arXiv:2611.32311v1 [cs.CV] 24 Nov 2026

SynAVID：动态场景合成沉浸式视听数据集

Jun Xu[1] SujingLi[1] ChunhaoBi[1] Zhengxue Cheng[1] Ronghua Wu[2]
Qunshan Gu[2] Li Song[1]1上海交通大学 [2]蚂蚁集团

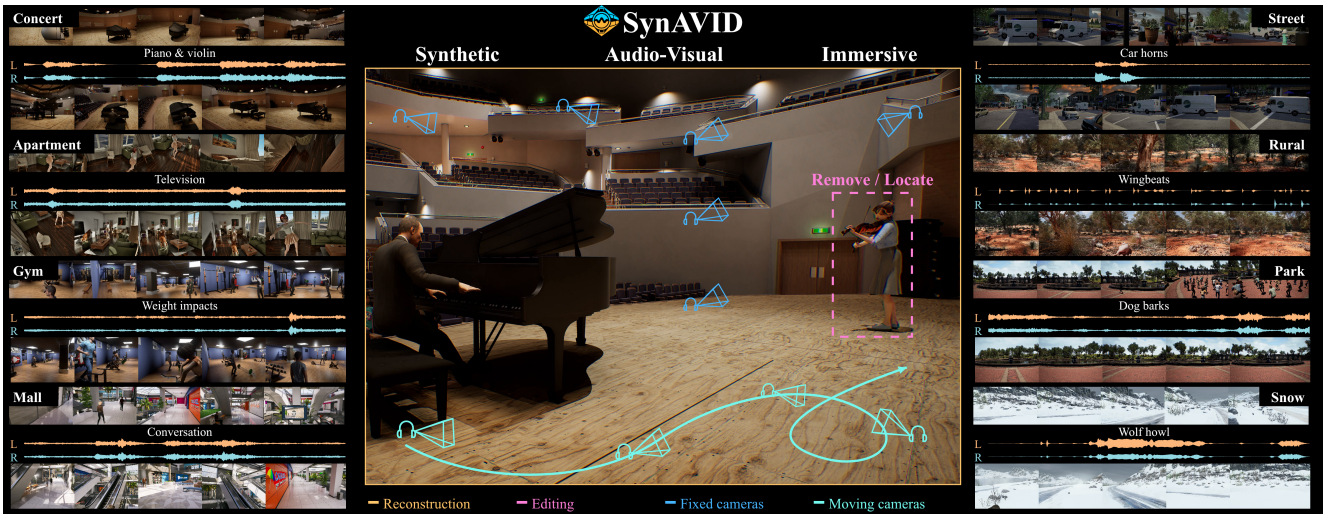

图 1。**SynAVID 概览**。侧面板展示了来自八个场景的多视角图像和选定的双耳（左/右）波形。中心展示了用于视听重建与编辑的固定和移动摄像机-听者对；位置和轨迹均为示意图。

摘要

沉浸式场景建模需要在视点和场景条件不断变化时，同步获取视觉与听觉观测。对于受控场景变化，获取此类观测及参考数据仍然颇具挑战。我们推出 ***SynAVID***，这是一个合成视听沉浸式数据集，包含八个室内和室外场景。每个分钟级片段都结合固定摄像机阵列和移动轨迹，并提供与视点相关且同步的双耳音频。共享的场景动态、声源事件和观察者姿态共同驱动视觉渲染与几何感知声学合成。我们评估了现有的视觉重建、双耳音频重建和视听编辑方法。混合观测配置通常能提升固定和移动测试视图上的视觉重建效果，但短暂被占据和观测稀疏的区域仍然难以处理。不同方法在频谱、双耳间和感知音频指标上的排名各不相同。几何与实例标注揭示了观测支持度分布不均，以及总体评分之外的局部高斯表示。成对参考数据用于评估目标编辑及其周边内容，而受控的几何变化则会产生依赖于听者的声学响应。项目页面位于 https://avery-xu.github.io/synavid-project/。

1. 引言

创建人们可以自由探索的数字环境，是沉浸式计算的核心目标，其应用涵盖虚拟现实、远程临场和交互媒体 [17]。这类体验依赖于用户在不断变化的场景中移动时，所看到和听到的内容之间保持连贯的关系 [16]。视觉能够呈现外观、运动和遮挡，而空间声音则传达声源的方向和活动信息，包括视野之外的声源，以及声源与周围环境之间的相互作用。这两种模态都会随着观察者的位置而变化
以及人和物体的运动与朝向变化 [18]。要从新视角体验动态场景，并保持视觉和声学线索的一致性，必须对这些关系进行建模。开发和评估此类能力，需要能够在统一时空框架内捕捉场景动态变化以及依赖于观察者的视觉和听觉响应的数据。

现有数据集推动了动态视觉重建的发展——

表1。 **与代表性数据集的比较**。 视图数量和时长概述了报告的采集配置。 **动态** 指场景运动； **固定/移动** 指摄像机。 **音频** 包括录音或脉冲响应。 **GeoGT**：参考深度/表面，不包括重建估计； **配对编辑**：匹配的原始/编辑后采集。 ✓/×：包括/不包括；—：不适用。

	数据集	来源	视图	时长	动态	固定	移动	音频	双耳	GeoGT	配对编辑
Dynamic visual	HyperNeRF [34]	Real	1–2	30–60 秒	✓	×	✓	×	×	×	×
	神经三维视频 [24]	Real	18–21	10–40 秒	✓	✓	×	×	×	×	×
	Google Immersive [6]	Real	41–46	0.3–30 秒	✓	✓	×	×	×	×	×
	DiVa-360 [29]	Real	53	5–200 秒	✓	✓	×	✓	×	×	×
	Charge [33]	合成	41	≈1–7 秒	✓	✓	✓	×	×	✓	×
	Syn4D [19]	合成	8	≈ 17秒	✓	✓	✓	×	×	✓	×
	MultiCamVideo [1]	合成	10	5.4 秒	✓	✓	✓	×	×	×	×
	StaDy4D [39]	合成	9	5–30 秒	✓	✓	✓	×	×	✓	✓
Spatial acoustic	RAF [10]	Real	图像	—	×	—	—	✓	×	×	×
	RWAVS [27]	Real	1	10–25 分钟	×	×	✓	✓	✓	×	×
	MRTD [13]	Real	360° 张图像	—	×	—	—	✓	✓	×	×
	AcousticRooms [28]	合成	深度图	—	×	—	—	✓	×	✓	×
	SoundSpaces-PanoIR [7]	合成	RGB-D全景图	—	×	—	—	✓	×	✓	×
	SoundSpaces-NVAS [8]	合成	4张图像	—	×	—	—	✓	✓	✓	×
Immersive audio-visual	FAIR-Play [12]	Real	1	10 s	✓	✓	×	✓	✓	×	×
	Tragic Talkers [2]	Real	22	≈13–43 秒	✓	✓	×	✓	×	×	×
	EasyCom [11]	Real	1	60 s	✓	×	✓	✓	✓	×	×
	ImViD [42]	Real	39	1–5 分钟	✓	✓	✓	✓	×	×	×
	Replay-NVAS [8]	Real	8	3–8 分钟	✓	✓	×	✓	✓	×	×
	Ego-Exo4D [14]	Real	5–6	8 秒–42 分钟	✓	✓	✓	✓	×	×	×
	MRSAudio [15]	Real	1–2	主要是 ＜ 10s	✓	✓	×	✓	✓	×	×
	SonoScene4D [20]	Real	1	8 s	✓	✓	×	✓	×	×	×
	SynAVID（本文方法）	合成	**60**	**120 秒**	✓	✓	✓	✓	✓	✓	✓

