UHD World Association
世界超高清视频产业联盟

UHD World Association
世界超高清视频产业联盟

UHD Word Association
www.theuwa.com

元宇宙时代超高清音视频技术白皮书

(征求意见稿)

世界超高清视频产业联盟

前 言

本文件由 UWA 联盟 xxx 组织制订，并负责解释。

本文件发布日期：xxxx 年 xx 月 xx 日。

本文件由世界超高清视频产业联盟提出并归口。

本文件主要起草单位：

中国移动通信集团有限公司、咪咕文化科技有限公司、中国移动通信有限公司研究院、中兴通讯股份有限公司、世界超高清视频产业联盟、中国信息通信研究院、腾讯计算机系统有限公司、华为技术有限公司、中国电子技术标准化研究院、工业和信息化部电子第五研究所、深圳思谋信息科技有限公司、凌云光技术股份有限公司、杭州当虹科技股份有限公司、北京全景声信息科技有限公司、广州视源电子科技股份有限公司、北京数码视讯科技股份有限公司、杭州海康威视数字技术股份有限公司、上海交通大学、中央广播电视总台、国家广播电视总局广播电视规划院、中国电信集团有限公司

本文件主要起草人：

李琳、徐嵩、王琦、单华琦、毕蕾、喻炜、杨蕾、郭勐、赵丽丽、黄成、李秋婷、刘耀东、张文刚、王斌、陈曦、王琼、胡颖、许晓中、刘杉、王志刚、谷晓军、赵晓莺、耿一丹、邱溥业、韦胜钰、蔡佳、赵轶、刘志杰、陈家兴、陈左乐、许舒敏、王虽然、潘兴德、熊伟、杜华、曾义、李淼、张玉兵，王乃洲、魏晔、刘利华、王福河、闫科锋、宋利、王子建、李厦、罗传飞

免责说明：

1， 本文件免费使用，仅供参考，不对使用本文件的产品负责。

2， 本文件刷新后上传联盟官网，不另行通知。

目录

1. 背景及意义

元宇宙孕育新产业新业态。随着新一代信息技术的迅猛发展，人们的生活生产处在一个现实世界和数字世界日益融合的进程中。2021 年以来，元宇宙概念快速升温，包括微软、谷歌、脸书、英伟达、腾讯等全球市值前十科技企业在内的国内外 ICT 巨头与初创企业纷纷发声。全球元宇宙第一股 Roblox 自 2021 上市当日突破 400 亿美元市值，全球最大社交网络平台 Facebook 宣布更名为 Meta，表示未来五年力争由传统社交平台转变为元宇宙公司，并承载数千亿美元的数字商务，为数百万创作者和开发者提供就业机会。微软表示元宇宙使计算嵌入到现实世界中，为数字空间带来真实临场感。英伟达发布工业元宇宙平台 Omniverse，旨在为影视、工业等行业应用提供模拟仿真与协同开发环境。我们认为，元宇宙是新一代信息技术的融合创新，是虚实相融的沉浸式互联空间。元宇宙的发展必定带动大量新兴产业并在现有产业中促成新业态，有望创造由数字“比特”与人类“原子”深度融合的新型社会景观。

各国政府开始布局元宇宙领域。美国从特定局部推动元宇宙概念下关键领域创新发展，美国国会通过《2021 年美国创新和竞争法案》，旨在扩大政府在科研中作用。该法案提出在五年内为美国国家科学基金会提供 1000 亿美元。其中，作为元宇宙主要支撑的“先进的通信和沉浸式技术”位列十大关键技术领域之一。韩国从技术创新、经济发展、社会民生全局开展元宇宙顶层设计，明确提出“元宇宙”发展的规划举措。2020 年底韩国总理在国家政策审查会议中宣布“沉浸式经济发展战略”。2021 年 11 月，韩国首尔市政府发布了《元宇宙首尔五年计划》，宣布从 2022 年起分三个阶段在经济、文化、旅游、教育、信访等市政府所有业务领域打造元宇宙行政服务生态。2022 年 1 月韩国政府公布《元宇宙新产业领先战略》，以“数字新大陆，迈向元宇宙的韩国”为愿景口号，提出截止到 2026 年，元宇宙产业规模全球前五，专业人才规模不少于四万，50 亿韩元以上元宇宙企业数量不低于 220 家，并围绕生产生活诸多领域挖掘落地 50 个“元宇宙+”创新应用场景。我国各地政府也于 2021 年开始布局元宇宙赛道。2021 年 8 月北京市启动元宇宙总体布局，考虑以互联网 3.0 指代元宇宙，并将其视为继互联网、移动互联网后下一代互联网的新业态，后续将开展元宇宙底层技术攻关。

2022 年 7 月，上海市政府印发《培育“元宇宙”新赛道行动方案》，成为我国首份较为系统具体的元宇宙专项政策，文件提出到 2025 年上海元宇宙产业规模达 3500 亿元。此外，自 2021 至今，合肥、武汉、杭州、成都、青岛、广州等地政府也相继发布政策，都主要从新一代互联网、数字经济、未来产业等视角编制元宇宙发展政策。

元宇宙将触发超高清音视频用户体验增量跃迁与技术产业持续演进。当前大众对美好生活的进阶需求对数字内容的体验方式提出了新要求，分辨率、高动态范围等视听质量维度的常规迭代难以带来用户体验的增量跃升。回顾对照移动互联网大众视听彼时创新业态的影响范围，如长视频带给用户“三高”体验，而画质提升带动了芯片、屏幕、编解码标准与拍摄器材等产业链迭代升级。短视频呈现用户全年龄、使用全时段、内容全题材的显著态势，且数字内容生产方式由以往专业机构产出（PGC）向大众用户产出（UGC）、人工智能内容生成（AIGC）的方式变革。在元宇宙概念下，超高清数字内容既需要做到让大众有明显感知，又能引领生活方式变革与产业结构创新。相比大众隔着手机、个人电脑上的 2D 屏幕点击浏览互联网，元宇宙中人们将“活”在互联时空里，学习、办公、娱乐、健身、购物、社交等人人交互、探索世界的共同体验发生在数字时空。现实时空与数字时空不再彼此分割，在虚实融合的互联时空下，未来超高清音视频生态的构建有赖于传统音视频向沉浸式、交互性与多维化方向发展演进。

2.元宇宙时代超高清音视频技术发展需求

元宇宙重新定义了人与空间的关系，**虚拟现实、人工智能、区块链、未来网络、数字孪生、先进计算**等技术搭建了通往元宇宙的通道，创造了虚拟与现实融合的交互方式，并正在改变和颠覆我们的生活。

2.1 应用场景

元宇宙处在初级阶段，正快速的改变着我们工作和生活，以下是一些典型场景。

1. 零售场景

在元宇宙时代的线上购物场景，通过超高清音视频技术呈现沉浸式的一维购物空间，升级顾客线上购物的消费体验。比如，创造线上数字购物空间，在线上直接模拟体验穿衣换衣服，用户通过交互式游戏化的体验完成消费。在线下购物场景，通过增强现实技术，提供数字化的商品 360 度悬浮展示，以及通过数字人营业员提供导购、导览服务，给用户提供特色创意的数字购物体验。比如，中国移动推出的 5G+AR 智慧营业厅和咪咕数字咖啡馆。

图 1 中国移动 5G XR 营业厅

2. 泛娱乐场景

元宇宙时代，**在体育和演艺直播**场景，赛事直播和线上演艺开拓了新的体验和交互方式。传统直播或演唱会中，粉丝的观看视觉有限，与偶像互动方式单一。而元宇宙中的演唱会则通过超高清技术呈现更有沉浸感的现场，并可以提供更灵活的游戏体验现场互动，如突破空间限制，粉丝能够通过虚拟分身的方式参与演出，通过互动游戏和歌手对应在虚拟世界的虚拟人、虚拟道具亲密互动、舞台效果千变万化。还可以提供更智能贴心的无障碍观赛体验，如卡塔尔世界杯直播中“为了听不到的你”智能字幕和数字人手语解说。

图 2 北京冬奥“集光之夜”数字人和演员互动演出和卡塔尔世界杯数字人手语主播

在游戏场景，元宇宙将现实生活的真实感带入游戏中，玩家可以有更沉浸的视听感受，通过穿戴设备实现触觉反馈，可以获得接近现实、超越现实的体验；此外在虚拟游戏中可以拥有类似现实世界实物资产的虚拟资产。元宇宙场景游戏的发展方向之一是降低硬件设备的门槛，进一步加快游戏云化，并且对超高清技术在沉浸式体验、实时互动等方面提出更高的要求。

图 3 北京冬奥“冰雪小镇”AR 小游戏

在运动健身场景，元宇宙相关技术把现实的健身融入到虚拟世界。元宇宙健身房支持用户随时随地利用碎片化时间健身，突破空间限制“面对面”与教练在线互动，用户可以使用虚拟身份加入虚拟社群，与健身爱好者们 PK 和交流。如咪咕善跑正在通过结合虚拟现实技术以虚拟骑行的形式打造 XR 骑行平台。基于虚拟现实的数智竞技场景，为运动健身提供更多沉浸的对抗性体验，VR 电竞已获奥委会、亚运会等认可。

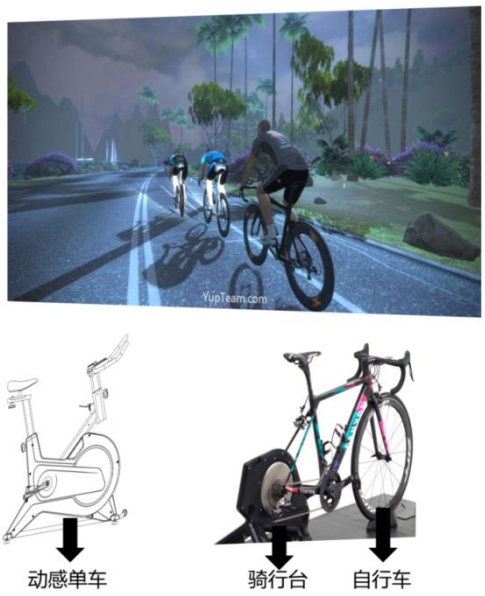

图 4 XR 骑行和数智竞技

3. 文旅场景

元宇宙的文旅场景，为游客重构了时空体验，让文化潮起来。线上旅游交互效果升级，增加沉浸感和参与感，游客只需要戴上虚拟现实设备便能够来一场“说走就走的旅行“。线下旅游通过增强现实技术，可以拓展时间、空间的体验，打造炫彩实景、梦幻虚拟、跨界互动。文旅景区可以利用 AR、VR、渲染等技术加持，观众可以真正进入到 XR 奇妙空间，第一视角沉浸感受璀璨的中国文化和美景。

图 5 厦门 98 投洽会 AR 夜景秀和鼓浪屿 AR 文化导览

4. 教育场景

在元宇宙的教育场景，教学不再受限于时空条件和真实设备，教育资源缺少和不均衡的问题得到改善，且各种创意课堂能够被激发，教学效果即将获得质和量的提升。如数学课堂中的图形和公式能够在几何空间中不断组合和变化，科学课堂中宇宙的产生和发展过程就在眼前模拟重现，历史课堂中学生甚至能够和历史人物互动。元宇宙的教育还能激发学生的创造性，突破现实约束展开各种创新实验等。

5. 政务场景

在元宇宙的办公场景，传统的远程办公缺少实时互动、沉浸感不足。而元宇宙技术能够使得虚拟办公以“面对面”互动的方式进行。元宇宙办公提升了临场感，让我们的互动方式更加自然、真实。未来，用户还可以通过 3D 化远程互动形式，从不同角度观察聊天对象，并进行更丰富的肢体和触感交流。

此外在学术、产业会议、活动等场景，通过元宇宙技术突破时空限制，实现线上线下联动的互动的体验。元宇宙中的集会将会是 3D 沉浸式的，通过 AR、VR 技术，线上和线下的参与者可以选择以虚拟形象、真实影

像出现，线上、线下齐聚一堂，互相交流，提升会议的参与感。目前，不少学术讲座论坛、毕业典礼等集会都以元宇宙的方式举行。

6. 工业场景

元宇宙工业生产场景，基于数字孪生技术，元宇宙生产车间提升了工厂设计和生产的精度、速度，并可实现节省成本和提高效率。元宇宙为工业生产打破时空的限制，促进生产研发涉及专家的远程协同，在虚拟空间中进行模拟测试，提前验证落地可能性，降低试错成本和风险。

此外在数字金融、数字农业等很多场景有很多应用，而且随着元宇宙的发展，新的应用场景还会不断拓展。这些场景对超高清音视频技术都提出了新的技术要求。

2.2 技术新需求

从元宇宙应用场景来看，为满足视觉、听觉以及触觉方面的体验要求，对超高清音视频技术在沉浸式体验、实时多维互动、高效内容生产和用户大规模在线能力等方面都有新的需求。

1. 沉浸式体验

打造面向元宇宙的视听体验，就必须要解决超高清音视频场景中的人、物、场实现问题以及沉浸式体验、可交互体验的实现问题。传统的超高清视觉和听觉呈现技术以二维内容为主，表现形式升级势在必行。以全景视频、自由视角视频、点云、空间音频等为代表的元宇宙音视频内容生产、高效编码，沉浸式影像的传输与呈现等方向进行技术演进，实现虚拟与现实融合的沉浸式体验。

1）沉浸式视频

而基于当前技术发展现状，人们所观看的视频类型仍以二维平面视频为主，并逐渐出现全景视频，虽然可提供初级的沉浸式体验，但是缺少互动，也不适合于长时间观看。自由视点视频、点云和光场视频，可以自由移动切换观看视角、具有临场感、如在体育直播场景可切换球员视角观赛、VIP 席位虚拟视角观赛等，更具有沉浸感、临场感，也更符合元宇宙体验的观赛体验。但当前还存在一些困难，一是场景的复杂性导致三维重

建的难度提高，从而影响内容生成质量；二是满足所要达到的效果需要的计算资源限制导致处理速度无法实现实时；三是重建产生的三维数据因为数据量、数据格式等问题，无法使用传统的流媒体方式以及移动网络进行传输；四是用户的观看体验端也需要强大的计算渲染能力，受限于当前的内容生产技术，成本很高，且较难保证用户的体验效果，当前仅能用在耗时较长的视频内容制作场景，无法应用于直播。

2）空间化音频

在视觉效果沉浸感的提升基础上，还需有空间化音频与之对应，空间化音频在元宇宙场景不是锦上添花而是必选。听觉是仅次于视频的重要感官，是人获取信息及定位的必要输入，基于没有空间感音频的虚拟世界会让人寸步难行。当前基于空间的音频体验在内容上以三维声为主要方式。和传统的多声道音频不同，空间化音频内容主要有两个重要要素，一是音频对象，声音场景中与发声体绑定的点声源或体积声源；二是环境音频，与场景相关的音频内容。这对音频的采集、制作、传输、到渲染和最终呈现都有新的技术需求。

