引用格式：林驰琛，刘晨鸣，范伟健，李梦柯，王永滨. 基于AIGC技术的视听新变化[J]. 中国传媒大学学报（自然科学版），2024，31(02)：01-08.
文章编号：1673-4793(2024)02-0001-08

基于AIGC技术的视听新变化

林驰琛[1]，刘晨鸣[2]，范伟健[1]，李梦柯[1]，王永滨[1*]
（1. 中国传媒大学媒体融合与传播国家重点实验室，北京 100024；2. 国家广播电视总局广播电视科学研究院，北京 100866）

摘要：随着科技创新和基础设施的完善，人工智能生成内容（AIGC）技术正在成为视听产业的新质生产力，为未来的视听体验带来了无限可能性。本文从AIGC领域中的视频生成和音频处理技术出发，结合个性化视听体验的应用场景，探讨AIGC技术助力新视听产业发展的可行性，并总结了在当前应用环境下AIGC技术面临的挑战和不足。在人工智能科技革命和视听产业变革的浪潮中，AIGC不仅扮演着推动产业发展的关键角色，更以其前瞻性的创新引领着视听模式的变革与融合。

关键词：AIGC；新视听；新质生产力；人工智能
中图分类号：TP18　**文献标识码：**A

New AIGC-based audiovisual changes

LIN Chichen[1], LIU Chenming[2], FAN Weijian[1], LI Mengke[1], WANG Yongbin[1*]
（1. State Key Laboratory of Media Convergence and Communication, Communication University of China, Beijing 100024, China; 2. Academy of Broadcasting Science, National Radio and Television Administration, Beijing 100866, China）

Abstract: With technological innovation and infrastructure improvement, Artificial Intelligence Generated Content (AIGC) technology is becoming a new quality of productivity in the audio-visual industry, bringing infinite possibilities for future audiovisual experiences. Starting from the video generation and audio processing technologies in the field of AIGC, this paper discussed the feasibility of AIGC technology to support the development of the new audio-visual industry in the light of the application scenarios of personalised audiovisual experience, and summarised the challenges and shortcomings of AIGC technology in the current application environment. In the wave of artificial intelligence technology revolution and audio-visual industry transformation, AIGC not only plays a key role in promoting the development of the industry, but also leads the transformation and integration of audiovisual mode with its forward-looking innovation.

Keywords: AIGC; new audiovisual; new quality productive forces; artificial intelligence

---

基金项目：北京市社科基金重点项目（23JCB002）

作者简介(*为通讯作者)：林驰琛（1997-），男，博士研究生，主要从事人工智能研究。Email：cuc_lcc@cuc.edu.cn；刘晨鸣（1981-），男，博士，主要从事超高清和人工智能研究。Email：liuchenming@abs.ac.cn；范伟健（1992-），男，博士研究生，主要从事大众认知安全研究。Email：fanwj@cuc.edu.cn；李梦柯（2001-）：女，硕士研究生，主要从事新闻来源分析研究。Email：limengke@cuc.edu.cn；王永滨（1963-），男，博士，教授，博士生导师，主要从事网络新媒体技术研究。Email：ybwang@cuc.edu.cn

1 引言

近年来,随着算力、预训练模型和多模态技术的不断汇聚发展,人工智能生成内容(Artificial Intelligence Generated Content,AIGC)已经成为视听领域的一大亮点。在聊天机器人、AI作画、虚拟主持人和新闻写作等视听应用场景中,AIGC均展现出了其独特的优势和潜力,不断推动着相关领域的创新与突破。AIGC是继专业生成内容(Professional Generated Content,PGC)和用户生成内容(User Generated Content,UGC)之后,利用人工智能技术自动或辅助生成内容的新型生产方式[1]。AIGC凭借先进的神经网络模型和庞大的训练数据规模,使其能够接受和处理更加复杂的语音、文本、图像等多模态数据,自动或辅助生成文本、图像、音频等多种形式的内容。AIGC还能通过融合知识检索、逻辑推理等手段,实现从感知、理解到生成、创作的跃迁。其发展历程从早期基于规则和统计的简单生成方法,逐步演进至如今基于强大预训练模型的多模态生成技术,如OpenAI的ChatGPT[2]。

同时,新视听领域正处于迅猛发展的阶段。数字媒体和互联网的普及使用户对视听内容的需求不断增长,个性化和多样化逐渐成为主流趋势。新兴媒体平台,如Netflix、YouTube等,运用智能算法精准满足用户的个性化需求。元宇宙、虚拟现实等场景为视听领域注入了新的活力,推动了内容创作和消费方式的创新变革。面对新视听领域的挑战与机遇,传统媒体行业也在积极调整策略,加大对数字化、智能化技术的投入,以适应新时代的发展潮流。AIGC技术在视听领域具有广泛的应用前景。在视频内容制作方面,AIGC技术已实现了更高效的视频智能编辑和生成功能,例如自动生成视频剪辑、特效和虚拟场景等。尤其是最近Sora[3]的公布,进一步颠覆了人们对视频生成技术的认知。在音频处理方面,AIGC技术也展现出其独特的优势,能够提升音频内容创作的智能化和个性化水平,加快音乐制作的速度,助力创作更具创意的音乐作品。

