



颠覆性重构

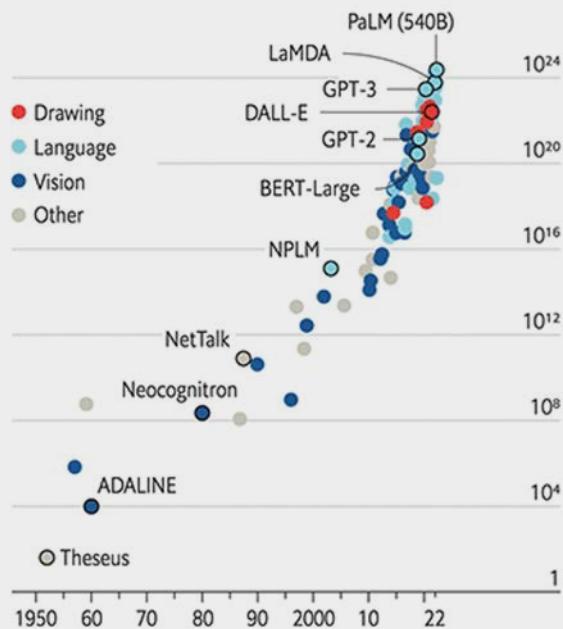
AI视频大模型的征程与展望

生数科技 & 计算机系 06级校友 唐家渝

拥抱规模效应：人工智能过去十年最大变革

The blessings of scale

AI training runs, estimated computing resources used
Floating-point operations, selected systems, by type, log scale



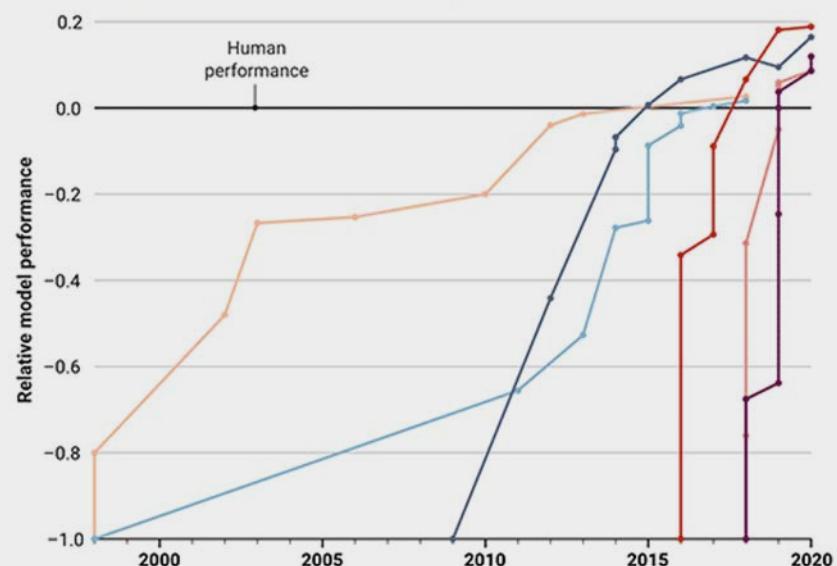
Sources: "Compute trends across three eras of machine learning", by J. Sevilla et al., arXiv, 2022; Our World in Data

Quick learners

The speed at which artificial intelligence models master benchmarks and surpass human baselines is accelerating. But they often fall short in the real world.

Benchmarks

- MNIST (handwriting recognition)
- Switchboard (speech recognition)
- ImageNet (image recognition)
- SQuAD 1.1 (reading comprehension)
- SQuAD 2.0 (reading comprehension)
- GLUE (language understanding)



(GRAPHIC) K. FRANKLIN/SCIENCE; (DATA) D. KIELA ET AL., DYNABENCH: RETHINKING BENCHMARKING IN NLP, DOI:10.48550/ARXIV.2104.14337

生成式人工智能为智能社会发展推开一扇新大门

2015-2022

小模型时代：通用性被广泛接受和研究之前，人工智能主要为特定任务或领域设计，例如图像识别、语音识别、文本分类等。

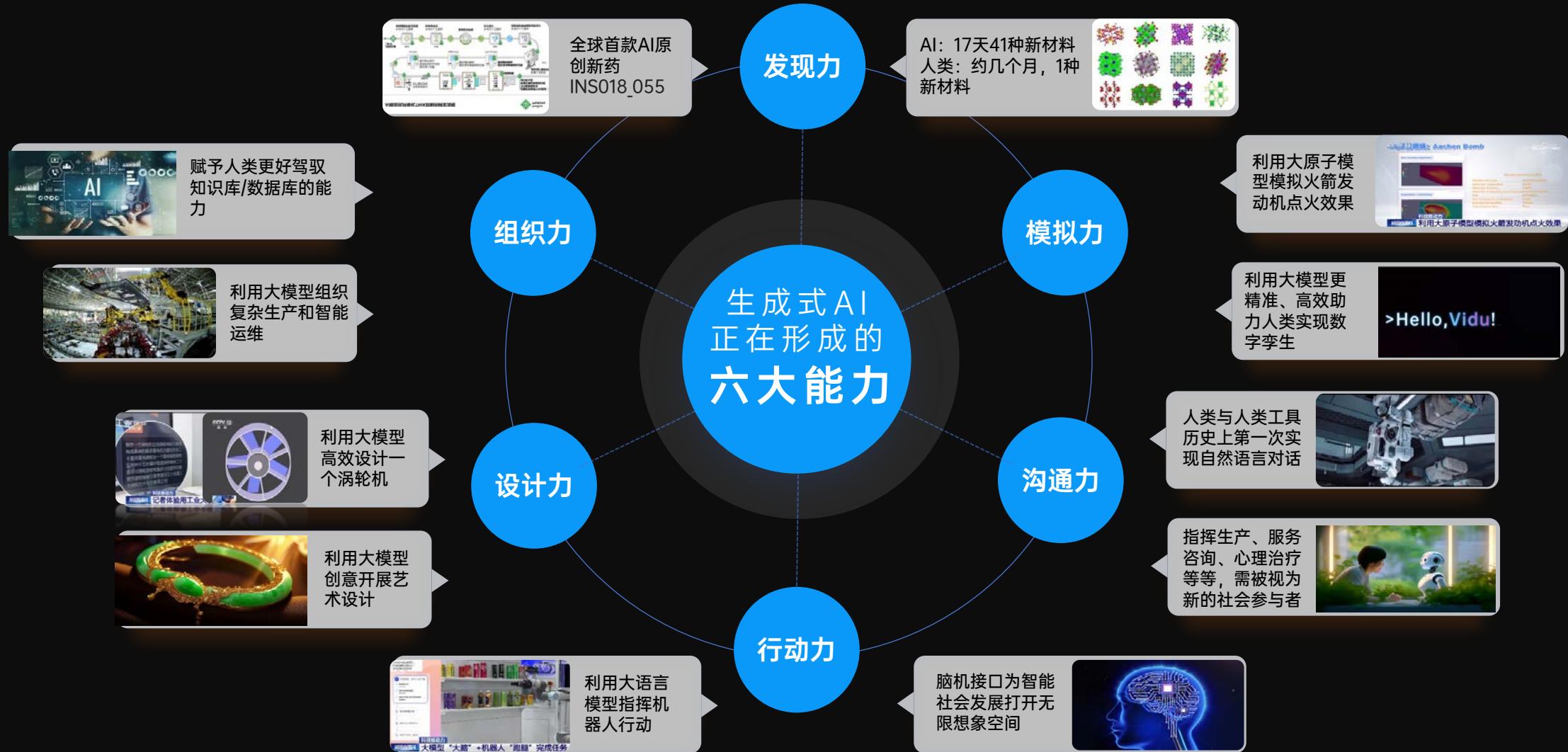
生成式人工智能：成为当前人类找到最靠近通用人工智能技术路径，推动了此轮技术革命和产业变革，成为发展新质生产力的重要引擎



2022~

大模型时代：通过新网络结构、新概率表示方法、大参数量、大数据等技术要素，大幅提升了AI的知识和逻辑推理能力，逐步推进人工智能从专用（只能处理特定任务）向通用（可以处理各种任务）转变。

