集、开放授权作品和著作权已到期的公共内容进行训练。6月，Adobe将Firefly的功能整合到Photoshop的beta版中。
- 6 月，Adobe 推出了 AI 图像生成工具 Firefly 企业版。新版本使得企业可以使用自身的数据资产训练企业自己的 Firefly 大模型，使得企业能够快速生成可安全商用的图像内容。
- 官网显示，企业版 Firefly 将于今年下半年正式发布，企业用户已经可以在企业版 Adobe Express 中使用 Firefly 的功能。

应用特色

- **多选项描述，降低画图门槛**，与单纯的文字描述（prompt）不同，Firefly细化多个维度的指标供用户选择，分层次、精细化地明晰了客户的制图需求，能够提供更让用户满意的图像。
- **AI工具Firefly结合已有明星产品Photoshop，方便图像的定向修改**，Adobe的产品Photoshop已经累积了非常可观的用户，通过与AI工具的结合，能够提升Photoshop的性能，同时AI工具更迅速地落地到应用当中，增加用户的粘性。
- **为文字套用不同样式和效果，有更丰富的商业应用场景**，Adobe产品对字体、矢量图有较深的技术积累，可以做出精美的效果，Stable Diffusion和Midjourney在这一方面没有很好的支持。
- **版权争议较小，**FireFly的训练数据基于Adobe自己的图库和公有数据。

Firefly：创意生成式人工智能（AI）模型集

不同的用户群体选择：部分专业创作者群体依然对Midjourney等生成型AI技术持抵触心理，**Firefly的策略是帮助创意设计者而不是替代。**

其他个性化示例：平面设计、营销内容、视频编辑、3D建模

百度：理解生成筛选三步走，不断优化文心一格的文生图效果

文心一格**基于文心大模型**的文生图系统，是百度依托飞桨和文心大模型于2022年8月推出的首款AI作画产品。用户只需输入自己的创想文字，并选择期望的画作风格，即可快速获取由一格生成的相应画作。此外，文心一格还支持文生图+图生图模式，用户输入绘画创意并上传参考图也可生成图片。

技术路径

- 文心知识增强跨模态理解大模型：基于多视角对比学习的ERNIE-ViL 2.0，在预训练过程能够同时学习模态间和模态内的多种关联性，提升图像和文本跨模态语义匹配效果
- **文心知识增强跨模态图文生成大模型：ERNIE-ViLG**，将文生图和图生文的任务融合到同一个模型进行端到端学习，从而增强文本和图像的跨模态语义对齐。通过渐进式扩散模型，生成空间由小及大、生成轮廓由粗到细，同时根据生成阶段自动选择最优生成网络，文生图的效果取得进一步提升

理解

基于知识的Prompt工程，理解用户需求并在此基础上丰富语义细节，降低用户输入描述成本

生成

基于扩散生成算法实现创意写实与恢弘构图的艺术画作生成

筛选

基于跨模态匹配大模型进行生成画作的结果排序，自动选出语义与美观度最佳的画作

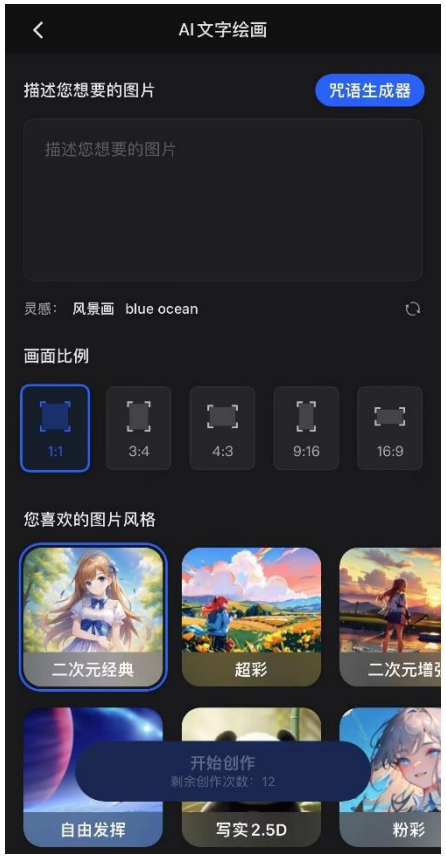
简介	
特色	• 完成创作后可以使用**AI编辑功能**，其中，**涂抹消除**能够结合画面内容在涂抹区域重新生成合理的内容，可用于局部内容消除、小范围画面修复；**涂抹编辑**在不填写的情况下默认做画面修复，填写指定内容将按照指定内容生成；**图片叠加**能够将基础图和另一张图片进行融合生成新的图片，两张图片各自的权重用户可以自行设置 • **实验室**目前推出以下三种功能：**人物动作识别再创作**能够通过识别人物图片中的动作，再结合输入的描述词，生成动作相近的画作；**线稿识别再创作**能够识别上传的图片，生成线稿图，再结合输入的描述词生成画作；**自定义模型**能够通过上传训练图片集，选择基础模型，调节参数训练自定义模型，训练完成的模型一经发布即成为用户的专属模型，可重复使用
收费模式	• **AI创作**：2电量/张（0.18~0.25元/张） • **AI编辑**：涂抹消除2电量/次，涂抹编辑2电量/次，图片叠加2电量/张 • 电量可通过每日签到、大赛投稿、画作分享、画作公开和充值（80电量/9.9元，200电量/23.9元，800电量/79.9元，1万电量/899元）等方式获得
会员服务	• **非会员**：仅支持单组画作生成，仅能使用AI创作功能 • **白银会员**：支持3组画作同时生成，能使用AI创作和AI编辑功能，最高可享900电量，送充电打折卡9折3张和白银排队加速 • **黄金会员**：支持5组画作同时生成，能使用AI创作、AI编辑和实验室功能，最高可享2300电量，送充电打折卡8.5折3张和黄金排队加速 • **铂金会员**：支持10组画作同时生成，能使用AI创作、AI编辑和实验室功能，最高可享6000电量，送充电打折卡8折3张和铂金排队加速

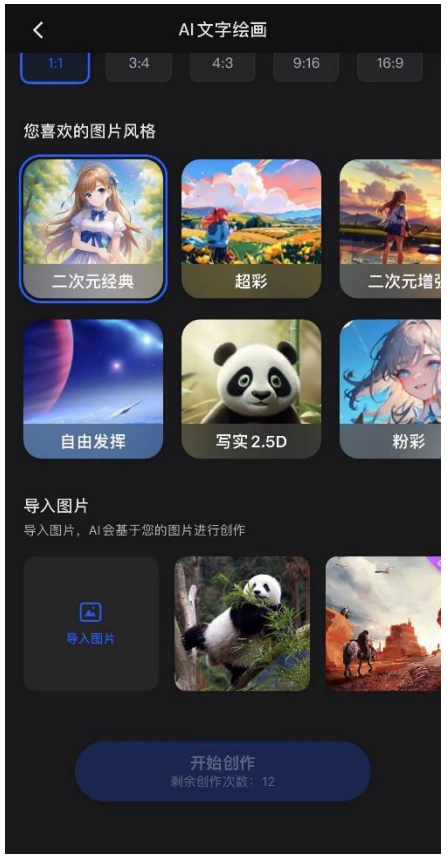
输入“一只在吃竹笋的大熊猫”，选择插画风格后生成的一组图片

万兴科技：持续加码AIGC，万兴爱画升级，Pixpic落地

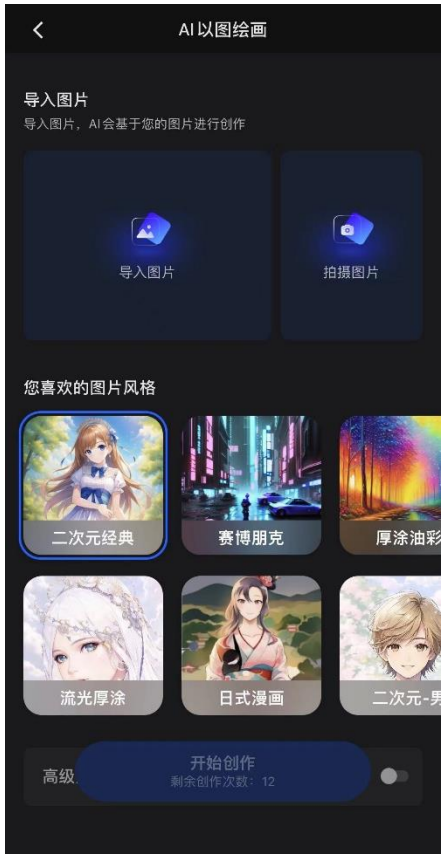
- 万兴爱画是万兴科技旗下的绘图创意类产品，作为公司探索AIGC领域的排头兵，于2022年11月以小程序形式首发上线，用户输入一段文字即可获得不同比例和多种艺术风格的AI绘画作品，经过多轮迭代升级，**目前已经全面支持AI文字绘画、AI以图绘图及AI简笔画三种创作模式和小程序、移动端及网页端多端畅享体验**。
- 近日，万兴科技在海外推出全新AIGC应用Pixpic，支持用户一键生成AI数字分身写真。作为一款**针对欧美人群的AI艺术照片生成器**，Pixpic集AI数字分身、多元化写真模板于一体，用户上传5张照片，即可快速生成专属数字分身，且支持艺术照、动画照、证件照、人像、工作照等写真模板风格。凭借简单的使用方法和对欧美人群照片偏好的特定生成优化等，Pixpic在Google Play一经上线，就吸引了广大用户下载体验和社交媒体分享。

文生图

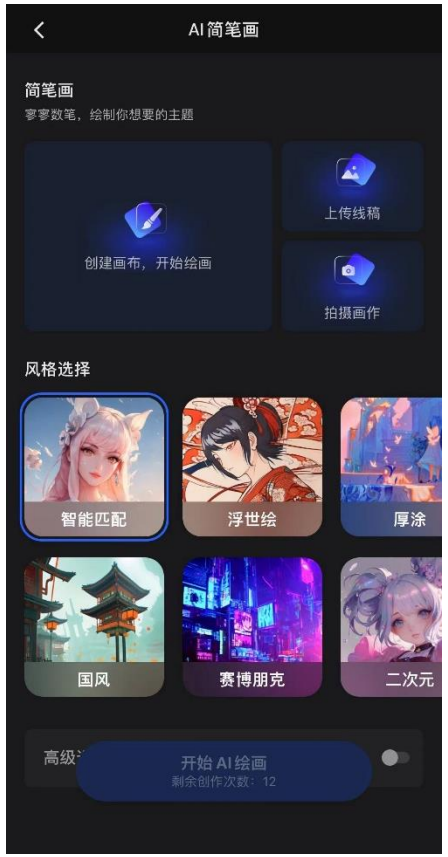




图生图



简笔画



特色

- 文生图：与手动输入文本不同，“咒语生成器”已预设超500个常用Tag，并提供人物&角色、五官、表情、头发、服装等十余个分类选项及多样化的风格效果，用户可以点选所需效果并对其权重进行调整，万兴爱画可据此自动生成一段丰富的文本内容并创作绘画作品
- 图生图：多元化图片风格选项
- 简笔画：用户只需简单描摹几笔，人机共创，5秒内即可绘制一幅高品质画作

收费模式

- 5元/10次（折合0.5元/次）
- 12元/30次（折合0.4元/次）
- 20元/100次（折合0.2元/次）

美图：着手布局B端市场，官宣自研视觉大模型，美图AI产品生态初步形成

2022年，美图的AIGC正式进入高速发展期。自2022年年底以来，美图快速迭代AIGC应用，平均每月有新功能上线，多元化程度远超竞品，覆盖生活场景与工作场景。2023年6月19日美图举办以“AI时代的影像生产力工具”为主题的第二届影像节，现场发布美图视觉大模型MiracleVision及6款新产品：WHEE、开拍、WinkStudio、美图设计室2.0、DreamAvatar数字人和美图AI助手RoboNeo，覆盖视觉创作、商业摄影、专业视频编辑、商业设计等领域，旨在全面提升影像行业的生产力，美图AI产品生态初步形成。相较于美图现有产品，这些新品部分面向B端用户，美图进一步释放布局ToB市场的信号。**AIGC+C端市场存量变现+B端市场增量拓展成为美图的主要战略。**