3）沉浸式终端

沉浸式音视频的呈现离不开沉浸式终端的支持，满足沉浸式体验仍有一些技术点需要突破。VR 头显，在视场角（FOV）方面当前设备还不能覆盖人的视野宽度 120°，未来需要更大的 FOV 提供更强的沉浸感；终端显示的时延会影响用户的注意力，延迟较大的时候会产生晕动症，MTP 时延控制在 20ms 以内，人体才不会有排斥反应；视觉效果需要强大的算力支撑，高精度模型、高清晰度 VR 视频、数字孪生、仿真设计的应用对设备的算力提出了越来越高的要求。AR 智能终端在硬件层面上，需要达到单眼屏幕分辨率大于 8K 超高分辨率，保证 120HZ 屏幕刷新率，FOV 角度大于 140 度，同时支持可变焦显；在软件层面上，需要支持超 8K 360 全景视频适配；同时在网络传输方面，就目前需求看，需要满足 MTP 时延至少 20ms 以内。

当前 VR/AR 设备越来越趋向于 MR 设备的方向发展，以及未来基于裸眼而不借助屏幕的全息呈像，可高度还原物体的三维特征，体验到与观看原物体时近似相同的视觉效果。这些对超高清音视频技术都提出了更多的要求，如实时空间定位识别、跟踪、实时渲染能力、呈现技术的提升等。

2. 实时多维互动

元宇宙时代的应用，仅视觉和听觉的沉浸技术提升还不够，还需要强交互，这对交互的自由度、实时性及多模态等方面提出了更高的要求。

1）更高自由度

配合沉浸式的音视频呈现，需要支持以用户为中心的视觉和听觉呈现，要有配套的更高自由度的交互能力。如全景场景三自由度（3DOF）交互，以及自由视点视频、点云视频等六自由度（6DOF）交互。在 3DOF 场景，需要通过传感器获得用户头部方位信息，视频通过显示设备为用户呈现对应视角的视频内容，音频也同样需要配合头部方位信息，渲染出对应效果的双耳音频。在 6DOF 场景，需要在头部方位信息的基础上，增加用户身体跟踪，如上下左右等平移信息，呈现对应效果的音视频体验。

2）多模态交互

传统基于输入设备触发指令或者输入文本的交互方式已经不能满足元宇宙沉浸式内容的需求。需要扩展到更多模态的交互，如语音交互，对身体动作、眼球等的捕捉，实现肢体、手势、表情等三维空间交互等。

3）实时渲染

为了实现以上灵活的交互方式和丰富的感官体验，需要在内容生产中提升配套的实时渲染能力。元宇宙场景下实时渲染将呈现以下几个发展方向，由单点渲染迈向分布式实时渲染，实现分布式物理、材质等仿真；由端侧渲染迈向云、边、端协同渲染，同时实现异构图形处理器（GPU）容器化调度渲染；由 1080P、4K 迈向 8K、16K 的超高清渲染；由光栅化为主迈向和光线追踪共存、最终实现路径追踪；由单人单视角渲染，迈向多人多视角共享渲染。

3. 高效内容生产

1）内容创作平台/工具

当前平面内容的创作平台已很成熟，然而面向元宇宙的内容创作平台还处于初期。如何支持空间音频、自由视点视频、点云视频等新媒体创作平台将会是新方向、新赛道。内容创作平台既需要提供热门、高质量素材和模版，包括更多影视级特效，提升内容生产效率，还需要支持各种热门内容格式编辑，包括 3D 内容格式。

2）智能内容生成/生产

面向元宇宙的内容数据结构复杂，数据量大，普通工具基于人工难以处理，利用 AI 技术自动生成内容的生产方式 AIGC 需要进一步发挥优势，如用于 3D 内容、可交互数字人的生产等，需要提升元宇宙时代数字内容的生产效率。

4. 实时及用户大规模在线

元宇宙的场景中大规模的用户可同时在线并进行实时互动，这对音视频的处理、通信、呈现等实时性、兼容性有较高的要求。当前实时音视频行业 400ms 左右端到端延时均值，难以满足元宇宙超强沉浸式体验的需求。在超高清传输方面，8K+超高清视频内容，对实时音视频通信传输、编解码、渲染技术能力都带来了更高的挑战。在终端设备兼容性方面，多类型 AR、VR 设备终端在元宇宙场景中的使用，同样给实时音视频通信的终端设备兼容性能力带来更多挑战。

当前实时音视频行业全互动能力的人数通常支持 30-50 人，难以满足元宇宙多维互动、全互动的社交场景，用户大规模在线是对算力和网络的双重考验，两者深度融合和智能化势在必行。类比电气时代的电网，元宇宙时代需要构筑一套算力网络，以算为中心、网为根基， 网、云、数、智、安、边、端、链等深度融合、为元宇宙时代的音视频相关业务提供一体化服务，达成“网络无所不达、算力无所不在、智能无所不及”的愿景。

3.元宇宙时代超高清音视频技术体系

3.1 概览

音视频技术是元宇宙产业应用的基石，其发展需要在单点技术的发展基础上，融合多种关联技术，形成针对横向技术栈的端到端解决方案，实现元宇宙时代音视频技术的有效连接和共同发展。如图 6 所示，关键技术簇主要包括内容生成、内容编码、网络传输、内容解码、交互与呈现、用户体验及版权保护。基于元宇宙时代的技术新需求，研究探讨音视频关键技术簇，对于提升元宇宙的视听体验，具有重要的技术支撑作用，本章后续内容将针对上述关键技术簇展开探讨。

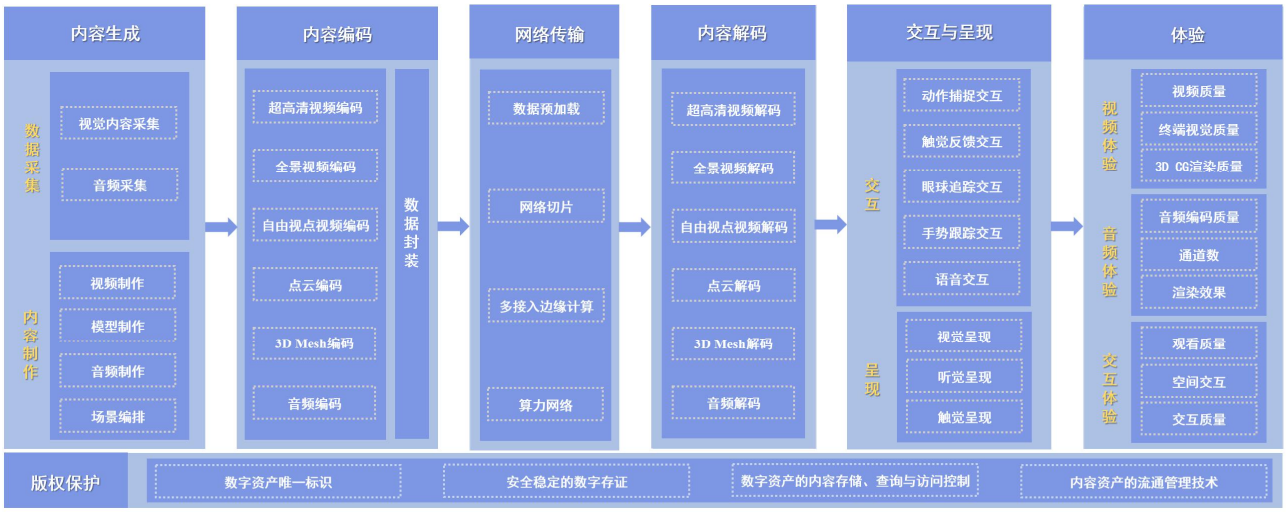

图 6 元宇宙时代超高清音视频技术全景图

3.2 内容生成

随着 4K/8K 相机传感器芯片的推广、AVS3 等编解码格式的升级、GPU 算力的突破、5G 带宽的提升，元宇宙时代的超高清音视频内容生成从采集、驱动、渲染的商业方案已经逐步成形，并在市场推广落地。新时代的内容制作，数据量要比传统的视频制作更加庞大，制作方式也更加复杂，耗时也更长。这不仅要求视频的分辨率、帧率、比特率、色阶满足超高清视频的制作规范，而且视频制作中会添加三维信息实现制作者与观众间的互动；声音方面会以 5.1、7.1、全景声等方式，提供声音的三维空间信息，给观众带来更强的沉浸感。

1. 数据采集

在元宇宙中，我们采用数字技术构建的数据空间来映射现实世界，构建元素包括空间、物品、人物等实体及声音。数据采集作为实体信息数字化的第一步，分为视觉信息和音频信息两个维度，采集实体的几何信息、颜色信息、材质信息、音频信息等数据。

1）视频内容采集

元宇宙通常的可采集方式有纯相机阵列扫描、结构光扫描、激光扫描、纯双目相机扫描、倾斜摄影等。在实际应用中，各大厂商也结合各类扫描方式的优点，致力研发以多模态采集融合、大数据辅助的方式，来优化采集流程、提高点云采集精度、材质贴图的清晰度。目前主流的研究方向是光场扫描技术和多模态遥感影像技术，前者主要应用于数字人扫描，后者主要应用于地形扫描。光场扫描技术结合了人工智能、计算摄影学、图形学等技术，联合利用结构光扫描空间信息，RGB 相机提供的超高清贴图纹理，辅助灯光集群仿真现实世界的光照信息，来提供最真实的三维模型数据采集。多模态遥感影像技术结合光学、红外、激光雷达等多种采集模式，辅助地形信息和 GPS 信息完成场景特征测量和数据采集。

• 纯相机阵列扫描

纯相机阵列扫描主要有两个方向的应用，分别是自由视点视频和三维建模。自由视点视频主要应用在电视转播和 VR 视频中，它的工作原理是利用多机位或环绕机位向用户提供环绕整个场地的任意角度视频信息，供用户以旋转切换的方式找到自己希望的视角和最佳追随观看位置。采集时使用高清相机阵列直接获取三维场景的纹理图，并通过主动深度传感（如 TOF 相机采集）和被动测距传感（算法计算）获取深度图。该模式能够提供 360 度的观看数据，但存在同步精度要求高，硬件成本高，存储数据量大的问题。三维建模方向主要针对柔体（衣服、人体）的数据采集，为人体建模提供精准的三维信息，它工作原理是采用在环形立柱群或球形结构体上，部署 360 度无死角的机位，来完成人体全方位的照片数据采集，再通过专业软件可以将采集的照片数据自动对齐、生成点云、添加纹理。该模式扫描速度快、同步精度能高达到毫秒级别，但该模式硬件构成复杂，成本高，并且不适用于镂空、反光等物体的扫描。

• 结构光扫描

结构光扫描主要应用在汽车、模具、文物、人脸扫描等场合，它的工作原理是将光源经过投射系统将光栅条纹投射到被测物体上，经过被测物体形面调制形成测量条纹，由相机采集测量条纹图像，进行解码和相位计算，最后利用外极线约束准则和立体视觉技术获得测量曲面的三维数据。该模式采集的信息分辨率较高，工作受环境光影响小，功耗低，但该模式硬件构成复杂，成本较高。

• 激光扫描

激光扫描主要应用于基础建设与空间三维测量等测绘领域，它的工作原理是用反射棱镜引导激光以均匀速度扫描物体的表面，同时接收物体表面反射的信号进行测距，用算法计算出深度数据。该模式扫描速度快、精度高、操作简单，但该模式对材质有限制，如果是黑色或透明的物体，需要在表面喷粉或者粘贴标记点。

• 纯双目相机

纯双目相机主要应用在 3D 电视转播、3D 立体 180°VR 视频等领域，它的工作原理是基于双目视差，利用成像设备从不同位置获取被测物体的两幅图像，通过计算图像对应点之间的位置偏差，获取物体三维信息。该模式采集流程相对成熟，硬件成本低，结构简单，可以在室内外使用，但该模式在抖动剧烈的情况下，存在双目相机画面缝合难的问题，容易造成拼接处的撕裂抖动等现象。

• 倾斜摄影

倾斜摄影主要应用在场景建模领域，它的工作原理是通过在同一飞行平台上搭载多台传感器，从一个垂直和四个倾斜，共五个角度同步采集影像，来获取场景的顶面和侧面的高分辨率纹理和点云信息。该模式成本低、数据量小、自带纹理信息、生产效率高，但该模式容易因为遮挡引起扫描漏洞，需要手工缝补。

2）音频内容采集

随着元宇宙场景多元化发展，传统单声道或立体声的音频格式已难以满足人类听觉系统的需求。物理空间对于听觉感知的影响并不局限于声源位置的感知，建立全景声这一具有沉浸式听觉感受的音频格式，探索更广泛的声音采集方式成为元宇宙音频的重要发展方向。元宇宙这一领域的快速发展正在重新建立声音和空间的关联，我们可以通过声音采集，即通过声学设备录制构建虚拟现实或整合增强现实中的声音。音频内容采集主要分为对象音频信号采集和环境音频信号采集。

传统录音方式在元宇宙音频采集中具有一定局限性，只适合于进行对象音源声音录制，即将某个声音对象作为单个点声源，进行声音录制，通常以单声道的声音文件呈现。元宇宙场景下的音频采集，更加注重声场信息拾取，当前主要采用 Ambisonics 拾音方式和人工头方式拾音方式。Ambisonics 拾音制式的优势是声场更加均衡，可直接进行后续信号处理和转换，主要分为一阶 Ambisonics（FOA）录音与高阶 Ambisonics (HOA)。目前，声场采集中，FOA 传声器发展较为成熟，图中最左所示为 FOA 传声器，它能够将声场还原扩展到较大区域，且所有方向等同无“焦点”声音存在，但 FOA 精度较低，不能较好表达声场信息；但 HOA 传声器，具有更高的空间准确性和解析度，将成为未来音频采集的主要发展方向。人工头录音方式包含了真实的头部相关函数，直接模拟出了人头及人耳对声音的干涉、衍射及反射效果，可以提供更高的真实性和沉浸式体验。

图 7　（从左向右依次为）FOA 传声器、HOA 传声器、人工头传声器

2. 内容制作

1）视频制作

元宇宙时代的视频具有三维立体和实时交互的特性，可以让用户感受到视觉冲击带来的沉浸式体验。视频制作将采集到的实体和实景的三维数据信息经过一系列处理制作成不同类型的视频内容，以满足不同元宇宙应用场景，主要包括全景视频、自由视点视频、点云视频、光场视频等，其中，自由视点视频和点云视频都是体积视频的表现形式之一。

• 全景视频制作

目前主流的全景视频分为 360 全景视频和 180 全景视频，两者的区别在于 360 全景视频可以看到 360°的全景画面，而 180 全景视频只能看到眼前的 180°视野。基于双目视差原理，180 全景视频模式可自带 3D 呈

