



**虚拟现实技术研究论文**

|  |  |
| --- | --- |
| 学 院： | 计算机学院 |
| 题 目： | AIGC技术赋能虚拟现实：语言生成模型在VR内容创作中的应用与挑战 |
| 学生姓名： | 陈松宇 冯敏恒 高艺菲 |
| 学 号： | 1120222222 1120220644 1120223193 |

2025 年 4 月 28 日

**目 录**

[1引言 3](#_Toc10693)

[2 AIGC在虚拟现实三维模型生成中的应用 5](#_Toc15168)

[2.1传统三维模型生成方法的局限性 5](#_Toc30246)

[2.2 AIGC在虚拟现实三维模型生成中的突破 5](#_Toc22367)

[2.3 AIGC在虚拟现实三维模型生成中的相关工具 7](#_Toc7719)

[2.4 AIGC 在虚拟现实三维模型生成中的现存问题 8](#_Toc27115)

[2.5 AIGC 在虚拟现实三维模型生成中的热点方向 9](#_Toc18424)

[3 AIGC在虚拟现实三维场景生成中的应用 9](#_Toc17586)

[3.1 AIGC 在虚拟现实三维场景生成中的关键技术应用 10](#_Toc14606)

[3.2 AIGC 在虚拟现实三维场景生成中存在的问题 13](#_Toc20173)

[3.3 未来研究热点 14](#_Toc24009)

[4 AIGC驱动的虚拟角色与对话系统 15](#_Toc1563)

[4.1 传统虚拟角色的局限性 15](#_Toc31621)

[4.2 基于ChatGPT的虚拟角色生成 15](#_Toc17818)

[4.3 基于ChatGPT的逼真对话实现 17](#_Toc18446)

[4.4 AIGC驱动的虚拟角色应用场景 18](#_Toc16264)

[4.5 面临的问题和未来发展方向 18](#_Toc14793)

[5 结论 19](#_Toc20841)

[6 附录 20](#_Toc15890)

[6.1 PPT汇报材料 20](#_Toc18793)

[6.2 参考文献 20](#_Toc12191)

# 摘要

随着以ChatGPT为代表的AIGC技术不断演进，VR内容生产正经历由手工建模向自动化生成的范式转变。本文围绕三维模型生成、虚拟场景构建与虚拟角色生成与对话三个核心方向，系统梳理了AIGC在VR中的技术体系与实际应用路径。在三维建模方面，分析了GAN、NeRF与扩散模型等生成机制及其代表性方法，指出当前存在生成质量不稳定与细节控制能力弱等问题；在场景构建中，探讨了基于语言指令的布局生成与多模态模型协同机制，并针对语义理解与逻辑一致性不足提出改进方向；在虚拟角色生成与对话方面，研究了多模态生成与情感计算等关键技术在角色外观、行为与语言交互中的集成应用。本文认为，未来应通过语义增强、多模态融合与交互优化等手段，进一步提升AIGC在虚拟现实领域中的实用性与可控性。

**关键词：**AIGC；VR；三维模型生成；三维场景生成；虚拟角色；对话系统

# 1引言

虚拟现实技术作为新一代人机交互的核心载体，其内容创作效率与质量直接制约着用户体验与行业应用。传统VR内容生产高度依赖人工建模与预设脚本，存在周期长、成本高、灵活性差等局限性。

近年来，以ChatGPT为代表的AIGC技术凭借其强大的生成能力与多模态理解优势，为VR内容创作提供了新的解决方案。通过语言生成模型与三维建模、场景布局、角色交互等技术的结合，AIGC能够实现“从文本到虚拟世界”的端到端生成，显著降低创作门槛并激发创新潜力。然而，当前技术仍面临生成可控性、逻辑连贯性、伦理安全等挑战。本文聚焦AIGC在VR内容生成中的关键应用场景，系统梳理技术进展与现存问题，并展望未来发展方向，旨在为AIGC与VR的深度融合提供理论参考与实践路径。

# 2 AIGC在虚拟现实三维模型生成中的应用

## 2.1传统三维模型生成方法的局限性

在虚拟现实领域中，传统三维模型生成主要依赖于人工和现有真实物体的数据。

手工建模利用3D软件（如3ds Max、Blender、Unity等）从零开始构建模型网格，需经历几何建模、细节雕刻、拓扑优化、UV拆分、贴图绘制、骨骼绑定等一系列精细复杂的流程。三维扫描则通过专业设备（如激光扫描仪、深度相机等）对真实对象进行多角度扫描拍照，利用算法重建点云或网格模型，再进行噪声清理与网格简化等处理。

上述方式的局限性在于：手工建模门槛高、周期长且成本昂贵，制约了虚拟场景中海量复杂模型的快速生产；扫描建模虽减少人工建模的工作量，但受制于设备和环境，难以大幅提升创作效率，更难以直接创造新的构型，且两者均对创作者的专业背景有较高要求，使得非专业人士几乎无法参与虚拟现实内容的自主创作。

## 2.2 AIGC在虚拟现实三维模型生成中的突破

AIGC技术基于大规模生成模型，从海量数据中学习数据的分布规律和潜在语义结构，进而能在高维潜在空间中自主生成、变换并优化符合人类指令的新型内容。故在其驱动下，三维模型生成变得更加高效普惠，为虚拟现实内容生产带来了全新的工作流程突破。

对于内容创作者而言，其大幅降低创作门槛，提高产出速度，并赋予内容生成以高度的多样性与创造性。非专业用户不再需要通过繁琐的软件操作来“设计”物体，仅需提供简单的描述（如文字、图片等），即可实现3D模型的“所想即所得”，节省大量寻找或制作素材的时间，多模态输入更进一步提高生成结果与用户想法的匹配度。同时，其远超人工建模的生产速度使得过去需要数周完成的模型现在以分钟甚至秒级产出，能够批量化生成海量对象，以缓解大规模虚拟环境的构建瓶颈。