不仅如此,AIGC技术还能为视听内容的个性化推荐和交互体验提供支持。该技术能够根据用户偏好和行为模式,智能推荐最符合用户需求的视听内容。同时,AIGC技术与游戏、虚拟现实(Virtual Reality,VR)等应用的融合,能够实现更加沉浸式的视听体验,提升用户的参与感和满意度。AIGC技术的发展不仅对视听创作和视听体验产生深远影响,并促使视听产业迎来内容生产和商业模式的重大变革。

尽管AIGC技术在处理和生成多模态内容方面取得了显著成就,但仍面临诸多亟待解决的问题。首先是模拟真实物理环境的难题。尽管AIGC模型可以生成逼真的场景,但在处理时序和因果关系方面仍存在挑战。此外,AIGC所依赖的大模型基座在边缘环境上受到计算资源和网络带宽的限制,使得其难以实现大规模边端部署。最后,AIGC技术与社会道德的对齐也是一个亟待解决的问题,因为其生成的内容可能会涉及到道德、隐私和版权等方面的问题,需要更多的研究和监管来确保其符合主流价值观和法律法规。综上所述,虽然AIGC技术在视听领域具有巨大的潜力,但需要继续努力克服这些挑战,才能实现其更广泛的应用和长足的发展。

2 新视听中的AIGC技术

2.1 基于AIGC的视频生成技术

2023年,商业化文生图产品相继发布,如Stable Diffusion[4]、Midjourney[5]、DALL-E 3[6]等。这些工具通过简洁的文本提示,便能生成高分辨率和高质量的新图像,充分展现了人工智能在创意图像生成方面的能力。然而,由于视频的时序性,从图像到视频的转换仍面临着巨大挑战。尽管工业界和学术界已付出大量努力,但大多数现有的视频生成工具,如Pika[7]和Gen-2[8],仍仅限于生成短暂的几秒钟视频片段。而Sora是第一个打破了这一限制的视频生成模型。它能够根据人类指令,生成长达一分钟的视频,这一突破对人工智能的生成研究与开发产生了深远影响。

AIGC的卓越表现离不开作为基座的大模型,而大模型上的尺度定律是其成功的关键。通过训练具有庞大规模参数的模型,使其表现出新的复杂行为或功能,这一现象通常被称作“涌现”。尽管以ChatGPT为代表的等众多大语言模型都展现出一定程度的涌现能力[9],但在视觉领域的大模型中,涌现能力的存在却鲜有确证。根据Sora公司所发布的技术报告[10],他们成功开发出首个展示出涌现能力的视觉模型。这一成果标志着视觉领域大模型取得重大突破。Sora的涌现能力来源于多方面的技术突破。首先,它借鉴了谷歌DeepMind在视觉模型NaViT上提出的Spacetime Patch[11]。该技术可以将不同分辨率和时长的视频直接转换成Patch序列,避免了放缩或切分等操作带来的信息损失,尽可能保留了视频中复杂的时空变化信息,从而赋予Sora强大的潜力。其次,Sora引入了创新的DiT架

构[12]。DiT采用Transformer的Encoder-Decoder架构来处理包含噪点的输入帧，并在每步预测更清晰的帧。这一架构给Sora带来了更优秀的特性，如可拓展性、鲁棒性和高效性等。结合Spacetime Patch，Sora能够捕捉视频的本质特征，从而提高输出质量。最后，Sora的成功得益于大量高质量的训练数据。它使用GPT-4来丰富视频描述的质量和细节，确保能够准确理解用户复杂的指令。这些关键要素的组合构成了Sora强大的技术基础，使其在相关领域表现出色。

Sora、GPT等预训练大模型不仅能用于个人视频生成，还在新视听领域中展现了巨大的潜力。在传统的电影行业，电影创作是艰苦且昂贵的，通常需要团队配合、设备基础和大量资金投入，但前沿的视频生成技术正在让一键实现电影制作从梦想走入现实。MovieFactory[13]使用ChatGPT制作的精心制作的脚本，基于扩散模型生成了电影风格的视频，代表着这一应用的重大飞跃。而MobileVidFactory[14]只需用户提供简单的文本即可自动生成垂直移动视频。之后，Vlogger[15]的出现，使得用户可以撰写一分钟长的视频博客。如今的Sora更是可以根据用户指令，只需要几秒就能生成富有创造力、逼近现实物理环境的特定风格长视频。能轻松制作电影内容的视频生成模型的迅速发展，标志着电影制作将向低成本化、大众化的方向转变。