现效果。全景视频的后期制作一般经过如下步骤：多机位的缝合拼接、色彩一致性的匹配、包装特效和转场应用。

在全景视频内容制作中引入高动态范围（HDR）技术使能 VR 超高清视频拥有更广的色彩容积和更高的动态范围，为图像保留更多细节。为了支持 HDR 标准的效果呈现，对 XR 终端的近场显示亮度要求要达到 400nit 以上，色域要达到 80% DCI-P 以上，目前主流的 VR 一体机的入眼实际亮度还达不到这个要求（以 80-40nit 为主），只是 Meta 公司在 2022 年中发布了对应的原型机。需要在 UWA 发布的 HDR VIVID 的 2D 视频技术标准之上，根据主流 VR、AR 眼镜的技术参数发展，增加和完善适配 VR 内容的 HDR 内容编码、格式、终端视频解码算法与显示的技术要求。

• 自由视点视频制作

自由视点视频本质是一种连续的三维数据信息的存储和重建。将纹理信息，深度信息，和角度信息保存为位图，以画面的形式进行编码和存储。在需要重建视角信息画面时，通过角度信息找到对应的纹理信息和深度信息，并进行纹理信息的 3D 重建，从而得到任意视角的画面信息。自由视点视频制作包括自由视点直播制作、 自由视点点播制作和精彩瞬间生成三种，以媒体流方式进行自由视点内容在线/离线制作。点播的制作，可以对转存的自由视点视频文件进行二次编辑等后期制作。精彩瞬间生成，选择某个机位的某个精彩时刻，以及所经历机位的运动轨迹，制作生成一段该精彩时刻的多视角连续视频片段。由于自由视点视频是通过多机位采集音视频数据制作而成，则如何保证所有画面帧级别对齐，以及多路音视频达到帧级同步是需要解决的难点。

• 点云视频制作

点云视频由 3D 数据组成的点云序列表示物体表面或场景大量的密集/稀疏点的三维坐标、纹理等信息。点云视频制作使用采集后的纹理信息和深度信息进点云序列存储（即点云格式），在需要重建视角画面信息时，直接使用 3D 数据信息进行对应角度的画面投影渲染。对于大尺寸物体或大范围场景，要经过多次多角度扫描，则需要解决点云去噪、简化、配准等处理中存在的问题。

• 光场视频制作

光场图像是通过密集的相机采集阵列或者光场相机所采集到的空间非常密集的多视点表达，采集的光场图像拥有远超二维图像的结构信息，实际编码和存储的信息是纹理信息和角度信息。光场技术能够提供更加逼

真的渲染场景和加光效果，由于光场相机拍摄的图像质量不高，因而光场图像畸变校正，光场图像的重聚焦，光场图像的增强是研究的技术方向。

2）模型制作

模型制作是真实世界中物体在元宇宙世界里的三维还原，主要包括角色、道具与场景三大部分，传统计算机动画（CG）流程中，模型制作由美术结合特定时空及风格设定原画，再由模型师进行三维制作还原，最终配合材质灯光渲染输出呈现。

随着视觉技术与计算机技术的进步，各种三维扫描技术如前文提到的 TOF、双目立体视觉、激光扫描、多视点矩阵扫描、摄影测量等手段被广泛应用于模型建模制作，视觉三维成像能有效解决物体 3D 信息获取，但是存在以下问题，一是扫描信息丢失，基于视觉的扫描技术以可视图像作为还原依据，受限于拍摄装备视点及物体本身遮挡关系存在信息丢失或误判，带来噪声、孔洞等固有缺陷，需后期人工干预处理；二是光谱信息丢失，受限于扫描环境与设备局限，扫描系统通常只能完成特定光照条件下的单一光谱信息采集，难以还原扫描物体尤其是人体在自然环境下的丰富细节，如高光、材质等细节信息。

目前主流发展趋势是构建更多视点、更高分辨率的扫描系统，辅助动态光谱，实现三维模型和材质、高光、法线、置换等系列贴图同步输出。同时，随着人工智能技术的发展，单帧图像三维重建方法也被广泛关注，此方法需要大量的先验真值数据库作为训练基础；构建多模态扫描矩阵是当前建模技术发展的关键，一方面能加速模型建模，另外一方面为 AI 算法提供更多有效的先验数据；AI 算法是建模技术的未来，通过人工智能降低三维建模对重资产、重流程、重人工的依赖，提升建模效率与品质。

- 3D 建模

目前应用较多的是人物建模，也是 3D 建模的主要发展方向，涉及材质、形状 运动学、毛发、布料等方面的建模。传统人物建模，往往采用手工建模的方式，建模质量高，可靠性高，广泛应用于游戏、电影、泛娱乐行业，也是当前技术水平下主流的建模方式。随着可微渲染、AI 生成技术的发展，人物建模方式正从传统的手工建模逐渐向算法自动建模过渡，并在很多任务上已达到较高水平。

虚拟人物形象是元宇宙的核心参与角色，对其进行高逼真建模一直是行业内的关注的核心技术难题。虚拟数字人逼真程度主要受两个因素影响，几何形状相似度和皮肤纹理逼真度。由于人脸皮肤的特殊物理特性，

传统的基于线性的人脸模型拟合方法，无法对面部细节，如皱纹、斑点、痣、肤色等，进行高逼真重建，且重建效果易受到遮挡、阴影、化妆等影响。目前，业内主要采用光场技术结合相机阵列采集人脸材质，人脸形状往往通过点云采集并结合点云配准实现。近几年，AI 建模技术展露头角，结合光场系统采集的高逼真纹理数据，可实现人脸高逼真重建。

人体建模主要致力于人体运动学、人体形状、肌肉变形建模。在材质方面，主要还是通过手工建模方式实现。人体建模较为著名的是马普研究所开发的 SMPL 模型，通过人体动力学建模，结合蒙皮技术，模拟人体运动、形状变形规律，其提出的人体骨骼拓扑结构为众多建模软件采纳并作为标准模板。在实际应用中，人体建模往往由人工实现。由于当前关注度较高的是人脸，因而对人体建模的迫切需求还未在应用中体现。

其他常见的 3D 建模有毛发建模，布料建模等等。近几年，毛发、布料建模技术的研究热度逐渐升高，技术路线与人脸材质建模类似，即通过光场系统/相机阵列采集逼真纹理，并重建 3D 点云，通过“AI+逆渲染”技术进行建模。

- 模型驱动

模型驱动以人物模型驱动应用较多，可以分为表情驱动、人体驱动、手势驱动等。

其中表情驱动可分为表情捕捉、语音驱动表情、文本驱动表情等。表情捕捉以图像、视频、点云或 Mesh 作为输入，实时重建 3D 人脸表情，这类技术主要用于直播、会议等场景。基于图像的表情捕捉，研究成果及应用较多，但从 2D 图像或视频恢复人脸三维结构往往缺少局部细节，即细微表情表现力不足。基于点云(或图像融合深度信息)的人脸表情捕捉，重建精度较高，往往需要深度相机或双目相机系统，应用成本相对较高。整体上，表情捕捉技术相对成熟，存在的主要问题是对脸部局部运动的重建，在这方面国内研究与应用案例相对较少，与国外先进水平，如：Unity Ziva，存在较大差距。语音/文本驱动表情，是指通过算法将语音或文本输入信号映射为表情系数，并通过表情系数驱动人脸表情，通常人脸表情模型采用 3DMM 这类线性模型，由于语音/文本到表情系数的映射往往是“一对多”或“多对多”，因而，嘴型驱动效果往往难以达到真人说话时的表情，即嘴型与语音的匹配度较低，语音/文本驱动表情技术是“虚拟数字人”应用所需的一项关键技术，在算法精度提升方面，依然有很大的研究潜力。

人体驱动、手势驱动技术，通常分别称为人体动作捕捉技术以及手势识别，这里将其统称为动作捕捉。动作捕捉从驱动形式上可分为基于动捕设备(惯性捕捉、光学捕捉)的动作捕捉和基于算法的动作捕捉，后者的输入与表情捕捉情况类似，可以是图像、视频、点云。整体上讲，基于设备的动作捕捉技术相对成熟，实时性以及捕捉精度可以满足一般应用场景需求，对于大角度、遮挡等情况下动作重建具有较好的鲁棒性。基于 AI 算法的动作捕捉，近几年一直保持较高的研究热度，且已有相关的产品及解决方案，其主要的难点在于大角度以及遮挡情况下如何对动作进行重建，以及对肌肉扭曲、拉伸等运动的模拟。

• 内容渲染

面向元宇宙时代更加复杂的场景制作及实时交互的特性，采用实时渲染的技术为主。相对于传统的非实时渲染，实时渲染所见即所得，能够更加直观的看到输出效果，方便制作人员进行调整。实时渲染技术主要在虚拟引擎厂家和显卡厂家搭建的平台上研发迭代，目前在自由视角、后期三维场景制作或者 XR 演播室等领域得到了广泛应用。

AI 超分辨率：元宇宙中超高清音视频需要处理巨量的数据信息，从而造成制作硬件和后期渲染制作周期的成本大幅提升。AI 超分辨率技术，是基于深度学习的方法，利用帧内和帧间的信息，在观众终端上呈现同样帧率、分辨率的视频画面前提下，通过让显卡渲染更少的像素来降低资源消耗。该项技术可以令制作方不再要求以全超高清的链路完成制作，转而采用高清链路完成前期制作，再通过 AI 超分辨率技术完成观众终端的超高清画面呈现。在元宇宙超高清没有完全普及前，这项技术是一种很好的过渡方式。

实时光线追踪：光线的仿真一直是元宇宙的一个重大研究课题，该项技术的突破，会给人物和三维场景的真实感带来飞跃式的提升。实时光线跟踪，是在构建的三维场景中，利用算法来模拟光线的路径，精确反映出物理世界中的阴影、反射、折射和全局光照等效果。目前该项技术对于硬件要求很高，如何将该技术应用到手机等消费级平台上，是各个厂家追逐的热点。

实时三维云渲染：元宇宙的超高清音视频渲染，需要占用大量的 CPU、GPU 资源，对终端用户的设备要求较高。实时云渲染技术，是以 5G+云的方式，在计算机集群上完成渲染的工作，再将处理的结果实时返回给终端用户，整个处理过程延时非常低。实时云渲染技术，一方面能够降低终端用户的部署成本，另一方面能够更好的实现协同制作的应用。

多屏实时渲染：在 XR 演播室系统中，三维场景要同时投射到地屏、背景屏和侧屏等多个屏幕上，最后再通过机内渲染或者机外渲染，实现虚实融合的效果。如何保证多屏能够实时同步，摄像机拍摄的画面、屏幕上呈现的画面和实际建模的画面颜色亮度如何能够保持一致，是多屏实时渲染重点解决的问题。

3）音频制作

元宇宙场景对音频制作技术提出新的要求，声音制作不再受扬声器或听者布局的限制，对基于声床、对象、场景三种信号进行空间信息解码和信号与处理，结合头相关变换函数（HRTF）、混响、FOA/HOA 等技术，进行渲染处理，获得具有 6DOF 空间特性的音频，输出符合现实人耳感知的声音信号，再结合真实场景的物理空间特性，声源指向性、衰减、衍射、反射等声学特性以及个性化 HRTF 等技术，实现更具有临场感和互动性的空间音频效果。

元宇宙场景下的音频制作，对于涉及的制作工具也提出更高要求，传统音频制作主要是依赖于 Protools、Nuendo 等音频制作站，而元宇宙时代的沉浸式音频制作，向着轻量化和便利化发展，沉浸式音频制作工具成为未来生产力的趋势。沉浸式音频制作工具应包含以下功能：

• 3D 声源建模

3D 声源建模是基于真实发声物体的体积大小、形状、不同方向声音衰减强弱的物理特性，通过理论分析和数值建模，结合远近场声学模型、去相关、实时计算等技术，还原具有体积的发声物体所具有的六自由度特性，使听众能通过音频感受到发声源的体积大小、形状、旋转等变化。

• 3D 声场建模

3D 声场建模是还原真实三维空间声场的重要技术手段，结合静态和动态场景对声音存在的空间环境建模，结合不同的材料吸声系数，利用基于几何和波形的声学建模方法，采用人工混响、卷积混响等技术获取室内冲击响应，最终呈现不同空间及声源位置的差异变化。目前，3D 声场建模技术向实现声笼、声障、动态开关门等场景演进，准确还原声波遇到障碍物时，产生的阻挡、投射听感，从而提升虚拟空间中声音的场景感和真实感。

• HRTF 双耳渲染

HRTF 双耳渲染是用于音频对象在耳机系统中重放，利用基于人头特性 HRTF 头相关传递函数重建音频的虚拟声像，结合头部追踪等技术，实现声源在虚拟空间的自由移动、旋转。

• 扬声器阵列渲染

扬声器阵列渲染是用于多扬声器系统中音频虚拟声像的重构。传统的渲染技术主要是通过多通道技术(channel-based)来现实，为了更好满足元宇宙环境中复杂的声学场景，基于对象(object-based) 和基于场景(scene-based)的渲染技术结合虚拟场景得到了加速发展，随着波场合成 (WFS) 、FOA/HOA、矢量平移(VBAP)等技术在不同场景的融合发展，沉浸空间音频渲染不再局限于固定的扬声器布局和数量面，具有更高的自由度和灵活性以及更沉浸的真实体验。

4）场景编排

元宇宙基于三维对象和空间场景等元素构建，对应的音视频数据和三维模型数据的类型多样、数据量大，并且在场景地图中分布离散，需要通用的三维模型格式对元宇宙场景进行统一、完整的场景描述，满足对融合不同类型数据的三维场景的编排、重用、共享与显示。

• 通用场景格式

元宇宙时代中不同软件平台制作内容的格式不同，不同软/硬件平台之间的互通和内容的共享，可通过统一的描述格式描述空间场景、空间中的对象模型及交互事件。国内，中国通信标准化协会（CCSA）正在研究面向电商、社交、游戏等领域的通用 3D 数据格式，支持跨硬件设备种类、跨软件平台的 3D 互动场景，同时满足适应元宇宙大场景的要求。国外，Khronos 发了基于场景的媒体的图像库传输格式（glTF），glTF 3D 统一数据格式是为了三维模型的网络传输而设计的文件格式，以 3D 场景为根节点组织数据，通过网格、材质、相机、动画、皮肤、纹理等节点的定义和关联，实现 3D 场景的描述。USD 是皮克斯开发的一种易于扩展的开源 3D 场景说明和文件格式，通过统一 3D 场景的格式，从而加速 3D 几何图形和着色的读取、写入、编辑、快速预览。考虑到元宇宙业务场景的复杂性、交互性等，通用的场景格式需要支持 2D/3D 混合的音视频数据、数字人、多模交互、通用渲染等信息描述，以满足沉浸的实时的、跨平台的交互。