目前美图主要的文生图产品包括WHEE（文生图、图生图）、美颜相机（AI写真、AI头像）和美图秀秀（AI绘画、AI简笔画）。

WHEE	
产品功能	• 文生图：输入创意描述（可智能补全）后还可选择性输入不希望呈现的内容，选择风格后即可进行生成，支持单次最高生成4张图片 • 图生图：可上传原图作为参考，其他部分基本与文生图相同

美图AI视觉大模型MiracleVision	
技术路径	**基于Diffusion模型。**在算法层面，MiracleVision**运用零样本学习算法**，利用类别的高维语义特征代替样本的低维特征，使得训练出来的模型具有迁移性（零样本学习使模型无需微调就能刻画事物特征，能较大程度地提高设计效率）
上线时间	2023年7月起陆续在美图各产品上线
参数量	10亿级别
竞争优势	• 相较于国外顶级视觉大模型，MiracleVision**在亚洲人像摄影以及国风国潮或者说中国美学上有更好的洞悉和理解**，因此在元素的识别、理解和生成上都具有更高的准确性和创造性 • 通过邀请艺术家、设计师等具有深厚美学背景的专业人士，研究探索美学趋势，能帮助MiracleVision不断更新和提高自己对美学的理解，**MiracleVision也是国内少数将专业人士纳入研发阶段的大模型** • 美学评估系统：通过基于机器学习的美学评估系统，能够帮助MiracleVision持续优化呈现给用户的效果 • **模型生态的构建：**未来美图将会为创作者提供创作支持，例如课程、社区和模型创作大赛等；创作者训练的模型可以在美图旗下产品进行分发，在分发的过程中还能持续进行模型优化；创作者在这一过程中能够获得分成 • 美图在**过去十几年做美颜相机的积累**，使得MiracleVision相较于其他视觉模型，**对人像的理解更深**，生成的图像质量更高
商业模式	MiracleVision将会通过API或SDK将自身能力输出给行业客户使用

美图：产品测评

	美颜相机
产品功能	• **AI写真：**用户上传8~20张多角度、多表情、多背景的照片用于面部档案生成，选择造型后即可生成写真 • **AI头像：**用户上传同一人的3~8张近距离自拍（正脸优先，尽量包含不同背景和表情，避免脸部被遮挡）后选择头像性别（女性/男性/小女孩/小男孩）即可生成专属AI头像
收费模式	• **AI写真：【女生】**芭比乐园3.9元/套，芭比3.9元/套，清冷白月光3.9元/套，女生证件照9.9元/套，Y2K牛仔画报3.9元/套，复古婚纱写真3.9元/套，女生美式证件照0.9元/次，职业形象照3.9元/套，毕业纪念照6.9元/套；**【男生】**男生证件照9.9元/套，男生美式证件照0.9元/次 • **AI头像：**50张专属头像（5种风格，每种10张）/6.6元，100张专属头像（10种风格，每种10张）/9.9元，订阅会员享100张免费头像

	美图秀秀
产品功能	• **AI绘画：【图生图】**导入原图后即可生成三种不同风格的图片；**【文生图】**输入关键词后即可生成4张图片 • **AI简笔画：**用户简单描摹几笔后即可生成6张图片，可选择性添加描述以帮助提高图片生成的准确性
收费模式	• **AI绘画：【图生图】**暂未商业化，可免费使用；**【文生图】**订阅制，每日可免费使用2次，继续使用需开通VIP（提升至120张/日） • **AI简笔画：**订阅制，可免费使用1次，之后需开通VIP • **VIP：**连续包月12元，连续包年98元（首年特惠88元），包年送一季128元，包月18元

文生图效果（输入文本：一只在吃竹笋的大熊猫）

图生图效果

WHEE

美图秀秀

资料来源：美颜相机App、美图秀秀App、国海证券研究所

妙鸭相机：多模板AI写真相机，新晋爆款产品，但成熟度仍待提高

妙鸭相机是一款由未序网络科技（上海）有限公司开发的应用程序，属于生成式AI在国内C端的商业化落地产品，用户只需支付9.9元并提交20张以上个人照片，即可快速在线生成一套质感媲美专业照相馆的写真集。对未序网络科技进行股权穿透可知其为**阿里系公司**，妙鸭相机实为**优酷旗下的内部创业项目**，该项目由2020年加入阿里巴巴的张月光带队开发，产品孵化团队暂未独立。

特色

妙鸭相机本质是一款图生图产品，用户上传一张正面照以及至少20张多光线、多视角、多表情的上半身照片，可先生成一个数字分身。基于数字分身，选择自己喜欢的模版，就可得到一套AI写真。目前妙鸭相机提供晚秋、羽翼、职场、回到童年等36个写真模板（女性模板28个，男性模板8个）

1.制作数字分身

上传20张以上包含人脸的合格照片

限时特惠：¥9.9 附赠10钻石

制作AI数字分身需要消耗昂贵的算力

我们需要收取一点费用保障服务的稳定性

2.生成写真

选择喜欢的模板

一键得到高质量写真造型

3.精修写真

选择喜欢的造型

进行高清化、更像我等精修操作

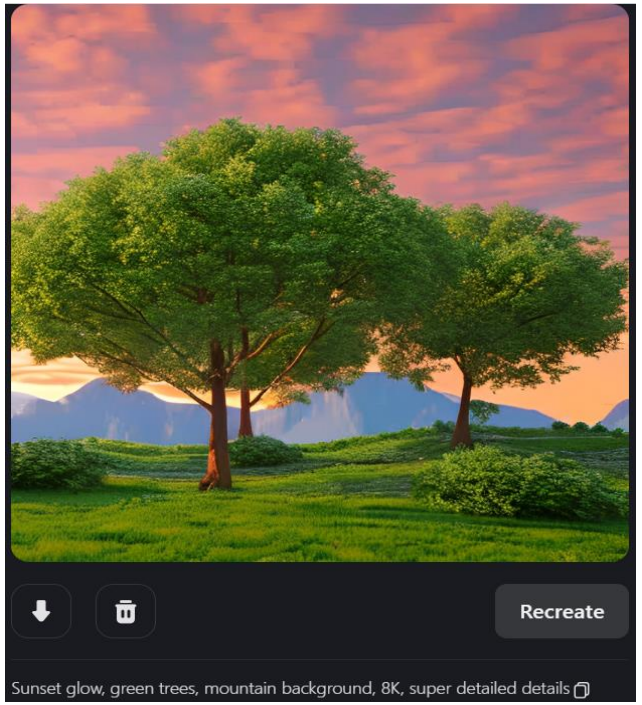
高清化、下载操作每次需2钻石

对比	Lensa	妙鸭相机
技术路径	Stable Diffusion模型+开源数据集LAION-5B	基于LoRA模型（Stable Diffusion的插件）微调的Diffusion模型
产品功能	用户通过上传一定数量（Lensa：10~20张，妙鸭相机：20张以上）的个人照片即可获得多张风格各异的AI肖像，**两者的区别在于Lensa侧重于生成头像，且支持许多非写实风格，而妙鸭相机则侧重于生成写真，相对而言更接近AI相机**	
收费模式	订阅会员制，35.99美元/年，提供一周免费试用期。同时，为了让用户更好地使用“魔法头像”功能，Lensa还提供了额外的付费选项，用户可以根据需求购买：50个头像/3.99美元，100个头像/5.99美元，或200个头像/7.99美元	限时特惠9.9元，附赠10颗钻石，钻石可用来高清化（2颗/张）和下载照片（2颗/张）。钻石可以通过邀请好友额外获得（5颗/邀请1位好友），也可以通过充值获得（60颗/6元，购买更多有折扣）
比较优势	目前已经是相对成熟稳定的应用，且是率先对文生图模型Stable Diffusion进行简化文本倒置过程创新的图生图应用	较低的定价，较为逼真的效果，以及对东亚脸型审美的优化；产品能力较强，用户生成单人模型后可便捷转化风格预览，用户体验好。
待解决问题	• **安全与隐私问题**：使用者担忧泄露隐私以及肖像权受到侵犯 • **对艺术家作品版权的侵犯**	• **安全与隐私问题**（或有，妙鸭相机于7月20日更新了部分条款，包括“您所上传的照片将仅用于本服务使用，我们仅提供图像处理服务，不会提取识别信息，不会用于识别用途，服务完成后，系统将自动删除上述信息，不予留存”） • 用户**需上传的照片数量过多** • 目前因使用人数过多而导致等待写真的**排队时间过长（算力资源不足）** • **产品成熟度不高**：写真生成的准确性有待提高（如上传儿童照片制作数字分身，生成的图片是儿童面容+成人体型）

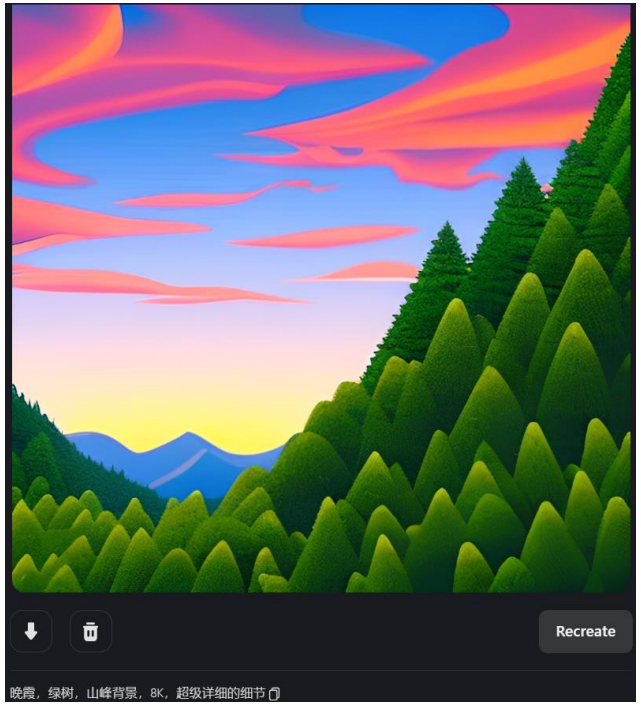
新国都：PicSo在海外率先上线，营收占比较小

PicSo是新国都子公司洞见科技有限公司于2022年四季度推出的文生图软件，同时支持移动端（iOS和安卓）和网页端，目前国内IP暂时无法使用，且相比中文文本输入，英文文本生成的图片质量更高。

英文文本输入



中文文本输入



特色

- AI Girl：聚焦女性形象生成，输入基本描述后可选择生成真人形象或动漫形象。描述角度包括动作、形体、衣着、发型、长相、场景、配件以及视角等，同时官网给出了一些tags供参考
- AI Art：基本流程同AI Girl，用户输入文字描述并选择风格后即可生成画作，支持风格包括动漫、素描、暗黑、赛博朋克等
- 用户可选择单次生成图片的数量，PicSo支持1、4、9的批量生成，后两者为付费选项

收费模式

- 用户每天可免费生成一张图片，更多的体验需要成为会员
- 会员专享权益包括每月100积分、专业风格、排队优先、无水印以及解锁多任务，价格为9.99美元/月或49.99美元/年，支付方式支持paypal、借记卡以及信用卡

重要版本更新

2022.10.20/1.3.1	2022.12.20/1.8.0	2023.2.14/1.10.1	2023.2.21/1.11.0	2023.6.21/1.15.0
软件上线App Store，可以根据文本生成图片	上线新功能：将照片和视频变成卡通风格	上线新功能：通过文本自定义生成各种动漫女孩；生成各类标签的真实女孩	上线新功能：开辟文本生成AI Girl板块	上线新功能：可以查看积分消耗及支付详情

