

人工智能生成内容技术综述

张旭龙¹, 瞿晓阳¹, 谢骏飞^{1,2}, 刘鹏程¹, 肖春光³, 王健宗¹

1. 平安科技(深圳)有限公司, 广东 深圳 518063;
2. 中国科学技术大学先进技术研究院, 安徽 合肥 230026;
3. 深圳市宝安区教育事业发展中心, 广东 深圳 518101

摘要

人工智能生成内容(AIGC)技术作为一种新的内容创作方式,为数字经济中的数字智能挑战提供了新的解决方案。AIGC利用人工智能技术辅助或替代人工创作,根据用户需求生成文本、图像、语音、视频等多种形式的内容。近年来,深度学习模型的快速发展显著提升了AIGC的生成能力,使其成为人工智能领域的研究热点。AIGC作为一项底层技术,拥有巨大的应用潜力,但也存在一些局限性。对AIGC技术进行了综述,首先介绍了AIGC的相关算法,包括生成对抗网络、变分自编码器扩散模型和大型生成模型等;然后探讨了AIGC在不同内容形式生成中的应用;最后总结了AIGC技术面临的挑战,并展望了未来的发展趋势。

关键词

人工智能生成内容;生成对抗网络;变分自编码器;扩散模型;大语言模型

中图分类号:TP391

文献标志码:A

doi:10.11959/j.issn.2096-0271.2025071

AIGC: a survey of AI-generated content technologies

Zhang Xulong¹, Qu Xiaoyang¹, Xie Junfei^{1,2}, Liu Pengcheng¹, Xiao Chunguang³, Wang Jianzong¹

1. Ping An Technology (Shenzhen) Co., Ltd., Shenzhen 518063, China
2. Institute of Advanced Technology, University of Science and Technology of China, Hefei 230026, China
3. Shenzhen Baoan District Education Information Center, Shenzhen 518101, China

Abstract

As a novel content creation method, AI-generated content (AIGC) technology provides new solutions to the challenges of digital intelligence in the digital economy. AIGC utilizes artificial intelligence to assist or replace manual content creation, generating various forms of content such as text, images, audio, and video based on user needs. In recent years, the rapid development of deep learning models has significantly enhanced the generative capabilities of AIGC, making it a research hotspot in the field of artificial intelligence. As a fundamental technology, AIGC has enormous application potential but also faces certain limitations. This paper provides a comprehensive overview of AIGC technology. Firstly, it introduces the related algorithms of AIGC, including generative adversarial networks, variational autoencoder, diffusion models, and large language models; then it explores the applications of AIGC in generating different content forms; finally, it summarizes the challenges faced by AIGC technology and discusses future development trends.

Key words

AI-generated content, generative adversarial networks, variational autoencoder, diffusion model, large language model

0 引言

作为一种新兴的内容创作方式,人工智能生成内容 (artificial intelligence generated content, AIGC) 是对传统内容生产模式,例如专业生产内容 (professional generated content, PGC)^[1-2] 和用户生成内容 (user-generated content, UGC)^[3-4] 的重要补充。AIGC 借助人工智能技术生成内容,以满足用户的个性化需求,被业界视为一项极具应用潜力的技术。因此,深入研究和应用 AIGC 技术具有重要的现实意义。

近年来,深度学习技术的飞速发展促使 AIGC 技术在算法、各类生成任务和技术产品方面取得了显著突破^[5]。在算法层面,基于深度学习的数据生成模型已经能够合成满足多样化需求的数据^[6-9]。例如生成对抗网络 (generative adversarial networks, GAN)^[10-11],它通过对抗学习的方式,使生成-判别模型能够将噪声直接转化为样本,这种特性使 GAN 可以被应用于众多领域。为了构建比 GAN 更灵活的模型架构并精确计算对数似然,一些研究者提出了一种基于扩散模型的生成方法^[12-13]。扩散模型包含两个主要过程:在前向扩散阶段逐步加噪,然后在反向阶段预测噪声,恢复出原始数据。训练完成后,模型可以利用这些去噪方法从随机输入生成新的“干净”数据。上述步骤使扩散模型在数据生成领域展现出卓越的性能,并引发了越来越多的关注。变分自编码器 (variational autoencoder, VAE) 是一种生成模型,它通过引入一种概率方法来描述数据生成过程,从而扩展了传统自编码器的概念^[14]。它的工作原理是,首先将输入编码到潜在

空间的分布中,然后从这个分布中采样以生成新数据。与仅压缩和重建数据的标准自编码器不同,VAE 能够生成新的样本^[15]。这是通过构建一个损失函数来实现的,该损失函数不仅最小化重建误差,还包括一个正则化潜在变量分布的项,通常会鼓励该分布近似于标准正态分布。近年来,人工智能生成任务不断发展,其中大模型迅速崛起,大模型相比其他深度学习模型在人工智能生成内容方面展现出更加显著的优势,例如 DeepSeek^[16]、GPT-4^[17] 和 DALL-E-2^[18]。这类模型凭借其庞大的参数量和出色的泛化能力,已成为当前人工智能生成内容领域的焦点。AIGC 模型的不断完善使其在各类生成任务 (包括图像、文本、语音、视频等生成) 中表现出色^[19]。随着各类生成任务的生成模型不断改进,将生成模型转化为技术产品已成为一种必然趋势,例如 ChatGPT^[20] 和 DreamStudio^[21]。尽管这些生成式人工智能方法能够生成更加逼真的样本,但是许多生成的样本并不符合预期要求,这导致将生成的样本应用于下游任务时,难以获得理想的结果。

本文旨在梳理 AIGC 技术的研究现状,对 AIGC 技术进行归纳总结,并分析该领域存在的挑战和未来发展方向。

1 基于 AIGC 技术的算法研究

AIGC 技术的算法研究可分为 4 类:基于 GAN 模型的算法研究、基于 VAE 模型的算法研究、基于扩散模型的算法研究、基于大型生成模型的算法研究。

1.1 基于 GAN 模型的算法研究

Goodfellow 等^[22]在其开创性工作中提

出了一种包含两个神经网络的框架，即生成器和判别器，如图1所示。生成器的目标是学习真实数据的分布，并生成逼真的样本；判别器则负责判断输入样本来自真实数据，还是生成器生成的伪造样本。这两个网络以相互对抗的方式进行训练，最终达到纳什均衡状态^[23]。该框架旨在通过生成模型更好地逼近真实数据分布，因此被称为GAN。GAN因其能够生成高质量数据样本的特性，在图像、文本、语音等领域受到了广泛关注^[24]。

在视觉生成领域，越来越多的研究者利用GAN的特性生成高质量的图像或视频内容。例如，Deng等^[25]提出了一种用于可控图像生成的三维感知条件生成模型，该模型能够同时渲染图像和像素对齐的标签特征图。Yin等^[26]提出了一种具有面部对称性先验的新型三维GAN反演方法，用于解决三维GAN反演中单目图像重建三维肖像的不适定问题。Xu等^[27]针对图像修复中存在的纹理模糊和结构失真等问题，提出了一种名为SH-GAN的网络模型。该网络引入了一个名为Spectral Hint Unit的谱处理模块，并提出了两种新颖的二维谱处理策略，即异质滤波和高斯分割，有效解决了纹理模糊和结构失真问题。Jain等^[28]提出了一种名为FcF（fourier coarse-to-fine）的图像修复框架，旨在同时生成更优的纹理和结构。该框架将快速傅里叶卷积的感知能力与粗到

细生成器的共同调制能力相结合，以捕捉全局重复纹理，并生成逼真的图像结构。

在文本生成领域，许多学者也利用GAN的特性进行文本生成或图像生成等相关研究。例如，Kanagawa等^[29]提出了一种带有生成对抗网络的风格预测器，以减少文本预测中全局样式标记与参考语音获得的风格嵌入之间的不匹配问题。Qu等^[30]提出了一种名为MOSTEL的端到端可训练场景文本编辑框架，用于生成清晰可读的编辑文本图像。

在语音生成领域，Yoneyama等^[31]提出了一种快速且音高可控的高保真神经声码器。该方法通过在噪声激励信息上分层条件化谐振滤波网络，并将源-滤波理论引入HiFi-GAN中，实现了语音质量和合成速度的提升。

虽然基于GAN的生成方法能够在上述领域通过较少的迭代次数生成高质量的样本数据^[32-33]，但这些方法在训练后期经常变得不稳定，甚至出现训练崩溃的现象。此外，基于GAN生成的数据并不总是能够很好地应用于下游任务，这可能导致训练成本显著增加。因此，研究者开始探索比GAN更灵活且具有鲁棒性的生成模型。

1.2 基于VAE模型的算法研究

在神经网络架构的发展历程中，自编码器是关键之一。自编码器主要由两个组件构成：编码器和解码器。编码器将输入转换为包含概念的数学表示，解码器则利用这些表示来生成与原始输入相似的新数据。如图2所示，VAE^[15]是一种特定类型的自编码器，它在编解码过程中增添了概率元素，从而生成多样且多变的数据样本。在语音生成领域，Melechovsky等^[34]提出了一种从文本到语音（text to speech，

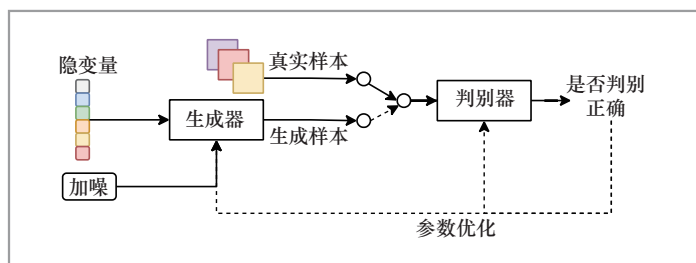


图1 GAN模型架构

TTS) 模型, 利用多级变分自编码器和对抗学习来解决 TTS 中的带口音语音合成和转换问题。Chen 等^[35]针对有声读物合成语音的表现力受到普遍的模型架构和训练数据中不平衡的风格分布的限制问题, 使用基于 VQVAE 预训练的自监督风格增强方法, 合成富有表现力的有声读物语音。Li 等^[36]提出了跨语句条件变分自编码器语音合成框架, 以增强语调并确保自然语音生成。该框架利用预训练语言模型的强大表征能力以及变分自编码器的再表达能力, 从周围句子中提取声学特征、说话人特征和文本特征, 以生成上下文敏感的语调特征, 更准确地模拟人类语调的生成过程。Xiang 等^[37]使用深度复数卷积循环网络与 VAE 结合进行语音增强, 他们假设信号的潜在变量遵循复高斯分布, 对这些分布建模能够更好地捕捉复杂信号的行为。