2.2 基于AIGC的音频处理技术

AIGC的音频处理技术可以被分为三类任务，分别是：文本生成语音、语音克隆和AI音乐生成。目前文本生成语音的发展已经相当成熟，言语质量已达到自然标准，只需要输入文本就可以输出特定说话者的语音。该任务主要应用于客服及硬件机器人、有声读物制作、语音播报等任务。而微软在2023年1月推出的VALL-E[16]以及多语言版本VALL-E X[17]将文本生成语音引导向了情感化、多语言的方向。VALL-E使用6万小时量级的英语语音数据进行预训练，将Encodec作为量化器，使用其中间层输出作为音频的离散表征，从而将音频变成长序列，压缩了表示空间，简化了对语音的建模难度，大大提高了生成质量。在零样本场景下，只需要3秒的音频作为提示，就能实现高自然度、高音色相似度和情感一致性的语音合成。VALL-E X则进一步支持中文和日语的语音合成。之后文本生成语音技术将向着更丰富情感、更多目标语言的方向发展，为视听行业提供助力。

语音克隆技术是利用给定的目标说话者语音模板，将输入语音转换为目标说话者的语音。此类技术的目的是合成特定说话者的语音，主要被应用于虚拟歌手演唱、自动配音等场景，在声音IP化日渐火热的背景下，此技术的实现对于动画、电影以及虚拟人行业有着重要意义。2023年9月，Spotify推出新的AI语音克隆工具Voice Translation[18]。该工具在OpenAI的自动语音识别（Automatic Speech Recognition，ASR）模型Whisper[19]的技术支持下，使用了语音转文本生成AI模型来翻译音频文件，并使用语音复制模型来匹配原始说话者的风格。此外，Voice Translation能够自动切换到西班牙语、法语和德语等各种语言，而且是“完全原声”，连说话节奏，语气都能还原，引起了许多主持人、主播和明星的惊叹。Whisper使用了680,000小时网络多语言和多任务监督数据作为训练集，如此庞大且多样化的数据集不仅提高了模型的鲁棒性，还使得模型对口音、背景噪音和行业术语也能保持很好的克隆效果。

而在AI音乐生成任务上，目前的AIGC技术可以支持基于开头旋律、图片、文字描述、音乐类型、情绪类型等生成特定乐曲。该技术最早的工作是谷歌在2022年9月发布的AudioLM[20]，其通过接收音频，就可以生成与提示风格类似的音乐。紧接着谷歌又发布了MusicLM[21]，相比于AudioLM，它可以额外支持纯文本的输入来生成相应的音乐。2023年11月16日，谷歌在前面实践的技术基础上，发布了Lyria[22]，并以“改变音乐创作的未来”作为这一AIGC音乐工具的注解。

DeepMind还与YouTube合作，为Lyria打造了两个重要的应用场景：Dream Track和Music AI tools，分别用于生成特定艺术家风格的短视频配乐和专业的音乐创作。只需给Dream Track输入主题与要选择授权的艺术家，它就能生成一段半分钟的短视频背景AI音乐，并且生成相应的音轨和歌词。而Music AI tools则面向音乐人、艺术家以及制作人，只需哼唱旋律，就能直接生成曲谱和演唱，还可以支持任意插入器乐伴奏，从而大大便利了音乐创作。另外，Lyria还加入了音频水印技术，可以在不影响收听体验的情况下嵌入，只有将声音转换成为二维可视化的频谱图才可以被算法识别，这也使得Lyria能够安全地用于音乐创作上。

3 AIGC助力独特视听体验

3.1 生成式视听推荐

除了丰富的视频和音频内容外，新视听产业的发展也离不开推荐系统的支持，即通过基于用户与内容

ID的协同过滤算法,为每个用户推荐合适的内容。然而,受到推荐系统传统范式和复杂领域知识的限制,构建强大的个性化推荐系统依然面临着诸多挑战与阻碍。首先,现有的推荐系统依赖于物品标识ID,而基于专家人工给定的物品标识ID不一定符合用户的新奇需求;其次,现有的推荐系统采用多级过滤范式,通过不断筛选来缩小决策的空间,可能面临次优化和全局视野有限的问题;最后,用户在调节推荐系统时,通常依赖于预先设计的反馈操作,如划过、拉黑或点击标签等,这种方式不仅限制了用户的主动交互体验,也降低了调节效率,使得个性化推荐的精准度和满意度难以达到理想水平。

为了解决上述问题,Wang等[23]首次提出了一个生成式推荐系统范式GeneRec。GeneRec使得推荐系统可以不再受限于人工生成的物品标识ID,能够自动地编辑原ID,或是通过多模态指令生成新ID,从而满足用户多样化的信息需求。Liu等[24]提出了一个基于LLM的生成式新闻推荐框架GENRE,利用已有的新闻数据来构建提示,并经由大语言模型来生成总结、用户画像和个性化新闻等信息。

