中图法分类号:TP335　　**文献标识码**: A　　**文章编号**: 1006-8961(2025)06-1953-32

论文引用格式:Zheng Y, Huang X, Qin F E, Liang Y, Huang Z X, Cao Y P, Tao X, Zheng M W, Yang H T, Wan P F, Chen J W and Huang X Y. 2025. AIGC 2D/3D technology development and creative industry applications. Journal of Image and Graphics, 30(6):1953-1984(郑屹, 黄向, 秦菲儿, 梁缘, 黄子瑄, 曹炎培, 陶鑫, 郑明悟, 杨皓天, 万鹏飞, 陈京炜, 黄心渊. 2025. 2D/3D生成式人工智能技术发展及创意产业应用. 中国图象图形学报, 30(6):1953-1984)[DOI:10.11834/jig.250005]

2D/3D生成式人工智能技术发展及创意产业应用

郑屹[1],黄向[1],秦菲儿[1],梁缘[2],黄子瑄[2],曹炎培[2],陶鑫[3],郑明悟[3],
杨皓天[3],万鹏飞[3],陈京炜[1],黄心渊[1*]

1. 中国传媒大学动画与数字艺术学院,北京　100024; 2. 北京哇嘶嗒科技有限公司,北京　100084;
3. 北京快手科技有限公司,北京　100085

摘　要: 随着数据和模型规模的增长,以大语言模型(large language model,LLM)和降噪扩散模型(denoising diffusion model)为代表的生成式技术的生成结果在多样性和品质上不断提高,这些大的生成式模型正在推动生成式人工智能(artificial intelligence generated content,AIGC)领域的快速发展。本文以创意产业的核心需求为切入点,对AIGC 2D/3D领域在2023—2024年间的技术和产业发展进行回顾与总结。首先总结生成式技术的发展背景和其市场应用价值。其次根据AIGC 2D/3D领域各自的技术发展进行总结,可以清晰地感受到该技术正在快速地发展和变化,从以生成式对抗网络(generative adversarial network,GAN)为核心快速发展为以降噪扩散模型和Transformer结构为核心。新结构具有更强的表达能力、更丰富的多样性和更灵活的控制能力。在AIGC 2D部分,以高质量生成基础、可控性生成技术和可编辑性生成技术为分类视角,对图像生成技术和视频生成技术分别进行梳理和总结;在AIGC 3D部分,在总结3D数据多样性的表达方式基础上,以输入数据的类型、输出数据的类型、生成方法3个不同视角对相关技术进行梳理。接下来分别总结两类技术当前遇到的技术问题和产业应用问题,如何提供更符合产业创作标准和需求的新技术是AIGC 2D/3D技术未来发展的重点方向。最后整体回顾过去20年间创意产业在技术发展带动下呈现出螺旋式向上发展的特征,并对技术发展的趋势提出一些思考与见解。

关键词: 生成式人工智能(AIGC);AIGC 2D;AIGC 3D;综述;创意产业应用;大语言模型(LLM)

AIGC 2D/3D technology development and creative industry applications

Zheng Yi[1], Huang Xiang[1], Qin Feier[1], Liang Yuan[2], Huang Zixuan[2], Cao Yanpei[2], Tao Xin[3],
Zheng Mingwu[3], Yang Haotian[3], Wan Pengfei[3], Chen Jingwei[1], Huang Xinyuan[1*]

1. School of Animation and Digital Arts, Communication University of China, Beijing　100024, China;
2. Beijing Wasida Technology Co., Ltd., Beijing　100084, China; 3. Beijing Yixiao Technology Co., Ltd., Beijing　100085, China

Abstract: The growth of data and model size has led to the emergence of new generative models, such as large language models and diffusion models, which produce increasingly high-quality and diverse results. These large generative models are driving the rapid development of artificial intelligence (AI)-generated content (AIGC). This article focuses on the core needs of the creative industry and reviews the technological and industrial developments in the AIGC 2D/3D field from 2023

收稿日期:2025-01-03;**修回日期**:2025-01-23;**预印本日期**:2025-01-30

*通信作者:黄心渊　hxy@cuc.edu.cn

基金项目:国社科艺术学重大项目(20ZD20)

Supported by:National Social Science Fund Major Art Studies Project(20ZD20)

to 2024. First, this article summarizes the development background of generative technologies and their market value. Second, given the technical development in the AIGC 2D/3D field, the pace of technological evolution is evidently fast. The focus has transitioned from primarily being on "generative adversarial networks" to focusing on "denoising diffusion models and Transformer structures". The new structure has stronger expressive power, greater diversity, and more flexible control capabilities. In the AIGC 2D review section, the article categorizes image and video generation technologies into three types: "high-quality generation foundation", "controllable generation technology", and "editable generation technology". AIGC 2D technology began around 2014 with models such as GANs and VAEs, which made breakthroughs in image generation. Researchers have focused on generating high-quality 2D images, improving resolution, and enhancing details. By 2018, the focus shifted to better image quality, with StyleGAN advancing realism and diversity, thereby driving AIGC's application in image processing and artistic creation. In 2020, diffusion models revolutionized 2D image generation, improving image quality and offering better control, such as text-to-image generation. By 2022, AIGC 2D technology evolved and expanded into advertising, content generation, and entertainment, with multimodal models such as CLIP and DALL-E enabling image creation from text descriptions. These innovations lowered the barrier for creators and improved generation quality. By 2024, AIGC 2D technology advanced to video generation, further expanding its applications. In the AIGC 3D review section, the article categorizes technologies based on "input data types", "output data types", and "generation methods". AIGC 3D development began in 2022, shifting from combining 2D generative models with implicit 3D representations to 3D diffusion models and mixed/explicit 3D data. Early research focused on traditional 3D models, such as voxels, meshes, and point clouds. With the success of generative adversarial network (GAN) and variational autoencoder (VAE) in 2D generation, these techniques were adapted to 3D. In 2023, breakthroughs in denoising diffusion models led to their application in 3D generation, showing potential for creating geometric structures and textures. However, the limitations of 2D diffusion models raised concerns about their ability to handle 3D data. By 2024, generative 3D technology matured, integrating with real-world applications and evolving from a 2D-based approach to a 3D-driven one. The use of multimodal models and natural language processing allowed users to generate complex 3D data from simple text or 2D images. This marked a shift from research to practical applications, with technological advancements and market applications progressing in tandem. Next, the article summarizes the current technical challenges and industry application issues faced by both types of technologies. The main emphasis for the future development of AIGC 2D/3D technologies will be how to provide new technologies that better meet industry creation standards and needs. For AIGC 2D, precision in controlling the structure, details, and style of generated images or videos to meet user expectations remains a bottleneck. This is particularly challenging in video generation, where camera movements, object motion, and scene composition are harder to control. Moreover, the lack of flexibility for fine-tuning and editing generated content remains an issue, and achieving a high match between generated results and multimodal inputs continues to be a challenge. Furthermore, content quality such as realism, accuracy, and aesthetic appeal needs significant improvement, especially in areas like hand and eye details and object motion continuity. AIGC 3D faces three key challenges: the gap between generated results and industry standards, differences in application requirements across industries, and the need to balance the divergent needs of industry experts and the general public. Industry-specific standards for 3D data, such as geometry features and texture resolution, demand high-quality outputs from AIGC 3D. Different industries require different technical adaptations of AIGC 3D technology. Furthermore, the differing needs of professional creators, who seek high precision, and general users, who prioritize ease of use and affordability, add complexity to AIGC 3D's real-world applications. Finally, the article provides a summary of the past 20 years, showing how the creative industry has experienced a "spiral upward development" driven by technological progress. It also offers some thoughts and perspectives on the future trends of technological development.

Key words: artificial intelligence generated content (AIGC); AIGC 2D; AIGC 3D; survey; creative industry application; large language model (LLM)

0 引 言

2023 年以来，生成式人工智能（artificial intelligence generated content，AIGC）技术在基础算法（如对抗生成网络、大型语言模型、去噪扩散模型等）以及计算硬件的飞速发展推动下，进入一个前所未有的高速发展阶段。随着技术的不断更新与迭代，新方法层出不穷，极大地促进了创意产业中内容生成的革新。然而，随着技术的飞速进步，也出现了一个亟待解决的问题：这些技术的未来发展方向是否与创意产业的实际需求相契合？它们是否真正解决了创作者和消费者在内容生产和消费过程中面临的核心问题？在创意产业快速发展的背景下，经过 20 年的变革，产业的趋势如何演变？未来生成式 AI 技术发展应重点关注哪些领域，才能真正推动产业的持续发展？

本文首先从创意产业的需求出发，回顾并总结 AIGC 2D/3D（AI-generated content in 2D/3D）技术在内容生成领域的需求基础、技术发展背景，以及这些技术在产业中所发挥的重要作用。在分析过程中，特别关注 AIGC 2D/3D 技术在创意产业中的具体应用，深入探讨专业化创作与大众化创作之间的异同点，特别是技术的普适性和可拓展性对创意产业的影响。

其次，从技术发展和产业应用两个层面，本文分别对 AIGC 2D 和 AIGC 3D 技术进行深入总结与分析。总体而言，AIGC 2D/3D 技术在过去一年多的进展中，经历了从基于对抗生成网络到基于降噪扩散模型，再到扩散模型与 Transformer 结构相结合的发展过程。技术的表现力也从初期相对简单、狭隘的生成效果，逐渐向精细、复杂且多样化的方向发展。然而，在技术逐步从理论研究向实际产业应用转化的过程中，其可控性弱、可编辑性差、生成品质难以达到行业标准等问题逐渐暴露，成为阻碍技术广泛应用的核心瓶颈。这些问题不仅影响了创意产业的实际生产力，也对技术的进一步发展提出更高的要求。因此，围绕这些技术短板展开的创新性解决方案，成为本阶段技术发展的主要脉络。

此外，本文还深入探讨了新兴 AIGC 技术在产业中的实际应用情况，详细介绍了当前行业内的一些领先企业及其创新实践。通过从产业需求的角度对现有技术的不足之处进行了反思。值得注意的是，虽然技术本身取得了诸多突破，但在如何呈现出更好的表现力，以及如何确保生成内容的可编辑性和可控性，仍然是亟待解决的关键问题。

最后，本文探讨了 AIGC 2D 和 AIGC 3D 技术未来可能面临的机遇与挑战，并结合过去创意产业所呈现的螺旋式向上发展特征，展望了 AIGC 2D/3D 技术在未来的发展趋势。随着技术和产业的不断演化，AI 生成技术有望进一步推动创意产业的多元化和个性化发展。

1 AIGC 2D/3D 技术的发展背景与价值

1.1 传统创意产业的生产模式

传统创意产业的生产模式通常依赖于手工创作和团队协作，涉及多个环节，如构思、设计、制作和发布。在这一过程中，创作团队需要投入大量时间、精力与资源，以实现富有创意的作品并最终呈现给消费者，从而实现产业的经济和文化价值。然而，传统的生产模式在某些方面存在显著的局限性，制约了创意的表达和实现效率。

首先，传统模式的生产流程烦琐，资源配置不均、团队沟通协调不畅等问题经常出现。这使得许多小型创意团队在与大型企业竞争时难以获得足够的支持和资源，限制了创新的广度和深度。

其次，创意作品在面向市场之前，往往难以准确预测其艺术价值和商业价值等多方面的成功或失败。例如，许多耗资巨大、精心制作的作品最终未能取得预期的市场效果，出现惨痛失败的案例屡见不鲜。

最后，随着移动互联网的普及和消费者生活方式的改变，内容消费呈现出碎片化趋势，消费者的选择越来越多样化，导致创意作品的市场风险增大。

因此，推动生产模式的转变，寻找更高效、更低成本的创作方式，已成为创意产业持续发展的核心任务之一。

1.2 生成式技术的发展背景

生成式技术的快速发展，得益于 2022 年以来在基础算法和计算硬件方面的显著进步。去噪扩散模型（Ho 等，2020）和大型语言模型（large language model，LLM）等先进算法的出现，为创意内容的生成提供了新的思路和方法。当依赖提示词（prompt）和

扩散模型生成多样化的数字内容成为可能时，AI创作的新时代悄然到来，创意内容的生产效率大幅提高。

这一系列技术的突破，不仅提升了内容创作的效率和质量，还衍生出更为丰富、多样化的创意形式。得益于计算能力的增强和生产成本的降低，生成式技术逐渐成为产业中可行的解决方案。随着更多创意作品通过AI生成，创作者和消费者的创作门槛也得到了有效降低，从而使得创意产业的生产模式逐步向更加开放和包容的方向转型。

1.3 AIGC 2D/3D技术的重要价值

自19世纪末摄影与摄像技术的诞生以来，图像和影视制作一直依赖于复杂的手工操作和高水平的专业技术。随着20世纪末计算机图形学的不断进步，数字特效和图像生成技术逐渐广泛应用于图像和影视制作领域，大大提升了作品的视觉效果。然而，这些进展依然面临着高成本和高技术门槛的限制，使得非专业人士难以制作出接近专业水准的作品。

进入2010年代，AIGC 2D技术的崛起和成熟改变了内容生产的方式。与传统人工创作相比，AIGC 2D技术能够通过AI自动生成高质量的2D视觉内容，大规模且高效地生产图像和素材。该技术不仅在专业生成内容（professionally generated content，PGC）领域得到了广泛应用，也极大推动了用户生成内容（user generated content，UGC）的发展，使得普通用户只需简单输入即可生成复杂的视觉内容。这种技术普及极大地扩展了创作者的参与范围，尤其在广告、教育和娱乐等多个领域，展现出巨大的潜力和广阔的应用前景。特别是在短视频和直播等领域，AIGC 2D技术将计算机视觉与计算机图形学结合，创造出新的特效应用形式，并获得了消费者的广泛认可。从而PGC和UGC进入了不同的发展轨道：专业创作者专注于为消费者提供高质量、高价值的内容，而普通创作者则侧重于创作个性化、娱乐性强的内容。这种技术创新不仅创造了巨大的社会和商业价值，也改变了社交和娱乐的模式，推动了整个产业的转型。

进入2020年代，AIGC 3D技术的兴起进一步拓展了计算机图形学的应用边界。通过结合AI和3D数据生成技术，AIGC 3D实现了从传统人工建模到自动化生成的革命，使得3D创意表达的可能性更加丰富。同时，AIGC 3D技术与其他前沿技术的交叉融合，也为其带来了新的发展动力。例如，AIGC 3D技术与XR（extended reality）技术结合，创造出即时生成内容的新体验；与机器人具身智能技术的结合，使得虚拟与现实的交互更加真实；与工业设计的结合，推动了生产设计的创新。AIGC 3D的快速发展不仅为创意产业带来了新的变革，还为跨行业应用开辟了广阔的前景。

2 AIGC 2D技术发展与产业应用

2.1 AIGC 2D综述

2.1.1 AIGC 2D技术的发展背景

AIGC 2D技术的发展可以追溯到2014年左右，得益于生成对抗网络（generative adversarial network，GAN）（Goodfellow等，2014）和变分自编码器（variational autoencoder，VAE）（Kingma和Welling，2022）等生成模型的提出，并首次在图像生成领域取得突破性进展。研究者逐渐探索如何通过深度学习生成高质量的2D图像，并涌现出大量基于生成模型的应用。早期研究聚焦于图像生成技术，如生成对抗网络、变分自编码器和自回归模型（autoregression model，AE），显著提升了低分辨率、特定领域图像（如人脸）的生成质量。

2018年，随着生成模型在图像生成上的表现不断提升，研究的重点方向聚焦于提升图像的分辨率和细节表现。伴随着StyleGAN（Karras等，2019）的推出，显著提升了2D图像生成的逼真度和多样性，并推动了AIGC技术在图像处理、艺术创作等领域的应用。

2020年，扩散模型（Ho等，2020）在2D图像生成领域取得突破，以其更强的生成质量和更稳定的训练过程引发了广泛关注。扩散模型通过逐步还原噪声生成清晰、高分辨率的图像，解决了GAN易陷入模式崩溃等问题。扩散模型在提升图像生成质量的同时，也为控制生成内容带来了更多可能性，例如通过条件扩散模型实现文本到图像的生成。

2022年，AIGC 2D技术继续发展，逐步应用于实际场景，如广告创作、内容生成和娱乐领域。基于大规模图像—文本数据的多模态生成模型，如CLIP（contrastive language-image pretraining）（Radford等，2021）、DALL·E（Ramesh等，2021），使用户可以通过

文本描述生成高质量的图像，极大地降低了图像创作的门槛，并赋予了创作者更多的表达自由。与此同时，随着更大规模数据集的推出和更多优化策略的提出，AIGC 2D 技术的生成效果和多样性进一步提升。

2024 年，AIGC 2D 技术从图像生成领域跨入了视频生成领域（如 VideoCrafter2（Chen 等，2024b）、CogVideoX（Yang 等，2024b）），技术的挑战也从单帧高质量图像生成，扩展到帧间的时序一致性与动态场景连贯性、物体运动的合理性等。同时，AIGC 2D 技术与自然语言处理、多模态生成技术深度融合，使用户不仅能够通过文本生成图像和视频，还能通过图像、音频等多模态输入生成更加复杂的内容。此外，随着技术的不断演进，使得结果的生成效果更具创意性、可控性。这一切进步均显著推动了 AIGC 2D 技术从实验研究向商业应用的发展进程。

2.1.2　AIGC 2D 技术的难点

尽管 AIGC 2D 技术在消费端取得了巨大影响力与广泛应用，但仍存在以下关键挑战。

首先，在生成图像或视频时，如何精确控制画面内容的结构、细节和风格符合用户的预期仍是技术瓶颈。特别是在视频生成领域，想要让画面的运镜、物体的运动完全符合创作意图，难度显著增加，如：符合设计师需求的画面元素构图可控性较弱；生成视频结果中的镜头运动推、拉、摇、移设置；特写、近景、中景、远景、全景等拍摄取景景别设置；角色运动的精确动态控制和节奏控制缺失等。

其次，当前生成的内容通常缺乏灵活的调整与编辑能力，对结果进行局部修改和微调时，尚无精准有效的方法，难以满足复杂的创作需求。尽管多模态输入的使用增加了生成的灵活性，但要实现生成结果与文本、图像、音频等输入信息之间的高度匹配，依然充满挑战，如：与输入文字、音频相匹配的角色口型动态和面部表情生成的不足；用输入角色的外形来修改已知视频的角色特征等；多层生成图像输出、图像光照控制输出等尚存在较多问题。

此外，在内容的真实性、准确性和美观性等内容品质方面，仍存在较大的可提升空间，如：角色的手指区域效果生成还存在错误；眼睛发丝等细节区域生成的细节和稳定性不足；生成画面的结果在人类主观审美评估中的缺陷；生成视频中物体运动和结构的连续性、稳定性等缺失。

最后，AIGC 2D 技术在实际应用中也面临计算资源的消耗与效率问题，特别是扩散模型等生成技术的高计算需求，仍制约着其大规模商业应用的可行性。

2.1.3　AIGC 2D 章节的综述范畴

本节技术综述的范围以 2D 数据端到端生成方法为主，并不包含经典的图像生成和编辑算法。

2.2　AIGC 2D 技术发展

随着去噪扩散模型技术的发展，AIGC 2D 技术进入到新的发展阶段。相比于基于生成对抗网络时代，基于新技术下的生成多样性和效果都得到了极大的提升，如图 1 所示，图中横轴代表技术的发表时间，纵轴代表技术采用的模型结构。

本节从 2D 图像生成技术（2.2.1 节）和 2D 视频生成技术（2.2.2 节）对 AIGC 2D 技术进行综述，以高质量生成、可控性生成和可编辑性生成 3 个分类方式为主进行综述，并整理了当前主流 AIGC 2D 技术方法和效果评估量化指标（2.2.3 节）。

AIGC 2D 技术在过去一年多的完整发展过程中，图像生成技术作为视频生成技术的基础，在诸多方面起到关键作用。同时，Transformer 也从大语言模型迁移至视频生成领域，成为核心结构之一。