• 内容编排

元宇宙时代下的音视频内容多样化，将素材、编排、预览等全部云化操作，可有效的提升制作效率，节省成本，以及共享编排内容。通过平台编排工具进行快速地定制化编排，支持全景视频、3D 模型、3D 场景等各种格式素材的编排，如 3D 空间场景生成、2D 全景视频与 3D 模型叠加，并可对热点、场景事件（如事件跳转）等进行编排。针对不同元宇宙的音视频业务，未来将进一步支持场景素材的多格式、场景的复杂交互设计、以及对素材内容的输入和管理、编排内容导出和共享，实现内容的自动化编排 。

3.3 编码及网络传输

1. 整体技术框架

音视频编码及网络传输的整体技术框架由音视频数据编码、数据封装、网络传输和解码等部分组成，如图 8 所示。

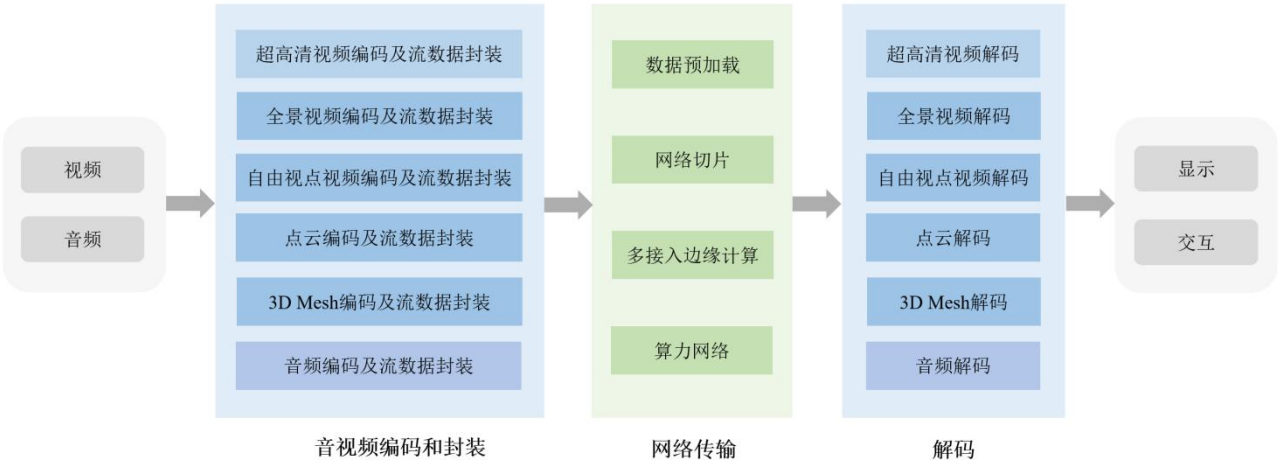

图 8 音视频编码及网络传输的整体技术框架

在音视频编码及网络传输的整体技术框架中，首先是内容源的获取，包含了元宇宙时代沉浸视听媒体信号，其中，视觉方面主要包含 4K/8K 超高清视频、全景视频、自由视点视频、点云、3D 网格（3D Mesh） 等；听觉方面包含 3D 音频、6DOF 音频等。然后，进行视听数据的编码封装，并分发传输至用户端。最后，用户终端解码后进行显示、观看。

2. 数据编码

1）视频编码

• 超高清视频编码

超高清视频将支持高帧率（HFR，如 100fps、120fps）、高动态范围（HDR）、广色域（WCG）。为了应对超高清视频的超高数据量和超大带宽需求，超高清视频的编码必须追求更高的压缩效率，从而实现低时延编码、实时编解码，以支持低时延的应用。基于此，国内外标准工作组织，如 ISO/IEC 运动图像专家组(MPEG) 和国内数字音视频编解码技术标准工作组（简称 AVS 工作组）分别提出了新一代视频编解码标准，开放媒体联盟 AOM 阵营也提出了面向互联网流媒体的开发编码标准 AV1。

MPEG 最新一代的编码标准 VVC 在码率节省 50%的情况下可以保持与上一代标准相似的主观质量。除压缩效率进一步提升之外，VVC 的通用性也可以支持如超高分辨率（4K/8K）、高动态范围（HDR）、屏幕内容编码、360 度沉浸式视频编码等多种新兴的视频内容与应用。AVS 工作组也发布了面向 8K 超高清视频(UHD)电视广播和 VR 等新兴应用场景的音视频信源编码标准 AVS3。最新一代 AVS3 标准提出了更灵活的扩展四叉树划分方式、更复杂的帧内预测模式，并在帧间预测环节围绕预测结构，预测单元粒度和预测模式等方面进行了优化提升。

另一方面，为了支持元宇宙时代更多样化的超高清视频应用场景，比如零售、工业场景等，当前业界也开始探索针对机器视觉的新兴编码标准，比如 ITU 与 ISO/IEC 正在开发的 VCM 技术，通过压缩视频或者是前序任务产生的特征图产生码流，以供机器分析直接使用。最终支持智能视频监控、智慧城市、智慧交通等多种机器视觉相关的任务。

• 全景视频编码

元宇宙时代，全景视频是常见的沉浸媒体格式之一。当下，由于全景视频具有高分辨率、高帧率、数据量大的特点，实现全景视频的低延迟传输需要较高的传输带宽，极大地阻碍了 VR 相关业务的发展。因此，高效的 VR 视频编码技术至关重要。MPEG 从 2015 年起开始制定相关标准，提出了全向媒体格式（OMAF），以表示 360 度媒体内容，包括 3DOF VR 视频、图像、音频、文本等。我国 AVS 标准工作组和 IEEE 1857 工作组联合制定的 VR 视频编码标准于 2018 年正式颁布。

目前，VR 视频编码可大致分为两类：传统编码方法和基于 FOV 的编码方法，前者通常将球面内容投影到二维平面上，再采用视频编码的方法进行编码。后者是指将输入的视频帧划分为相同大小的矩形，每个矩形编码为一个片（Tile）以便并行处理，同时可以 Tile 为单位拆分码流进行切片分发。

未来，高效的球面-平面的投影、高效的片划分、提高并行处理速度、实现用户姿态变化快速响应及视角自适应动态切换等将是未来全景视频编码技术的重要发展方向。

• 自由视点视频编码

自由视点视频实现了 6DOF 观看体验，相较于其他沉浸媒体，其具有低成本制作、应用场景多等优势。面向自由视点视频，MPEG 制定了沉浸式视频（MIV）标准，旨在提供沉浸式、6DOF 立体视觉场景的高效编码，可用于虚拟现实（VR)、增强现实（AR）和混合现实（MR)。MPEG 计划基于视觉体积视频编码 V3C 的框架进行技术集成和扩展。国内 IEEE 1857 和 AVS 联合成立的虚拟现实标准制定联合工作组 VRU 于 2019 年启动了自由视点视频标准制定工作，该标准采用超高清编码基础设施进行自由视点视频压缩，在编码端将多视点图和多深度图直接进行拼接，加上辅助的元数据进行 2D 视频编码，在终端将基于深度图的处理合成虚拟视点的合成。

MPEG MIV 编码器的输入是多组视频，由一组无序的具有任意位姿（源视角）的真实或虚拟摄像机捕获。来自每个源视角的视频集包含了投影几何信息（深度以及可选的占用图）和属性信息（例如纹理、表面法线、材质贴图、反射率、透明度等）。此外，还提供了每个源视角的元数据，例如相机内外参、投影格式、投影平面尺寸和源视频的位深度（几何和属性）。MIV 编码器将基于几何和属性信息生成属性图集和几何图集，以及基于每个源视角的元数据生成描述图集的元数据。通过生成图集将减少视图间冗余，从而降低编码比特率，同时保持呈现给观众的内容质量，以实现实时的沉浸式视频服务。生成的属性和几何图集使用 2D 视频编码器编码为视频比特流，而元数据使用 MIV 标准编码。

更高效的补丁块生成、深度图生成方法、兼容更多的视频编码器、基于率失真优化的联合比特分配、面向全景内容的非同构视图选择、支持更宽的深度动态范围并实现占用误差校正、降低视点切换的延迟等将是未来自由视点视频编码的技术的发展方向。

• 点云编码

元宇宙时代下新一代视觉沉浸媒体 3D 点云由一系列点的 3D 坐标和其对应的属性信息（颜色、反射强度、法向量等）组成。MPEG 基于点云编码的需求，相继发布了基于视频的点云编码（V-PCC）和基于几何的点云编码（G-PCC）。AVS 工作组也成立了点云技术工作组，发布了点云编码参考软件（PCRM），其是在 3D 空间对点云进行编码处理。Google 也发布了开源库 Draco，用于压缩和解压缩 3D 几何网格和点云。

目前，点云编码主要分为两种技术路线：基于视频的点云编码方法和基于几何的点云编码方法。基于视频的点云编码基于视觉体积视频编码 V3C 的框架，其方法主要涉及块（Patch）划分重组、几何和纹理图像的生成、填充及编码、辅助划分信息以及占位图的编码等新技术。基于几何的点云编码方法主要涉及八叉树表征、属性预处理、属性变换、变换/预测、属性量化、属性熵编码等新技术。

未来，点云编码技术发展将围绕下列方面展开：更高效的几何表征方法及属性预测技术、更灵活的编码速度配置，支持多种延迟场景，比如离线转码、云游戏、视频直播、视频会议等，兼容各终端设备，实现场景自适应，提高编码速度（帧间并行加速、预分析和后处理加速），智能码率控制等。

• 动态 3D 网格编码

3D 网格已成为视觉沉浸媒体主要的数据格式之一。一个 3D 网格由以下部分组成：拓扑信息、几何信息、映射信息、顶点属性、属性映射。上述五部分中，若任一部分包含时变，即为动态网格。相较于静态网格，一个动态网格序列的数据量更为庞大，因为它包含大量随时间变化的信息。MPEG 正在计划开发一种新的 3D 网格压缩标准，用以压缩具有时变拓扑信息和可选时变属性映射的动态网格。该标准面向各类应用场景，如实时沉浸式通信、自由视点视频、AR 和 VR 等。MPEG 计划基于视觉体积视频编码（V3C）的框架进行技术集成和扩展。目前，国内标准工作组暂未开设相关专题组。

未来，3D 网格编码技术的发展特点大致如下：支持静态和动态网格编码、支持有损和无损压缩、支持随机访问、低延迟、具有容错能力、支持并行编码和解。

2） 音频编码

元宇宙时代的沉浸式音频编码，瞄准基于强互动技术的沉浸式体验和虚实融合场景的应用需求，从个性化音频制作到基于场景的个性化渲染回放，除要兼顾传统的音质和传输外，对交互、渲染、场景体验等提出了更高要求，对音频元素的建模和实时表达，即音频元数据编解码传输提出了更高要求。国际上，3D 音频 MPEG-H

和沉浸式音频 MPEG-I 是 MPEG 在研的两个标准系列。其中，MPEG-H 3D Audio 可以提供沉浸式和个性化音频服务，而 MPEG-I 是为虚拟和增强现实应用程序开发的全新标准，旨在创造自然、逼真的 VR 体验。相比 MPEG-H，MPEG-I 能够提供更优秀的互动性和沉浸感，并将支持 6DOF，用户可以在虚拟空间内移动，并与虚拟空间物品进行交互。 3GPP 沉浸式语音及音频服务（IVAS）已于 2017 年立项，从版本 15 开始已经演进到版本 18，其中包括 VR 流媒体服务、沉浸式语音及音频编码、提升 VR 用户体验、边缘计算、场景探索、触觉和媒体服务，SA4 沉浸式多媒体类型和内容格式，XR 相关业务，QoS、QoE 指标等。国内，AVS 工作组从 2016 年起，开始对沉浸式音频和互动式音频开展研讨和技术征集，与国际主流标准组织的技术和规划处于并跑阶段，AVS3 音频和虚拟现实音频标准在同步制定中。

元宇宙时代的音频技术创新可以体现在以下三个方面：

第一，结合大数据和深度学习等人工智能技术，能够显著提高音频编码的压缩码率，降低音频编码的复杂度。一方面，深度学习和神经网络技术可以应用于特征变换、窗型判断、心理声学、熵编码等编码模块中，与传统音频编码方式相结合，作为混合人工智能的音频编码，提高音频编码效率；另一方面，也可以完全使用大数据和深度学习技术，完全替代传统音频编码方法，做全智能的音频压缩编码，显著提升音频编码质量和编码效率。

第二，提供沉浸式和个性化音频，增加用户互动性体验。沉浸式音频可以实现 6DOF 的声音感受，即在任意空间声场中任意活动，得到该声场中的所有音频对象信息，并与任意音频对象进行任意交互，获得与现实相同的沉浸式音频体验。可支持多种音频元数据的编码和传输，除需要兼容国际标准 ITU-R BS.2076 音频定义模型中涉及到的音床、对象、Ambisonics、HRTF 等元数据信息外，还需要可扩展的音频元数据架构，可以支持 6DOF 的多元音频元数据扩展，如使用多种传感器（如陀螺仪等）获得音源与用户的相对位置、静态或动态声源的辐射和在不同空间场的混响参数等。

第三，灵活适配各类回放场景，对应扬声器声场和耳机回放，提供高质量、高保真、定位准确、低延时的音频技术服务。运用 Ambisonics、HRTF、混响等多种空间渲染技术，实现不同回放环境的沉浸式音频体验。

同时，降低音频编码、渲染的计算复杂度，支持自适应可变速率，降低音频系统延时，探索真实声场的全息重现技术和音频、视频和传感等多模态数据的联合编码与协同表达模式，是元宇宙时代沉浸式音频实现的主要研究方向。

3. 数据封装

元宇宙时代音视频数据的格式多样，数据封装的媒体类型涉及二维场景和三维场景，包括传统超高清视频、自由视点视频、点云数据、音频及元数据等媒体数据。

对于二维场景的音视频，基于不同的应用场景采用相应的封装格式，如 TS，MP4，MOV，FLV，MXF，MKV 等。其中，TS 格式扩展性比较友好，可以支持多种流媒体协议，比如 HLS，SRT，UDP，MPEG-DASH，同时支持多种音视频编码，视频如 H265/AVS2/AVS3/AV1/VVC 等，音频 Dolby/DTS/AVS2/AVS3 等。MP4 格式主要应用在 MPEG-DASH，CMAF 协议中，在直播场景中采用 MP4 格式中的 fragment-MP4 的封装格式，既支持丰富的音视频编码类型，也支持 Low-Latency HLS，CMAF 等超低时延的流媒体协议，从而解决存储、分发，编码等工作，降低成本，提高效率。针对制播域的 IP 化改造，主要采用美国电影电视工程师协会(SMPTE)发布的 SMPTE ST 2110 协议，是一套规范现场制作流程中不同 IP 实体流的承载、同步和描述的标准，用于现场制作、播放以及其他专业媒体应用的协议，把 4K/8K 的信号以 IP 协议输出，采用 MXF，MOV 等格式封装 JPEG-XS 编码视频，可达到极低的延迟和极低的复杂度，视觉无损的质量。

