一、数字人技术

数字人是指以数字形式存在于数字空间中，具有拟人或真人的外貌、行为和特点的虚拟人物，也称之为虚拟形象、数字虚拟人、虚拟数字人等。数字人的核心技术主要包括计算机图形学、动作捕捉、图像渲染、AI 等。数字人可以打造更完美的人设，为品牌带来正向价值。互联网、金融、电商平台、消费品牌、汽车出行等领域纷纷推出数字人，用于品牌营销、智能客服等方向。 数字人可以按照不同维度进行分类：

- 根据人物图形资源的维度，数字人可分为 2D 和 3D 两大类，从外形上又可分为 2D 真人、2D 卡通、3D 卡通、3D 风格化、3D 写实、3D 超写实、3D 高保真等多种。
- 根据驱动的维度，可分为真人驱动和 AI 驱动两种。
- 根据商业和功能维度，可分为内容/IP 型、功能服务型和虚拟分身等三种

数字人在拟人化程度、自动化水平、应用场景三方面的表现水平，将数据责任分为 L1~L5 五个等级

L1 级：数字人形象写实，以 CG 人工建模为主，主要用于传统动画制作以及平面展示，应用场景非常局限。

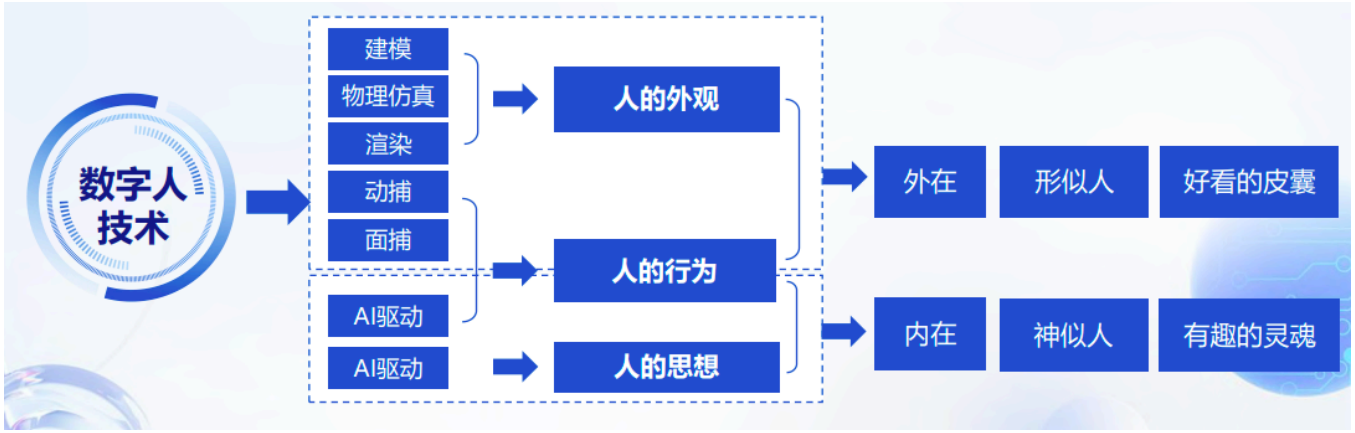
L2 级：数字人实现了写实形象的生动展现和动作表情交互，但主要依赖外部动作捕捉，实现口型、表情、肢体动作的驱动和表达，应用场景有锁拓展，但依然局限于传统的视屏录播和影像制作上。

L3 级：在 L2 写实在形象和动作表情交互的基础上，能够通过大数据和算法来驱动数字人完成口型、表情、肢体动作的驱动和表达，应用场景开始往部分实时驱动交互的动态场景延伸。

L4 级：在这个阶段的数字人以写实形象和动作表情实时生成及驱动为核心，也拥有了一定程度的理解智能能力，但依然“真假可辨”，主要以被动感知和人工指令输入驱动为主，主要用于垂直领域，比如在规范化的客服或者虚拟人直播领域，能够替代人工完成一些程序性工作

L5 级：完美形态下的数字人，既拥有“好看的皮囊”，形象精美高度写实，表情动作驱动流畅自然，还拥有了“有趣的灵魂”，能够完全理解用户意图并主动表达，不断适应环境变化，做到主动感知及驱动，完成自我学习成长。