最后，讨论了当前 AIGC 2D 生成技术存在的局限性（2.2.4 节），包括数据问题和算法问题。

2.2.1　图像生成技术

1）图像高质量生成技术。图像生成领域因深度学习技术的快速发展而涌现出一系列方法，包括变分自编码器（VAE）、标准化流（normalizing flow）（Dinh 等，2017）、生成对抗网络（GAN）以及去噪扩散模型。在这些模型中，存在三难问题（Xiao 等，2022），即一个生成模型通常难以同时满足高样本质量、高模式覆盖和高速采样 3 个关键要求。这是因为受限于现有技术发展的时代局限性，一个技术方案往往会为了满足某些要求而牺牲其他方面。在上述方法中，变分自编码器和标准化流在高模式覆盖和高速采样方面表现优异，但生成样本的质量相对较低。

生成对抗网络通过引入对抗机制，实现了高质量样本生成，且具有强大的灵活性，能够针对不同任务进行定制和部署。另外其所需资源相对较少，广泛应用于定制垂类的图像生成和编辑任务，例如人像表情驱动 LivePortrait（Guo 等，2025），关键点可控

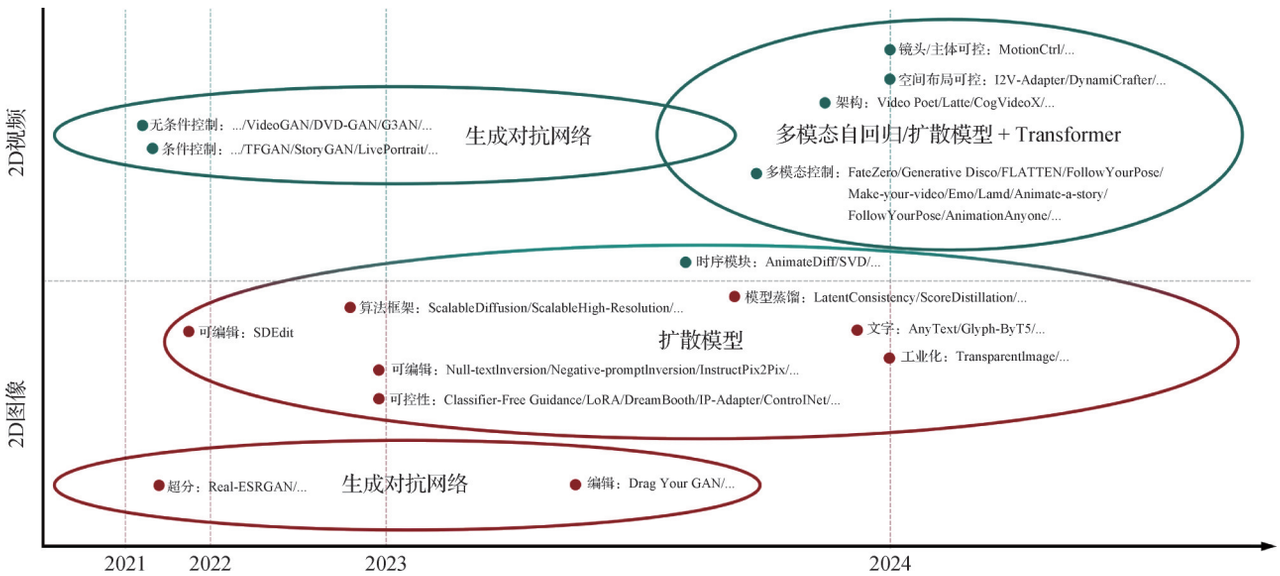

图1　AIGC 2D的技术发展图谱

Fig. 1　AIGC 2D technology development chart

生成DragYourGAN（Pan等，2023），图像超分辨率Real-ESRGAN（Wang等，2021b）等。但它常常以牺牲模式覆盖为代价，容易出现模式崩溃（mode collapse）问题。

扩散模型则以其高品质和多样性著称，广泛应用于高质量内容生成领域，但其面临资源消耗巨大的问题，针对这一问题，研究人员通过改进算法框架进行优化，如ScalableDiffusion（Peebles和Xie，2023）和ScalableHigh-Resolution（Crowson等，2024），模型蒸馏LatentConsistency（Luo等，2023）和ScoreDistillation（Lukoianov等，2024）等方向都有大量相关研究。

2）图像可控性生成技术。图像的可控性生成技术一直是学术界和工业界的重点研究方向之一。由于从零开始训练扩散模型是一项复杂且昂贵的工作，因此研究者通常采用在预训练模型基础上，对模型的部分结构进行优化、微调以及增加新结构的方法实现目标。

按照应用场景，图像的可控生成可以分为语义信号控制、特定物体和风格生成、空间布局控制3大类。

语义信号控制指通过自然语言的输入实现。利用自然语言经过文本编码器后得到的特征，通过无分类器引导Classifier-Free Guidance（Ho和Salimans，2022）的方式控制生成内容，其优点在于相比分类引导器，不需要外部分类器且具有更多的灵活性与多样性。缺点是需要反复调整影响强度来达到生成目标；近期的研究通过优化文本编码的质量Exploring The Role Of LLM（Ma等，2024a）和文本特征与图像特征的交互方式Scaling Rectified Flow Transformers（Esser等，2024）提升语义响应的准确性。

特定物体和风格生成指用新输入的图像特征控制模型生成指定人物、物体或艺术风格的图像，由于具有易用且直观的使用效果，因此在影视制作等方面有广泛应用。该思想在技术上可以分为两类：一种是使用在线微调技术，例如，LoRA（low-rank adaptation）（Hu等，2021）通过引入额外的权重矩阵对原有模型进行微调，实现了更精细的生成控制，并且减少了训练所需的资源和参数的数量，并且能实现任意组合可插拔，便于模型分享。DreamBooth（Ruiz等，2023）允许进行模型权重的个性化微调，用户可以选择全面训练所有参数，或采用诸如LoRA这样的技术对部分参数进行训练。两种技术作为对比，LoRA则在实现特定生成效果的控制上展现了更高的精确度，而DreamBooth在对数据集的适应性上提供了更大的宽容度。另一种是通过增加适配器，实现物体或人物信息的注入，如IP-Adapter（Ye等，2023）通过对文本特征和图像特征采用了解耦的交叉注意机制实现图像和文本信息对生成结果的控制。

空间布局控制指允许用户生成指定生成图像的布局结构。基于ControlNet（Zhang等，2023a）通过对网络添加额外的多模态控制信号（如边缘图、深度图或姿态等先验信息）影响生成结果，显著提升了生成

图像的空间布局可控性。该方法由于可以与LoRA、IP-Adapter等方法混合使用，且因为其强大的图像布局控制能力而得到广泛使用。

3）图像可编辑性生成技术。在图像的可编辑性生成技术方面，技术的发展主要围绕着如何在生成图像的基础上进行局部或全局的调整，以及设计新的输出特征以更符合行业需求两大类。

在生成基础上进行局部或全局调整方向上，研究者提出一系列基于逆向编辑的方案，目前较为成功的路线分为“反演（inversion）”和“构造配对数据训练指令编辑”两条路。反演的核心思路是将待编辑图像映射回生成过程中的潜在空间，再通过重新施加条件或控制信号，实现对图像的精确编辑。这一方法的优势在于能够保留图像原有的视觉特征，同时进行局部或全局的调整。扩散模型的引入，尤其是其逐步生成的特性，为在生成过程中施加修改和编辑提供了便利。通过在生成过程中的中间步骤施加编辑约束或局部调整，可以实现复杂的、分阶段的内容修改。如SDEdit（Meng等，2022）是这一方法的开创性工作，基于DDIM（denoising diffusion implicit models）Inversion的方法，扩散模型反演技术的改进Null-textInversion（Mokady等，2023）和Negative-promptInversion（Miyake等，2024）也进一步提升了真实图像的编辑效果。另一类则是通过构造配对数据训练指令编辑模型，避免了反演的过程。例如，InstructPix2Pix（Brooks等，2023）和EmuEdit（Sheynin等，2024）等方法，实现了根据用户输入的自然语言指令进行图像编辑，交互简单灵活。

在关于如何让生成结果更符合行业应用需求上，TransparentImage（Zhang和Agrawala，2024）探索了前景和背景的分层生成方法、VectorFusion（Jain等，2023）探索了矢量化结果生成，以及Glyph-ByT5-v2（Liu等，2024h）和AnyText（Tuo等，2024）探索了文字生成。

2.2.2　视频生成技术

继扩散模型在图像生成领域取得广泛成功之后，自2023年以来视频生成领域也开始逐渐升温，吸引了大量的学者和企业投入到相关的工作中去。早期的工作通常是基于预训练的图像生成模型，一大类是将其对视频进行在线微调，如Tune-a-Video（Wu等，2023），这类方法通过在大规模图像生成模型的基础上进行少量视频数据的微调，从而使模型能够生成连贯的视频内容。这些方法通常依赖于图像生成模型的稳定性和强大的特征表达能力，并通过短时视频数据的补充学习，以生成具有时序一致性的视频。另外一类方法是在网络结构中添加可学习的时序模块以支持视频生成。例如，AnimateDiff（Guo等，2024b）和SVD（singular value decomposition）（Blattmann等，2023）等方法通过在模型中引入时序建模模块，能够生成数十帧长度的中小运动的视频。

2023年下半年，尤其是2024年初OpenAI的Sora放出的技术报告页面，学术界和工业届中逐渐开始出现一些效果更为成熟的模型。这些模型主要沿着两条路线发展：一条是以VideoPoet（Kondratyuk等，2024）为代表的多模态自回归路线，它们通常是结合文本、图像和音频等多种输入模态，按照LLM的训练方式对视频内容进行自回归式地生成；另一条是与Sora一样，依然沿用图像生成里的生成模型范式（扩散模型（Ho等，2020）/ Flow Matching（Lipman等，2023）模型），如CogVideoX（Yang等，2024b）、Latte（Ma等，2024c），主要改进是将生成模型的主干网络换成了扩展性更强的Transformer模型，并投入了更多的视频数据和算力资源进行训练。除此之外，VAE、扩散路径等组件也都有对应的改进并应用在视频生成中。

在经历了视频生成技术所带来的短暂新鲜感后，内容创作者们对该技术能否在产业落地提出一些关键问题：生成内容高品质问题（如真实感、美感和清晰度等）与生成内容可控性和可编辑问题（如物体特征控制、视听语言控制、运动控制、局部内容编辑和关键帧控制等）。

1）视频高质量生成技术。视频生成的高品质问题是一个本质问题，目前主流的做法是遵循大语言模型的成功经验。首先，提高数据源的品质，剔除低质数据，例如摄影摄像领域中的专业知识，对数据源品质进行分类与筛查。其次，提高对数据的标记品质，针对“审美”具有的强烈民族性、文化性、主观性、时代性和视听语言专业性等特征，将高品质数据进行更详尽的数据整理与标记工作，并结合计算美学的一些方法加速数据处理的效率和效果，以期提高模型在这一方面的能力。最后，参考大语言模型的后训练部分，收集一批高质量数据并标注后，加入监督微调阶段（supervised fine-tuning，SFT）和基于人类

反馈的强化学习(reinforcement learning with human feedback, RLHF)阶段进一步提升模型的真实感、美感和清晰度等,如MovieGen(Polyak等,2025)所述。

2)视频可控性生成技术。相比于图像生成领域,视频生成领域中,由于增加了时序信息,因此用户不仅需要模型具有语义和多模态信号控制、特定物体和风格生成、空间布局控制的能力,还需要具有镜头和主体运动控制的能力,一些基于图像生成控制的思想和方法,也被扩展到视频生成控制中。

(1)语义和多模态信号控制。现代视频生成模型本质上就是通过多模态训练完成的,这使得多模态信号成为实现可控性和编辑的重要手段之一。常用的控制信号例如,文本信号实现视频内容编辑FateZero(Qi等,2023),音频信号可以用于控制背景音乐或角色的口型同步Generative Disco(Liu等,2023c)和EMO(Tian等,2025),图像信号则可用于调整生成视频的风格、场景布局,甚至控制特定角色的外貌特征,如Lamd(Hu等,2023)和Generative Image Dynamics(Li等,2024e)。除此之外,深度图Make-your-video(Xing等,2025a)和Animate-a-story(He等,2023)、光流图FLATTEN(Cong等,2024)或3D姿态信息Animate Anyone(Hu,2024)和Follow Your Pose(Ma等,2024d)也能够为视频生成提供精确的空间布局和动作引导,使生成的内容更符合物理逻辑和时空一致性。通过多模态信号的控制,用户能够更加灵活、直观地编辑和定制视频内容,满足复杂生成场景中的各类需求。

(2)特定物体和风格生成。用户希望在生成视频时,能够明确控制视频中角色的身份、外貌以及物品的特定属性。例如,保持某个特定人物的面部特征、衣着风格或其他显著特征一致,不论是在多帧视频生成中,还是在局部视频的编辑时,用户都需要这种精确的控制能力。此外,生成过程中物品的ID也需要保持不变,尤其在涉及产品展示或品牌推广的场景中,确保生成的视频内容与实际物体一致。现有方案LivePortrait(Guo等,2025)和AnimateAnyone(Hu,2024)通常是设计专门的ID特征提取器,并将ID特征向量以交叉注意力等方式注入到生成网络中。

(3)空间布局控制。生成视频不仅需要内容连贯,场景的空间布局同样至关重要。用户希望能够精确控制生成视频中各个对象在空间中的位置、距离及其相互关系,生成符合设想场景的画面布局。一大类方法,如I2V-Adapter(Guo等,2024a)和DynamiCrafter(Xing等,2025b),则采用一幅图像作为额外的引导信息,使用交叉注意力机制或拼接机制输入视频生成模型来使得视频的内容和图像一致,从而实现空间和场景布局可控的效果。另一类方法通过显式控制注意力位置实现VideoTetris(Tian等,2024)。

(4)镜头和主体运动控制。镜头和主体运动是视频内容中的核心元素,创作者希望在生成过程中能够精细的控制运动,从而能够模拟不同的镜头语言效果,以满足电影、广告或游戏制作中的特定需求。镜头控制常通过提取现有视频镜头的运动信息,或直接根据预设的运动合成渲染数据后,将成对的相机轨迹信息和视频输入给生成模型从而实现精细的控制MotionCtrl(Wang等,2024b)。运动控制则关注于通过指定的运动路径或速度场引导视频生成,使得生成内容可以精确遵循预定的运动模式。其中的代表性工作为DragNUWA(Yin等,2023)。DragNUWA则通过引入轨迹采样器进行开放域的轨迹控制,并结合多尺度融合和自适应训练策略,实现了多尺度的轨迹控制和一致的视频生成效果。但这些方法与影视产业中导演关于角色、场景、镜头的调度方法和专业的视听语言之间还有不小差异,以及生成视频中的物理真实性等问题,都需要研究者后续进行持续探索。

3)视频可编辑性生成技术。在视频编辑方向中,有效保持被编辑视频的帧间连续性和一致性是一个具有挑战的目标。Video-P2P(Liu等,2024d)介绍了一种调优一个文本到集合模型以完成反演和一种新的注意力解耦引导策略的方法在早期实现了一种视频编辑能力;CoDeF(Ouyang等,2024)将输入视频分解为2D内容规范场和3D时间形变场,从而提高了帧间一致性,降低了视频生成中的闪烁问题;CCEdit(Feng等,2024)设计了三分支神经网络结构,将结构控制和外观控制分离,实现了精确且多样性的编辑能力;CAMEL(Zhang等,2024b)通过添加运动提示词和因果运动增强注意力机制,实现了被编辑对象的外观调整及运动模式保持;Video256(Li等,2024b)依靠时空一致性约束,通过时空期望最大化反演实现了视频编辑能力;VidToMe(Li等,

2024d)通过跨帧令牌合并以增强生成视频的时间一致性;RAVE(Kara等,2024)采用了一种噪声重排策略,利用帧之间的时空交互,更快速地生成时间一致的视频;COVE(Wang等,2024a)提出一种基于滑动窗口的策略来计算跨帧之间令牌的相似性和相关性;NaRCan(Chen等,2024d)集成了混合变形场和扩散先验,分别采用单应性建模和多层感知器解决全局与局部问题,增强了模型对动态视频的处理能力。

2.2.3　AIGC 2D技术方法分类和效果评估

表1列举了本综述中的AIGC 2D技术的部分相关方法。

当前针对AIGC 2D的评价方法,可分成3个大维度:视觉质量、语义相关性和动态质量。评估方法通常会利用技术质量指标、预训练多模态模型或是专用的数据驱动的模型进行打分。随着技术的发展,开始出现一些使用多模态语言模型进行评级的工作。此外还有一些最新的工作,尝试评价其物理合理性(PhyGenBench(Meng等,2024))。以下为相关机评榜单:FETV(Liu等,2023d)、EvalCrafter(Liu等,2024g)、VBench(Huang等,2024b)、AIGCBench(Fan等,2023)、T2VScore(Wu等,2024a)、VideoGen-Eval(Zeng等,2024)和PhyGenBench(Meng等,2024)。

2.2.4　AIGC 2D技术现存问题

1)AIGC 2D技术的数据问题。在AIGC中,数据问题是一个关键挑战,涉及数据合规性、数据规模、质量、图像视频数据管理的复杂性以及成本和数据偏见等多个方面。确保数据符合隐私和版权法规的合规性需要严格的监管,而大规模高质量数据的收集和处理要求巨大的存储和计算资源,增加了项目成本。此外,视频数据的管理比静态图像更加困难,因为它涉及到更大的文件尺寸和更复杂的时间依赖性。最后,避免数据集中的种族、风格等偏见,确保生成内容的多样性和公正性,也是不可忽视的重要问题。

2)AIGC 2D技术的算法问题。尽管近两年图像和视频生成算法取得了突破性进展,其在生成质量、可控性和可编辑性,以及生成效率等多个关键领域仍存在挑战。

首先,在生成质量和多样性方面,尽管现有技术能够创造出越来越逼真的图像,但它们在表现细腻的几何纹理细节、光影效果以及复杂场景的深度和动态性方面仍有限。算法在理解和渲染不同材质的光线反射和透视变形等高级视觉属性时,尤其容易出现问题,这影响了最终生成物的真实性和视觉吸引力。文字的准确生成,也是当前的关键痛点。

其次,在可控性和可编辑性方面,虽然近期的模型开始支持基于文本的描述和其他条件输入以可控生成图像和视频,但距离产业应用的要求仍然存在差距。例如,当前的图像和视频通常都是单帧和单片段生成,而影视制作需要维持长视频中的人物、物体、镜头和场景之间的一致性,以及按照指定的艺术风格生成内容,这时当前模型往往难以精确地执行。

表1　AIGC 2D部分关键方法对比

Table 1　Some examples of AIGC 2D methods

方法	输出类型	模型结构
SDEdit	图像	Diffusion
LoRA 2206	图像	Transformer
CFG	图像	Diffusion
DreamBooth	图像	Diffusion
FlowMatching	图像	Diffusion
InstructPix2Pix	图像	Diffusion
ControlNet	图像	Diffusion
DragYourGAN	图像	GAN
IP-Adapter	图像	Diffusion
LatentConsistency	图像	Diffusion
ScoreDistillation	图像	Diffusion
Transparent图像	图像	Diffusion
AnimateDiff	视频	Diffusion
DragNUWA	视频	GAN
SVD	视频	Diffusion
FollowYourPose	视频	Diffusion
Make-your-视频	视频	Diffusion
I2V-Adapter	视频	Diffusion
FateZero	视频	Diffusion
LivePortrait	视频	GAN
视频Poet	视频	Transformer
视频Tetris	视频	Diffusion
Cog视频X	视频	Diffusion + Transformer
MovieGen	视频	Diffusion

特别是维持故事线索一致性、角色表情和动作的自然过渡等方面对控制准确性有非常高要求，算法还远远未能达到完全令人满意的水平。这一点在需要精确控制细节以适应特定叙事或品牌要求的商业应用中尤为重要。

最后，在生成效率方面，虽然生成单一图像的速度有所提升，但在处理长视频序列时，效率仍然是一个重大瓶颈。这不仅因为视频数据量本身就大，还因为要保持时间上的连贯性和逻辑性，需要算法进行更加复杂的计算。此外，高分辨率和高质量的内容生成要求巨大的计算资源，不仅增加了成本，也限制了技术的广泛应用，尤其在资源有限的环境中。2024年以来，随着Scaling Laws（Kaplan等，2020）和Transformer架构在视觉内容生成领域的越来越广泛应用，进一步显著增大了对计算资源的要求。