对于三维场景的音视频，MPEG、AVS 等组织已制定或正在制定对全景视频数据、自由视点视频和点云数据的文件封装存储，封装基于 ISO BMFF 格式进行扩展。由于音视频数据具有三维的视听效果，并满足 3DOF/6DOF 体验，其数据量巨大，为支持应用对音视频数据高效实时的访问、解码、渲染处理等，针对不同的场景和数据类型采用灵活的封装方式，如全景视频和沉浸式音频采用基于分片的封装方式，点云视频采用基于组件、组件分片、打包视频的封装方式。同时，为支持更丰富的业务场景，在文件封装中添加视窗信息、动态的视点信息、时间线信息、音频元数据信息等。

未来，音视频数据封装将基于不同的元宇宙场景及其编码数据格式进行封装格式的扩展，如 3D 视频/沉浸式音频/模型数据的封装，基于超低时延音视频自适应传输的封装、基于特定应用场景的元数据扩展。

4. 网络传输

1）网络技术

元宇宙时代视频业务从高清向超高清、沉浸式发展，时延要求视频流量越来越大，时延要求越来越苛刻，从点播的数秒，到直播的秒级，再到沉浸交互所需的毫秒级。千兆光网络、Wi-Fi6 和 5G、5G-A 网络的发展为元宇宙业务提供了坚实的网络支撑。由于元宇宙业务需求的差异性，采用 5G 切片技术定制元宇宙业务的低时延和大流量切片网络，为其分配独立的网络资源，并允许动态扩缩容网络资源，以及管理资源的生命周期。同时，5G 切片结合 MEC 边缘云，实现端到端低延时高可靠，支持关键业务。

算力网络作为联结需求侧及供给侧的桥梁，通过确定性网络使算力连接成网，通过云间/云边协同提供灵活、高效算力，实现存算资源优化，在提升业务体验的同时，极大降低成本。算力与网络深度融合的特性，算力网络各层注入视频能力，可构建面向视频业务的全场景算力网络，为元宇宙时代的视频业务创新提供坚实支撑。

面向视频业务算力网络是在基础算力网络上叠加了视频业务的算网逻辑。在资源层面，指基于基础算力进行视频处理逻辑的使能，在编排控制层面（算网大脑）增加面向视频业务层面的编排控制。元宇宙的视频业务要保证用户实时沉浸、无卡顿流畅的交互体验，因此面向视频业务的算力网络需要分层级获取算力，逐层处理，分级服务。

图 9 面向视频业务的算力网络架构

在基础算网设施层，需要在端、边、云构建算力网络泛在、异构的视频类算力，满足视频业务日益增长的算力需求，按需分层部署增强的层次化异构算力，如现场视频网关、边缘的视频业务算力及视频云，满足多场景算力需求。在视频算网层，集成编解码、渲染、分发和分析多种视频能力提供视频算力服务，使能应用开发和能力共享，同时增强 CDN 网络，实现 CDN/RTN/VSN 多网融合，构建统一低时延交互式视频网络。在编排调度层，需要提供就近、就算、就碳智能算力路由，实现视频应用的云端、边缘算力资源统一调度。

2）流媒体协议

元宇宙业务场景对流媒体传输的实时性和互动性提出了更高的要求，这就需要在传统的 RTMP、SRT、HLS 等基础上增加实时互动的支持。实时互动，指在远程条件下沟通、协作，可随时随地接入、实时地传递虚实融合的多维信息，身临其境的交互体验。实时互动作为下一代互联网基础设施，实现了从“在线”到“在场”的重要转变，将推动互联网向以“临场感”为主要特征的元宇宙方向的升级变革，当前几个主流的技术方向如下。

MPEG-DASH 是一项基于 HTTP 的动态自适应流传输技术，由 MPEG 在 2012 年推出。它不限制编码格式及内容，能够根据当前带宽容量、网络性能等情况自适应地实现不同码率之间的灵活切换，在为用户提供低卡顿体验的同时保证播放内容的质量。当前，MPEG-DASH 协议已成为全景视频的主要传输协议。

WebRTC（Web Real-Time Communications）是一项实时通讯技术，早期由 Google 开源，实现了基于网页的实时通讯能力，并于 2021 年被万维网联盟（W3C）和互联网工程任务组（IETF）采纳为官方标准。WebRTC 可以实现超低延时、低卡顿的实时通讯效果，但是对虚拟现实内容面向元宇宙新的媒体类型媒体传输和交互支持不足。WebRTC-NV（Next Version）是下一代 WebRTC，是当前 WebRTC1.0 之后的标准，意在支持当前 WebRTC API 不可能或很难实现的新用例，比如 VR。主要是从通道扩展性、模块成熟和完善性、采集扩展性、独立的标准等 4 方面能力提升。

QUIC 是一项基于 UDP 的低时延通用传输协议，由 Google 推出，它从可靠传输、安全机制、时延等方面对 UDP 协议进行了优化，通过加密、流量控制、拥塞控制等技术，实现了更灵活、更安全、低时延的传输。目前，多个浏览器已支持 QUIC，比如 Google Chrome 浏览器、Microsoft Edge、Firefox 等。同时，该协议已广泛应用于移动端直播、短视频、高速图片文件下载等业务场景。

综上所述，面向未来元宇宙沉浸式体验的需求，3D 视觉媒体信息的低时延高效传输是亟需解决的问题。因此，如何基于 3D 视觉信息的特点对传输协议进行优化以实现低时延传输将是传输协议进一步发展的方向。

3.4 交互与呈现

在元宇宙时代的音视频呈现中，交互是衡量用户沉浸感的一个重要指标。

1. 用户交互

当前业界支持的用户交互手段通常包括动作捕捉交互、触觉反馈交互、眼球追踪交互、手势跟踪交互以及语音交互。

1）动作捕捉交互

动作捕捉交互通过捕捉用户的动作作为用户输入的交互信号。其中动作捕捉可分为穿戴式和非穿戴式动作捕捉。穿戴式动作捕捉包括惯性捕捉、光学捕捉、激光捕捉等几种方式。

- 惯性捕捉：通过惯性导航传感器 AHRS(航姿参考系统)、IMU(惯性测量单元)测量用户的运动加速度、方位、倾斜角等特性；
- 光学捕捉：通过对目标上特定光点的监视和跟踪来完成运动捕捉的任务；
- 激光捕捉：利用定位光塔，对定位空间发射横竖两个方向扫射的激光，在被定位物体上放置多个激光感应接收器，通过计算两束光线到达定位物体的角度差，计算定位物体的坐标

非穿戴式捕捉即利用摄像头采集人体运动视频数据，通过 AI 算法计算关键点数据完成人体运动捕捉。

2）触觉反馈交互

触觉反馈交互主要通过用户利用按钮和震动进行交互反馈，常见的支持触觉反馈交互的手柄通常为两手分立的、6 个自由度空间跟踪的（3 个转动自由度 3 个平移自由度）、带按钮和震动反馈的手柄。

3）眼球追踪交互

眼球追踪交互，又称为注视点追踪，是利用传感器捕获、提取眼球特征信息对眼睛的运动情况进行测量并估算视线方向或眼睛注视点位置的技术。随着摄像技术、红外技术和计算机技术的发展，当前采用的追踪方法主要有巩膜一虹膜边缘法、瞳孔追踪方法、瞳孔一角膜反射法。

4）手势跟踪交互

手势跟踪交互通过手势的识别方法可以分为两种方式：使用光学跟踪的手势跟踪和使用传感器的手势跟踪。前者通过在一体化移动 VR 头显上直接集成光学手部跟踪设备进行手势跟踪；后者则通过可穿戴数据手套/指环集成的惯性传感器来跟踪用户的手指乃至整个手臂的运动。

5）语音交互

智能语音交互是基于语音识别、语音合成、自然语言处理等技术，使之实现“能听、会说、懂你”的智能人机交互功能，可以用于语音指令、实时记录、智能客服、内容播报、在线教育、自学习语言模型等多种元宇宙应用场景。听感自然的语言交流、个性化音色变化、多语种支持、高识别率实时的语音交互，是元宇宙未来的发展方向。

总体看，当前用户交互主要存在两方面的问题：一方面，体验者需要额外穿戴或手持设备才能完成对交互动作的感知，交互体验不好；另一方面，交互带来的场景变化需要有足够算力支撑高精度场景实时渲染，当前已经有一些端侧芯片渲染，但是算法难以达到预期效果。

未来解决问题的主要途径在于交互方式的演进以及渲染逻辑的更新，采用视觉加人工智能的方式实现人体行为动作识别并进行语义理解分析，无需穿戴即可实现正常交互，采用云渲染方式降低交互对端侧渲染硬件算力需求。

2. 媒体呈现

元宇宙时代下媒体呈现的方式有视觉、听觉、触觉等，通过将多种呈现方式的融合，为用户提供沉浸、真实的视听体验。

1）视觉呈现

随着元宇宙时代的到来，消费者对视觉的呈现要求越来越高。这里不仅是对底层显示技术 LCD 和 OLED 的挑战，也是对内容显示效果的挑战，并呈现出以下特点：

- 显示特性越来越高

元宇宙时代的到来，消费者对近眼显示需求不断提升，对 LCD、LED 等底层显示面板技术升级需求越来越迫切。以 Mini LED 背光 Fast -LCD，以及硅基 OLED 的诞生拉开帷幕。为了更好的呈现“元宇宙”视觉效果，完善 VR 产品近眼超清细腻画质需求。就需要进一步提升显示面板在高对比度、高刷新率、高亮度等方面的性能，同时辅以 HDR 功能。虽然硅基 OLED 采用单晶硅芯片基底，改善了纱窗效应，具有更高分辨率和对比度。但单体成本相对较高，因此为了高端 VR 产品的推广，硅基 OLED 要通过技术降低单体成本。除传统显示屏技术外，微显示屏技术也需要大突破。即满足用户的便携性，又能在眼前显示 60 寸以上超大屏幕感，将 3D 内容渲染到真实世界，给用户沉浸式体验。微显示屏配合光波导镜片未来将单颗像素尺寸逐渐变小，整体分辨率不断提升。

- 展现效果多维化、立体化

元宇宙时代的到来是对传统内容显示方式的一场变革，从原来传统的平面显示到立体呈现，从二维显示到三维显示。为了更好的提供“元宇宙”的视觉体验，提供更加接近真实世界的图像内容。计算全息三维显示的技术突破必不可少。我们需要解决计算全息图重建质量不足、波前调制器件和全息显示系统性能受限以及三维内容源匮乏等问题。提供高质量全息动态三维显示，低噪声全息图的获取、像质优化和畸变校正技术的开发以及三维内容源的构建，实现高质量、低噪声、无畸变、高刷新、真三维的动态全息三维显示，是计算全息显示发展的必由之路，也是元宇宙显示的必然要求。

2）听觉呈现

元宇宙时代的沉浸式音频在听觉呈现的角度上，主要强调在虚拟环境中模拟真实的音频体验。因此，从贴近真实世界的角度，元宇宙时代的听觉呈现主要体现在以下三个方面。

• 真实的空间音频渲染

基于声音传播的理论模型和人耳听觉的理论模型，当前业界采用的 Ambisonics、空间渲染等沉浸式音频渲染算法可以用于共同构造更逼真的用户听觉呈现效果。比如通过构建不同的声源模型，可以模拟出有向声源和体积声源；通过房间声学渲染可以提供真实的各类环境重现；通过分析场景的结构和几何障碍物进而构造声音透射和衍射的障碍物效果；通过分析用户或者场景的移动来构造多普勒效应。

• 六自由度音频呈现

在元宇宙的音视频消费场景中，用户自由度是十分重要的一个体验维度。因此，听觉呈现上的 6DOF 效果在音视频消费中也十分重要。基于人耳的听觉感受理论，结合用户在场景中的头部位置和朝向，即可通过调节音频的增益、构造声场等方法来模拟音频的定向、衰减、传播、遮挡和阻塞以及混响等效果，最终通过扬声器排布或者双耳渲染算法，为用户提供 6DOF 的听觉呈现效果。

• 支持可交互的音频呈现

对于元宇宙时代的音视频消费来说，可交互性也是一个重要的体验维度。听觉呈现的可交互特征主要体现在以下两个方面：同一场景内的多用户交互以及场景对象交互。

在音频呈现中，同一场景内不同的用户互相对话是社交场景下的必然需求。用户与用户之间的声音交互除了低时延的对话需求外，还支持根据用户位置进行相应的空间音频渲染呈现。

除了与用户交互外，音频呈现中还支持与场景内的对象进行交互，比如通过交互操作激发、关闭、调节场景内的某些声音对象，使用基于应用场景和场景内对象的音频特征建模和互动生成的音频元数据可以提供更好的交互体验。

3）触觉呈现

在元宇宙时代的超高清音视频消费中，媒体内容的呈现往往伴随着各种各样的可穿戴设备或者可交互设备。因此，元宇宙时代的沉浸式媒体在呈现方式上，除了传统的视觉和听觉方面的呈现外，还具备触觉呈现这种新的呈现方式。

触觉呈现依托于触觉反馈技术，触觉反馈技术通过硬件与软件结合的触觉反馈机制，允许用户通过他们的身体接收信息，提供一种嵌入式的身体感觉，传递关于用户正在使用的系统的关键信息。在当前业界，触觉反馈技术已经有了较为广泛的应用，通常分为振动触觉反馈系统和电触觉反馈系统。振动触觉反馈系统通过模拟物体的重量、压力、阻力等对用户进行基于振动的触觉反馈。电触觉反馈使用电脉冲向用户皮肤的神经末梢提供触觉刺激来模拟温度变化、压力变化、潮湿感等真实体验。

从当前业界的触觉反馈技术不难看出，从日常普遍应用的设备振动反馈，到细分领域的多样化触觉反馈，触觉反馈本身已经是一种用户习惯的呈现方式。而在元宇宙时代，随着可穿戴设备和可交互设备的普及以及对音视频媒体消费体验的更高要求，触觉呈现将成为视觉与听觉呈现的补充维度。用户在消费音视频内容时可感知的触觉反馈将不再局限于基础的振动反馈，而是包括振动、压力、速度、加速度、温度、湿度、嗅觉等全方位体感的更逼近真实世界的触觉呈现体验。

3.5 体验评测

元宇宙应用的沉浸式体验包括视觉、听觉、触觉、味觉、嗅觉，其中主要以视觉和听觉两种感官维度占总信息输入的 90%以上，目前从传统 2D 音视频已经有了相对成熟的评测指标和评测手段，但对 3D 视频质量和三维音频质量的评测指标与手段还在发展阶段。此外，元宇宙时代的内容强调实时互动性，所以在音视频互动性上也要增加对应的评测手段和指标。所以当前我们重点投入 3D 视频、三维音频的音视频质量和交互响应