生成式人工智能正重塑人类社会生产力





生生

数数

任何足够先进的技术都等同于魔术。

—— 亚瑟·C·克拉克



与互联网时代相比，AI 正生猛地
改变这个世界

多模态领域特别是
AI 视频生成
正是这其中颠覆性的重要变化之一

这是一支正片中没出现过的哪吒花絮



裂空爪围困陈塘关画面



申公豹变豹子头

(过去) (现在)

三个月 vs 10秒



这是一支关于“回家”的短片



(过去)

(现在)

20人团队

1人

1个月

12天

制作 1 分钟

制作 3 分钟

《家的回响》



这是一支好莱坞电影宣传片 中国水墨动画风格



0.5 天 生成上千支视频素材

10 天 完成整个片子的制作

90 % 替代近90%的后期投入

◀ 《毒液：最后一舞》中国宣传片

这是《熊猫计划》“呼呼”宣传片段



批量生成 短视频素材

低成本打造 “IP 短视频账号”

为影视宣发提供新思路

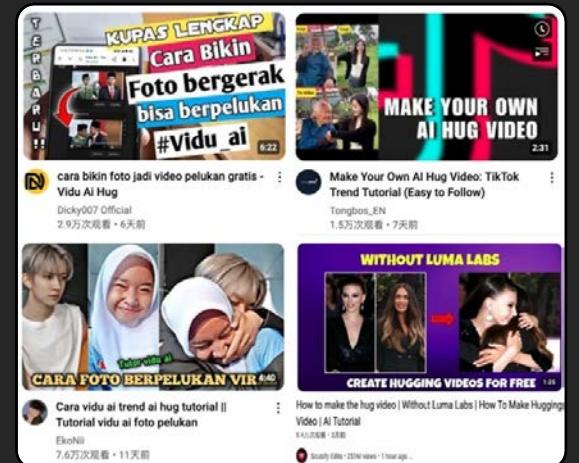
这是在 TikTok 上获得百万点赞的视频



一位海外素人博主，她用 Vidu 制作了一段跨越时空拥抱的视频，获得**近百万**点赞。



像这样的 Vidu 视频在 TikTok 上有**数十万条**。然后在 YouTube，也涌现了一批教大家用 Vidu 如何制作拥抱视频的教学视频，单个视频观看量都接近破 **10w**。





1. 多模态大模型的技术发展
2. 多模态大模型的行业背景
3. 视频大模型的创新与实践
4. 视频大模型 2025 趋势预测





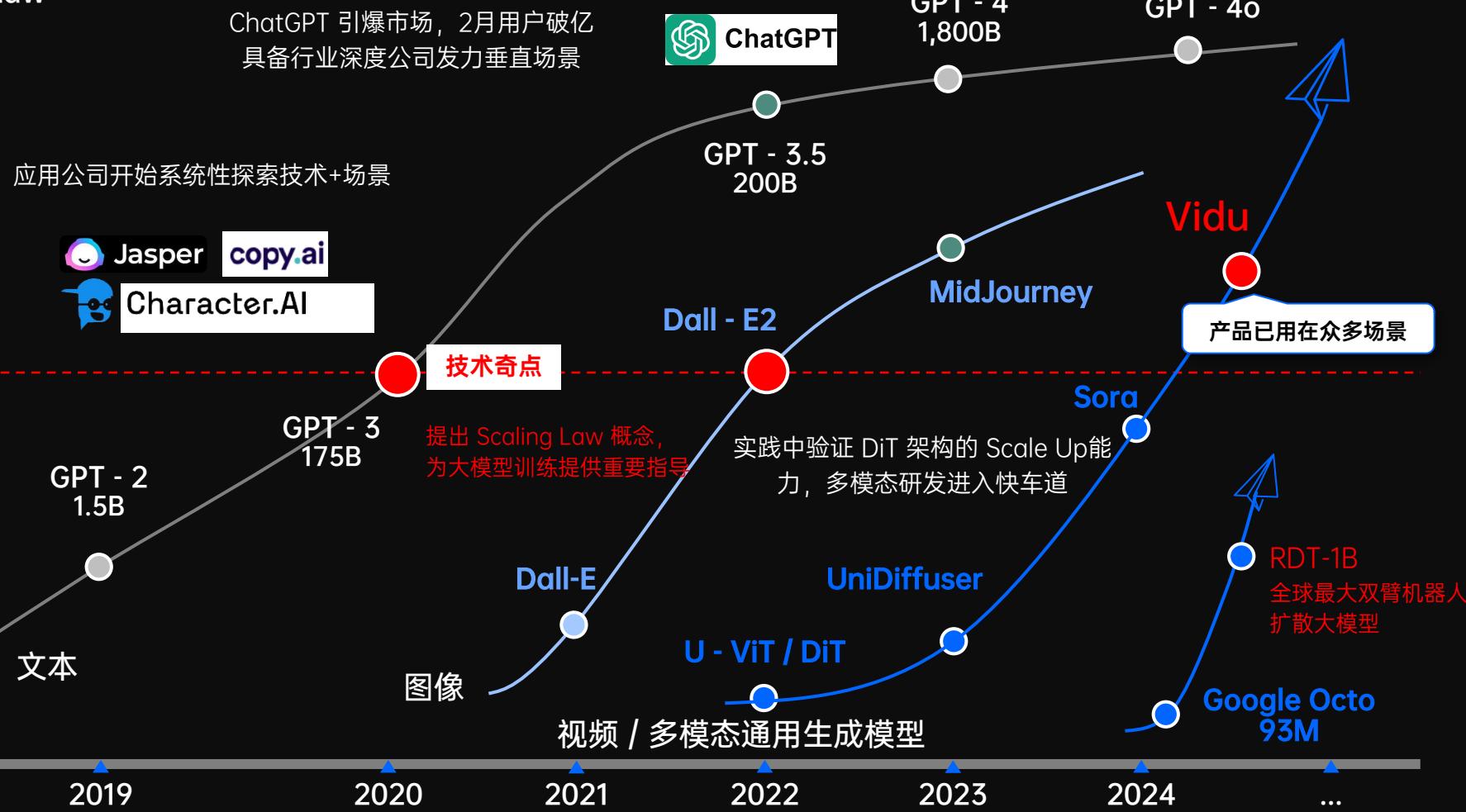
Part 1 多模态大模型的技术发展

多模态大模型发展总体脉络

Scaling Law

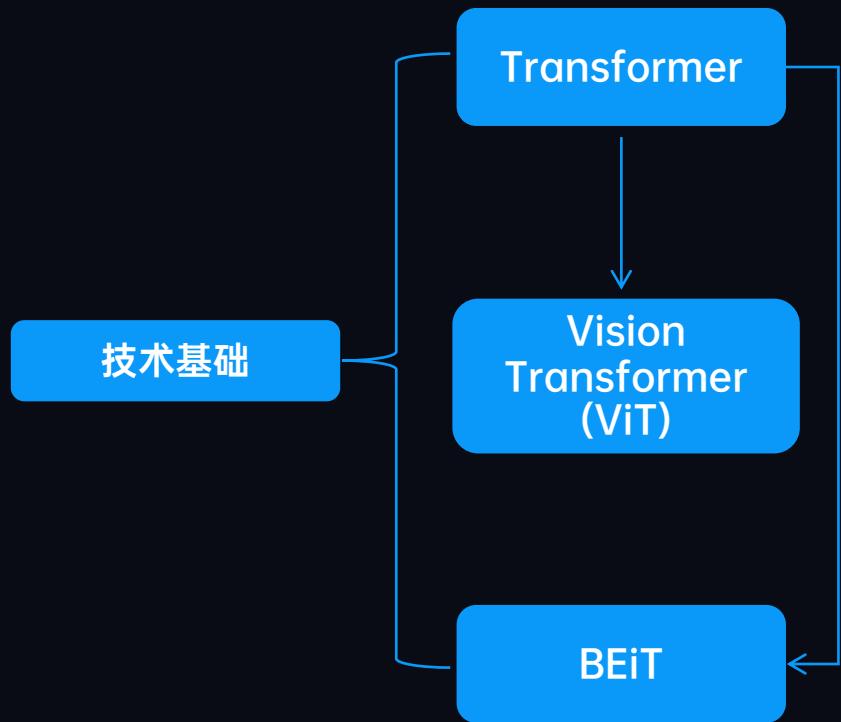
场景探索

基础研发



多模态大模型已完成技术论证，在拥有更为丰富、更多模态的训练数据基础上，将产生更高维度的“智能涌现”，解锁更多应用场景

三大技术基础支撑多模态大模型的发展起步



2017年，基于 Transformer 架构的大模型取得很好的效果，例如BERT、GPT，但仅局限于文本单一模态上，无法将 self-attention 中良好的泛化性迁移到图像、视频等其他模态中

