面向人到虚拟形象重建、基于话语与情感数据的精简面部数据采集

Seoyoung Kang Seokhwan Yang、Hail Song、BoramYoon、Jinwook Kim、Kangsoo Kim、Woontack Woo

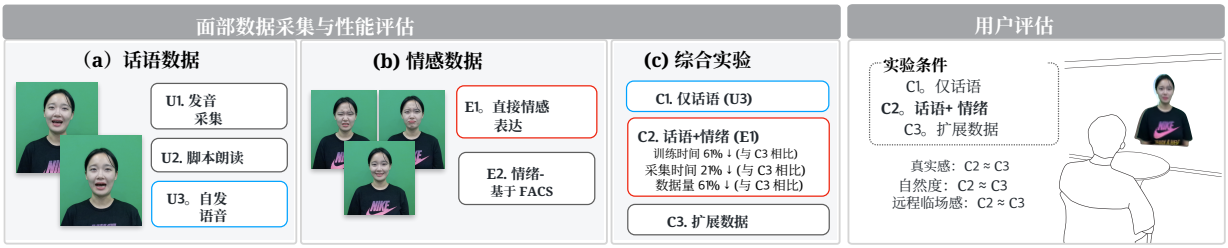

图 1：（左）通过识别关键表情来优化面部数据采集：将（a）话语数据实验中的自发语音和（b）情绪数据实验中的直接情感表达整合到（c）综合实验中，分别对应仅话语（C1）和话语 + 情绪（C2）。（右）用户评估模拟了真实的面对面交谈。

摘要——本研究探索一种面向对话情境的精简面部数据采集方法，以应对现有方法往往需要大规模数据集、并优先考虑技术指标而非用户感知与体验的局限。我们系统研究哪些面部表情数据对于重建照片级真实感虚拟形象至关重要，以及如何高效采集这些数据。本研究采用两阶段方法，以确定高效的数据收集策略并评估其有效性。在第一阶段，我们进行面部数据采集，并使用话语数据和情感数据评估重建性能。在第二阶段，我们开展全面的用户评估，比较三种递进条件：仅话语数据、话语数据与情感数据，以及涉及扩展数据的控制条件。来自24名参与模拟面对面对话的参与者的结果表明，有针对性的话语与情感数据在感知真实感、自然度和远程临场感方面达到了与扩展数据采集方法相当的水平，同时减少了训练时间和数据使用量。这些结果证明，有针对性的数据输入能够实现高效的虚拟形象面部重建，为AR/VR远程临场感等实时应用提供实用指导，并凸显数据量与感知质量之间的权衡。

索引词—虚拟化身，照片级真实感虚拟化身，面部重建，面部表情，面部数据采集，真实感，自然度，远程临场感，话语，情绪。

arXiv:2602.01729v1 [cs.HC] 2 Feb 2026

1 引言

随着增强现实（AR）和虚拟现实（VR）技术不断发展，三维化身已成为沉浸式环境的重要组成部分，使用户能够展现自我并保持身份认同 [25,42]。对逼真且富有表现力的虚拟形象的需求日益增长，尤其是在远程临场感应用中 [40,47]，例如 *Project Starline*[1]，在这些应用中，真实自然的交互至关重要。然而，当前生成照片级真实感虚拟形象的方法仍然计算密集型，通常依赖大量的面部数据采集 [32]。尽管近期方法能够实现较高的技术保真度，但它们并未充分回答：究竟需要哪些类型的面部数据，才能支持有效的会话交互。因此，重建过程通常通常需要大规模数据集和强大的硬件，因此对普通用户而言不太实用 [1,7]。

尽管准确性和相似度等技术指标 [22,49,53]对于评估虚拟形象至关重要，但它们无法全面反映用户体验。照片级真实感虚拟形象研究的一项关键空白，在于尚未充分了解数据收集策略的变化如何影响交流过程中的用户感知 [36]。出色的技术表现并不保证用户满意度；外观无懈可击的虚拟形象仍可能引起不适感，或触发恐怖谷效应 [38], , 最终降低参与度和沉浸感 [21]。本研究聚焦于对话虚拟形象重建所需的最少但在感知上足够的面部表情数据，考察两项关键输入：话语与情感数据，以应对这一挑战。我们不认为数据越多结果就一定越好，而是强调精简数据收集过程的优势。通过确定话语与情感输入的最佳组合，我们旨在证明，无需依赖大规模数据集或过多的计算资源，也能构建高质量的虚拟形象。