文生视频代表模型及应用

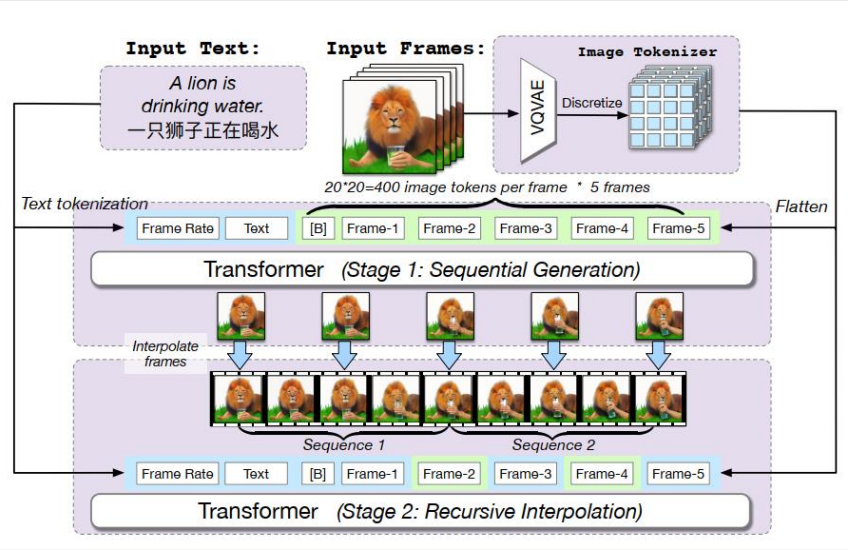
视频生成模型：行业迎来小幅高潮，生成质量仍有提升空间

在一定程度上，文本生成视频可以看作是文本生成图像的进阶版技术，同样是以Token为中介，关联文本和图像生成，逐帧生成所需图片，最后逐帧生成完整视频。据量子位发布的《AIGC/AI生成内容产业展望报告》，视频生成将成为近期跨模态生成领域的中高潜力场景，其背后逻辑是不同技术带来的主流内容形式的变化。

视频生成模型														
模型名称	GODIVA	NüWA（女娲）	Cogvideo	Make-A-Video	Imagen Video	Phenaki	MagicVideo	Tune-A-Video	Text2Video-Zero	Runway Gen-1	NUWA-XL	VideoLDM	PYoCo	**发展趋势**
发布时间	2021.04	2021.11	2022.05	2022.09	2022.10	2022.10	2022.11	2022.12	2023.03	2023.02	2023.03	2023.04	2023.05	
研发机构	微软、杜克大学	微软、北京大学	智源、清华大学	Meta	Google	Google	字节跳动	新加坡国立大学、腾讯	Picsart AI Resarch (PAIR)，UT Austin，U of Oregon，UIUC	Runway	微软	Nvidia	Nvidia	**整体厂商数量较少；最新研究以国外厂商为主**
支持语言	英语	英语	中文、英语	英语	英语	英语	英语	英语	英语	英语	英语	英语	英语	**以英语为主**
底层算法	AR(自回归模型)	AR(自回归模型)	Transformer		Diffusion Model	Transformer	Diffusion Model	Diffusion Model	Diffusion Model	Diffusion	Diffusion Model	Diffusion Model	Diffusion Model	**Transformer和Diffusion Model并存，Diffusion Model占据主流**
参数量		0.9B	9.4B		11.6B	1.8B						3.1B		**参数量↑**
训练数据集大小	Howto100M（包含超过1.36亿个文本-视频对的大规模文本-视频数据集）	2.9M个文本-图像对；727K个视频；241K个文本-视频对	5.4M文本-视频对	.3B文本-图像对；0M文本-视频对	14M个文本-视频对；60M个文本-图像对；LAION-400M文本-图像对	5M文本-视频对；0M文本-图像对；LAION-400M文本-图像对	Laion 5B；10M视频；0M Webvid10M视频；M 视频-文本对；UCF-101，MSR-VTT	DAVIS数据库的42个视频；140个Prompt		240M 图像；6.4M 视频 clips	166集，平均38000帧，1440×1080分辨率	10.7M 视频-文本对-52K视频小时	1.2B文本-图像对；22.5M图像-视频对	**训练集大小、种类↑**
是否开源	否	是	是	否	否	否	否	否	否	否	否	否	否	**国产厂商开源；国外厂商新模型不开源**
生成视频分辨率	128*128		480*480	1280*768	1280*768	较低	256*256	512*512			256*256	1280*2048		**视频分辨率无明显提升**
生成视频时长			4秒	5秒	5秒	2-2.5分钟					11分钟			**视频时长↑**
Zero-Shot	否	否	是	是			是	否				是	是	**Zero-Shot的能力↑**
帧率			32帧/秒	24帧/秒	24帧/秒							30帧/秒		**帧率维持24-32帧左右，无较大突破**
UCF-101 IS（↑）			23.55（中文）/25.27（英文）	33			-					33.45	47.76	**视频图像质量↑**
UCF-101 FVD（↓）			751.34（中文）/701.59（英文）	367.23			699					550.61	355.19	

清华CogVideo：首个开源的中文文本生成视频模型，基于自回归模型

自回归模型



生成步骤：
1.基于VQ-VAE，将每帧标记为图像token；
2.基于低帧率和文本顺序**生成关键帧**；
3.基于文本、帧率以及已知的帧递归插值，逐步生成中间帧。

CogVideo是由清华团队2022年发布的基于预训练的CogView2（文本生成图像模型）9B-参数转换器。**CogVideo是当时最大的、首个开源的文本生成视频模型，支持中文prompt，参数高达94亿。**CogVideo采用的Transformer结构，和CogView的几乎一致，例如使用夹层范数（Sandwich LayerNorm）和PB-Relax来稳定训练。

模型创新

- **多帧率分层训练策略：**能够更好地对齐文本和视频剪辑，显著地提高视频生成的准确性，这种训练策略赋予了CogVideo在复杂语义运动的生成过程中控制变化强度的能力。
- **基于预训练的文本生成图像模型：**通过微调预训练的文本生成图像模型，节省了从头开始预训练的花费，提高了生成的效率。

数据集：CogVideo在包含 540 万个字幕视频的数据集上预训练模型，空间分辨率为 160 × 160（CogView2 可以上采样到 480 × 480）。

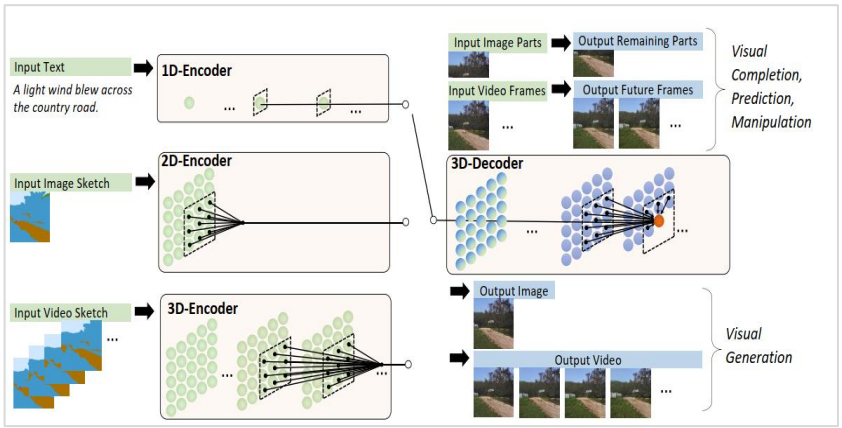
优点：
- 模型拥有较好的对齐文本和视频剪辑的能力，生成的视频质量及准确性有大幅提高。
- 相比于先前模型，能够生成较高分辨率（480×480）的视频。

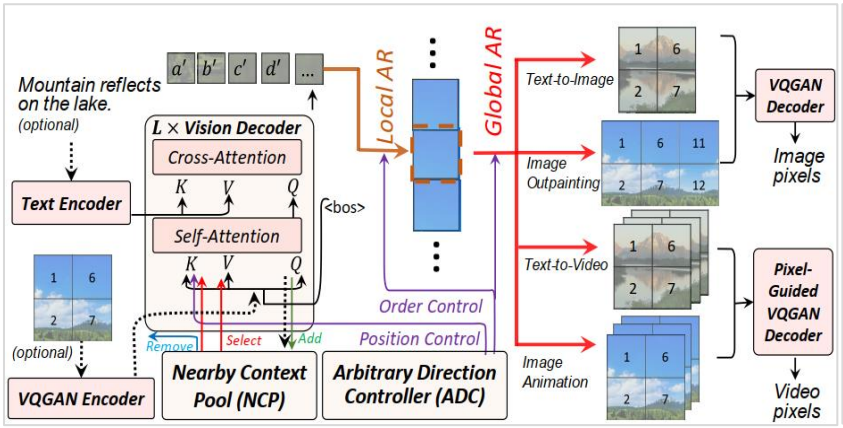
挑战：
- 输入序列长度受限于模型的规模和GPU内存。

资料来源：《CogVideo: Large-scale Pretraining for Text-to-Video Generation via Transformers》Wenyi Hong等，国海证券研究所