2.3 AIGC 2D技术在产业中的应用

2.3.1 AIGC 2D产业报告

随着人工智能技术的不断成熟，AIGC 2D技术在图像和视频生成领域迅速崛起。自2023年以来，AI与图像、视频结合的应用在全球市场上表现出色，成为盈利能力最强的AI细分赛道。在2024年1月至8月期间，AI + 图像视频类应用内购收入总和接近11亿美元，占整个AI应用市场总收入的53%，成为AI细分市场中最具盈利潜力的领域（来自SensorTower的报告，网址https://sensortower.com/Zh-CN/blog/state-of-ai-apps-2024-report-CN）。

在全球范围内，AIGC 2D技术的主要驱动力来自于生成式模型的不断优化与创新。海外平台方面，Pika、Runway Gen-3和LumaAI等平台表现尤为突出。Runway Gen-3 Alpha和Luma AI是当前视频生成领域的两大领先平台，分别凭借其高保真度的视频生成能力和3D重建技术受到广泛关注。Runway Gen-3通过简单的文本提示即可快速生成复杂动态场景，特别擅长处理复杂的场景和人物动作，生成结果接近现实，并广泛应用于广告和娱乐等领域。而Luma AI则以其高效的3D重建和复杂材质渲染能力著称，尤其适用于虚拟场景构建和影片制作，保证了动态场景中的一致性和高质量表现。这些平台正推动AIGC技术在广告、影视制作等领域的革新发展。

国内市场同样涌现出一批优秀的AIGC 2D技术平台，如快手可灵（I2V-Adapter（Guo等，2024a）、LivePortrait（Guo等，2025））、字节跳动即梦、Minimax海螺、生数科技Vidu、爱诗科技PixVerse。在国内的AIGC 2D视频生成领域，快手可灵平台（图2）作为国内首个可开放体验的大规模文生视频大模型，展现了极强的技术实力。可灵通过支持多模态输入和复杂场景生成，特别适用于短视频、直播带货和创意广告制作，成为国内首批广泛应用于视频生成的顶尖平台之一，且在国内外社交平台引发了大众的广泛关注与使用；字节跳动即梦展现了流畅的运镜、动态生成及多样的风格化效果，近期的持续更新也表现出强大的生成能力；MiniMax海螺以其精美的镜头运动和自然的美学表现为特色，适合娱乐和广告内容的创作；生数科技Vidu则专注于广告和品牌宣传的自动化视频生成，提供高精度的场景生成能力；爱诗科技PixVerse主打创意场景生成和风格化处理技术，在社交媒体和创意广告中展现了不俗的表现。

这些平台的成功离不开各自的技术创新和市场定位。海外平台通常依托于强大的计算资源和开放的技术生态，推动AIGC技术的快速迭代与应用扩展；而国内平台更多依赖于强大的大模型技术人才、海量平台数据和用户需求的深刻理解，形成了独特的竞争优势。例如可灵、即梦等模型通过结合本地化的数据和场景，实现了更精准的本地化内容生成与优化，具有更强的中国文化特征。此外，各大平台都在不断探索更加可控、可编辑以及多模态融合的方向，这是当前技术进化的关键路径，也必将成为未来AIGC视频生成领域的核心竞争点。随着生成技术和生成效果的不断升级，AIGC 2D技术将在实时性和互动性上进一步突破，推动视频生成技术向更高质量和更广泛的应用场景迈进。

2.3.2 AIGC 2D技术在产业中的应用和问题

AIGC 2D技术在产业中的应用涵盖了多个领域，主要包括以下方面：

1）广告与营销。利用AIGC 2D技术，广告创作者可以快速生成高质量的视觉内容，如广告片段、品牌宣传视频和动态特效，显著提高内容生产效率。同时，AIGC 2D技术还能实现广告内容的个性化定制，根据不同用户群体的偏好生成定制化广告，提升广告的精准投放效果。

2）影视制作。AIGC 2D技术在影视制作中的应用主要体现在场景设计和特效制作等环节。在场景设计方面，AIGC 2D技术能够根据导演的需求快速

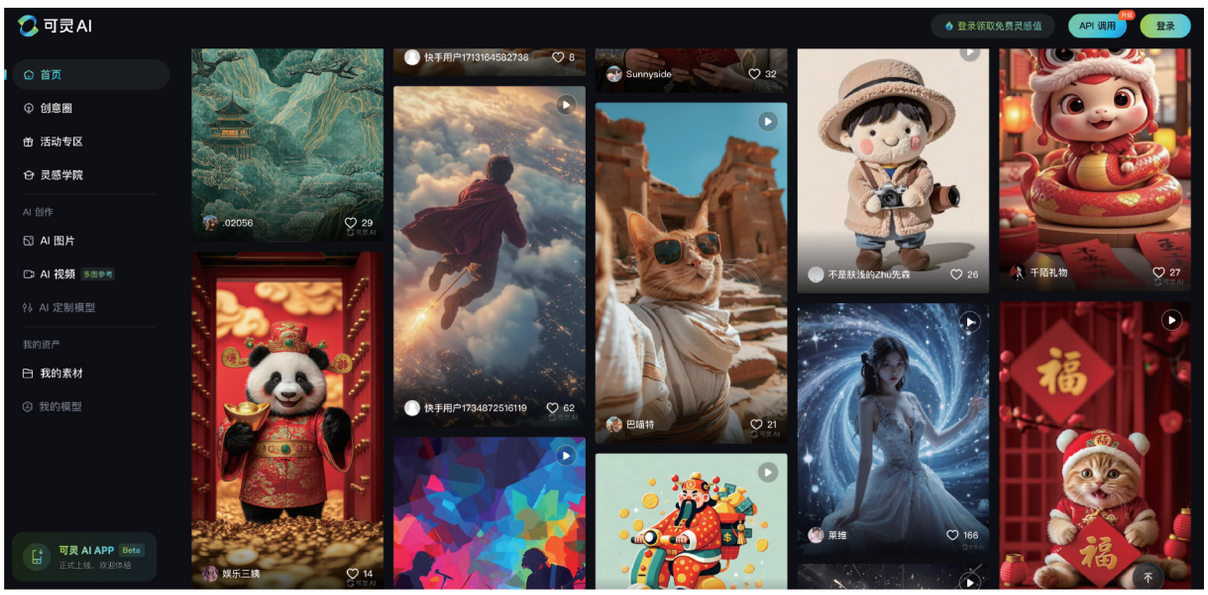

图 2　快手可灵的产品主页

Fig. 2　KeLing's product homepage

生成多样化的场景布局和视觉效果；在特效制作中，AIGC 2D 技术则能够实现复杂的视觉特效和动态效果，降低制作成本并缩短制作周期。

3）电商与直播。在电商和直播领域，AIGC 2D 技术主要应用于产品展示视频的自动生成和直播内容的智能优化。通过生成高质量的产品展示视频，电商平台可以更直观地展示产品特点，提升用户购买欲望；在直播过程中，AIGC 2D 技术能够实时生成互动内容和动态特效，增强直播的观赏性和互动性，提升用户体验和粘性。例如 Pika 近期推出的多种不同特效类视频生成案例，获得了大众社交媒体的广泛关注和传播，并成为 AIGC 2D 视频生成领域当前阶段内标杆性技术应用方式之一。

虽然 AIGC 2D 技术已经在行业中有着广泛应用，但依然面临 3 个方面的问题：1）生成结果与工业标准还存在有较大差距；2）工具链的成熟度不足；3）生成数据的版权和隐私问题。

首先，基于当前技术生成的图像/视频在输出的分辨率（包括 dpi（dots per inch）和物理尺寸）、色彩模式、色彩 Bit 深度以及文件格式等，都无法达到产业应用的基本标准（如工业打印要求达到 300 dpi、CMYK、矢量格式等）。

其次，尽管 AIGC 技术在图像和视频内容的自动生成方面取得了显著进步，但相关的工具链尚未完善和成熟。在产业应用中，图像视频生成技术需要与上下游工具链无缝集成，这包括图层编辑、绿幕技术、影视工业的流水线、视觉特效以及调色等环节。目前，这些工具和系统往往是分离的，缺乏标准化和互操作性。如，AIGC 生成的内容需要适配不同的特效软件和后期制作流程，但由于工具链的兼容性和扩展性不足，使得集成过程耗时且复杂。虽然近期 Adobe 等工具链厂家开始逐步支持 AIGC 内容，学术界也开始出现将 AIGC 技术适配产业流程的研究，但对于复杂的影视制作，现有的工具往往难以提供足够精细的控制级别，以符合专业制作的要求，这限制了其在高端影视制作中的应用。

最后，随着 AIGC 技术的快速发展，生成内容所涉及的版权、安全、伦理和偏见问题也逐渐引起广泛关注。这些问题不仅影响着内容创作者的权益，也关系到社会整体的道德标准和安全保障。

3　AIGC 3D 技术发展与产业应用

3.1　AIGC 3D 综述

3.1.1　AIGC 3D 技术的发展背景

AIGC 3D 技术自 2022 年起开始发展，其发展的整体趋势是从 2D 生成模型和 3D 数据隐式表达方式相结合，逐步向 3D 扩散模型和 3D 数据混合/显式表达方式相结合的方向在快速发展。

在早期研究中，研究者主要聚焦于探索以传统

3D模型显式表达方式(如体素、网格或点云)为主的生成方法。伴随着生成对抗网络(GAN)和变分自编码器(VAE)在图像生成领域的成功,研究者们开始逐步尝试将这些2D生成技术扩展到3D生成领域。

2023年,由于去噪扩散模型在2D图像生成领域取得突破,呈现出丰富的多样性和高品质,因此该技术也被迅速应用于3D数据的生成。研究者们对扩散模型在几何空间结构生成和纹理多样性生成所具有的潜力寄予厚望,然而,2D扩散模型固有的局限性也引发了对其在3D数据生成领域能力上限的质疑。

2024年,生成式3D技术进一步成熟,并逐渐与实际应用场景深度融合。伴随高品质3D数据集的出现,生成技术也从最初的仅依靠2D数据和模型驱动生成,逐步发展为2D + 3D数据及方法混合驱动生成,并最终迈向基于原生3D数据驱动生成的方向。同时,多模态生成模型与自然语言处理的结合,使得用户仅通过简单的文本描述或2D图像的引导便能生成复杂3D数据。至此,生成式3D技术开始从实验室逐步迈向"技术研究"与"市场化应用"双轨并行、相互促进的新阶段。

3.1.2 AIGC 3D技术的难点

尽管AIGC 3D技术取得了显著进展,但仍面临诸多关键挑战:1)3D数据的获取与标注过程复杂且成本高昂,现有的大规模3D数据集质量参差不齐,限制了模型进一步提升。尤其在特殊应用场景中,拥有高品质骨骼蒙皮属性并支持动画效果的数据极为稀缺。2)生成高质量的3D数据与纹理依然充满挑战,尤其是复杂形状和细腻材质的表现,现有技术难以全面满足高精度与逼真度的要求,并且尽管多模态生成模型具有高度灵活性,但如何确保输入的文本或图像与生成的3D数据之间的精准匹配,仍是一个重要的技术难点。3)生成模型对计算资源的需求依然巨大,特别是扩散模型的训练与推理效率问题亟待解决,以便更好地适应实际应用。4)AIGC 3D生成的数据与工业界的标准在细节质量、规范性和可编辑性等方面还存在显著差异,这成为该技术在产业中成功应用的关键挑战。

3.1.3 AIGC 3D的综述范畴

本节技术综述的范围仅以完整的3D数据生成方法为主,并不包含单独的材质/纹理生成、动作生成以及物理仿真生成等方向。

3.2 AIGC 3D技术发展

首先,3D数据的表达方式与AIGC 3D算法密切相关,因此本节首先对常见的三维数据表达方式进行整理和分类(3.2.1节)。

其次,从基于训练数据特征(3.2.2节)、基于输出数据表达方式(3.2.3节)和基于不同生成技术方法(3.2.4节)3个角度综述当前AIGC 3D的生成技术。其中,又以基于训练数据特征的技术分类为主进行展开,详细综述了技术的3个发展阶段,分别是基于2D扩散模型先验的生成方法阶段、基于2D + 3D特征混合的生成方法阶段和基于原生3D特征的生成方法阶段。之后,又从基于输出数据表达方式的技术分类和基于不同生成方法的技术分类两个角度对已有技术进行梳理和总结。除去以上技术梳理部分,还整理了主流技术方法的量化指标(3.2.5节)。

从以上技术发展变化的过程可以感知,AIGC 3D技术已经从早期依赖预训练2D生成模型、低几何材质精度、低产业可用的NeRF表达方法,快速向以2D + 3D混合/原生3D为训练数据、具有高几何材质精度、光影材质解耦以及高产业可用的3DMesh表达方式发展,算法框架也从GAN或扩散模型,发展为扩散模型 + Transformer相组合的结构发展。

最后,讨论当前技术方法存在的局限性(3.2.6节),列举了数据问题、算法问题以及产业应用局限性3个核心问题。

3.2.1 3D数据的表达方式

在AIGC 3D的研究领域,3D几何数据的有效表示是生成3D数据的关键。3D数据的表示方式大致可分为3类:显式表达、隐式表达和混合表达。相比于AIGC 2D体系中统一的2D格式,多样的3D数据表达方式显然增加了研究的复杂度,如图3所示。

1)显式表达。显式表达是对3D数据的几何形状或结构进行直接明确表达的方法。该表达方法将场景描绘为基本图元的集合,使用多种形式(包括多边形网格、点云、多层表示等)表示3D数据的表面或内部特征,具有表达能力强、易于可视化和可编辑等特点。

(1)多边形网格。作为最常用的几何表示方式,多边形网格通过一组点(顶点)、边和面的集合及它们之间的拓扑关系,构建和表达3D几何结构。该表达方式的优势在于其能够高效地进行几何处理与变

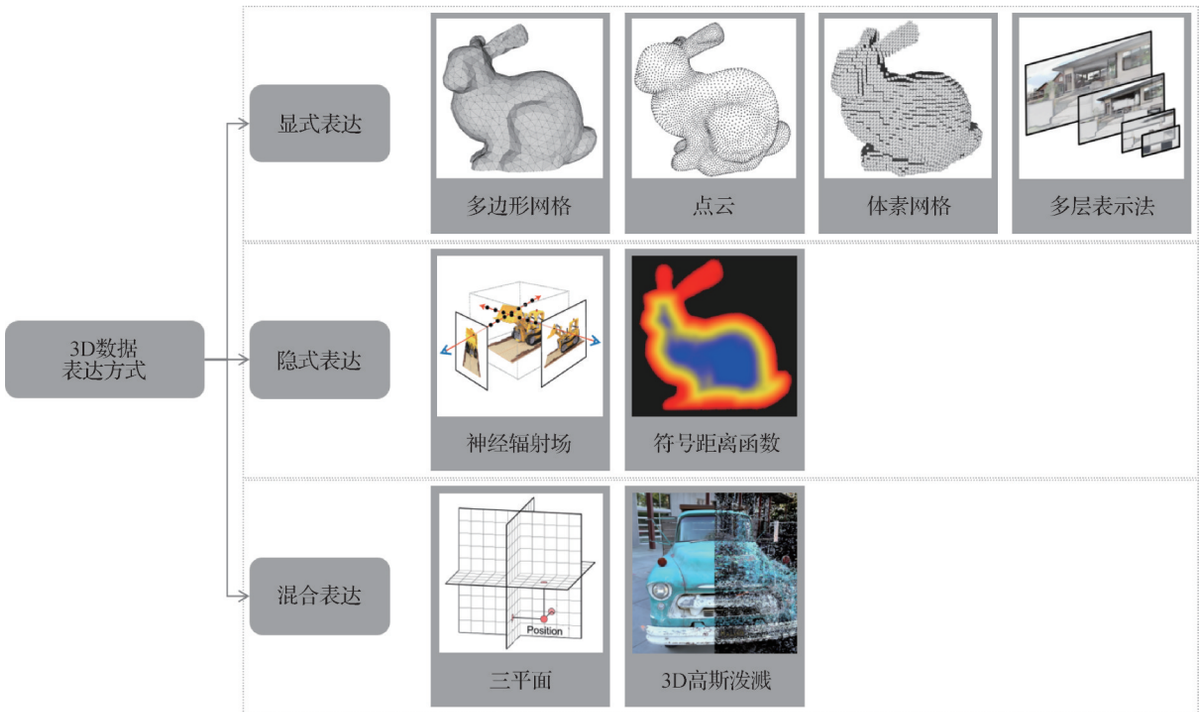

图 3　3D 数据的多种表达方式

Fig. 3　Mutiple representations methods of 3D data

形操作，具有灵活性和兼容性，与现代图形处理器（graphics processing unit，GPU）高效兼容，具有极高的渲染效率与表现效果（Akenine-Möller 等，2018）。

（2）点云。点云（吴一全 等，2024）通过在 3D 空间中绘制大量离散点表达 3D 图形的空间分布和表面特性。点云数据通常含有噪声，且由于其不规则性和无序性，与多边形网格相比，在某些应用场景中的适用性略显不足。随着技术的发展，当点云与更强大的特征重建方法（如 3DGS（3D Gaussian splatting）（Kerbl 等，2023））结合后，其对复杂场景的重建与渲染能力展现出强大潜力。

（3）体素网格。3D 空间被划分成一系列规则排列的立方体，即体素。由于体素的规律性，它与卷积神经网络（convolutional neural network，CNN）的兼容性非常高，使得它在 3D 重建和合成方面中展现出卓越性能。

（4）多层表示法。多层表示法通过多层半透明彩色图层表达 3D 场景特征，包括多层深度图像（layered depth images，LDI）（Shade 等，1998；Penner 和 Zhang，2017；Choi 等，2019；Shih 等，2020）和立体放大技术（stereo magnification）（Zhou 等，2018）等。这些方法在特定的应用场景中可以提供良好的视觉呈现效果，特别是在需要表示复杂几何结构或透明物体时表现尤为突出。

2）隐式表达。隐式表达不依赖于 3D 图形的表面几何信息，通常采用水平集或数学函数表达 3D 物体表面。隐式表示具有紧凑灵活、微分渲染管线灵活嵌入等特点。以下为常见的隐式表达方法：

（1）神经辐射场（neural radiance field，NeRF）（Mildenhall 等，2021）是一种利用隐式神经网络表示 3D 数据的方法。NeRF 支持从任意视角生成新视图，继承了体渲染技术的优势，渲染质量显著提升。在实际应用中，NeRF 通常作用于全局表示，然而其依然面临训练时间成本昂贵、渲染速度缓慢和泛化能力不足以及仅适用于静态场景等局限性。针对这个问题，NSVF（neural sparse voxel field）（Liu 等，2020）通过神经稀疏体素场的方法，Instant-NGP（Müller 等，2022）通过哈希网格的方法，实现加速训练和渲染。

（2）符号距离函数（signed distance function，SDF）是一种以 3D 物体表面作为距离场的零水平集，为该有限空间内的每个点分配一个实数值，表示该点到目标 3D 物体表面的有符号最短距离的隐式表达法。SDF 不直接存储几何表面信息，而是通过一个连续的标量场来隐式定义物体的表面，具有表面光滑且连续、高效的表面查询和适合几何操作等

特点。随着深度学习技术的发展，研究人员也探索了将两者相结合的方法，如：仅通过一个神经网络即可刻画一类物体SDF值的DeepSDF（Park等，2019）、可呈现更光滑几何表面及渲染效果的VolSDF（Yariv等，2021）以及可以更好地刻画复杂结构或自遮挡结构特征的NeuS（Wang等，2021a）。

3）混合表达。目前多数隐式表达依赖于神经辐射场（NeRF）或有向符号距离场（SDF），这种间接描述物体形状的方式难以从目标物体的显式监督中获得提升。相反，显式表达直接描述物体的几何形状、结构等属性，便于进行有效的显式监督。混合表达充分结合了这两种表达的互补优势，利用渲染算法进行视图合成，并嵌入特征以更准确地描述物体的形状和属性。以下为几种常见的混合表达方法：

（1）三平面（Chan等，2022）表示将3D体积分解为3个正交平面（*XY*、*XZ*和*YZ*平面），并在这些平面上表示3D形状的特征。现有技术（Shue等，2022；Wang等，2023b；Chen等，2023a；Gu等，2024；Zou等，2024）将三平面法和扩散模型相结合，从而生成高质量、多样性的3D数据。