这两个方面的体验指标和评测方法。而当前触觉、嗅觉、味觉等新的感官体验的技术还在早期探索阶段，目前在评测指标与评测方法上还没有系统化方法。

1. 视频体验

视频输入是在元宇宙时代用户获取信息的最主要途径，图像呈现技术目前正处于超高清平面视频向立体化的自然视频逐步发展的阶段，除了原有的平面视频在分辨率、帧率、声音、色彩等方面在立体视频上的延伸之外，3D 视频带来的视场角、CG 渲染等质量也是影响用户体验的重要指标。

1）视频质量：在传统 2D 视频基础上增加 3D 视频的评测指标

3D 视频的评测指标在传统 2D 视频频码率、分辨率、帧率、HDR（位深/色域/最大亮度/动态元数据）上，需要增加 3D 视频源、视频 FOV 角度、6DOF 能力等衡量 3D 视频能力的指标。如 3D 视频源的支持，为了形成带有深度的 3D 影像，一个场景需要同时拍摄两个影片，即左眼视频源和右眼视频源。视场角 FOV，90°视场角往往是 VR 沉浸体验的及格线，未来的理想沉浸式终端的 FOV 应该在 110°-120°，当前的 VR 设备主要在 105°-110°，主流的 AR 设备主要在 40°-55°。

2）CG 生成的 3D 内容：目前主要集中在数字人的质量评测上

元宇宙时代的视频内容，除了根据真实环境拍摄之外，还有大量通过计算机生成（CG），包括数字世界的“人物场”，对于 CG 生成的内容质量的主要衡量指标有 3D 内容的面片数、纹理、色彩、光照真实性等。其中数字人的评价指标和评测方法已经有了一定基础，如 CCSA 的《数字人系统基础可信能力要求及评估方法》中对数字人的形象、语音、驱动、交互、人设、工程化能力、安全性上都做了评估指标和测试评估的方法梳理。但目前在 CG 内容质量上还有两方面的不足，一个是对物和场景的评测上没有展开研究，一个是 CG 内容的质量还有大量主观的艺术评价成分存在，难以完全客观化和指标化。

3）3D 沉浸式视频内容的交互性：增加身体动作带来的交互体验的衡量指标

在视频交互方面，除了 2D 视频本身的播放流畅性之外，通过捕捉用户在观看视频中的 3DOF 和 6DOF 动作信息和设备输入信息，如头戴 HMD 的动作、XR 手柄输入、手势&身体动作识别的信息，视频响应的及时性与物理真实性都是影响用户体验的重要维度。此外，会增加 3DOF/6DOF、物理真实性、MTP 时延（身体、头部）、操作交互时延等新的衡量指标。（1）物理真实性：一般指在三维世界中物体除了视觉效果之外，在刚

性、重力、速度、热度、摩擦等物理维度，与真实世界物体的一致程度。（2）MTP 时延：运动到成像时延（Motion-to-Photon（MTP） Latency），是指从 IMU 或视觉传感器检测头部/手部的运动，到图像引擎渲染出对应的新画面并显示到屏幕的时延。VR 应用场景中，理想值是小于 20ms，在 AR 应用中，理想值为 5ms。（3）操作互动时延：当消费者对互动内容发出指令到接收到音视频回应的端到端时延 RTT，入门级为小于 100ms，舒适级为小于 70ms，理想级为小于 50ms。 当前在交互响应时延上有系统的评测指标和方法，但是对物理真实性、3D 物体的黑边率、穿模率等还需加强研究。

2. 音频体验

在元宇宙场景的音频体验，主要体现在为 XR 一体机、VR 头显、AR 眼镜、手机、智能网联汽车等设备打造身临其境的听觉体验。音频作为视听感受的重要感知系统，可以实现在元宇宙场景下，用户与用户、用户与场景、用户与对象之间的沉浸式交互体验，通过还原声音的方向、距离、音色、体积，打造真实的现场氛围感，增强用户的数字体验。

对沉浸式音频体验的评测，除需支持更高的音频采样率、位深、编码码率、支持声道数等传统音频评测维度外，还需要增加对空间音频渲染技术，即 HRTF 双耳渲染、Ambisonics 声场渲染、空间混响等技术的评测，客观测试包括频响曲线、谐波失真、延迟、频率范围等指标；主观测试应以沉浸感为主，应至少达到以下指标：（1）音频音色响度一致；（2）低声音延迟；双耳声音同步再现；（3）声音定位准确，轨迹清晰；（4）还原真实声场宽度；（5）系统整体协调等。在测试方法上，需要考虑空间音频渲染造成的无参考音频的情况，并区分扬声器和耳机等不同重放设备，区分 3DOF 和 6DOF 下不同体验以及有无视觉配合情况下的测试效果的不同。

3.6 版权保护

版权保护并不单指一项技术，而是针对源自文学、艺术作品扩展开来的各种类型传播媒体内容而展开的，用于解决著作权、访问权、使用权、所属权等一系列问题，而需要应用实施权限证明、密码保护和鉴别追溯等手段的技术体系。传统的版权保护更多是面向线下和实体的作品和形象，所以采用的手段以人工介入为主，在

元宇宙时代的超高清数字内容，则是结合了密码技术对数字媒体内容进行数据信息的加密处理，但对于具备数字作品流通和交易的场景的元宇宙环境下的版权保护手段还需要探索和持续发展。

元宇宙环境下超高清音视频将升级为全景视频、自由视角视频、点云视频等为代表的新媒体，新的元宇宙数字媒体内容和以虚拟人或吉祥物为代表的人工智能生成物，将不断涌现在文学艺术、新闻资讯、互动娱乐、体育赛事、工业互联等不同领域场景中。这些虚拟与现实结合的交互应用将为大众的生活带来更多新的数字资产和资源配置方式，并催生出新的商业形态。

元宇宙虚拟环境由各种虚拟场景下的物体形象、人物形象、建筑及环境等，以及释放在虚拟场景中的数字媒体内容来共同组成。可以想到的是，提供和构建这些虚拟场景的主体必然是来自不同平台、团体或个人。而在元宇宙新的商业形态下，数字媒体内容所具备的新的商业价值必然隐藏着新的安全风险。所以基于版权保护的思路，需要针对来自于不同主体之间，对这些海量数字内容资产的相互访问和流通的商业需求，提供基于密码和区块链等技术手段的安全访问控制、权利保障和存证鉴别的技术能力，在确保数字媒体内容拥有者的利益，甚至一定程度上满足个体隐私空间的功能需求的同时，又要能兼顾国家对网络空间安全的监管和可控的政策需求。这些也是版权保护在未来元宇宙场景下面向超高清音视频技术应用时需要解决的一系列新课题。

1. 元宇宙场景下版权保护的对象定义范围

相较于依赖广播电视或互联网流媒体等传播的传统视音频作品或媒体内容，在元宇宙场景下呈现和流通的数字媒体内容有更加多样的类型：

传统视音频的数字媒体内容：基于传统视音频采集编码技术，在元宇宙场景下仍可应用于影视点播、直播，视频会议，在线教育等视频应用场景。这种情况如果继续沿用国内外的 DRM 等相关技术手段，在目前阶段还能满足基本的版权保护需求，却较难适应未来元宇宙虚拟世界场景的新需求。比如，对于虚拟影院、虚拟会议室以及个人空间等应用，现有技术还无法满足对上述场景中涉及的视角移动、旋转、远近等新型音视频媒体形式的适配需求。未来技术发展不仅仅要面向数字媒体内容产生、传输和存储的环节，更可能会涉及到在云端或终端的计算机图像实时渲染的环节，在网络环境上也会对应在云、边、端的各种节点类型中。

增强元宇宙的新数字媒体内容：如全景视频、自由视角视频、点云视频等为代表的元宇宙音视频，以及未来发展出来的新视频技术，将会采用与传统视音频不同的采集和编码技术，带来新数据类型、数据格式，以及数据量大、实时性和画面同步性要求高等一系列问题。所以版权保护针对这些新视频内容的产生、传输、分发、存储、展示的全生命周期过程，也需要研究高效安全的技术方案。

更加灵活的数字作品：影视、艺术、生活以及商业周边的美术作品、数字形象、个性肖像、音乐、短视频、二创作品等已经开始以数字藏品或 NFT 的形式出现在互联网上，作为新一代互联网的元宇宙场景，这些数字媒体内容作为数字资产也是元宇宙虚拟场景的重要组成部分之一。针对这种情况，目前主要基于区块链和智能合约技术进行数字存证的版权保护方式基本还处于探索阶段，诸如数字资产的唯一性标识、版权归属有效性的界定和延伸、多人创作和二次创作对数字版权的权利归属等问题，还需要在业务开展和相关司法定义的进步过程中来完善。

人工智能衍生的虚拟主体：元宇宙的虚拟世界环境给基于人工智能的虚拟人、吉祥物提供了生存环境，并且随着人工智能、大数据和相关技术的进一步发展，这种能具备一定自主创作的虚拟主体也许有可能创造出属于“他们”的作品，这是未来需要进一步考虑和延伸的版权保护定义范围。

2. 元宇宙场景下版权保护相关技术

版权保护涉及到数字媒体内容全生命周期下的多种业务环节，而元宇宙场景下的数字媒体内容既是信息载体，也是数字资产，与虚拟世界下的各种虚拟主体会产生愈加复杂的相互联系，因此对相关技术的研究和发展提出相应的需求。

- 内容生成和编码阶段

包括超高清音视频在内的元宇宙下的数字媒体内容，在内容生成阶段是数据的最原始状态。

首先，可以按针对超高清音视频等各种元宇宙下新的数字媒体内容的数据格式要求和标准规范，对内容进行统一性的分类和唯一性的数字资产标识，而确定数字资产唯一标识的方法是超高清音视频内容后续生命周期各环节进一步分析处理的基础，并且随着数字媒体内容的格式不断发展，数字资产标识也需要相应适配。

其次，这里可以通过数字特征提取、区块链智能合约及其他安全可信的数字存证技术，将数字内容的原始数据、创作数字内容的主体、权利归属信息等原始的版权信息关联起来，以类似 NFT/NFR 的数字存证方式在元宇宙的虚拟世界中注册一个数字媒体的“身份证”。

然后，在必要的场景下，在适配了新的元宇宙视频技术的条件下，可以通过数字水印技术将标识、时间戳、单位等描述信息注入和隐藏到视音频媒体内容中，为后续生命周期环节提供一个“随身携带”的可以检验的数字证据。

最后，在有确定商业需求的情况下，还需要通过密码技术对视频媒体内容进行加密处理后再存储，这里会涉及到密钥管理服务系统、授权管理服务系统等基础平台在元宇宙网络环境下的部署，从而为后续的内容审查、发布流通、媒资管理等环节提供基础安全服务。

- 网络传输与平台间流通阶段

一方面，当超高清音视频等元宇宙下的数字媒体内容生成并发布在业务服务的网络节点之后，根据业务需要会进行平台间的数字资产交换和流通，必要时还需进行内容审核和监管。目前较成熟的方式是通过数字身份认证、内容加密、数字水印、授权访问控制等技术手段，保证数字媒体内容在网络传输过程中远离信息泄露和盗取等安全问题，并在必要时追溯查询流通传播的路径。但同时，目前的技术还需要对超高清音视频等元宇宙下的数字媒体内容进行适配，提高内容加密数字水印的计算效率，同时也存在不同版权保护框架下安全平台间密钥交换、身份互认等问题需要完善。

另一方面，数字媒体内容在生产阶段已经具备的唯一标识和存证信息，在平台间交换流通数字内容资产的过程中会发挥重要的作用。目前的趋势是通过区块链和智能合约技术，将交换流通的各个环节都存证上链，基于区块链技术的密钥管理平台服务也在探索之中。然而关于区块链技术需要说明的是，作为目前被认可的安全存证技术，具有一定安全稳定不可篡改的特性，但国内相关研究单位也亟待在降低分布式存储资源、提高共识计算效率、提升链上数据查询效率、增加智能合约在跨链互通等情况下的安全性等方面做进一步研究。

- 终端呈现阶段

超高清音视频等元宇宙下的数字媒体内容最终呈现在用户终端时，需要获取内容媒体的访问控制权限，根据实际业务需要可能包括身份认证、授权获取、解密密钥等，由于终端对实时性、低延迟等用户体验的需要，版权保护技术也需要进一步在新媒体内容格式的适配，密码计算效率等方面研究和提升。

• 安全执行环境的问题

还需要考虑的是，以上版权保护所涉及的各个阶段，都离不开安全的执行环境，否则极易造成敏感信息的泄露。因此对于版权保护体系来说，不管是数字存证的签名和验证，还是中级环节的密码计算，需要进一步解决在云、边、端各个计算节点中尚不完全具备安全可信执行环境的问题。

4.元宇宙时代超高清音视频技术标准化及建议

4.1 标准化需求

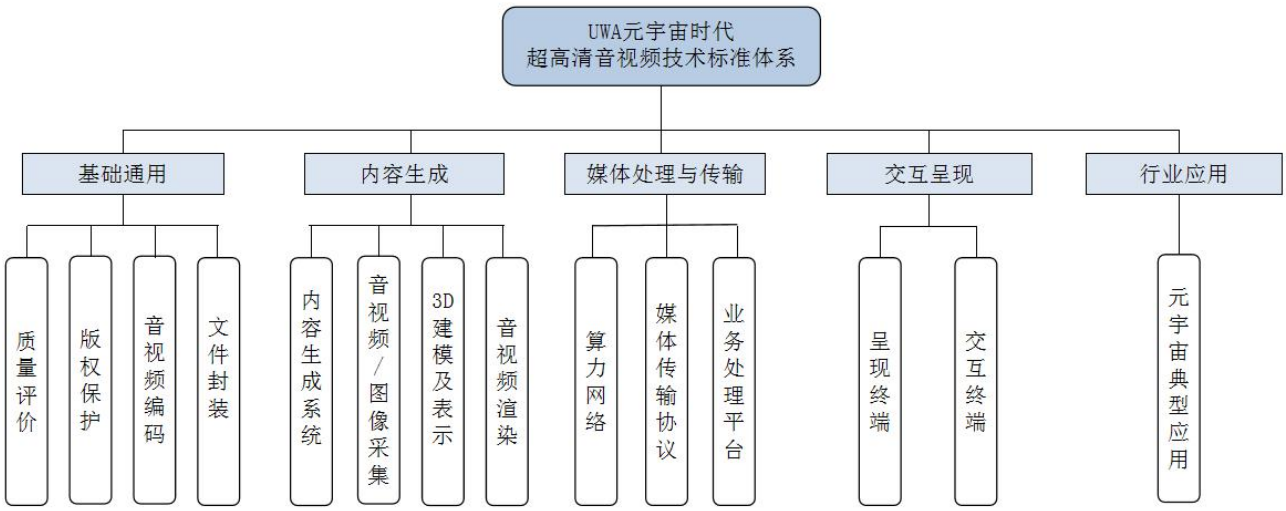

图 10 UWA 元宇宙时代超高清音视频技术体系框架

元宇宙时代音视频技术涉及非常丰富的业务应用场景，音视频技术作为元宇宙的基础技术，不同的场景有音视频相关的新技术需求。根据本白皮书中分析的元宇宙时代超高清音视频技术现状及发展趋势，从音视频采集、制作、传输、呈现、应用等各环节提取出标准化需求，提出元宇宙时代超高清音视频技术的标准体系，包括基础通用、内容生成、媒体传输与处理、交互呈现、行业应用五个方面，给出了初步的标准化建议。