技术发展是数字人第一推动力



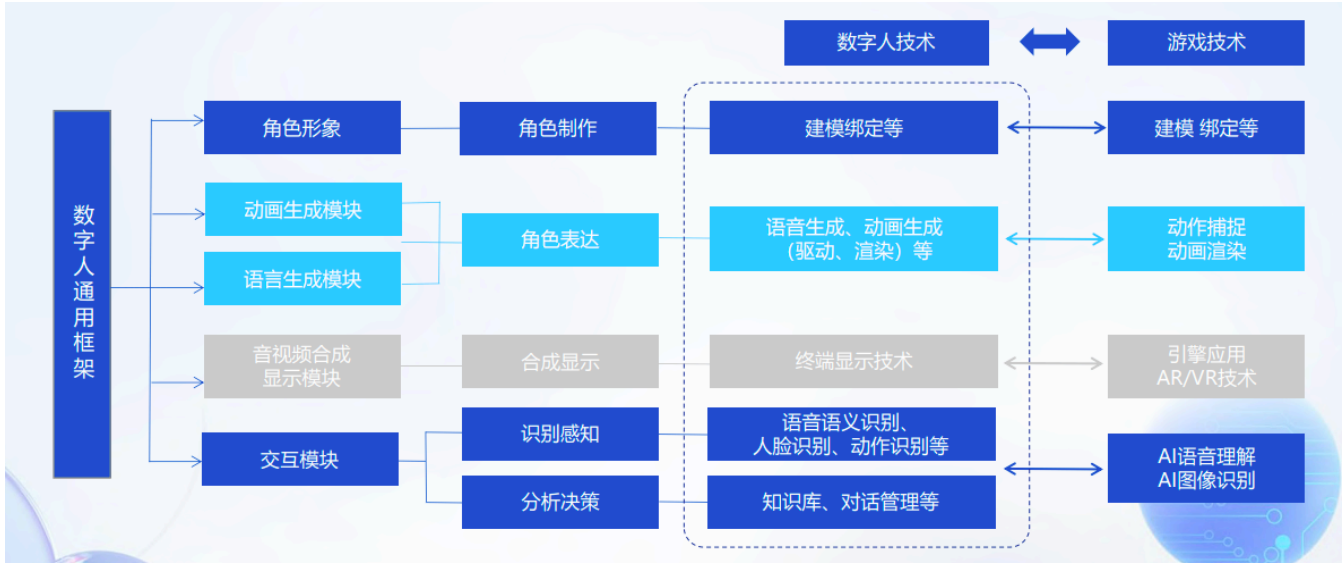
数字人具有以下三方面特征,分别是由建模、物理仿真、渲染、动捕、面捕和AI等技术支持,各项技术不断迭代,推动数字人制作效能和智能水平提升，其中通过AI技术实现高度拟人化的“思想和行为”,进而给用户带来亲切感、参与感、互动感和沉浸感是未来发展的重要方向。

1. 人的外观,具有人的相貌、性别和性格等人物特征。

2. 人的行为,具有语言、面部表情和肢体动作的能力。

3. 人的思想,具有识别外界环境,并与人交流互动的能力。

游戏科技不断赋能数字人制作



二、数字人技术

数字人制造和运营服务的 B 端市场不断扩大，将面向更广大的 C 端用户提供服务，各类数字人价值定位和商业模式有差异

1. 三种数字人（内容/IP 型、功能服务型和虚拟分身型）在产品定位、应用行业、核心价值、竞争力等方面存在显著差异。
2. 其中，内容/IP 型主要应用于影视、文娱和市场营销等场景，功能服务型主要应用于行业服务场景。这两种类型数字人制作方式以 PGC 为主，从数字人制作厂商角度，更多是面向 B 端。
3. 此外，虚拟分身类型数字人（Avatar）一般为 C 端用户制作虚拟形象，应用于 C 端用户在虚拟空间中的形象分身和代理。

产品分类	内容/IP型	功能服务型	虚拟分身（Avatar）
制作方式	PGC为主	PGC为主	UGC为主
产品定位	偶像，用于娱乐、科教等	数字员工，提供拟人化服务	在虚拟空间中的身份代理
代表应用	虚拟偶像、数字航天员、品牌代言人	虚拟主播、数字客服、数字专家	游戏身份如VRChat、虚拟演唱会观众
应用行业	电影、电视、综艺、科普纪录片等	金融、文旅、零售、直播	游戏、VR应用
核心价值	虚拟IP/偶像身份的具象化	人工作的增强和辅助，降低企业成本，提供自动化、智能化和标准化服务	虚拟空间中的交互入口，推动虚拟内容生产
核心竞争力	形象的艺术性、IP的打造和运营能力	智能交互能力	沉浸化、实时化、体验感
驱动方式	中之人驱动	AI驱动/中之人驱动	中之人驱动/指令驱动