（2）3D高斯泼溅（3D Gaussian splatting，3DGS）（Kerbl等，2023）通过在3D空间中的一系列加权高斯分布实现对3D场景的隐式表示，并能通过相机捕获的稀疏分布集表示复杂场景。该方法能够保证低训练成本下的高质量视觉生成效果，并且支持实时、快速和高质量的渲染。该方法也存在属性编辑与修改困难的问题，因此也成为当前的研究热点之一。

3.2.2　基于训练数据特征的技术分类

AIGC 3D技术经历了从早期的生成式对抗网络、变分自编码器等生成模型的探索，到基于Transformer架构和扩散模型技术的飞跃。早期的方法在3D数据生成上取得了一定成果，但在生成复杂场景和生成丰富细节方面存在局限。2023年以来，随着扩散模型的引入，为AIGC 3D方向的发展注入了强大的生命力。该方法不仅能生成更复杂的场景和更丰富的细节，还提高了生成结果的稳定性和泛化能力。快速成为当前最主流的技术方案之一。

以训练数据特征进行分类的技术路线，大致可分为3个方法类型：基于2D扩散模型先验的生成方法、基于2D + 3D特征混合的生成方法和基于3D特征的生成方法，如图4所示。图中，横轴为技术发表的时间，纵轴为输入的数据类型。

此外，考虑到不同行业对3D数据的需求及其特征的差异性，这些方法可以细分为针对特定对象类型的生成类别，包括单品物体（难点包括：多样、精致、几何一致、高性能等）、场景（难点包括：尺寸大、布局合理、精致等）及单品人像（在满足单品物体的基本要求上增加了可动性和表情驱动下的ID一致性），每个类别各自面临不同的技术挑战。同时，随着大语言模型等技术的发展，跨模态驱动3D生成也正在成为一种新的趋势，并与前者交叉融合。

1）基于2D扩散模型先验的生成方法。随着2D数据驱动的图像生成技术的显著进步，研究者们开始探索将这些方法扩展至3D数据的生成领域。这类方法通常是利用预训练的2D生成模型得到多视角图像，然后综合推断出这些图像所唯一对应的3D数据。由于其基于大量丰富的2D数据集，该方法相比于当时的其他方向的方法（如基于GANs/VAEs/自回归模型等）具有显著的多样性及品质优势。

（1）单品物体。在2D数据生成领域，文本到图像生成的突破性进展是通过在训练数十亿图像—文本对上训练的扩散模型实现的。然而，由于3D数据集的稀缺性，这些方法不能直接迁移到3D数据的生成上。为了应对这一挑战，出现了一种新的研究方向，即基于大规模配对图像文本数据集训练的2D扩散模型来构建3D数据。一个代表性方法是DreamFusion（Poole等，2022），其使用分数蒸馏采样（score distillation sampling，SDS）损失优化NeRF，但该方法在处理速度和渲染品质两方面存在局限性。不少研究人员针对其采用的SDS和NeRF方法分别展开新的工作。针对SDS应用效果与理论存在差距的改进研究中，ProlificDreamer（Wang等，2023c）基于变分评分蒸馏（variational score distillation，VSD）框架替代SDS方法实现了更高质量的表面纹理和细节生成效果；Text-to-3DwithCSD（Yu等，2023）发现仅使用分类器评分蒸馏（classifier score distillation，CSD）替代SDS，就足以有效完成文本到3D的生成任务。LucidDreamer（Liang等，2024）使用区间得分匹配（interval score marching，ISM）替换SDS，并与3DGS相结合的方法，提高了生成结果的纹理细节和渲染速度。针对NeRF训练过程耗时长的问题，Magic3D（Lin等，2023）通过引入双阶段优化框架优化计算速度。DreamGaussian（Tang等，2024b）和GaussianDiffusion（Li等，2024c）则通过引入基于3DGS的生成方

法提高效率。Instant3D(Li 等,2024a)提出可以仅通过一次神经网络的前向传播生成3D数据的方法,省略反复迭代优化的步骤,从而提高效率。除以上方法,同期也有其他巧妙的方法来探索更复杂效果的生成。如Fantasia3D(Chen 等,2023b)通过解耦几何和外观的建模与学习,实现效果的提升;DreamDissector(Yan 等,2025)探索了一种将多物体进行创意组合为复杂物体的方法。

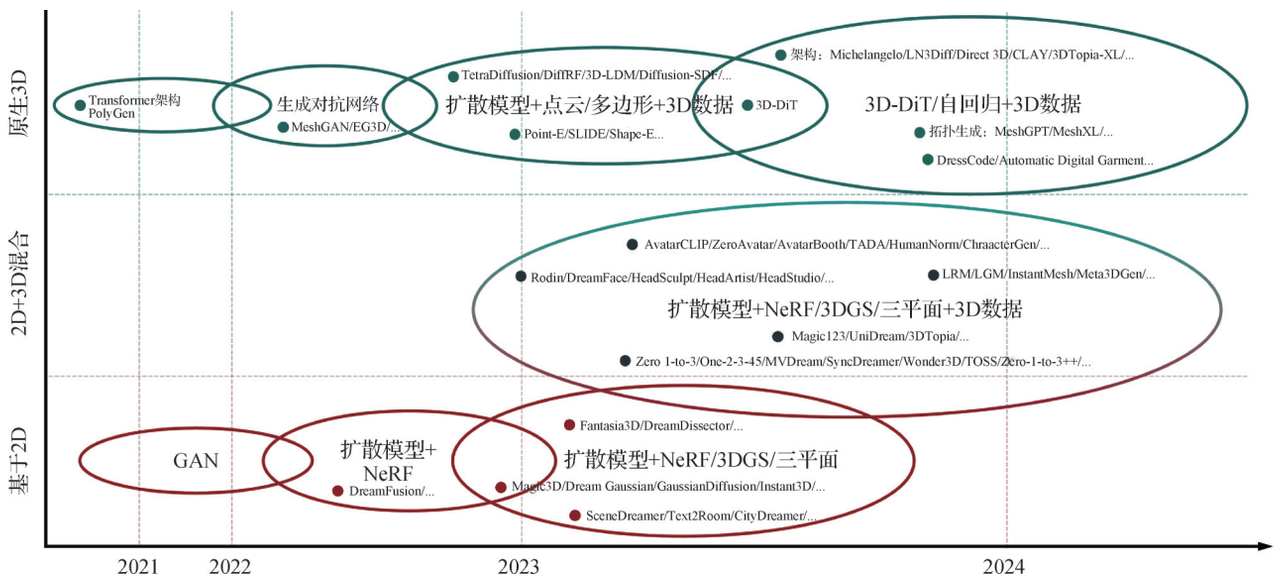

图4　AIGC 3D技术发展图谱

Fig. 4　AIGC 3D technology development chart

(2)场景。在探索2D数据到3D场景生成的领域中,主要关注从2D文本或图像信息生成3D场景几何结构与纹理的方法。这些方法旨在确保生成场景不仅在全局一致性上表现出色,而且在几何细节的准确性上也能达到高水准。具体而言,可以把场景构建任务分为两个主要类别:无边界大场景生成和室内场景的生成。无边界大场景的生成着眼于创建开阔的3D环境,更注重场景的生成效率、场景的整体一致性以及大规模数据处理时的资源优化水平。SceneDreamer(Chen 等,2023c)通过鸟瞰视角来高效处理几何和语义信息,并使用体积渲染技术生成3D场景。CityDreamer(Xie 等,2024)通过解耦建筑物和背景生成,将自然场景生成提升到城市级别。室内场景的生成则专注于从文本描述中生成3D数据。Text2Room(Höllein 等,2023)通过文本约束、多角度图像生成、深度估计和连续对齐策略方法生成带纹理的3D室内场景。进一步地,SceneTex(Chen 等,2024a)通过引入多分辨率纹理场和跨注意力编码器,改进了Text2Room在纹理细节、几何一致性和风格统一性上的不足,显著提升了生成的室内3D场景的纹理生成质量。

(3)单品人像。相对普通物体的生成,生成3D角色的任务在技术上更加复杂,因为它们不仅需要具有细致的纹理、复杂的几何结构以及角色ID的一致性,还需要能够具备逼真的动画表现。StyleSDF(Or-EI 等,2022)将SDF方法和StyleGAN2的图像生成能力相结合,相比之前的方法具有更好的生成效果。AvatarCraft(Jiang 等,2023)和DreamAvatar(Cao 等,2024)结合NeRF和扩散模型,引入对生成角色位姿的指导,显著提高了生成角色的质量。然而,这些方法存在生成过程计算资源消耗较大,且尚未能实现完整的动态动画效果。为了解决这些问题,DreamWaltz(Huang 等,2023)提出一种新的框架,通过SDS优化复杂角色的生成,并利用SMPL(a skinned multi-person linear model)模型实现角色的动画化。DreamHuman(Kolotouros 等,2023)则通过在隐式3D人体模型(imGHUM)的语义标签距离空间中优化NeRF,并结合多个细粒度损失,生成具有实例特定表面变形的可动3D角色。AvatarVerse(Zhang 等,2024c)进一步结合了ControlNet和SMPL,为角色生成提供视角一致的2D人体骨架指导,进而在生成角色的姿态和几何一致性方面取得显著提升。

HumanGaussian（Liu 等，2024e）和 HumanGaussianSplatting（Moreau 等，2024）则将 3DGS 优化技术应用于 3D 角色生成中，显著提升了生成效率。

2）基于 2D + 3D 特征混合的生成方法。这类方法旨在克服仅依赖单一 2D 数据进行生成的多种局限性，即通过结合 2D 和 3D 数据进行结合训练，提升生成内容质量。作为对比，早期的基于原生 3D 数据的生成方法受到 3D 数据集稀缺的限制，难以充分捕捉复杂的几何信息；而基于 2D 数据升维的方法生成结果缺乏一致性，在不同视角下存在几何失真和空间结构不准确的问题。

（1）单品物体。为了提升多视角一致性，Zero-1-to-3（Liu 等，2023b）通过微调扩散模型实现给定视角图像的 3D 数据生成方法。在其推动下，One-2-3-45（Liu 等，2023a）进一步利用其生成的多视图图像，经过升维方法得到高质量的纹理网格。MVDream（Shi 等，2024）、SyncDreamer（Liu 等，2024f）和 Wonder3D（Long 等，2024）引入了注意力机制。TOSS（Shi 等，2023b）引入文本作为语义信息约束生成过程，Zero-1-to-3++（Shi 等，2023a）引入条件生成和噪声调度。为了提升生成几何与材质的质量，Magic123（Qian 等，2023）、UniDream（Liu 等，2025）、3DTopia（Hong 等，2024a）等方法则将生成管线拆分为多阶段生成过程，但这些方法普遍存在对计算资源的需求较高的问题。LRM（Hong 等，2024b）采用了 Transformer 架构和三平面法，首次端到端地从模型单视图生成其所对应的 3D 表示。LGM（Tang 等，2025）将 3DGS 优化技术引入预训练过程，实现了高分辨率、快速的生成。InstantMesh（Xu 等，2024b）将 LRM 和 Zero-1-to-3++相结合，在生成速度和质量上取得了提升。SketchDream（Liu 等，2024b）将手绘草图和文本信息相结合，同时利用利用深度引导解决 2D 到 3D 的歧义问题。Meta3DGen（Bensadoun 等，2024b）通过融合 Meta3DAssetGen（Siddiqui 等，2024b）和 Meta3DTextureGen（Bensadoun 等，2024a），生成了拥有高分辨率纹理和基于 PBR 材质的 3D 数据。

（2）场景。在通过混合方法生成场景的算法中，大致又可以分为两个方向：第 1 类是通过生成场景内物体之间的布局关系，对 3D 数据进行布局规划并放置的方法。该类方法包括基于规则驱动的方法、基于自回归模型方法（Yeh 等，2012）、基于数据学习内在关系的方法（Wu 等，2019；Kar 等，2019；Devaranjan 等，2020）、基于生成式模型 GAN 的方法（Yang 等，2021）以及基于扩散模型的方法（Tang 等，2024a；Hu 等，2024a）等。此类方法的优势是生成的场景布局合理性高；缺陷是可生成场景物体的多样性相对较少（受限于预制作的 3D 数据集）。第 2 类是通过结合 2D 图像和 3D 空间的先验知识进行生成。例如，MVDiffusion（Tang 等，2023）结合全局感知方式以及像素到像素的对应关系，解决了常见的误差累积问题。ControlRoom3D（Schult 等，2024）通过输入几何和语义信息生成高质量 3D 房间网格。SceneWiz3D（Zhang 等，2024e）通过粒子群优化和 RGBD 全景扩散模型自动优化场景布局生复杂 3D 场景。此类方法结合了 2D 数据和 3D 数据的优势，适用于生成静态、具有丰富细节的场景，但也受到生成速度和计算资源的限制。

（3）单品人像。在头像生成子方向中，Rodin（Wang 等，2023b）依托于高品质 3D 半身头像数据集，结合新的扩散网络结构，实现可文本驱动的高质量的 3D 半身头像数据生成。DreamFace（Zhang 等，2023b）通过结合扩散先验模型和基于高质量 PBR 纹理数据集优化的纹理扩散模型，能够生成具有物理真实的 3D 头像。HeadSculpt（Han 等，2023）通过将 3D 头部先验知识、文本信息与扩散模型相结合，并引入头像 ID 保持，在实现细节编辑的同时，也保持了头像的 ID 特征。HeadArtist（Liu 等，2024c）则通过引入面部关键点指导 ControlNet 作为生成先验，构建了一个高效的生成流程，且该方法支持 3D 几何及纹理编辑。HeadStudio（Zhou 等，2025）通过利用语义驱动 3DGS 生成真实、灵动的头像动态效果。为了实现更加真实的 3D 人像生成，研究者们也将 2D + 3D 数据混合的方法，用于合成具有高精度细节的 3D 人像。AvatarCLIP（Hong 等，2022）基于 CLIP 引导，初步实现了人像和动态效果的生成。ZeroAvatar（Weng 等，2023）利用 3D 人体先验，实现了单图零样本的人像生成。AvatarBooth（Zeng 等，2023）采用了分别生成头像和身体并最终融合的双阶段法。Sculpt（Sanyal 等，2024）使用一种无监督学习方法，生成具有不同姿势和丰富服装类型的 3D 人像。TADA（Liao 等，2024）使用文本驱动 3D 数据和动画生成，且保持与 SMPL-X 规范一致，提高了该方法下游任务的可应用性。HumanNorm（Huang 等，2024a）提出一个双阶段基于法线自适应扩散模型和法线对

齐扩散模型的流程。StructLDM（Hu 等，2024b）是一个从图像、视频中学习的前馈 3D 生成模型。CharacterGen（Peng 等，2024）通过将输入图像信息转化为产业标准的 A-Pose 姿态，实现了角色蒙皮生成，从而更好地支持了下游商业工具链的使用，该技术方法从一定意义上讲对提高产业的生产效率具有很大帮助，第一次让人像生成的结果具有较强的实用性。

3）基于原生 3D 特征的生成方法。这类方法直接利用原生 3D 数据对生成模型进行训练，生成高质量的 3D 数据。传统的显式表示方式（如点云、体素、网格）通过几何结构直接生成具备高几何精度的 3D 场景。随着隐式表示（如神经隐式场）和混合表示的引入，这类方法得到了进一步扩展。隐式表示通过学习连续的空间函数，直接生成 3D 数据，具有生成灵活、空间压缩效率高的优势。混合表示则结合了显式和隐式方法的优点，在生成速度和精度之间取得平衡。

此方法相比于 2D 数据升维方法和 2D + 3D 混合训练方法，具有几何细节精度更高、无需考虑视角一致性问题以及处理复杂场景能力更强的显著优势。

（1）单品物体。在基于原生 3D 特征方法发展的早期，PolyGen（Nash 等，2020）采用 Transformer 架构顺序预测多边形网格顶点。EG3D（efficient geometry-aware 3D）（Chan 等，2022）、NFD（Shue 等，2022）、Control3Diff（Gu 等，2024）和 Single-Stage DiffusionNeRF（Chen 等，2023a）采用了三平面提升了生成结果的一致性和渲染效果。Neural Wavelet-domain Diffusion（Hui 等，2022）在多尺度和局部区域生成方面取得了显著进展。TetraDiffusion（Kalischek 等，2025）则使用四面体分解表示 3D 空间，有效提升了 3D 数据生成的效率。Point-E（Nichol 等，2022）使用基于点云数据训练的扩散模型生成 3D 数据。DiffRF（Müller 等，2023）在保持局部不变的情况下结合遮罩，实现被遮盖区域多样化的生成。Shap-E（Jun 和 Nichol，2023）则采用了双阶段训练方法，相比基于显式生成点云的模型 Point-E（Nichol 等，2022），在收敛速度上表现更好。3D-LDM（Nam 等，2022）、Diffusion-SDF（Li 等，2023）和 Locally Attentional SDF Diffusion（Zheng 等，2023）探索了使用符号距离场生成方法。基于 3D-DiT（Mo 等，2023）和自回归方法的架构开启了原生 3D 特征生成的新阶段。MeshGPT（Siddiqui 等，2024a）是第一种采用基于序列的方法，以自回归方式，结合图卷积编码器学习几何特征与约束，从而生成具有锐利边缘的紧凑三角形网格。在此基础之上，MeshXL（Chen 等，2024c）采用了将显式坐标表示与隐式神经嵌入相结合的表示形式。此类方法生成的 3D 数据在拓扑特征上，非常接近产业标准下制作的结果，具有很大的发展潜力。Spice-E（Sella 等，2024）则通过将文本信息与引导形状相结合，提高了生成结果的可控性。DressCode（He 等，2024）通过将大语言模型与服装缝合片生成相结合的方法，实现了简单易用的服装数据生成。Style3D（Liu 等，2024a）通过 AI 分类、启发式方法和数值优化实现了高自动化生成服装切片的初始生成方法。在原生 3D 的训练框架和训练方法的改进上，Michelangelo（Zhao 等，2023）探索了一种基于 Transformer 的形状—图像—文本对齐的 3D 生成框架；LN3Diff（Lan 等，2025）提出一种基于 3D-VAE 结构和 3D-DiT 结构的双阶段通用 3D 生成框架；Direct3D（Wu 等，2024b）提出 D3D-VAE 和 D3D-DiT，摒弃了通过生成多视图驱动 3D 生成的方法，且实现了任意图像输入均可生成高品质 3D 数据的效果；CLAY（Zhang 等，2024d）通过对大规模原生 3D 数据的精细化处理，结合 3D-DiT 结构和渐进式训练方法，完成 15 亿参数量的大模型训练，得到了极高的生成品质效果，并正快速向工业化标准的要求迈进；3DTopia-XL（Chen 等，2024e）提出一种称为 PrimX 的几何 Patch 特征编码器，结合 DiT 结构，有效提高了模型、材质及纹理的生成效果。

（2）场景。Diffusion Probabilistic Models（Lee 等，2023）通过离散类别标签表示场景。XCube（Ren 等，2024）通过分层体素潜在扩散模型生成高分辨率的稀疏 3D 体素网格。SemCity（Lee 等，2024）通过引入三平面表示，将 3D 场景压缩到 3 个正交的 2D 平面中，在提高生成效率的同时，仍能在复杂场景中保持细节和精度。

3.2.3 基于输出数据表达方式的技术分类

3D 数据的表达方式具有多种不同方法，展现出不同的技术优势和应用场景优势。

在 AIGC 3D 技术的早期，受限于对基础模型的研究，大多数方法采用的表达方法为隐式表达。隐式表达法的优势在于能够捕捉到数据多样化的特征。DreamFusion（Poole 等，2022）是此类方法的代表，在此之后，Magic3D（Lin 等，2023）、Neural

Wavelet-domain Diffusion（Hui 等，2022）、TetraDiffusion（Kalischek 等，2025）、DiffRF（Müller 等，2023）以及 Shap-E（Jun 和 Nichol，2023）等采用了各种神经辐射场的表示方法。3D-LDM（Nam 等，2022）、Diffusion-SDF（Li 等，2023）和 Locally Attentional SDF Diffusion（Zheng 等，2023）使用了符号距离场，而 Surf-D（Yu 等，2025）则使用无符号距离场，解决了符号距离场在非闭合几何数据生成结果上的不足。PVD（Zhou 等，2021）和 VolumeDiffusion（Tang 等，2024c）通过结合体积编码方法进行 3D 数据生成。