真实与合成多视角数据集提供了关于动态外观和几何结构的丰富观测；合成数据还额外提供密集的几何标注，但其主要侧重点仍是视觉重建与理解 [19, 24, 33]。声学数据集和仿真器刻画了声音如何随声源和接收器配置而变化，但通常聚焦于静态场景几何 [7, 10, 27]。动态视听录音通过捕捉不断演变场景的视觉与听觉观测，将这两个方向联系起来 [8, 42]。然而，物理采集难以独立控制场景演变和观测条件，因此难以系统地研究这些因素对视觉与声学重建的影响。上述方向共同促使我们探索能够联合捕捉动态视觉与声学行为，同时允许系统控制场景内容和观测条件的数据。

我们介绍**SynAVID**，这是一个合成视听沉浸式数据集，涵盖动态的室内和室外环境（图1）。SynAVID 通过密集的固定摄像机阵列和移动轨迹，捕捉分钟尺度的活动，为同一不断演变的场景提供互补观测。每台摄像机都与一名位于同一位置的双耳听者配对，生成与观察者位置和朝向相对应的同步视频和空间音频。共享的场景配置和事件时间线驱动视觉渲染及感知几何结构的声学合成，将可见运动、声源活动以及声音传播的变化联系起来。这种共享配置使观测轨迹能够独立于场景活动而变化，同时，几何标注和配对编辑参考支持区域重建诊断以及对目标编辑的评估。

我们的贡献主要有三点：

• SynAVID，一个包含固定和移动多视角视频以及同步、依赖听者的双耳音频的动态室内与室外场景合成数据集。

• 一种生成工作流：沿不同的摄像机—听者轨迹重放共享的场景与声音事件时间线，并为受控的场景编辑生成参考观测。

• 使用标准指标对视觉重建、双耳音频重建和视听编辑进行评估，并辅以区域诊断和受控的动态遮挡案例。

2. 相关工作

表1比较了视觉、声学和联合视听观测方面的代表性数据集。

动态视觉数据集。动态高斯重建中常用的数据集包括适用于单目设置的 D-NeRF [35] 和 HyperNeRF [34]。Neural 3D Video（N3V） [24], 、Google I mmersive [6], 、Technicolor [37], 和 DiVa-360 [29] 提供了互为补充的动态场景同步多视角录音。合成数据集则可更灵活地控制场景内容和观测条件：

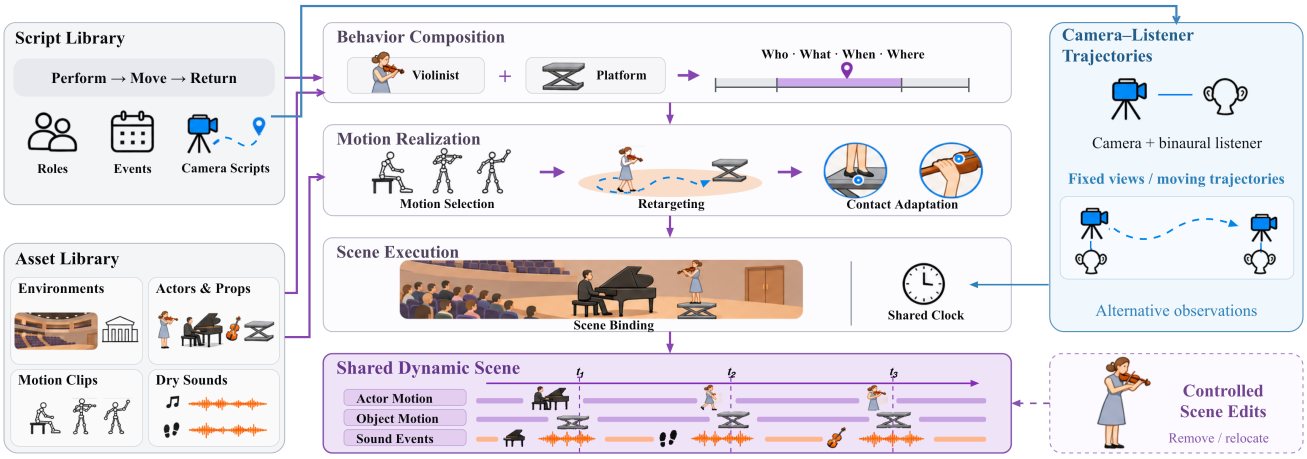

图2. **SynAVID 中的场景组织**。 资产库和脚本库支持行为组合、动作实现和场景执行，并在共享时间线上生成演员动作、物体运动和声音事件轨迹。摄像机—听者轨迹指定对同一片段进行固定或移动观测。虚线所示的干预表示对选定场景元素进行受控更改；观测轨迹和场景内容可以分别调整。

Charge [33] 和 Syn4D [19] 结合了多种摄像机配置与几何监督，而 Multi-CamVideo [1]则着重于多样化的同步摄像机轨迹。StaDy4D [39] 还提供了有动态演员和无动态演员的配对拍摄素材。这些资源通过互补的拍摄布局、场景动态和参考标注，为视觉重建提供支持。

空间声学数据集与仿真。 实测资源涵盖 RAF [10] 中与视觉对齐的单声道房间响应，以及 RWAVS [27] 中与图像配对的移动听者双耳录音。仿真通过可控的声源、接收器和环境配置，补充这些实测数据。SoundSpaces 2.0 [7] 支持连续声学查询，SoundSpaces-NVAS [8] 则将渲染视图与双耳语音配对。AcousticRooms [28] 改变房间几何结构和材料，而 AudioWorldSim [43] 则生成双耳导航序列。这些资源主要刻画静态场景几何中的声学变化，即便信号或听者随时间变化亦是如此。

沉浸式视听数据集。 联合录制通过互补的视觉与声学观测，捕捉动态活动。ImViD [42] 提供分钟级的多视角视频及摄像机音频，而 Replay-NVAS [8]则提供同步视角和经测量的双耳目标音频。Ego-Exo4D [14] 结合了移动的第一视角视图、固定的第三视角视图以及设备音频。MRSAudio [15]的各个子集中都包含与双耳录音和Ambisonics录音配对的视频；SonoScene4D [20] 则将全景视频与在另一个聆听位置录制的一阶Ambisonics录音配对。这些数据集的参考数据均与录制时的配置绑定，因此场景变化后难以进行受控比较。SynAVID通过共享的场景和时间线，协调动态几何结构、声源事件以及固定和移动的摄像机–听者对，为这些资源提供补充。其同步视频和建模的双耳参考音频，支持在已知条件和受控场景变化下进行联合评估。

3. 数据集与制作流程

SynAVID 将动态片段的构建（图 2）与通过成对的相机和听者进行观测（图 3）区分开来。

3.1. 场景组织

场景内容、脚本化行为和观测轨迹共用统一时钟。一个片段定义了环境、演员、物体和预定声源；独立指定的摄像机—听者轨迹决定了场景如何被观看和听见。

图 2 将可复用资产和脚本组织为三个阶段。 行为组合 在空间和时间上分配参与者、动作和声音事件。 动作实现 选择并调整动作片段和接触。 场景执行将动作和发射器绑定到场景实例，并沿时间线评估其状态和声源活动。