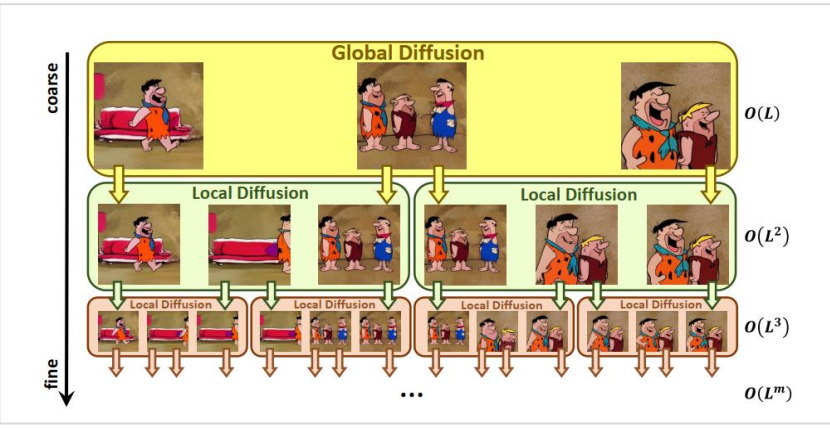
微软：NUWA系列从自回归到扩散模型，视频生成长度增加

自回归模型





扩散模型



NUWA

与NUWA相比，NUWA-Infinity在分辨率和可变尺寸的生成中，有的卓越视觉合成能力

NUWA-Infinity

NUWA-XL通过16句简单描述能够生成11分钟动画

NUWA-XL

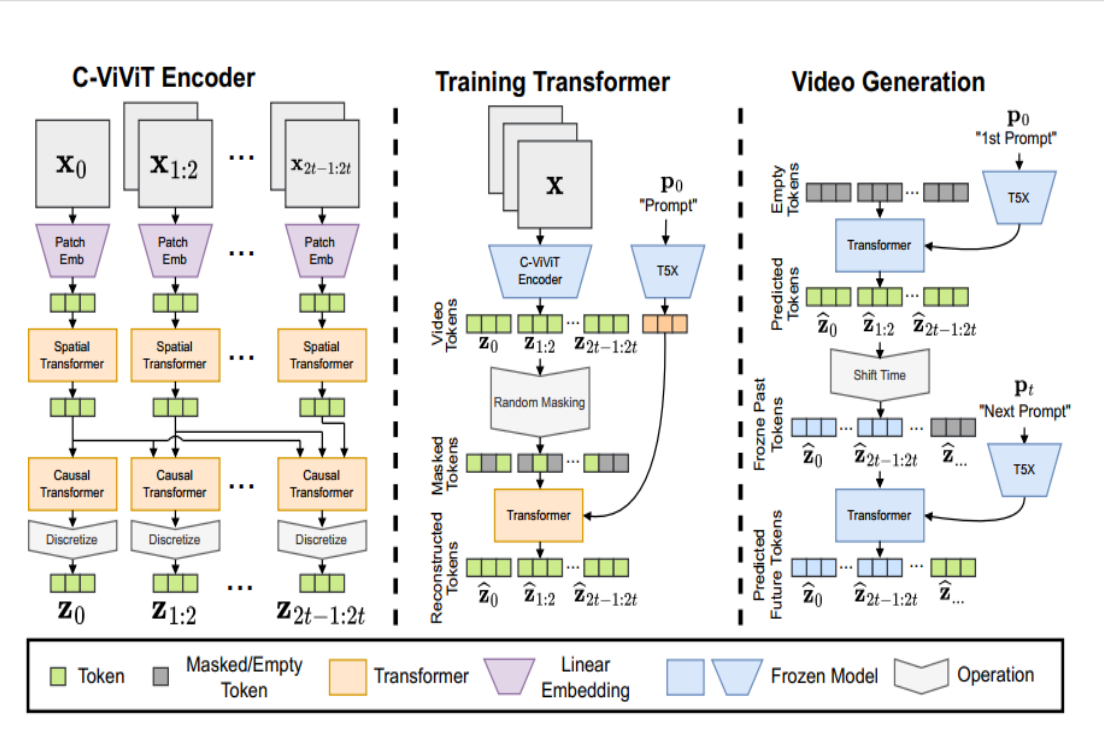
	NUWA	NUWA-Infinity	NUWA-XL
定义	统一多模态预训练模型	用于无限视觉合成的生成模型，生成任意大小的高分辨率图像或长时间视频。	用于超长视频生成的新型扩散基于扩散架构。
底层模型	Transformer框架-自回归模型	自回归模型（autoregressive over autoregressive）	扩散模型（Diffusion over Diffusion）
模型特点	由图像和视频预训练的模型；采用能够同时覆盖语言、图像和视频以及不同场景任务的三维变压器编码器-解码器框架；既能将视频作为三维数据处理，又能将文本和图像分别作为一维和二维数据进行处理；通过三维邻近注意力（3DNA）机制考虑视觉数据的性质并降低计算复杂度。	全局补丁级自回归模型考虑补丁之间的依赖关系，局部令牌级自回归模型考虑每个补丁内可视令牌之间的依赖关系；邻近上下文池 （NCP） 引入到已生成的缓存相关补丁中，作为正在生成的当前补丁的上下文；任意方向控制器（ADC）用于为不同的视觉合成任务确定合适的生成顺序，并学习阶次感知位置嵌入。	采用“从粗到细”的过程，应用全局扩散模型生成整个时间范围内的关键帧，然后局部扩散模型递归填充附近帧之间的内容，视频可以以相同的粒度并行生成。

资料来源：《NÜWA: Visual Synthesis Pre-training for Neural visUal World creAtion》Chenfei Wu等，《NUWA-Infinity: Autoregressive over Autoregressive Generation for Infinite Visual Synthesis》Chenfei Wu等，《NUWA-XL: Diffusion over Diffusion for eXtremely Long Video Generation》Shengming Yin等，NUWA官网，新智元，国海证券研究所

谷歌 Phenaki：首个可生成长视频的自回归模型

国海证券 SEALAND SECURITIES

自回归模型



Phenaki由Google Research开发制作，**该模型是第一个能够从开放域时间变量提示中生成视频的模型，能够根据一系列开放域文本提示生成可变长度的视频**。通过将视频压缩为离散的令牌的小型表示形式，词例化程序使用时间上的因果注意力，允许处理可变长度的视频。转换器以预先计算的文本令牌为条件，使用双向屏蔽转换器使得文本生成视频令牌，生成的视频令牌随后被取消标记化以创建实际视频。

特色功能

- **交互视频：**通过选择上下文词组合来创建有关主题的视频。
- **图像+文本描述生成视频（a still image + a prompt）：**输入第一帧和文本描述，Phenaki即能输出视频
- **长视频：**Phenaki是第一个可以通过一长串的文本描述（a long sequences of prompts）、并且描述可以随着时间的推移而变化以生成长达2分钟连贯视频的模型。
- **C-ViViT：**是ViViT的变体，通过视频生成的额外架构更改，它可以在时间和空间维度上压缩视频，同时在时间上保持自动回归，此功能允许自动回归生成任意长度的视频。

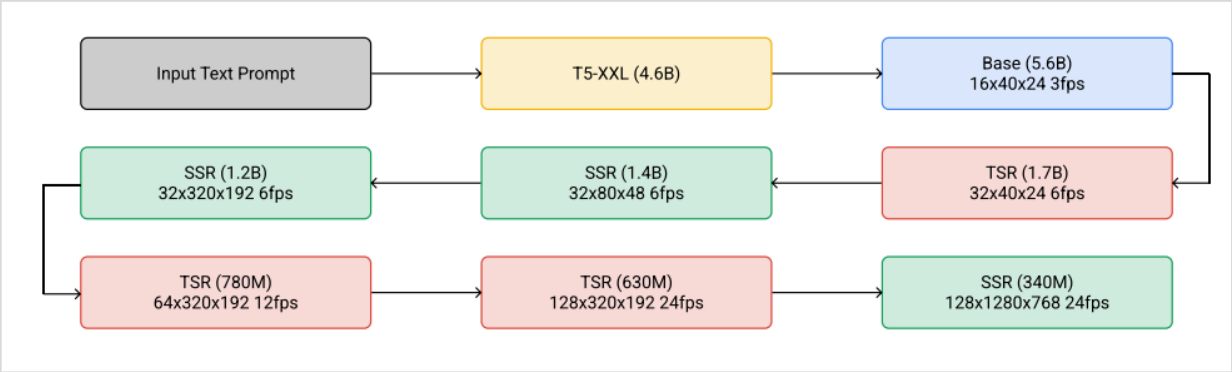
优点：

- Phenaki可以生成以开放域提示为条件的时间连贯性和多样性的视频，甚至能够处理一些数据集中不存在的新概念。
- Phenaki展示了图像和视频的联合训练来提高生成质量和多样性的方式，在大量图像文本对以及少量视频文本示例进行联合训练产生超出视频数据集中可用的泛化能力。

挑战：

- 生成的图像分辨率较低。

谷歌 Imagen Video：应用级联模型和渐进式蒸馏加速提升视频质量



SSR和TSR模型：模型使用时间卷积而不是时间注意力。基础模型中的时间注意力使 Imagen Video 能够对长期时间依赖性进行建模，而 SSR 和 TSR 模型中的时间卷积允许 Imagen Video 在上采样过程中保持局部时间一致性。与时间注意力相比，使用时间卷积降低了内存和计算成本 - 这一点至关重要，因为 TSR 和 SSR 模型的真正目的是在高帧速率和空间分辨率下运行。

Imagen Video由7个子模型组成（1个T5文本编码器、1个基础视频扩散模型、3个SSR扩散模型、3个TSR扩散模型），分别执行文本条件视频生成、空间超分辨率和时间超分辨率。

生成步骤：文本输入至级联采样管道开始生成，逐步SSR用以提高视频的分辨率，TSR用以提高视频的帧数。

扩散模型

Imagen Video是一个基于级联视频扩散模型的文本条件视频生成系统，由谷歌团队提出。Imagen Video使用冻结的T5文本编码器、基本视频生成模型和一系列交错的空间和时间视频超分辨率模型生成高清视频，将以前基于扩散的图像生成工作的结果转移到视频生成设置中。

模型特色

- **渐进式蒸馏：**视频模型中应用了渐进式蒸馏，无需分类器指导，以实现快速、高质量的采样。蒸馏在采样时间和感知质量之间提供了非常有利的权衡：蒸馏级联的速度提高了约18×，同时产生的视频质量与原始模型的样品相似。就FLOP而言，蒸馏模型的效率提高了约36×。
- **U-Net：**基本视频模型从扩大视频 U-Net 的参数计数中受益匪浅，通过增加网络的基本通道数和深度来执行此扩展。
- **多种方法从图像域转移到视频：**例如v参数化、条件反射增强和无分类器指导，并且发现这些在视频设置中也很有用

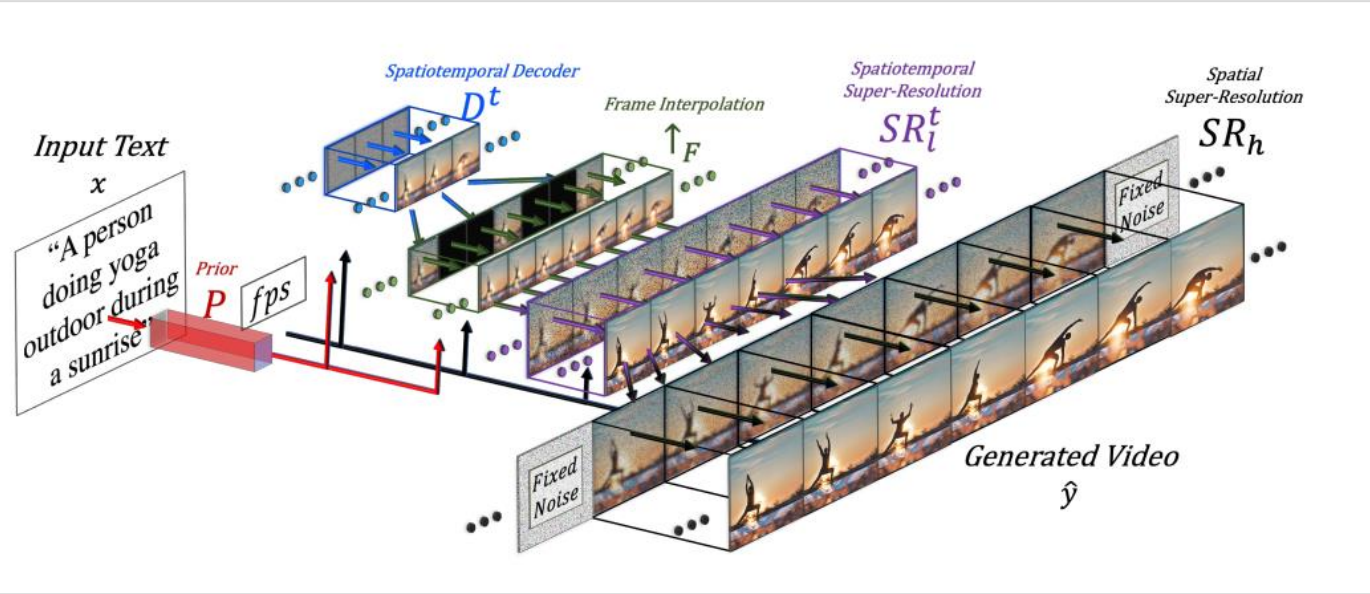
优点：

- 不仅能够生成高保真视频，而且具有高度的可控性和世界知识，包括能够生成各种艺术风格的各种视频和文本动画、能够理解3D结构、能够呈现具有不同样式和动态的各种文本

挑战：

- 在当前的模型大小下，性能尚未饱和，未来进一步扩展视频生成的模型的仍有空间。
- 模型的输入和输出仍然存在社会偏见和刻板印象，难以检测和过滤。

Meta Make-A-Video：创新采用无监督学习，加速模型训练



生成步骤：

1、通过Prior P输入文本，转换为图像embedding以及帧率；

2、解码器生成16帧64×64的图像；

3、按插值到更高的帧速率，并通过SR将分辨率提高，从而生成高时空分辨率的视频。

Make-A-Video是Meta旗下的基于文本生成视频的模型，**从配对的文本图像数据中了解世界的样子以及描述的方法，并从无监督的视频片段中了解世界是如何移动的。**Make-A-Video由三个主要组件组成：（i）在文本图像对上训练的基本文本生成图像模型；ii）时空卷积和注意力层，将网络的构建块扩展到时间维度；（iii）由两个时空层组成的时空网络，以及文本生成视频所需的用于高帧率生成的帧插值网络。

模型创新

- **使用无监督学习：**从数量级更多的视频中学习世界动态有助于研究人员摆脱对标记数据的依赖。
- **时空管道：**能够通过新设计的时空管道（包含视频解码器、插值模型和两个超分辨率模型）去生成高分辨率和帧率的视频，并且能够实现除文本生成视频以外的应用。

优点：

- 由于模型不需要从scratch中学习视觉和多模态表示，因此加速了文本生成视频模型的训练
- 模型成功将基于扩散模型的文字生成图像模型扩展到文字生成视频模型，无需配对的文字-视频数据
- 生成的视频继承了图像生成模型的广阔性（美学多样性，幻想描绘等）。