而上述突破的实现，依赖于AIGC底层算法与技术体系的持续迭代。

### 2.2.1 AIGC技术赋能三维模型生成初期突破

三维模型生成的早期技术主要依赖于离散体素表示，生成效果受限于分辨率与细节表达能力。随着生成对抗网络（GAN）在图像领域的成功应用，研究者开始探索其在三维建模中的潜力。GAN通过生成器与判别器的对抗训练机制，推动生成样本逐步逼近真实数据分布。Wu等人提出3D-GAN[[[1]](#endnote-0)]，首次将对抗生成机制引入三维体素数据建模，实现从低维潜空间直接生成高质量三维对象，并在无监督条件下有效学习三维形状的结构特征。

为进一步突破体素表示带来的分辨率与表面连续性瓶颈，Chen等人提出IM-NET[[[2]](#endnote-1)]，通过引入隐式场建模方式，学习点坐标与形状特征之间的映射，在连续空间中刻画三维边界，从而显著提升生成形状的细节表达与灵活采样能力。其探索推动了三维生成方法由传统离散网格向连续空间隐式建模转变，作为早期将神经网络应用于3D几何建模的重要工作之一，为后续的NeRF、DeepSDF等技术提供了理论启发。

### 2.2.2 基于自然语言驱动的三维模型生成演进

随着自然语言理解与三维建模技术的交融，研究者开始探索从文本直接生成三维模型的新范式。

Chen等人提出Text2Shape[[[3]](#endnote-2)]，首次通过学习自然语言描述与三维体素模型的联合嵌入空间，实现了简单物体类别下的文本到彩色三维形状检索与生成。

此后，Sanghi等人提出CLIP-Forge[[[4]](#endnote-3)]，该方法结合了OpenAI提出的CLIP模型（通过在大规模图文对上进行对比学习，构建统一的视觉-语言嵌入空间，具备强大的跨模态理解与零样本推理能力）。该方法首先将三维形状渲染为图像并提取特征，再通过条件归一化流将文本特征映射到三位潜在空间，最终实现了无需配对训练数据、支持零样本推理且具有多样性输出能力的文本到三维模型生成。

然而，上述方法在ShapeNet之外的类别泛化能力较弱，为进一步提升生成质量，Jain等人提出Dream Fields[[[5]](#endnote-4)]，首次将神经辐射场（NeRF）引入文本到三维生成任务中。NeRF[[[6]](#endnote-5)]是一种基于神经网络的隐式三维场景标识方法，通过在连续空间中映射每个三维坐标与视角方向的颜色与体密度，从而实现高保真、多视角一致的图像渲染。Dream Fields通过优化NeRF，使其渲染图像在预训练视觉语言模型特征空间中与文本描述高度匹配，从而实现了无需三维监督、连续空间细粒度表达和多视角一致性的三维模型生成。

### 2.2.3 基于扩散模型的高质量三维模型生成方法

随着扩散模型在图像生成领域取得突破，研究者开始探索以预训练二维扩散模型指导三维建模的新路径，其通过在数据分布上逐步添加噪声并学习逆过程进行去噪，能够在采样时从随机噪声中逐步恢复出高质量样本。该机制使其在生成质量和稳定性方面渐渐表现出优于 GAN 的性能。

Poole等人提出DreamFusion[[[7]](#endnote-6)]，首次引入分数蒸馏采样（SDS）方法，利用预训练扩散模型内部的去噪得分作为监督信号，优化NeRF以生成符合文本描述的三维对象，实现了无需三维数据、端到端驱动的高质量三维建模。

Chen等人提出Magic3D[[[8]](#endnote-7)]，在DreamFusion框架基础上引入两阶段优化机制，先以低分辨率扩散模型引导 NeRF 优化生成粗略结构，再将其转化为显式网格并通过高分辨率扩散模型细化几何与纹理。相较于 DreamFusion，Magic3D 将引导分辨率提升至 512×512，整体优化时间缩短约2倍（在40min内得到模型），在保证生成质量的同时显著提升效率。

而针对现有方法优化时间长、计算成本高的问题，Nichol等人提出Point·E[[[9]](#endnote-8)]，采用两阶段扩散生成流程，先从文本生成合成图像，再由图像条件生成三维点云，实现在单个GPU上运行1-2分钟即得到模型，尽管细节表现仍受限于点云稀疏性。随后，Jun和Nichol提出Shap·E[[[10]](#endnote-9)]，旨在解决了现有生成方法依赖于显示表示（点云、体素等）即无法兼顾质量与速度的问题。Shap·E通过两阶段建模框架，首先将三维数据编码为隐式神经表示，随后在潜空间中利用扩散模型进行条件采样，直接生成可用于 NeRF 渲染或网格重建的可微结构，相比Point·E在细节连续性与表达能力上表现更优，在数秒内即可生成三维资产。

近期，Li 等人提出 CraftsMan[[[11]](#endnote-10)]生成式三维建模系统，采用基于三维潜空间的原生扩散模型，结合多视角一致性条件，在约5秒内生成结构规则的粗略网格，随后通过法线图引导的细化模块，在10-20秒内完成高质量集合增强与交互式局部编辑，显著提升生成速度、拓扑合理性与细节控制能力，推动三维模型生成迈向高保真、低延迟、可编辑的新阶段。

### 2.2.4 三维模型生成新范式探索动态

近期，北京大学团队提出OctGPT[[[12]](#endnote-11)]，它是一种使用基于序列化八叉树表示的多尺度自回归模型生成三维模型的方法，突破了近年来扩散模型主导的生成路径。该方法以八叉树结构对三维模型进行多尺度编码，以保留空间局部性与层次信息，并此基础上构建基于 Transformer 的自回归生成器，引入膨胀注意力与移位窗口机制进行高效跨尺度建模，设计尺度感知的 Teacher Forcing Mask 与并行采样策略，实现长序列建模加速与高效推理。OctGPT 在多个数据集上（如ShapeNet、Objaverse等）在生成质量、结构一致性等方面表现出对现有扩散与自回归方法的显著竞争力。