数字人技术一般情况情况下由人物形象、语音生成、动画生成、音视频合成显示、交互等 5 个模块构成

1. 人物形象：人物形象根据人物图形资源的维度，可分为 2D 和 3D 两大类，从外形上又可分为卡通、拟人、写实、超写实等风格
2. 语音生成、动画生成：语音生成模块和动画生成模块可分别基于文本生成对应的人物语音，以及与之相匹配的人物动画
3. 音视频合成显示：音视频合成显示模块将语音和动画合成视频，再显示给用户
4. 交互模块：交互模块使得数字人具备交互功能，即通过语音识别等智能技术识别用户的意图，并根据用户当前意图决定数字人后续的语音和动作，驱动人物下一轮的交互

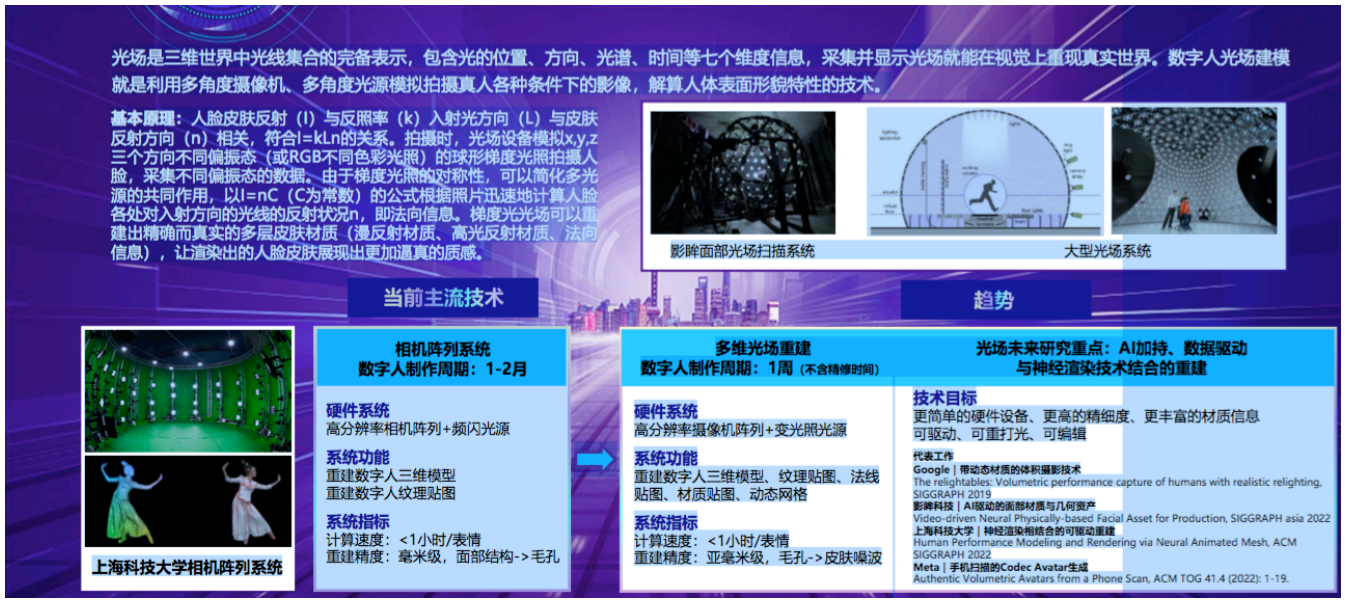
AI 技术驱动数字人多模态交互更神似人，并逐步覆盖数字人全流程当前数字人对语言理解还是以文本为主，动作合成上声唇同步较为完善：

1. AI 驱动数字人是指数字人等语音表达、面部表情和动作形态等通过深度学习模型进行运算，并将其结果实时或者离线驱动，并进行渲染。目前主流的方式是围绕 NLP 能力通过文本驱动，本质是通过 ASR-NLP-TTS 等 AI 技术进行感知-决策-表达的闭环来驱动数字人交互，同时需要预先设置相关的知识图谱或问答库等，与数字人的对话系统对接，但目前 NLP 在通用性场景的能力还需要进一步完善。
2. 计算机视觉(CV)目前数字人声唇同步技术相对完善，在游戏中已经大量应用；而其他表情和动作还需要描述性的数据或者标签驱动，尚未智能合成，表情动作也是是 AI 驱动未来发展的重点方向。

	技术	阶段	作用和目的	发展趋势
语音理解	ASR	感知阶段	将人的语音转换为文本	相对成熟
	NLP	决策阶段	处理并理解文本，以对话能力为核心，为数字人的大脑	配合知识图谱，应用于特定场景，未来通用型模型还需要完善
	TTS	表达阶段	将需要输出的文本合成为语音	相对成熟，未来方向增加断句、多音字的准确度，增加情感，更加似人
动作合成	AI驱动嘴形动作	表达阶段	建立输入文本到输出音频与输出视觉信息的关联映射，主要是对采集到的文本到语音和嘴形视频（2D）/嘴形动画（3D）的数据进行模型训练，得到相关模型，并智能合成	随着写实度的提高，微表情更多，超写实对精度要求更高，超写实还需要进一步完善
	AI驱动其他动作	表达阶段	动作是采用随机策略或者脚本进行预设，需要人工配制描述性的数据或者标签	尚未实现智能合成