在图像领域, VAE 也被广泛用于图像生成、虚拟人等任务。利用变分自编码器对视频进行时空压缩, 在 OpenAI 的 SORA 以及众多其他视频生成模型中发挥着关键作用。例如, 许多类似 LLM 的视频模型在 VQVAE 框架内学习源自三维 VAE 的离散标记的分布, 而大多数基于扩散的视频模型则捕捉由二维 VAE 提取的未经量化的连续潜在变量的分布。Zhao 等^[38]提出了一种用于训练潜在视频模型的视频 VAE 的方法, 使其潜在空间与给定的图像 VAE 兼容。在无监督学习中, 提取表征学习空间是机器学习中的一个开放性挑战。基于连续潜在表示的变分自编码器和基于离散潜在表示的矢量量化-变分自编码器是无监督表征学习的重要方法。VQ-VAE 是一种离散潜在变量模型, 能够在无监督学习中学习图像的显著特征表示, 它是连续潜在变量模型 VAE 的一个可行替代方案。然而, 训练深度离散变量模型是具有挑战性

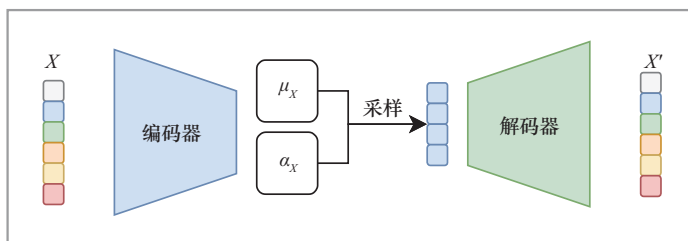


图2 变分自编码器模型架构概览

的, 因为离散化操作本身是非可微的。Pucci 等^[39]提出了胶囊向量-变分自编码器, 由量化码本表示的离散瓶颈被一个胶囊层取代。胶囊可以成功地应用于聚类过程, 重新引入了模型瓶颈的可微性。

1.3 基于扩散模型的算法研究

Ho 等^[40]提出了一种名为扩散模型的生成模型, 该模型不仅具备比 GAN 更灵活的模型架构, 而且展现出更强的鲁棒性。扩散模型的设计灵感源于非平衡热力学, 如图 3 所示, 其过程可分为前向扩散和反向扩散两个阶段。在前向扩散阶段, 模型逐步向数据添加高斯噪声, 直至最终数据服从各向同性高斯分布。反向扩散阶段则逆转这一过程, 从噪声分布中逐步恢复原始数据。扩散模型能够生成比 GAN 更高质量的数据, 因此在图像、文本、语音等生成领域受到了广泛关注。

在视觉生成领域, 越来越多的研究者利用扩散模型的特性生成高质量的图像或视频内容。例如, Bhunia 等^[41]提出了一种基于扩散的方法——人物形象扩散模型 (person image diffusion model, PIDM), 用于处理生成图像过程中可能遇到的复杂变形和严重遮挡问题。PIDM 将复杂的转移问题分解为一系列简单的前向-后向去噪步骤, 有助于学习合理的源到目标转换轨迹, 从而生成具有真实纹理

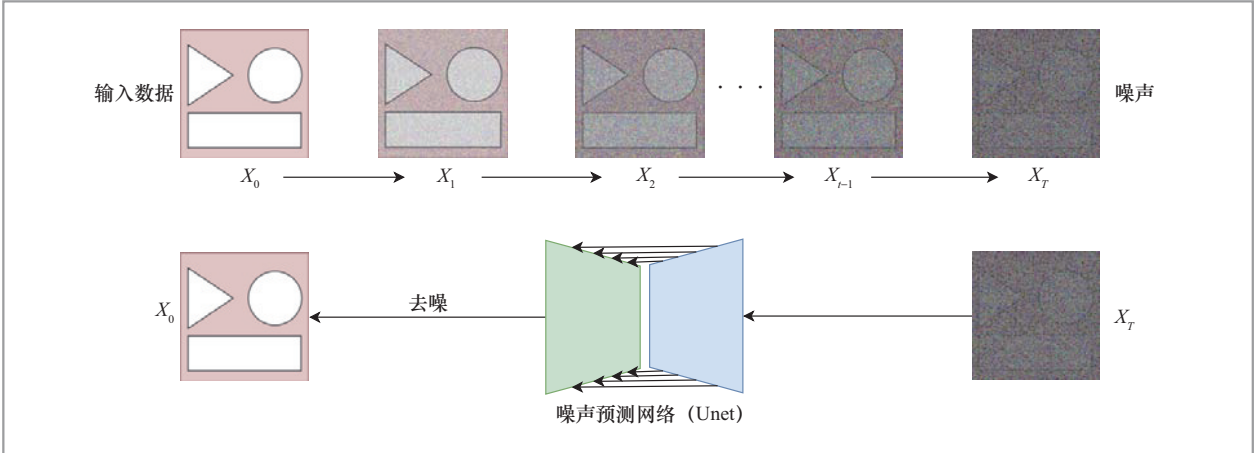


图3 扩散模型架构概览

和未失真外观细节的图像。Blattmann等^[42]提出了用于高效高分辨率视频生成的视频潜在扩散模型。该方法的关键在于选择基于预训练的图像扩散模型，并通过引入时间对齐层在时间上微调它们，将它们转变为视频生成器，从而制作出高质量的视频内容。

在文本生成领域，许多学者也利用扩散模型生成高质量的文本数据。例如，为了解决文本生成问题，He等^[43]提出了DiffusionBERT，这是一种基于离散扩散模型的新型生成式掩码语言模型，用于代替自回归模型，并能够生成更加丰富和多样化的文本。Lin等^[44]提出了一种新颖的扩散语言模型GENIE，采用了序列到序列的结构。该模型中，双向编码器对源序列进行编码，而去噪解码器以非自回归的方式预测并去除目标序列中的噪声。通过这种设计，模型能够逐步从初始的嘈杂状态中精炼输出，生成更具多样性的文本。此外，该模型引入了一种名为连续段落去噪的新型预训练方法，旨在将整个段落作为目标序列进行去噪，从而生成更具多样性的文本样本。

同样地，在语音生成领域，Takahashi

等^[45]为了生成高质量的歌声，提出了一种用于歌声神经声码器的分层扩散模型，实验结果表明该方法能够为多位歌手生成高质量的歌声。另一项类似的研究^[46]也证明了扩散模型在生成高质量语音方面的有效性。

尽管在上述领域，基于扩散模型的生成方法凭借其强大的鲁棒性能够生成高质量的数据样本^[47]，然而该模型依赖于长马尔可夫链的扩散步骤来生成样本，导致计算资源和时间消耗较高。因此，一些学者开始探索其他类型的生成式模型，以突破这些限制。

1.4 基于人工智能生成式大模型的算法研究

自ChatGPT问世以来，基于人工智能生成式大模型的崛起引发了广泛关注，越来越多的科技从业者和企业开始重视其应用价值。这类模型的兴起并非偶然，而是得益于算法研究的突破。本节将从人工智能生成式大模型的算法研究角度出发，探讨当前生成式大模型的研究现状，并总结这类模型的优势和不足。

生成式大模型的雏形可以追溯到基于Transformer的BERT模型。BERT模型采用基于注意力的Transformer结构和自监督学习方法,通过对无标签文本数据集进行预训练,使其生成高质量文本。尽管存在多种基于BERT模型的变体^[48-50],但与以GPT为代表的另一类生成式大模型相比,它们在未经微调的情况下难以有效地应用于下游任务。而以GPT为代表的生成式大模型在少样本甚至零样本学习的文本生成方面表现出更具竞争力的性能。

Radford等^[51]开创性地提出了GPT-1模型。该模型采用了一个12层的Transformer结构。虽然GPT-1模型仅包含1.17亿个参数,但它在自然语言理解、问答、语义相似性和文本分类等方面均表现出色,证明了GPT模型的可行性。GPT-2^[52]则专注于在没有明确监督的情况下学习自然语言处理任务。尽管相比于GPT-1,GPT-2的改动幅度并不大,但它在不同类别的词汇性能识别、捕捉长期依赖关系、常识理解、阅读理解、摘要和翻译等7个生成任务方面均取得了显著进步,进一步验证了GPT模型的可行性。

真正使人工智能生成式大模型呈现爆发式增长的原因是GPT-3模型的出现。GPT-3拥有1750亿个参数、96个注意力层以及320万的批处理大小,规模大幅增加^[53]。通过对小说、论文、网站等各种在线内容进行广泛训练,GPT-3在文本生成和语言翻译方面表现出显著的性能提升。GPT-3.5是广受欢迎的GPT-3的一个变体,而ChatGPT则是GPT-3.5经过精细调整的版本。相比于GPT-3,GPT-3.5引入了额外的精细调整过程,使其能够更好地理解和遵循人类指令,生成更符合预期效果的内容,从而具有更好的实用性。2023年3月,OpenAI发布了GPT-4^[17],

这是一个大型多模态模型,能够接收文本和图像作为输入,并生成文本作为输出。与GPT-3.5仅限于文本到文本生成相比,GPT-4具有更强大的能力。GPT模型在文本生成领域的出色表现,使其在其他领域也受到了广泛关注。

2025年1月发布的DeepSeek-V3^[54]是一个拥有6710亿参数的MoE语言模型,它采用了多头潜在注意力和DeepSeekMoE架构,并且首创了无辅助损失的负载平衡策略和多令牌预测训练目标。DeepSeek-V3在14.8万亿个多样化的质量标记上进行了预训练,随后通过监督微调和强化学习阶段充分发挥其能力。尽管性能出色,但其完整训练仅需278.8万H800 GPU小时,训练过程非常稳定。另外,DeepSeek-R1是一个强大的开源强化学习模型,专为复杂决策和优化而设计。它在推理和生成任务中表现出色,能够处理复杂的逻辑和数学问题。

在视觉生成领域,基于GPT模型的方法取得了令人瞩目的成就。例如,OpenAI开发的DALL-E 2^[55]是对GPT模型的改进,能够在几分钟内根据文本描述创建独特且高质量的图像。Stable Diffusion则是另一种改进方法,能够快速将文本转换为图像,并提供AI绘画等相关功能。Bai等^[56]提出了一种能够在不使用任何语言数据的情况下学习大型视觉模型的方法,用于解决多种视觉生成任务。

在语音生成领域,基于GPT模型的思想也得到了越来越广泛的应用。例如,Wang等^[57]提出了VALLE,这是一种从文本到语音(TTS)的大模型,VALLE模型能够合成高质量的个性化语音。Rubenstein等^[58]引入了一种名为AudioPaLM的模型,这是一个用于语音理解和生成的大语言模型。该模型将基于文

本的语言模型 PaLM-2^[59]和基于语音的语言模型 AudioLM^[60]融合成一个统一的多模态架构，可处理和生成文本和语音，包括语音识别和语音到语音的翻译等应用。

在多模态生成领域，Google 发布了一款名为 Gemini 的多模态生成式大模型。Gemini 专为多模态应用而设计，能够在文本、图像、视频、音频和代码之间进行无缝推理。实验结果表明，在上述任务的一系列基准测试中，Gemini 的性能超越了目前最先进的水平。