## 2.3 AIGC在虚拟现实三维模型生成中的相关工具

近年来，虚拟现实三维模型生成工具在市场上迅速涌现，部分产品已达到较高的成熟度，面向大众直接提供在线或插件式的服务。例如，**Meshy**、Hyper3d等提供基于文本或图像输入的在线三维模型生成服务，用户可直接通过网页操作；再如**Unity等主流建模软件**也逐步推出生成式三维内容辅助创作工具（如Unity Muse），适用于开发者本地部署使用。

为更系统地评估此类工具的应用效果，我们选择了四个代表性的在线工具（Meshy、Hyper3d、Tripo、Sudo AI）进行统一使用测试。在文本输入模态下，统一采用Prompt：“一个真实感很强的陶瓷花瓶，带有裂纹，表面是蓝色的花纹图案，釉面有光泽反射”；在图像输入模态下，输入统一参考图像，观察各工具生成结果的表现差异。结果详见图1。

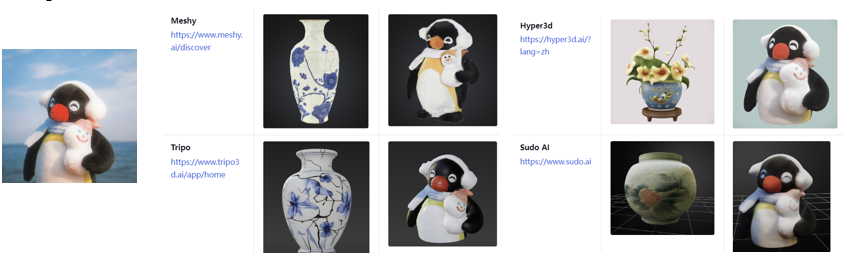


图1：四种在线三维模型生成的测试结果

对比测试结果，我们发现：

在文本输入模态下，生成的三维模型在结构完整性与形体准确度方面表现相近，但呈现出视觉的多样性，Meshy 的生成速度最快且在整体视觉效果上最为出色。

在图像输入模态下，各工具生成的三维形状均与输入图像的轮廓较为一致，但纹理、色彩与细节还原差异显著：Meshy 在该模态下的表现不及文本模态，色彩失真明显，面部细节不准确；Hyper3D 能较好还原正面外观，但背部纹理几乎丢失；SudoAI 输出结果整体粗糙，形体简化更为严重；Tripo 模型全貌接近原图，但放大后手部结构等细节存在不合理性。

此外，我们亦尝试使用了其他三维生成工具，但部分平台在实际使用过程中出现生成过程长时间停滞、无输出结果，或因注册流程繁琐、访问权限受限而在短期内无法完成测试。

综上可见，尽管现有工具在一定程度上能够支持虚拟现实内容创作者实现个性化三维建模需求，但在多模态一致性、纹理还原质量及结构细节保真等方面仍存在较为显著的性能瓶颈。

## 2.4 AIGC 在虚拟现实三维模型生成中的现存问题

### 2.4.1 可控性差

模型生成过程存在随机性，用户在写prompt时对细节控制有限，尤其在复杂对象描述中易产生歧义，导致生成结果与预期不符。此外，由于三维模型通常作为场景搭建的基础单元，若在用户逐一生成的过程中存在风格不统一，将直接削弱整体视觉协调性，增加后期人工协调的成本。

### 2.4.2 生成模型质量偏低

许多AI生成的模型在拓扑结构与纹理精细度上尚不及人工建模，缺乏真实感或细节的合理性，导致生成结果往往需经过大量额外修整后才能应用于VR场景。

### 2.4.3 计算成本高

高质量3D生成通常需较大的计算资源与时间，高分辨率的模型生成可能仍需数十分钟GPU运算，推理开销大，且批量生成时算力与时间成本进一步放大。

## 2.5 AIGC 在虚拟现实三维模型生成中的热点方向

### 2.5.1 增强交互设计

开发更直观的控制机制，允许用户通过局部编辑、语言指令或草图范例引导生成过程，并支持多轮交互式细化，兼顾AI生成速度与人工设计精度，提升VR内容创作的自主性与创造性。

### 2.5.2 提升生成质量

通过持续改进网络架构（如融合显式与隐式混合表示）、引入更精细的判别器与评估指标，以及采用高分辨率训练数据，以进一步提升模型的细节刻画能力与拓扑合理性。同时，结合几何约束与物理先验，尽可能减少常见缺陷的产生，实现接近手工建模水准的高质量输出。

### 2.5.3 推动实时生成与沉浸式创作

加速算法优化与算力提升，使用户能够以语音、文本等自然交互方式即时指挥AI生成虚拟三维模型，并在VR环境中实时调整与体验，打破传统建模流程，促进内容生成与沉浸编辑的深度融合。

# 3 AIGC在虚拟现实三维场景生成中的应用

在科技飞速发展的当下，ChatGPT 引领的人工智能生成技术（AIGC）正逐渐渗透到各个领域，为诸多行业带来前所未有的变革与机遇。AIGC 技术在虚拟现实三维场景生成中具有多方面的潜在影响，从技术实现到应用拓展，从现存问题到未来研究方向，都值得深入探讨。

## 3.1 AIGC 在虚拟现实三维场景生成中的关键技术应用

### 3.1.1 智能算法驱动的场景元素生成与布局

#### 传统方式概述​

传统场景元素创建依赖设计师手动建模。例如虚拟城市场景搭建，需逐一构建建筑模型并手动布局，自然元素生成也多依赖插件或预设库，耗时且对复杂场景布局优化能力有限。

#### AIGC 相关理论研究

AIGC 中的生成对抗网络（GANs）和变分自编码器（VAEs）在场景生成中表现突出。生成对抗网络由生成器和判别器组成，生成器负责生成场景元素，判别器负责对生成的模型进行评估。通过两者不断对抗训练，生成器能够生成越来越逼真的场景元素模型。

高斯泼溅（Gaussian Splatting）以 3D 高斯分布代替传统网格，无需复杂网格拓扑结构就能实现高质量渲染。用三维高斯分布来表示场景，通过投影到2D屏幕并做加权融合实现渲染。