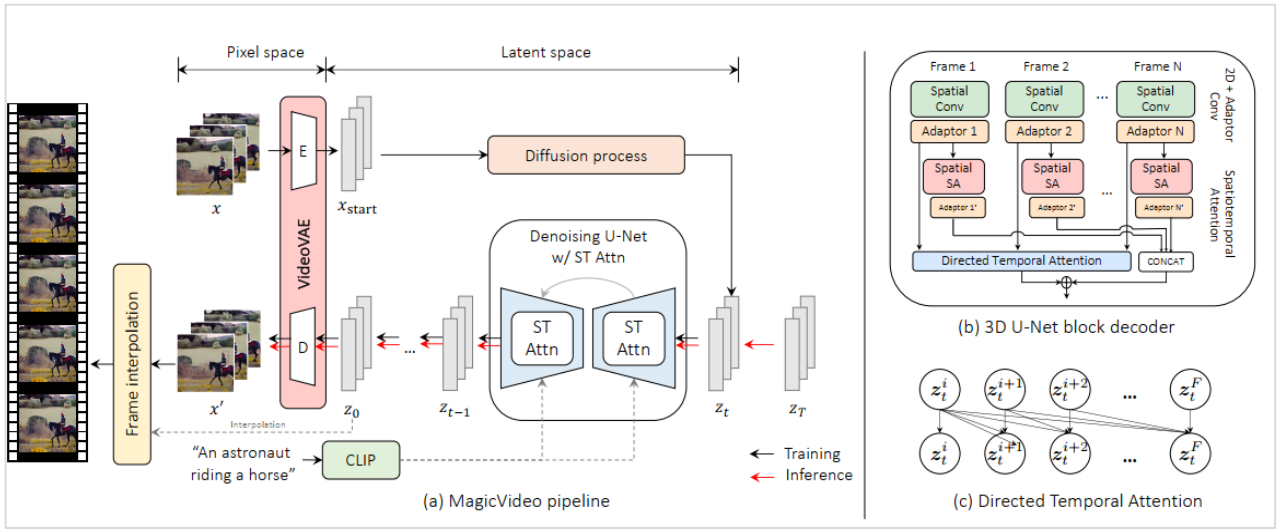
挑战：

- 无法学习文本与现象之间的关联，只能从视频推断
- 在整合以及生成包含多个场景和事件的长视频方面仍待后续完善

字节跳动MagicVideo：平滑视频剪辑助力商业应用

国海证券 SEALAND SECURITIES

扩散模型



MagicVideo是字节跳动提出的一种基于潜在扩散模型的高效文本到视频生成框架，**MagicVideo可以生成与给定文本描述一致的平滑视频剪辑**。MagicVideo的核心在于关键帧生成，通过扩散模型来近似低维潜在空间中16个关键帧的分布，结合具有高效的视频分配适配器和定向时间注意力模块的3D U-Net解码器，用于视频生成。

生成步骤：

1、使用预先训练的VAE将视频片段映射到低维潜在空间，对视频片段在低维潜在空间中的分布进行建模；

2、在推理阶段，首先在潜在空间中生成关键帧，然后插入关键帧以暂时平滑帧序列；

3、将潜在序列映射回RGB空间，并将获得的视频上采样到高分辨率空间，以获得更好的视觉质量。

模型创新

- **视频训练加速：**使在图像任务上训练的U-Net降噪器适应视频数据：用于图像到视频分布调整的帧轻量级适配器和用于捕获跨帧时间依赖性的定向时间注意力模块，因此可以利用文本到图像模型中卷积运算符的信息权重来加速视频训练。
- **像素抖动改进：**为了改善生成的视频中的像素抖动，提出了一种新颖的VideoVAE自动编码器，以实现更好的RGB重建。

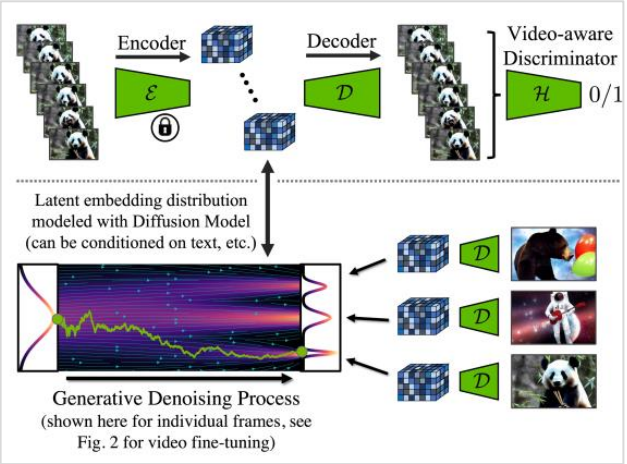
数据集：Laion 5B；10M视频；50M Webvid10M视频；7M 视频-文本对；UCF-101，MSR-VTT

优点：

- MagicVideo可以生成具有现实或虚构内容的高质量视频剪辑，文本生成视频剪辑，商用场景丰富，可以在剪映、抖音等软件里应用。
- 由于新颖高效的3D U-Net设计和低维空间中的视频分布建模，MagicVideo可以在单个GPU卡上合成具有256×256空间分辨率的视频剪辑，就FLOP而言，计算量比视频扩散模型（VDM）少约64倍。

NVIDIA：侧重扩散模型，实现高质量视频合成

Video LDM

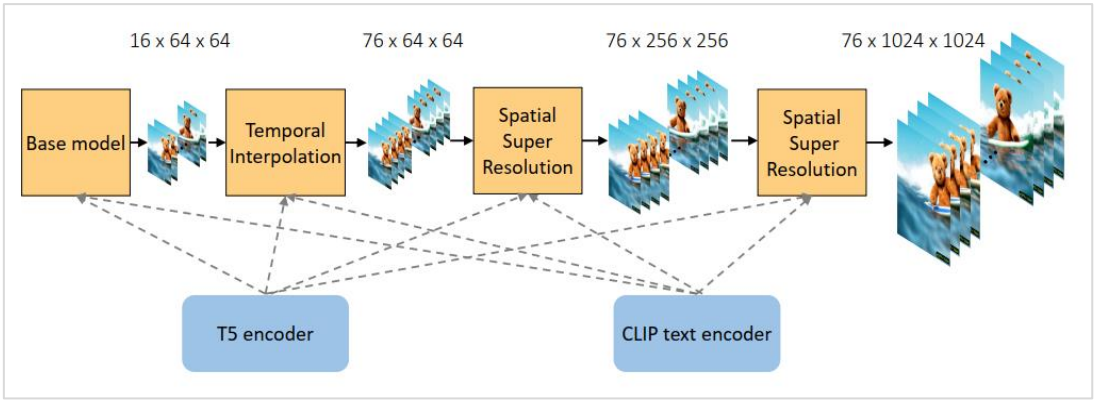


潜在扩散模型（LDM）通过在压缩的低维潜在空间中训练扩散模型，实现高质量的图像合成，同时避免过多的计算需求。

生成步骤：
1、先仅在图像上预训练 LDM；
2、通过向潜在空间扩散模型引入时间维度并对编码的图像序列（即视频）进行微调，将图像生成器转换为视频生成器；
3、在时间上对齐扩散模型上采样器，将它们转换为时间一致的视频超分辨率模型。

两个实际应用：
1、模拟野外驾驶数据：已经在分辨率为512×1024的真实驾驶视频上验证了的视频LDM，实现了最先进的性能。
2、 使用文本到视频建模创建创意内容：可以将公开可用的、最先进的文本到图像LDM稳定扩散转换为高效且富有表现力的文本到视频模型，分辨率高达1280×2048。

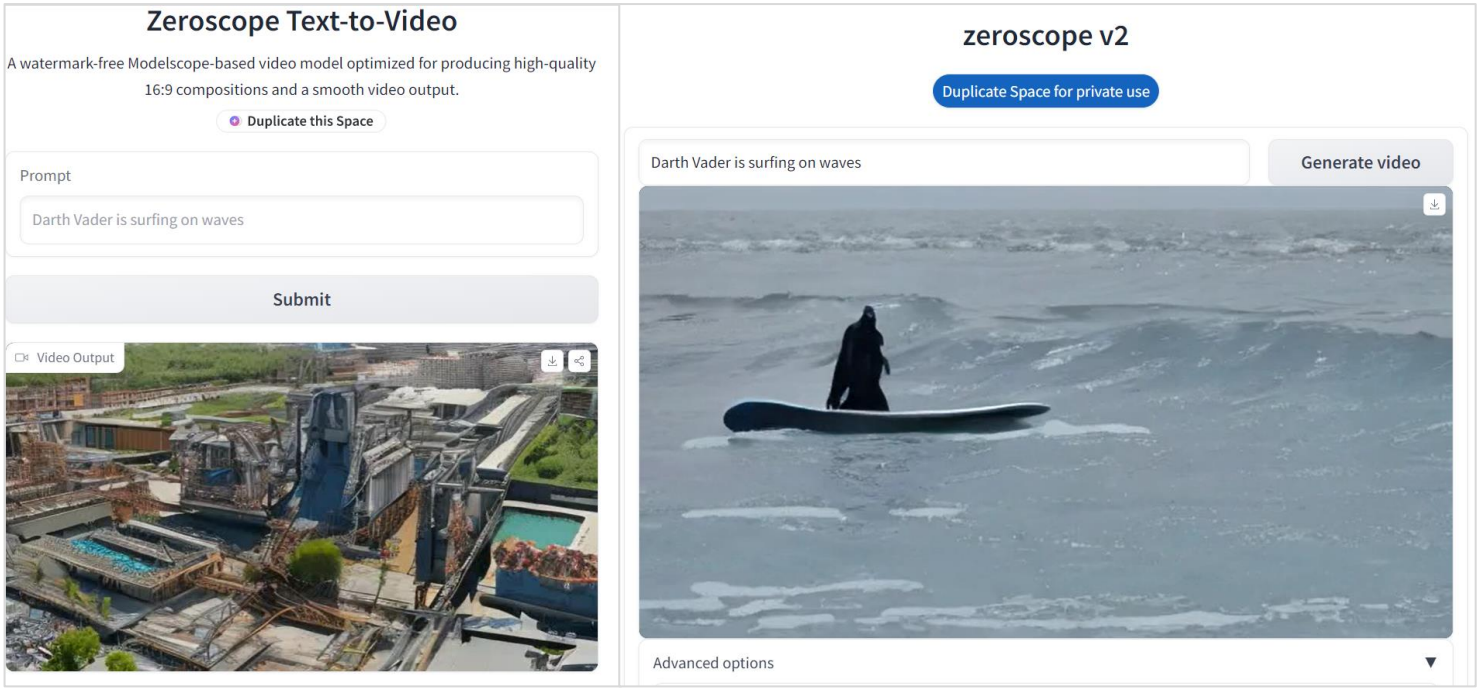
PYoCo



模型使用视频数据微调预训练图像扩散模型作为视频合成任务的解决方案，**研究了适合顺序视频帧生成的混合噪声和渐进噪声先验，应用渐进式噪声先验来微调基于扩散的文本到图像模型**。管道由四个网络级联组成：一个基本模型和三个上采样模型。所有四种模型都将输入作为从 T5 编码器和 CLIP 文本编码器获取的文本嵌入。

生成步骤：
1、基本模型生成 16 个空间分辨率为 64 × 64 的视频帧，跳帧为 5。
2、第一个上采样模型执行时间插值，生成大小为 76 × 64 × 64 的视频；
3、随后的两个超分辨率模型执行空间超分辨率以生成大小为 76 × 256 × 256 和 76 × 1024 × 1024 的视频。

Zeroscope：拥有较高质量输出的中国开源模型



使用说明：通过简单的文字描述输入（prompt），在数分钟内，用户可以免费得到视频输出，目前只有约4s的视频输出，画面比较单一，运动轨迹不丰富。

扩散模型

Zeroscope是魔搭社区（ModelScope）里文本生成视频的模型，其中Zeroscope_v2大模型在Hugging Face上开源，该模型是基于17亿参数量ModelScope-text-to-video-synthesis模型进行二次开发。

图像质量

- Zeroscope生成的视频没有水印，适配16:9的宽高比，有着较高质量和流畅的视频输出。
- 从V1到V2，Zeroscope的视频生成画面质量、速度和逻辑性都有较大的提高。