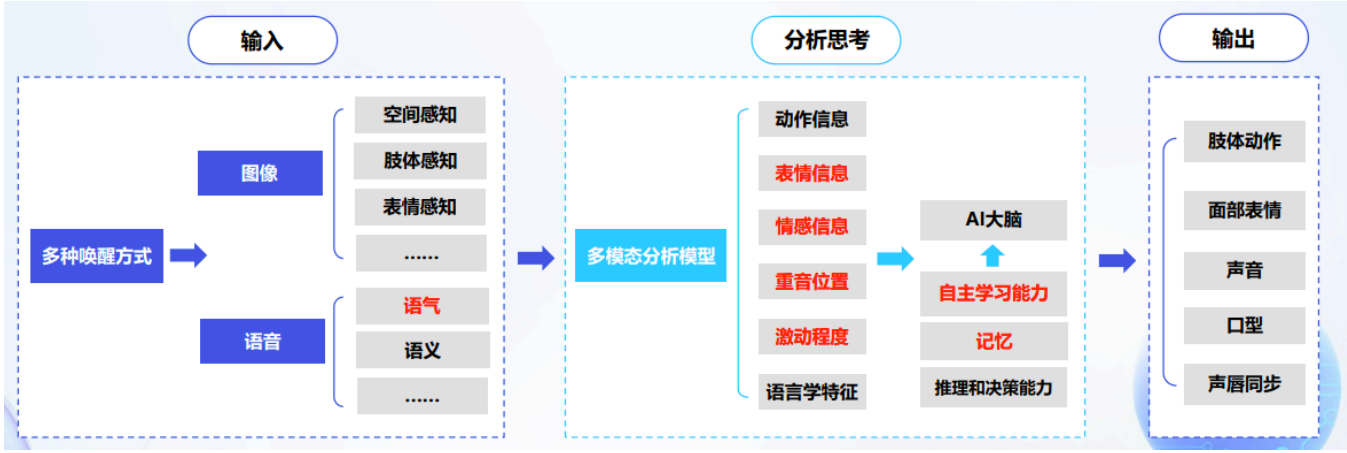
光场建模维度提升带来影视级数字人制作周期大幅缩减,AI 技术是未来发展重点

光场是三维世界中光线集合的完备表示,包含光的位置、方向、光谱、时间等七个维度信息,采集并显示光场就能在视觉上重现真实世界。数字人光场建模就是利用多角度摄像机、多角度光源模拟拍摄真人各种条件下的影像,解算人体表面形貌特性的技术。



多模态 AI 技术是未来数字人发展的最大推动力，驱动数字人“思想”更像人：

未来 AI 技术的重点方向是在输入端实现多模态感知输入，在输出端提升多模态交互能力，综合提升数字人的表现力，从目前的基于文本的交互，转化为基于语义的交互，特别是需要强化对人情绪的感知和表达



一个简单的案例，在百度飞桨平台如何生成一个数字人

生成虚拟数字人总共需要调用三个模型，分别是 First Order Motion（表情迁移）、Text to Speech（文本转语音）和 Wav2Lip（唇形合成）。

1.把图像放入 First Order Motion 模型实现面部表情迁移，让虚拟主播的表情更加逼近真人。

- **表情迁移**

通过FOM模型，输入图像和驱动视频，让人像动起来。

In [ ]

```
import paddlehub as hub

FOM_Module = hub.Module(name="first_order_motion")
FOM_Module.generate(source_image="input_data/test.jpg", # 输入图像
                    driving_video="input_data/zimeng.mp4", # 输入驱动视频
                    ratio=0.4,
                    image_size=256,
                    output_dir='./output/', # 输出文件夹
                    filename='FOM.mp4', # 输出文件名
                    use_gpu=True)
```

2.输入你想让数字人说的话，通过 Text to Speech 模型，将输入的文字转换成音频输出。

In [ ]
```
sentences = ['开发者你好，欢迎使用飞桨，我是你的专属虚拟人。'] # 输入说话内容

TTS_Module = hub.Module(
    name='fastspeech2_baker',
    version='1.0.0')
wav_files =  TTS_Module.generate(sentences)
print(f'声音已生成，音频文件输出在{wav_files}')
```