浙江大学研发了一个室内场景自动灯光设计系统，系统可以在三维室内场景中自动放置不同类型的光源美化照明效果。利用深度神经网络技术从专业设计师的灯光设计中隐式地学习灯光设计的原则，比基于规则的方法更优。[[[13]](#endnote-12)]

#### 实际应用案例​

Unreal Engine（UE）的 RealityCapture基于计算机视觉和深度学习技术，将实拍照片、视频或激光扫描数据转换为高精度三维模型，在写实感建模与纹理还原方面具有极高价值。

英伟达 Omniverse 平台借助智能算法快速生成场景元素并自动布局。可通过简单操作，利用平台中的生成对抗网络模型生成各类高质量的场景元素，同时借助强化学习算法自动布局场景。

一些游戏公司在开发开放世界游戏时，利用智能算法快速生成多样化的游戏场景，如随机生成不同风格的城镇、森林等场景元素，并进行合理布局。如Wonder Dynamics 技术可以将视频中多镜头拍摄角度的序列转换为 3D 空间中的虚拟场景，重建摄像机和角色动画。

|  |  |  |
| --- | --- | --- |
| **对比维度** | **传统建模方式** | **AIGC技术** |
| **场景生成效率** | 依赖手动建模，耗时长 | 基于GANs/VAEs自动生成，效率高 |
| **模型精细度** | 依赖设计师经验，细节可控但优化能力有限 | 通过对抗训练提升逼真度，但对复杂细节仍需优化 |
| **动态布局能力** | 需手动调整，缺乏自适应布局 | 强化学习算法自动布局，支持动态场景生成 |
| **典型案例** | 3ds Max、Maya等工具 | UE RealityCapture、英伟达Omniverse、Wonder Dynamics |

表1：传统建模方式与AIGC技术的差异对比

### 3.1.2 自然语言处理相关的场景描述生成

#### 传统方式概述​

传统场景构建需专业人员使用 3ds Max、Maya 等软件手动建模与布局，对技术与经验要求高，计算量大且硬件依赖性强。

#### AIGC 相关理论研究

自然语言处理（NLP）是 AIGC 的核心技术之一，能将用户自然语言描述转化为场景构建信息。基于大规模预训练语言模型，结合三维模型库与生成算法，可生成三维场景，降低创作门槛，提升生成效率。

北京大学VDIG 研究团队提出了 LLM 引导的复杂三维场景可控生成框架 GALA3D，能够生成高质量、高一致性、具有多物体和复杂交互关系的 3D 场景，支持对话式交互的可控编辑。GALA3D 利用大型语言模型（LLMs）生成初始布局，并提出布局引导的生成式 3D 高斯表示构建复杂 3D 场景，还设计了通过自适应几何控制优化 3D 高斯的形状和分布，以生成具有一致几何、纹理、比例和精确交互的 3D 场景。此外，GALA3D 还提出了一种组合优化机制，结合条件扩散先验和文生图模型，协作生成具有一致风格的 3D 多物体场景。[[[14]](#endnote-13)]

天津大学团队的Narrator 模型从文本描述中自然可控地生成逼真且多样的交互。研究者提出了一个基于关系推理的生成模型，通过场景图分别对场景和描述中的空间关系进行建模，并引入一种将交互动作表示为原子身体部位状态的部位级交互机制。还提出了一种简单但有效的多人生成策略，这是对可控的多人场景交互生成的首次探索。实验结果证明 Narrator 能够可控地生成多样化的交互，效果明显优于现有工作。[[[15]](#endnote-14)]

#### 实际应用案例

Artsteps 是一个基于网页的应用程序，允许用户创建虚拟展览、活动和品牌故事。​用户可以设计逼真的三维空间，并通过自然语言描述来布置场景元素，适用于教育、艺术展示等领域。​

WonderWorld 是一个研究项目，能够从单张图像生成交互式的三维场景，支持用户在虚拟环境中进行实时交互和探索。​该项目的代码和软件已公开，供研究和实验使用。

|  |  |  |  |
| --- | --- | --- | --- |
| **团队/工具** | **核心技术** | **应用领域** | **创新点** |
| **北京大学GALA3D** | LLM引导的3D高斯表示生成 | 复杂多物体交互场景 | 支持对话式编辑，自适应几何控制优化 |
| **天津大学Narrator** | 关系推理+部位级交互机制 | 多人交互场景 | 首次实现可控多人交互生成 |
| **Artsteps** | 自然语言驱动的场景布置 | 虚拟展览、教育展示 | 用户友好，无需专业建模知识 |
| **WonderWorld** | 单图生成交互式3D场景 | 虚拟探索与交互 | 支持实时交互，代码开源 |

表：2 不同研究团队/工具的NLP场景生成技术

### 3.1.3 基于深度学习的场景材质与光照模拟

#### 传统方式概述​

传统材质与光照效果依赖手动设置与传统渲染算法，需精准设置材质参数，计算复杂，动态光照实时性差，难以满足 VR 交互需求。

#### AIGC 相关理论研究

Bruno A. D. Marques 等人提出了一种基于卷积神经网络（CNN）的模型，用于估计混合现实环境中的复杂光照。模型通过输入RGB图像，实时预测环境光照，使用球面调和函数（SH）表示环境光照，能够在没有先验场景信息的情况下高效地表示区域光照。[[[16]](#endnote-15)]

Jiayang Bai 等人提出了一种基于图神经网络（GCN）的框架，用于从单张低动态范围图像预测室内空间变化的光照。核心是一个新的光照模型DSGLight，基于深度增强的球面高斯（SG）和图卷积网络（GCN），能够更真实地编码直接光照和间接环境光照，并实现空间变化光照下的稳定阴影和遮蔽。[[[17]](#endnote-16)]