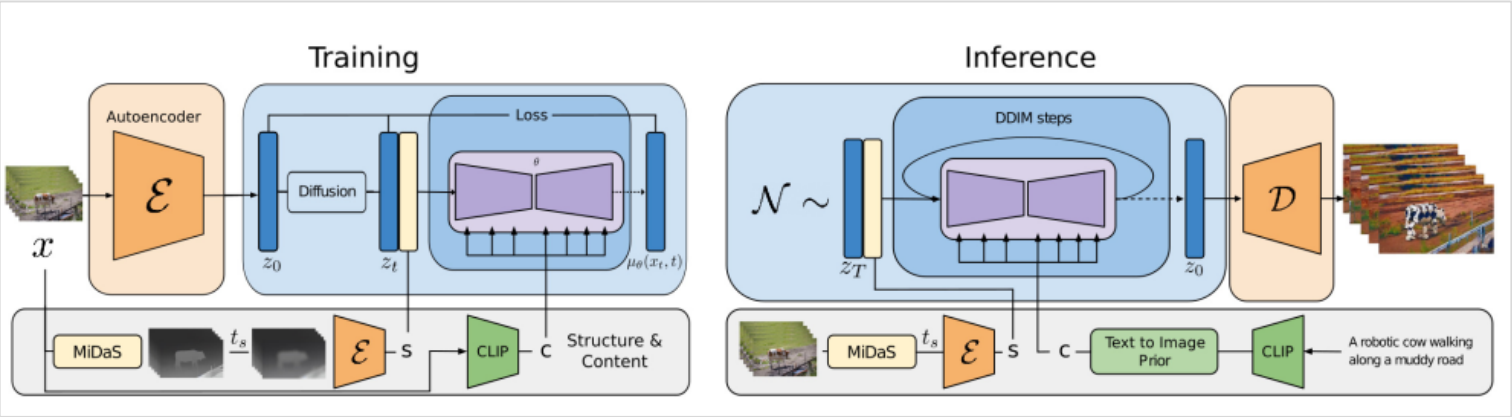
数据集： Zeroscope_v2_576w 采用 24 帧、576x320 分辨率的 9923 个剪辑和 29769 个标记帧进行训练。

优点：
- 开源模型能够充分集思广益，加速模型的发展与迭代，增加社区内用户的参与。
- 和Runway的Gen-1、Gen-2一样直达C端用户，目前其他文本生成视频模型并未开放给C端用户。

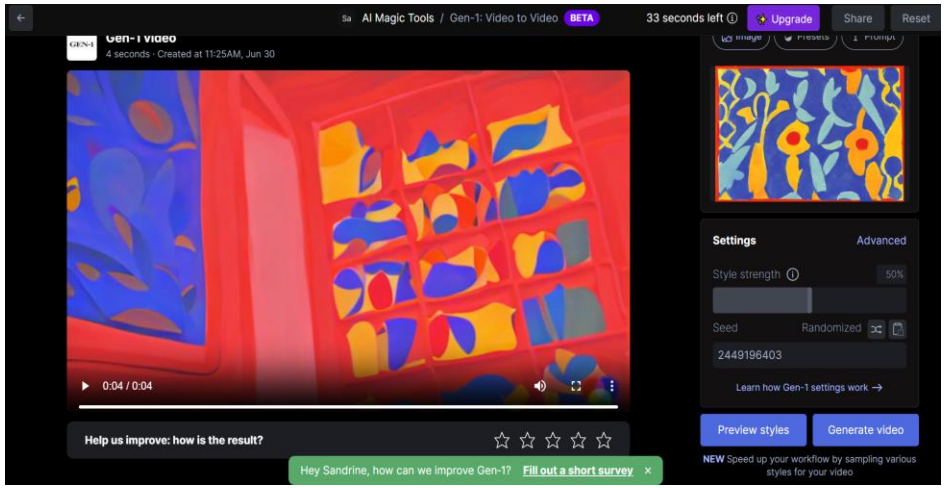
挑战：
- 由于模型开源，无公司及团队支撑，无明确的商业化路径。
- 无明确的团队支撑产品的迭代及研发，后续发展的形势不明确。

*Zeroscope无会议期刊

Runway Gen-1：基于潜在扩散模型，助力商用发展



生成步骤：
1、训练期间：输入视频后使用固定编码器编码并扩散。提取一个结构表示和一个内容表示，模型学习在潜在空间反转扩散过程
2、推理过程：输入视频的结构，通过先验将 CLIP 文本嵌入转换为图像嵌入，通过文本指定内容。



Gen-1使用说明：
用户基于自己对于预期产出视频的想象，可以从图像、预设、prompt（文字描述）等方面设置转化后视频的风格，可以预览生成的视频，能在5分钟内生成视频，并对生成的视频打分，提升平台对用户的理解。但能够生成的视频时长较短。（在Pro Plan的订阅下，Gen-1最多15秒）

扩散模型

Gen-1是Runway提出的转换视频风格的模型，于2023年2月推出商用，同期发布论文。Gen-1将潜在扩散模型扩展到视频生成，通过将时间层引入到预训练的图像模型中并对图像和视频进行联合训练。模型通过推断来自输入视频的形状表示，并修改它基于描述编辑的文本提示，目的是编辑视频内容保留其结构。

模型创新

- **无需额外训练和预处理：**结构和内容感知模型根据示例图像或文本修改视频，编辑完全在推理时执行，无需额外的每个视频训练或预处理。
- **控制与结构一致性：**模型展示了对时间、内容的完全控制和结构的一致性，对图像和视频数据的联合训练可以对时间一致性进行推理时间控制。为了结构一致性，对展示中的不同细节级别的训练允许在推理过程中选择所需的设置。
- **部分微调：**证明经过训练的模型可以进一步通过对一小组图像进行微调，进行定制以生成特定主题的更准确的视频

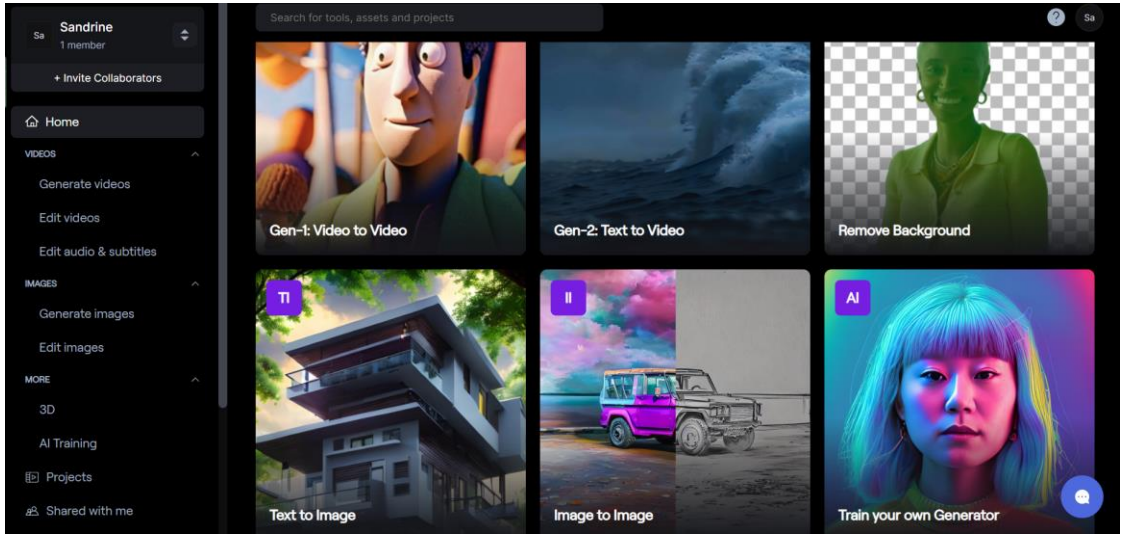
资料来源：《Structure and Content-Guided Video Synthesis with Diffusion Models》Patrick Esser等，Runway公司官网，国海证券研究所

Runway Gen-1 & Gen-2：商用文生视频的明星应用

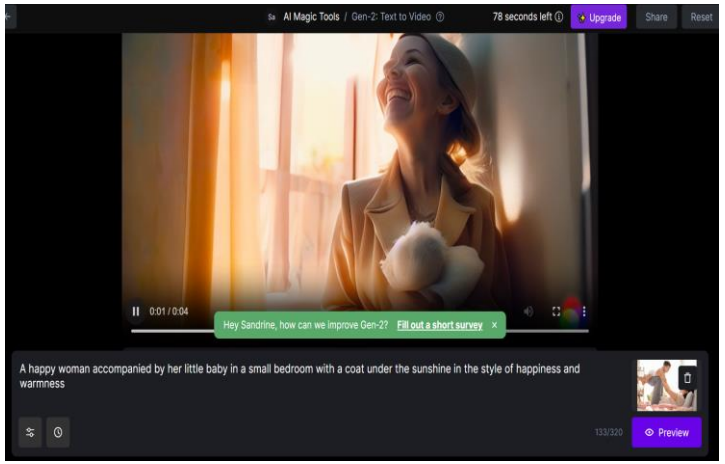
Runway 是基于生成式AI的图像和视频编辑软件供应商，是目前面向C端客户商业化的公司，由Cristóbal Valenzuela, Alejandro Matamala和Anastasis Germanidis创立于2018年，是福布斯AI50榜单：最有前途的人工智能公司之一，其总部位于美国纽约。公司坚持在AIGC领域，细分领域从原来的图片转换到视频的编辑与生成。Runway可以支持用户进行图像处理、文本生成图像、更改视频风格、文生视频等多项服务。

融资情况

- 2018年12月，Lux Capital领投Runway，种子轮融资200万美元；2020年12月Runway获得由Amplify Partners 领投的850万美元的A轮融资；2021年底，Runway以 2 亿美金的估值完成了 B 轮融资 3500 万美元融资；它在2022年12月从Felicis领投的C轮融资中获得了5000万美元的投资，公司估值为5亿美元。
- 6月，Runway宣布**完成了新一轮1.41亿美元的融资**，由谷歌领投，英伟达和Salesforce Ventures等跟投，目前**公司的估值已达15亿美元**。



Gen-2基于文字生成视频画面



产品版本

2021年
Latent Diffusion发布

2022年
Stable Diffusion发布

2023年2月
Gen-1：通过应用文本或参考图像所指定的风格将现有视频转换成新视频的模型

2023年3月
Gen-2：可以通过输入文本或参考图像从零生成视频。

商业化

- Runway的技术被广泛应用于电影、电视与广告等领域，奥斯卡获奖电影《瞬息全宇宙》背后的视觉效果团队使用了Runway的技术来帮助创建某些场景，比如用AI工具去除背景、放慢视频、制作无限延伸的图片等。
- 6月，Runway推出首批商业化的**文本转视频**的模型Gen-2，该模型能够根据文本和图像生成视频，目前可以**免费体验**。
- Runway为用户提供**免费试用额度**（125点/约26个视频），在付费订阅方面，通过销售月度“点数”（credits）供用户使用Gen-1、Gen-2等产品及增值服务，分别有**标准版**（$12/月-625点）和**高级版**（$28/月-2250点）。

Synthesia：海外领先的AI视频应用，已开启商业化

Synthesia是一个人工智能视频创作平台。该平台素材丰富，支持120多种语言，提供140+个AI化身；制作时间短，不需要视频剪辑技巧，可以在几分钟内创建带有AI化身和声音的视频。目前，Synthesia的服务的企业客户超过5万家，35%的财富100强企业是其忠实客户。Synthesia已自动生成了超过1200万个视频，用户增长率超过400%。

- **创始团队：创始人学历背景突出，来自多所知名大学。**Synthesia于 2017 年由来自伦敦大学学院、斯坦福大学、慕尼黑工业大学和剑桥大学的人工智能研究人员和企业家团队创立。
- **融资情况： 从2019年起进行多轮融资，跻身独角兽企业。**2023年6月13日，Synthesia 正式宣布完成了 9,000 万美金的 C 轮融资，估值达 10 亿美金正式晋升独角兽，本轮融资由Accel领投，NVentures、Kleiner Perkins、GV、Firstmark capital、Alex Wang、Olivier Pomel、Amjad Masad参投。