ViT 的出现打通了 NLP 和 CV 的壁垒，Transformer 的限制在于其输入数据的大小，ViT 模型通过将图片 patch 化，实现了将 Transformer 应用于图像领域的问题，同时利用 patch embedding 方法实现高效提取图像特征

生成式预训练是自监督学习重要方法和训练目标，生成式预训练在自然语言处理中取得较大成功。BEiT 模型的出现，将生成式预训练从 NLP 迁移到 CV 上，就是将 BERT 中的掩码语言学习（MIM）方法应用到图像领域

多模态大模型的技术发展阶段

第一阶段

处理信息的基础能力
逻辑推理和生成

语言模态

第二阶段

提高信息密度
增加图像理解和生成能力

图像模态

第三阶段

模态发散
应对多模态理解和生成场景

三维模态

视频模态

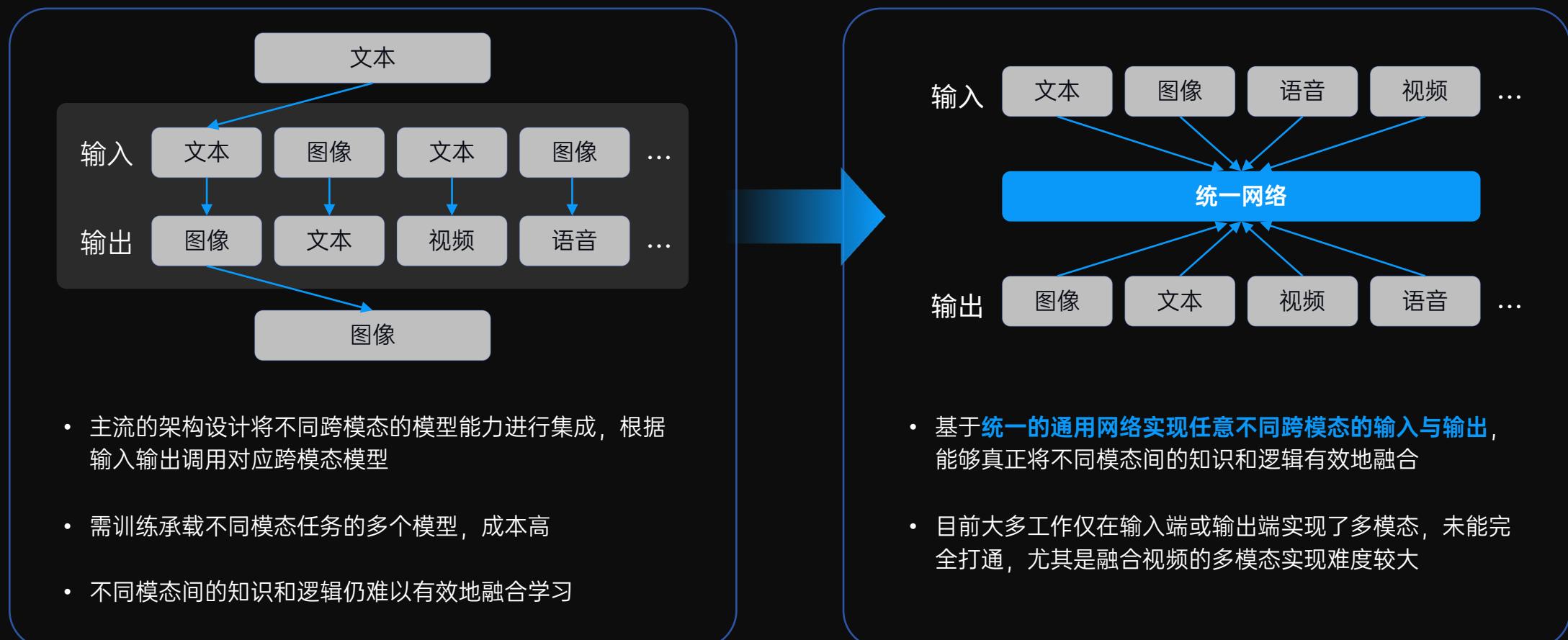
语音模态

其他.....

整体发展趋势呈现，最开始局限于文图的单一跨模态，其次在输入端实现了文图等多模态，在逐步过渡到输入、输出端均实现多模态，不仅模态数增多，模态融合程度在逐步增强

多模态大模型的发展趋势是统一化与通用化

设计统一的、跨任务、跨模态的多模态预训练模型，包括**统一的结构、统一的参数、统一的训练方式**，应对多个多模态任务或单模态任务





视频 的信息最为丰富
使其成为实现多模态统一的
关键模态

Sora的出现，使业界真正重视起视频生成



《Sora 宣传视频》

Sora来了,我们该如何应对

2024年2月12日 过去数天的视频制作过程,对Sora来说只是简单几句话的事。OpenAI发布Sora模型几天后,300位创始人及机构表示,“AGI(通用人工智能)的实现将从10年缩短至两三年”,国内视频生成软件Pika创始人郭文强,开始筹备对Sora的新产品:Stability AI CEO马特·莫斯塔克不由一...
由中国青年报

Sora未公开发布已掀起“掘金热潮”?专家:警惕这些不良现象

2024年2月12日 近日,OpenAI发布了首款文本视频模型——Sora,仅需要素即可自动生成一段长达60秒的高质量视频。Sora的横空出世,在AIGC领域引起不小的轰动。然而,在Sora还未公测前,已有商家以售卖相关...
红星新闻

Sora目前唯一体验方式,原来藏在了官网里

不久前,有博主通过观看OpenAI官方生成的Sora视频,建立个人网站Sora FM吸引流量,并计划在OpenAI正式发布Sora内测后,通过套壳充值成为“付费”方式进行变现。但很快,该博主发现,已有多家商家以售卖相关...
虎嗅APP

新闻透视:“Sora爆火 国产大模型如何迎头赶上”

2024年2月12日 随着聊天机器人ChatGPT之后,近日,由美国人工智能公司OpenAI近日推出的文本视频工具Sora横空出世,震动科技圈的同时,也在春节后引爆了AIGC(AI+生成内容)板块。Wind数据显示,春节后,自2月19日至...
经济观察报

制作门槛大幅降低:深度伪造引发担忧!“Sora来了”视频业何去何从?

2024年2月12日 【钛晨报】记者 丁雅涵 环球时报驻美特别约稿记者 郭亚凡 编译者 张锐 美国人工智能公司OpenAI近日推出的生成式人工智能模型Sora一夜之间刷屏,仍进去一段文字,很快生成“大片”质量的...
环球时报

The Conversation
<https://theconversation.com/openais-new-generative-model-sora-could-revolutionize-video-generation-157111> · 翻译此页

OpenAI's new generative tool Sora could revolutionize video generation

2024年2月21日 —— Sora is a text-to-video model that significantly advances the integration of deep learning, natural language processing and computer vision to transform textual ...

CBS News
<https://www.cbsnews.com/news/openai-sora/> · 翻译此页

OpenAI's new text-to-video tool, Sora, has one artificial ...

2024年2月16日 —— Sora, an AI application that takes written prompts and turns them into original videos, is already so powerful that one AI expert says it has him “terrified.”

YouTube - The Wall Street Journal
54.5万+ 次观看 · 1条评论

OpenAI's Sora Made Me Crazy AI Videos—Then the CTO ...

OpenAI CEO Sam Altman and CTO Mira Murati on the Future of AI and ChatGPT | WSJ Tech Live 2023 The Wall Street Journal

The Conversation
<https://theconversation.com/what-is-sora-a-new-generative-ai-tool-could-transform-video-generation-157111> · 翻译此页

What is Sora? A new generative AI tool could transform ...

2024年2月19日 —— OpenAI announced a new generative AI system named Sora, which produces short videos from text prompts.

Nature
<https://www.nature.com/news/openai-sora-could-change-science.html> · 翻译此页

How OpenAI's text-to-video tool Sora could change science

2024年3月11日 —— Text-to-video AI tools such as Sora could help researchers to wade through huge data sets, such as those produced by the European particle-physics laboratory ...