不同的摄像机—听者轨迹可观测同一未改变的片段。受控编辑会修改选定元素，以便与原始配置进行比较。

3.2. 动态场景构建

在每个片段中，持续活动与局部运动和交互相结合，同时改变可见性和声音传播。

动作片段被重定向到角色上，并通过短暂的混合连接成更长的表演。路径将这些动作置于环境中；局部逆运动学和接触调整使脚、手以及手持

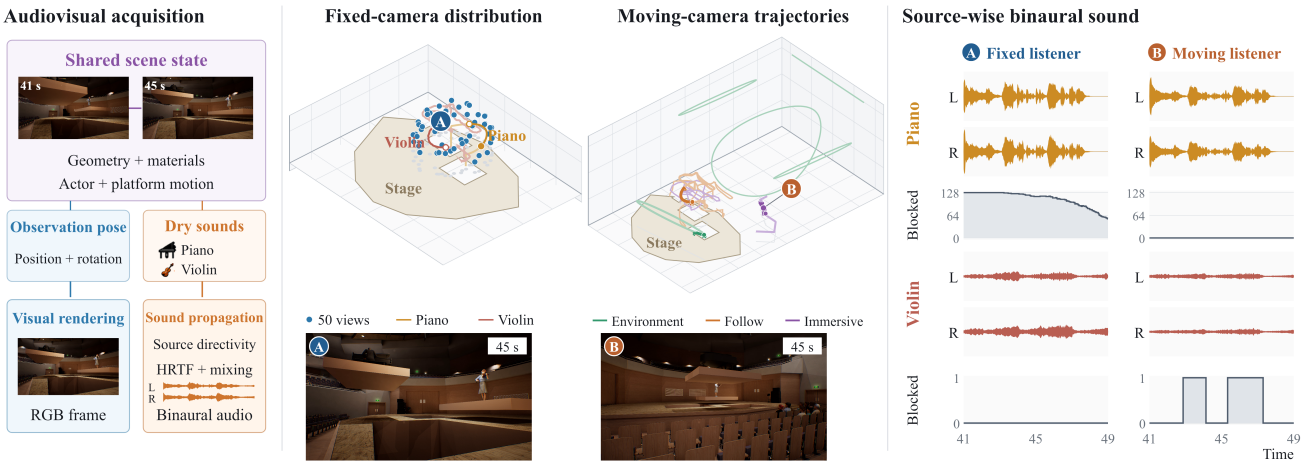

图3. **视听采集与示例观测。** 共享场景状态驱动 RGB 渲染和空间声音合成（左）。不同缩放级别下，固定摄像机、移动路径和声源运动均以音乐会舞台为参照展示（中）。浅色路径覆盖整个片段；深色线段突出显示代表性轨迹中 41–49 秒的部分。绿色/橙色/紫色分别表示环境/跟随/沉浸式摄像机。A 和 B 将 45 秒时的 RGB 画面与 41–49 秒期间按声源区分的左右声道声音配对展示（右）。镜像 RMS 包络在共享振幅刻度上显示接收声音总量；遮挡 计数表示遮挡可见性样本。建模细节见补充材料。

物体与支撑表面对齐。物体运动遵循相同的时间线，其中也包括承载表演者的平台。

干信号绑定到演员、物体或固定发射器，通过预定片段和增益定义发出什么声音及其发出时间。发射器轨迹标示各声音的位置；脚步声和接触声音随动作变化，而音乐和语音则占据指定时段。音乐播放在事件层面进行协调。

3.3. 同步视听采集

互补视角。 图 3 （左）中的共享场景状态为每个摄像机-听者对提供视觉和声学输入。固定阵列从多个方向反复观测活动区域。环境、跟随和沉浸式轨迹带来不断变化的观看距离、角度和场景覆盖。它们使更多表面进入视野，并以不同的时长出现，从而在一个片段中形成持续和间歇性的观测。图 3 （中）展示了音乐会舞台周围的固定阵列，以及延伸至周边空间的移动路径，同时还展示了钢琴和小提琴声源位置的变化。Unreal Engine 以30 fps渲染原生4K RGB，并提供相应的摄像机内参和位姿。

听者相关声音。 声源和听者的位姿、指向性以及中间几何体共同塑造了各个视点接收到的声音。Steam Audio [40] 通过一系列准静态场景状态，模拟直达和反射传播。声学材料取自 Acoustic- Rooms [28]及其预设；钢琴采用测得的指向性模型 [4]。在图3 （右）中，尽管直达路径受到严重遮挡，听者 A 仍能听到钢琴声；而听者 B 则会遇到小提琴遮挡的时段。这些总声包络包含反射，因此直达路径受阻并不一定意味着

表2。**场景覆盖。** 四个室内（上）和四个室外（下）环境中的代表性动态与声源。每个120秒的片段结合了50个固定视角和10条移动摄像机-听者轨迹。

场景	动态	主要声音
音乐会	表演，平台	钢琴，小提琴
公寓舞蹈，家具移动 电视，狗，机器人		
Gym	锻炼，器材使用 音乐，语音，撞击声	
Mall	步行，自动扶梯	语音，脚步声
街道	行人、交通	发动机、喇叭声、语音
乡村	动物 ，车辆	鸣叫，车辆
Park	锻炼、社交活动	音乐、语音、狗
Snow	动物 ，车辆	鸣叫，接触雪 ，车辆

不会产生静音。各声源的专属波形揭示了每种声音如何影响听者听到的混合音频。

配对视听观测。 听者与每台摄像机的位置和朝向一致。FABIAN头相关传递函数 [5]将声源响应映射到听者的耳朵，从而得到视点 c 和耳朵 $e \in \{L, R\}$处的混合音：

$$y_c^e[n] = \sum_s g_s \sum_k h_{s,c}^e[n,k] x_s[n-k], \qquad (1)$$

这里， x_s 是预定的干信号， $h_{s,\ c}^e$ 是随时间变化的双耳滤波器，而 g_s 是各视角共用的声源增益。图像和48kHz 音频对应相同的场景事件，而接收到的声音保留了听众之间的传播延迟、混响尾音和相对声级。实现细节见补充材料。

3.4. 数据集概述

SynAVID涵盖四个室内和四个室外环境（表2）。每个环境包含一个从50个固定摄像机和10条移动轨迹中观测到的120秒片段，



图4。 **跨场景的新视角重建**。 各列展示不同场景，各行展示真实值及各方法的结果；每列中的测试摄像机和时间点均相同。F/FM 分别表示固定/混合训练视角。帧由压缩视频解码而来；表 3 使用浮点渲染结果。

共计16分钟的不同场景时长、480个视图序列和16个视图小时。

对于每个视角，基础版本提供 RGB视频、混合双耳波形、摄像机和听者参数、时间戳以及划分元数据。源数据包括干信号、发射时间线和发射器绑定信息。

根据请求，同一批片段还支持导出其他轨迹以及深度、表面标签和几何结构。补充材料详细介绍了发布内容、参考覆盖范围、声学设置和实现检查。

4. 实验

我们的实验涵盖动态视觉重建（第4.1节）、双耳音频重建（第 4.2节）以及视听编辑（第 4.3节）。除重建与编辑结果外，我们还利用区域标注和一项受控几何干预，考察局部视觉表征以及听者相关的声音传播。

4.1. 视觉重建

我们在全部八个场景中评估 FreeTimeGS++(FTGS++) [23]，时空高斯（STG） [26]，和 4C4D [44] 。每种方法都针对每段 120 秒的序列拟合一个模型。FTGS++使用仅固定视角（F）或固定与移动视角混合（FM）的训练视图；STG 和 4C4D 使用 FM。我们在固定和移动的留出视图上进行评估。

表3 报告了两个测试组的峰值信噪比、结构相似性指数和学习感知图像块相似度。FM FTGS++ 在七个场景的所有指标上均优于 F，其中室外移动视角下的峰值信噪比提升了 2.13–6.95 分贝。音乐会仍是一个例外：STG 的表现最佳，且 FM 并未始终优于 FTGS++。在图4中，FM 减少了街道和公园中的大面积背景伪影，但人物仍然缺失或呈现重影。所选重建结果在公寓、健身房和乡村场景中保留了更多可辨认的主体，而音乐会和购物中心中的人物则严重模糊或几乎完全缺失。这些结果表明，在混合观测条件下，重建效果有所改善，但局部失效仍然存在；我们将在购物中心场景中对此进行更深入的考察。