借助大语言模型的生成能力,AIGC技术驱动的推荐系统有能力将多级过滤范式重塑成单级过滤,即大语言模型本身可以作为一个完整的推荐流程,直接生成要推荐的视听内容,从而消除多级过滤的复杂性和次优性。在AIGC驱动的生成式推荐系统中,基于大语言模型的推荐算法隐式地作用于系统中的所有物品,以决定要推荐哪些内容。

此外,在推荐系统调节上,Zhang等[25]尝试构建可分析基于自然语言形式的用户需求指令的推荐系统新范式,通过形式化推荐指令的偏好、意图和任务形式来准确理解用户。Bao等[26]则尝试将现有数据转化为多种形式的自然语言指令,使得LLM可以很好地适配到推荐系统,提升推荐系统的泛化能力。未来,生成式推荐系统有望取代统治长达10年之久的传统范式,而且生成式推荐系统的潜力还将随着大模型的进步而提高。

随着AIGC技术的发展,基于AIGC的生成式推荐技术有望提供个性化和精准的视听内容推荐,重塑推荐交互体验。它能准确捕捉用户的兴趣和偏好,为用户量身定制内容,使用户能更轻松地发现和享受符合口味的视听作品,从而提高用户满意度和参与度。

3.2 虚拟现实交互

随着预训练模型、多模态生成上的进步,AIGC技术也被用于在虚拟现实中描摹出逼真的人物、场景和环境。除了素材上的堆叠,还可以通过构造优秀的交互体验来增加用户的视听沉浸感,这也是虚拟现实的独特之处。

利用如眼动、手势和姿态等交互技术,并结合图像采集、音频采集等硬件,经由预训练模型进行数据收集、分析和决策后,交互设备能够直接理解人类的各种面部表情、手势和姿态,进而判别用户的真实意图,创造个性化的交互体验。

在台北的Computex 2023展会上,英伟达携手Convai展示了一款名为Kairos的游戏Demo,让玩家可以用自然语言和游戏角色对话[27]。Convai是一款专为虚拟世界而设计的对话人工智能平台,可以让非玩家角色(NPC)具有人类般的智能和反应能力,被英伟达称之为"对未来游戏的一瞥"。Convai不仅仅让角色拥有丰富的知识和专业性,还可以感知场景中的实体,并根据用户的交互产生对应的动作或语言。

而这一切都是基于英伟达的Avatar Cloud Engine(ACE) for Games服务所构建[28],这是一套基于生成式人工智能的模型加工流水线,可以与任何游戏引擎或应用平台集成,让任何用户都能轻松地创建出令人惊叹的虚拟世界体验。它包括三个部分:Nvidia NeMo定制语言模型,并可以根据角色设定进行微调;Nvidia Riva用于语音识别和语音合成,实现实时的语音对话;Nvidia Omniverse Audio2Face则用于根据语音生成相应的面部动画。

此外,Apple公司发布的VR头显产品Vision Pro也集成了AIGC技术[29],用以创造全新的虚拟现实交互体验。Vision Pro可以通过前置摄像头扫描人的面部信息,再基于AIGC技术为用户生成一个"数字分身"。并且当用户正通过FaceTime通话时,数字分身可以动态模仿用户的面部和手部的动作,保留数字分身的体积感和深度。为了避免使用VR时用户与周围的人产生隔绝,Vision Pro中还加入了Eyesight功能,当检测到有人处于附近,Eyesight可以让Vision Pro的外壳上生成用户的眼睛,并真实描摹用户的眼神进行互动。

随着AIGC技术的发展,更多全新视听交互模式即将出现,相信它将成为未来虚拟现实体验的重要推动力,为用户带来更加丰富和多样化的沉浸式体验。

4 AIGC助力独特视听体验

AIGC技术在视频生成、音频处理、个性化推荐和虚拟交互等新视听场景上都取得的令人惊叹的发展,具备强大的应用潜力和研究价值。随着这些AIGC技

术的更新，人们会不由自主地期待马上出现一个可以完美模拟现实的生成式模型。然而，就现实情况而言，AIGC技术在应用于新视听领域时，仍面临着一系列亟待深入探索和有效解决的挑战。

4.1 现实物理模拟

Sora在数据处理和模型架构上的突破，打破了人们对于视频生成模型的上限预估，不仅可以理解和执行人类复杂指令，还能创建具有各种角色和丰富细节的视频，可以在制作长视频的同时，保持“一镜到底”的视觉叙事体验。