▼ Synthesia创始人团队

▼ Synthesia融资情况

2019年4月
获得310万美元种子轮融资

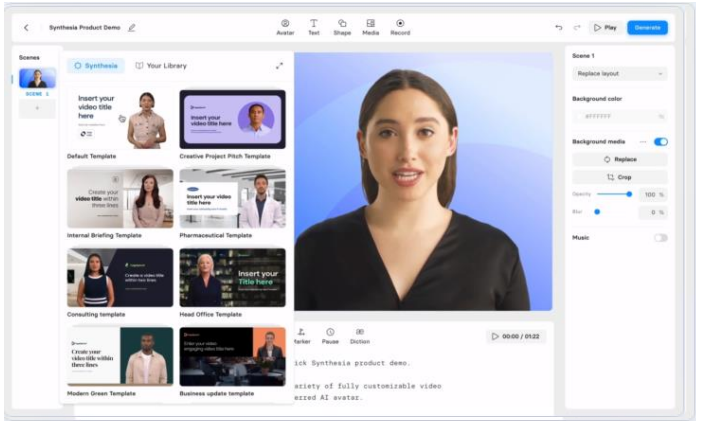
2021年4月
获得1,250万美元A轮融资

2021年12月
获得5,000万美元B轮融资

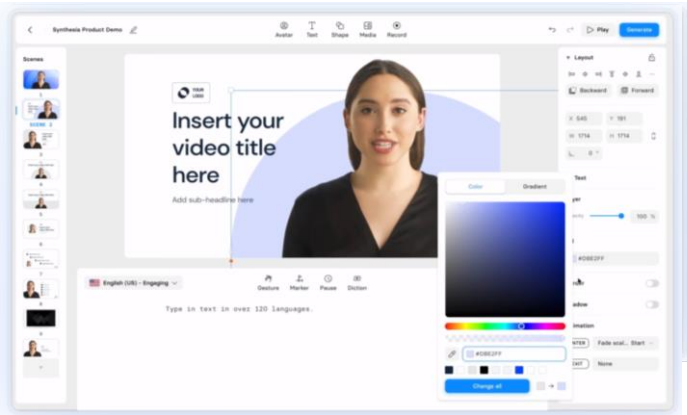
2023年6月
获得9,000万美元C轮融资

- **产品更新：更新速度快，更新内容丰富度高。**Synthesia的产品更新速度较快，日常更新主要为丰富素材库。更新的内容包括改进、新功能发布、AI形象、STUDIO、API、提醒、声音、模板等方向，并将不同方向设置成标签，标注在每次产品更新之后，帮助快速检索相关更新情况。
- **商业模式：以订阅费和定制费用为主。** Synthesia的产品分为个人版与企业版两个版本，**个人版本收取固定订阅费用，价格为22.5美元/月，全年订阅享受25%折扣；企业版本根据座位数的不同费用不同。**两种版本均可体验基础的视频制作功能，但在素材丰富度、特殊功能以及优先级上存在较大差距。

▼ 选择模板界面



▼ 输入文本界面

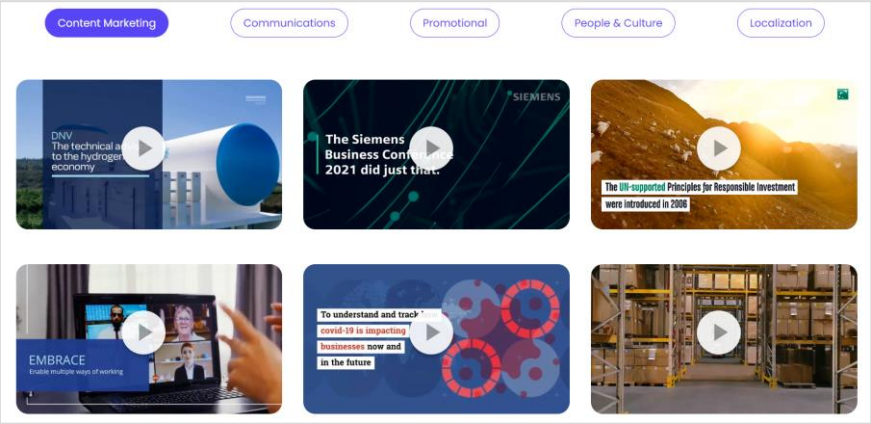


资料来源：AIGC开放社区， Synthesia公司官网，Hashdock，投中网，巴比特资讯，国海证券研究所

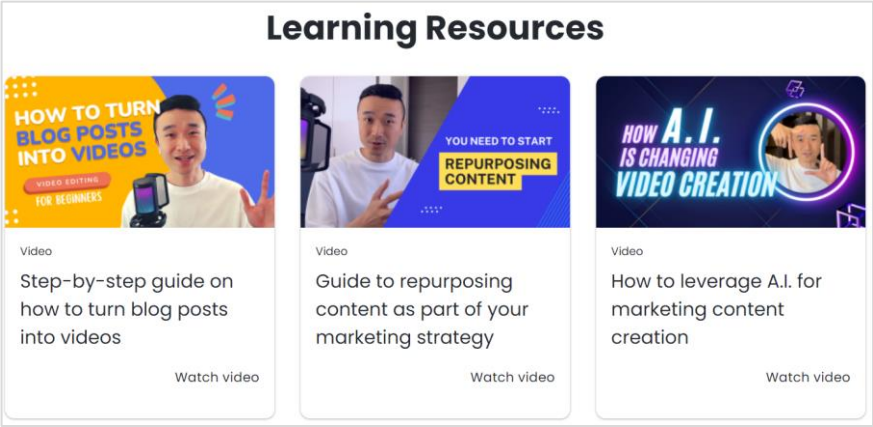
Lumen5：可将文本转化为视频，自动生成对应的场景和角色

Lumen5是一款视频创建软件，可以帮助营销人员、发布商和品牌创建视频内容，可以将博客文章（blog posts）转换为视频、头部说话内容（Talking Head Video），超过一百万家公司使用Lumen5来讲述他们的故事。

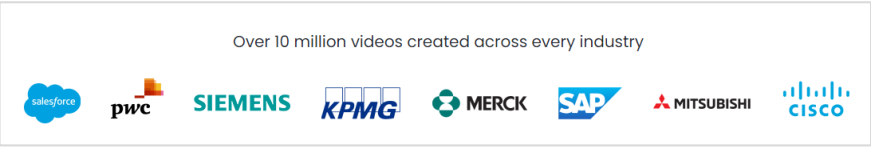
拥有丰富的应用场景

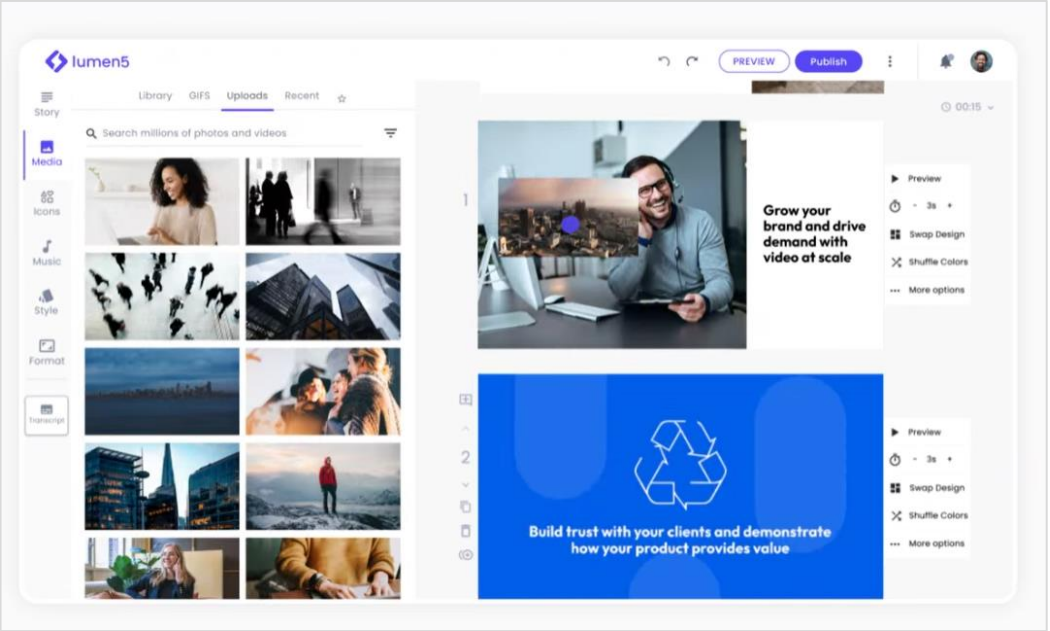


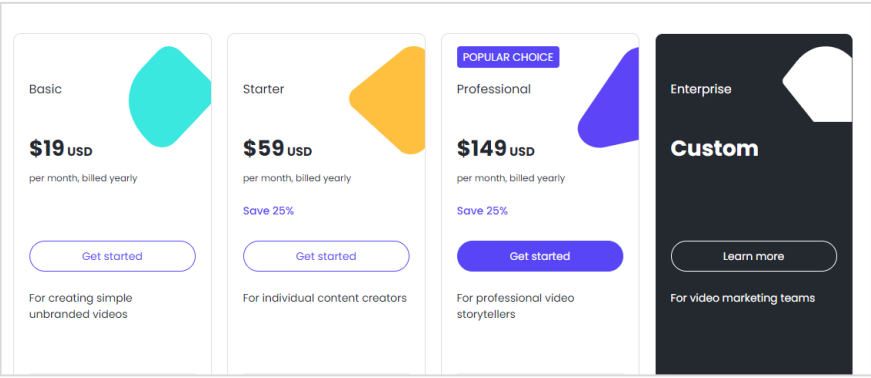
丰富的学习资源，让非专业人士轻松上手



客户群体丰富，商业化成熟







将博客文章转换为视频：

- 只需弹出博客链接即可开始，或者复制、粘贴内容至Lumen5
- AI自动生成视频；，Lumen5会总结文章，并且将场景与相关素材进行匹配；
- 能够将创造力和机器学习相结合。

创建头部说话内容：

- 通过标注和剪辑增强视频、视觉叠加来吸引观众；
- 自动为所有视频剪辑生成字幕。
- 通过转录进行编辑变得更容易。

成熟的定价策略：

- 分别有基础版、初学者版、专业版以及企业版，在不同的层次提供不同的级别的功能，用户也可以加入社群（community），免费使用部分功能；
- 能够提供1080p的视频分辨率、无水印的视频、500M的图像及视频库存，为企业版能提供定制化的品牌模板和设计团队

文生图和文生视频底层技术不断演进、模型持续迭代，涌现出一批优质原生AI应用，在C端开创了全新的应用体验，同时在B端游戏、营销、影视制作、文旅、电商等多个行业均开启应用，实现降本增效，长期有望进一步打开商业化空间。我们看好AI多模态行业投资机会，维持行业“推荐”评级，建议关注微软、Meta、Adobe、谷歌、百度、阿里巴巴、美图、万兴科技、新国都等相关标的。

风险提示

- 竞争加剧风险：文生图行业应用涌现，生成效果较为接近，存在竞争风险
- 内容质量不佳风险：文生图部分应用生成效果相对有限
- 用户流失风险：C端应用用户留存率不稳定，存在流失风险
- 政策监管风险：人工智能生成内容存在监管风险
- 变现不及预期风险：应用在商业化付费上存在不及预期风险
- 估值调整风险等：板块行业存在估值调整风险