观测支持度。以往研究指出，动态重建中的监督在时间上分布不均 [38]。图5 对比了购物中心一个视角在 68 秒时的三种支持度模式：短暂有人停留的位置、持续被观测到的店铺招牌，以及主要由移动摄像机拍摄到的走廊墙面。在 2 Hz 采样下，几何结构和实例标签估计这些固定世界区域中可见的像素；对各视图求和并随时间积分，即得到像素秒。走廊墙面约 92% 的投影像素支持度来自多个时间段内的移动视图，但其累计总量仍远低于招牌。上排并列展示了在这些不同支持度模式下，FM 改善的墙面外观与缺失的人物。

最终表示。 近期工作还通过高斯的分配来考察缺失的动态内容 [36]。我们通过对合成权重求和来衡量局部表示 $T\alpha$ ，这些权重来自中心位于每个固定身体区域或表面邻域内的高斯，并在目标核心像素上取平均。尽管 FM 改善了背景外观，但人物和墙壁的局部贡献仍远低于店铺招牌（图5，右）。墙壁上大部分渲染不透明度来自其定义邻域之外的高斯，因此外观改善可能与较弱的局部贡献并存。这些最终模型诊断描述了所选区域；这些区域的大小和时间占用率各不相同，完整定义见补充材料。

表 3。**固定和移动测试相机上的视觉重建**。F：46 个固定训练视角；FM：53 个室内或 54 个室外固定和移动训练视角。分数先对每条测试路线中的帧取平均，再分别对固定组或移动组中的路径取平均。**粗体**/下划线：每个场景和指标中的最佳/次优结果。FTGS++ F/FM 在训练得到的初始化方面也有所不同；完整设置见补充材料。

方法	固定测试			移动测试			固定测试			移动测试		
	峰值信噪比↑	结构相似性指数↑	学习感知图像块相似度↓	峰值信噪比↑	结构相似性指数↑	学习感知图像块相似度↓	峰值信噪比↑	结构相似性指数↑	学习感知图像块相似度↓	峰值信噪比↑	结构相似性指数↑	学习感知图像块相似度↓
	音乐会						**公寓**					
FTGS++^{F} [23]	22.422	0.6218	0.5414	16.868	0.5306	0.6153	20.570	0.7810	0.1888	21.422	0.7798	0.1742
FTGS++^{FM} [23]	20.284	0.5691	0.6643	17.749	0.5866	0.6684	**22.001**	**0.8351**	**0.1518**	**22.908**	**0.8466**	**0.1324**
STG^{FM} [26]	**25.443**	**0.7775**	**0.2641**	**19.889**	**0.6587**	**0.3821**	21.462	0.8024	0.2028	22.178	0.8080	0.1810
4C4D^{FM} [44]	20.864	0.6812	0.4119	18.478	0.6178	0.5218	17.884	0.6239	0.4722	18.753	0.6303	0.4492
	Gym						**Mall**					
FTGS++^{F} [23]	19.877	0.7503	0.2261	21.858	0.7917	0.1883	13.012	0.3555	0.5071	12.860	0.3812	0.5027
FTGS++^{FM} [23]	**23.346**	**0.8553**	**0.1329**	**25.403**	**0.8869**	**0.1163**	**18.023**	**0.6337**	**0.3029**	**17.629**	**0.6770**	**0.2849**
STG^{FM} [26]	22.713	0.8240	0.2134	24.018	0.8404	0.2069	17.572	0.5994	0.3846	16.899	0.6269	0.3791
4C4D^{FM} [44]	15.030	0.6235	0.5720	18.190	0.6976	0.4929	11.788	0.4106	0.8021	11.631	0.4368	0.8133
	街道						**乡村**					
FTGS++^{F} [23]	25.151	0.8082	0.1685	14.874	0.3555	0.5376	17.949	0.4686	0.3468	16.710	0.3636	0.4683
FTGS++^{FM} [23]	**26.792**	**0.8501**	**0.1377**	**21.823**	**0.7308**	**0.3115**	**19.905**	**0.5584**	**0.2999**	**18.841**	**0.4698**	**0.4363**
STG^{FM} [26]	24.973	0.7907	0.2505	20.532	0.6663	0.4596	19.267	0.4978	0.4320	17.818	0.4113	0.5458
4C4D^{FM} [44]	19.419	0.5868	0.5497	17.654	0.5442	0.7249	16.833	0.3110	0.7600	15.957	0.3104	0.7440
	Park						**Snow**					
FTGS++^{F} [23]	21.885	0.6144	0.2551	15.577	0.3631	0.5034	20.309	0.7914	0.2307	13.985	0.5714	0.4713
FTGS++^{FM} [23]	**22.132**	**0.6436**	**0.2468**	**19.407**	**0.5266**	**0.4068**	22.216	**0.8649**	**0.1677**	**17.970**	**0.7626**	**0.2992**
STG^{FM} [26]	21.092	0.6402	0.3306	18.745	0.5002	0.5068	**22.659**	0.8217	0.2300	17.864	0.7408	0.3241
4C4D^{FM} [44]	15.643	0.4319	0.6635	15.158	0.3648	0.8225	18.100	0.6984	0.4192	14.775	0.6308	0.5234

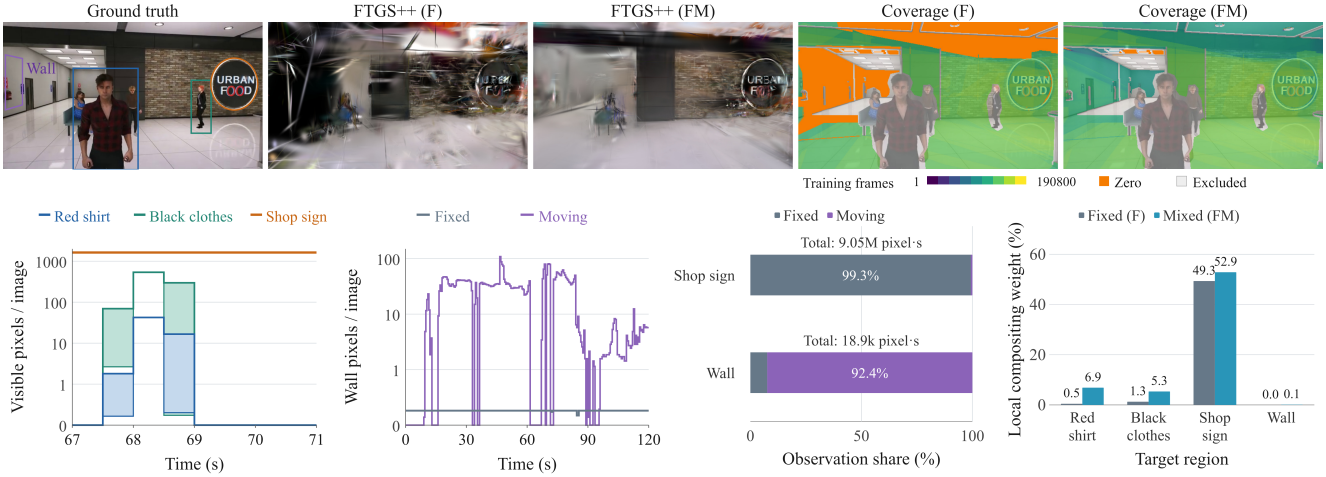

图5. **购物中心观测与表征诊断**。 上：共享测试视图、FTGS++ 重建结果以及静态表面覆盖率（对数刻度；橙色：零；褪色部分：已排除）。下方从左至右：固定人物/标志位置处的支持度、固定视图和移动视图中的墙面支持度、累计观测占比与像素秒总量，以及目标像素上邻近高斯的平均合成权重。支持度曲线以 symlog 刻度显示每张训练图像对应的像素数；阴影表示较低和较高的几何归因界限。在采样时间中，67–71 秒以外的人物支持度为零；总量则对各视图和时间上的支持度进行积分。