然而，Sora仍存在诸多缺点，首先是它对复杂场景中物理原理的处理存在混乱，导致无法准确建模现实物理表现。例如，吃饼干的动作结束后并未出现相应的咬痕，这说明系统偶尔会偏离物理规律。而在运动的模拟上则更为严重，其中Sora生成的运动有时会违反了现实的物理模型，例如物体的不自然变换或对椅子等刚性结构的不正确模拟，从而导致不真实的物理交互。

除了运动，Sora有时会误解提示中与放置或排列相关的指令，从而导致混乱（例如，混淆左右方向）。此外，它在更细节、精确的时间序列设计上依然存在失误，特别是依据指定好的序列来移动镜头时，可能会导致视频时间流发生偏差。而在涉及大量角色或元素的复杂场景中，Sora倾向于插入不相关的动物或人，改变最初设想的场景构图和氛围，偏离了原始计划的叙事或视觉布局。这个问题不仅阻碍了模型准确重建特定场景或叙述的能力，更对其产生与用户期望紧密契合的内容的可靠性以及生成内容的连贯性造成了严重影响。

这种不真实或偏差的生成可能源于分布外(Out-of-Distribution，OOD)场景鲁棒性上的不足[30]。Sora的成功源于使用了超大规模的原始训练数据集，而上述场景可能是训练数据集中的长尾信息导致学习到了不鲁棒的特征。但这种罕见的场景往往是人类主观评判模型是否具备强大现实物理模拟能力的基准，因此对于视频生成模型的分布外鲁棒性依然是一个研究蓝海。

现有的技术通过对抗性训练[31-32]、模型集成[33]、对比学习[34]和知识检索[35]来提高模型的鲁棒性，然而在针对下游具体任务的微调时，往往过分强调任务目标，与鲁棒性目标背道而驰，这也会导致灾难性遗忘[36]和鲁棒性损失[37]。因此存在的一个重大的挑战就是开发具备任务目标适应性的鲁棒性增强方法，在充分解决目标任务的前提下，保持模型的鲁棒性。

此外，为了适配本地的视听应用场景，减少存储和计算代价，模型往往会经由低秩分解[38]、剪枝[39]、量化[40-41]和知识蒸馏[42]等压缩技术的处理，从而得到更小、更快但性能相当的模型。因此鲁棒性和压缩技术的结合也是支撑大模型落地到新视听产业的重要策略，现有工作已经成功对普通深度神经网络进行压缩且保持高鲁棒性[43]，然而如何构建压缩大模型参数量的同时保持其涌现能力和高鲁棒性的策略尚需探索。

4.2 资源利用效率

大模型的成功来源于其在训练数据、参数量上的可拓展性，并呈现了显著的尺度定律——更大的模型总能带来更好的性能。然而，这一可拓展性的代价是对资源的极度饥渴，不论是算力硬件、内存、服务器还是电力，大模型的资源需求都对产业的高质量发展、绿色发展带来的巨大挑战，同时也阻碍了新技术的普及。因此有许多研究致力于提高大模型的资源利用效率，包括算法结构的优化、训练方法的改进和更好的数据管理策略。

首先是对于Transformer模型的结构优化。一些工作专注于优化Transformer中的注意力机制，重新设计近似于点积注意力的新算子，以实现更低的时间/内存复杂度[44-46]。此外，还有工作聚焦到硬件层面，通过改进CUDA代码中内核融合、gemm优化等方面来减少内存需求和提高计算速度[47-48]。其次是引入了新的训练策略，例如使用混合精度训练[49]、跳层计算[50-51]和输入剪枝[52]策略来提高模型的计算速度。最后则是使用重要性采样[53]、数据增强[54]等方式来减少对数据规模的需求。

然而以上资源效率优化策略主要针对的是通用场景下的大模型部署，对特定场景下的大模型资源利用效率优化仍需进一步考虑。新视听领域为大模型的使用带来了更多的开放性挑战，例如不同模态数据资源使用的权衡、多种资源效率指标的综合优化和多模型集成策略等，我们需要进一步构建大模型资源利用效率评估的统一基准、探索高效的数据调度方法和洞察大模型的尺度定律。

当前，绿色低碳已成为社会经济发展的重要价值取向，社会生产生活方式进行全面绿色低碳转型，新视听产业的高质量发展、助力“双碳”目标实现的需求都要求我们构建高效的AIGC资源使用范式，通过融合数智化和低碳化，新视听将呈现更具竞争力的产业生态。

4.3 技术的道德焦虑

如今的AIGC的技术迭代频率在反复冲击着所有人

的认知,AIGC的真实上限也许已经超出了人类一开始最大胆的设想,对AIGC的研究也许已经离“万物皆可AIGC”不远,但在媒体报道的推动下,人们对生成人工智能的快速发展所带来的不利后果的担忧与日俱增,这些担忧涉及多个方面,且有可能形成持久的社会焦虑和舆论。在这一背景下,理性考量AIGC技术在视听领域面临的伦理和法律挑战,是一个非常紧迫的任务。