- GAU-GAN
- GAU-GAN-2
- ViT-VQGAN
- 《Zero-Shot Text-to-Image Generation》Aditya Ramesh等
- 《Diffusion Models Beat GANs on Image Synthesis》Prafulla Dhariwal等
- 《CogView: Mastering Text-to-Image Generation via Transformers》Ming Ding等
- 《ERNIE-VILG: UNIFIED GENERATIVE PRE-TRAINING FOR BIDIRECTIONAL VISION-LANGUAGE GENERATION》Han Zhang等
- 《GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models》Alex Nichol等
- 《Make-A-Scene: Scene-Based Text-to-Image Generation with Human Priors》Oran Gafni等
- 《Hierarchical Text-Conditional Image Generation with CLIP Latents》Aditya Ramesh等
- 《High-Resolution Image Synthesis with Latent Diffusion Models》Robin Rombach等
- 《Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding》Chitwan Saharia等
- 《CogView2: Faster and Better Text-to-Image Generation via Hierarchical Transformers》Ming Ding等
- 《Scaling Autoregressive Models for Content-Rich Text-to-Image Generation》Jiahui Yu等
- 《ERNIE-ViLG 2.0: Improving Text-to-Image Diffusion Model with Knowledge-Enhanced Mixture-of-Denoising-Experts》Zhida Feng等
- 《Scaling Autoregressive Multi-Modal Models: Pretraining and Instruction Tuning》Lili Yu等

附录2-P8

- 《Generating Videos with Scene Dynamics》Carl Vondrick等
- 《VideoGPT: Video Generation using VQ-VAE and Transformers》Wilson Yan等
- 《Temporal Generative Adversarial Nets with Singular Value Clipping》Masaki Saito等
- 《MoCoGAN: Decomposing Motion and Content for Video Generation》Sergey Tulyakov等
- 《ADVERSARIAL VIDEO GENERATION ON COMPLEX DATASETS》Aidan Clark等
- 《GENERATING VIDEOS WITH DYNAMICS-AWARE IMPLICIT GENERATIVE ADVERSARIAL NETWORKS》Sihyun Yu等
- 《GODIVA: Generating Open-DomaIn Videos from nAtural Descriptions》 Chenfei Wu等
- 《NÜWA: Visual Synthesis Pre-training for Neural visUal World creAtion》Chenfei Wu等
- 《CogVideo: Large-scale Pretraining for Text-to-Video Generation via Transformers》Wenyi Hong等
- 《MAKE-A-VIDEO: TEXT-TO-VIDEO GENERATION WITHOUT TEXT-VIDEO DATA》Uriel Singer等
- 《IMAGEN VIDEO: HIGH DEFINITION VIDEO GENERATION WITH DIFFUSION MODELS》Jonathan Ho等
- 《PHENAKI: VARIABLE LENGTH VIDEO GENERATION FROM OPEN DOMAIN TEXTUAL DESCRIPTIONS》Ruben Villegas等
- 《MagicVideo: Efficient Video Generation With Latent Diffusion Models》Daquan Zhou等
- 《Tune-A-Video: One-Shot Tuning of Image Diffusion Models for Text-to-Video Generation》Jay Zhangjie Wu等
- 《Text2Video-Zero:Text-to-Image Diffusion Models are Zero-Shot Video Generators》Levon Khachatryan等
- 《NUWA-XL: Diffusion over Diffusion for eXtremely Long Video Generation》Shengming Yin等
- 《Structure and Content-Guided Video Synthesis with Diffusion Models》Patrick Esser等
- 《Align your Latents: High-Resolution Video Synthesis with Latent Diffusion Models》Andreas Blattmann等
- 《Preserve Your Own Correlation: A Noise Prior for Video Diffusion Models》Songwei Ge等

- 《Generating Videos with Scene Dynamics》Carl Vondrick等
- 《VideoGPT: Video Generation using VQ-VAE and Transformers》Wilson Yan等
- 《Temporal Generative Adversarial Nets with Singular Value Clipping》Masaki Saito等
- 《MoCoGAN: Decomposing Motion and Content for Video Generation》Sergey Tulyakov等
- 《ADVERSARIAL VIDEO GENERATION ON COMPLEX DATASETS》Aidan Clark等
- 《GENERATING VIDEOS WITH DYNAMICS-AWARE IMPLICIT GENERATIVE ADVERSARIAL NETWORKS》Sihyun Yu等

- 《Video Diffusion Models》Jonathan Ho等
- 《MAKE-A-VIDEO: TEXT-TO-VIDEO GENERATION WITHOUT TEXT-VIDEO DATA》Uriel Singer等
- 《IMAGEN VIDEO: HIGH DEFINITION VIDEO GENERATION WITH DIFFUSION MODELS》Jonathan Ho等
- 《Tune-A-Video: One-Shot Tuning of Image Diffusion Models for Text-to-Video Generation》Jay Zhangjie Wu等
- 《Structure and Content-Guided Video Synthesis with Diffusion Models》Patrick Esser等
- 《Dreamix: Video Diffusion Models are General Video Editors》Eyal Molad等
- 《NUWA-XL: Diffusion over Diffusion for eXtremely Long Video Generation》Shengming Yin等
- 《Text2Video-Zero:Text-to-Image Diffusion Models are Zero-Shot Video Generators》Levon Khachatryan等
- 《Align your Latents: High-Resolution Video Synthesis with Latent Diffusion Models》Andreas Blattmann等
- 《Preserve Your Own Correlation: A Noise Prior for Video Diffusion Models》Songwei Ge等

- 《Zero-Shot Text-to-Image Generation》Aditya Ramesh等
- 《Diffusion Models Beat GANs on Image Synthesis》Prafulla Dhariwal等
- 《CogView: Mastering Text-to-Image Generation via Transformers》Ming Ding等
- 《ERNIE-VILG: UNIFIED GENERATIVE PRE-TRAINING FOR BIDIRECTIONAL VISION-LANGUAGE GENERATION》Han Zhang等
- 《GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models》Alex Nichol等
- 《Make-A-Scene: Scene-Based Text-to-Image Generation with Human Priors》Oran Gafni等
- 《Hierarchical Text-Conditional Image Generation with CLIP Latents》Aditya Ramesh等
- 《High-Resolution Image Synthesis with Latent Diffusion Models》Robin Rombach等
- 《Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding》Chitwan Saharia等
- 《CogView2: Faster and Better Text-to-Image Generation via Hierarchical Transformers》Ming Ding等
- 《Scaling Autoregressive Models for Content-Rich Text-to-Image Generation》Jiahui Yu等
- 《ERNIE-ViLG 2.0: Improving Text-to-Image Diffusion Model with Knowledge-Enhanced Mixture-of-Denoising-Experts》Zhida Feng等
- 《Scaling Autoregressive Multi-Modal Models: Pretraining and Instruction Tuning》Lili Yu等
- 《Improving Image Generation with Better Captions》James Betker等

- 《GODIVA: Generating Open-DomaIn Videos from nAtural Descriptions》 Chenfei Wu等
- 《NÜWA: Visual Synthesis Pre-training for Neural visUal World creAtion》Chenfei Wu等
- 《CogVideo: Large-scale Pretraining for Text-to-Video Generation via Transformers》Wenyi Hong等
- 《MAKE-A-VIDEO: TEXT-TO-VIDEO GENERATION WITHOUT TEXT-VIDEO DATA》Uriel Singer等
- 《IMAGEN VIDEO: HIGH DEFINITION VIDEO GENERATION WITH DIFFUSION MODELS》Jonathan Ho等
- 《PHENAKI: VARIABLE LENGTH VIDEO GENERATION FROM OPEN DOMAIN TEXTUAL DESCRIPTIONS》Ruben Villegas等
- 《MagicVideo: Efficient Video Generation With Latent Diffusion Models》Daquan Zhou等
- 《Tune-A-Video: One-Shot Tuning of Image Diffusion Models for Text-to-Video Generation》Jay Zhangjie Wu等
- 《Text2Video-Zero:Text-to-Image Diffusion Models are Zero-Shot Video Generators》Levon Khachatryan等
- 《NUWA-XL: Diffusion over Diffusion for eXtremely Long Video Generation》Shengming Yin等
- 《Structure and Content-Guided Video Synthesis with Diffusion Models》Patrick Esser等
- 《Align your Latents: High-Resolution Video Synthesis with Latent Diffusion Models》Andreas Blattmann等
- 《Preserve Your Own Correlation: A Noise Prior for Video Diffusion Models》Songwei Ge等

研究小组介绍

海外小组介绍

陈梦竹，南开大学本科&硕士，6年证券从业经验，现任国海证券海外研究团队首席分析师，专注于全球内容&社交互联网、消费互联网、科技互联网板块研究。
尹芮，康奈尔大学硕士，中国人民大学本科，2年证券从业经验，现任国海证券海外互联网分析师，主要覆盖内容&社交互联网方向。
张娟娟，上海财经大学硕士，三年产业工作经验，曾任职于阿里、美团，现任国海证券海外互联网研究助理，主要覆盖生活互联网方向。
罗婉琦，伦敦政治经济学院硕士，现任国海证券海外研究团队研究助理，主要覆盖消费互联网方向。

分析师承诺

陈梦竹, 尹芮, 本报告中的分析师均具有中国证券业协会授予的证券投资咨询执业资格并注册为证券分析师，以勤勉的职业态度，独立，客观的出具本报告。本报告清晰准确的反映了分析师本人的研究观点。分析师本人不曾因，不因，也将不会因本报告中的具体推荐意见或观点而直接或间接收取到任何形式的补偿。

国海证券投资评级标准

行业投资评级

推荐：行业基本面向好，行业指数领先沪深300指数；
中性：行业基本面稳定，行业指数跟随沪深300指数；
回避：行业基本面向淡，行业指数落后沪深300指数。

股票投资评级

买入：相对沪深300 指数涨幅20%以上；
增持：相对沪深300 指数涨幅介于10%～20%之间；
中性：相对沪深300 指数涨幅介于-10%～10%之间；
卖出：相对沪深300 指数跌幅10%以上。

免责声明和风险提示

免责声明

本报告的风险等级定级为R4，仅供符合国海证券股份有限公司（简称“本公司”）投资者适当性管理要求的的客户（简称“客户”）使用。本公司不会因接收人收到本报告而视其为客户。客户及/或投资者应当认识到有关本报告的短信提示、电话推荐等只是研究观点的简要沟通，需以本公司的完整报告为准，本公司接受客户的后续问询。

本公司具有中国证监会许可的证券投资咨询业务资格。本报告中的信息均来源于公开资料及合法获得的相关内部外部报告资料，本公司对这些信息的准确性及完整性不作任何保证，不保证其中的信息已做最新变更，也不保证相关的建议不会发生任何变更。本报告所载的资料、意见及推测仅反映本公司于发布本报告当日的判断，本报告所指的证券或投资标的的价格、价值及投资收入可能会波动。在不同时期，本公司可发出与本报告所载资料、意见及推测不一致的报告。报告中的内容和意见仅供参考，在任何情况下，本报告中所表达的意见并不构成对所述证券买卖的出价和征价。本公司及其本公司员工对使用本报告及其内容所引发的任何直接或间接损失概不负责。本公司或关联机构可能会持有报告中所提到的公司所发行的证券头寸并进行交易，还可能为这些公司提供或争取提供投资银行、财务顾问或者金融产品等服务。本公司在知晓范围内依法合规地履行披露义务。

风险提示

市场有风险，投资需谨慎。投资者不应将本报告为作出投资决策的唯一参考因素，亦不应认为本报告可以取代自己的判断。在决定投资前，如有需要，投资者务必向本公司或其他专业人士咨询并谨慎决策。在任何情况下，本报告中的信息或所表述的意见均不构成对任何人的投资建议。投资者务必注意，其据此做出的任何投资决策与本公司、本公司员工或者关联机构无关。

若本公司以外的其他机构（以下简称“该机构”）发送本报告，则由该机构独自为此发送行为负责。通过此途径获得本报告的投资者应自行联系该机构以要求获悉更详细信息。本报告不构成本公司向该机构之客户提供的投资建议。

任何形式的分享证券投资收益或者分担证券投资损失的书面或口头承诺均为无效。本公司、本公司员工或者关联机构亦不为该机构之客户因使用本报告或报告所载内容引起的任何损失承担任何责任。

郑重声明

国海证券 SEALAND SECURITIES

国海证券 · 研究所 · 海外研究团队

心怀家国，洞悉四海

国海研究上海

上海市黄浦区绿地外滩中心C1栋国海证券大厦

邮编：200023

电话：021-61981300

国海研究深圳

深圳市福田区竹子林四路光大银行大厦28F

邮编：518041

电话：0755-83706353

国海研究北京

北京市海淀区西直门外大街168号腾达大厦25F

邮编：100044

电话：010-88576597