混合方法中的代表是三平面表示法和 3D 高斯泼溅法。三平面法的优势是能够与 2D 生成架构进行集成，EG3D（Chan 等，2022）、Rodin（Wang 等，2023b）、Zero-1-to-3（Liu 等，2023b）以及 Single-Stage DiffusionNeRF（Chen 等，2023a）等方法均通过三平面法生成 3D 数据并进行渲染呈现。DreamGaussian（Tang 等，2024b）、GaussianDiffusion（Li 等，2024c）、LucidDreamer（Liang 等，2024）、GVGEN（He 等，2025）、GaussianCube（Zhang 等，2024a）以及 AtlasGaussians（Yang 等，2024a）则设计了不同的高斯拟合方法，从而有效提高了 3D 数据的外观表现精度。

显式表达法的优势在于直观且结果接近产业标准。在生成式 3D 技术发展的早期，3dAAE（Zamorski 等，2020）、PointFlow（Yang 等，2019）、tree-GAN（Shu 等，2019）以及 Point-E（Nichol 等，2022）等采用了点云的数据表达方式，PolyGen（Nash 等，2020）采用了多边形的表达方式，但这些方法的表达精度还有待提升。随着技术的发展，SLIDE（Lyu 等，2023）和 3D-DiT（Mo 等，2023）使用基于点云数据训练的扩散模型生成 3D 数据。MeshGPT（Siddiqui 等，2024a）、MeshXL（Chen 等，2024c）以及 CLAY（Zhang 等，2024d）则展现出新技术架构下的惊人的生成效果。

在新技术方案的推动下，其生成结果逐渐符合于 3D 创意产业生产标准。

3.2.4　基于不同生成方法的技术分类

前馈式 3D 生成模型，包括了生成对抗网络、变分自编码器、归一化流模型、扩散模型、自回归模型以及最新的 3D-DiT 模型。tree-GAN（Shu 等，2019）通过生成对抗网络探索了点云生成，EG3D（Chan 等，2022）是一个精彩的基于生成对抗网络的研究，gDNA（Chen 等，2022）探索了生成多样性穿衣角色的方法、SurfGen（Luo 等，2021）探索了生成多样拓扑结构的方法。PointFlow（Yang 等，2019）提出一种两级分布层次结构，基于归一化流模型的点云生成方法。SDM-NET（Gao 等，2019）探索了一种生成结构化可变形网格的生成方法，3dAAE（Zamorski 等，2020）是首个端到端的点云模型生成方法，它们都基于基于自编码器。PointGrow（Sun 等，2020）、PolyGen（Nash 等，2020）和 AutoSDF（Mittal 等，2022）等方法探索了自回归模型的方法来生成 3D 数据。DreamFusion（Poole 等，2022）、Rodin（Wang 等，2023b）、DreamFace（Zhang 等，2023b）、Zero-1-to-3（Liu 等，2023b）、Fantasia3D（Chen 等，2023b）、Single-Stage DiffusionNeRF（Chen 等，2023a）、ProlificDreamer（Wang 等，2023c）、HeadSculpt（Han 等，2023）、HeadArtist（Liu 等，2024c）、HeadStudio（Zhou 等，2025）以及 DreamDissector（Yan 等，2025）等则充分探索了扩散模型的生成方法。3D-DiT（Mo 等，2023）方法创新性地将扩散模型和 Transformer 方法结合，扩展了新的模型架构，随后诞生了诸如 MeshGPT（Siddiqui 等，2024a）、MeshXL（Chen 等，2024c）以及 CLAY（Zhang 等，2024d）等新方法。

3.2.5　AIGC 3D 技术方法分类和效果评估

表 2 列举了本综述中的 AIGC 3D 技术的部分相关信息，以输入数据类型、输出数据类型和模型结构进行了分类整理。

表 3 列举了 AIGC-3D 部分关键方法的量化指标，对比项包括语义相关、3D 合理性、几何与纹理相关性、纹理细节和几何细节 5 个关键维度。

3.2.6　AIGC 3D 技术现存问题

1）AIGC 3D 技术的数据问题：

（1）3D 数据采集成本高、采集效率低。目前已知的数据采集方式包括 3D 数据采集、人工建模和程序化生成等。与 2D 数据采集方式相比，3D 数据的采集方式复杂，且采集结果的规范不统一，导致使用时需要进行大量前置处理工作，费时费力。

（2）3D 数据规格标准化程度不高，数据规范不统一。3D 数据的格式与应用场景密切相关，不同行业（如娱乐、工业、医疗、军事等）对数据的特征需求差异大。这与 2D 平面/视频数据统一的标准相比，具有显著的差异。

（3）3D 数据品质普遍不高，数据的类型分布不平衡、数据的多样性不足，开源数据的总体品质和规模还有待提高。现有的开源数据集主要可分为“2D

表 2　AIGC 3D 部分关键方法对比

Table 2　Comparison of some key AIGC 3D methods

方法	输入类型	输出类型	模型结构
DreamFusion	2D	NeRF	Diffusion
ProlificDreamer	2D	NeRF/Mesh	Diffusion
LucidDreamer	2D	3DGS	Diffusion
Magic3D	2D	Textured Mesh	Diffusion
DreamGaussian	2D	Textured Mesh	Diffusion
GaussianDiffusion	2D	3DGS	Diffusion
Instant3D	2D	Triplane	Diffusion
Fantasia3D	2D	DMTet	Diffusion
DreamAvatar	2D	NeRF	Diffusion
HumanGaussian	2D	3DGS	Diffusion
Zero-1-to-3	2D+3D	Image	Diffusion
One-2-3-45	2D+3D	Textured Mesh	Diffusion
MVDream	2D+3D	NeRF	Diffusion
Magic123	2D+3D	Textured Mesh	Diffusion
3DTopia	2D+3D	Triplane	Diffusion
LRM	2D+3D	NeRF	Transformer
MVDiffusion	2D+3D	Image	Diffusion
Rodin	2D+3D	Triplane	Diffusion
CharacterGen	2D+3D	Textured Mesh	Diffusion
PolyGen	3D	Mesh	Transformer
EG3D	3D	Triplane	GAN
Point-E	3D	Point Cloud	Diffusion
Shap-E	3D	STF	Diffusion
3D-DiT	3D	Point Colud	3D-DiT
MeshGPT	3D	Mesh	Transformer
MeshXL	3D	Mesh	Transformer
CLAY	3D	Textured Mesh	VAE+ 3D-DiT
3DTopia-XL	3D	Textured Mesh	3D-DiT

多视角数据集”：Gobjaverse（Xu 等，2024a）、Human-Normal（Ran 等，2024）；“3D 原生数据集：ShapeNet（Chang 等，2015）、SDF-StyleGAN（Zheng 等，2022）、Objaverse（Deitke 等，2022）、ShapeNetV2（Thai 等，2021）等，但这些数据从规模、品类、品质和多样性等角度来衡量，远远无法满足新的生成模型训练所需。随着行业从业者的关注和投入，这方面的情况正在不断改善，一些更符合工业标准的 3D 原生数据集正在相继出现，例如大规模 3D 静态数据集 Objaverse-XL（Deitke 等，2023）、符合产业标准且带有角色绑定属性的 3D 角色数据集 CharacterGen（Peng 等，2024）以及高精度 LightField 人像数据集 OLAT（one light at a time）（Zhang 等，2021）等。

2）AIGC 3D 技术的算法问题：

（1）生成质量。3D 模型的生成质量依赖于输入数据的精度和多样性。目前，3D 生成模型的质量存在几何形状、纹理和材质等方面的缺陷，尤其是较难以生成具有表面结构整齐清晰、合理的拓扑结构、或是高精度的细节表达等。在面向娱乐、工业等应用，要求生成的 3D 模型既具备视觉上的逼真感，又要保证物理上的一致性，但当前技术往往只能在特定任务上优化部分质量维度。

（2）生成效率。3D 模型生成的计算开销较大，尤其是涉及复杂几何形态和高分辨率纹理时。现有生成技术往往需要较高的算力支持，导致生成效率不高。程序化生成和部分基于学习的 3D 生成方法已提升了效率，但在大规模场景中应用时，生成时间仍然较长。实时生成 3D 数据的能力还有待进一步突破。

（3）生成可控性。当前的生成技术在可控性方面较为有限，特别是在生成特定场景需求的 3D 数据时，难以精确控制细节。例如：具有特定风格和一致性的模型、复杂物体的高品质 PBR 材质、骨骼蒙皮权重生成以及动画等属性常需要后期手动调节。尽管已有基于条件输入的生成模型（如通过文本、图像等控制生成结果），但在生成过程中的细节和参数控制仍需进一步细化，以满足更高层次的定制化需求。

3.3　AIGC 3D 技术在产业中的应用

3.3.1　AIGC 3D 技术的产业报告

当前，AIGC 3D 已成为生成式 AI 领域中最具前瞻性的方向之一，这一技术广泛体现在世界模型（world models）、空间智能（spatial intelligence）等概念中，成为全球科技界判断下一代人工智能发展的重要指标。从投资热度来看，美国斯坦福大学教授李飞飞的 World Labs 筹集 2.3 亿美元投入世界大模型研发，北京哇嘶嗒科技有限公司（VAST）完成数亿元人民币融资，而 A16Z 则通过投资 Yellow、Kaedim、CSM 等公司持续布局 3D 生成领域，显示出资本市场对该技术的强烈信心。

表3　AIGC 3D部分关键方法的量化数据对比

Table 3　Quantitative comparison of some key AIGC 3D methods

方法	语义相关	3D合理性	几何纹理相关性	纹理细节	几何细节
TripoSG	**1 225.4**	**1 177.2**	**1 273.4**	**1 221.1**	**1 247.2**
DreamFusion	1 000.0	1 000.0	1 000.0	1 000.0	1 000.0
DreamGaussian	1 103.5	983.1	1 108.5	1 096.0	1 122.9
Fantasia3D	966.6	882.9	977.9	1 018.8	997.6
Instant3D	1 116.2	1 052.4	1 145.7	1 115.6	1 169.0
latent-NeRF	1 168.1	1 071.3	1 173.9	1 169.5	1 179.7
Magic3D	1 018.8	908.6	986.4	1 020.4	977.5
MVDream	1 166.9	1 103.9	1 113.1	1 201.0	1 142.6
Point-E	749.5	715.9	847.8	791.6	782.0
ProlificDreamer	1 149.9	1 012.6	1 112.9	1 180.1	1 098.2
Shap-E	857.9	890.8	912.2	859.4	778.4
SJC	1 094.1	1 056.8	1 085.7	1 150.1	1 128.8
SyncDreamer	1 030.7	1 010.9	1 002.5	1 150.1	1 011.1
Wonder3D	959.6	956.3	979.7	987.8	943.8

注：加粗字体表示各列最优结果。

在技术发展方面，开源生态和商业产品均取得显著进展。作为领域的标杆企业之一，VAST与Stability AI开源的TripoSR在ComfyUI生态中表现优异，而基于TripoSR的Fast3D API则因其优异的生成速度获得市场认可，并率先推出秒级3D生成技术，提升设计生成体验，其提供的3D角色自动化资产生成+绑定+动画功能（CharacterGen（Peng等，2024））也广受好评，上线半年即促成全球用户生成超600万个模型，如图5所示；太极图形Meshy首创了高精度雕塑模式（含PBR（physically based rendering）贴图）、多种贴图风格控制以及局部贴图修复等功能，广受用户好评，据SimilarWeb的公开数据，截至2024年9月，Meshy月访问量已突破100万，深受全球游戏、影视和3D打印等行业用户的喜爱；影眸（Deemos）科技以产业化标准为目标，于2024年3月在GDC首发、6月正式上线的3D生成工具Rodin（基于CLAY（Zhang等，2024d）），在两个月内获得了超20万用户，其背后技术也获得了2024年度SIGGRAPH（Special Interest Group for Computer GRAPHICS）最佳论文提名。

以上这些新锐AIGC 3D企业在不同领域都展现出巨大的技术与产品竞争力。

从行业生态来看，AIGC 3D技术正推动形成新型的内容创作、分发和消费平台。预计未来将出现更多专注于特定垂直领域的解决方案，同时基础设施服务商、内容创作平台和应用服务提供商的协同也将更加紧密。这一技术不仅将重塑传统3D内容生产方式，更有望催生全新的商业模式和产业形态，成为驱动数字经济发展的重要引擎。

图5　通过VAST技术驱动的3D模型生成（Li等，2025）

Fig. 5　3D model generated by VAST's technology (Li et al., 2025)

3.3.2　AIGC 3D技术在产业中的应用及问题

传统3D建模行业面临多个显著的痛点和挑战。首先，是成本高昂的问题；其次，3D建模是一个技术

与艺术密集型的领域；最后，以个人需求定制的3D模型成本非常昂贵。

因此，AIGC 3D技术在产业中的应用涵盖了多个领域，展现出显著的创新价值和发展潜力，可以称为产业+3D生成。

游戏开发领域，AIGC 3D技术正在重塑传统的开发流程，开发者可以从简单的文本描述或2D概念原画快速生成高质量的3D角色、场景和道具模型，显著提升开发效率。统计数据显示，采用AIGC 3D技术的游戏公司资产制作效率平均提升200%，开发成本降低40%。在游戏行业，同时还催生了融合进化、战斗塔防等创新玩法，为游戏产业注入新的活力。如VAST与网易、字节跳动、腾讯、微软等企业合作；Meshy与SEGA、Snap、Supercell等游戏公司合作；影眸科技与Unity中国合作。

3D打印领域，AIGC 3D技术正在推动家用消费级市场的快速发展。通过将2D图像或文本描述直接转换为可打印的3D模型文件，建模时间缩短了70～80%，极大地降低了使用门槛。例如VAST和影眸分别与拓竹科技的合作，通过3D打印技术，激发用户创意并扩大市场影响力。

除此之外，AIGC 3D技术凭借其展现出其高品质自动化生成的特点，还应用于工业设计、具身智能—机器人等严肃领域，如VAST与高砖积木的合作及与机器人公司优必选的合作等。

随着技术的不断进步，AIGC 3D在各个领域的应用将更加深入，推动多个产业的升级和创新发展。

虽然AIGC 3D技术已经在行业中有了广泛应用，但依然面临着3个方面的问题：1）生成结果与工业标准还存在有较大差距；2）不同产业需求对AIGC 3D技术的应用要求存在较大差异；3）满足行业专家与大众需求之间的差异性问题。

首先，成熟的产业需求对AIGC 3D技术生成结果的高标准和高要求。传统的3D制作流程具有明确且多样的规范与品质要求。通过AIGC 3D技术生成的3D数据在满足多样的数据标准和生成品质上仍有待提高。例如几何结构特征（如建模领域的软/硬边效果）、几何拓扑（如生物建模需要拓扑与其运动特征相符合）、几何UV坐标（要求具有分布均匀、边界整齐、分布无明显扭曲等特征）、几何蒙皮权重（蒙皮设计是一种高度依赖设计师经验的低效率设计工作）、材质设定（材质语义与材质参数要求高度匹配）、纹理分辨率及细节（分辨率达到4 K且细节丰富、规则）以及纹理分层（符合设计师设计工作标准规范）等。

其次，多样的产业需求对AIGC 3D技术提出高适配性要求。不同产业对AIGC 3D技术的需求存在较大的行业差异（如游戏行业、影视行业、3D打印行业之间存在的差异性），这导致该技术必须针对不同行业间的需求差异实现不同的技术适配。

最后，满足行业专家需求与满足大众创意需求之间的差异性。当前的AIGC 3D技术面临着来自于专业创作者（追求高专业性、高可控化和高品质）与普通创作者（追求高易用性、高个性化和高性价比）需求之间差异性的挑战，这种显著需求差异为AIGC 3D技术的行业落地带来了挑战。

4 AIGC 2D/3D技术发展趋势

4.1 内容生产技术的发展历史与趋势

在互联网早期阶段，内容创作主要由专业团队主导，媒体公司、广播电视机构和电影制作公司通过传统渠道制作与分发内容，这种模式称为PGC。随着Web 2.0技术的兴起，互联网进入一个交互式和社交化的新时代，博客、论坛和社交媒体平台迅速崛起，UGC开始占据重要位置。普通用户通过这些平台创作并分享视频、图像和文章，打破了传统内容创作的壁垒。然而，UGC的黄金期并未持续太久，传统媒体公司迅速调整策略，在UGC平台上发布内容，以吸引更多受众。与此同时，UGC创作者逐步走向专业化，创作出接近专业水准的作品，模糊了专业与业余创作之间的界限。进入2010年代后期，智能手机的普及和社交媒体应用的兴起，使用户能够随时随地制作和分享内容，短视频与即时分享模式成为主流。这种即时性与社交化的内容创作方式改变了用户消费内容的习惯，生产效率和分发速度大幅提升。然而，与Web 2.0时代的趋势相似，专业创作者的涌入提升了消费者的审美标准，并再一次夺回了大众的注意力与消费价值。值得注意的是，这一时期的内容消费开始分化为两股不同的力量：专业创作者提供的高品质高附加值的内容精品和非专业创作者提供的高个性化内容快销品。到了2020年代，生成式人工智能技术的迅速发展，为非专业创作者提供了全新的创作方式，拓展了创作潜力。

回顾互联网时代内容创作的发展历程，可以发现其呈现出螺旋上升的趋势特征：1）新技术的出现、发展为创意内容的品质提升提供新的空间，主要表现在生产力提升、生产成本下降、传播成本下降、创意形态的增多等多个方面。2）非专业用户群体广泛使用新技术进行快速创作与试错，并由此获得早期的技术红利，少部分非专业创作者由此逐步发展为专业创作者。3）专业用户群体虽起步较晚，但在受到市场影响后会进行反馈，最终凭借高品质内容吸引消费者注意。4）产业进入技术的成熟期，等待下一轮新技术的出现。这个过程构成了完整技术产业化的发展循环。

4.2 AIGC 2D/3D技术的发展趋势

结合创意产业近10年间的发展特征（如短视频平台、即时通信平台以及专业化工具平台等），可以预见，AIGC 2D/3D技术的发展正在经历或即将经历以下几个阶段，其中阶段1和阶段2是已经发生的，当前正处于阶段3，而阶段4和阶段5则是对该技术发展趋势的展望。

1）阶段1。伴随着2023年AIGC 2D在技术上的革新，其多样性、高质量的生成能力，引发了产业与消费者的广泛关注及不同角度不同层面的讨论，其中既包含了对其效果的赞誉，也包含了大量关于版权保护问题、社会伦理问题以及从业者就业风险问题等质疑和批评。但不可否认的是，越来越多的用户开始主动或被动地接纳这一新技术的出现，并在日常生活和工作中尝试使用该技术，各种作品层出不穷并在社交媒体上获得了广泛关注。同年，AIGC 3D技术获得了生成品质上的突破，也开始逐步被3D内容创作者所关注，但整体的关注度弱于AIGC 2D。

2）阶段2。随着使用者的增多，尤其专业用户的加入并在专业场景下的使用，是AIGC 2D/3D技术开始正式进入产业应用的标志。不同用户群的需求也开始发生分化：针对专业用户，须满足其在工业化生产过程中的基本需求：标准化、高品质和稳定可控等。同时，新技术需要逐步融入到现有的工作流程中，避免成为生产流程中新的瓶颈点；针对非专业用户，须能够围绕其个人专属特征，快速且廉价地生成富含个性化特征的内容，以满足其社交属性的需求。同时，需要提供易用且稳定的技术使用方式，降低技术使用门槛，提高技术的容错性。

3）阶段3。相比于传统的计算机图形图像学技术，AIGC 2D/3D技术在生成结果上，目前缺少必要的可控性、可编辑性和高品质性，这导致了该技术在泛产业中的应用进展速度普遍不高，行业整体反馈逐步回归理性，对该技术的评价态度也趋于合理。在这一阶段中，AIGC 2D/3D的技术推动者们在努力提升该技术存在的显著不足，使用者们则在不断尝试更好的技术应用方法，以期减少上述3方面的问题。