The New York Times
<https://www.nytimes.com/2024/02/15/technology/openai-sora.html> · 翻译此页

OpenAI Unveils Sora, an A.I. That Generates Eye-Popping ...

2024年2月16日 —— Sora generates videos in response to short descriptions, like “a gorgeously rendered papercraft world of a coral reef, rete with colorful fish and sea creatures ...

What Is Sora? A New Generative AI Tool Could Transform ...

2024年2月19日 —— OpenAI announced a new generative AI system named Sora, which produces short videos from text prompts.

在海内外引发大量关注和讨论



Sora 推动视频生成逼近“GPT-3”时刻

指标	Diffusion Transformer 视频模型 (Sora为例)	Diffusion 视频模型
视频时长	60s	2 - 4s
世界理解能力	强	弱
数字世界模拟	支持	不支持
物体一致性	强	弱
物体连续性	强	弱
架构	Transfromer	U - Net
文本理解能力	强	一般
清晰度	1080p (端到端生成)	最高4k
扩展视频生成	前/后	后
驱动方式	数据驱动	图片驱动
视频到视频剪辑	强	弱
世界互动能力	强	弱
原生纵横比	强	弱
无缝连接能力	强	弱
3D运动连贯性	强	弱

Diffusion Transformer 架构

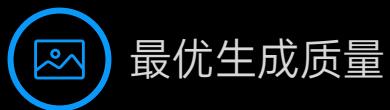
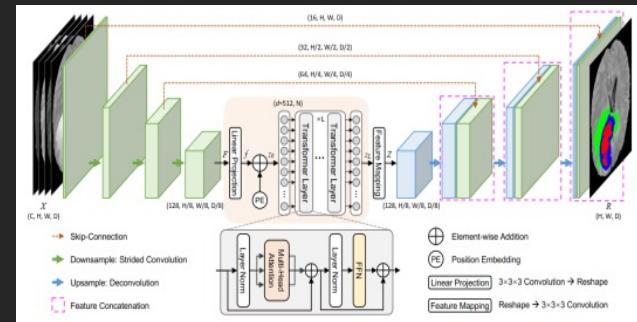
以扩散模型 (Diffusion Model) 为底座

模拟物理学中的“扩散现象”，先加噪再通过去噪来生成图像
天然更适合视觉数据的处理



Transformer 替代传统 U-Net 网络

在噪音预测环节用 Transformer 替换常用的 U-Net
天然可扩展性更强



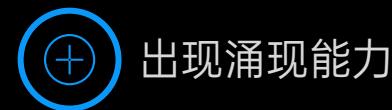
最优生成质量



计算开销可控



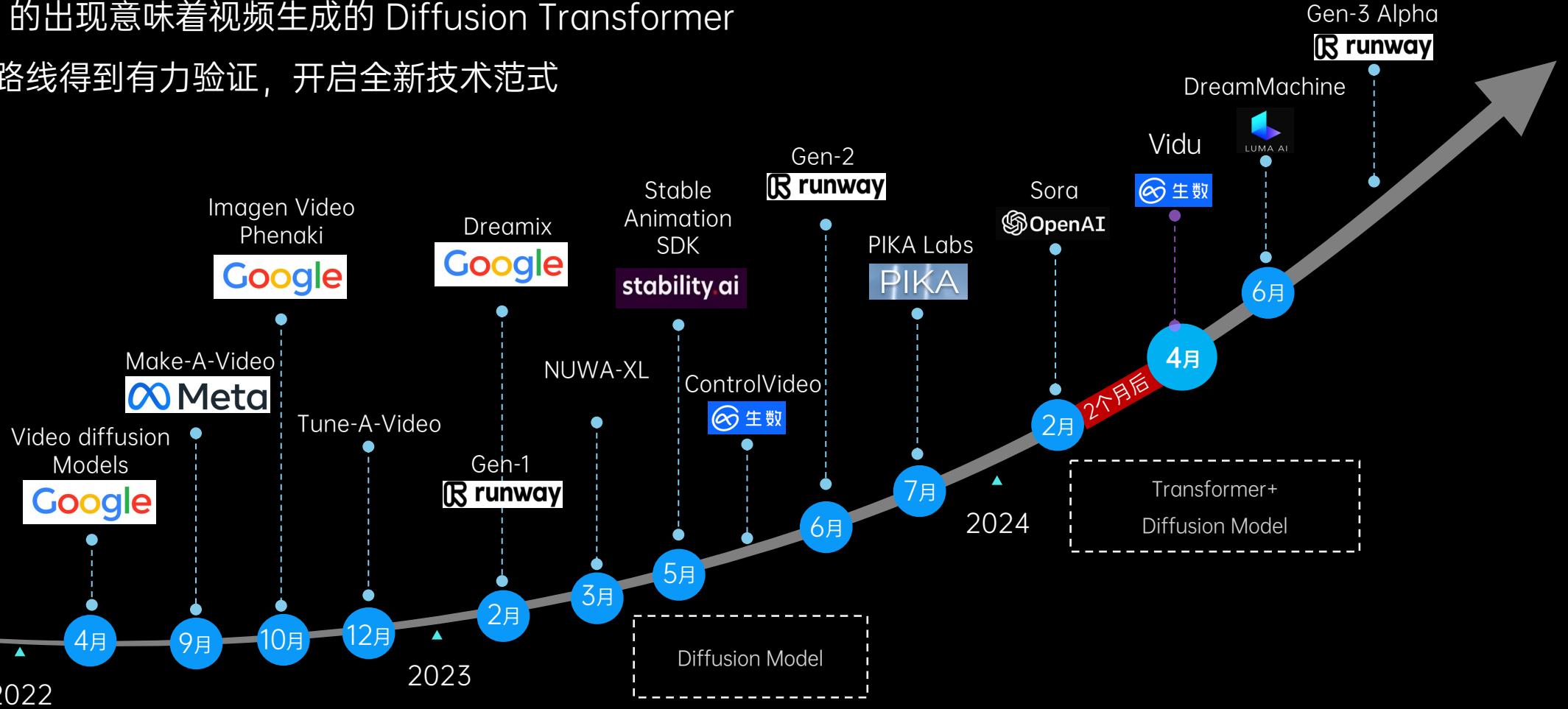
参数规模扩展



出现涌现能力

视频大模型的路线统一

Sora 的出现意味着视频生成的 Diffusion Transformer 技术路线得到有力验证，开启全新技术范式





视频大模型的迭代不再只是暴力美学

算力竞赛

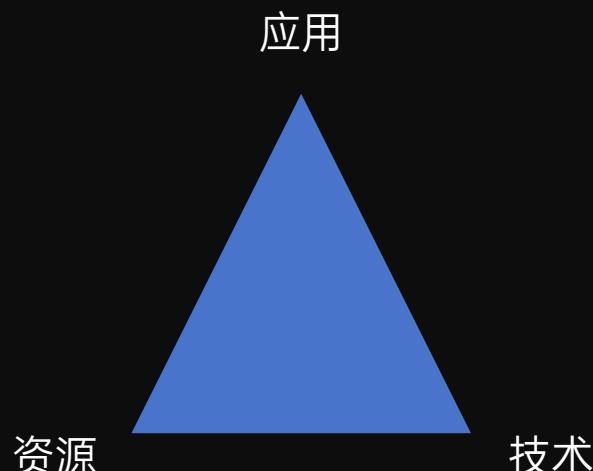
算法优化



Part 2 多模态大模型的行业背景

商业应用是 AI 可持续发展的加速器

技术、应用、资源在 AI 的发展中需要互相协同，相互促进，单独发展某个方面，都无法实现飞轮式的发展



01

商业应用为 AI 提供使用场景，在AI与各场景不断的实践中，促进技术不断迭代

02

资源在应用的促进下，向技术倾斜，给予技术进步的空间

03

技术在应用中让企业/行业在生产过程中不断优化流程，最大化利用人才和资本等资源

AI 工具的应用意愿普遍存在

使用 AI 最希望达到的是
提高制作效率

2024 年对海外媒体娱乐公司
应用 AI 意愿的问卷调查

行业	将在全环节应用AI	应用于简单的重复的任务	定量分析	提高制作效率	生成更多可选项以供参考	辅助设计过程
媒体娱乐	32%	37%	33%	46%	36%	34%
广告视觉设计	33%	35%	31%	47%	35%	34%
影视	33%	41%	36%	44%	36%	35%
游戏	30%	37%	32%	47%	39%	32%