4.2. 双耳音频重建

在八个场景中，我们根据在另一视角录制的同步空间混合音频，预测目标视角的双耳音频。模型针对每个场景分别进行训练，使用 46 个固定视角，并在另外 4 个保留的固定视角上进行评估；对于室内和室外场景，采用统一的预处理和评分方式。对于每个场景，各方法均使用一段来自训练视角的输入录音；完整设置见补充材料。

我们将 AudioGS [3], ViGAS [8], 和 AV-Cloud [9]（Full）进行比较，并以双耳复制作为参考基线。AudioGS 使用音频和听者位姿，而 ViGAS 和 AV-Cloud 还会使用视觉观测。输入不包括干声源、测试视图图像以及引擎几何结构或声学响应。

表4 每个场景平均包含 480 个一秒测试窗口。STFT 表示复频谱 L2 误差；RTE 是一种辅助

表 4. **双耳音频重建**。 在留出的固定听者位置处的误差，输入为来自另一视角的同步混合音频。双耳复制保留输入的左/右声道，不进行空间适配；DPAM/CDPAM 使用单声道下混。数值越低越好。LRE：分贝；RTE：秒；MSE：$\times 10^3$。**粗体**/下划线：每个场景和指标下的最佳/次佳结果。

方法	MAG↓	ENV↓	LRE↓	RTE↓	STFT↓	均方误差↓	DPAM↓	CDPAM↓	MAG↓	ENV↓	LRE↓	RTE↓	STFT↓	均方误差↓	DPAM↓	CDPAM↓
	音乐会								**公寓**							
AudioGS [3]	0.1447	0.06541	1.871	0.0337	206.15	3.200	**3.6573**	**0.15566**	0.0918	0.04138	0.796	0.0298	148.14	1.966	3.6566	**0.15483**
ViGAS [8]	**0.1132**	**0.05708**	**1.008**	**0.0281**	**189.30**	**2.725**	3.8276	0.66120	**0.0772**	**0.03684**	**0.555**	**0.0272**	148.68	1.989	3.7772	1.04410
AV-Cloud [9]	0.1310	0.06193	1.246	0.0288	202.92	3.046	3.7004	0.15615	0.0921	0.03882	0.698	0.0310	**140.07**	**1.746**	3.7567	0.25978
双耳复制	0.1634	0.07386	1.971	0.0358	224.69	3.747	3.7031	0.16021	0.1040	0.04383	0.949	0.0299	152.88	2.108	**3.6552**	0.19700
	Gym								**Mall**							
AudioGS [3]	0.0732	0.02380	0.743	0.0338	71.61	0.484	**3.7549**	**0.21796**	0.1314	0.04377	1.449	**0.0714**	105.83	0.949	3.9014	**0.24507**
ViGAS [8]	**0.0615**	**0.02180**	0.658	0.0283	**66.99**	**0.432**	4.0117	1.25600	**0.1055**	**0.03582**	1.165	0.0816	**80.79**	**0.587**	3.9555	1.22420
AV-Cloud [9]	0.0667	0.02383	**0.623**	**0.0268**	71.85	0.480	3.8734	0.23328	0.1087	0.03629	**1.066**	0.1027	85.34	0.628	4.0022	0.26779
双耳复制	0.0758	0.02444	0.771	0.0333	72.92	0.499	3.8107	0.23213	0.1338	0.04577	1.568	0.0776	108.30	0.998	**3.8639**	0.24631
	街道								**乡村**							
AudioGS [3]	0.0280	0.00675	2.359	0.0523	16.94	0.02177	**3.9018**	0.38975	0.0222	0.00530	3.702	**0.0432**	12.34	0.02271	3.6817	0.19429
ViGAS [8]	**0.0199**	**0.00519**	**1.168**	**0.0382**	**13.36**	**0.01448**	3.9517	**0.22257**	**0.0182**	**0.00439**	2.872	0.0540	**10.42**	**0.01566**	**3.6695**	**0.18888**
AV-Cloud [9]	0.0219	0.00540	1.419	0.0439	14.71	0.01709	4.0737	0.44165	0.0196	0.00467	**2.544**	0.0632	11.34	0.01763	3.7695	0.26065
双耳复制	0.0312	0.00787	2.420	0.0561	18.83	0.02610	3.9252	0.42361	0.0256	0.00607	3.724	0.0435	13.77	0.02899	3.7129	0.19979
	Park								**Snow**							
AudioGS [3]	0.0490	0.01338	1.961	0.0613	32.73	0.09613	3.8707	0.23156	0.0181	0.00743	4.897	**0.0496**	16.95	0.17798	**3.7080**	**0.20177**
ViGAS [8]	**0.0344**	**0.00913**	1.374	**0.0498**	**23.49**	**0.05265**	**3.8645**	**0.14499**	**0.01210.00425**		3.805	0.0794	**11.26**	0.06485	3.8747	0.22305
AV-Cloud [9]	0.0388	0.00970	**1.270**	0.0601	26.04	0.06055	3.9648	0.21445	0.0157	0.00483	**3.550**	0.0841	11.91	**0.06000**	3.8664	0.37259
双耳复制	0.0635	0.01941	2.120	0.0629	41.62	0.14419	3.8834	0.26267	0.0211	0.00946	5.292	0.0522	20.14	0.26047	3.7585	0.22046

表5. **音乐会视觉编辑**。四个固定视角，0–10秒， 1280×720；未编辑对照与编辑后真值共用。掩码峰值信噪比：目标轮廓或移动并集，不包括阴影。ProPainter/ReVideo：移除/移动，其中 ReVideo 已调整尺寸。

方法	输出	移除				移动			
		峰值信噪比↑	结构相似性指数↑	学习感知图像块相似度↓	掩码峰值信噪比↑	峰值信噪比↑	结构相似性指数↑	学习感知图像块相似度↓	掩码峰值信噪比↑
输入视频	未编辑	23.182	0.8784	0.0684	8.391	22.879	0.8747	0.0763	8.597
ProPainter [45]/ ReVideo [32]	已编辑	25.405	0.8848	0.0599	12.196	25.252	0.7369	0.1287	18.505
FTGS++ [23]	未编辑	22.474	0.8018	0.1692	8.554	22.237	0.7984	0.1759	8.763
	已编辑	23.727	0.8061	0.1649	11.151	23.566	0.8038	0.1669	11.444
TRASE [25]	未编辑	22.448	0.8209	0.1301	8.474	22.189	0.8173	0.1371	8.670
	已编辑	25.482	0.8373	0.1042	17.787	25.311	0.8350	0.1058	17.770

指标，用于衡量从预测音频 RT_{60} 和参考音频 [8]估计结果之间的差异。ViGAS 在全部八个场景中均取得最低的 MAG 和 ENV，并在其中七个场景中取得最低的 STFT 误差，表明其在室内和室外场景下都能持续改善频谱和包络增益。其他属性的排名则有所不同：AV-Cloud 在五个场景中的双耳对数能量比（LRE）方面领先，而 AudioGS 或 binauralcopying 在六个场景中的 DPAM 方面领先。DPAM [30]和 CDPAM [31]衡量单声道折叠感知相似度；LRE 衡量双耳平衡。

4.3. 视听编辑

我们的音乐会案例研究考察目标编辑和依赖几何结构的声学变化。视觉参考通过重新渲染场景，移除小提琴演奏者、小提琴和琴弓，或将它们水平移动 1.5 米，同时保持其他演员和时间线不变。音频移除的评估以原始钢琴音轨为基准；另一次单独的面板干预则重新计算传播。