虽然基于人工智能生成式大模型的方法能够在上述领域中生成符合人们预期的数据样本，并广泛应用于文本、图像、视频、音频等领域，但这类模型仍然存在一些局限性。例如，在某些特定领域，尤其涉及数据隐私和安全等方面，生成式大模型的表现仍有待提升。此外，在一些需要逻辑推理的任务上，例如专业资格考试等，这类模型与人类水平还存在一定差距。尽管存在这些不足，但生成式大模型强大的泛化能力和通用性使其仍然备受关注，并具有广阔的应用前景。

2 AIGC 技术在各种生成式任务中的研究进展

ChatGPT 的横空出世以及 DeepSeek 模型的火爆，标志着基于 AIGC 技术的各类生成式任务取得了突破性进展。本节从文本生成、图像生成、语音生成和视频生成 4 个核心任务，对其进行详细阐述。

2.1 AIGC 在文本生成式任务中的研究进展

AIGC 技术在文本生成任务中的卓越进展使其在机器翻译、对话生成、创意写作

等应用方面取得显著成效。

(1) 机器翻译

机器翻译是指自动将文本从一种语言翻译成另一种语言的技术^[61-62]。随着深度学习逐渐取代基于规则^[63]和统计^[64-65]的方法，神经机器翻译（neural machine translation, NMT）凭借其对语言专业知识的最低要求^[66-67]和在捕获句子长距离依赖关系方面的优势^[68]，已成为机器翻译的主流方法。深度学习机器翻译的成功主要归功于语言模型的使用^[69]，该模型能够根据前面的单词预测下一个单词出现的概率。

Seq2Seq^[70]是将编码器-解码器 RNN 结构^[71]应用于机器翻译的开创性工作。然而，当句子很长时，Seq2Seq 的性能会下降。为了解决这个问题，Gehring 等^[72]提出了一种基于全卷积的 Seq2Seq 模型。虽然上述方法在文本翻译方面取得了不错的效果，但受限于模型架构以及只能根据前面的单词预测下一个单词的机制，导致翻译结果较为死板。随着 Transformer 模型架构的流行，利用预训练语言模型进行机器翻译成为可能。Rothe 等^[73]对使用预训练检查点进行序列生成进行了广泛研究，实验结果表明，使用 BERT^[74]或 RoBERTa^[75]初始化模型权重，能够显著提升从英语到德语的翻译性能。

虽然上述方法展现了较为出色的翻译水平，但很多翻译结果显得生硬，且过于依赖前文设定的条件，尤其在处理一些复杂长句时，难以准确表达句子的含义。因此，一些研究者尝试借助通用大型模型来解决长句翻译的问题。

随着 ChatGPT 在机器翻译领域展现出卓越的翻译能力，一些学者致力于利用生成式大模型使机器翻译更符合人类的逻辑和自然语言表达方式。例如，Dorosti 等^[76]提出了一种融合 Transformer 模型与具有

GPT 架构的波斯语言模型的方法，以解决从英语到波斯语翻译时数据不足的问题；此外，他们还将 CNN 模型与自编码器进行融合，以改善特征选择并降低维度。Chen 等^[77]提出了一种创新方法，通过上下文学习来增强 GPT-4 的翻译能力。该方法的核心思想是利用 TF-IDF 矩阵和余弦相似度分数构建一个检索器，实验结果表明，相较于不进行上下文学习的 GPT-4，通过上下文学习增强 GPT-4 翻译能力的方法翻译准确性更高。

近年来，机器翻译领域在多语言支持和偏见缓解方面取得了显著进展。例如，GemmaX2-28-9B 模型^[78]的发布标志着开源大型语言模型在多语言翻译中的突破。该模型支持 28 种语言，在 WMT-24 基准测试中表现出色。研究者通过引入并行第一单语第二（parallel-first monolingual-second, PFMS）数据混合策略，显著提升了低资源语言的翻译性能，表明了开源模型在实用性上的潜力。与此同时，微软研究团队^[79]提出了一种利用大语言模型缓解语音翻译中性别偏见的方法。他们发现现有系统在翻译女性说话者内容时存在男性偏见，通过修正训练数据并微调模型，可确保翻译尊重说话者性别身份。这一方法在减少不准确或冒犯性翻译方面显示出成效，尽管研究也指出上下文感知模型可能在某些情况下放大偏见，需进一步优化。

（2）知识利用

知识利用是智能系统完成基于支持性事实证据的知识密集型任务（例如常识问答和事实补全）的关键能力。具体来说，这就要求 LLM 能够有效地利用来自预训练语料库的丰富事实知识，或者在需要时检索外部数据。问答（QA）和知识补全一直是评估这种能力的两个常见任务。根据测试任务（问答或知识补全）和评估设置

（是否使用外部资源），可以将现有的知识利用任务划分为 3 类：封闭式问答（closed-book QA）、开放式问答（open-book QA）和知识补全。

封闭式问答^[80]任务评估 LLM 从预训练语料库中获取事实知识的能力。在这些任务中，LLM 仅根据给定的上下文回答问题，而不使用任何外部资源。为了评估这种能力，研究人员通常使用 Natural Questions^[81]、Web Questions^[82]和 TriviaQA^[83]等数据集，并采用准确率指标进行评估。实验证明，LLM 在封闭式问答任务中表现出色，甚至能够超越最先进的开放域问答系统^[84]。此外，LLM 在这些任务中展现出与模型和数据大小相关的扩展定律模式：增加模型参数和训练数据量，可以提升 LLM 的容量，使其从预训练数据中学习更多知识。然而，尽管封闭式问答任务能够帮助 LLM 获取更多知识，但在依赖于细粒度知识的问答任务上，LLM 的表现仍然较差。这是因为 LLM 在预训练过程中学习到的知识可能不够全面或不够精确，无法满足这类任务的需求。

2025 年，KidneyTalk Team^[85]提出了一种无须编码即可部署的私有 LLM，利用增强的医疗文档知识数据库提升封闭式问答能力。研究表明，该模型在肾脏疾病相关问答任务中，通过生成合成医疗问答数据提升性能，准确率从 74.2% 提升至 82.1%，显著弥补了细粒度知识的不足，在医疗领域的应用中表现突出。

与封闭式问答任务不同，在开放式问答任务中，LLM 可以从外部知识库或文档集合中提取有用的证据，然后基于提取的证据回答问题^[86-88]。典型的开放式问答数据集（例如 Natural Questions^[81]、OpenBookQA^[89]和 SQuAD^[90]）与封闭式问答数据集存在重叠，但它们整合了外部

数据源。在开放式问答任务的评估中，准确率和F1分数是常用的评估指标。

为了从外部资源中选择相关知识，LLM通常与文本检索器配对使用，该检索器可以独立训练，也可以与LLM联合训练^[86,91]。评估主要关注LLM如何利用提取的知识回答问题，研究表明，检索到的证据能够显著提高生成答案的准确性。然而，尽管开放式问答任务能够帮助LLM更快、更准确地生成答案，但如果LLM用过时的知识资源进行预训练或检索，可能会导致对时效性敏感问题生成错误的答案。这是因为外部知识库可能无法及时更新，导致LLM无法获取最新的信息。

在知识补全任务中，LLM在某种程度上被视为知识库^[92]，具备填补或预测知识单元缺失部分的能力。这些任务的目标是深入研究和评估LLM从预训练数据中学到了多少知识以及学到了哪些类型的知识。现有的知识补全任务大致可以分为知识图谱补全任务（如FB15k-237^[93]和WN18-RR^[94]）和事实补全任务（如WikiFact^[95]），分别旨在补全知识图谱中的三元组和关于特定事实的不完整句子。实证研究表明，现有的LLM在完成涉及特定关系类型的知识补全方面存在困难^[96]。例如，在WikiFact的评估结果中，LLM在一些预训练数据中频繁出现的关系上表现良好，而在一些罕见的关系上表现较差。有趣的是，在相同的评估设置下，InstructGPT在WikiFact的所有子集中的表现均优于GPT-3。

虽然基于知识补全任务的LLM在频繁出现的数据上表现良好，但在一些较少见或罕见的数据上，这类模型的表现不佳，这可能会导致整体补全效果较差。这是因为LLM的知识主要来自预训练数据，如果某些知识在预训练数据中出现频率较低，

LLM就难以有效地学习和掌握这些知识。

（3）生成写作

AIGC技术在生成式写作方面的应用使机器能够独立进行文章创作。如今，借助AIGC技术，机器可以自动生成新闻稿、诗歌等多种类型的文本。例如，智能写作算法能够从结构化数据中提取关键信息并生成新闻报道。由于机器的处理速度远超人类，AIGC在生成具有时效性的新闻方面更具优势。

在撰写稿件方面，AIGC能够有效地挖掘相关信息^[97]。它可以从大量材料中选择合适的内容，并通过内容分析提取关键信息，最终按照写作逻辑组织成文。此外，AIGC还能生成符合特定格式要求的创意作品，例如诗歌和对联^[98]。由于模型的输出具有不可预测性，AIGC在创造性方面与人类写作颇为相似。例如，如果我们想用AIGC创作古诗，可以输入大量的古诗作为训练数据来训练模型。

尽管AIGC能够轻松地创建高质量的文本内容，但其创作的内容仍然缺乏深度思考，无法生成具有创新性、情感共鸣和复杂逻辑思维的文本。在这个领域，AIGC暂时还无法取代人类在写作实践中的主观能动性。

2.2 AIGC在图像生成式任务中的研究进展

AIGC技术在图像生成任务方面取得的显著进展，使其在图像编辑、图像修复、多模态图像生成等应用领域大放异彩。

（1）图像编辑

图像编辑是指根据特定需求修改图像，例如风格迁移。从技术角度来看，一些图像恢复任务（如图像上色）也可以归类为图像编辑，因为上色可以被视为一种特定需求^[99]。早期的工作主要采用基于优化的

方法^[100-101]，这些方法建立在预训练的CNN编码器的基础上，但其迭代性质导致处理时间较长。为了克服对属性注释标签的依赖，研究者们引入了无监督学习来解耦不同的属性^[102]。

另一类图像编辑则通过合并两张图像来改变图像语义。例如，图像变形^[103]对两张图像的内容进行插值，而风格迁移^[104]则生成一张新图像，其中内容来自一张图像，风格来自另一张图像。图像变形的一种常见方法是在像素空间进行插值，但这容易产生伪影。相比之下，在潜在空间进行插值可以更好地考虑视图变化，生成更平滑的图像。这些潜在空间表示通常通过GAN反演方法^[105]获得。许多研究^[106-107]探索了利用预训练GAN的潜在空间进行图像变形的方法。对于风格迁移任务，StyleGAN及其变种^[108]已成为解决风格迁移问题的主流方法。虽然上述方法能够快速生成图像，但生成的图像质量不够高，且容易产生伪影。因此，需要使用泛化能力和通用性更强的大模型来生成更高质量的图像。