数据来源：Autodesk

备注：该问卷调查受访者来自 1579 家影视及游戏公司，其工作年限平均为 11 年，70% 为决策人员。
这些公司的注册地主要为欧美

多模态技术在影视行业的渗透正在加深



2022年3月美国上映

仅 **2500万** 美元预算的特效大电影
用 AI 辅助影片特效制作

“仅仅**几次点击**就让我**节省几个小时**，
我可以用这些时间尝试三四种不同的效果，
让影片效果更好。”

—— 导演兼编剧 Evan Halleck

瞬息全宇宙



2024年10月23日中国上映

好莱坞“五大”首次在华拥抱 AI

首映日释出中国独家 AI 赋能新水墨动画，黑白水墨背景下，毒液丝滑变身

毒液：最后一舞



Our T2 Remake



Trailer: Genesis



Air Head



玩具反斗城的起源



传说





视频大模型渗透率提升的驱动力

企业 / 行业对 **降本增效** 的诉求

娱乐工具变成生产力工具的关键

AI 成本 < 人力成本

且还在迅速下降

AI 生成效果 ≥ 人力制作

视频生成技术持续进步

视频生成技术逐步渗透 加速新文化浪潮到来



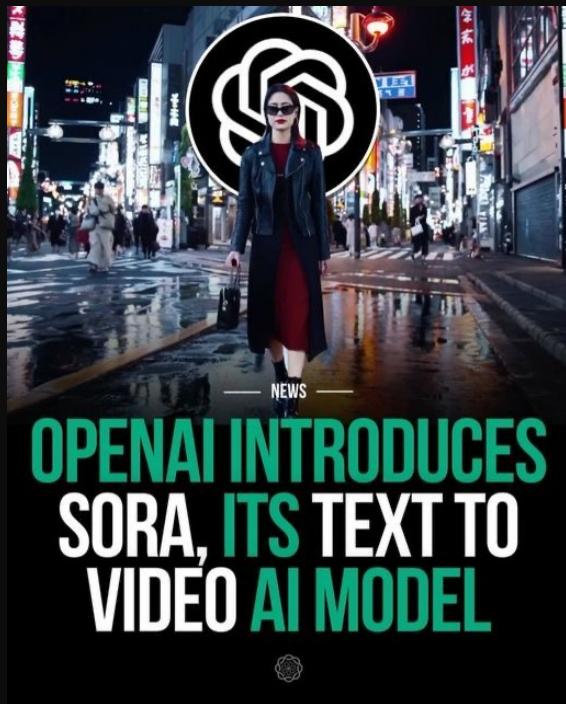


Part 3 视频大模型的创新与实践

视频大模型带来影视行业变革新动力

2024年2月 Sora发布

引发影视行业对AI技术的广泛关注和讨论



2024年4月国内首个全自研长视频模型 Vido发布

引发海内外广泛关注和报道

新闻联播

东方时空

微博热搜

海外媒体广泛报道

新闻联播

东方时空

微博热搜

海外媒体广泛报道



累计生成**过亿**视频，到目前为止

Vidu 是全球**增速最快**的 AI 视频模型工具



百万用户

20天

千万用户

100天

超90%

海外用户
占比

中国站 vidu.cn 国际站 vidu.com

2024 年 7 月全球上线，全球 200 多个国家和地区用户广泛使用



公司发展历程

生数科技成立

2023年3月

生数科技成立，致力于打造全球领先的多模态大模型及应用产品

使命：提升全人类的创造力和生产力



Vidu 产品历程

2024年4月

国内首发全栈自研、长时长、高一致性、高动态性的视频大模型 Vidu



2024年9月

Vidu 全球首发主体参照功能



2025年1月

Vidu 2.0 上线，推进视频生成“人人可用”



2024年7月

Vidu 全球上线，覆盖 200 多个国家和地区

2024年11月

Vidu 1.5 上线，全球首发多主体一致性功能

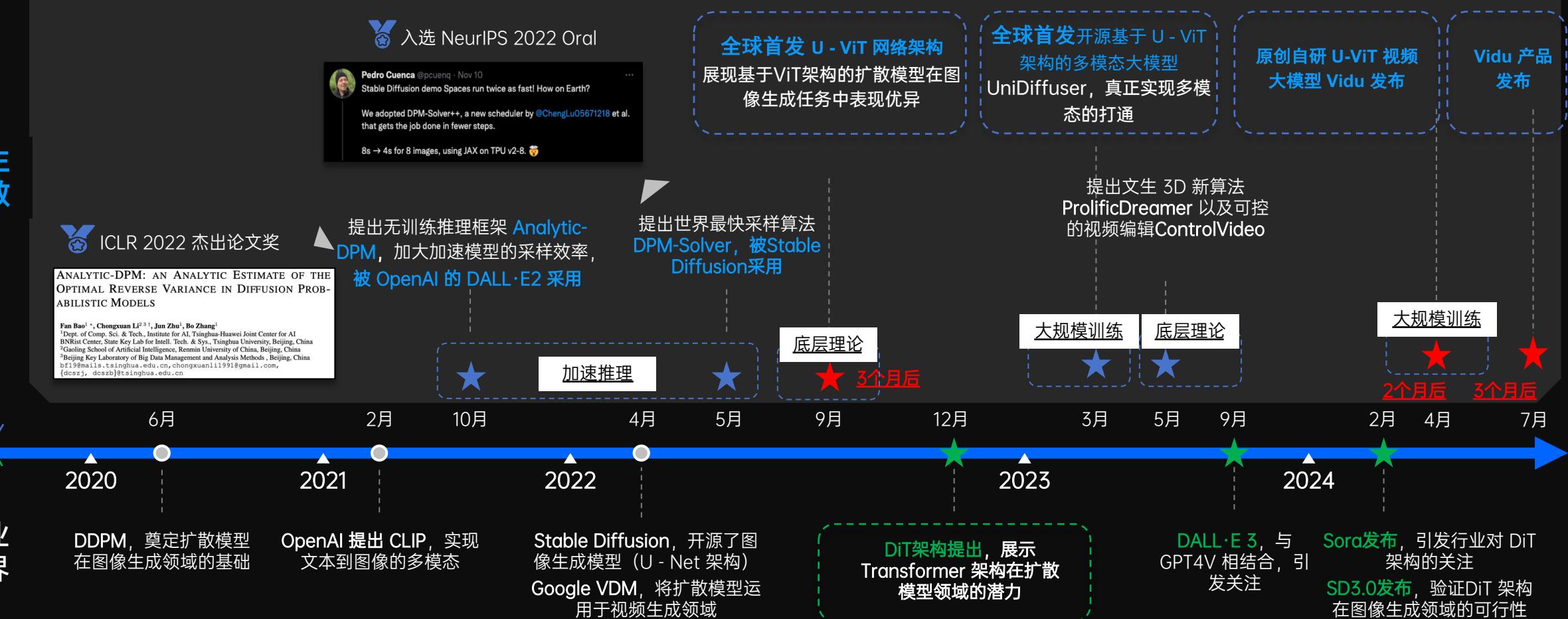


研究多模态生成最早、最懂的团队之一 在算法和工程领域展现超群技术能力

坚持生成式模型研究近 20 年，基于深厚积累敏锐识别扩散模型趋势，是国内**最早**开启该领域研究的团队
涉及底层理论、加速推理、大规模训练等**全栈领域**，相关顶会论文 30+ 篇