4）阶段4。AIGC 2D/3D技术通过解决可控性、可编辑性和高品质性3大问题，并最终殊途同归发展为一套统一的AIGC技术，这一系列技术进步则再次激发了创作者的热情，各类作品层出不穷。然而，随着技术发展带来生产力提升，也衍生出了技术滥用的负面效应：大量低质作品充斥市场，并且这一现象将持续一段时间。面对泛滥的低质作品，观众最终会逐渐厌倦，并重新回归对艺术作品内在美的价值追求。在这一阶段的中后期，中高水平的专业艺术作品将重新占据主流消费市场。

5）阶段5。当大部分算法问题得到解决后，AIGC技术将面临一个更为根本且严峻的挑战：人的艺术原创性从何而来？如果这一问题得不到有效解决，AI所创作的作品终将缺乏灵魂与深度，难以真正触动人心。因为人类的艺术原创性，源自于个人的独特经历、情感、思想以及与社会和文化的互动，正是这种独特的生命与情感体验积累，赋予了艺术作品深刻的意义与价值，这是当前AI无法简单复制的。

5 结 语

AIGC技术的发展使内容创作变得更加自动化、多样化，同时也将推动新的互动体验模式的诞生。可以预见，随着AIGC 2D/3D技术的进一步发展，未来的内容创作生态将发生深刻变革。

致谢：本文由中国图象图形学学会数码艺术专业委员会组织撰写，该专业委员会链接为https://www.csig.org.cn/16/201612/49318.html，在此表示感谢。

参考文献（References）

Akenine-Möller T, Haines E, Hoffman N, Pesce A, Iwanicki M and Hillaire S. 2018. Real-Time Rendering. 4th ed. Boca Raton: CRC

Press

Bensadoun R, Kleiman Y, Azuri I, Harosh O, Vedaldi A, Neverova N and Gafni O. 2024a. Meta 3D TextureGen: fast and consistent texture generation for 3D objects [EB/OL]. [2025-01-03]. https://arxiv.org/pdf/2407.02430.pdf

Bensadoun R, Monnier T, Kleiman Y, Kokkinos F, Siddiqui Y, Kariya M, Harosh O, Shapovalov R, Graham B, Garreau E, Karnewar A, Cao A, Azuri I, Makarov I, Le E T, Toisoul A, Novotny D, Gafni O, Neverova N and Vedaldi A. 2024b. Meta 3D gen [EB/OL]. [2025-01-03]. https://arxiv.org/pdf/2407.02599.pdf

Blattmann A, Dockhorn T, Kulal S, Mendelevitch D, Kilian M, Lorenz D, Levi Y, English Z, Voleti V, Letts A, Jampani V and Rombach R. 2023. Stable video diffusion: scaling latent video diffusion models to large datasets [EB/OL]. [2025-01-03]. https://arxiv.org/pdf/2311.15127.pdf

Brooks T, Holynski A and Efros A A. 2023. InstructPix2Pix: learning to follow image editing instructions//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE: 18392-18402 [DOI: 10.1109/CVPR52729.2023.01764]

Cao Y K, Cao Y P, Han K, Shan Y and Wong K Y K. 2024. DreamAvatar: text-and-shape guided 3D human avatar generation via diffusion models//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE: 958-968 [DOI: 10.1109/CVPR52733.2024.00097]

Chan E R, Lin C Z, Chan M A, Nagano K, Pan B X, de Mello S, Gallo O, Guibas L, Tremblay J, Khamis S, Karras T and Wetzstein G. 2022. Efficient geometry-aware 3D generative adversarial networks//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE: 16102-16112 [DOI: 10.1109/CVPR52688.2022.01565]

Chang A X, Funkhouser T, Guibas L, Hanrahan P, Huang Q X, Li Z M, Savarese S, Savva M, Song S R, Su H, Xiao J X, Yi L and Yu F. 2015. ShapeNet: an information-rich 3D model repository [EB/OL]. [2025-01-03]. https://arxiv.org/pdf/1512.03012.pdf

Chen D Z, Li H X, Lee H Y, Tulyakov S and Nießner M. 2024a. SceneTex: high-quality texture synthesis for indoor scenes via diffusion priors//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE: 21081-21091 [DOI: 10.1109/CVPR52733.2024.01992]

Chen H S, Gu J T, Chen A P, Tian W, Tu Z W, Liu L J and Su H. 2023a. Single-stage diffusion NeRF: a unified approach to 3D generation and reconstruction//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE: 2416-2425 [DOI: 10.1109/ICCV51070.2023.00229]

Chen H X, Zhang Y, Cun X D, Xia M H, Wang X T, Weng C and Shan Y. 2024b. VideoCrafter2: overcoming data limitations for high-quality video diffusion models//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE: 7310-7320 [DOI: 10.1109/CVPR52733.2024.00698]

Chen R, Chen Y W, Jiao N X and Jia K. 2023b. Fantasia3D: disentangling geometry and appearance for high-quality text-to-3D content creation//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE: 22189-22199 [DOI: 10.1109/ICCV51070.2023.02033]

Chen S J, Chen X, Pang A Q, Zeng X F, Cheng W, Fu Y J, Yin F K, Wang Y R, Wang Z B, Zhang C, Yu J Y, Yu G, Fu B and Chen T. 2024c. MeshXL: neural coordinate field for generative 3D foundation models [EB/OL]. [2025-01-03]. https://arxiv.org/pdf/2405.20853.pdf

Chen T H, Chan J W, Shiu H S, Yen S H, Yeh C H and Liu Y L. 2024d. NaRCan: natural refined canonical image with integration of diffusion prior for video editing [EB/OL]. [2025-01-03]. https://arxiv.org/pdf/2406.06523.pdf

Chen X, Jiang T J, Song J, Yang J L, Black M J, Geiger A and Hilliges O. 2022. gDNA: towards generative detailed neural avatars//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE: 20395-20405 [DOI: 10.1109/CVPR52688.2022.01978]

Chen Z X, Tang J X, Dong Y H, Cao Z A, Hong F Z, Lan Y S, Wang T F, Xie H Z, Wu T, Saito S, Pan L, Lin D H and Liu Z W. 2024e. 3DTopia-XL: scaling high-quality 3D asset generation via primitive diffusion [EB/OL]. [2025-01-03]. https://arxiv.org/pdf/2409.12957.pdf

Chen Z X, Wang G C and Liu Z W. 2023c. SceneDreamer: unbounded 3D scene generation from 2D image collections. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(12): 15562-15576 [DOI: 10.1109/TPAMI.2023.3321857]

Choi I, Gallo O, Troccoli A, Kim M H and Kautz J. 2019. Extreme view synthesis//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea (South): IEEE: 7780-7789 [DOI: 10.1109/ICCV.2019.00787]

Cong Y R, Xu M M, Simon C, Chen S F, Ren J W, Xie Y P, Perez-Rua J M, Rosenhahn B, Xiang T and He S. 2024. FLATTEN: optical flow-guided attention for consistent text-to-video editing [EB/OL]. [2025-01-03]. https://arxiv.org/pdf/2310.05922.pdf

Crowson K, Baumann S A, Birch A, Abraham T M, Kaplan D Z and Shippole E. 2024. Scalable high-resolution pixel-space image synthesis with hourglass diffusion transformers [EB/OL]. [2025-01-03]. https://arxiv.org/pdf/2401.11605.pdf

Deitke M, Liu R S, Wallingford M, Ngo H, Michel O, Kusupati A, Fan A, Laforte C, Voleti V, Gadre S Y, VanderBilt E, Kembhavi A, Vondrick C, Gkioxari G, Ehsani K, Schmidt L and Farhadi A. 2023. Objaverse-XL: a universe of 10M+ 3D objects//Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc.: 35799-35813 [DOI: 10.5555/3666122.3667676]

Deitke M, Schwenk D, Salvador J, Weihs L, Michel O, VanderBilt E, Schmidt L, Ehsani K, Kembhavi A and Farhadi A. 2022. Objaverse: a universe of annotated 3D objects [EB/OL]. [2025-01-03]. https://arxiv.org/pdf/2212.08051.pdf

Devaranjan J, Kar A and Fidler S. 2020. Meta-Sim2: unsupervised learning of scene structure for synthetic data generation//Proceedings of the 16th European Conference on Computer Vision —— ECCV 2020. Glasgow, UK: Springer: 715-733 [DOI: 10.1007/978-3-030-58520-4_42]

Dinh L, Sohl-Dickstein J and Bengio S. 2017. Density estimation using real NVP [EB/OL]. [2025-01-03]. https://arxiv.org/pdf/1605.08803.pdf

Esser P, Kulal S, Blattmann A, Entezari R, Müller J, Saini H, Levi Y, Lorenz D, Sauer A, Boesel F, Podell D, Dockhorn T, English Z, Lacey K, Goodwin A, Marek Y and Rombach R. 2024. Scaling rectified flow transformers for high-resolution image synthesis [EB/OL]. [2025-01-03]. https://arxiv.org/pdf/2403.03206.pdf

Fan F D, Luo C J, Gao W L and Zhan J F. 2023. AIGCBench: comprehensive evaluation of image-to-video content generated by AI. BenchCouncil Transactions on Benchmarks, Standards and Evaluations, 3(4): #100152 [DOI: 10.1016/j.tbench.2024.100152]

Feng R Y, Weng W M, Wang Y H, Yuan Y H, Bao J M, Luo C, Chen Z B and Guo B N. 2024. CCEdit: creative and controllable video editing via diffusion models//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE: 6712-6722 [DOI: 10.1109/CVPR52733.2024.00641]

Gao L, Yang J, Wu T, Yuan Y J, Fu H B, Lai Y K and Zhang H. 2019. SDM-NET: deep generative network for structured deformable mesh. ACM Transactions on Graphics, 38(6): #243 [DOI: 10.1145/3355089.3356488]

Goodfellow I J, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A and Bengio Y. 2014. Generative adversarial networks [EB/OL]. [2025-01-03]. https://arxiv.org/pdf/1406.2661.pdf

Gu J T, Gao Q Z, Zhai S F, Chen B Q, Liu L J and Susskind J. 2024. Control3Diff: learning controllable 3D diffusion models from single-view images//Proceedings of 2024 International Conference on 3D Vision (3DV). Davos, Switzerland: IEEE: 685-696 [DOI: 10.1109/3DV62453.2024.00030]

Guo J Z, Zhang D Y, Liu X Q, Zhong Z Z, Zhang Y, Wan P F and Zhang D. 2025. LivePortrait: efficient portrait animation with stitching and retargeting control [EB/OL]. [2025-01-03]. https://arxiv.org/pdf/2407.03168.pdf

Guo X, Zheng M W, Hou L, Gao Y, Deng Y F, Wan P F, Zhang D, Liu Y F, Hu W M, Zha Z J, Huang H B and Ma C Y. 2024a. I2V-Adapter: a general image-to-video adapter for diffusion models// Proceedings of 2024 ACM SIGGRAPH Conference Papers. Denver, USA: Association for Computing Machinery: #112 [DOI: 10.1145/3641519.3657407]

Guo Y W, Yang C Y, Rao A Y, Liang Z Y, Wang Y H, Qiao Y, Agrawala M, Lin D H and Dai B. 2024b. AnimateDiff: animate your personalized text-to-image diffusion models without specific tuning [EB/OL]. [2025-01-03]. https://arxiv.org/pdf/2307.04725.pdf

Han X, Cao Y K, Han K, Zhu X T, Deng J K, Song Y Z, Xiang T and Wong K Y. 2023. HeadSculpt: crafting 3D head avatars with text// Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc.: 4915-4936

He K, Yao K X, Zhang Q X, Yu J Y, Liu L J and Xu L. 2024. DressCode: autoregressively sewing and generating garments from text guidance. ACM Transactions on Graphics, 43(4): #72 [DOI: 10.1145/3658147]

He X L, Chen J Y, Peng S D, Huang D, Li Y G, Huang X S, Yuan C, Ouyang W L and He T. 2025. GVGEN: text-to-3D generation with volumetric representation//Proceedings of the 18th European Conference on Computer Vision-ECCV 2024. Milan, Italy: Springer: 463-479 [DOI: 10.1007/978-3-031-73242-3_26]

He Y Q, Xia M H, Chen H X, Cun X D, Gong Y, Xing J B, Zhang Y, Wang X T, Weng C, Shan Y and Chen Q F. 2023. Animate-a-story: storytelling with retrieval-augmented video generation [EB/OL]. [2025-01-03]. https://arxiv.org/pdf/2307.06940.pdf

Ho J, Jain A and Abbeel P. 2020. Denoising diffusion probabilistic models//Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: 6840-6851

Ho J and Salimans T. 2022. Classifier-free diffusion guidance [EB/OL]. [2025-01-03]. https://arxiv.org/pdf/2207.12598.pdf

Höllein L, Cao A, Owens A, Johnson J and Nießner M. 2023. Text2Room: extracting textured 3D meshes from 2D text-to-image models//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE: 7875-7886 [DOI: 10.1109/ICCV51070.2023.00727]

Hong F Z, Tang J X, Cao Z A, Shi M, Wu T, Chen Z X, Yang S, Wang T F, Pan L, Lin D H and Liu Z W. 2024a. 3DTopia: large text-to-3D generation model with hybrid diffusion priors [EB/OL]. [2025-01-03]. https://arxiv.org/pdf/2403.02234.pdf

Hong F Z, Zhang M Y, Pan L, Cai Z G, Yang L and Liu Z W. 2022. AvatarCLIP: zero-shot text-driven generation and animation of 3D avatars. ACM Transactions on Graphics, 41(4): #161 [DOI: 10.1145/3528223.3530094]

Hong Y C, Zhang K, Gu J X, Bi S, Zhou Y, Liu D F, Liu F, Sunkavalli K, Bui T and Tan H. 2024b. LRM: large reconstruction model for single image to 3D [EB/OL]. [2025-01-03]. https://arxiv.org/pdf/2311.04400.pdf

Hu E J, Shen Y L, Wallis P, Allen-Zhu Z, Li Y Z, Wang S A, Wang L and Chen W Z. 2021. LoRA: low-rank adaptation of large language

models［EB/OL］.［2025-01-03］. https://arxiv.org/pdf/2106.09685.pdf

Hu L. 2024. Animate anyone: consistent and controllable image-to-video synthesis for character animation//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE: 8153-8163［DOI: 10.1109/CVPR52733.2024.00779］

Hu S Y, Arroyo D M, Debats S, Manhardt F, Carlone L and Tombari F. 2024a. Mixed diffusion for 3D indoor scene synthesis［EB/OL］.［2025-01-03］. https://arxiv.org/pdf/2405.21066.pdf

Hu T, Hong F Z and Liu Z W. 2024b. StructLDM: structured latent diffusion for 3D human generation［EB/OL］.［2025-01-03］. https://arxiv.org/pdf/2404.01241.pdf

Hu Y S, Chen Z Z and Luo C. 2023. LaMD: latent motion diffusion for video generation［EB/OL］.［2025-01-03］. https://arxiv.org/pdf/2304.11603.pdf

Huang X, Shao R Z, Zhang Q, Zhang H W, Feng Y, Liu Y B and Wang Q. 2024a. HumanNorm: learning normal diffusion model for high-quality and realistic 3D human generation//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE: 4568-4577［DOI: 10.1109/CVPR52733.2024.00437］

Huang Y K, Wang J N, Zeng A L, Cao H, Qi X B, Shi Y K, Zha Z J and Zhang L. 2023. DreamWaltz: make a scene with complex 3D animatable avatars//Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc.: 4566-4584

Huang Z Q, He Y N, Yu J S, Zhang F, Si C Y, Jiang Y M, Zhang Y H, Wu T X, Jin Q Y, Chanpaisit N, Wang Y H, Chen X Y, Wang L M, Lin D H, Qiao Y and Liu Z W. 2024b. VBench: comprehensive benchmark suite for video generative models//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE: 21807-21818［DOI: 10.1109/CVPR52733.2024.02060］

Hui K H, Li R H, Hu J Y and Fu C W. 2022. Neural wavelet-domain diffusion for 3D shape generation//Proceedings of 2022 SIGGRAPH Asia Conference Papers. Daegu, Republic of Korea: Association for Computing Machinery: #24［DOI: 10.1145/3550469.3555394］

Jain A, Xie A and Abbeel P. 2023. VectorFusion: text-to-SVG by abstracting pixel-based diffusion models//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE: 1911-1920［DOI: 10.1109/CVPR52729.2023.00190］

Jiang R X, Wang C, Zhang J B, Chai M L, He M M, Chen D D and Liao J. 2023. AvatarCraft: transforming text into neural human avatars with parameterized shape and pose control//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE: 14325-14336［DOI: 10.1109/ICCV51070.2023.01322］

Jun H and Nichol A. 2023. Shap.E: generating conditional 3D implicit functions［EB/OL］.［2025-01-03］. https://arxiv.org/pdf/2305.02463.pdf

Kalischek N, Peters T, Wegner J D and Schindler K. 2025. TetraDiffusion: tetrahedral diffusion models for 3D shape generation//Proceedings of the 18th European Conference on Computer Vision. Milan, Italy: Springer-Verlag: 357-373［DOI: 10.1007/978-3-031-73668-1_21］

Kaplan J, McCandlish S, Henighan T, Brown T B, Chess B, Child R, Gray S, Radford A, Wu J and Amodei D. 2020. Scaling laws for neural language models［EB/OL］.［2025-01-03］. https://arxiv.org/pdf/2001.08361.pdf

Kar A, Prakash A, Liu M Y, Cameracci E, Yuan J, Rusiniak M, Acuna D, Torralba A and Fidler S. 2019. Meta-Sim: learning to generate synthetic datasets//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea (South): IEEE: 4550-4559［DOI: 10.1109/ICCV.2019.00465］

Kara O, Kurtkaya B, Yesiltepe H, Rehg J M and Yanardag P. 2024. RAVE: randomized noise shuffling for fast and consistent video editing with diffusion models//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE: 6507-6516［DOI: 10.1109/CVPR52733.2024.00622］

Karras T, Laine S and Aila T. 2019. A style-based generator architecture for generative adversarial networks//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE: 4396-4405［DOI: 10.1109/CVPR.2019.00453］

Kerbl B, Kopanas G, Leimkühler T and Drettakis G. 2023. 3D Gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics, 42(4): #139［DOI: 10.1145/3592433］

Kingma D P and Welling M. 2022. Auto-encoding variational bayes［EB/OL］.［2025-01-03］. https://arxiv.org/pdf/1312.6114.pdf

Kolotouros N, Alldieck T, Zanfir A, Bazavan E G, Fieraru M and Sminchisescu C. 2023. DreamHuman: animatable 3D avatars from text//Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc.: 10516-10529

Kondratyuk D, Yu L J, Gu X Y, Lezama J, Huang J, Schindler G, Hornung R, Birodkar V, Yan J, Chiu M C, Somandepalli K, Akbari H, Alon Y, Cheng Y, Dillon J, Gupta A, Hahn M, Hauth A, Hendon D, Martinez A, Minnen D, Sirotenko M, Sohn K, Yang X, Adam H, Yang M H, Essa I, Wang H S, Ross D A, Seybold B and Jiang L. 2024. VideoPoet: a large language model for zero-shot video generation［EB/OL］.［2025-01-03］. https://arxiv.org/pdf/2312.14125.pdf

Lan Y S, Hong F Z, Yang S, Zhou S C, Meng X Y, Dai B, Pan X G and Loy C C. 2025. LN3Diff: scalable latent neural fields diffusion for speedy 3D generation//Proceedings of the 18th European Confer-

ence on Computer Vision-ECCV 2024. Milan, Italy: Springer-Verlag: 112-130 [DOI: 10.1007/978-3-031-73235-5_7]

Lee J, Im W, Lee S and Yoon S E. 2023. Diffusion probabilistic models for scene-scale 3D categorical data [EB/OL]. [2025-01-03]. https://arxiv.org/pdf/2301.00527.pdf

Lee J, Lee S, Jo C, Im W, Seon J and Yoon S E. 2024. SemCity: semantic scene generation with triplane diffusion//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE: 28337-28347 [DOI: 10.1109/CVPR52733.2024.02677]

Li M, Zhou P, Liu J W, Keppo J, Lin M, Yan S C and Xu X Y. 2024a. Instant3D: instant text-to-3D generation. International Journal of Computer Vision, 132(10): 4456-4472 [DOI: 10.1007/s11263-024-02097-5]