重要的是，我们的研究纳入了用户评估，因为我们认识到，仅凭技术指标不足以了解虚拟形象在现实世界场景中的效果。通过综合考量用户对真实感、自然度、沟通同步性和远程临场感的感知，我们确保虚拟形象不仅易于构建，也能让用户感到舒适并乐于参与。因此，我们的研究聚焦于以下问题：

研究问题1. 不同的话语数据采集方法如何影响对话情境中虚拟形象面部重建的准确性和效率？**研究问题2.** 哪种话语与情感数据组合能够实现

• *Seoyoung Kang* 就职于 *KAIST UVR*实验室。电子邮箱：*seoyoung.kang@kaist.ac.kr*。
• *Seokhwan Yang* 就职于 *KAIST UVR*实验室。电子邮箱：*ysshwan147@kaist.ac.kr*。
• *Hail Song* 就职于 *KAIST UVR*实验室。电子邮箱：*hail96@kaist.ac.kr*。
• *Boram Yoon* 就职于 *KAIST KI-ITC ARRC*。电子邮箱：*boram.yoon1206@kaist.ac.kr*。
• *Jinwook Kim* 就职于 *KAIST GSCT*。电子邮箱：*jinwook.kim31@kaist.ac.kr*。
• *Kangsoo Kim* 就职于卡尔加里大学。电子邮箱：*kangsoo.kim@ucalgary.ca*。
• *Woontack Woo* 就职于 *KAIST UVR*实验室和 *KAIST KI-ITC ARRC*。通讯作者。电子邮箱：*wwoo@kaist.ac.kr*。
稿件收到日期：*xx xxx. 201x*；录用日期：*xx xxx. 201x*。发布日期：*xx xxx. 201x*；当前版本日期：*xx xxx. 201x*。如需获取本文的转载件，请发送电子邮件至：*reprints@ieee.org*。数字对象标识符：*xx.xxxx/TVCG.201x.xxxxxxx*
[1]Google Project Starline：https://starline.google/ (访问日期：2026- 01-04)

既能保持较高真实感，又能最大限度降低数据采集和计算成本的虚拟形象？

RQ3. 不同的数据收集策略（仅话语、话语 + 情绪以及扩展数据）如何影响用户对虚拟化身自然度、真实感、同步性和远程临场感的感知？

为回答这些问题，我们设计了一项两阶段研究（图1），旨在确定实现高保真虚拟化身重建所需的关键数据。第一阶段通过比较不同类型的话语（U）和情绪（E）数据，系统地评估面部数据采集方式，以确定最高效的数据组合。第二阶段是一项针对RQ3 的验证研究，比较三种条件：**仅话语**（C1）、**话语 + 情绪**（C2），以及**广泛数据组合**（C3），以评估这些数据收集策略如何影响模拟面对面对话中的用户感知。研究结果表明，策略性地结合自发语音与直接情绪数据，其感知效果可与大规模数据集相当。通过减少数据需求而不牺牲真实感，我们证明，有针对性的方法比单纯收集更多数据更有效，这表明精简的数据收集策略能够有效支持高质量的化身通信。

2 相关工作

2.1 增强现实/虚拟现实中的远程临场感虚拟形象

沉浸式三维远程临场系统的最新进展显著增强了远程增强现实/虚拟现实用户参与共享环境的能力，支持跨越物理距离的无缝沟通与协作 [15, 23, 40]。在这些环境中，虚拟形象作为数字化身，使用户能够以第一人称视角进行实时互动 [3], , 从而实现远程临场感——即借助某种通信媒介，在某个环境中产生身临其境的心理感受 [48]。这些虚拟形象的视觉真实感和行为真实感对于沟通质量至关重要，因为它们是互动的主要媒介 [2, 29, 46]。早期研究引入了具有可自定义特征的通用化身 [8, 50],，但这些形象未能体现每位用户独特且逼真的面部特征。转向能够更准确反映现实外貌的虚拟形象带来了新的挑战，尤其是捕捉和处理面部表情所需的时间与精力。例如，Carrigan 等人 [6], 提出了 Expression Packing，这是一种高效的混合变形传输方法，能够尽可能减少所需的训练表情数量，同时仍支持个性化。