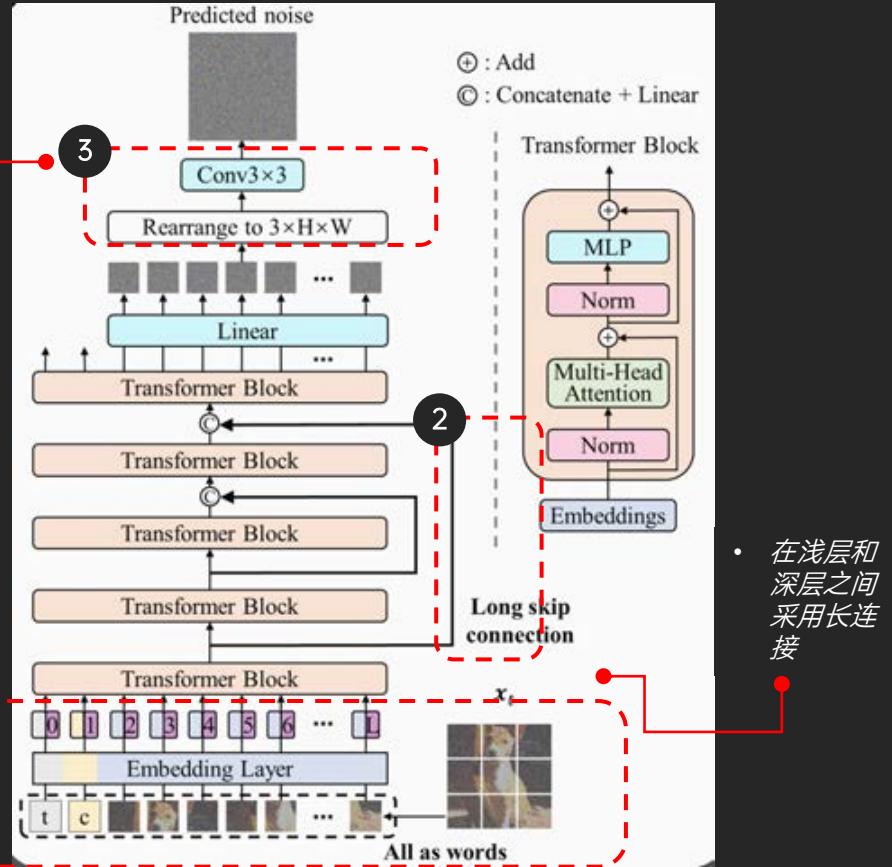
生数

业界



全球首发Diffusion Transformer架构U-ViT，并论证其具备极强的Scale Up能力

- 在输出噪声预测之前添加额外的 3×3 卷积块，按位置信息组合成一幅噪声图，以获得更好的视觉质量
- 将所有输入（时间、条件和噪声图像块）视为token



将所有输入统一为序列

把图片、时间、条件等所有输入都转化为 token，形成统一的序列

统一图文的输入和生成

可在文字和图像模态上实现任意输入与生成

更快的训练收敛速度

使用“长连接”技术，训练速度比 Stable Diffusion 提升 **7倍**以上

极强的涌现能力

在大规模参数量级上验证其具备极强的 Scale Up 能力



打造并开源全球首个基于 U-ViT 的通用多模态模型 UniDiffuser，领先 SD 1年！

首次将 ViT 架构成功应用于大规模训练，可对标依旧采用传统 U-Net 架构的世界领先模型 SD，性能 SOTA



[One Transformer Fits All Distributions in Multi-modal Diffusion at Scale, ICML 2023]

首先在图像领域成功Scale

图像生成语义理解强、美学性突出

率先成功将首个 Diffusion Transformer 模型 Scale 至大参数，在图像生成任务上表现出卓越效果

2023年

3月

Q2

Q3

Q4

Q1

2024年

率先开源
1B参数模型

数倍参数 Scale
美学性极强

Scale至更大量级
多元风格拓展

持续 Scaling
增强语义理解

持续迭代

Prompt: 一扇窗户前的红色桌子上放着一个茶杯，茶杯子外面写着字母"A"，还有插满鲜花的花瓶，安迪·沃霍尔，波普艺术风格



语义理解增强，文字生成精准



美学性提升，艺术风格突出

Prompt: Portrait of an old man sitting in a cafe, photographed through the window



- 精准语义理解
- 艺术级美学水准
- 支持多元风格
- 改善人脸、人手等难题
- 支持中国元素生成
- ...

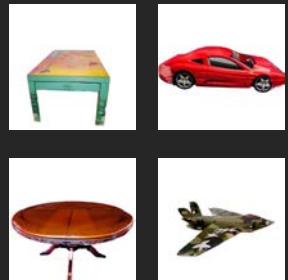
空间维度拓展至 3D资产 / 4D 动画生成能力， 具备极致分辨率

文生 3D 模型 Prolific Dreamer
极致分辨率、细致纹理、逼真程度



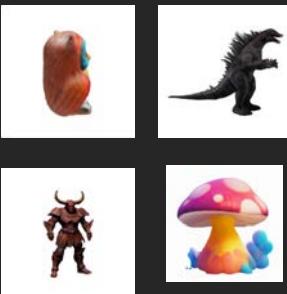
3D 资产生成

文本生成 3D
几何规整、UV Map 连续性好

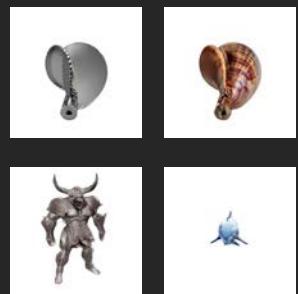


分钟级模型生成，最快秒级生成

图像生成 3D
可控、通用性强，多视角一致

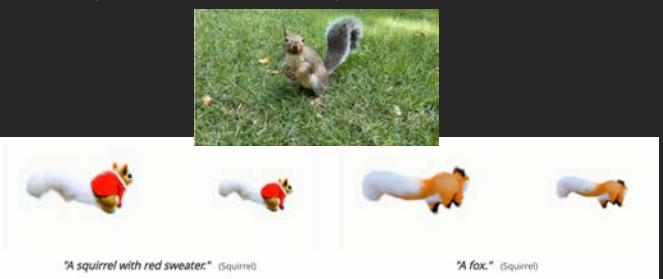


文本生成贴图
色彩还原、质感真实



4D 动画生成

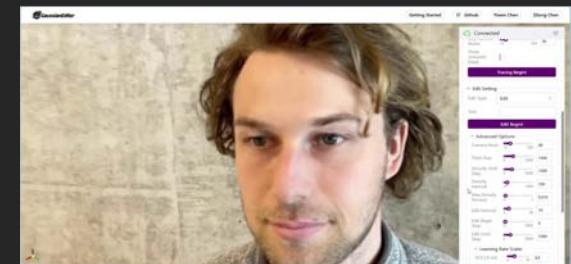
输入一段视频加 text prompt，输出逐帧的运动 3D
动画，自动绑定骨骼动作，支持 360° 全景视角