随着GPT模型的兴起，越来越多的学者开始使用GPT模型进行图像编辑。例如，Brooks等^[109]提出了一种根据人类指令编辑图像的模型InstructPix2Pix。为了使模型能够理解人类指令并据此编辑图像，他们结合了两个大型预训练模型（GPT-3和Stable Diffusion^[121]），并生成了一个大规模的图像编辑示例数据集。然后，他们使用InstructPix2Pix在生成的数据上进行训练，使其能够在推理时泛化到真实图像和用户指令。实验结果表明，该模型可以在几秒钟内快速编辑图像，并生成高质量的编辑结果。Fu等^[110]则发现，有时人类的指令对现有研究方法来说过于简略，导致模型难以捕捉或遵循。因此，他们提出了一种基于多模态大语言模型

（multimodal LLM，MLLM）的图像编辑方法，该方法能够提取富有表现力的指令并提供明确的引导。实验结果表明，该模型能够生成高质量的图像编辑结果。Huang等^[111]提出了SmartEdit，这是一种将LLM与视觉输入相结合的新方法，旨在增强基于指令的图像编辑的理解和推理能力。实验结果表明，SmartEdit在处理感知数据和复杂指令的图像编辑方面表现出色。

近期研究进一步扩展了LLM在三维场景编辑中的应用。Kathare等^[112]提出的Instructive3D模型，创新性地将文本指令与大型重建模型（large reconstruction model，LRM）结合，实现了3D对象的细粒度编辑。该方法通过在三维潜在空间中引入适配器模块，允许用户通过文本提示直接修改3D模型的几何与材质属性（如颜色、反射率），而无须重新生成完整的3D对象。实验表明，Instructive3D在Objaverse LVIS数据集上生成的编辑结果，在几何一致性和细节精度上均优于传统两步法（先生成3D模型再编辑）。这一进展为增强现实、动画制作等领域提供了更高效的3D内容创作工具。

虽然上述方法证明了LLM在图像编辑方面的可行性，但要实现用户随心所欲地编辑图像的目标，还有很长的路要走。

（2）图像恢复

图像恢复旨在从退化图像恢复清晰图像。退化图像到清晰图像之间存在无数种可能的映射关系，因此这是一个复杂的反问题。图像退化主要源于两个方面：原始图像信息缺失和对清晰图像添加不良扰动。随着GAN及其变体的流行，基于GAN的方法在图像恢复领域得到了广泛应用^[113-114]。

近期研究通过结合扩散模型与大规模

预训练技术，显著提升了图像恢复的可控性与真实感。董超团队提出的 SUPIR^[115]创新性地 将 SDXL (StableDiffusion-XL) 与 130 亿参数量的 MLLM 结合，通过设计 6 亿参数量的适配器模块，在 2 000 万高质量图像数据上训练，实现了复杂场景下的超分辨率、去噪和语义引导修复。该方法引入负样本和负提示机制，增强了模型对低质量图像的理解，在无参考指标（如感知质量）上显著优于传统方法。实验表明，SUPIR 在真实场景图像恢复中能有效保留纹理细节，并支持文本引导的可控修复（如调整材质、增强特定对象）。

此外，扩散模型在 3D 场景恢复中也取得突破。罗格斯大学团队提出的 ScoreHMR^[116]将扩散模型作为先验，通过得分引导的迭代优化，实现了单目/多视角图像中的 3D 人体姿态与形状恢复。该方法通过去噪扩散隐式模型（denoising diffusion implicit model, DDIM）采样和循环优化，在保持与图像观测对齐的同时，提升了 3D 重建的准确性和时间连贯性，在 Human3.6M、MPI-INF-3DHP 等数据集上的表现超越传统优化方法，尤其在多视角融合和视频序列重建中表现突出。

尽管上述基于扩散模型的方法能够实现高质量的图像恢复，但对于一些特定的特征细节修复效果并不理想，只能处理常规情况下的图像恢复任务，并且计算成本较高。因此，采用通用的 LLM 来解决上述问题已成为一种趋势。

随着 GPT 模型在图像生成领域的广泛应用，许多研究人员开始利用 GPT 模型进行图像恢复。例如，Wei 等^[117]指出先前的图像恢复方法缺乏交互性机制且泛化能力较差。为了解决这些问题，他们提出了一种名为 Clarity ChatGPT 的方法，该方法将 ChatGPT 的对话智能与多种图像编辑方

法相结合，能够自动检测图像退化类型，并选择合适的编辑方法来恢复图像，或者根据用户反馈迭代生成令人满意的结果。Lu 等^[118]则发现现有的超分辨率方法存在两个问题：未充分利用有价值的语义信息来提升超分辨率质量，以及在面对图像退化时使用的先验知识不够准确。为了解决这些问题，他们提出了一种基于 SAM^[119]大模型的更强大且具有语义感知先验的方法，该方法集成了 SAM 引导细化模块（SAM-guided refinement module, SEEM），实现了更好的图像对齐和融合效果。

尽管上述基于大型模型的方法在图像恢复方面表现出色，但在特定场景下的图像恢复效果仍有待提高，例如在医学图像恢复等领域。这表明当前的人工智能生成式大模型与人们对图像恢复的要求还有一定的差距。尽管如此，生成式大模型的泛化能力和交互性使其在图像恢复领域仍然备受关注，并具有广阔的应用前景。

（3）多模态图像生成

多模态图像生成是一项涉及多个感官模态（如视觉、文本、音频等）的复合任务，其目标是生成能够满足多种感官模态需求的图像。这项研究旨在通过整合不同感官信息，提高生成图像的质量、多样性和真实性。多模态图像生成可以根据引导方式划分为文本引导、音频引导、视觉引导和其他引导等类别。

文本引导图像生成是一种基于文本描述生成图像的方法，其目标是根据自然语言描述生成与之相符的图像。这一领域的研究旨在实现从文本到图像的有意义的转换，使计算机能够理解并呈现文本中描述的视觉概念。常见的文本引导图像生成方法通常使用 GAN 模型^[120-121]、扩散模型^[122-123]等深度学习架构。这些模型通过

学习从文本到图像的映射关系，生成高质量、逼真的图像。

然而，这些方法生成的图像较为简单，这是因为它们的输入文本通常缺乏复杂的逻辑，导致生成的图像背景单一，只适用于特定情境。为了解决这个问题，一些学者开始研究如何将多模态信息与LLM结合起来，以满足人们对更复杂、更具通用性的图像生成的需求。例如，Zhong等^[124]提出了一种名为语义理解与推理适配器（SUR-adapter）的方法，该方法融合了LLM和扩散模型，能够理解复杂的叙述内容，并生成高质量且内容丰富的图像。Wen等^[125]则提出了一种大型视觉语言模型（LVLM），用于解决输入文本与生成图像的对齐问题，有效提升了图像生成质量。Brade等^[126]提出了一种名为Promptify的交互式系统，旨在帮助用户以交互方式探索和创建用于文本生成图像的文本提示。

随着MLLM的发展，研究人员进一步探索了其在图像生成中的潜力。例如，Wu等^[127]提出的DiffSensei框架，创新性地将MLLM与扩散模型结合，实现了定制化漫画的生成。实验表明，DiffSensei在角色一致性、文本适配性和图像质量上显著优于现有方法（如StoryDiffusion^[128]），其FID指标在MangaZero数据集上达到0.407，对比语言-图像预训练（contrastive language-image pre-training, CLIP）相似度提升至0.235^[127]。虽然在某些特定任务上，这些LLM的效果可能不尽如人意，但与传统的文本引导图像生成方法相比，它们生成的图像内容更丰富，并具有更强的泛化能力。

视觉引导图像生成利用特定的图像作为引导，通过像素级别的细节传达图像的空间和结构特征。这种引导方式为图像生成提供了极高的控制度，使用户能够进行

交互式操控和精准处理，从而在图像生成过程中实现更精确和个性化的结果。

常见的视觉引导类型包括分割特征图^[129-130]、关键点^[131-132]、场景布局^[133-134]等。这些引导类型在像素空间中承载特定的图像属性，为生成过程提供丰富的信息。此外，一些研究还探索了基于深度图^[135]、法线图^[135]、轨迹图^[136]等条件的图像合成。

近年来，参考图像引导技术取得显著突破，如He等^[137]提出的Imagine Yourself模型通过三大创新实现了无需微调的个性化图像生成：采用可训练CLIP ViT-H视觉编码器提取参考图像身份特征，并通过掩码视觉嵌入聚焦面部区域以避免背景干扰，设计并行注意力机制融合CLIP、UL2、ByT5 3个文本编码器与视觉编码器特征，以动态平衡文本与视觉引导权重，从而提升文本对齐精度，构建合成配对数据（SynPairs），通过MLLM生成多样化训练数据以缓解参考图像不足的问题，并增强模型对复杂提示（如表情、姿势变化）的适应性。实验表明，该模型在文本对齐（+27.8%）、视觉吸引力（+31.6%）等指标上显著优于IP-Adapter^[138]、InstantID^[139]等现有方法，且不需要针对每个用户进行模型微调。

视觉引导可以通过多种方式进行，例如使用预训练模型（如分割模型、深度预测器、姿势预测器）、应用算法（如Canny边缘检测）以及手动标注和涂鸦。这种多样化的引导方式使视觉引导图像生成技术具有广泛的适用性和灵活性。视觉引导图像生成不仅在图像合成领域具有突出贡献，而且通过对引导元素的调整，还为各种图像编辑任务提供了一种直观而强大的方法。

音频引导图像生成是一种基于语音描述生成图像的方法。与文本和视觉引导不同，音频引导提供了时间信息，这在生成

动态或序列化视觉内容时非常有用。通常情况下，音频信号与图像之间的关系^[140]比文本或视觉引导图像生成更抽象。例如，与某些动作或环境相关联的音频可能更倾向于提供启示，而不是明确定义视觉内容^[141]；声音可以传达情感色彩，并根据音频的情感色彩生成相应的视觉内容。

音频引导图像生成在说话人脸生成领域得到了广泛应用^[142]。例如，Tan 等^[143]提出了一种新颖的情感活动记忆网络（EMMN），用于生成情感说话人脸，该方法不仅能够生成准确的嘴唇运动，还能够生成面部表情动态。Stypułkowski 等^[144]提出了一种自回归扩散模型，只需一个身份图像和音频序列就能生成逼真的说话头部视频。Zhong 等^[145]提出了一种两阶段通用个体方法，用于基于音频的说话人脸生成。该方法首先从音频中获取准确的嘴唇和下颌关键点，然后将多个参考图像与目标表情和姿势进行对齐，并使用声学特征增强嘴唇同步效果，最终生成更逼真、同步且保持身份的说话人脸视频。

随着扩散模型的发展，研究人员进一步突破传统方法的限制，提出了更先进的音频驱动人像动画技术。

- EchoMimic 框架^[146]：Cao 等创新性地结合音频和面部地标双模态输入，通过多模态融合技术（支持单模态或联合驱动生成）、空间损失与音频增强（直接学习像素级面部结构并提升鲁棒性）以及部分感知运动同步（分区域对齐解决全脸变换失真问题），在 HDTF、CelebV-HQ 等数据集上的效果显著优于现有方法，FID 达到 29.136、SSIM 达到 0.812。