还有工作提出，利用循环神经网络（RNN）及其变体长短期记忆网络（LSTM），可根据场景的时间、天气等设定，生成相应的光照效果。模型能够学习到不同时间、天气条件下光线的变化规律，如白天阳光的角度、强度变化，夜晚月光、灯光的效果等，并且考虑光线在物体间的反射与折射，使场景光照更真实自然。同时，通过强化学习技术，模型可根据用户在虚拟场景中的交互行为，实时调整光照效果，提升用户体验。

#### 实际应用案例​

一些游戏引擎如 Unity 和 Unreal Engine 已开始集成基于深度学习的材质与光照模拟技术。游戏开发者可利用这些引擎中的深度学习模型，快速为游戏场景中的物体生成逼真的材质效果，并且实现动态光照模拟。

目前也有专门的材质生成工具，如 Substance 3D，它利用人工智能技术，可根据用户输入的材质风格描述生成高质量的材质纹理，广泛应用于游戏、影视等领域。

## 3.2 AIGC 在虚拟现实三维场景生成中存在的问题

### 3.2.1 场景生成的准确性与精细度不足

AIGC 技术在虚拟现实三维场景生成虽有进展，但生成场景的准确性与精细度欠佳。以自然语言生成场景为例，语言的模糊性和多样性使大模型难以精准理解用户意图，导致生成场景与预期有偏差。在场景元素、材质和光照模拟方面，智能算法生成的模型细节粗糙，对复杂材质和特殊光照条件的模拟不够准确。

### 3.2.2 缺乏语义理解与逻辑连贯性​

当前 AIGC 技术在语义理解和场景逻辑连贯性上存在局限。基于自然语言生成场景时，模型常简单组合关键词对应模型，缺乏对语义和场景元素逻辑关系的深入把握，生成的场景生硬、不自然。

### 3.2.3 数据依赖与版权问题

AIGC 技术高度依赖大量数据训练，数据获取和使用存在问题。数据收集整理成本高且质量不一，影响模型训练效果；同时，数据版权归属不明，未经授权使用易引发纠纷，AIGC 生成场景的版权归属也不明确，阻碍其应用和商业化。

## 3.3 未来研究热点

### 3.3.1 强化语义理解与场景逻辑推理

未来研究将改进自然语言处理模型，开发基于知识图谱的模型，提升语义理解能力；开发智能算法，运用强化学习，优化场景逻辑推理，让生成场景更符合用户意图且逻辑连贯。

### 3.3.2 多模态融合与交互技术研究

探索融合图像、语音、手势等多模态信息，实现自然高效人机交互；研究用户与场景实时自然交互，提升用户参与感和体验感。

### 3.3.3 提升生成场景的个性化与创新性

构建用户画像分析模块，结合GAN等生成网络，打造风格化场景定制能力。开发创意激发算法，通过对抗训练突破模式化生成，在科幻和超现实等特殊场景中展现创新性。

### 3.3.4 解决数据与版权相关问题

采用迁移学习降低数据依赖，研究小样本学习提升模型效率。建立区块链存证体系，实现数据溯源与版权追踪，配合数字水印技术构建版权保护双机制。推动行业标准制定，明确AIGC内容权属规则。

随着语义理解、多模态交互等关键技术的突破和数据治理体系的完善，AIGC将推动虚拟场景生成向智能化、个性化方向演进，为教育、文旅、工业仿真等领域提供高效场景构建工具，加速VR/AR技术的普及应用。

# 4 AIGC驱动的虚拟角色与对话系统

在虚拟角色领域，ChatGPT和人工智能生成技术的快速发展和广泛应用也为VR行业带来了新的可能。AIGC赋能虚拟角色，使其具有更完善的人设和更灵活的交互，虚拟角色从电子傀儡走向赛博生命，从机械式走向智能化。

## 4.1 传统虚拟角色的局限性

传统虚拟角色长期受限于预设脚本和有限状态机技术，其交互行为呈现明显的机械性特征。以游戏NPC为例，角色对话通常采用树状对话结构，缺乏动态响应能力。很多用户因机械式对话在一段时间后弃用脚本型聊天机器人。此外，传统虚拟角色的情感识别能力存在显著缺陷，基于规则的情感识别系统对情绪的识别准确率不高。用户情感被忽视，使用体验糟糕。这些局限性在元宇宙和数字孪生快速发展的背景下愈发凸显，亟待解决。

## 4.2 基于ChatGPT的虚拟角色生成

新一代虚拟角色生成技术借助大语言模型（LLM）实现了多维人格建模。通过多种技术的整合，虚拟角色不只拥有外貌，还可以拥有神态、动作、声音，甚至是自己独特的性格。

### 4.2.1 技术介绍

结合多模态技术，虚拟角色生成涵盖外貌（CLIP驱动的图像合成）、神态（Unreal Engine MetaHuman的微表情控制）、声音（VALL-E的3秒语音克隆）、动作等要素。下面对技术进行介绍。

CLIP（Contrastive Language-Image Pre-training）是由OpenAI推出的、将图像和文本映射到同一语义空间多模态模型，具有零样本能力，无需微调即可实现跨模态检索、图像分类等功能，同时还支持通过文本语义调整图像生成方向[[[18]](#endnote-17)]。以抖音虚拟博主“柳夜熙”为例，GPT-4生成角色设定文本，CLIP将其转换为视觉特征向量，驱动图像生成模型输出符合设定的形象。用户与GPT-4进行对话，CLIP实时调整生成图像的属性，形成闭环迭代。

在虚拟角色的生成过程中，除外貌之外，角色的表情也是情感传递的关键。Unreal Engine MetaHuman是一个用于虚拟角色生成的工具，其强大之处在于其对微表情的精细控制。创作者除定制角色外貌外，还可以调整角色的面部表情、眼神、嘴型等细节，甚至能够模拟出非常细微的情感波动。当虚拟角色在进行对话时，能够通过表情和眼神的变化来传递情感信息，大大增强虚拟角色的沉浸感和互动性。此外，Pan Y等人2025年的论文中引入了一个基于绑定参数的全新情感面部表情数据集，并提出了 Latents2Rig 模型，声称这一方法在真实感和准确性方面均显著优于现有的最佳方法。[[[19]](#endnote-18)]