3.得到面部表情迁移的视频和音频之后，将音频文件和动态视频输入到 Wav2Lip 模型，并根据音频内容调整唇形，让唇形根据说话的内容动态改变，使得虚拟人更加接近真人效果。

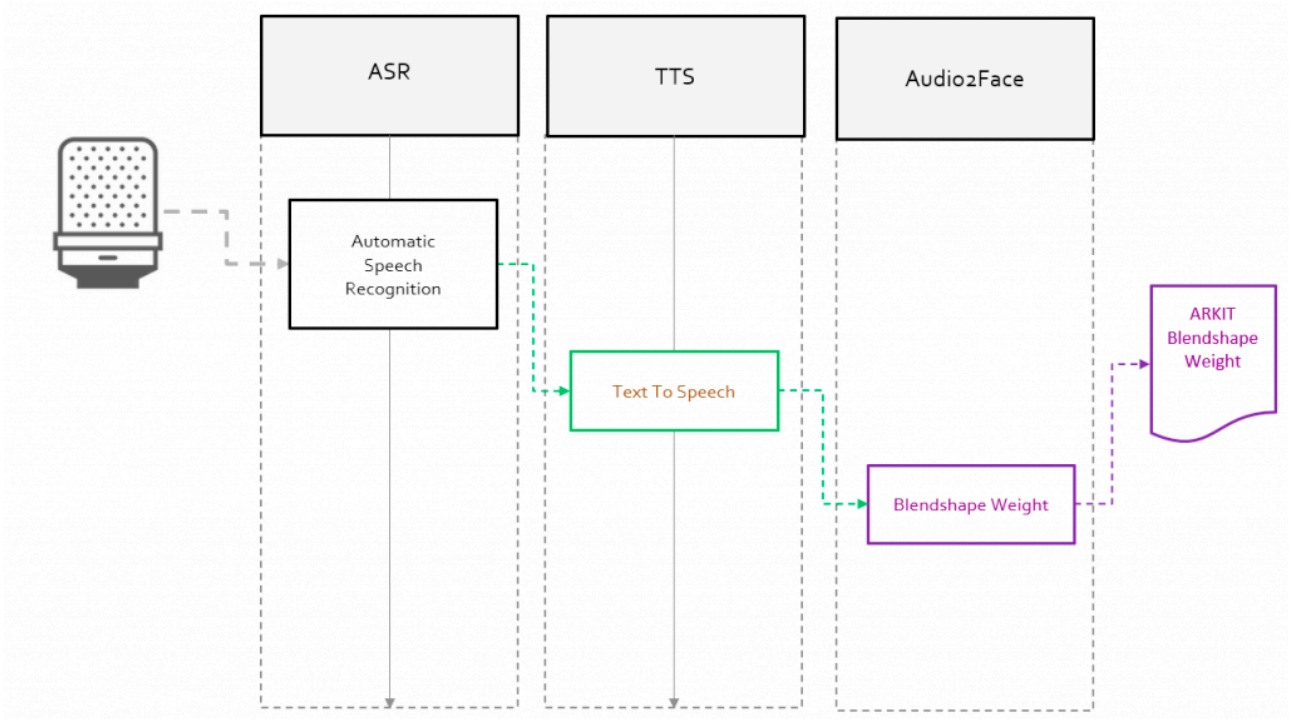
In [ ]
```
W2F_Module = hub.Module(name="wav2lip")

W2F_Module.wav2lip_transfer(face='output/FOM.mp4',
                            audio='wavs/1.wav',
                            output_dir='./transfer_result/',
                            use_gpu=True)
```

这样，就得到了一个简单的数字人模型


飞桨
AI虚拟主播

三、播报数字人人技术

数字人合成主流的三个技术



开源产品介绍

FACEGOOD：

主要完成 Voice2Face 部分，ASR，TTS 由思必驰智能机器人完成。如果你想用自己的声音，或者第三方的 ASR，TTS 可以自行进行替换。当然 FACEGOOD Audio2Face 部分也可根据自己的喜好进行重新训练，比如你想用自己的声音或其它类型的声音，或者不同于 FACEGOOD 使用的模型绑定作为驱动数据，都可以根据下面提到的流程完成自己专属的动画驱动算法模型训练