喻国明教授认为,“AIGC弥合了数字文明社会的‘能力沟’”[55]。AIGC技术在各个任务的发光发热,不仅仅让每个人能利用其完成工作,还有可能让人沦为技术的附庸,忽略了其中隐藏的风险。

AIGC由人类文化浇灌而成,其本身无差别地学习训练语料的知识,因此也会具备一定的偏见和错误知识,当处于事物的成长期时,人类对于这种缺陷可以保持一定程度上的容忍,但当AIGC被真正普及用于新视听内容的生产时,这种语义理解上的偏差和谬误就成为了重要的风险因素。

最为焦点的工作就是AI生成有害内容的检测,其最常见方法是考虑使用自动核查来辅助我们捕获网络上传播的AI生成有害信息。然而随着AIGC技术的精进,AI生成的内容在情感、心理学和表达上愈发接近人类,且有害内容并非以易于发现的关键词呈现,而是以潜在的、扭曲的社会偏见来危害人类认知。现有工作发现了生成式大模型可能产生歧视性、排斥性的语言,并有可能放大人类社会数据中的刻板印象[56]。此外,多模态模型也同样可能表现出对性别、种族和宗教方便的社会偏见[57]。虽然已经有一些方法可以减轻生成时所伴随的偏见[58-59],但是依然没有改变大模型可能生成有害、偏见内容的本质。而且高度拟真化的AIGC内容有可能混淆虚拟视听与现实的边界,数字角色生成、语音克隆和视频生成都有可能成为不法分子的帮凶,难辨真假的信息让人杯弓蛇影。

基于此,将AIGC技术与社会主义核心价值观对齐是一个有助于减少AIGC技术危害性、追求安全视听体验和助力主流价值观传播的大模型优化策略。其重点在于构建具备社会主义价值观先验的人类反馈强化训练策略和多模态训练数据。通过对齐主流价值观的AIGC技术,新视听可以给全社会带来福祉,寻求社会道德的最大公约数。

5 结论

随着数据资源的流通开放、全国算力一体化体系的构建以及数字基础设施的不断完善,AIGC具备替代内容创作者完成更多信息挖掘、素材加工等基础性劳动的潜力,并更进一步提供强大的内容生产范式,为更具想象力和丰富性的内容创作提供可能。

现有的AIGC技术正在迈向高生成内容质量、强指令语义理解和多领域泛化,在视频生成、音频处理和个性化推荐和虚拟交互视听体验上,AIGC将赋能新视听产业发展、丰富人民生活、提升社会现代化水平。

此外,AIGC技术也势必会给广播电视和网络视听行业带来一定挑战。现有数字基础设施和算力产业尚未形成规模,算法仍未完全释放潜能,且存在偏见放大、助长虚假信息传播和侵犯数字版权等道德不良面,亟需加快相关监管法律的完善和技术创新的推进。

人工智能等新兴技术的出现正在深刻改造现有的产业,展现其创新驱动非线性成长的巨大潜力。探究以AIGC为代表的创新视听技术驱动新质生产力的一般性和特殊性规律,催生相关新兴产业、未来产业,以融合而非破坏的形式完成创新技术范式转变,是视听产业将要面临的挑战和机遇。

中国的视听产业正站在一个由追赶向超越追赶转变的重要节点,新兴的AIGC技术范式宛如一股强大的动力源泉,为加速视听产业新质生产力的孕育和成长提供了前所未有的战略机遇。探索这一技术范式不仅深度赋能新视听产业的全链条、全过程,更能成为推动社会经济发展的强劲变速器,引领着整个产业迈向更加广阔和深邃的未来。

参考文献(References):

[1] Wu J, Gan W, Chen Z, et al. Ai-Generated Content (AIGC): a survey [DB/OL]. arXiv:2304.06632, 2023.

[2] OpenAI. Introducing ChatGPT[EB/OL]. [2024-1-08].https://openai.com/blog/chatgpt.

[3] Xie S. Take on SORA technical report[EB/OL]. (2024-2-16)[2024-1-08].https://twitter.com/sainingxie/status/1758433676105310543, 2024.

[4] Rombach R, Blattmann A, Lorenz D, et al. High-resolution image synthesis with latent diffusion models[C]//Conference on Computer Vision and Pattern Recognition(CVPR), 2022: 10684-10695.

[5] M AI. Midjourney: Text to image with AI art generator[EB/OL]. [2024-1-08]. https://www.midjourneyai.ai/en.

[6] Betker J, Goh G, Jing L, et al. Improving image generation with better captions[J]. Computer Science, 2023, 2(3): 8.

[7] Pika. Pika is the idea-to-video platform that sets your creativity in motion[EB/OL]. [2024-1-08].https://pika. art/home.