- LOOPY 模型^[147]：Jiang 等提出端到端音频驱动扩散模型，通过跨剪辑与剪辑内时间模块（扩展时间感受野至 100 帧以上）和音频到潜空间模块（增强音频与

运动关联），在无约束条件下生成自然动态的说话人视频，在 Openset 测试集上主观评分显著优于现有方法，标志着音频驱动人像动画向更自由、更自然的方向迈出重要一步。

除了上述引导方式，还有一些其他类型的引导方式^[148]，例如草图引导、图形引导、类别引导等。这些引导方式各有特点，可以根据不同的应用场景选择合适的引导方式。总而言之，多模态图像生成是一个充满活力的研究领域，它将不同感官模态的信息整合到图像生成过程，为生成更丰富、更逼真、更具表现力的图像提供了新的可能性。

2.3 AIGC 在语音生成式任务中的研究进展

AIGC 在语音生成任务上的研究涵盖了多个领域，其中语音合成是重要研究方向之一，旨在使机器能够根据文本生成自然流畅、易于理解的语音^[149]。传统的共振峰合成仍然在开源 NVDA 屏幕阅读器中使用，然而这些方法生成的语音不够自然，并且存在一些瑕疵，影响了语音的可懂度。

早期的语音合成研究工作主要包括 3 个模块：文本分析、声学模型和声码器。WaveNet^[150]的出现为语音合成领域带来了一场革命，它能够直接从语言特征生成原始波形，显著提升了合成语音的质量。为了进一步提升语音质量和声音的多样性，研究者引入了生成模型，如 GAN。然而，GAN 的训练过程不稳定。相比之下，扩散模型不需要判别器，训练更加稳定和简单，因此在语音合成领域逐渐流行起来。

近期研究显著拓展了语音生成的边界，涵盖多模态融合、情感生成、人类反馈优化以及高效低资源合成等方向。在多模态驱动合成方面，Coron 等^[151]提出的

VLOGGER 结合单张图像和音频输入，利用扩散模型生成情感丰富且动作协调的虚拟化身视频。该方法通过音频驱动 3D 人体运动参数生成头部动作、表情和手势，并结合扩散模型生成时空一致的高质量视频，为增强现实、在线教育等场景提供了高效解决方案。与此同时，ICAGC 2024 技术报告^[152]展示了情感语音生成的最新进展，其挑战赛要求模型在克隆说话人身份的基础上调制情感表达（如小说朗诵、古诗演绎），并通过背景音效融合增强音频的真实性。在语音合成自然度与一致性优化方面，复旦大学提出的 SpeechAlign^[153]创新性地构建了基于人类反馈的迭代优化框架。该方法通过对比真实语音的黄金编解码令牌与模型生成的合成令牌，构建偏好数据集，并结合强化学习（RLHF-PPO）、直接偏好优化（DPO）等策略，迭代校准模型输出分布。实验表明，SpeechAlign 在 LibriSpeech 和 VCTK 数据集上显著降低了词错误率（WER）并提升了说话人相似度（SIM），其中 DPO-Iter3 将 WER 从基线的 7.2 降至 6.0，SIM 从 0.87 提升至 0.90。通过人类验证（71% 胜率）和 t-SNE 可视化分析，该框架有效弥合了训练与推理阶段的分布差距，实现了语音生成质量的持续自我提升，为神经编解码器语言模型的偏好对齐提供了新范式。此外，为实现高效与低资源语音合成，部分研究^[154]通过计划采样、得分网络训练和联合训练 GAN 等方法加速推理过程并缩短生成时间，而另一些研究^[155]则聚焦端到端模型，直接从文本生成波形，简化训练推理流程并减少对人工标注数据的依赖。这些进展共同推动了语音生成技术向更高效、更自然、更具情感表现力的方向迈进。

总而言之，语音合成算法研究正朝着深度学习和生成模型的方向不断发展，以

期提高生成语音的自然度、多样性和效率。

2.4 AIGC 在视频生成式任务中的研究进展

相较于图像生成领域的飞速发展，视频生成的技术进步显得相对缓慢，这主要归因于视频数据的高维度特性，使建模过程更复杂。视频生成不仅需要生成逼真的像素级画面，还需确保帧间语义的连贯性和动态一致性。当前的视频生成研究主要分为无引导生成和有引导生成两大类，其中文本引导的视频生成方法因其灵活性和应用潜力备受瞩目。近年来，以扩散模型为基础的 AIGC 技术在视频生成领域取得了突破性进展，显著提升了生成质量和时序一致性。

早期研究尝试将图像生成技术从单帧扩展到多帧，例如基于 VAE 或 GAN 的方法^[156-157]，但生成的视频内容往往比较单调且重复，如海浪等动态纹理^[158]，其空间模式重复，时间变化有限。这些方法局限于简单场景的低分辨率短视频生成，难以满足复杂场景的需求。随着生成式模型的进步，部分研究^[159-161]开始探索更真实的视频生成，如通过改进 GAN 或 VAE 生成具有一定动态变化的视频，但仍受限于分辨率和场景复杂度。近年来，扩散模型的引入显著推动了视频生成技术的发展。例如，Mei 等^[162]基于扩散模型的视频生成方法在质量和一致性上取得了显著成果，但文本引导的视频生成仍面临挑战，包括对长序列的建模能力不足和对复杂动态的控制比较困难。

2024 年，Song 等^[163]提出的 DFoT（diffusion forcing transformer）为视频生成带来了新的突破。DFoT 是一种基于扩散模型的视频生成框架，通过引入灵活的历史帧条件（history conditioning），克服了传统视频扩散模型（如 DiT^[164]）仅

支持固定长度条件的限制。DFoT采用噪声即掩码（noise-as-masking）的训练范式，对每帧独立施加不同噪声水平，使模型能够适应任意长度的历史帧输入。这种方法不仅提升了生成视频的质量，还通过自回归扩展实现了超长视频生成，例如从单帧图像生成862帧的导航视频，远超训练集中的最大视频长度^[163]。与此同时，DFoT支持的历史引导（history guidance, HG）方法进一步增强了生成效果。HG包括基础的香草历史引导（HG-v）和更高级的时间与频率历史引导（HG-tf）。HG-v通过对任意长度的历史帧进行分类器自由引导，显著提高了视频质量和时序一致性；HG-tf则通过组合不同历史窗口的得分和部分噪声掩码历史帧，增强了动态表现，并实现了对分布外（out-of-distribution, OOD）历史的鲁棒性^[163]。

相比之下，早期基于VAE或GAN的文本引导视频生成方法^[156-157]在复杂场景和高分辨率视频生成中表现不足，而基于扩散模型的工作（如Stable Video Diffusion^[165]）虽然改进了质量，但仍受限于固定条件帧数。DFoT通过其灵活性和历史引导策略，不仅在标准数据集（如Kinetics-600）上取得了与工业级模型（如W.A.L.T^[166]）媲美的性能，还在长视频生成和动态控制上展现了新的可能性。此外，研究表明，DFoT可以通过微调现有视频扩散模型高效实现，使其具备历史引导能力，为未来将HG整合进大型基础模型提供了可能。

3 总结与展望

AIGC技术近年来发展迅猛，其应用已渗透到文本、图像、音频、视频等多个领

域。然而，AIGC在发展过程中也面临着一系列挑战。

- 可解释性不足：尽管AIGC模型可以生成令人惊艳的作品，但其生成过程如同一个“黑匣子”，难以理解。特别是当模型生成不良内容时，这种缺乏可解释性使问题溯源和控制变得异常困难。

- 创造性与创新性有限：目前的AIGC技术大多依赖于对现有数据的学习和模仿，缺乏真正的创造力和创新性。生成的内容往往是已有信息的重组或变体，难以突破限制，产生真正原创的内容。

- 理解与推理能力欠缺：AIGC模型在理解语境、进行逻辑推理方面仍有待提升。这导致生成的内容可能缺乏深度和逻辑性，难以胜任复杂的分析和推理任务。

- 数据偏见与不公平性：AIGC模型的训练依赖于大规模数据集，而这些数据可能存在着社会、文化、性别等方面的偏见和不公平。模型在学习过程中可能会吸收这些偏见，进而生成带有歧视性的内容，加剧社会不平等。

未来的研究应致力于提高AIGC模型的透明度，例如引入可解释性模块，揭示模型生成内容的机制和原理。这有助于增强模型的可信度和安全性，并方便用户进行干预和控制。AIGC技术的未来发展方向应更加注重培养模型的创造力和创新性，使其能够生成真正新颖、独特、具有艺术价值的内容，而不仅仅是对现有信息的简单模仿。研究人员需要不断改进AIGC模型的语境理解和逻辑推理能力，使其能够更好地理解用户的意图，进行更深层次的分析和推理，生成更具逻辑性和深度的内容。为了减少AIGC模型中的偏见和不公平性，需要改进训练数据、优化算法，并加强对模型的伦理审查。未来的AIGC技术发展应更加注重社会责任和公平性，避

免生成带有歧视性的内容。总而言之，AIGC 技术在发展过程中面临诸多挑战，但同时也拥有广阔的应用前景。相信通过持续的研究和创新，AIGC 技术将在各个领域发挥更大的作用，为人类社会带来更多福祉。

参考文献：

- [1] Huang D J, Kao Y T, Chuang T H, et al. SB-VQA: a stack-based video quality assessment framework for video enhancement[C]//Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Piscataway: IEEE Press, 2023: 1613-1622.
- [2] Lin Y J, Gao Z P, Du H Y, et al. Blockchain-aided AI-generated content services: Stackelberg game-based content caching approach[C]//Proceedings of the 2023 IEEE International Conference on Web Services. Piscataway: IEEE Press, 2023: 186-195.
- [3] Wu H N, Zhang E L, Liao L, et al. Exploring video quality assessment on user generated contents from aesthetic and technical perspectives[C]//Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2023: 20087-20097.
- [4] Advani M, Gokhale N. Influence of brand-related user generated content (UGC) and brand engagement on Instagram[C]//Proceedings of the AIP Conference. [S.l.]: AIP Publishing, 2023.
- [5] 李祎恒, 张峥. AIGC 独创性标准的构建[J]. 大数据, 2025, 11(1): 36-44.
Li Y H, Zhang Z. Construction of the originality criterion for AIGC[J]. Big Data Research, 2025, 11(1): 36-44.
- [6] Hao Y, Liu Y, Mou L. Teacher forcing recovers reward functions for text generation[J]. Advances in Neural Information Processing Systems, 2022, 35: 12594-12607.
- [7] Avrahami O, Hayes T, Gafni O, et al. SpaText: Spatio-Textual representation for controllable image generation[C]//Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2023: 18370-18380.
- [8] Yang J K, Peng W X, Li X T, et al. Panoptic video scene graph generation [C]//Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2023: 18675-18685.
- [9] Pascual S, Bhattacharya G, Yeh C, et al. Full-band general audio synthesis with score-based diffusion[C]// Proceedings of the 2023 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE Press, 2023: 1-5.
- [10] Pan X G, Tewari A, Leimkühler T, et al. Drag your GAN: interactive point-based manipulation on the generative image manifold[C]//Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Proceedings. New York: ACM, 2023: 1-11.
- [11] Yang G X, Fu F F, Fei N Y, et al. DiST-GAN: distillation-based semantic transfer for text-guided face generation[C]// Proceedings of the 2023 IEEE International Conference on Multimedia and Expo. Piscataway: IEEE Press, 2023: 840-845.
- [12] Kumari N, Zhang B L, Wang S Y, et al. Ablating concepts in text-to-image dif-