Wang等人于2023年提出 VALL-E模型，将TTS视为一种条件语言建模任务，训练数据规模扩大到60K小时的英语语音。VALL-E模型具备上下文学习能力，能使用仅3s的录音作为声学提示合成高质量个性化语音，并保留说话者情感和声学环境。Vall-E在语音自然性和说话者相似性方面显著优于最先进的零-shot TTS系统[[[20]](#endnote-19)]。借助VALL-E，用户可以自由为虚拟角色克隆声音使其以该声音进行对话，大大增强了角色的亲和力和真实感，提高沉浸度。

动作生成方面，北京理工大学的团队提出了一种新的方法框架，通过结合指令和场景信息，风格化生成虚拟角色的动作[[[21]](#endnote-20)]。方法预定义一组有限的原子动作集合，在指令解析部分运用大模型将文本指令解析成由原子动作组成的子任务，在动作生成部分基于条件变分自编码器（conditional variational autoencoder，cVAE）设计逐帧动作生成网络。在动作生成阶段，考虑不同的风格特征实现风格化动作生成任务。

### 4.2.2 技术融合案例​

随着虚拟角色技术的不断发展，单一技术往往难以满足复杂、逼真且富有表现力的数字人物需求。下面以NVIDIA Audio2Face技术与Soul Machines的生物神经网络系统为例探讨如何实现技术整合使虚拟角色更加逼真。

NVIDIA的Audio2Face技术识别音频中的语音内容、语调、节奏和情感特征，并将这些特征映射到虚拟角色的面部动作上，将声音转化为与之匹配的面部表情，尤其是在口型同步方面表现突出。Soul Machines则专注于通过生物神经网络模拟人类自然行为，尤其是面部动态反应。它不仅停留在单一面部表情，还能模拟复杂面部反应，如自然真实地眨眼等微小细节，极大提升了虚拟角色的生动性和人类化。当这两项技术结合时，音频输入实时驱动虚拟角色的面部表情，Soul Machines在基础面部动作上加入更加细致的神经网络驱动，如自然的眨眼、面部肌肉的轻微波动等，增强角色的自然表现力。

### 4.2.3 个性化设定

在AIGC技术的推动下，虚拟角色的性格、行为模式可以被用户定义。例如，微软小冰基于用户输入的背景故事和关键词（如“傲娇学霸”“社恐艺术家”），通过知识图谱和情感计算模型动态生成角色的语言风格与互动模式，支持长篇连贯人设塑造；《AI Dungeon》游戏NPC采用GPT-3+世界模型生成具有复杂动机的虚拟角色，玩家可通过文本指令实时修改角色的性格倾向（如“让反派突然产生同情心”）。

个性化设定的关键技术包括人格嵌入、多模态行为生成、动态人设演化等。人格嵌入（Persona Embedding）是指将性格特质（如MBTI类型）编码为向量，结合对话生成模型（如BlenderBot），实现可控性格输出。多模态行为生成是指通过StyleGAN-V + 语音驱动模型同步生成角色的表情、手势与语调，强化性格表现。动态人设演化则基于强化学习（RL）的角色行为优化框架，使虚拟角色在长期互动中自主调整性格。Madotto[[[22]](#endnote-21)]等人在2019年的论文中指出，现有个性化对话模型依赖人工设计的角色描述来提升对话一致性，成本高昂且需手工设计特征。论文提出将模型无关元学习（MAML）框架扩展至个性化对话任务，无需使用任何角色描述，仅通过同一用户的少量对话样本即可快速适应新角色。实验表明该方案在多方面表现优秀，或将成为虚拟角色性格设定的新范式。

## 4.3 基于ChatGPT的逼真对话实现

### 4.3.1 上下文理解与记忆机制

传统的对话系统通常会因为上下文信息丢失或短期记忆限制，导致多轮对话中准确性下降。而ChatGPT基于Transformer-XL架构（Dai等，2019）[[[23]](#endnote-22)]，可以捕捉长期依赖关系，解决了上下文碎片化问题，在多轮对话场景下表现良好。

短期记忆方面，ChatGPT动态保留最近N个token的上下文，并通过键值缓存避免重复计算历史文本，另外还具备层次化记忆压缩功能，对早期对话内容进行语义摘要，生成用户偏好。长期记忆则依赖Lewis等人（2020）提出的 检索增强生成（RAG）[[[24]](#endnote-23)]框架，历史对话编码为向量存储于Pinecone等数据库，检索时基于余弦相似度注入相关上下文，使多轮对话轮次提升。记忆管理方面，通过LRU算法自动淘汰低权重记忆，避免噪声干扰；通过置信度评分（如BERT矛盾检测[[[25]](#endnote-24)]）应对冲突，更新数据库。

### 4.3.2 情感计算

情感计算使虚拟角色在理解用户语义内容的基础上进一步识别并响应用户的情感状态。Zhang等人（2024）系统综述了大语言模型（LLM）在情感计算中的进展[[[26]](#endnote-25)]。在 LLM 出现之前，情感计算的主导范式是微调预训练语言模型 (PLM) [[[27]](#endnote-26)]，例如 BERT和RoBERTa[[[28]](#endnote-27)]，但其受数据集限制多且泛化能力差。具有强大泛化能力的LLM则为情感计算带来新的机遇。论文总结了LLM在AC中的能力和局限性相关的早期研究，并在此基础上探索了两种使LLM适应AC任务的关键技术：指令调整和提示工程。论文对LLM用于情感计算的前景表示了认可。

实际应用方面，ChatGPT整合BERT情感分析模型与微软Custom Neural Voice，实时检测用户情感信号（如文本负面词频、语音基频上升），并生成适配响应（如悲伤时降低语速、增加安慰性措辞）；微软小冰在对话中添加呼吸声和停顿，当检测到用户情绪低落时，语速放慢、语气变得温和。情感计算和反馈机制使用户与虚拟角色的互动更加贴合人类期待的情感反应，提升了对话的逼真度和亲和力。