[8] Pika. Gen-2: the next step forward for generative AI[EB/OL].

[2024-1-08].https://research. runwayml.com/gen2.

[9] Wei J, Tay Y, Bommasani R, et al. Emergent abilities of large language models[EB/OL]. arXiv:2206.07682, 2022.

[10] OpenAI. Video generation models as world simulators[EB/OL]. [2024-1-08]. https://openai. com/research/video-generation-models-as-world-simulators.

[11] Dehghani M, Mustafa B, Djolonga J, et al. Patch n'pack: Navit, a vision transformer for any aspect ratio and resolution[C]// Advances in Neural Information Processing Systems 36 (NeurIPS), 2023.

[12] Peebles W, Xie S. Scalable diffusion models with transformers[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR), 2023: 4195-4205.

[13] Zhu J, Yang H, He H, et al. Moviefactory: automatic movie creation from text using large generative models for language and images[C]//Proceedings of the 31st ACM International Conference on Multimedia, 2023: 9313-9319.

[14] Zhu J, Yang H, Wang W, et al. Mobilevidfactory: automatic diffusion-based social media video generation for mobile devices from text[C]//MM '23: Proceedings of the 31st ACM International Conference on Multimedia, 2023: 9371-9373.

[15] Zhuang S, Li K, Chen X, et al. Vlogger: make your dream a vlog[DB/OL]. arXiv:2401.09414, 2024.

[16] Wang C, Chen S, Wu Y, et al. Neural codec language models are zero-shot text to speech synthesizers[DB/OL]. arXiv: 2301.02111, 2023.

[17] Zhang Z, Zhou L, Wang C, et al. Speak foreign languages with your own voice: cross-lingual neural codec language modeling [DB/OL]. arXiv:2303.03926, 2023.

[18] Spotify. Spotify's AI voice translation pilot means your favorite podcasters might be heard in your native language [EB/OL]. [2024-1-08].https://newsroom.spotify.com/.

[19] Radford A, Kim J W, Xu T, et al. Robust speech recognition via large-scale weak supervision[C]// ICML'23: Proceedings of the 40th International Conference on Machine Learning, 2023: 28492-28518.

[20] Borsos Z, Marinier R, Vincent D, et al. Audiolm: a language modeling approach to audio generation [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2022, 31: 2523-2533.

[21] Agostinelli A, Denk T I, Borsos Z, et al. Musiclm: generating music from text[DB/OL]. arXiv:2301.11325, 2023.

[22] Google. Transforming the future of music creation [EB/OL]. [2024-1-08]. https://deepmind. google/discover/blog/transforming-the-future-of-music-creation/.

[23] Wang W, Lin X, Feng F, et al. Generative recommendation: towards next-generation recommender paradigm[DB/OL]. arXiv:2304.03516, 2023.

[24] Liu Q, Chen N, Sakai T, et al. A first look at LLM-powered generative news recommendation[DB/OL]. arXiv:2305.06566, 2023.

[25] Zhang J, Xie R, Hou Y, et al. Recommendation as instruction following: a large language model empowered recommendation approach[DB/OL]. arXiv:2305.07001, 2023.

[26] Bao K, Zhang J, Zhang Y, et al. Tallrec: an effective and efficient tuning framework to align large language model with recommendation[C]//RecSys '23: Proceedings of the 17th ACM Conference on Recommender Systems, 2023: 1007-1014.

[27] Nvdia. Introducing NVIDIA ACE for games-spark life into virtual characterswith generative AI[EB/OL]. [2024-01-08] https://www.nvidia.cn/geforce/news/nvidia-ace-architecture-ai-npc-personalities/.

[28] Nvdia. NVIDIA ACE [EB/OL]. [2024-1-08]. https://developer.nvidia.com/ace, 2023.

[29] Apple. Apple Vision Pro (2023) [EB/OL]. [2024-1-08]. https://www.apple.com/apple-vision-pro/.

[30] Wang J, Hu X, Hou W, et al. On the robustness of ChatGPT: an adversarial and out-of-distribution perspective[DB/OL] arXiv:2302.12095, 2023.

[31] Ni S, Li J, Kao H Y. R-AT: regularized adversarial training for natural language understanding[C]//Findings of the Association for Computational Linguistics: EMNLP2022, 2022: 6427-6440.

[32] Liu X, Cheng H, He P, et al. Adversarial training for large neural language models[DB/OL]. arXiv:2004.08994, 2020.

[33] Artetxe M, Bhosale S, Goyal N, et al. Efficient large scale language modeling with mixtures of experts[C]//Conference on Empirical Methods in Natural Language Processing (EMNLP), 2022: 11699-11732.

[34] Gao T, Yao X, Chen D. SimCSE: simple contrastive learning of sentence embeddings [C]//Conference on Empirical Methods in Natural Language Processing (EMNLP), 2021: 6894-6910.