- fusion models[C]//Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2023: 22634–22645.
- [13] Wu Z J, Wang Y N, Feng M T, et al. Sketch and text guided diffusion model for colored point cloud generation[C]//Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2023: 8895–8905.
- [14] Rumelhart D E, Hinton G E, Williams R J. Learning representations by back-propagating errors[J]. *Nature*, 1986, 323: 533–536.
- [15] Kingma D P, Welling M. Auto-encoding variational bayes[Z]. 2013.
- [16] Bi X, Chen D L, Chen G T, et al. Deep-Seek LLM: scaling open-source language models with longtermism[EB]. arXiv preprint, 2024, arXiv: 2401.02954.
- [17] OpenAI. GPT-4 technical report[EB]. arXiv preprint, 2023, arXiv:2303.08774.
- [18] Ramesh A, Dhariwal P, Nichol A, et al. Hierarchical text-conditional image generation with CLIP latents[EB]. arXiv preprint, 2023, arXiv: 2204.06125.
- [19] 施敏, 杨海军. 生成式人工智能的算法伦理难点分析与探索[J]. *大数据*, 2025, 11(1): 167–174.
- Shi M, Yang H J. Analysis and exploration of algorithmic ethical difficulties in generative artificial intelligence[J]. *Big Data Research*, 2025, 11(1): 167–174.
- [20] Wei X, Cui X Y, Cheng N, et al. ChatIE: zero-shot information extraction *via* chatting with ChatGPT[EB]. arXiv preprint, 2023, arXiv: 2302.10205.
- [21] Rombach R, Blattmann A, Lorenz D, et al. High-resolution image synthesis with latent diffusion models[C]//Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2022: 10674–10685.
- [22] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets[C]//Advances in neural Information Processing Systems. [S.l.:s.n.], 2014.
- [23] Joshi I, Grimmer M, Rathgeb C, et al. Synthetic data in human analysis: a survey[EB]. arXiv preprint, 2022, arXiv: 2208.09191.
- [24] 司亚中, 张旭龙, 杨帆, 等. 基于生成对抗网络的多特征融合去雾技术[J]. *大数据*, 2024, 10(4): 77–88.
- Si Y Z, Zhang X L, Yang F, et al. Multiple-feature fusion based generative adversarial network for image dehazing[J]. *Big Data Research*, 2024, 10(4): 77–88.
- [25] Deng K L, Yang G S, Ramanan D, et al. 3D-aware conditional image synthesis [C]//Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2023: 4434–4445.
- [26] Yin F, Zhang Y, Wang X, et al. 3d gan inversion with facial symmetry prior[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.:s.n.], 2023: 342–351.
- [27] Xu X Q, Navasardyan S, Tadevo-Syan V, et al. Image completion with heterogeneously filtered spectral hints[C]//Proceedings of the 2023 IEEE/CVF Winter Conference on Applications of Computer Vision. Piscataway: IEEE Press, 2023: 4580–4590.
- [28] Jain J, Zhou Y Q, Yu N, et al. Keys to better image inpainting: structure and texture go hand in hand[C]//Proceedings of the 2023 IEEE/CVF Winter Conference on Applications of Computer Vi-

- sion. Piscataway: IEEE Press, 2023: 208–217.
- [29] Kanagawa H, Ijima Y. Enhancement of text-predicting style token with generative adversarial network for expressive speech synthesis[C]//Proceedings of the ICASSP 2023 – 2023 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE Press, 2023: 1–5.
- [30] Qu Y D, Tan Q F, Xie H T, et al. Exploring stroke-level modifications for scene text editing[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2023, 37(2): 2119–2127.
- [31] Yoneyama R, Wu Y C, Toda T. Source-filter HiFi-GAN: fast and pitch controllable high-fidelity neural vocoder[C]//Proceedings of the ICASSP 2023 – 2023 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE Press, 2023: 1–5.
- [32] Xu L, Song Z Y, Wang D L, et al. ActFormer: a GAN-based transformer towards general action-conditioned 3D human motion generation[C]//Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2023: 2228–2238.
- [33] Xu Y Y, He S F, Wong K K, et al. RIGID: recurrent GAN inversion and editing of real face videos[C]//Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2023: 13645–13655.
- [34] Melechovsky J, Mehrish A, Sisman B, et al. Accent conversion in text-to-speech using multi-level VAE and adversarial training[EB]. arXiv preprint, 2024, arXiv: 2406.01018.
- [35] Chen X Y, Wang X, Zhang S F, et al. StyleSpeech: self-supervised style enhancing with VQ-VAE-based pre-training for expressive audiobook speech synthesis[EB]. arXiv preprint, 2023, arXiv: 2312.12181.
- [36] Li Y, Yu C, Sun G Z, et al. Cross-utterance conditioned VAE for speech generation[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2024, 32: 4263–4276.
- [37] Xiang Y, Tian J G, Hu X H, et al. A deep representation learning-based speech enhancement method using complex convolution recurrent variational autoencoder[C]//Proceedings of the 2024 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE Press, 2024: 781–785.
- [38] Zhao S, Zhang Y, Cun X, et al. CV-VAE: a compatible video VAE for latent generative video models[J]. Advances in Neural Information Processing Systems, 2025, 37: 12847–12871.
- [39] Pucci R, Micheloni C, Foresti G L, et al. CVGAN: image generation with capsule vector-VAE[M]//Image Analysis and Processing – ICIAP 2022. Cham: Springer, 2022: 536–547.
- [40] Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models[C]//Advances in Neural Information Processing Systems. [S.l.:s.n.], 2020, 33: 6840–6851.
- [41] Kumar Bhunia A, Khan S, Cholakal H, et al. Person image synthesis *via* denoising diffusion model[C]//Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2023: 5968–5976.
- [42] Blattmann A, Rombach R, Ling H, et al. Align your latents: high-resolution video synthesis with latent diffusion

- models[C]//Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2023: 22563–22575.
- [43] He Z F, Sun T X, Tang Q, et al. DiffusionBERT: improving generative masked language models with diffusion models[C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics. Stroudsburg: USAACL, 2023: 4521–4534.
- [44] Lin Z W, Gong Y Y, Shen Y L, et al. Text generation with diffusion language models: a pre-training approach with continuous paragraph denoise[C]//Proceedings of the International Conference on Machine Learning. [S.l.: s.n.], 2023: 21051–21064.
- [45] Takahashi N, Kumar M, Singh, et al. Hierarchical diffusion models for singing voice neural vocoder[C]//Proceedings of the 2023 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE Press, 2023: 1–5.
- [46] Yamamoto R, Yoneyama R, Toda T. NNSVS: a neural network-based singing voice synthesis toolkit[C]//Proceedings of the 2023 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE Press, 2023: 1–5.
- [47] Yang L, Zhang Z L, Song Y, et al. Diffusion models: a comprehensive survey of methods and applications[J]. ACM Computing Surveys, 2024, 56(4): 1–39.
- [48] Cui Y M, Che W X, Liu T, et al. Revisiting pre-trained models for Chinese natural language processing[C]//Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2020. Stroudsburg: USAACL, 2020: 657–668.
- [49] Nguyen X B, Duong C N, Li X, et al. Micron-BERT: BERT-based facial micro-expression recognition[C]//Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2023: 1482–1492.
- [50] Ravi S, Chinchure A, Sigal L, et al. VLC-BERT: visual question answering with contextualized commonsense knowledge[C]//Proceedings of the 2023 IEEE/CVF Winter Conference on Applications of Computer Vision. Piscataway: IEEE Press, 2023: 1155–1165.
- [51] Radford A, Narasimhan K, Salimans T, et al. Improving language understanding by generative pre-training[Z]. 2018.
- [52] Radford A, Wu J, Child R, et al. Language models are unsupervised multi-task learners[J]. OpenAI Blog, 2019, 1(8): 9.
- [53] Brown T B, Mann B, Ryder N, et al. Language models are few-shot learners[C]//Proceedings of the 34th International Conference on Neural Information Processing Systems. New York: ACM, 2020: 1877–1901.
- [54] Liu A, Feng B, Xue B, et al. DeepSeek-v3 technical report[EB]. arXiv preprint, 2024, arXiv:2412.19437.
- [55] Liu Z, Janowicz K, Currierk, et al. Uncovering regional defaults from photo-realistic forests in text-to-image generation with DALL-E 2[EB]. arXiv preprint, 2024, arXiv:2410.17255.
- [56] Bai Y T, Geng X Y, Mangalam K, et al. Sequential modeling enables scalable learning for large vision models[EB]. arXiv preprint, 2023, arXiv:2312.00785.
- [57] Wang C Y, Chen S Y, Wu Y, et al. Neu-

- ral codec language models are zero-shot text to speech synthesizers[EB]. arXiv preprint, 2023, arXiv: 2301.02111.
- [58] Rubenstein P K, Asawaroengc-Hai C, Nguyen D D, et al. AudioPaLM: a large language model that can speak and listen [EB]. arXiv preprint, 2023, arXiv: 2306.12925.
- [59] Anil R, Dai A M, Firat O, et al. PaLM 2 technical report[EB]. arXiv preprint, 2023, arXiv: 2305.10403.
- [60] 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, 2023, 31: 2523–2533.
- [61] Yang S H, Wang Y X, Chu X W. A survey of deep learning techniques for neural machine translation[EB]. arXiv preprint, 2020, arXiv: 2002.07526.
- [62] Duan S F, Zhao H, Zhang D D. Syntax-aware data augmentation for neural machine translation[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2023, 31: 2988–2999.
- [63] Sethi N, Dev A, Bansal P. A novel neural machine translation approach for low-resource Sanskrit-Hindi language pair[J]. ACM Transactions on Asian and Low-Resource Language Information Processing, 2023: 3591207.
- [64] Ranathunga S, Lee E A, Prifti Skenduli M, et al. Neural machine translation for low-resource languages: a survey[J]. ACM Computing Surveys, 2023, 55(11): 1–37.
- [65] Tonja A L, Kolesnikova O, Gelbukh A, et al. Low-resource neural machine translation improvement using source-side monolingual data[J]. Applied Sciences, 2023, 13(2): 1201.
- [66] Li Y C, Li J H, Jiang J, et al. P-transformer: towards better document-to-document neural machine translation [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2023, 31: 3859–3870.
- [67] Gupta D, Kharbanda S, Zhou J W, et al. CLIPTrans: transferring visual knowledge with pre-trained models for multi-modal machine translation[C]//Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2023: 2863–2874.
- [68] Girdhar R, El-Nouby A, Liu Z, et al. ImageBind one embedding space to bind them all[C]//Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2023: 15180–15190.
- [69] Han S, Shin M, Park S, et al. Unified neural topic model *via* contrastive learning and term weighting[C]//Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics. Stroudsburg: USAACL, 2023: 1794–1809.
- [70] Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks[J]. Advances in Neural Information Processing Systems, 2014, 4 (January): 3104–3112.
- [71] Kalchbrenner N, Blunsom P. Recurrent continuous translation models[C]//Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: USAACL, 2013: 1700–1709.
- [72] Gehring J, Auli M, Grangier D, et al. Convolutional sequence to sequence learning[EB]. arXiv preprint, 2017, arXiv: 1705.03122.
- [73] Rothe S, Narayan S, Severyn A. Lever -