设置。 ProPainter [45] 和 ReVideo [32] 在四个评估视角下编辑观测视频；FTGS++ 删除/平移和 TRASE [25]则从 46 个固定训练相机视角渲染这些视图。移除操作使用从 RGB 派生的 SAM [21] 掩码，并在各方法对应的输入视角下共享；引擎标签仅用于评估。每种方法均设有未编辑对照；全帧、目标掩码和互补区域的评分均使用编辑后真值。

视觉移除。 所有编辑器的表现均优于各自的未编辑对照（表 5）。ProPainter 的全帧结构相似性指数和学习感知图像块相似度得分最高，而 TRASE 的掩码峰值信噪比最高。图6（第一行）显示了 ProPainter 和 FTGS 留下的目标碎片++；TRASE 更彻底地移除了表演者，但保留了其阴影。

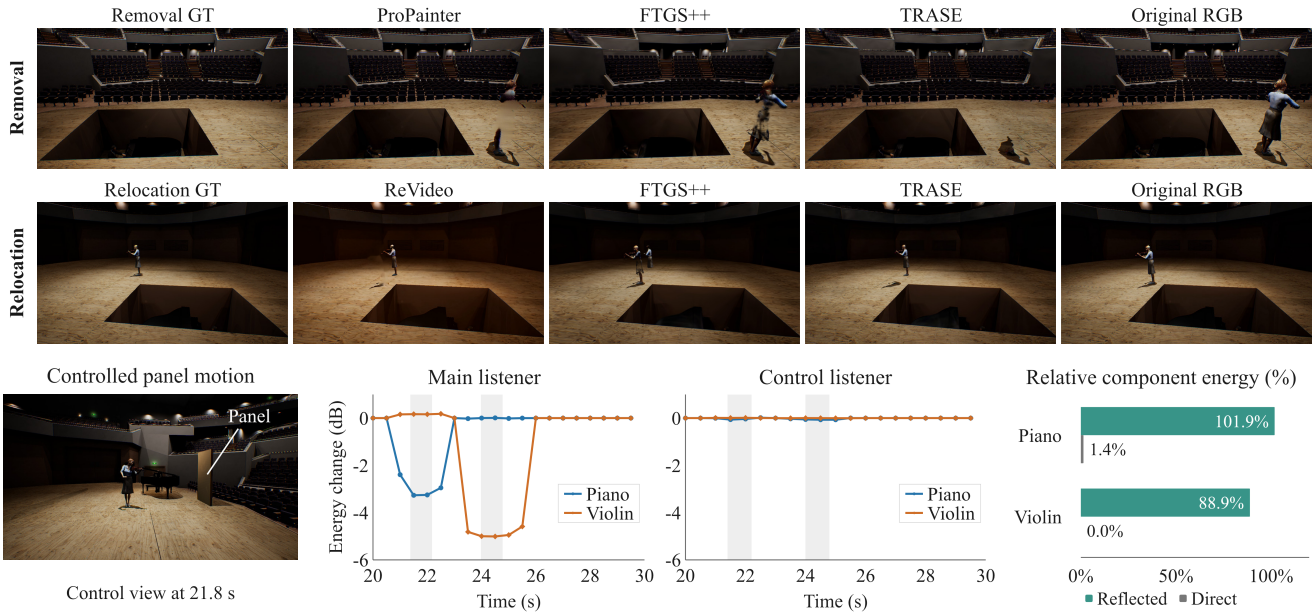

图6. **音乐会中的编辑与受控遮挡**。 上：时间匹配的移除（后方）与移动（前方斜向），对比编辑后真值、视频编辑器、FTGS++、TRASE 和原始 RGB。下：移动面板、相对于静止面板的两位听者处单位响应总能量变化（共用刻度），以及主听者处的相对直达/反射能量（100%：静止面板下的各分量）。阴影区域：钢琴/小提琴保持段；点：成对的 2赫兹响应。

表6. **音乐会音频移除**。 相对于原始钢琴音轨的四视角均值。未更改/静音：输入混合音频/输出为零；静音没有双耳评分。数值越低越好；LRE/ILD：分贝，IPD：弧度。

时间	方法	MAG↓	LRE↓	多分辨率短时傅里叶变换↓	ILD↓	IPD↓
0–10 s	未更改	0.10090	1.30209	6.78894	1.62532	0.13382
	SmartDJ [22]	0.07734	3.85581	5.40598	6.54081	1.41593
	Banquet [41]	0.10072	1.30032	6.78272	1.62508	0.13381
	静音	0.01296	–	1.05214	–	–
20–30 s	未更改	0.06298	0.21442	1.30935	0.44705	0.05603
	SmartDJ [22]	0.18376	3.56959	1.60936	5.53048	1.36872
	Banquet [41]	0.06297	0.21449	1.30920	0.44736	0.05605
	静音	0.23699	–	2.57097	–	–

视觉重定位。 ReVideo 的掩码峰值信噪比最高，但在全帧峰值信噪比相近的情况下，TRASE 的结构相似性指数和学习感知图像块相似度表现更好（表 5）。图 6 （第二行）显示了 FTGS 中的内容重复++ 以及 ReVideo 对表演者外观和背景的改变。与各自未编辑的输出相比，ReVideo 在目标并集之外的峰值信噪比降低了 1.91 分贝，而 FTGS++和 TRASE 的变化小于 0.2 分贝（补充材料）。

音频移除。 我们将文本引导的 SmartDJ [22]与干查询条件下的 Banquet [41]进行比较。多分辨率短时傅里叶变换衡量多分辨率频谱和/差误差；ILD 和 IPD 衡量双耳声级和相位误差。在 0–10 秒，静音会使 MAG 和多分辨率短时傅里叶变换误差最小，此时钢琴较晚才进入；而在 20–30 秒，当乐器声重叠时，二者误差均最大（表6）。SmartDJ 降低了早期频谱误差，却加剧了双耳误差；在声音重叠期间，五项误差全部相较于输入混合音频均有所增加，而 Banquet 则几乎未更改。综合来看，这些参照有助于区分目标编辑准确度与对周边画面及保留声音的保真度。

受控动态遮挡。我们在音乐会场景中渲染配对视频，并针对静止或移动的面板，在 20–30 秒内重新计算直达声和反射声；同时固定干声发射、表演者运动以及两组摄像机—听者位姿。面板依次穿过通往主听者的小提琴和钢琴声线路径；对照听者则始终保持直达视线。图6（底部）测量了各声源对相同单位脉冲的响应能量，从而将传播变化与音乐动态区分开来。在相应的停留阶段，主听者处的总响应能量分别下降 3.26 和 5.00 分贝，而对照听者处的变化低于 0.07 分贝。直达声受到显著衰减或被阻挡，而反射能量仍然存在。这一案例将可见运动与依赖听者的声学响应联系起来；统计数据和片段见补充材料。

5. 结论

SynAVID 将分钟级动态、固定与移动多视角视频以及同步双耳音频结合于一个可控的合成数据集之中。混合观测配置通常能够改善视觉重建，但局部失败现象持续存在；在频谱、双耳间和感知指标上，音频方法排名各不相同。几何结构和实例标注将这些观察结果与区域支撑度及最终表示联系起来，而配对编辑参照则评估目标变化与场景保真度—

进一步地，受控的面板干预将几何结构变化与依赖听者的声学响应联系起来。SynAVID 为开发沉浸式视听模型提供了资源；扩大场景覆盖并迁移至真实数据是未来工作的方向。

大语言模型使用声明

我们使用 OpenAI ChatGPT 和 Codex 支持论文写作与编辑、相关工作与引文准备、实验设计与结果分析、图表准备，以及数据集构建与注释。在这些工作中，作者通过交互式使用（生成）审阅、编辑或批准工具生成的输出。对于代码和实现，Codex 还在有限的逐步人工指导下自主执行多步骤任务（智能体）。对于研究构想、假设、主张和结论，AI 工具协助编辑或润色作者撰写的内容（辅助）。此外，我们还将论文的主要部分提供给 OpenAI（ChatGPT/GPT），以获取审稿人式反馈。无论是否使用大语言模型或 AI 工具，作者均对所有提交内容承担全部责任。