国际首发 4D 动画生成

3D 场景编辑

通过文本对话或手动调参的方式，灵活编辑 3D 场景，
包括添加物体、删除物体等，可实时查看变化



率先突破长视频生成瓶颈，推出中国首个长时长、高一致性、高动态性视频大模型Vidu



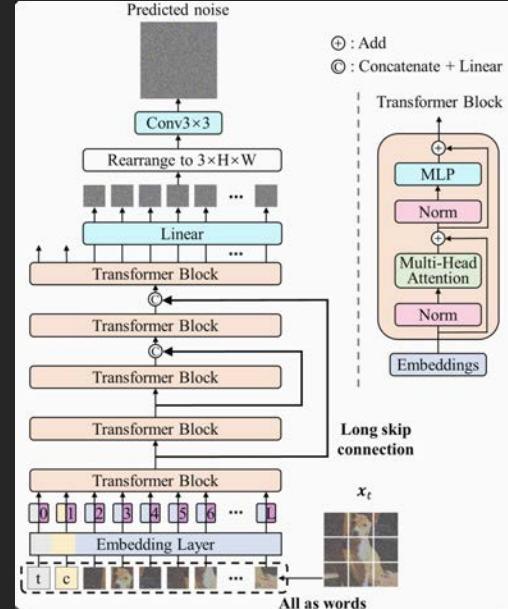
逐帧质量高

- 视频生成可以有效继承 UniDiffuser 在图像生成上的优势，画面美观度高，单帧质量好

时间一致性好

- 基于 Diffusion Transformer 思想设计视频生成模型，Patch 间强交互的架构有利于实现时间一致性，视频连贯度好

23年底具备短视频生成能力



将 U-ViT 架构运用于长视频生成模型

24年4月推出中国首个长视频生成模型并持续迭代



基于算法及工程的积累，叠加持续攻坚长视频数据收集与处理等难点，率先突破长视频生成瓶颈



实现多种效果 具备多维优势

极速推理



30 秒生成 4 秒片段

语义理解强



视频连贯 人物和场景在时空中保持一致

视频时长出色



一镜到底 32 s

单一大模型

端到端生成

动作幅度大，画面真实



主体运动幅度大 真实感强，充满视觉冲击力

画面表现力强



虚构场景能力 超现实主义画面

具备多镜头语言



复杂动态镜头 远/近/中景/特写
长镜头/追焦等效果

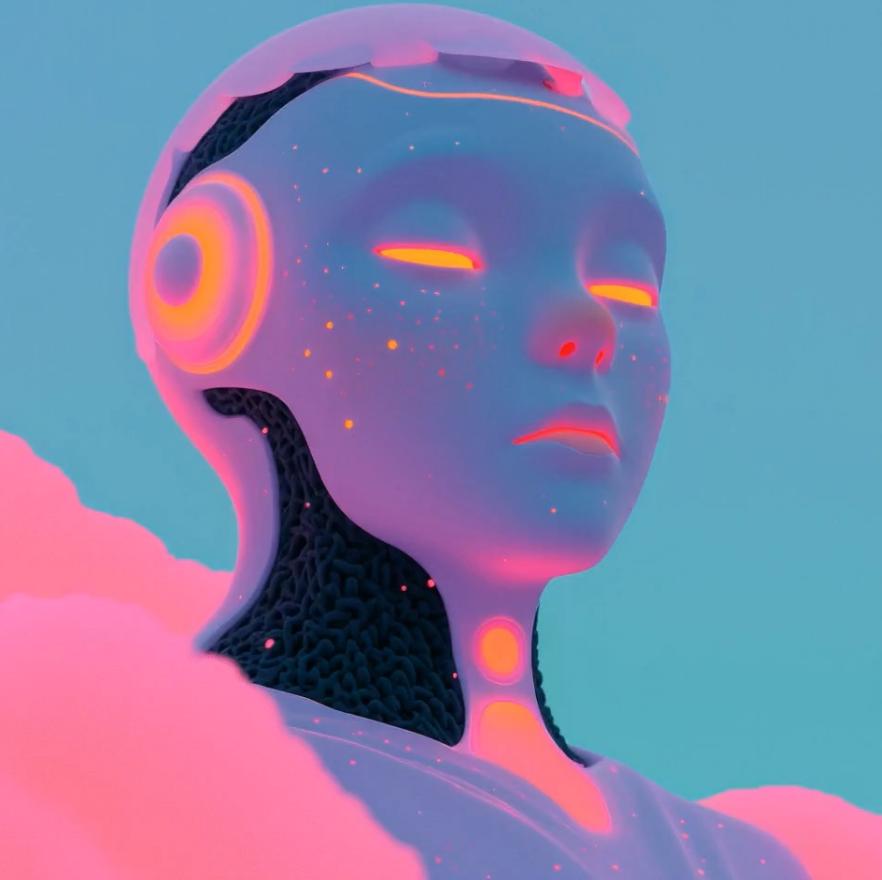
模拟真实物理世界



复杂、细节丰富的场景 光影效果、人物表情

多能力实现视频生成——动画风格优异

动作



氛围



情景



动物



人物



多能力实现视频生成——首尾帧

首帧



尾帧



首帧



尾帧



提示词：玫瑰变成女人，花瓣落下，人从花中出现



高一致性创新能力 —— 单主体一致

全球首个“参考生视频”功能，实现对“任意主体”的一致性生成，让视频生成具有高一致性

人物形象一致



- 输入形象 -



商品形象一致



- 输入形象 -



动物形象一致



- 输入形象 -



虚构形象一致



- 输入形象 -





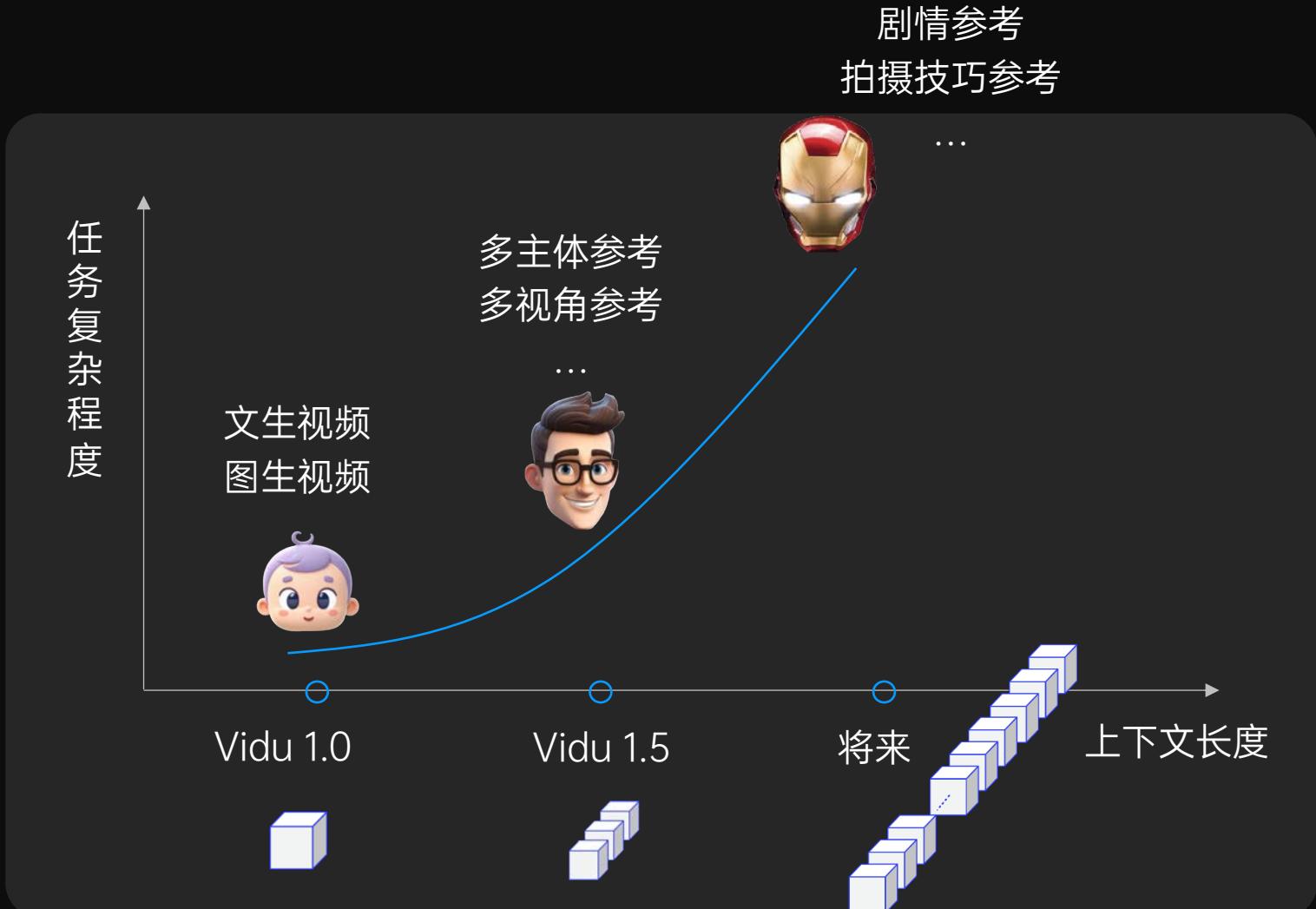
Vidu视频大模型迭代升级 开启视觉上下文时代

Vidu 1.5 将模型设计 向前迈进一大步

智能涌现

更高效

更一致



高一致性功能迭代 —— 多主体一致

由单图迭代至三图，通过上传多个主体及目标场景，即可实现让多主体在指定场景中互动



+



主体

主体

+



场景



提示词：男孩手里拿着蛋糕在水晶场景里

落地行业案例



动漫行业：细腻塑造动漫形象，差异化风格无限扩展

动漫风格生成：行业领先的动漫风格生成，万物皆可二次元



- 人物表情、动作 -



- 宫崎骏 -



- 大友克洋 -



- 武内直子 -



- 场景变换 -



- 音乐主题 -



- 80年代风格 -



- 人物说话 -

还有更多...