## 4.4 AIGC驱动的虚拟角色应用场景

AIGC驱动的虚拟角色在现实生活各个领域如旅游业、服务业、金融行业等已经得到了广泛应用。敦煌研究院的AI导游整合GLAM知识图谱，为游客提供高精度的相关知识解答；Replika应用认知行为疗法框架，使用户与AI对话进行心理疏导，取得卓越成效；摩根大通的COIN系统部署虚拟财务顾问，处理合规查询的耗时大幅减少；网易《逆水寒》手游NPC支持6.5万种动态对话组合，任务路径多样性提升70%。

## 4.5 面临的问题和未来发展方向

当前，AIGC虚拟角色面临三大核心挑战。首先是伦理安全，研究表明，AIGC虚拟角色生成的回答可能具有道德偏差，甚至于引发社会和法律问题。其次，语言泛化问题较为凸显，跨领域对话准确率存在显著下降。最后，实时响应仍是关键瓶颈，复杂场景下虚拟角色响应延迟较长，高于理想的人类交互阈值，影响用户体验。

面对这些挑战，未来的发展方向呈现出三大趋势：1.多模态融合。Google的PaLM-E模型联合训练视觉与语言，使物理交互成功率显著提升，虚拟角色的智能化与跨界能力得到推进。2.神经渲染技术突破。NERF（Mildenhall等，2021）[[[29]](#endnote-28)]技术大幅提升3D角色建模速度，增强了虚拟角色在实时交互中的表现能力。3.情感计算的深化。MIT Affectiva团队通过微表情识别技术的提升大幅提高情绪判断的准确率，使虚拟角色能够更精确地响应用户情感，更加人性化。

AIGC正在重塑人机交互范式，虚拟角色从"数字傀儡"进化为"智能伙伴"。当技术突破与伦理框架同步完善时，我们将迎来真正具有社会价值的数字生命形态。

# 5 结论

本文围绕AIGC技术在虚拟现实内容创作中的应用与挑战，系统分析了三维模型生成、虚拟场景构建与虚拟角色生成与对话三个关键方向，并对相关前沿方法和代表性工具进行了深入探讨。

在三维模型生成方面，本文详细回顾了从传统手工建模向AIGC自动生成的转变路径，分析了从GAN、IM-NET、NeRF到扩散模型的发展演进，重点剖析了Text2Shape、DreamFusion、Magic3D、Shap·E等典型方法的技术机制与优势。然而，当前存在生成质量不稳定、控制粒度不足与几何细节保真性差等问题，未来可从“增强交互式设计机制”“加速算法优化与算力提升”等方向进一步突破。

在三维场景构建方面，本文聚焦AIGC在场景元素生成、空间布局自动化与自然语言驱动建构中的应用潜力，探讨了GAN、LLM及多模态模型在复杂环境生成中的优势与不足。通过分析GALA3D、Narrator等代表性框架，揭示当前语义理解有限、场景逻辑缺失与缺乏个性化调控等问题，并提出了“融合语义图谱的生成机制”“基于多模态推理的布局控制算法”“面向用户交互反馈的场景生成系统”等未来研究方向。

在虚拟角色生成与对话方面，本文研究深入分析了ChatGPT驱动下虚拟角色的生成、塑形与交互表现能力，涵盖外貌建构（如CLIP）、语音合成（如VALL-E）、表情控制（如MetaHuman、Audio2Face）与对话响应。本文指出当前仍面临情感识别粗糙、行为反应延迟与语境记忆不足等难点，建议持续探索“多模态情感计算体系”“个性化记忆增强机制”“动态人设演化模型”等前沿方向，以实现更具智能感与亲和力的虚拟交互主体。

综上所述，AIGC正在为虚拟现实内容生产提供智能化、自动化的新范式，但其仍面临生成可控性、语义一致性与伦理治理等挑战。我们认为，未来应进一步推动语义增强、多模态融合、边缘计算优化与知识驱动机制的深度结合，同时构建统一的数据规范与版权保护体系，推动AIGC与VR技术实现高质量融合与落地。随着这些瓶颈的逐步解决，AIGC将深刻改变VR内容创作范式，为教育、医疗、工业仿真、文化传播等领域开启更具创造力与包容性的虚拟现实时代。