- aging pre-trained checkpoints for sequence generation tasks[J]. Transactions of the Association for Computational Linguistics, 2020, 8: 264–280.
- [74] Devlin J, Chang M W, Lee K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[EB]. arXiv preprint, 2018, arXiv: 1810.04805.
- [75] Liu Y H, Ott M, Goyal N, et al. RoBERTa: a robustly optimized BERT pretraining approach[EB]. arXiv preprint, 2019, arXiv: 1907.11692.
- [76] Dorosti M, Pedram M M. Improving English to Persian machine translation with GPT language model and autoencoders[C]//Proceedings of the 2023 9th International Conference on Web Research. Piscataway: IEEE Press, 2023: 214–220.
- [77] Chen Y F. Enhancing machine translation through advanced in-context learning: a methodological strategy for GPT-4 improvement[EB]. arXiv preprint, 2023, arXiv: 2311.10765.
- [78] Cui M L, Gao P Z, Liu W, et al. Multilingual machine translation with open large language models at practical scale: an empirical study[EB]. arXiv preprint, 2025, arXiv: 2502.02481.
- [79] Slator. How Microsoft wants to address gender bias in AI speech translation[Z]. 2025.
- [80] Roberts A, Raffel C, Shazeer N. How much knowledge can you pack into the parameters of a language model[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: USAACL, 2020: 5418–5426.
- [81] Kwiatkowski T, Palomaki J, Redfield O, et al. Natural questions: a benchmark for question answering research[J]. Transactions of the Association for Computational Linguistics, 2019, 7: 453–466.
- [82] Berant J, Chou A, Frostig R, et al. Semantic parsing on freebase from question-answer pairs[C]//Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. [S.l.:s.n.], 2013: 1533–1544.
- [83] Joshi M, Choi E, Weld D, et al. TriviaQA: a large scale distantly supervised challenge dataset for reading comprehension[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: USAACL, 2017: 1601–1611.
- [84] Chowdhery A, Narang S R, Devlin J, et al. PaLM: scaling language modeling with pathways[EB]. arXiv preprint, 2022, arXiv: 2204.02311.
- [85] Long Y C, Yang C, Tang G Z, et al. KidneyTalk-open: No-code deployment of a private large language model with medical documentation-enhanced knowledge database for kidney disease [EB]. arXiv preprint, 2025, arXiv: 2503.04153.
- [86] Izacard G, Lewis P, Lomeli M, et al. Atlas: few-shot learning with retrieval augmented language models[EB]. arXiv preprint, 2022, arXiv: 2208.03299.
- [87] Asai A, Min S, Zhong Z X, et al. Retrieval-based language models and applications[C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics. Stroudsburg: USAACL, 2023: 41–46.
- [88] Lan Y S, He G L, Jiang J H, et al. Complex knowledge base question answering: a survey[J]. IEEE Transactions on Knowledge and Data Engineering, 2023, 35(11): 11196–11215.
- [89] Mihaylov T, Clark P, Khot T, et al. Can

- a suit of armor conduct electricity: a new dataset for open book question answering[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: USAACL, 2018: 2381–2391.
- [90] Rajpurkar P, Zhang J, Lopyrev K, et al. SQuAD: 100, 000+ questions for machine comprehension of text[C]//Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: USAACL, 2016: 2383–2392.
- [91] Borgeaud S, Mensch A, Hoffmann J, et al. Improving language models by retrieving from trillions of tokens[EB]. arXiv preprint, 2021, arXiv: 2112.04426.
- [92] Petroni F, Rocktäschel T, Riedel S, et al. Language models as knowledge bases [C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Stroudsburg: USAACL, 2019: 2463–2473.
- [93] Toutanova K, Chen D Q. Observed versus latent features for knowledge base and text inference[C]//Proceedings of the 3rd Workshop on Continuous Vector Space Models and Their Compositionality. Stroudsburg: USAACL, 2015: 57–66.
- [94] Dettmers T, Minervini P, Stenetorp P, et al. Convolutional 2D knowledge graph embeddings[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2018, 32(1): 1462–1469.
- [95] Goodrich B, Rao V, Liu P J, et al. Assessing the factual accuracy of generated text[C]//Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York: ACM, 2019: 166–175.
- [96] Liang P, Bommasani R, Lee T, et al. Holistic evaluation of language models [EB]. arXiv preprint, 2022, arXiv: 2211.09110.
- [97] Zhang Y, Chen M D, Liu L Z. A review on text mining[C]//Proceedings of the 2015 6th IEEE International Conference on Software Engineering and Service Science. Piscataway: IEEE Press, 2015: 681–685.
- [98] Wu J Y, Gan W S, Chen Z F, et al. AI-generated content (AIGC): a survey[EB]. arXiv preprint, 2023, arXiv: 2304.06632.
- [99] 司亚中, 张旭龙, 杨帆, 等. 沙尘图像视觉增强技术综述 [J]. 大数据, 2025, 11(1): 117–134.
- Si Y Z, Zhang X L, Yang F, et al. A survey of sand-dust image visual enhancement techniques[J]. Big Data Research, 2025, 11(1): 117–134.
- [100] Gu S Y, Bao J M, Yang H, et al. Mask-guided portrait editing with conditional GANs[C]//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2019: 3431–3440.
- [101] Upchurch P, Gardner J, Pleiss G, et al. Deep feature interpolation for image content changes[C]//Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2017: 6090–6099.
- [102] Xu C C, Zhou M, Ge T Z, et al. Unsupervised domain adaption with pixel-level discriminator for image-aware layout generation[C]//Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2023: 10114–10123.
- [103] Fu B Y, Damer N. Face morphing at-

- tacks and face image quality: the effect of morphing and the unsupervised attack detection by quality[EB]. arXiv preprint, 2023, arXiv: 2208.05864.
- [104]Huang S Y, An J, Wei D L, et al. QuantArt: quantizing image style transfer towards high visual fidelity[C]//Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2023: 5947–5956.
- [105]Pernuš M, Štruc V, Dobrišek S. Mask-FaceGAN: high-resolution face editing with masked GAN latent code optimization[J]. IEEE Transactions on Image Processing, 2023, 32: 5893–5908.
- [106]Zhu J P, Yang C Y, Shen Y J, et al. LinkGAN: linking GAN latent to pixels for controllable image synthesis[C]//Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2023: 7622–7632.
- [107]Aoshima T, Matsubara T. Deep curvilinear editing: commutative and nonlinear image manipulation for pretrained deep generative model[C]//Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2023: 5957–5967.
- [108]Fu F L, Li R. Based on StyleGAN for face image cartoon style transfer[C]//Proceedings of the 3rd International Conference on Control and Intelligent Robotics. New York: SPIE, 2023: 858–862.
- [109]Brooks T, Holynski A, Efros A A. InstructPix2Pix: learning to follow image editing instructions[C]//Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2023: 18392–18402.
- [110]Fu T J, Hu W Z, Du X Z, et al. Guiding instruction-based image editing via multimodal large language models[EB]. arXiv preprint, 2023, arXiv: 2309.17102.
- [111]Huang Y Z, Xie L B, Wang X T, et al. SmartEdit: exploring complex instruction-based image editing with multimodal large language models[EB]. arXiv preprint, 2023, arXiv: 2312.06739.
- [112]Kathare K, Dhiman A, Gowda K V, et al. Instructive3D: editing large reconstruction models with text instructions[EB]. arXiv preprint, 2025, arXiv: 2501.04374.
- [113]Ni M H, Li X M, Zuo W M. NÜWA-LIP: language-guided image inpainting with defect-free VQGAN[C]//Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2023: 14183–14192.
- [114]Shen L L, Yan J, Sun X C, et al. Wavelet-based self-attention GAN with collaborative feature fusion for image inpainting[J]. IEEE Transactions on Emerging Topics in Computational Intelligence, 2023, 7(6): 1651–1664.
- [115]Yu F H, Gu J J, Li Z Y, et al. Scaling up to excellence: practicing model scaling for photo-realistic image restoration in the wild[C]//Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2024: 25669–25680.
- [116]Stathopoulos A, Han L G, Metaxas D. Score-guided diffusion for 3D human recovery[C]//Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscat-

- away: IEEE Press, 2024: 906–915.
- [117] Wei Y Y, Zhang Z, Ren J H, et al. Clarity ChatGPT: an interactive and adaptive processing system for image restoration and enhancement[EB]. arXiv preprint, 2023, arXiv: 2311.11695.
- [118] Lu Z H, Xiao Z Y, Bai J W, et al. Can SAM boost video super-resolution[EB]. arXiv preprint, 2023, arXiv: 2305.06524.
- [119] Kirillov A, Mintun E, Ravi N, et al. Segment anything[EB]. arXiv preprint, 2023, arXiv: 2304.02643.
- [120] Qiao Y Y, Chen Q, Deng C R, et al. R-GAN: exploring human-like way for reasonable text-to-image synthesis via generative adversarial networks[C]//Proceedings of the 29th ACM International Conference on Multimedia. New York: ACM, 2021: 2085–2093.
- [121] Liao W T, Hu K, Yang M Y, et al. Text to image generation with semantic-spatial aware GAN[C]//Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2022: 18166–18175.
- [122] Zhou Y F, Liu B C, Zhu Y Z, et al. Shifted diffusion for text-to-image generation[C]//Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2023: 10157–10166.
- [123] Kumari N, Zhang B L, Zhang R, et al. Multi-concept customization of text-to-image diffusion[C]//Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2023: 1931–1941.
- [124] Zhong S S, Huang Z Z, Wen W, et al. SUR-adapter: enhancing text-to-image pre-trained diffusion models with large language models[C]//Proceedings of the 31st ACM International Conference on Multimedia. New York: ACM, 2023: 567–578.
- [125] Wen S, Fang G A, Zhang R R, et al. Improving compositional text-to-image generation with large vision-language models[EB]. arXiv preprint, 2023, arXiv: 2310.06311.
- [126] Brade S, Wang B, Sousa M, et al. Promptify: text-to-image generation through interactive prompt exploration with large language models[C]//Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology. New York: ACM, 2023: 1–14.
- [127] Wu J Z, Tang C, Wang J B, et al. Diff-Sensei: bridging multi-modal LLMs and diffusion models for customized manga generation[EB]. arXiv preprint, 2024, arXiv: 2412.07589.
- [128] Zhou Y P, Zhou D Q, Cheng M M, et al. Storydiffusion: consistent self-attention for long-range image and video generation[J]. Advances in Neural Information Processing Systems, 2024, 37: 110315–110340.
- [129] Lyu Z Y, Li X M, Niu Z X, et al. Semantic-shape adaptive feature modulation for semantic image synthesis[C]//Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2022: 11204–11213.
- [130] Zeng Y, Lin Z, Zhang J M, et al. SceneComposer: any-level semantic image synthesis[C]//Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2023: 22468–