VideoReTalking：

让视频中的人物的嘴型与输入的声音同步。你只需要输入任意一个视频和一个音频文件，它能给你生成一个新的视频，在这个视频里，人物的嘴型会与音频同步。VideoReTalking 不仅可以让嘴型与声音同步，还可以根据声音改变视频中人物的表情。整个过程不需要用户干预，都是自动完成的。工作流程：整个系统的工作流程分为三个主要步骤：面部视频生成、音频驱动的嘴型同步和面部增强。所有这些步骤都是基于学习的方法，并且可以在一个顺序的流程中完成，无需用户干预。1、面部视频生成：首先，系统会使用表情编辑网络来修改每一帧的表情，使其与一个标准表情模板相符，从而生成一个具有标准表情的视频。2、音频驱动的嘴型同步：然后，这个视频和给定的音频一起被输入到嘴型同步网络中，生成一个嘴型与音频同步的视频。3、面部增强：最后，系统通过身份感知的面部增强网络和后处理来提高合成面部的照片真实性。

项目及演示：opentalker.github.io/video-retalking/

so-vits 声音克隆技术

1. So-VITS-SVC 是由 B 站 UP 主共同开发的一个开源项目，其目标是实现端到端的人声克隆，只要输入一段人声音频和一段歌词，就可以生成相同或者相似的人声唱出歌词的音频。该项目基于 VITS 模型，但做了一些改进和优化，例如增加了 SVC（speaker Verification Classifier）模块来提升音色的相似度。以及使用了更高采样率来提升音质。该项目目前已经发布了 4.0 版本，并提供了多种语言(中文，日文，英文等)和多种音色（碧蓝档案，初音未来，洛天依等）的预训练模型提供用户下载和使用

2. So-VITS-SVC 基于 Soft 编码的 VITS 声音转换模型，通过 Hubert 的 Soft 编码输入来替换 VITS 中的 ppg 从而实现的声学转换模型，而 VITS 是一种基于对抗网络训练的声音合成模型，就是效果比传统的 VITS 模型更好的声音转换模型

微软 GAIA

GAIA 能够从语音和单张肖像图片合成自然的会说话的头像视频。也就是只需要你的一张照片就能让它开口说话。它甚至支持诸如“悲伤”、“张开嘴”或“惊讶”等文本提示，来指导视频生成。
GAIA 还允许你精确控制虚拟人物的每个面部动作，比如微笑或惊讶的表情。可以接受语音、视频或文字指令创建会说话的人物头像视频。

GAIA 揭示了两个关键洞见：

1. 用语音来驱动虚拟人物运动，而虚拟人物的背景和外貌（appearance）在整个视频中保持不变。受此启发，本文将每一帧的运动和外貌分开，其中外貌在帧之间共享，而运动对每一帧都是唯一的。为了根据语音预测运动，本文将运动序列编码为运动潜在序列，并使用以输入语音为条件的扩散模型来预测潜在序列；

2. 当一个人在说出给定的内容时，表情和头部姿态存在巨大的多样性，这需要一个大规模和多样化的数据集。因此，该研究收集了一个高质量的能说话的虚拟人物数据集，该数据集由 16K 个不同年龄、性别、皮肤类型和说话风格的独特说话者组成，使生成结果自然且多样化。

项目地址：https://microsoft.github.io/GAIA/

四、数字人产业应用现状

《虚拟数字人深度产业报告》预计，到2030年我国虚拟数字人整体市场规模将达到2700亿元，其中，“服务型虚拟人”总规模也将超过950亿元。
如同秃鹰盯上腐肉,嗅到万亿商机的各方势力,都欲分一杯羹,这也直接导致了目前的虚拟人玩家格局陷入了“混战”状态。
「自象限」根据各方数据不完全统计，目前国内虚拟数字人核心厂商约有6000家。而按天眼查的数据显示，相关厂商数量甚至超过6万家。
同时，随着大模型(Large Model)的兴起，虚拟人的产业格局也在发生深刻变化。

千亿市场，厂商“混战”

如果说元宇宙时期的虚拟人已经是一把大火，那大模型就相当于在这之上又烹上了一勺油。一瞬间，铺天盖地的数字人厂商涌来，将本就复杂的行业搅得愈发浑浊。这其中，既包括从元宇宙时期就一直坚持虚拟数字人的厂商，也有依靠全栈技术优势轻松迈出第一步的大厂，更不乏闻风而来的换道厂商。
「自象限」初步了统计核心厂商的类型后发现，这些厂商大致可以分为四类：

市场技术能力分析

1. **百度虚拟人：**百度旗下的百度智能云定位于全场景、大生态的数字人场景，目前虚拟人大概有二十余位，在实际应用当中，主要落地于智能陪伴、综艺、代言人等，比如知名明星“龚俊数字人”，无论是商业价值排名，还是企业排名，百度的虚拟人各方面都是“带头大哥”般的存在，但是相应的在使用价格方面也是有较高的成本。