4.2 基础通用标准化建议

1. 音视频编码

1）全景视频

国际上，ITU-R SG6 研究组 2018 年 4 月，发布了报告 Report ITU-R BT.2420-0，对近年 VR 视频的发展进行了梳理。随后 2019 年 1 月发布了 Recommendation ITU-R BT.2123-0，规范了 VR 视频的基本参数。针对 VR 沉浸式音视频数据格式、编码、传输等关键技术，MPEG 开启了“Coded Representation of Immersive Media”（简称 MPEG-I）系列标准（标准号：ISO/IEC 23090）的制定工作。3GPP 组织目前已经研究发布了 3GPP TS 26.118 和 3GPP TR 26.999。JPEG 组织正在研发一种描述 360°全景视频图像编码的技术标准 JPEG XT 和一种具有极低延迟和极低复杂度的压缩算法 JPEG XS。IEEE 1857 工作组针对沉浸视频内容发布 IEEE 1857.9 标准，用于压缩、解压缩和重建沉浸式视频内容。

国内方面，2016 年初，中国数字音视频编解码技术标准工作组（简称“AVS 工作组”）启动了 VR 国家标准的制定工作（以下简称 AVS VR）。2019 年 7 月，AVS VR 视频编码标准正式被国标委立项，项目名称：《信息技术 虚拟现实内容表达第 2 部分：视频》，项目编号：20192086-T-469。

未来，针对“更高技术格式、更新应用场景、更美视听体验”的高新视频新业态发展需求，围绕全景视频节目制作和交换等环节，面对全景视频关键技术发展方向，制定有关节目制作的基本技术流程、视频参数、视频编解码、全景摄像机的主要技术要求和测量方法等相关技术标准和规范；面对更新应用场景需求，制定适用于全景视频节目在 5G 网络、有线电视网络、IPTV、互联网中的传输分发、云服务平台和系统、各类全景视频终端等的通用技术要求、技术规范以及质量测评标准。

2）自由视点视频

国际上，SC29/WG4 目前正在重点研制 ISO/IEC 23090-12“沉浸式视频（MIV）”标准，同时也在制定 DAM“与 V3V 协同的扩展原理”和相关验证试验。WG4 正在起草 23090-23“MIV 的符合性和参考软件”的相关研究草案。另外，SC29/WG4 还对 MIV 进行了 2 项探索试验（在 MIV 中使用低复杂度增强视频编码、基于目标的编码）。

国内方面，IEEE 1857 和 AVS 成立的虚拟现实标准制定联合工作组 VRU 在国标《信息技术 虚拟现实内容表达第 2 部分：视频》中增加自由视点视频档次，该标准转化成 IEEE 1857.9 沉浸式视频标准， 作为国际标准于 2021 年颁布。

未来，围绕新兴沉浸式生态系统的关键需求，将制定针对光滑、透明或高反射表面的编解码和渲染标准；面对实时沉浸式视频播放需求，制定支持更灵活地捕获、编码和呈现沉浸式体积内容的标准。

3）3D 点云

国际上，ISO/IEC JTC 1/SC29/WG7 正在开展 23090-9“基于视频的点云压缩数据的传输”、ISO/IEC 23090-21“G-PCC 参考软件”、ISO/IEC 23090-22“G-PCC 符合性”、ISO/IEC 23090-20“V-PCC 符合性”、ISO/IEC 23090-19“V-PCC 参考软件”。

国内方面，2019 年 3 月 AVS 工作组启动了我国自主的点云压缩专题组工作，填补了国内在点云压缩标准技术领域的空白，制定具有自主知识产权的点云压缩标准已经成为自动驾驶及数字孪生等新兴领域技术竞争和行业竞争的必要之路。

未来， 立足服务国家产业发展及科学研究，围绕**数字孪生产业及空间科学研究**对**海量时空图形数据存储使用**的需要，以数据为基础，以算法为核心，制定发布高效的点云编解码系列标准：针对**自动驾驶、数字博物馆**等产业应用及天文等科学研究，实现对点云数据的几何空间信息及多种属性信息的高效表示，兼顾通用软硬件实现复杂度，制定发布高效的点云编解码系列标准，包括系统部分、点云编解码部分、参考软件及符合性测试；利用神经网络轻量化方法实现点云编解码网络的推理加速，以便于后期利用成熟的深度学习处理器和芯片设计制造流程，实现软硬件系统支持和应用系统服务。

4）3D Mesh

动态 3D 网格编码标准方面，国际上 ISO/IEC JTC 1/SC29/WG7 针对网格视频编码发布了测试模型 V-Mesh TMC（目前已更新至 v1.1），并进行了新的探索试验。未来纳入标准考虑的探索方向为帧间预测参考 mesh 的生成、压缩失真主观评价指标、高效的熵编码、基于 AI 的端到端编码等。

未来，围绕开放协同的虚拟现实生态建设和数字孪生产业发展，面向海量实时采集和生成的高精度网格模型数据，制定发布高效的网格编解码系列标准：针对虚拟/增强现实应用和实景三维中国建设，重点探索实景

采集网格及重建网格与编解码的关联性，实现面向网格数据压缩的通用性高效表示；面向混合现实和数字孪生等应用场景中的可变速率、可伸缩渐进式传输和抗噪等需求，制定兼顾灵活性和鲁棒性的网格深度编码解决方案。

5）Audio Vivid 三维声

世界超高清视频产业联盟(UWA)牵头，与 AVS 编解码标准协同，联合产业端到端生态，于 2022 年 4 月推动发布三维菁彩声（Audio Vivid）技术团体标准草案，并被国家广播电视总局接纳为行业标准。

对比业界现有的三维声技术，三维菁彩声（Audio Vivid）技术标准的目标是面向全球，技术先进，是一个更加开放的、具备产业安全要求的技术标准和方案，同时产业生态政策友好，更加适合超高清产业生态各方进行端到端的产业部署。

2. 文件封装

音视频文件封装主要以国际标准化组织(ISO)早期制定的 ISO/IEC 13818-1 (TS) 和 ISO/IEC 14496-12（MP4）标准作为基础规范，进行各种音视频编码的封装标准制定版本，标准并仍在持续更新，如今年 VVC 和 EVC 的 TS/MP4 的封装格式规范均已经制定完成。 同时，国内外标准组织已发布或正在制定适用于元宇宙时代应用场景的音视频数据的文件封装标准，如 ISO/IEC 23090-2（OMAF）。

国际上，MPEG 已制定了支持 H.264/HEVC/VVC 编码标准的文件封装标准，也同样制定了全景视频、自由视点视频、点云数据的文件封装标准，并且还在制定沉浸音频、触觉数据的文件封装标准。国内方面，AVS 工作组基于上述标准制定了 AVS3 视频编码的文件封装标准 AVS3-P6，并正在第二版中支持对 AVS3 音频编码的文件封装。AVS 工作组也很早就开始了音视频编码标准化的工作。AVS3-P6 在 2022 年 6 月已经完成第一阶段的工作，制定了 AVS3 视频编码的文件封装标准，第二阶段草案也已经完成，主要是增加了 AVS3 Audio Vivid 音频编码的支持。同时，AVS 也正在制定支持沉浸式媒体的文件封装标准，包括对全景视频、自由视点视频、点云等多种媒体数据格式的支持。随着 AVS3-P6 标准的逐步发布，AVS3 音视频的生态系统将更加多样化。

3. 质量评价

元宇宙最佳体验状态是用户无法区分现实世界和虚拟世界，需要满足沉浸感、低延迟、多样性、随时随地、身份、朋友、经济、文明等八大指标特征。基于音视频技术对元宇宙服务用户体验的影响因素及原理分析，结合当前音视频技术发展情况，将从沉浸体验、交互体验、安全保障三方面构建质量指标标准化体系。

1）沉浸体验指标标准化

根据当前技术发展情况，将沉浸感分成三个维度，第一个层次为沉浸在能够关注到的信息之中，这是目前移动互联网已能达到的初级沉浸效果；第二个层次为感官上的沉浸，这也是当前元宇宙初级阶段要达到的效果；第三个层次为大脑的沉浸，是指我们的全脑全然地认为我们沉浸在某个世界里，这是元宇宙将要发展的终极状态。结合元宇宙在不同应用场景的特点，需音视频技术上满足一定的指标方能实现第二层以上的高质量的沉浸效果。

元宇宙沉浸效果的上限由**内容源**决定，对元宇宙内容提出内容分辨率、内容帧率、码率、音质、内容生成及时性等参数指标标准化要求、并制定体验评价方法可极大提升用户的沉浸体验。元宇宙沉浸体验还与**终端设备**的解码和显示性能密切相关，对元宇宙终端设备提出解码性能及显示性能的标准化要求及评价方法可有力保障用户的感官体验。目前主要采用 VR/AR 等近眼显示设备作为元宇宙的接入端口，这类近眼显示设备直接由用户佩戴在头上使用，若佩戴不舒适将对沉浸体验产生干扰。对近眼显示终端提出重量、尺寸、发热及散热、面部贴合度、透气性等舒适度质量指标可优化用户的使用体验，增强沉浸感体验效果。元宇宙沉浸体验也受**平台处理能力**的影响。为实现显示终端轻量化的发展，元宇宙的内容渲染将交由平台进行处理，高质量的渲染效果能提供更真实的视听感受，但对平台的处理能力提出更高的要求。此外，由于元宇宙的强社交应用场景需要平台满足**多用户同时在线**的功能，如元宇宙交互用户过多，超出平台的渲染处理能力，可能导致 GPU 渲染的帧率降低，从而影响元宇宙服务的流畅度体验，因此对元宇宙平台提出指标标准化要求可有效提升用户的沉浸体验。元宇宙服务的沉浸体验也与**传输网络**的性能有关。传输网络的带宽、丢包、时延及抖动都可能导致服务体验受损。

2）交互体验指标标准化

交互体验主要受交互维度数、交互响应时延以及交互精度的影响，从这三个维度开展指标标准化工作能有效保障用户的交互体验质量。

交互维度数主要指 HMD 和操控手柄支持的交互维度数，目前主要为 3DOF/6DOF 技术，后续随着应用场景的发展与变化，可能需要有更多的肢体节点支持多维度交互，应根据技术发展，对交互维度、定位精度、反馈准确性、决策响应准确度等方面开展标准化研究工作。

交互精度主要指虚拟形象的表情、动作、情感等交互反馈的真实度，采用多模态交互技术提供支持。如在强社交应用场景中，基于眼动跟踪、表情捕获、智能语音生成、肢体动作还原等技术能生成更真实的数字人形象。可根据不同的应用场景分别开展标准化研究工作，根据技术发展制定交互精度质量指标要求，提升数字身份的交互真实性。

交互响应时延则主要包括头部 MTP 时延、肢体 MTP 时延（可以是多节点）、操作性响应时延（如扣动扳机）等指标参数。交互响应时延越低则用户体验越真实。从 MTP 时延（身体、头部）、操作交互时延等项目开展交互体验质量技术标准化研究，能有效保障元宇宙的社交属性，促进虚拟社会的建立。

4. 内容安全与版权保护

制定内容安全、版权安全等质量标准化指标，为元宇宙服务中身份、文明等特征提供安全保障。

内容安全质量标准化指标主要包括：（1）制作内容及画面的安全性，保障元宇宙社交内容中不出现反动、血腥或造成观感不适的相关内容；（2）传输内容的安全性，保障内容在传输、存储等过程中不被修改或删除。

版权安全质量标准化指标主要包括：（1）肖像安全性，对云宇宙社交个人肖像或 NFT 头像的安全保障；（2）内容版权安全，对传输内容的版权保护及版权管理标准化。

4.3 内容生成的标准化

1. 内容生成关键环节

元宇宙时代下音视频的内容多样并支持实时交互，内容生成过程复杂，包括音视频/图像采集、音视频制作、3D 建模、渲染等多个环节。

1）音视频/图像采集

音视频/图像采集作为内容生成的第一个环节。国家广播电视总局办公厅印发《4K 超高清电视节目制作技术实施指南（2020 版）》，推荐了 4K 超高清电视节目拍摄制作方法与流程，中国电影电视技术学会制定了《广播级 4K 超高清摄像机的技术要求和测量方法》标准。音视频/图像采集标准涉及采集设备能力、布设、运动、精度等方面。

2）3D 建模及表示

3D 模型是元宇宙场景中的基本元素，MPEG 正在研究场景的通用描述格式以及 3D 图像媒体表述，通过对 3D 图像统一描述可现实 3D 场景的渲染、呈现、交互。3D 建模及表示标准涉及 3D 模型内容制作平台、3D 模型/场景通用格式等方面。

3）音视频渲染

音视频渲染是内容制作的关键环节之一，对用户视听体验有直接影响。CCSA 研制了《面向 AR/VR 的分布式云渲染系统技术要求》《面向渲染业务的云平台能力要求》标准以支持大数据量、高质量、实时性的渲染技术标准。音视频渲染标准涉及通用的渲染参考架构、渲染能力平台、渲染引擎等方面。

2. 内容生成系统

为提高元宇宙端到端的内容生产效率，降低生产成本，完善及加速推进元宇宙产业发展，需要进行内容生产系统整体的标准化工作。当前已在 XR 虚拟融合制作及数字人制作的系统进行了标准探索研究，后续智能的内容生成系统、针对特定行业应用的内容生成系统等都将是标准化的方向。

1）XR 虚实融合制作系统

元宇宙时代下音视频内容的制作是虚实融合的制作，在中央广播电视总台超高清视音频制播呈现国家重点实验室的发起下，世界超高清视频产业联盟已发布了国内首个 XR 虚实融合制作系统，该标准分别对 XR 虚实融合制作系统包括 LED 大屏系统、拍摄系统、目标跟踪系统、渲染系统这四个子系统制定了技术要求并明确了测试方法，为指导 XR 虚实融合制作的规范开展提供了明确的技术指导，也为开展面向后续虚拟影棚、虚拟演播室等具体应用场景的标准化工作奠定了基础。