- 22478.
- [131] Gong J, Foo L G, Fan Z P, et al. Diff-Pose: toward more reliable 3D pose estimation[C]//Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2023: 13041–13051.
- [132] Tripathi S, Müller L, Huang C P, et al. 3D human pose estimation *via* intuitive physics[C]//Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2023: 4713–4725.
- [133] Chai S, Zhuang L S, Yan F Y. LayoutDM: transformer-based diffusion model for layout generation[C]//Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2023: 18349–18358.
- [134] Xue H, Huang Z, Sun Q, et al. Free-style layout-to-image synthesis[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2023: 14256–14266.
- [135] Zhang L M, Rao A Y, Agrawala M. Adding conditional control to text-to-image diffusion models[C]//Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2023: 3813–3824.
- [136] Koh J Y, Baldrige J, Lee H, et al. Text-to-image generation grounded by fine-grained user attention[C]//Proceedings of the 2021 IEEE Winter Conference on Applications of Computer Vision. Piscataway: IEEE Press, 2021: 237–246.
- [137] He Z C, Sun B, Juefei-Xu F, et al. Imagine yourself: tuning-free personalized image generation[EB]. arXiv preprint, 2024, arXiv: 2409.13346.
- [138] Ye H, Zhang J, Liu S B, et al. IP-adapter: text compatible image prompt adapter for text-to-image diffusion models[EB]. arXiv preprint, 2023, arXiv: 2308.06721.
- [139] Wang Q X, Bai X, Wang H F, et al. InstantID: zero-shot identity-preserving generation in seconds[EB]. arXiv preprint, 2024, arXiv: 2401.07519.
- [140] Li J G, Zhang X F, Jia C M, et al. Direct speech-to-image translation[J]. IEEE Journal of Selected Topics in Signal Processing, 2020, 14(3): 517–529.
- [141] Aytar Y, Vondrick C, Torralba A. SoundNet: learning sound representations from unlabeled video[EB]. arXiv preprint, 2016, arXiv: 1610.09001.
- [142] 张冰源, 张旭龙, 王健宗, 等. 数字说话人脸生成技术综述[J]. 大数据, 2024, 10(5): 74–95.
- Zhang B Y, Zhang X L, Wang J Z, et al. Survey of audio-driven talking face generation technology[J]. Big Data Research, 2024, 10(5): 74–95.
- [143] Tan S, Ji B, Pan Y. EMMN: emotional motion memory network for audio-driven emotional talking face generation [C]//Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2023: 22089–22099.
- [144] Stypułkowski M, Vougioukas K, He S, et al. Diffused heads: diffusion models beat GANs on talking-face generation [C]//Proceedings of the 2024 IEEE/CVF Winter Conference on Applications of Computer Vision. Piscataway: IEEE Press, 2024: 5089–5098.
- [145] Zhong W Z, Fang C W, Cai Y Q, et al. Identity-preserving talking face generation with landmark and appearance

- priors[C]//Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2023: 9729–9738.
- [146] Chen Z Y, Cao J J, Chen Z Q, et al. EchoMimic: lifelike audio-driven portrait animations through editable landmark conditions[EB]. arXiv preprint, 2024, arXiv: 2407.08136.
- [147] Jiang J W, Liang C, Yang J Q, et al. Loopy: taming audio-driven portrait avatar with long-term motion dependency[C]//Proceedings of the 13th International Conference on Learning Representations. [S.l.:s.n.], 2024.
- [148] Zen H G, Sak H. Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis[C]//Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE Press, 2015: 4470–4474.
- [149] 施昊翔, 张旭龙, 王健宗, 等. 情感语音合成综述[J]. 大数据, 2024, 10(5): 56–73.
- Shi H X, Zhang X L, Wang J Z, et al. A summary of emotional speech synthesis [J]. Big Data Research, 2024, 10(5): 56–73.
- [150] Van Den Oord A, Dieleman S, Zen H G, et al. WaveNet: a generative model for raw audio[EB]. arXiv preprint, 2016, arXiv: 1609.03499.
- [151] Corona E, Zanfira A, Bazavan E G, et al. VLOGGER: multimodal diffusion for embodied avatar synthesis[EB]. arXiv preprint, 2024, arXiv: 2403.08764.
- [152] Fu R B, Liu R, Qiang C Y, et al. ICAGC 2024: inspirational and convincing audio generation challenge 2024[C]//Proceedings of the 2024 IEEE 14th International Symposium on Chinese Spoken Language Processing. Piscataway: IEEE Press, 2024: 626–630.
- [153] Zhang D, Li Z W, Li S M, et al. SpeechAlign: aligning speech generation to human preferences[EB]. arXiv preprint, 2024, arXiv: 2404.05600.
- [154] Huang R J, Zhao Z, Liu H D, et al. ProDiff: progressive fast diffusion model for high-quality text-to-speech [C]//Proceedings of the 30th ACM International Conference on Multimedia. New York: ACM, 2022: 2595–2605.
- [155] Chen Z H, Tan X, Wang K, et al. In-fergrad: improving diffusion models for vocoder by considering inference in training[C]//Proceedings of the ICASSP 2022 – 2022 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE Press, 2022: 8432–8436.
- [156] Mittal G, Marwah T, Balasubramanian V N. Sync-draw: Automatic video generation using deep recurrent attentive architectures[C]//Proceedings of the 25th ACM international conference on Multimedia. New York: ACM, 2017: 1096–1104.
- [157] Liu Y, Wang X, Yuan Y T, et al. Cross-modal dual learning for sentence-to-video generation[C]//Proceedings of the 27th ACM International Conference on Multimedia. New York: ACM, 2019: 1239–1247.
- [158] Wei L Y, Levoy M. Fast texture synthesis using tree-structured vector quantization[C]//Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques. [S.l.: s.n.], 2000: 479–488.
- [159] Clark A, Donahue J, Simonyan K. Adversarial video generation on complex datasets[EB]. arXiv preprint, 2019,

arXiv: 1907.06571.

[160]Ohnishi K, Yamamoto S, Ushiku Y, et al. Hierarchical video generation from orthogonal information: optical flow and texture[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2018, 32(1).

[161]Saito M, Saito S, Koyama M, et al. Train sparsely, generate densely: memory-efficient unsupervised training of high-resolution temporal GAN[J]. In - ternational Journal of Computer Vision, 2020, 128(10): 2586–2606.

[162]Mei K F, Patel V. VIDM: video implicit diffusion models[J]. Proceedings of the AAAI Conference on Artificial Intelli - gence, 2023, 37(8): 9117–9125.

[163]Song K, Chen B Y, Simchowit M, et al. History-guided video diffusion[EB]. arXiv preprint, 2025, arXiv: 2502.06764.

[164]Peebles W, Xie S N. Scalable diffusion models with transformers[C]//Proceed - ings of the 2023 IEEE/CVF International Conference on Computer Vision. Piscat - away: IEEE Press, 2023: 4172–4182.

[165]Blattmann A, Dockhorn T, Kulal S, et al. Stable video diffusion: scaling latent video diffusion models to large datasets [EB]. arXiv preprint, 2023, arXiv: 2311.15127.

[166]Gupta A, Yu L J, Sohn K, et al. Photo - realistic video generation with diffusion models[M]//Computer Vision - ECCV 2024. Cham: Springer Nature Switzer - land, 2024: 393–411.

作者简介



张旭龙（1988–），男，博士，平安科技（深圳）有限公司，高级算法研究员，复旦大学计算机理学博士，2023年入选上海市东方英才计划青年项目。兼任深圳清华大学研究院及中国科学技术大学先进技术研究院校外导师，中国自动化学会联邦数据与联邦智能专业委员会委员。目前是IEEE、中国自动化学会以及中国计算机学会会员。主要研究方向为大模型、具身智能、跨模态智能计算以及模型上下文协议（MCP）等。发表学术论文80余篇，申请国家发明专利授权20余项。



瞿晓阳（1988–），男，博士，平安科技（深圳）有限公司前沿机器学习算法分组负责人，清华大学深圳国际研究生院校外导师，中国科技大学校先进技术研究院校外导师，中佛罗里达大学访问学者，从事机器学习、大数据、体系结构方面的研究工作，在语音语义分析、自动化机器学习、零样本和小样本学习、高性能计算与存储等方面经验丰富。在体系结构方向（如INFOCOM、DAC、IPDPS、TPDS）和人工智能方向（如NeurIPS、IJCAI、ICASSP、Interspeech）等国际会议和期刊发表过近50篇文章，其中1篇论文荣获会议最佳学生论文奖提名；担任多个国际期刊的评委，已授权专利70篇，已出版专著2本。



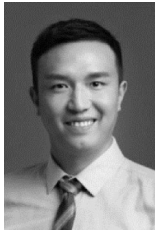
谢骏飞（2002–），男，中国科学技术大学硕士生，平安科技（深圳）有限公司算法实习生，主要研究方向为人工智能、具身智能和迁移学习等。



刘鹏程（1998–），男，西安工程大学硕士生，平安科技（深圳）有限公司算法实习生，主要研究方向为X光违禁品检测、AIGC图像生成等。



肖春光（1976–），男，深圳市宝安区教育事业发展中心高级教师，主要研究方向为教育数字化、网络安全、信息化教学创新。



王健宗（1983–），男，博士，平安科技（深圳）有限公司副总工程师，资深人工智能总监，联邦学习技术部总经理，智能金融前沿技术研究院院长。美国佛罗里达大学人工智能博士后，美国莱斯大学和华中科技大学联合培养博士，中国计算机学会资深会员，中国计算机学会大数据专家委员会委员，中国自动化学会联邦数据和联邦智能专业委员会副主任。主要研究方向为大模型、联邦学习和深度学习等。

收稿日期: 2025-03-31
通信作者: 王健宗, jzwang@188.com
基金项目: 广东省重点领域研发计划“新一代人工智能”重大专项(No.2021B0101400003)
Foundation Item: The Key Research and Development Program of Guangdong Province (No.2021B0101400003)