参考文献

[1] Jianhong Bai、Menghan Xia、Xiao Fu、Xintao Wang、Lianrui Mu、Jinwen Cao、Zuozhu Liu、Haoji Hu、Xiang Bai、Pengfei Wan 和 Di Zhang。ReCamMaster：基于单个视频的摄像机控制生成式渲染。载于 *IEEE/CVF*国际计算机视觉大会（*ICCV*）论文集，第 14834–14844 页，2025年。 2，3[2] Davide Berghi、Marco Volino 和 Philip J. B. Jackson。Tragic Talkers：用于视听机器学习研究的莎士比亚风格声场与光场数据集。载于 第*19*届 *ACM SIGGRAPH*欧洲视觉媒体制作大会论文集，2022年。 2[3] Chunhao Bi、Houqiang Zhong、Zhixin Xu、Li Song 和 Zhengxue Cheng。AudioGS：基于频谱图的音频高斯溅射，用于声场重建。 *arXiv*预印本 *arXiv:2604.08967*，2026年。 6，7[4] Manuel Brandner、Nils Meyer-Kahlen 和 Matthias Frank。用于增强声学现实的三角钢琴指向性模式测量。载于声学进展 – *DAGA*，2020 年。 4[5] Fabian Brinkmann 、Alexander Lindau、Stefan Weinzierl、Steven van de Par、Markus Müller-Trapet、Rob Opdam 和 Michael Vorländer。适用于不同头部相对于躯干朝向的高分辨率全球面头相关传递函数数据库。 音频工程学会期刊，65(10):841–848，2017年。 4

[6] Michael Broxton, John Flynn, Ryan Overbeck, Daniel Erickson, Peter Hedman, Matthew DuVall, Jason Dourgarian, Jay Busch, Matt Whalen 和 Paul Debevec。采用分层网格表示的沉浸式光场视频。 *ACM* 图形学汇刊，39(4):86:1–86:15，2020年。 2

[7]Changan Chen、Carl Schissler、Sanchit Garg、Philip Kobernik、Alexander Clegg、Paul Calamia、Dhruv Batra、Philip Robinson 和 Kristen Grauman。SoundSpaces 2.0：用于视觉声学学习的仿真平台。载于神经信息处理系统进展，2022年。 2，3

[8] Changan Chen、Alexander Richard、Roman Shapovalov、Vamsi Krishna Ithapu、Natalia Neverova、Kristen Grauman 和 Andrea Vedaldi。新视角声学合成。载于《IEEE/CVF计算机视觉与模式识别会议》（CVPR）论文集，第 6409–6419 页，2023 年。2、3、6、7[9] Mingfei Chen 和 Eli Shlizerman。AV-Cloud：通过视听云点溅射实现空间音频渲染。载于《神经信息处理系统进展》，2024年。6、7[10] Ziyang Chen、Israel D. Gebru、Christian Richardt、Anurag Kumar、William Laney、Andrew Owens 和 Alexander Richard。真实声场：视听房间声学数据集与基准测试。载于《IEEE/CVF计算机视觉与模式识别会议》（CVPR）论文集，第 21886–21896 页，2024年。2、3[11] Jacob Donley、Vladimir Tourbabin、Jung-Suk Lee、Mark Broyles、Hao Jiang、Jie Shen、Maja Pantic、Vamsi Krishna Ithapu 和 Ravish Mehra。EasyCom：一个增强现实数据集，旨在支持嘈杂环境下便捷沟通的算法。arXiv预印本 arXiv:2107.04174，2021年。2[12] Ruohan Gao 和 Kristen Grauman。2.5D视觉声音。载于《IEEE/CVF计算机视觉与模式识别会议》（CVPR）论文集，第 324–333 页，2019年。2[13] Philipp Götz、Georg Götz、Nils Meyer-Kahlen、Kyung Yun Lee、Karolina Prawda、Emanuël A. P. Habets 和 Sebastian J. Schlecht。用于盲估计能量衰减的多房间过渡数据集。载于《2024年第18届声学信号增强国际研讨会》（IWAENC），第 299–303 页。IEEE，2024 年。2[14] Kristen Grauman 、Andrew Westbury、Lorenzo Torresani、Kris Kitani、Jitendra Malik、Triantafyllos Afouras、Kumar Ashutosh、Vijay Baiyya、Siddhant Bansal、Bikram Boote、Eugene Byrne、Zach Chavis、Joya Chen、Feng Cheng、Fu-Jen Chu、Sean Crane、Avijit Dasgupta、Jing Dong、Maria Escobar、Cristhian Forigua、Abrham Gebreselasie、Sanjay Haresh、Jing Huang、Md Mohaiminul Islam、Suyog Jain、Rawal Khirodkar、Devansh Kukreja、Kevin J Liang、Jia-Wei Liu、Sagnik Majumder、Yongsen Mao、Miguel Martin、Effrosyni Mavroudi、Tushar Nagarajan、Francesco Ragusa、Santhosh Kumar Ramakrishnan、Luigi Seminara、Arjun Somayazulu、Yale Song、Shan Su、Zihui Xue、Edward Zhang、Jinxu Zhang、Angela Castillo、Changan Chen、Xinzhu Fu、Ryosuke Furuta、Cristina Gonzalez、Prince Gupta、Jiabo Hu、Yifei Huang、Yiming Huang、Weslie Khoo、Anush Kumar、Robert Kuo、Sach Lakhavani、Miao Liu、Mi Luo、Zhengyi Luo、Brighid Meredith、Austin Miller、Oluwatumininu Oguntola、Xiaqing Pan、Penny Peng、Shraman Pramanick、Merey Ramazanova、Fiona Ryan、Wei Shan、Kiran Somasundaram、Chenan Song、Audrey Southerland、Masatoshi Tateno、Huiyu Wang、Yuchen Wang、Takuma Yagi、Mingfei Yan、Xitong Yang、Zecheng Yu、Shengxin Cindy Zha、Chen Zhao、Ziwei Zhao、Zhifan Zhu、Jeff Zhuo、Pablo Arbelaez、Gedas Bertasius、Dima Damen、Jakob Engel、Giovanni Maria Farinella、Antonino Furnari、Bernard Ghanem、Judy Hoffman、C. V. Jawahar、Richard Newcombe、Hyun Soo Park、James M. Rehg、Yoichi Sato,