Li M H, Duan Y Q, Zhou J and Lu J W. 2023. Diffusion-SDF: text-to-shape via voxelized diffusion//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE: 12642-12651 [DOI: 10.1109/CVPR52729.2023.01216]

Li M M, Li Y, Yang T Y, Liu Y F, Yue D X, Lin Z H and Xu D. 2024b. A video is worth 256 bases: spatial-temporal expectation-maximization inversion for zero-shot video editing//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE: 7528-7537 [DOI: 10.1109/CVPR52733.2024.00719]

Li X H, Wang H B and Tseng K K. 2024c. GaussianDiffusion: 3D Gaussian splatting for denoising diffusion probabilistic models with structured noise [EB/OL]. [2025-01-03]. https://arxiv.org/pdf/2311.11221.pdf

Li X R, Ma C, Yang X K and Yang M H. 2024d. VidToMe: video token merging for zero-shot video editing//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE: 7486-7495 [DOI: 10.1109/CVPR52733.2024.00715]

Li Y G, Zou Z X, Liu Z X, Wang D H, Liang Y, Yu Z P, Liu X C, Guo Y C, Liang D, Ouyang W L and Cao Y P. 2025. TripoSG: High-Fidelity 3D Shape Synthesis using Large-Scale Rectified Flow Models [EB/OL]. [2025-01-03]. https://arxiv.org/pdf/2502.06608.pdf

Li Z Q, Tucker R, Snavely N and Holynski A. 2024e. Generative image dynamics//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE: 24142-24153 [DOI: 10.1109/CVPR52733.2024.02279]

Liang Y X, Yang X, Lin J T, Li H D, Xu X G and Chen Y C. 2024. LucidDreamer: towards high-fidelity text-to-3D generation via interval score matching//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE: 6517-6526 [DOI: 10.1109/CVPR52733.2024.00623]

Liao T T, Yi H W, Xiu Y L, Tang J X, Huang Y Y, Thies J and Black M J. 2024. TADA! Text to animatable digital avatars//Proceedings of 2024 International Conference on 3D Vision (3DV). Davos, Switzerland: IEEE: 1508-1519 [DOI: 10.1109/3DV62453.2024.00150]

Lin C H, Gao J, Tang L M, Takikawa T, Zeng X H, Huang X, Kreis K, Fidler S, Liu M Y and Lin T Y. 2023. Magic3D: high-resolution text-to-3D content creation//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE: 300-309 [DOI: 10.1109/CVPR52729.2023.00037]

Lipman Y, Chen R T Q, Ben-Hamu H, Nickel M and Le M. 2023. Flow matching for generative modeling [EB/OL]. [2025-01-03]. https://arxiv.org/pdf/2210.02747.pdf

Liu C, Xu W W, Yang Y and Wang H M. 2024a. Automatic digital garment initialization from sewing patterns. ACM Transactions on Graphics, 43(4): #74 [DOI: 10.1145/3658128]

Liu F L, Fu H B, Lai Y K and Gao L. 2024b. SketchDream: sketch-based text-to-3D generation and editing. ACM Transactions on Graphics, 43(4): #44 [DOI: 10.1145/3658120]

Liu H Y, Wang X, Wan Z Y, Shen Y J, Song Y B, Liao J and Chen Q F. 2024c. HeadArtist: text-conditioned 3D head generation with self score distillation//Proceedings of 2024 ACM SIGGRAPH Conference Papers. Denver, USA: Association for Computing Machinery: #16 [DOI: 10.1145/3641519.3657512]

Liu L J, Gu J T, Lin K Z, Chua T S and Theobalt C. 2020. Neural sparse voxel fields//Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: 15651-15663

Liu M H, Xu C, Jin H A, Chen L H, T M V, Xu Z X and Su H. 2023a. One-2-3-45: any single image to 3D mesh in 45 seconds without per-shape optimization//Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc.: 22226-22246

Liu R S, Wu R D, van Hoorick B, Tokmakov P, Zakharov S and Vondrick C. 2023b. Zero-1-to-3: zero-shot one image to 3D object//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE: 9264-9275 [DOI: 10.1109/ICCV51070.2023.00853]

Liu S T, Zhang Y C, Li W B, Lin Z and Jia J Y. 2024d. Video-P2P: video editing with cross-attention control//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE: 8599-8608 [DOI: 10.1109/CVPR52733.2024.00821]

Liu V, Long T, Raw N and Chilton L. 2023c. Generative disco: text-to-video generation for music visualization [EB/OL]. [2025-01-03]. https://arxiv.org/pdf/2304.08551.pdf

Liu X, Zhan X H, Tang J X, Shan Y, Zeng G, Lin D H, Liu X H and Liu Z W. 2024e. HumanGaussian: text-driven 3D human generation with Gaussian splatting//Proceedings of 2024 IEEE/CVF Con-

ference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE: 6646-6657 [DOI: 10.1109/CVPR52733.2024.00635]

Liu Y, Lin C, Zeng Z J, Long X X, Liu L J, Komura T and Wang W P. 2024f. SyncDreamer: generating multiview-consistent images from a single-view image [EB/OL]. [2025-01-03]. https://arxiv.org/pdf/2309.03453.pdf

Liu Y F, Cun X D, Liu X B, Wang X T, Zhang Y, Chen H X, Liu Y, Zeng T Y, Chan R and Shan Y. 2024g. EvalCrafter: benchmarking and evaluating large video generation models//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE: 22139-22149 [DOI: 10.1109/CVPR52733.2024.02090]

Liu Y X, Li L, Ren S H, Gao R D, Li S C, Chen S S, Sun X and Hou L. 2023d. FETV: a benchmark for fine-grained evaluation of open-domain text-to-video generation//Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc.: 62352-62387 [DOI: 10.5555/3666122.3668845]

Liu Z X, Li Y G, Lin Y T, Yu X, Peng S D, Cao Y P, Qi X J, Huang X S, Liang D and Ouyang W L. 2025. UniDream: unifying diffusion priors for relightable text-to-3D generation//Proceedings of the 18th European Conference on Computer Vision. Milan, Italy: Springer-Verlag: 74-91 [DOI: 10.1007/978-3-031-72652-1_5]

Liu Z Y, Liang W C, Zhao Y M, Chen B H, Liang L, Wang L J, Li J and Yuan Y H. 2024h. Glyph-ByT5-v2: a strong aesthetic baseline for accurate multilingual visual text rendering [EB/OL]. [2025-01-03]. https://arxiv.org/pdf/2406.10208.pdf

Long X X, Guo Y C, Lin C, Liu Y, Dou Z Y, Liu L J, Ma Y X, Zhang S H, Habermann M, Theobalt C and Wang W P. 2024. Wonder3D: single image to 3D using cross-domain diffusion//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE: 9970-9980 [DOI: 10.1109/CVPR52733.2024.00951]

Lukoianov A, de Ocáriz Borde H S, Greenewald K, Guizilini V C, Bagautdinov T, Sitzmann V and Solomon J. 2024. Score distillation via reparametrized DDIM [EB/OL]. [2025-01-03]. https://arxiv.org/pdf/2405.15891.pdf

Luo A, Li T Q, Zhang W H and Lee T S. 2021. SurfGen: adversarial 3D shape synthesis with explicit surface discriminators//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada: IEEE: 16218-16228 [DOI: 10.1109/ICCV48922.2021.01593]

Luo S M, Tan Y Q, Huang L B, Li J and Zhao H. 2023. Latent consistency models: synthesizing high-resolution images with few-step inference [EB/OL]. [2025-01-03]. https://arxiv.org/pdf/2310.04378.pdf

Lyu Z Y, Wang J Y, An Y W, Zhang Y, Lin D H and Dai B. 2023. Controllable mesh generation through sparse latent point diffusion models//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE: 271-280 [DOI: 10.1109/CVPR52729.2023.00034]

Ma B Q, Zong Z F, Song G L, Li H S and Liu Y. 2024a. Exploring the role of large language models in prompt encoding for diffusion models [EB/OL]. [2025-01-03]. https://arxiv.org/pdf/2406.11831.pdf

Ma J, Liang J H, Chen C and Lu H N. 2024b. Subject-diffusion: open domain personalized text-to-image generation without test-time fine-tuning//Proceedings of 2024 ACM SIGGRAPH Conference Papers. Denver, USA: Association for Computing Machinery: #25 [DOI: 10.1145/3641519.3657469]

Ma X, Wang Y H, Jia G Y, Chen X Y, Liu Z W, Li Y F, Chen C J and Qiao Y. 2024c. Latte: latent diffusion transformer for video generation [EB/OL]. [2025-01-03]. https://arxiv.org/pdf/2401.03048.pdf

Ma Y, He Y Q, Cun X D, Wang X T, Chen S R, Li X and Chen Q F. 2024d. Follow your pose: pose-guided text-to-video generation using pose-free videos//Proceedings of the 38th AAAI Conference on Artificial Intelligence and the 36th Conference on Innovative Applications of Artificial Intelligence and the 14th Educational Advances in Artificial Intelligence. Vancouver, Canada: AAAI Press: 4117-4125 [DOI: 10.1609/aaai.v38i5.28206]

Meng C L, He Y T, Song Y, Song J M, Wu J J, Zhu J Y and Ermon S. 2022. SDEdit: guided image synthesis and editing with stochastic differential equations [EB/OL]. [2025-01-03]. https://arxiv.org/pdf/2108.01073.pdf

Meng F Q, Liao J Q, Tan X Y, Shao W Q, Lu Q F, Zhang K P, Cheng Y, Li D Q, Qiao Y and Luo P. 2024. Towards world simulator: crafting physical commonsense-based benchmark for video generation [EB/OL]. [2025-01-03]. https://arxiv.org/pdf/2410.05363.pdf

Metzer G, Richardson E, Patashnik O, Giryes R and Cohen-Or D. 2023. Latent-NeRF for shape-guided generation of 3D shapes and textures//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE: 12663-12673 [DOI: 10.1109/CVPR52729.2023.01218]

Mildenhall B, Srinivasan P P, Tancik M, Barron J T, Ramamoorthi R and Ng R. 2021. NeRF: representing scenes as neural radiance fields for view synthesis. Communications of the ACM, 65(1): 99-106 [DOI: 10.1145/3503250]

Mittal P, Cheng Y C, Singh M and Tulsiani S. 2022. AutoSDF: shape priors for 3D completion, reconstruction and generation//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE: 306-315 [DOI: 10.1109/CVPR52688.2022.00040]

Miyake D, Iohara A, Saito Y and Tanaka T. 2024. Negative-prompt inversion: fast image inversion for editing with text-guided diffusion models [EB/OL]. [2025-01-03]. https://arxiv.org/pdf/2305.16807.pdf

Mo S T, Xie E Z, Chu R H, Yao L W, Hong L Q, Nießner M and Li Z G. 2023. DiT-3D: exploring plain diffusion transformers for 3D

shape generation//Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc.: 67960-67971

Mokady R, Hertz A, Aberman K, Pritch Y and Cohen-Or D. 2023. Null-text inversion for editing real images using guided diffusion models//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE: 6038-6047 [DOI: 10.1109/CVPR52729.2023.00585]

Moreau A, Song J F, Dhamo H, Shaw R, Zhou Y R and Pérez-Pellitero E. 2024. Human Gaussian splatting: real-time rendering of animatable avatars//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE: 788-798 [DOI: 10.1109/CVPR52733.2024.00081]

Müller N, Siddiqui Y, Porzi L, Bulò S R, Kontschieder P and Nießner M. 2023. DiffRF: rendering-guided 3D radiance field diffusion//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE: 4328-4338 [DOI: 10.1109/CVPR52729.2023.00421]

Müller T, Evans A, Schied C and Keller A. 2022. Instant neural graphics primitives with a multiresolution hash encoding. ACM Transactions on Graphics, 41(4): #102 [DOI: 10.1145/3528223.3530127]

Nam G, Khlifi M, Rodriguez A, Tono A, Zhou L Q and Guerrero P. 2022. 3D-LDM: neural implicit 3D shape generation with latent diffusion models [EB/OL]. [2025-01-03]. https://arxiv.org/pdf/2212.00842.pdf

Nash C, Ganin Y, Eslami S M A and Battaglia P W. 2020. PolyGen: an autoregressive generative model of 3D meshes//Proceedings of the 37th International Conference on Machine Learning. Onling: JMLR.org: 7220-7229

Nichol A, Jun H, Dhariwal P, Mishkin P and Chen M. 2022. Point.E: a system for generating 3D point clouds from complex prompts [EB/OL]. [2025-01-03]. https://arxiv.org/pdf/2212.08751.pdf

Or-EI R, Luo X, Shan M Y, Shechtman E, Park J J and Kemelmacher-Shlizerman I. 2022. StyleSDF: high-resolution 3D-consistent image and geometry generation//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE: 13493-13503 [DOI: 10.1109/CVPR52688.2022.01314]

Ouyang H, Wang Q Y, Xiao Y X, Bai Q Y, Zhang J T, Zheng K C, Zhou X W, Chen Q F and Chen Q F. 2024. CoDeF: content deformation fields for temporally consistent video processing//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE: 8089-8099 [DOI: 10.1109/CVPR52733.2024.00773]

Pan X G, Tewari A, Leimkühler T, Liu L J, Meka A and Theobalt C. 2023. Drag your GAN: interactive point-based manipulation on the generative image manifold//Proceedings of 2023 ACM SIGGRAPH Conference Proceedings. Los Angeles, USA: Association for Computing Machinery: #78 [DOI: 10.1145/3588432.3591500]

Park J J, Florence P, Straub J, Newcombe R and Lovegrove S. 2019. DeepSDF: learning continuous signed distance functions for shape representation//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE: 165-174 [DOI: 10.1109/CVPR.2019.00025]

Peebles W and Xie S N. 2023. Scalable diffusion models with transformers//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE: 4172-4182 [DOI: 10.1109/ICCV51070.2023.00387]

Peng H Y, Zhang J P, Guo M H, Cao Y P and Hu S M. 2024. CharacterGen: efficient 3D character generation from single images with multi-view pose canonicalization. ACM Transactions on Graphics, 43(4): #84 [DOI: 10.1145/3658217]

Penner E and Zhang L. 2017. Soft 3D reconstruction for view synthesis. ACM Transactions on Graphics, 36(6): #235 [DOI: 10.1145/3130800.3130855]

Polyak A, Zohar A, Brown A, Tjandra A, Sinha A, Lee A, Vyas A, Shi B W, Ma C Y, Chuang C Y, Yan D, Choudhary D, Wang D K, Sethi G, Pang G, Ma H Y, Misra I, Hou J, Wang J L, Jagadeesh K, Li K P, Zhang L X, Singh M, Williamson M, Le M, Yu M, Singh M K, Zhang P Z, Vajda P, Duval Q, Girdhar R, Sumbaly R, Rambhatla S, Tsai S, Azadi S, Datta S, Chen S Y, Bell S, Ramaswamy S, Sheynin S, Bhattacharya S, Motwani S, Xu T, Li T H, Hou T B, Hsu W N, Yin X, Dai X L, Taigman Y, Luo Y Q, Liu Y C, Wu Y C, Zhao Y, Kirstain Y, He Z C, He Z J, Pumarola A, Thabet A, Sanakoyeu A, Mallya A, Guo B S, Araya B, Kerr B, Wood C, Liu C, Peng C, Vengertsev D, Schonfeld E, Blanchard E, Juefei-Xu F, Nord F, Liang J, Hoffman J, Kohler J, Fire K, Sivakumar K, Chen L, Yu L C, Gao L Y, Georgopoulos M, Moritz R, Sampson S K, Li S K, Parmeggiani S, Fine S, Fowler T, Petrovic V and Du Y M. 2025. Movie gen: a cast of media foundation models [EB/OL]. [2025-01-03]. https://arxiv.org/pdf/2410.13720.pdf

Poole B, Jain A, Barron J T and Mildenhall B. 2022. DreamFusion: text-to-3D using 2D diffusion [EB/OL]. [2025-01-03]. https://arxiv.org/pdf/2209.14988.pdf

Qi C Y, Cun X D, Zhang Y, Lei C Y, Wang X T, Shan Y and Chen Q F. 2023. FateZero: fusing attentions for zero-shot text-based video editing//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE: 15886-15896 [DOI: 10.1109/ICCV51070.2023.01460]

Qian G C, Mai J J, Hamdi A, Ren J A, Siarohin A, Li B, Lee H Y, Skorokhodov I, Wonka P, Tulyakov S and Ghanem B. 2023. Magic123: one image to high-quality 3D object generation using both 2D and 3D diffusion priors [EB/OL]. [2025-01-03]. https://arxiv.org/pdf/2306.17843.pdf

Radford A, Kim J W, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, Krueger G and Sutskever I. 2021. Learning transferable visual models from natural language

supervision//Proceedings of the 38th International Conference on Machine Learning. [s.l.]: PMLR: 8748-8763

Ramesh A, Pavlov M, Goh G, Gray S, Voss C, Radford A, Chen M and Sutskever I. 2021. Zero-shot text-to-image generation [EB/OL]. [2025-01-03]. https://arxiv.org/pdf/2102.12092.pdf

Ran Q, Xu C and Dong Z L. 2024. Normal estimation for human-body images [EB/OL]. [2025-01-03]. https://aigc3d.github.io/HumanNormal/

Ren X C, Huang J H, Zeng X H, Museth K, Fidler S and Williams F. 2024. XCube: large-scale 3D generative modeling using sparse voxel hierarchies//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE: 4209-4219 [DOI: 10.1109/CVPR52733.2024.00403]

Ruiz N, Li Y Z, Jampani V, Pritch Y, Rubinstein M and Aberman K. 2023. DreamBooth: fine tuning text-to-image diffusion models for subject-driven generation//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE: 22500-22510 [DOI: 10.1109/CVPR52729.2023.02155]

Sanyal S, Ghosh P, Yang J L, Black M J, Thies J and Bolkart T. 2024. SCULPT: shape-conditioned unpaired learning of pose-dependent clothed and textured human meshes//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE: 2362-2371 [DOI: 10.1109/CVPR52733.2024.00229]

Schult J, Tsai S, Höllein L, Wu B C, Wang J L, Ma C Y, Li K P, Wang X F, Wimbauer F, He Z J, Zhang P Z, Leibe B, Vajda P and Hou J. 2024. ControlRoom3D: room generation using semantic proxy rooms//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE: 6201-6210 [DOI: 10.1109/CVPR52733.2024.00593]

Sella E, Fiebelman G, Atia N and Averbuch-Elor H. 2024. Spice·E: structural priors in 3D diffusion using cross-entity attention//Proceedings of 2024 ACM SIGGRAPH Conference Papers. Denver, USA: Association for Computing Machineryz: #15 [DOI: 10.1145/3641519.3657461]

Shade J, Gortler S, He L W and Szeliski R. 1998. Layered depth images//Proceedings of the 25th Annual Conference on Computer Graphics and Interactive Techniques. [s.l.]: Association for Computing Machinery: 231-242 [DOI: 10.1145/280814.280882]

Sheynin S, Polyak A, Singer U, Kirstain Y, Zohar A, Ashual O, Parikh D and Taigman Y. 2024. Emu edit: precise image editing via recognition and generation tasks//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE: 8871-8879 [DOI: 10.1109/CVPR52733.2024.00847]

Shi R X, Chen H S, Zhang Z Y, Liu M H, Xu C, Wei X Y, Chen L H, Zeng C and Su H. 2023a. Zero123++: a single image to consistent multi-view diffusion base model [EB/OL]. [2025-01-03]. https://arxiv.org/pdf/2310.15110.pdf

Shi Y C, Wang P, Ye J L, Long M, Li K J and Yang X. 2024. MVDream: multi-view diffusion for 3D generation [EB/OL]. [2025-01-03]. https://arxiv.org/pdf/2308.16512.pdf

Shi Y K, Wang J N, Cao H, Tang B S, Qi X B, Yang T Y, Huang Y K, Liu S L, Zhang L and Shum H Y. 2023b. TOSS: high-quality text-guided novel view synthesis from a single image [EB/OL]. [2025-01-03]. https://arxiv.org/pdf/2310.10644.pdf

Shih M L, Su S Y, Kopf J and Huang J B. 2020. 3D photography using context-aware layered depth inpainting//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE: 8025-8035 [DOI: 10.1109/CVPR42600.2020.00805]