近年来的技术创新推动了高保真体积渲染技术的发展，使动态面部表情的采集更加高效 [33]。与此同时，高斯泼溅方法应运而生，能够以更高效率渲染复杂结构 [26, 37, 52]。设备端计算也取得了显著进展，使实时化身生成无需依赖强大的外部硬件。例如，*Apple Vision Pro* 的 *Persona*（测试版）[2] 等应用展示了集成系统如何让面向远程临场感的虚拟化身创建变得更加便捷 [34]。然而，尽管照片级真实感虚拟形象能够增强远程临场感中的沉浸感 [14], , 实时创建它们在计算上仍然要求很高。从头戴式摄像头采集并复现面部表情需要大量计算能力，而当前流程往往依赖冗长而复杂的数据采集环节。这些要求限制了可扩展性，也使虚拟化身对日常用户而言不够实用。

尽管取得了进展，数据采集阶段本身往往仍是瓶颈。许多方法仍忽视时间和复杂度等实际问题，而这些问题可能影响普通用户的采用意愿。我们的工作通过探索对话场景并引入简化数据捕获流程的技术来应对这一问题，从而缩短流程所需时间并降低其复杂度。通过简化这一阶段，我们的方法在保持真实感和表现力的同时降低了采用门槛，最终助力打造更实用、更易获得的基于头像的远程呈现体验。

[2]Apple Vision Pro 的 Persona：https://support.apple.com/guide/apple-vision-pro/dev934d40a17/visionos（访问日期：2026-01-04）

2.2 照片级真实感虚拟形象与面部重建

自三维可变形模型（3DMM） [5], 问世以来，照片级真实感虚拟形象面部生成取得了显著进展。该模型利用主成分分析（PCA）在低维子空间中表示面部外观和表情。通过将面部模型的复杂性简化为易于处理的向量，3DMM成为面部重建技术的基石；此后，研究者又将其扩展为FLAME等参数化头部模型 [31], , 纳入表情、形状和姿态，从而实现更全面的面部表示。神经辐射场（NeRF）彻底改变了虚拟化身重建，使得仅凭单目RGB视频等有限输入即可生成高质量三维场景 [16, 17, 54]。Gafni等人 [16] 提出了用于四维头像的动态NeRF，提升了运动和视角的一致性。Grassal等人 [17] 和 Zheng等人 [54] 进一步改进了基于隐式可变形模型的神经头部虚拟形象，提升了真实感和通用性。近期研究如高斯头部头像 [52] 和 HyperGaussians [43] 以更高的真实感推进了虚拟化身表示，而 RGBAvatar [30]、SplattingAvatar [44],和FlashAvatar [51] 等方法则致力于降低计算需求，同时保持虚拟形象的高保真度。

尽管取得了这些技术进步，但我们仍未充分了解这些改进如何影响用户感知。大多数研究关注准确性和相似度等技术指标 [22, 49, 53], , 却往往忽视这些指标如何转化为实际应用中基于头像的交互体验，例如感知真实感、自然度和表现力。保真度的细微提升未必能显著改善整体用户体验，导致技术性能高与用户期望之间存在脱节。我们通过聚焦日常对话场景来应对这一问题，在这些场景中，虚拟形象创建效率更为关键。通过优先采用话语和情感数据等必要输入，我们能够生成富有表现力、逼真的虚拟形象，同时降低计算需求。这种兼顾技术性能与用户体验的平衡，使我们的虚拟形象非常适用于可扩展的实时AR/VR应用，确保实现沉浸式交流，同时避免过度消耗资源。

3 方法

本研究通过确定能够平衡真实感、表现力和计算效率的必要输入，探讨如何优化照片级真实感虚拟形象面部重建的数据采集。我们采用两阶段设计：第一阶段评估不同话语与情感数据采集方法的表现，第二阶段比较用户对这些方法生成的虚拟形象的感知。通过这一设计，我们能够评估降低数据复杂度是否会导致虚拟形象质量出现可感知的差异，从而在不损害用户体验的前提下，找出具有成本效益的策略。

3.1 面部数据采集与评估

本研究聚焦于考察两项关键输入，即**话语数据** 和**情绪数据**，以确定哪些面部表情对对话虚拟形象至关重要。我们旨在测试不同的采集方法，并找出捕捉自然且富有表现力的面部运动的最有效技术。