2. **科大讯飞：**科大讯飞是一家知名的智能语音和人工智能上市企业，在亚太地区拥有广泛的影响力。他们的虚拟人产品以其高度准确的语音识别能力和快速响应的特点而闻名，其虚拟数字人也具有较高的逼真程度。然而，科大讯飞的虚拟人

产品需要倚赖强大的算法和计算资源，这可能导致在使用时对设备资源的占用较多。因此，对于部分低配设备或网络环境较差的用户而言，可能难以获得良好的用户体验。

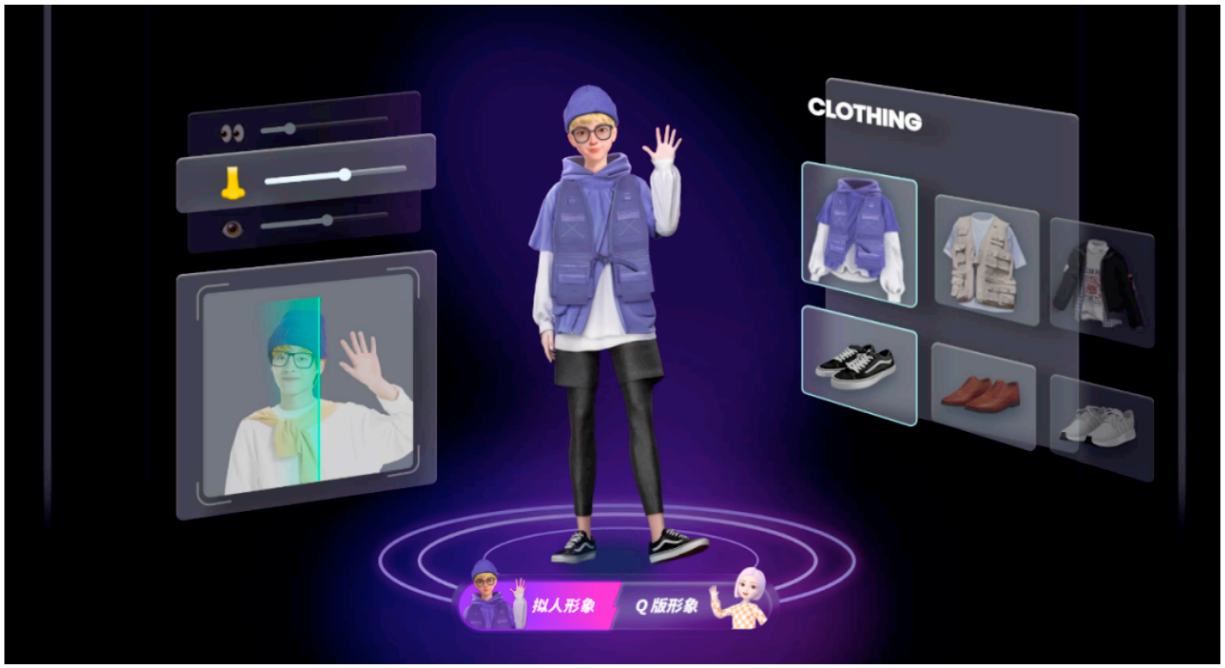
3. **小冰：**小冰在智能化和人性化方面表现出色，与用户之间可以进行有趣而自然的互动，用户体验非常出色。小冰也拥有许多成功的数字人案例，比如万科员工"崔筱盼"作为一名"数字员工"获得了年度万科新人奖的荣誉。然而，与其他虚拟数字人相同，小冰也存在语音识别准确性、资源占用等问题需。

参考项	百度	科大讯飞	小冰	sytheia	魔珐科技	世优科技	相芯科技	硅基智能	风平智能
技术实力	强	强	强	强	强	强	一般	强	强
3D虚拟人	强	强	一般	一般	强	强	一般	一般	一般
数字真人	一般	强	强	强	一般	一般	一般	强	强
逼真程度	强	强	强	强	一般	一般	一般	强	强
建模难易程度	复杂	复杂	复杂	复杂	复杂	复杂	复杂	简单	简单
AI视频	✓	✓	✓	✓	✓	✓	✓	✓	✓
AI直播生产	✓	✓	✓	X	✓	✓	✓	✓	✓
AI创作	✓	X	✓	✓	X	X	X	X	✓
AI图片	✓	✓	✓	✓	X	X	X	✓	✓
AI一键成片	✓	X	✓	✓	X	X	X	X	✓
素材库	强	强	强	强	X	X	X	X	较强
多语言能力	✓	✓	✓	✓	✓	X	X	✓	✓
主要应用场景	文娱、传媒	智能客服、智慧交通	数字员工、虚拟专家	企业传播、数字视频营销和广告本地化	代言、影视	演艺、文娱	虚拟化身、虚拟助手	医疗、教育、金融、本地生活、直播带货、短视频	医疗、教育、金融、本地生活、直播带货、短视频
直播互动性	高	高	高	X	高	X	X	高	高
稳定性	高	高	高	X	X	高	X	高	高
多平台分发体系	X	✓	✓	✓	X	X	X	X	✓
定制能力	✓	✓	✓	✓	✓	✓	✓	✓	✓
隐私保护	✓	✓	✓	✓	✓	✓	✓	X	✓
品牌知名度	强	强	强	一般	强	强	强	强	弱
价格	高	高	高	高	高	高	高	高	低