Shu D W, Park S W and Kwon J. 2019. 3D point cloud generative adversarial network based on tree structured graph convolutions//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea (South): IEEE: 3858-3867 [DOI: 10.1109/ICCV.2019.00396]

Shue J R, Chan E R, Po R, Ankner Z, Wu J J and Wetzstein G. 2022. 3D neural field generation using triplane diffusion//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE: 20875-20886 [DOI: 10.1109/CVPR52729.2023.02000]

Siddiqui Y, Alliegro A, Artemov A, Tommasi T, Sirigatti D, Rosov V, Dai A and Nießner M. 2024a. MeshGPT: generating triangle meshes with decoder-only transformers//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE: 19615-19625 [DOI: 10.1109/CVPR52733.2024.01855]

Siddiqui Y, Monnier T, Kokkinos F, Kariya M, Kleiman Y, Garreau E, Gafni O, Neverova N, Vedaldi A, Shapovalov R and Novotny D. 2024b. Meta 3D AssetGen: text-to-mesh generation with high-quality geometry, texture, and PBR materials [EB/OL]. [2025-01-03]. https://arxiv.org/pdf/2407.02445.pdf

Sun Y B, Wang Y, Liu Z W, Siegel J E and Sarma S E. 2020. PointGrow: autoregressively learned point cloud generation with self-attention//Proceedings of 2020 IEEE Winter Conference on Applications of Computer Vision (WACV). Snowmass, USA: IEEE: 61-70 [DOI: 10.1109/WACV45572.2020.9093430]

Tang J P, Nie Y Y, Markhasin L, Dai A, Thies J and Nießner M. 2024a. DiffuScene: denoising diffusion models for generative indoor scene synthesis//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE: 20507-20518 [DOI: 10.1109/CVPR52733.2024.01938]

Tang J X, Chen Z X, Chen X K, Wang T F, Zeng G and Liu Z W. 2025. LGM: large multi-view Gaussian model for high-resolution 3D content creation//Proceedings of the 18th European Conference on Computer Vision. Milan, Italy: Springer: 1-18 [DOI: 10.1007/

978-3-031-73235-5_1]

Tang J X, Ren J W, Zhou H, Liu Z W and Zeng G. 2024b. DreamGaussian: generative Gaussian splatting for efficient 3D content creation [EB/OL]. [2025-01-03]. https://arxiv.org/pdf/2309.16653.pdf

Tang S T, Zhang F Y, Chen J C, Wang P and Furukawa Y. 2023. MVDiffusion: enabling holistic multi-view image generation with correspondence-aware diffusion//Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc.: 51202-51233

Tang Z C, Gu S Y, Wang C Y, Zhang T, Bao J M, Chen D and Guo B N. 2024c. VolumeDiffusion: flexible text-to-3D generation with efficient volumetric encoder [EB/OL]. [2025-01-03]. https://arxiv.org/pdf/2312.11459.pdf

Thai A, Stojanov S, Upadhya V and Rehg J M. 2021. 3D reconstruction of novel object shapes from single images//Proceedings of 2021 International Conference on 3D Vision (3DV). London, United Kingdom: IEEE: 85-95 [DOI: 10.1109/3DV53792.2021.00019]

Tian L R, Wang Q, Zhang B and Bo L F. 2025. EMO: emote portrait alive generating expressive portrait videos with Audio2Video diffusion model under weak conditions//Proceedings of the 18th European Conference on Computer Vision. Milan, Italy: Springer: 244-260 [DOI: 10.1007/978-3-031-73010-8_15]

Tian Y, Yang L, Yang H T, Gao Y, Deng Y F, Chen J M, Wang X T, Yu Z C, Tao X, Wan P F, Zhang D and Cui B. 2024. VideoTetris: towards compositional text-to-video generation [EB/OL]. [2025-01-03]. https://arxiv.org/pdf/2406.04277.pdf

Tuo Y X, Xiang W M, He J Y, Geng Y F and Xie X S. 2024. AnyText: multilingual visual text generation and editing [EB/OL]. [2025-01-03]. https://arxiv.org/pdf/2311.03054.pdf

Wang H C, Du X D, Li J H, Yeh R A and Shakhnarovich G. 2023a. Score jacobian chaining: lifting pretrained 2D diffusion models for 3D generation//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE: 12619-12629 [DOI: 10.1109/CVPR52729.2023.01214]

Wang J S, Ma Y, Guo J Y, Xiao Y C, Huang G and Li X. 2024a. COVE: unleashing the diffusion feature correspondence for consistent video editing [EB/OL]. [2025-01-03]. https://arxiv.org/pdf/2406.08850.pdf

Wang P, Liu L J, Liu Y, Theobalt C, Komura T and Wang W P. 2021a. NeuS: learning neural implicit surfaces by volume rendering for multi-view reconstruction//Proceedings of the 35th International Conference on Neural Information Processing Systems. [s.l.]: Curran Associates Inc.: 27171-27183

Wang T F, Zhang B, Zhang T, Gu S Y, Bao J M, Baltrusaitis T, Shen J J, Chen D, Wen F, Chen Q F and Guo B N. 2023b. RODIN: a generative model for sculpting 3D digital avatars using diffusion//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE: 4563-4573 [DOI: 10.1109/CVPR52729.2023.00443]

Wang X T, Xie L B, Dong C and Shan Y. 2021b. Real-ESRGAN: training real-world blind super-resolution with pure synthetic data//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). Montreal, Canada: IEEE: 1905-1914 [DOI: 10.1109/ICCVW54120.2021.00217]

Wang Z X, Yuan Z Y, Wang X T, Li Y W, Chen T S, Xia M H, Luo P and Shan Y. 2024b. MotionCtrl: a unified and flexible motion controller for video generation//Proceedings of 2024 ACM SIGGRAPH Conference Papers. Denver, USA: Association for Computing Machinery: #114 [DOI: 10.1145/3641519.3657518]

Wang Z Y, Lu C, Wang Y K, Bao F, Li C X, Su H and Zhu J. 2023c. ProlificDreamer: high-fidelity and diverse text-to-3D generation with variational score distillation//Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc.: 8406-8441

Weng Z Z, Wang Z Y and Yeung S. 2023. ZeroAvatar: zero-shot 3D avatar generation from a single image [EB/OL]. [2025-01-03]. https://arxiv.org/pdf/2305.16411.pdf

Wu J Z, Fang G A, Wu H N, Wang X T, Ge Y X, Cun X D, Zhang D J, Liu J W, Gu Y C, Zhao R, Lin W S, Hsu W, Shan Y and Shou M Z. 2024a. T2VScore: towards a better metric for text-to-video generation [EB/OL]. [2025-01-03]. https://arxiv.org/pdf/2401.07781.pdf

Wu J Z, Ge Y X, Wang X T, Lei S W, Gu Y C, Shi Y F, Hsu W, Shan Y, Qie X H and Shou M Z. 2023. Tune-A-Video: one-shot tuning of image diffusion models for text-to-video generation//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE: 7589-7599 [DOI: 10.1109/ICCV51070.2023.00701]

Wu S, Lin Y T, Zhang F H, Zeng Y F, Xu J X, Torr P, Cao X and Yao Y. 2024b. Direct3D: scalable image-to-3D generation via 3D latent diffusion transformer [EB/OL]. [2025-01-03]. https://arxiv.org/pdf/2405.14832.pdf

Wu W M, Fu X M, Tang R, Wang Y H, Qi Y H and Liu L G. 2019. Data-driven interior plan generation for residential buildings. ACM Transactions on Graphics, 38(6): #234 [DOI: 10.1145/3355089.3356556]

Wu Y Q, Chen H X and Zhang Y. 2024. Review of 3D point cloud processing methods based on deep learning. Chinese Journal of Lasers, 51(5): #0509001 (吴一全, 陈慧娴, 张耀. 2024. 基于深度学习的三维点云处理方法研究进展. 中国激光, 51(5): #0509001) [DOI: 10.3788/CJL230924]

Xiao Z S, Kreis K and Vahdat A. 2022. Tackling the generative learning trilemma with denoising diffusion GANs [EB/OL]. [2025-01-03]. https://arxiv.org/pdf/2112.07804.pdf

Xie H Z, Chen Z X, Hong F Z and Liu Z W. 2024. CityDreamer: compositional generative model of unbounded 3D cities//Proceedings of

2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE: 9666-9675 [DOI: 10.1109/CVPR52733.2024.00923]

Xing J B, Xia M H, Liu Y X, Zhang Y C, Zhang Y, He Y Q, Liu H Y, Chen H X, Cun X D, Wang X T, Shan Y and Wong T T. 2025a. Make-your-video: customized video generation using textual and structural guidance. IEEE Transactions on Visualization and Computer Graphics, 31(2): 1526-1541 [DOI: 10.1109/TVCG.2024.3365804]

Xing J B, Xia M H, Zhang Y, Chen H X, Yu W B, Liu H Y, Liu G Y, Wang X T, Shan Y and Wong T T. 2025b. DynamiCrafter: animating open-domain images with video diffusion priors//Proceedings of the 18th European Conference on Computer Vision. Milan, Italy: Springer: 399-417 [DOI: 10.1007/978-3-031-72952-2_23]

Xu C, Dong Y, Zuo Q, Zhang J F, Ye X D, Geng W B, Zhang Y X, Gu X D, Qiu L T, Zhao Z Y, Ran Q, Jiang J Y, Dong Z L and Bo L F. 2024a. G-buffer objaverse: high-quality rendering dataset of objaverse [EB/OL]. [2025-01-03]. https://aigc3d.githwb.io/gobjaverse/

Xu J L, Cheng W H, Gao Y M, Wang X T, Gao S H and Shan Y. 2024b. InstantMesh: efficient 3D mesh generation from a single image with sparse-view large reconstruction models [EB/OL]. [2025-01-03]. https://arxiv.org/pdf/2404.07191.pdf

Yan Z Z, Zhou J P, Meng F P, Wu Y S, Qiu L T, Ye Z S, Cui S G, Chen G Y and Han X G. 2025. DreamDissector: learning disentangled text-to-3D generation from 2D diffusion priors//Proceedings of the 18th European Conference on Computer Vision. Milan, Italy: Springer-Verlag: 124-141 [DOI: 10.1007/978-3-031-73254-6_8]

Yang G D, Huang X, Hao Z K, Liu M Y, Belongie S and Hariharan B. 2019. PointFlow: 3D point cloud generation with continuous normalizing flows//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 4540-4549 [DOI: 10.1109/ICCV.2019.00464]

Yang H T, Dong Y, Jiang H W, Xu D J, Pavlakos G and Huang Q X. 2024a. Atlas Gaussians diffusion for 3D generation [EB/OL]. [2025-01-03]. https://arxiv.org/pdf/2408.13055.pdf

Yang M J, Guo Y X, Zhou B and Tong X. 2021. Indoor scene generation from a collection of semantic-segmented depth images//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada: IEEE: 15183-15192 [DOI: 10.1109/ICCV48922.2021.01492]

Yang Z Y, Teng J Y, Zheng W D, Ding M, Huang S Y, Xu J Z, Yang Y M, Hong W Y, Zhang X H, Feng G Y, Yin D, Gu X T, Zhang Y X, Wang W H, Cheng Y, Liu T, Xu B, Dong Y X and Tang J. 2024b. CogVideoX: text-to-video diffusion models with an expert transformer [EB/OL]. [2025-01-03]. https://arxiv.org/pdf/2408.06072.pdf

Yariv L, Gu J T, Kasten Y and Lipman Y. 2021. Volume rendering of neural implicit surfaces//Proceedings of the 35th International Conference on Neural Information Processing Systems. [s.l.]: Curran Associates Inc.: 4805-4815

Ye H, Zhang J, Liu S B, Han X and Yang W. 2023. IP-Adapter: text compatible image prompt adapter for text-to-image diffusion models [EB/OL]. [2025-01-03]. https://arxiv.org/pdf/2308.06721.pdf

Yeh Y T, Yang L F, Watson M, Goodman N D and Hanrahan P. 2012. Synthesizing open worlds with constraints using locally annealed reversible jump MCMC. ACM Transactions on Graphics, 31(4): #56 [DOI: 10.1145/2185520.2185552]

Yin S M, Wu C F, Liang J, Shi J, Li H Q, Ming G and Duan N. 2023. DragNUWA: fine-grained control in video generation by integrating text, image, and trajectory [EB/OL]. [2025-01-03]. https://arxiv.org/pdf/2308.08089.pdf

Yu X, Guo Y C, Li Y G, Liang D, Zhang S H and Qi X J. 2023. Text-to-3D with classifier score distillation [EB/OL]. [2025-01-03]. https://arxiv.org/pdf/2310.19415.pdf

Yu Z M, Dou Z Y, Long X X, Lin C, Li Z K, Liu Y, Müller N, Komura T, Habermann M, Theobalt C, Li X and Wang W P. 2025. Surf-D: generating high-quality surfaces of arbitrary topologies using diffusion models//Proceedings of the 18th European Conference on Computer Vision. Milan, Italy: Springer: 419-438 [DOI: 10.1007/978-3-031-72933-1_24]

Zamorski M, Zięba M, Klukowski P, Nowak R, Kurach K, Stokowiec W and Trzciński T. 2020. Adversarial autoencoders for compact representations of 3D point clouds. Computer Vision and Image Understanding, 193: #102921 [DOI: 10.1016/j.cviu.2020.102921]

Zeng A L, Yang Y H, Chen W D and Liu W. 2024. The dawn of video generation: preliminary explorations with SORA-like models [EB/OL]. [2025-01-03]. https://arxiv.org/pdf/2410.05227.pdf

Zeng Y F, Lu Y X, Ji X Y, Yao Y, Zhu H and Cao X. 2023. AvatarBooth: high-quality and customizable 3D human avatar generation [EB/OL]. [2025-01-03]. https://arxiv.org/pdf/2306.09864.pdf

Zhang B W, Cheng Y J, Yang J L, Wang C Y, Zhao F, Tang Y S, Chen D and Guo B N. 2024a. GaussianCube: a structured and explicit radiance representation for 3D generative modeling [EB/OL]. [2025-01-03]. https://arxiv.org/pdf/2403.19655.pdf

Zhang G W, Zhang T Y, Niu G L, Tan Z C, Bai Y L and Yang Q. 2024b. CAMEL: causal motion enhancement tailored for lifting text-driven video editing//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE: 9079-9088 [DOI: 10.1109/CVPR52733.2024.00867]

Zhang H C, Chen B W, Yang H, Qu L, Wang X, Chen L, Long C, Zhu F D, Du D and Zheng M. 2024c. AvatarVerse: high-quality and stable 3D avatar creation from text and pose//Proceedings of the 38th AAAI Conference on Artificial Intelligence. Vancouver, Canada: AAAI: 7124-7132 [DOI: 10.1609/aaai.v38i7.28540]

Zhang L M and Agrawala M. 2024. Transparent image layer diffusion using latent transparency. ACM Transactions on Graphics, 43(4): #100 [DOI: 10.1145/3658150]

Zhang L M, Rao A Y and Agrawala M. 2023a. Adding conditional control to text-to-image diffusion models//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE: 3813-3824 [DOI: 10.1109/ICCV51070.2023.00355]

Zhang L W, Qiu Q W, Lin H Y, Zhang Q X, Shi C, Yang W, Shi Y, Yang S B, Xu L and Yu J Y. 2023b. DreamFace: progressive generation of animatable 3D faces under text guidance. ACM Transactions on Graphics, 42(4): #138 [DOI: 10.1145/3592094]

Zhang L W, Wang Z Y, Zhang Q X, Qiu Q W, Pang A Q, Jiang H R, Yang W, Xu L and Yu J Y. 2024d. CLAY: a controllable large-scale generative model for creating high-quality 3D assets. ACM Transactions on Graphics, 43(4): #120 [DOI: 10.1145/3658146]

Zhang L W, Zhang Q X, Wu M Y, Yu J Y and Xu L. 2021. Neural video portrait relighting in real-time via consistency modeling//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada: IEEE: 782-792 [DOI: 10.1109/ICCV48922.2021.00084]

Zhang Q H, Wang C Y, Siarohin A, Zhuang P Y, Xu Y H, Yang C Y, Lin D H, Zhou B L, Tulyakov S and Lee H Y. 2024e. SceneWiz3D: Towards text-guided 3D scene composition//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE: 6829-6838 [DOI: 10.1109/CVPR52733.2024.00652]

Zhao Z B, Liu W, Chen X, Zeng X F, Wang R, Cheng P, Fu B, Chen T, Yu G and Gao S H. 2023. Michelangelo: conditional 3D shape generation based on shape-image-text aligned latent representation//Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc.: 73969-73982

Zheng X Y, Liu Y, Wang P S and Tong X. 2022. SDF-StyleGAN: implicit SDF-based StyleGAN for 3D shape generation. Computer Graphics Forum, 41(5): 52-63 [DOI: 10.1111/cgf.14602]

Zheng X Y, Pan H, Wang P S, Tong X, Liu Y and Shum H Y. 2023. Locally attentional SDF diffusion for controllable 3D shape generation. ACM Transactions on Graphics, 42(4): #91 [DOI: 10.1145/3592103]

Zhou L Q, Du Y L and Wu J J. 2021. 3D shape generation and completion through point-voxel diffusion//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada: IEEE: 5806-5815 [DOI: 10.1109/ICCV48922.2021.00577]

Zhou T H, Tucker R, Flynn J, Fyffe G and Snavely N. 2018. Stereo magnification: learning view synthesis using multiplane images. ACM Transactions on Graphics, 37(4): #65 [DOI: 10.1145/3197517.3201323]

Zhou Z L, Ma F, Fan H H, Yang Z X and Yang Y. 2025. HeadStudio: text to animatable head avatars with 3D Gaussian splatting//Proceedings of the 18th European Conference on Computer Vision — ECCV 2024. Milan, Italy: Springer-Verlag: 145-163 [DOI: 10.1007/978-3-031-73411-3_9]

Zou Z X, Yu Z P, Guo Y C, Li Y G, Liang D, Cao Y P and Zhang S H. 2024. Triplane meets Gaussian splatting: fast and generalizable single-view 3D reconstruction with transformers//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE: 10324-10335 [DOI: 10.1109/CVPR52733.2024.00983]

作者简介

郑屹，男，副教授，主要研究方向为计算机图形学、生成式人工智能、智能创作与人机交互。
E-mail：zhengyi7592@cuc.edu.cn
黄心渊，通信作者，男，博士生导师，主要研究方向为计算机图形学、虚拟现实、计算机动画和人工智能。
E-mail：hxy@cuc.edu.cn
黄向，男，硕士研究生，主要研究方向为计算机图形学。
E-mail：huangxiang2228@163.com
秦菲儿，女，硕士研究生，主要研究方向为计算机图形学。
E-mail：feiqin2323@gmail.com
梁缘，男，博士，主要研究方向为通用三维人工智能生成内容(3D AIGC)场景下的计算机图形学应用，包括动作建模与生成、材质建模与增材制造。E-mail：liangyuan@vastai3d.com
黄子瑄，女，硕士，主要研究方向为AIGC 3D产品设计与产业应用。E-mail：huangzixuan@vastai3d.com
曹炎培，男，博士，主要研究方向为计算机图形学与三维视觉、3D生成式AI的技术研发与应用。
E-mail：caoyanpei@gmail.com
陶鑫，男，博士，主要研究方向为计算摄影学、计算机视觉、图形学、图像增强与还原、图像视频生成算法、数字人和多模态大模型。E-mail：taoxin@kuaishou.com
郑明悟，男，算法工程师，主要研究方向为基于计算机视觉和计算机图形学的视觉内容创作技术，包括视频生成和3D重建渲染。E-mail：zhengmingwu@kuaishou.com
杨皓天，男，算法工程师，主要研究方向为基于计算机视觉和计算机图形学的视觉内容创作技术，包括视频生成和3D重建渲染。E-mail：yanghaotian@kuaishou.com
万鹏飞，男，博士，主要研究方向为计算机视觉、计算机图形学、机器学习、人机交互。E-mail：wanpengfei@kuaishou.com
陈京炜，女，教授，主要研究方向为游戏心理学、游戏设计及电子竞技理论与实践。E-mail：chenjw@cuc.edu.cn