# 6 附录

## 6.1 PPT汇报材料

<https://github.com/MHFeng907/VR_Final_Paper.git>

## 6.2 参考文献

1. [] Jiajun Wu, Chengkai Zhang, Tianfan Xue, William T. Freeman, and Joshua B. Tenenbaum. 2016. Learning a probabilistic latent space of object shapes via 3D generative-adversarial modeling. In Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS'16). Curran Associates Inc., Red Hook, NY, USA, 82–90. [↑](#endnote-ref-0)
2. [] Z. Chen and H. Zhang, "Learning Implicit Fields for Generative Shape Modeling," 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 2019, pp. 5932-5941, doi: 10.1109/CVPR.2019.00609. [↑](#endnote-ref-1)
3. [] Chen, Kevin & Choy, Christopher & Savva, Manolis & Chang, Angel & Funkhouser, Thomas & Savarese, Silvio. (2018). Text2Shape: Generating Shapes from Natural Language by Learning Joint Embeddings. 10.48550/arXiv.1803.08495. [↑](#endnote-ref-2)
4. [] A. Sanghi et al., "CLIP-Forge: Towards Zero-Shot Text-to-Shape Generation," 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 2022, pp. 18582-18592, doi: 10.1109/CVPR52688.2022.01805. [↑](#endnote-ref-3)
5. [] A. Jain, B. Mildenhall, J. T. Barron, P. Abbeel and B. Poole, "Zero-Shot Text-Guided Object Generation with Dream Fields," 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 2022, pp. 857-866, doi: 10.1109/CVPR52688.2022.00094. [↑](#endnote-ref-4)
6. [] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. 2021. NeRF: representing scenes as neural radiance fields for view synthesis. Commun. ACM 65, 1 (January 2022), 99–106. https://doi.org/10.1145/3503250 [↑](#endnote-ref-5)
7. [] Poole, Ben & Jain, Ajay & Barron, Jonathan & Mildenhall, Ben. (2022). DreamFusion: Text-to-3D using 2D Diffusion. 10.48550/arXiv.2209.14988. [↑](#endnote-ref-6)
8. [] C. -H. Lin et al., "Magic3D: High-Resolution Text-to-3D Content Creation," 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 2023, pp. 300-309, doi: 10.1109/CVPR52729.2023.00037. [↑](#endnote-ref-7)
9. [] Nichol, Alex & Jun, Heewoo & Dhariwal, Prafulla & Mishkin, Pamela & Chen, Mark. (2022). Point-E: A System for Generating 3D Point Clouds from Complex Prompts. 10.48550/arXiv.2212.08751. [↑](#endnote-ref-8)
10. [] Jun, Heewoo & Nichol, Alex. (2023). Shap-E: Generating Conditional 3D Implicit Functions. 10.48550/arXiv.2305.02463. [↑](#endnote-ref-9)
11. [] Li, Weiyu & Liu, Jiarui & Chen, Rui & Liang, Yixun & Chen, Xuelin & Tan, Ping & Long, Xiaoxiao. (2024). CraftsMan: High-fidelity Mesh Generation with 3D Native Generation and Interactive Geometry Refiner. 10.48550/arXiv.2405.14979. [↑](#endnote-ref-10)
12. [] Wei, Si-Tong & Wang, Rui-Huan & Zhou, Chuan-Zhi & Chen, Baoquan & Wang, Peng-Shuai. (2025). OctGPT: Octree-based Multiscale Autoregressive Models for 3D Shape Generation. 10.48550/arXiv.2504.09975. [↑](#endnote-ref-11)
13. [] Zhang, J., Wang, H., Liu, Y., & Li, K. (2021). Data-driven digital lighting design for residential indoor spaces (arXiv:2108.07903). arXiv. https://arxiv.org/abs/2108.07903 [↑](#endnote-ref-12)
14. [] Zhou, X., Ran, X., Xiong, Y., He, J., Lin, Z., Wang, Y., Sun, D., & Yang, M.-H. (2024). GALA3D: Towards text-to-3D complex scene generation via layout-guided generative Gaussian splatting (arXiv:2402.07207). arXiv. https://arxiv.org/abs/2402.07207 [↑](#endnote-ref-13)
15. [] Xuan, H., Li, X., Zhang, J., Zhang, H., Liu, Y., & Li, K. (2023). Narrator: Towards natural control of human-scene interaction generation via relationship reasoning (arXiv:2303.09410). arXiv. https://arxiv.org/abs/2303.09410 [↑](#endnote-ref-14)
16. [] Marques, B. A. D., Clua, E. W. G., Montenegro, A. A., & Vasconcelos, C. N. (2021). Spatially and color consistent environment lighting estimation using deep neural networks for mixed reality. Computers & Graphics, 100, 75–85. [https://arxiv.org/abs/2108.07903](https://arxiv.org/abs/2108.07903" \t "_new) [↑](#endnote-ref-15)
17. [] Bai, J., Guo, J., Wang, C., Chen, Z., He, Z., Yang, S., Yu, P., Zhang, Y., & Guo, Y. (2023). Deep graph learning for spatially-varying indoor lighting prediction. Science China Information Sciences, 66(3), 132106. [https://doi.org/10.1007/s11432-022-3576-9](https://doi.org/10.1007/s11432-022-3576-9" \t "_new) [↑](#endnote-ref-16)
18. [] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision[C]//International conference on machine learning. PmLR, 2021: 8748-8763. [↑](#endnote-ref-17)
19. [] Pan Y, Liu C, Xu S, et al. VASA-Rig: Audio-Driven 3D Facial Animation with ‘Live’Mood Dynamics in Virtual Reality[J]. IEEE Transactions on Visualization and Computer Graphics, 2025. [↑](#endnote-ref-18)
20. [] Wang C, Chen S, Wu Y, et al. Neural codec language models are zero-shot text to speech synthesizers[J]. arXiv preprint arXiv:2301.02111, 2023. [↑](#endnote-ref-19)
21. [] 梁玮, 黄一帆, 沈雨欣, 等. 指令驱动的虚拟角色风格化动作生成[J]. 北京理工大学学报自然版, 2024, 44(11): 1199-1207. [↑](#endnote-ref-20)
22. [] Madotto A, Lin Z, Wu C S, et al. Personalizing dialogue agents via meta-learning[C]//Proceedings of the 57th annual meeting of the association for computational linguistics. 2019: 5454-5459. [↑](#endnote-ref-21)
23. [] Dai Z, Yang Z, Yang Y, et al. Transformer-xl: Attentive language models beyond a fixed-length context[J]. arXiv preprint arXiv:1901.02860, 2019. [↑](#endnote-ref-22)
24. [] Lewis P, Perez E, Piktus A, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks[J]. Advances in neural information processing systems, 2020, 33: 9459-9474. [↑](#endnote-ref-23)
25. [] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers). 2019: 4171-4186. [↑](#endnote-ref-24)
26. [] Zhang Y, Yang X, Xu X, et al. Affective computing in the era of large language models: A survey from the nlp perspective[J]. arXiv preprint arXiv:2408.04638, 2024. [↑](#endnote-ref-25)
27. [] M. M. Amin, E. Cambria, and B. W. Schuller, “Will affective computing emerge from foundation models and general artificial intelligence? a first evaluation of chatgpt,” IEEE Intelligent Systems, vol. 38, no. 2, pp. 15–23, 2023. [↑](#endnote-ref-26)
28. [] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, “Roberta: A robustly optimized bert pretraining approach,” 2019. [↑](#endnote-ref-27)
29. [] Mildenhall B, Srinivasan P P, Tancik M, et al. Nerf: Representing scenes as neural radiance fields for view synthesis[J]. Communications of the ACM, 2021, 65(1): 99-106. [↑](#endnote-ref-28)