今年8月、9月开始，虚拟人厂商产品开始加速迭代。据不完全统计，两个月内至少有10家厂商发布了新的虚拟人产品。产品的高度迭代意味着虚拟人正在飞快得适应市场需求，而这也意味着虚拟人第一阶段的赛点已经走入关键阶段。

1. 从类型来看，虚拟人厂商分为两类，一类直接交付虚拟人产品，包括通用虚拟人产品、行业垂直场景的虚拟人产品，比如电商、零售、营销、直播等，客户即拿即用，或标准化或定制化；另一类则提供虚拟人制作平台，客户通过使用平台提供的工具，自主生产虚拟人。
2. 相比之下，产品交付类型更适合企业探索虚拟人初期，几乎不需要技术团队配合，门槛更低，也是目前较多企业选择的方式。
3. 针对这类产品形式，虚拟人厂商也提供了多样的购买方案。如汽车试驾一样，品牌在购买虚拟人之前，可以先可进行 Demo 的试用，真实感受虚拟人的表情、动作、交互等等。除此之外，品牌在购买前还可以进行方案咨询，厂商会根据客户情况，制定具体的虚拟人传播方案，并有多种不同风格的虚拟人可以选择

以即构虚拟人 Avatar 为例，企业可选择拟人形象和 Q 版形象，



大模型让虚拟人“长了脑子，有了思考和推理能力，AIGC 技术让虚拟人能够有想法，TTS 技术则让虚拟人能够表达”

几天前，在 GPT-4 版本更新，TTS 实现了进度，文本驱动语音有了语气和口吻，在停顿、重音和自然交互程度上有了极大的提升。不仅可以模仿不同的口吻，甚至设定“渣女”时还学会了“夹子音”。

市场热度不断，增量资金驱动行业超速发展

数字人浪潮将驱动大量行业出现创新变革，今后 3 年将呈现高速发展趋势

数字人产业发展十大趋势

价值定位 数字人制造和运营服务的B端市场不断扩大，将面向更广大的C端用户提供服务，各类数字人价值定位和商业模式有差异。

技术迭代 技术集综合迭代驱动数字人形似人，制作效能将继续提升

AI 赋能 AI技术驱动数字人多模态交互更神似人，并逐步覆盖数字人全流程

融合发展 数字人技术与SLAM、3D交互、体积视频、空间音频等技术深度融合，渲染将从本地到云端

行业应用 千行千面的数字人将成为人机交互新入口，但深度上仍需挖掘

C端模式 UGC数字人将加速出现，成为未来产业的增量空间

硬件载体 数字人仍以2D显示设备为主，3D显示设备成为特定领域的新解法

发展路径 在场是数字人发展的高级阶段，将与应用场景深度耦合

产业聚集 艺术和技术双轮驱动，北京有望成为产业新高地

合规前置 数字人版权保护及行业合规体系需同步建设，推动实现可用、可靠、可知、可控

数字人行业应用全景图

功能服务型 | IP/内容型 | 虚拟分身型

银行	资管	保险	影视	传媒	市场营销	文旅	政务	教育	家居	交通出行	互联网
服务一体机											
沉浸服务											
视频客服											
理财引导											
服务导航	内部培训			手语融媒体直播		游览导游助手					
远程开户授权	智能投教	代理人培训		新闻热点问答		线上VR展厅					
信审面签	智能开户	代理人营销助手		一体机大屏	明星虚拟分身代言	线下问询指引	政务微信助手				Avatar
信用卡开卡	线下投教展厅	视频回访		3D虚拟演播厅	虚拟偶像代言	线下会议主持	一网通办一体机	AI智能助教	家庭管家-物联		虚拟偶像
理财产品介绍	VR云展厅	信保面签		手语新闻制作	企业直播	线下展厅讲解	服务引导	课件生产	冰箱助手	接待一体机	游戏
资讯早报	基金播报	培训课件	电影动画	2D虚拟主播视频生产	数智导购	文化短视频	警务大屏	校警大屏	电视助手	车载助手	虚拟直播
金融			影视传媒和市场			公共服务			生活服务		互联网