Manolis Savva、Jianbo Shi、Mike Zheng Shou 和 Michael Wray。Ego-Exo4D：从第一人称和第三人称视角理解熟练人类活动。载于《IEEE/CVF计算机视觉与模式识别会议》（CVPR）论文集，第 19383–19400 页，2024年。2、3[15] Wenxiang Guo、Changhao Pan、Zhiyuan Zhu、Xintong Hu、Yu Zhang、Li Tang、Rui Yang、Han Wang、Zongbao Zhang、Yuhan Wang、Yixuan Chen、Hankun Xu、Ke Xu、Pengfei Fan、Zhetao Chen、Yanhao Yu、Qiange Huang、Fei Wu 和 Zhou Zhao。MRSAudio：一个带精细标注的大规模多模态录制空间音频数据集。载于《神经信息处理系统进展》，2025年。2、3[16] 国际电信联盟。先进沉浸式视听系统使用场景及当前状态汇编。ITU-R BT.2420-1号报告，ITU，2020年。1[17] ISO/IEC JTC 1/SC 29/AG 3。MPEG沉浸式视频白皮书。白皮书 N0059，ISO/IEC，2022年。1[18] ISO/IEC JTC 1/SC 29/AG 3。MPEG-I沉浸式音频白皮书。白皮书 N150，ISO/IEC，2024年。1[19] Zeren Jiang、Yushi Lan、Yihang Luo、Yufan Deng、Zihang Lai、Edgar Sucar、Christian Rupprecht、Iro Laina、Diane Larlus、Chuanxia Zheng 和 Andrea Vedaldi。Syn4D：一个多视角合成 4 05207、2026 2605 2、3 D 数据集。arXiv预印本 arXiv: . .[20]Derong Jin、Xiyi Chen、Ming C. Lin 和 Ruohan Gao。SonoWorld4D：通过扩散同步，从单个视频生成4D视听场景。arXiv预印本 arXiv:2610.40640，2026年。2、3[21] Alexander Kirillov、Eric Mintun、Nikhila Ravi、Hanzi Mao、Chloe Rolland、Laura Gustafson、Tete Xiao、Spencer Whitehead、Alexander C. Berg、Wan-Yen Lo、Piotr Dollár 和 Ross Girshick。分割一切。载于《IEEE/CVF国际计算机视觉大会》（ICCV）论文集，第 4015–4026 页，2023 年。7[22] Zitong Lan 、 Yiduo Hao 和 Mingmin Zhao。SmartDJ：使用音频语言模型进行声明式音频编辑。载于《国际学习表征会议》，2026年。8[23] Lucas Yunkyu Lee、Soonho Kim、Youngwook Kim、Sang-min Kim 和 Jaesik Park。FreeTimeGS++：动态高斯点溅射的奥秘及其原理。arXiv预印本 arXiv:2605.03337，2026年。5、6、7[24] Tianye Li、Mira Slavcheva、Michael Zollhöfer、Simon Green、Christoph Lassner、Changil Kim、Tanner Schmidt、Steven Lovegrove、Michael Goesele、Richard Newcombe 和 Zhaoyang Lv。基于多视角视频的神经三维视频合成。载于《IEEE/CVF计算机视觉与模式识别会议》（CVPR）论文集，第 5521–5531 页，2022年。2[25] Yun-Jin Li、Mariia Gladkova、Yan Xia 和 Daniel Cremers。TRASE：无跟踪4D分割与编辑。arXiv预印本 arXiv:2411.19290，2024年。7[26] Zhan Li、Zhang Chen、Zhong Li 和 Yi Xu。用于实时动态视图合成的时空高斯特征点溅射。载于《IEEE/CVF计算机视觉与模式识别会议》（CVPR）论文集，2024年。5、6[27] Susan Liang、Chao Huang、Yapeng Tian、Anurag Kumar 和 Chenliang Xu。AV-NeRF：学习用于真实世界视听场景合成的神经场。载于《神经信息处理系统进展》，2023年。2、3[28] Xiulong Liu、Anurag Kumar、Paul Calamia、Sebastià V. Amengual、Calvin Murdock、Ishwarya Ananthabhotla、Philip Robinson、Eli Shlizerman、Vamsi Krishna Ithapu 和 Ruohan Gao。在任意环境中的任意位置听觉。载于 IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，第5732–5741页，2025年。2、3、4[29] Cheng-You Lu、Peisen Zhou、Angela Xing、Chandradeep Pokhariya、Arnab Dey、Ishaan Nikhil Shah、Rugved Mavidipalli、Dylan Hu、Andrew I. Comport、Kefan Chen 和 Srinath Sridhar。DiVa-360：用于沉浸式神经场的动态视觉数据集。载于 IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，第22466–22476页，2024年。2[30] Pranay Manocha、Adam Finkelstein、Richard Zhang、Nicholas J. Bryan、Gautham J. Mysore 和 Zeyu Jin。从刚可察觉差异中学习的可微感知音频度量。载于 Interspeech 论文集，第2852–2856页，2020年。7[31] Pranay Manocha、Zeyu Jin、Richard Zhang 和 Adam Finkelstein。CDPAM：用于感知音频相似度的对比学习。载于 IEEE国际声学、语音与信号处理会议（ICASSP）论文集，2021年。7[32] Chong Mou、Mingdeng Cao、Xintao Wang、Zhaoyang Zhang、Ying Shan 和 Jian Zhang。ReVideo：通过运动与内容控制重制视频。载于《神经信息处理系统进展》，2024年。7[33] Michal Nazarczuk、Thomas Tanay、Arthur Moreau、Zhensong Zhang 和 Eduardo Pérez-Pellitero。Charge：全面的新视角合成基准测试与数据集——将它们统一起来。载于 IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，第15323–15333页，2026年。2、3[34] Keunhong Park、Utkarsh Sinha、Peter Hedman、Jonathan T. Barron、Sofien Bouaziz、Dan B Goldman、Ricardo Martin-Brualla 和 Steven M. Seitz。HyperNeRF：用于拓扑变化神经辐射场的高维表示。载于《ACM图形学汇刊》，40（6），2021年。2[35] Albert Pumarola、Enric Corona、Gerard Pons-Moll 和 Francesc Moreno-Noguer。D-NeRF：用于动态场景的神经辐射场。载于 IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，第10318–10327页，2021年。2[36] Minhyeok Roh、Minseo Lee、Jun Young Jeong、Byeonghyeon Lee、Lucas Yunkyu Lee、Jaesik Park 和 Eunbyung Park。清除幻影：用于渐隐动态内容的高斯分配。arXiv预印本 arXiv:2610.21302，2026年。5[37] Neus Sabater、Guillaume Boisson、Benoit Vandame、Paul Kerbiriou、Frederic Babon、Matthieu Hog、Remy Gendrot、Tristan Langlois、Olivier Bureller、Arno Schubert 和 Valerie Allie。多视角光场的视频数据集与处理流程

视频。载于 IEEE计算机视觉与模式识别会议（CVPR）研讨会论文集，第30–40页，2017年。2[38] Vikram Sandu、Mayurdeep Pathak 和 Rajiv Soundararajan。用于动态3D高斯泼溅的时序感知稠密化。arXiv预印本 arXiv:2606.23212，2026年。5[39] Hao-Tang Tsui、Yu-Rou Tuan、Ethan Lai 和 Chen-Yu Wang。StaDy4D：借助 SIGMA 实现完整的4D静态—动态重建。载于 IEEE/CVF计算机视觉与模式识别会议（CVPR）研讨会论文集，第4697–4707页，2026年。2、3[40] Valve公司。Steam Audio。https : / / valvesoftware.github.io/steam-audio/。版本4.8.1。访问日期：2026年11月。4[41] Karn N. Watcharasupat 和 Alexander Lerch。超越四个声部的与声部无关的单解码器音乐源分离系统。载于国际音乐信息检索学会会议论文集，2024年。8[42] Zhengxian Yang、Shi Pan、Shengqi Wang、Haoxiang Wang、Li Lin、Guanjun Li、Zhengqi Wen、Borong Lin、Jianhua Tao 和 Tao Yu。ImViD：用于增强虚拟现实沉浸感的沉浸式体积视频。载于 IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，第16554–16564页，2025年。2、3[43] Luis Vitor Zerkowski 和 Luiz Velho。AudioWorldSim：面向世界模型的真实双耳音频数据集。arXiv预印本 arXiv:2608.21075，2026年。3[44] Junsheng Zhou、Zhifan Yang、Liang Han、Wenyuan Zhang、Kanle Shi、Shenkun Xu 和 Yu-Shen Liu。4C4D：4摄像机4D高斯泼溅。载于 IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，2026年。5、6[45] Shangchen Zhou、Chongyi Li、Kelvin C. K. Chan 和 Chen Change Loy。ProPainter：改进视频修复中的传播与变换器。载于 IEEE/CVF国际计算机视觉大会论文集，2023年。7