3.1.1 话语数据采集

话语相关表情对于生成能够准确复现日常会话行为中面部动态的虚拟形象至关重要 [41]。话语数据是指在主动说话过程中采集的由语音驱动的面部运动，反映与音素产生相关的协调发音动态 [20]。这三种方法均包含头部和眼部运动的基线记录，以捕捉关键的会话动态。为探索不同的采集策略，我们测试了三种方法：

- **发音采集（U1）：** 参与者发出了韩语中的关键辅音和元音（例如“Ga-Na-Da-Ra”和“A-E-I-O-U”）。此方法考察了仅用核心发音动作构成的最小数据集是否能够生成逼真的对话虚拟形象。

• **脚本朗读（U2）：** 参与者朗读了预先编写的新闻文章脚本，以采集结构化且一致的语音模式。为确保所有参与者的内容一致，我们提供了统一脚本。这种受控方法有助于我们评估标准化数据是否能提高重建质量。

• **自发语音（U3）：** 参与者围绕自行选择的主题发表未经事先准备、内容丰富的演讲，呈现出更加生动且自然多变的面部表情。这种方法捕捉了对话的自然流动，旨在考察基于自由话题自发语音数据训练的虚拟形象能否捕捉现实世界中的多样性。

3.1.2 情感数据采集

为补充话语数据，我们开展了情感表达实验。每种话语方法都搭配两种情绪捕获方式——直接情感表达和基于情感的面部动作编码系统——最终形成六种数据采集策略：U1E1、U1E2、U2E1、U2E2、U3E1 和 U3E2。

• **直接情感表达（E1）：** 参与者表达基本情绪 [11], 时，以保罗·艾克曼的面部表情图像作为视觉参考，引导他们做出目标表情。所涉及的情绪包括快乐、悲伤、愤怒、恐惧、厌恶、惊讶和中性，以捕捉即时且直接的面部表情 [9]。

• **基于情感的面部动作编码系统（E2）：** 参与者根据 GIF动画的指引，重现与基本情绪相关的面部动作单元（AU） [10, 13], 。此方法针对眉毛、眼睛、鼻子、脸颊、嘴唇和下颌的细微肌肉运动，捕捉构成情绪表达基础的细腻动态。

3.2 综合实验

研究的第二阶段比较了基于阶段 1（见第3.1节）中最有效采集方法的三种条件。该实验使我们能够检验精简方法在技术评估和以用户为中心的评估中，是否能达到与数据密集型方法相当的表现。

• **仅话语（C1）：** 根据话语数据实验的结果，我们采用了捕捉自然面部表情效果最佳的方法。此条件评估了仅使用话语数据生成的虚拟形象，能否在对话交互中达到足够的真实感和自然度。

• **话语与情感（C2）：** 此条件结合了话语与情感数据实验中表现最佳的方法。我们检验了这种最优组合能否在保持计算效率的同时，提升虚拟化身表现。

• **全面数据组合（C3）：** 该条件涵盖了广泛的面部表情，超出了 C1 和 C2 中采用的针对性方法。它作为比较基线，帮助我们评估采集显著更多的数据是否能较精简的方法带来有意义的改进。

3.3 实验设备

数据使用 Microsoft Azure Kinect 采集[3], , 设备放置在参与者前方 0.6 米处，以捕捉清晰的面部运动（如图 2 所示）。虽然该设备支持 RGB 和深度数据，但为符合重建流程的要求，我们仅录制了 RGB 视频 [55]。所有视频最初均以 3840×2160 像素的分辨率和 30 fps 的帧率采集，之后在预处理过程中调整为 512×512 像素，以满足重建流程的输入要求。我们对摄像机进行了校准，以确保所有采集环节的数据捕获一致且准确。我们

[3]Microsoft Azure Kinect：https://learn.microsoft.com/en-us/azure/kinect-dk/（访问日期：2026-01-04）

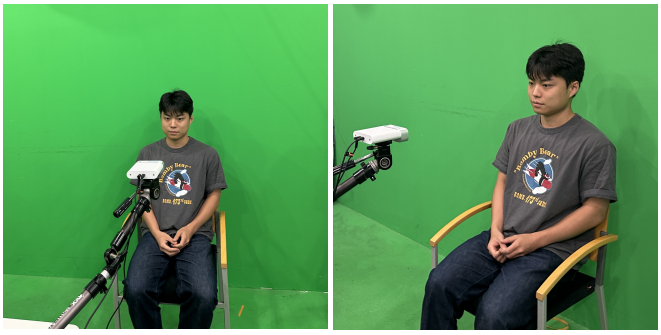

图 2：（左）数据采集设置，展示参与者正面朝向镜头的正视图；（右）侧视图。

录制了面部表情视频数据，并将其组合生成各个实验条件，随后利用这些数据从视频帧中重建虚拟化身的面部。此外，我们还录制了用于定量评估和用户研究的测试视频。这些视频呈现了模拟日常交流的对话场景，使我们能够评估虚拟化身的重建性能。