[35] Lewis P, Perez E, Piktus A, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks[J]. Advances in Neural Information Processing Systems, 2020, 33: 9459-9474.

[36] Thanh-Tung H, Tran T. Catastrophic forgetting and mode collapse in GANs[C]//2020 International Joint Conference on Neural Networks (IJCNN), 2020.

[37] Suprem A, Pu C. Evaluating generalizability of fine-tuned models for fake news detection[DB/OL]. arXiv:2205.07154, 2022.

[38] Hu E J, Shen Y, Wallis P, et al. LoRA: low-rank adaptation of large language models[C]//International Conference on Learning Representations. (ICLR), 2021.

[39] Ma X, Fang G, Wang X. Llm-pruner: on the structural pruning of large language models[J]. Advances in Neural Information Processing Systems, 2023, 36: 21702-21720.

[40] Dettmers T, Lewis M, Belkada Y, et al. Gpt3. int8 (): 8-bit

matrix multiplication for transformers at scale[C]//Advances in Neural Information Processing Systems, 2022, 35: 30318-30332.

[41] Frantar E, Ashkboos S, Hoefler T, et al. Gptq: accurate post-training quantization for generative pre-trained transformers [DB/OL]. arXiv:2210.17323, 2022.

[42] Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network[DB/OL]. arXiv:1503.02531, 2015.

[43] Phan H, Yin M, Sui Y, et al. Cstar: towards compact and structured deep neural networks with adversarial robustness[C]//AAAI'23/IAAI'23/EAAI'23: Proceedings of the 37th AAAI Conference on Artificial Intelligence and 35th Conference on Innovative Applications of Artificial Intelligence and 13th Symposium on Educational Advances in Artificial Intelligence, 2023, 230: 2065-2073.

[44] Kitaev N, Kaiser L, Levskaya A. Reformer: the efficient transformer [C]// International Conference on Learning Representations (ICLR), 2020.

[45] Katharopoulos A, Vyas A, Pappas N, et al. Transformers are RNNs: fast autoregressive transformers with linear attention[C]// Proceedings of the 37th International Conference on Machine Learning, 2020: 5156-5165.

[46] Shen Z, Zhang M, Zhao H, et al. Efficient attention: attention with linear complexities[C]// Winter Conference on Applications of Computer Vision (WACV), 2021: 3531-3539.

[47] Dao T, Fu D Y, Ermon S, et al. Flashattention: fast and memory-efficient exact attention with io-awareness[C]//36th Conference on Neural Information Processing Systems (NeurIPS), 2022, 35: 16344-16359.

[48] Dao T. FlashAttention-2: faster attention with better parallelism and work partitioning[DB/OL]. arXiv:2307.08691, 2023.

[49] Micikevicius P, Narang S, Alben J, et al. Mixed precision training [C]//International Conference on Learning Representations (ICLR), 2018.

[50] Din A Y, Karidi T, Choshen L, et al. Jump to conclusions: short-cutting transformers with linear transformations[DB/OL]. arXiv:2303. 09435, 2023.

[51] Wang J, Chen K, Chen G, et al. Skipbert: efficient inference with shallow layer skipping [C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2022: 7287-7301.

[52] Wang H, Zhang Z, Han S. SpAtten: efficient sparse attention architecture with cascade token and head pruning[C]//International Symposium on High-Performance Computer Architecture (HPCA), 2021: 97-110.

[53] Csiba D, Richtárik P. Importance sampling for minibatches[J]. Journal of Machine Learning Research, 2018, 19(1): 962-982.

[54] Rebuffi S A, Gowal S, Calian D A, et al. Data augmentation can improve robustness[C]//35th Conference on Neural Information Processing Systems (NeurIPS), 2021, 34: 29935-29948.

[55] 喻国明,苏健威.生成式人工智能浪潮下的传播革命与媒介生态——从ChatGPT到全面智能化时代的未来[J].新疆师范大学学报(哲学社会科学版), 2023, 44(05): 81-90.

[56] Weidinger L, Mellor J, Rauh M, et al. Ethical and social risks of harm from language models [DB/OL]. arXiv:2112.04359, 2021.

[57] Janghorbani S, Melo G D. Multi-modal bias: introducing a framework for stereotypical bias assessment beyond gender and race in vision–language models[C]// 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL), 2023: 1725-1735.

[58] Coda-Forno J, Witte K, Jagadish A K, et al. Inducing anxiety in large language models increases exploration and bias [DB/OL]. arXiv:2304.11111, 2023.

[59] Yu Y, Zhuang Y, Zhang J, et al. Large language model as attributed training data generator: a tale of diversity and bias[C]// 37th Conference on Neural Information Processing Systems (NeurIPS), 2024, 36.

编辑:王谦