领先的动画表现力

扩展多种画家及导演风格

助力短剧、动画、动画短剧行业

传统3D动画
制作流程



传统2D动画
制作流程



基于Vidu流程

节省后期 90 % 的后期制作时间

相较传统流程提效 3 倍





已被众多创作者、艺术家使用，创作出绝妙作品

Vidu: 操作易上手、效果拔群，已被融入内容生产链条，释放无尽创造力

Vidu X AimateLab



一只北极熊历尽艰难险阻，跨越整个地球送来的礼物究竟是什么？

Vidu X 闲人一坤



这是一个闲人拯救地球的故事

Vidu X Huaranse



《聊斋》改编《耳中人》

Vidu X Game Over



AI古风短剧

“动画制作效率提升 **40** 倍，
生产周期从 **4** 个月缩短至 **1** 周”

广告行业应用：颠覆商业广告生产模式





效果广告 —— 高效制作动态海报，让产品宣传 更具吸引力

快消品动态海报

IP宣传动态海报

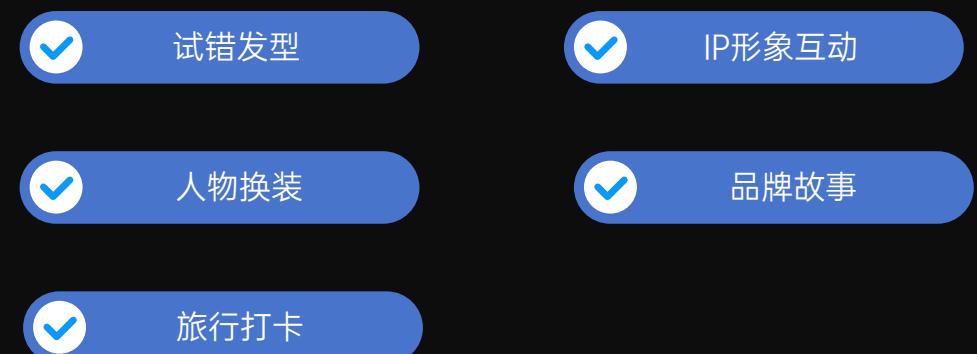


互动广告——多主体参照物的出圈营销

与现象级爆款作品的互动



与特定商品的日常互动



文旅行业：城市旅游推介的特效玩法， 打造不一样的视觉感受



大美安康



春来桃源



多场景画面自然融合，一键生成高质量、
多题材的地方推广视频



关键照片输入

动态视频输出

历史文物与搞笑艺术融合，搭建呈现沉浸式文博与文化场景



历史文物与搞笑艺术融合，搭建呈现沉浸式文博与文化场景



历史文物与搞笑艺术融合，搭建呈现沉浸式文博与文化场景



关键照片输入



动态视频输出

Part 4 视频大模型 2025 趋势预测

我们的预判

2025年是视频生成内容爆发元年



AI 动画

AI 音乐MV

AI 宣传片

趋势一

视频生成技术继续高速发展，达到甚至超过MJ V5时刻



图生视频

趋势一

视频生成技术继续高速发展，达到甚至超过MJ V5时刻



参考道具和人物生成



左图基础上增加了服装元素的参考

参考生视频

趋势一

视频生成技术继续高速发展，达到甚至超过MJ V5时刻



直出带音频的视频

趋势二

专业人群规模化涌入，出现消费价值高的爆款破圈内容



美克美家



@柔树

趋势三

使用成本大幅降低、体验大幅上升，人人可用、易用



角→分

Vidu 2.0 每秒单价成本降至
最低 4 分钱



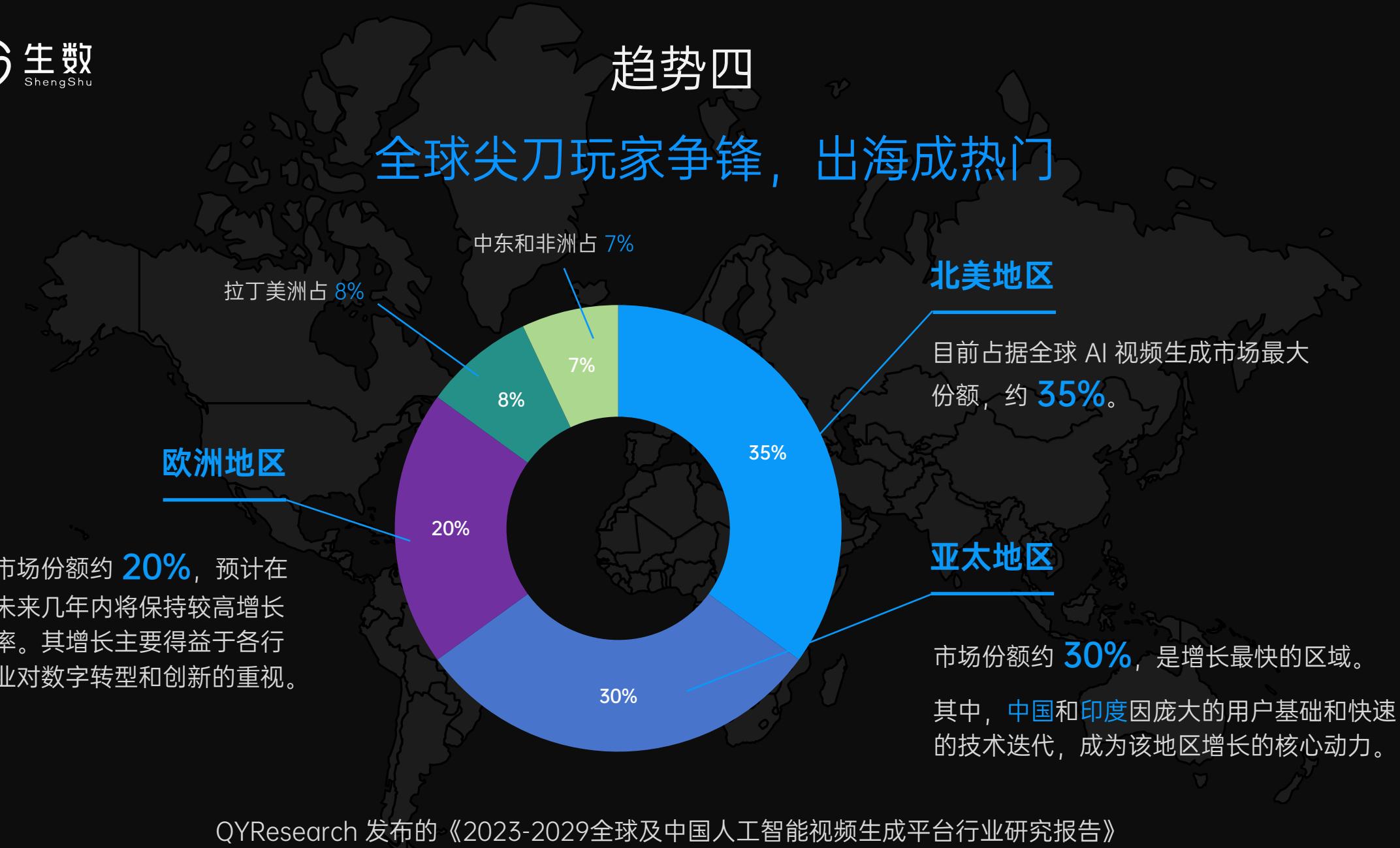
模板、灵感、一键生成



大幅缩短视频创作流程

趋势四

全球尖刀玩家争锋，出海成热门



A background image showing a group of diverse individuals, possibly models or performers, with their faces and bodies illuminated by vibrant, multi-colored lights in shades of red, orange, yellow, green, blue, and purple. They are looking directly at the camera with serious expressions. The lighting creates a dramatic, high-contrast effect.

以Vidu为例，在首个千万用户中
海外用户占比超过 90 %



当 Vidiu 上线，不只中国用户使用
很快在海外200多个国家和地区风靡



视频大模型出海的一点心得

战略

想清楚为什么出海

合规

从第一天就思考

组织

要有熟悉对应
地区的人才

多与对应地区的
用户沟通
实地拜访

产品

没有全球化
只有本地化

市场运营

传播裂变秘诀:
AI 视频工具
+
内容运营
+
全球社媒

看得更远一点

实时互动式内容体验



交互式电影游戏《夜班 (Late Shift)》

数字世界 → 物理世界



RDT-1B:
Robotics Diffusion
Transformer
Tsinghua University
October 2024



1. 多模态大模型的技术发展
2. 多模态大模型的行业背景
3. 视频大模型的创新与实践
4. 视频大模型 2025 趋势预测





One more thing...



一点个人的建议

平和

而不是焦虑

上手

而不是谈谈

热爱+AI

而不是 AI + 热门



谢谢

答疑环节