2）数字人制作系统

虚拟数字人的生产也离不开内容制作系统的支持。当前相关标准还处于起步阶段，但是 ITU-T、ISO 等国际标准组织和 CCSA 等国内标准组织都已经认识到虚拟数字人标准的重要性，积极推进虚拟数字人标准的制定。但是只涵盖了虚拟数字人关键技术、产品与服务、行业应用、等领域的一部分，制作系统还是空白，远远不能满足虚拟数字人高效生产的需求。因而需要我国相关单位、公司和行业人员积极推动数字人制作系统相关标准的制定。如形象生成（三维建模、美妆捏脸、服饰装扮…）、多模态生成（语音生成、动作生成、表情生成…）、集成显示（显示标准、跨平台显示…）等内容生产和评测标准的制定。

4.4 媒体传输与处理

1. 算力网络

算力网络作为一种新型网络架构，算力网络将分布的算力资源进行连接并提供算力和网络的统一编排，同时为应用屏蔽异构算力资源，提供统一的算力服务。从国内外标准进展来看，算力网络整体架构目前在国内达成了初步共识，但相关关键技术及业务场景的研究还处于初级阶段。在国内，中国通信标准化协会 CCSA TC3 研究了算力网络的网络功能架构以及接口和模块的技术要求的行标，TC1 正在研究面向视频业务的算力要求应用场景的行标。在国际上，国际电信联盟电信标准分局 ITU-T SG11 和 SG13 进行研究，其中 SG11 研究了 Q.CPN 标准（算力网络的信令需求）与 Q.BNG-INC 标准（算力网络边界网关的信令要求），SG13 研究了 Y.CPN-arch（算力网络架构与框架）标准、Y.CAN（算力感知网络）等系列标准。算力网络的标准目前主要是从架构、安全和服务等几方面进行研究。元宇宙下面向视频业务的算力网络建议进一步对业务使能、算网大脑、算力使能、网络使能等组件进行标准化研究，提供对音频数据超大规模、超低时延传输，以及提供对各类业务的处理、融合和运营管理。

2. 媒体传输

元宇宙场景中需要支持多种类型的音视频数据传输，以及对实时性、互动性有较高的要求 。3GPP SA4 正在进行 5G_RTP、iRTCW 、FS_ eiRTCW 等标准研究项目，将针对沉浸的实时业务（如 XR 业务）的沉浸媒体和相关元数据的实时传输，以及沉浸的实时通信。同时， MPEG 已制定或正在制定支持元宇宙场景的全景视频、多视点视频、点云数据等沉浸媒体的传输标准，使用扩展的 DASH/MMT 协议传输 MPEG 的沉浸媒体封装文件。

IETF 和 W3C 组织于 2021 年将 WebRTC 采纳为官方标准，目前也正在研究下一代 WebRTC 标准。WebRTC 工作组正在开发媒体捕获和媒体流以及屏幕捕获等规范，同时审阅支持 WebRTC 新用例的技术提案；探索边缘计算对 Web 平台的影响以及有关用例和需求，在 Web 浏览器中整合网络质量监测和预测。

为支持元宇宙中不同场景的媒体传输，将对潜在的媒体传输协议的功能扩展并优化使用（如 RTP 协议、WebRTC 协议）、传输的功能组件等进行标准研究，以及对元宇宙中潜在的、新兴的沉浸媒体数据格式提供灵活的传输/访问机制（如基于空间的媒体访问，基于视角的媒体传输）进行标准研究，以提高传输效率，减少终端开销，增加沉浸体验，满足不同的业务场景。

3. 业务平台

元宇宙时代的音视频业务平台将基于基础算力网络构建，提供各种视频基础业务能力，并对上层应用服务和终端开放应用能力，包括能力平台（如视频平台、元宇宙引擎、云渲染平台、AI 算法平台）、管理运营平台。基于元宇宙的业务的多样性和开放性，建议研究业务平台架构、功能要求、接口、技术流程、网络需求等技术指标。

4.5 呈现与交互标准化

从显示、音效、交互、健康四个维度说明对虚拟显示终端产品的标准化要求。

1. 呈现终端

1）显示性能标准化指标

虚拟现实硬件技术、软件技术和交互技术是影响视觉体验的核心因素，可从分辨率、双眼视差、3D 串扰、屈光度调节以及系统延迟等项目开展标准化工作。

2）音效性能标准化指标

VR 音频质量受包括音频质量、音源方位、定位准确度和混响在内等参数影响，清晰度、音色平衡影响音质体验，沉浸感、定位准确度则影响空间感。通常来说高阶立体混响信号具有最佳的综合性能。5.1 声道和高阶环境立体混响信号在还原双耳信号在沉浸感方面表现更好，而基于对象的双耳信号在定位精度方面表现更好，即基于声道和基于场景的双耳 3D 音频擅长场景渲染，基于对象的双耳 3D 音频擅长对象渲染。因此，

在电影和游戏制作中，用于实现特定的声音效果的基于对象的音频，通常与用于实现场景声音的基于通道或基于场景的音频结合使用。此外，减少视觉信息也会显着降低音频质量分数。

建议从音频质量、音源方位、定位准确度和混响等项目对沉浸式听觉产生影响的项目开展标准化工作。

3）健康标准化指标

过度使用 VR/AR（一般大于 30 分钟）可能对眼健康造成一定伤害，具体临床表现包括：头痛、头晕、恶心、同时伴有眼睛干涩、甚至出现复视、流泪、眼涨疼、视力疲劳、精力无法集中等症状。类似于视疲劳、干眼、视频终端综合症等功能性眼病的症状。因此急需开展近眼显示产品的健康标准化研究工作。

由于显示画质，纱窗、拖尾、闪烁等过低的画面质量引发的视觉疲劳会进一步引发眩晕。而提高辨率、响应时间、刷新率，降低 MTP 时延等性能可有效缓解上述症状。

视觉与其他感官通道的冲突。如前庭功能障碍、幽闭恐惧症。可通过强化眼动追踪定位，提高广视角和角分辨率，改善畸变等，强化视觉与听觉、触觉、前庭系统、动作反馈的协同一致，进而缓解冲突症状。

辐辏调节冲突。佩戴 VR/AR 头显设备后，双眼在产生立体视的同时，双目焦点调节与视觉景深并不匹配。因而 VR/AR 头显设备难以如实反映类似现实世界中观看远近物体的清晰或者模糊变化。因此可选择非固定焦深的多焦点显示（Multifocal Display）或可变焦显示（Varifocal Display）缓解上述问题。

此外，由于可穿戴产品具有使用时贴近人体的特性，材料安全就尤为重要。

因此从设备外壳及与人体接触部分材料是否会引起人体过敏现象、是否会刺激皮肤、是否含有有毒有害物质及元素等方面开展标准化工作；并应跟进技术发展，及时提高对分辨率、响应时间、刷新率，降低头动和视野的延迟（MTP）、视角、畸变、多焦点显示/可变焦点显示等项目的要求，为消费者提供更舒适健康的观看体验。

4）信息安全

具有云服务功能的交互与呈现设备可收集的数据主要涉及用户的生理活动、生活偏好、行为习惯、行动轨迹等，这些数据通常会通过蓝牙、无线网或局域网上传至云端或移动设备，若一旦发生信息泄露，将极大的侵害到用户利益。提升设备保护信息和数据的能力，以使未授权的人员或系统不能阅读修改信息和数据，允许

授权人员或系统进行访问，将有效提升用户使用终端设备时的安全性。建议从平台信息安全、终端固件信息安全、传输信息安全等项目开展标准化研究工作。

2. 交互终端

在元宇宙时代的音视频呈现中，交互是衡量用户沉浸感的一个重要指标，建议从角度漂移、采样频率、跟踪范围及误差、灵敏度、延迟、捕获精度、反馈准确性等交互性能参数开展标准化研究工作。

4.6 产业应用标准化

音视频技术作为元宇宙时代的重要基础之一，将音视频技术与重点行业领域融合，推动云宇宙的产业应用规模发展。产业应用标准涉及文化旅游、教育培训、零售购物、娱乐游戏、政务办公，工业生产等。

5. 总结与展望

在全球经济形势复杂多变和新冠疫情的影响下，人类社会生活和生产方式面临新的挑战，信息消费与产业数字化转型也随之迎来新的机遇。作为新一代信息技术融合创新的典型领域， 超高清音视频在未来互联网发展中不可或缺。

超高清音视频将撬动新一代信息技术产业链上下游新的增长空间。在终端硬件方面，2017 年全球智能手机出货量首次下降，虚拟现实等新型超高清终端成为智能终端发展的重点领域，预计 2025 年终端出货量达亿部。高分辨率、大视场角、轻薄小型化的发展趋势催生超高清显示器件在近眼显示、微显示等领域市场需求，高性能图形渲染与感知交互技术推动集成电路在超高清沉浸影音市场拓展。**在数字内容方面，**在音视频质量优化这一既有发展轨道上的迭代难以触发超高清用户体验的增量跃升。在元宇宙概念下，新型超高清音视频将推动覆盖“采、编、播、传、显”的传统超高清音视频向三维沉浸影音方向演进升级，用户数字体验日益身临其境。**在基础设施方面，**超高清应用侧群体互动的巨大变化需算力与通信等基础设施随之升级，算力聚焦分布式、实时性支持沉浸媒体的计算能力，传输网络聚焦低延迟、高稳定、高流畅的全球跨域确定性通信能力。

超高清音视频新型终端有望成为新的互联网入口。随着终端技术在感知、计算、传输、显示等方面的发展迭代，以手机为代表的传统智能终端难以承载新型超高清音视频人机交互的需求，桌面计算针对信息处理，智能手机聚焦沟通互动，元宇宙概念下终端入口或将是“体验式的”，未来移动互联网智能终端将由以信息处理为中心向以适人体验为中心方向发展，业界聚力发展 XR 等新型超高清智能终端，探索布局后移动互联网时代终端入口主阵地。

新型超高清音视频平台将有望重新界定现有互联网信息模型与价值模型。当前互联网公司主要提供交易、文娱等功能平台，用户基于功能类别形成相应的产业生态。元宇宙概念下，超高清音视频生态竞争或将更加激烈，主要表现在：**一是，**超高清音视频将带来新的信息形态与交互形式。传统互联网业务在实现用户虚实融合、身临其境的数字体验方面存在局限，平面的数字世界与立体的物理世界难以直接融合，3D 技术产业链条的垂直突破与水平贯通将驱动超高清虚实融合数字空间成型，在这一背景下，各大 ICT 巨头纷纷积极布局三维数据格式、传输、存储、渲染、建模、仿真与交互等细分领域。此外，当前数字世界中“浏览旁观”的交互方式难以承载大众在现实世界中的“自主体验”，元宇宙 3D 虚实融合的数字空间将解锁“加强版”交互体验的新形态，除既有交互精度、频度与广度外，业界正对交互深度这一新方向大力投入。**二是，**超高清音视频将在云、网、边、端、用、人等融为一体的创新体系下重构现有系统架构，触发产业跃迁，进而在这一深度融合创新的框架下，重新界定并迭代优化一批新技术、新产品、新标准、新市场与新业态。其发展难点与焦点在于如何解构重组流媒体、交互应用及机器视觉等传统任务流程，即结合超高清音视频沉浸化、互动化、三维化的趋势特性，探索云网边端各环节计算负载分担的“最优配方”。**三是，**用户可进行独立的内容生产和对数字世界的创造，超高清音视频内容主导权或将由企业向大众创作者转移，价值收益模式由网红流量货币化向影音内容货币化方向转变，有望推动互联网平台经济的价值重构。

新型超高清音视频将向生产生活领域加速渗透，赋能千行百业。在文娱休闲场景，针对传统文娱体验互动性有限、社交性不足、体验形式单一等现状问题，新型超高清音视频支持融合型、分享型和沉浸型数字内容与服务，有助于围绕信息技术融合创新应用，打造信息消费升级版，培育中高端消费领域新增长点。**在教育培训场景，**针对传统教学过程中部分课程内容难于记忆、难于实践、难于理解等现状问题，新型超高清音视频将推动教学模式由被动接受向自主体验升级，有助于提升教学质量与行业培训效果。**在工业生产场景，**针对产品

复杂度的不断提升、技能娴熟工人的紧缺、设计开发与规划生产的协同、营销与销售绩效的压力等问题，新型超高清音视频可为开发设计、生产制造、营销销售、运营维护等人员连接起数字世界和现实世界，提升企业数字化转型过程中，从多元数据获取洞察的能力与水平。**在商贸会展场景，**针对线下会展参与可行性与便利性、固有组织成本、传统线上活动感官认知与互动体验受限等现状问题，新型超高清音视频有助于实现会展组织由以活动议程为中心向与会体验为中心的方向转变。

6. 附录

6.1 缩略语

下列术语和定义适用于本文件：

波场合成（WFS，Wave Field Synthesis）

超高清视频（UHD，Ultra High Definition）

高动态范围（HDR，High-Dynamic Range）

高阶立体声场信号（HOA，Higher-Order Ambisonics）

高帧率（HFR，High Frame Rate）

广色域（WCG，Wide Color Gamut）

混合现实（MR，Mixed Reality）

计算机图形学（CG，Computer Graphics）

基于 HTTP 的动态自适应流（DASH，Dynamic Adaptive Streaming over HTTP）

基于几何的点云编码（G-PCC，Geometry-based Point Cloud Compression）

基于视频的点云编码（V-PCC，Video-based Point Cloud Compression）

扩展现实（XR，Extended Reality）

全向媒体格式（OMAF, Omnidirectional MediA Format）

人工智能生成内容（AIGC，AI Generated Content）

视场角（FOV，Field of View）

数码电影视频技术（MOV，Movie digital video technology）

数字化影院系统（DTS，Digital Theatre System）

素材交换格式（MXF，Material eXchange Format）

通用场景描述（USD，Universal Scene Description ）

通用媒体应用程序格式（CMAF，Common Media Application Format）

通用视频编码标准（VVC，Versatile Video Coding）

头戴式显示器（HMD，Head-Mounted Display）

头相关传递函数（HRTF，Head-related Transfer Function）

图形语言传输格式（glTF，Graphics Language Transmission Format）

往返时延（RTT，Round-Trip Time）

网际互连协议（IP，Internet Protocol）

文本格式字幕（SRT，SubRip Text）

虚拟现实（VR，Virtual Reality）

移动边缘计算（MEC，Mobile Edge Computing）

一阶立体声场信号（FOA，First-Order Ambisonics）

用户生成内容（UGC，User Generated Content）

用户数据报协议（UDP，User Datagram Protocol）

运动显示时延（MTP Latency，Motion-To-Photons Latency）

增强现实（AR，Augmented Reality）

专业生成内容（PGC，Professional Generated Content）

自由度（DOF，Degrees of Freedom）

HTTP 的自适应码率流（HLS，HTTP Live Streaming）

ISO 媒体文件格式（ISOBMFF，ISO base media file format）

MPEG 沉浸式视频（MIV，MPEG Immersive Video）

UHD World Association
世界超高清视频产业联盟