视频采集使用一台台式机完成，该设备配备 NVIDIA GeForce RTX 3090 GPU（24 GB 显存）和 Intel(R) Core(TM) i9-10900K CPU，专用于数据记录与存储。录制的视频随后传输至服务器，用于虚拟化身重建及后续处理。虚拟化身重建与渲染在一台配备 NVIDIA A40 GPU（40 GB 显存）和 Intel Xeon Gold 6326 CPU 的服务器上进行。在重建过程中，分配的 GPU 显存最高约占可用容量的 98%；系统内存使用量则较低，最多约使用了总计 256 GB 内存的 5%。

3.4 实验流程

在多种条件下采集数据集后，我们使用 PointAvatar [55], 一种基于神经点云的方法重建虚拟形象，该方法非常适合会话场景中的高保真表情建模。根据仅以 RGB视频为输入的 PointAvatar 框架，未记录深度数据。重建过程始于预处理，包括背景去除、分割和视频归一化。所有视频均被调整为 512×512 像素的分辨率，并将帧长度标准化，以匹配数据集中最长的序列。这些步骤确保数据既干净，又可在不同条件之间进行比较。为构建实验条件，我们利用 PointAvatar 的配置系统，在训练过程中系统地纳入或排除特定数据成分。通过有选择地纳入话语或情感数据，我们确保仅数据构成发生变化，而时间同步和计算环境保持一致，从而实现受控的消融实验，并保留自然的会话动态。

接下来，使用 FLAME [31] 进行参数化建模，从预处理后的帧中提取面部结构和表情。随后，将这些参数应用于标准点云，使其变形，从而复现丰富多样的面部运动。然后，将变形后的点云渲染成高质量图像。此阶段采用着色和细节保留技术来捕捉精细的面部细节，从而生成高质量的渲染输出。最后，将重建后的帧集成到 Unity 环境中。每个帧序列都导出为 .ply 文件，并设置为以每秒 30 帧的速度播放，以确保时间一致性。此步骤实现了流畅的动画和逼真的虚拟化身呈现，以便在对话场景中进行评估。总体而言，该流程为比较不同数据收集策略提供了统一而严谨的框架，涵盖定量分析和用户评估。

3.5 度量指标

3.5.1 性能评估

为了评估不同方法的性能，我们使用定量图像比较指标和计算成本度量指标，对重建质量和计算成本进行了评估。这种双重方法既能全面比较重建质量，也能评估各方法的效率与可行性。

在图像比较中，我们采用了四种广泛使用的度量指标：平均绝对误差（MAE）、结构相似性指数（SSIM） [49], 峰值信噪比（PSNR） [22], 以及学习感知图像块相似度（LPIPS） [53]。MAE 通过测量渲染图像与原始图像之间的平均绝对差异来反映像素级精度。SSIM 评估原始图像的结构细节保留情况，得分越高表示相似度越高。PSNR 以信噪比衡量图像质量，数值越高表示重建图像越清晰、噪声越少。LPIPS 基于深度学习特征评估感知相似度，得分越低表示感知上与原图越接近。

对于计算成本，我们测量了训练时间、数据采集时间和采集数据量。训练时间记录训练每个化身模型所需的时长。数据采集时间记录每种条件下面部视频数据的采集时长。采集数据量以捕获视频的总文件大小衡量，从而直观反映存储和处理需求。

3.5.2 用户评估

除性能指标外，我们还开展了一项用户研究，以收集采用不同数据收集方法生成的虚拟形象的定量评分和定性反馈。研究的主要目标是了解用户如何感知这些虚拟形象，并确定影响整体体验的因素。我们借鉴 Fraser 等人 [12], 的方法，评估了面部动画真实感，以衡量虚拟化身的面部表情和动作。我们使用该量表测量真实感（“虚拟化身的面部表情看起来很真实”）和自然度（“虚拟化身的面部表情看起来很自然”）。为便于更细致地分析用户感知，我们将真实感和自然度作为彼此独立的子维度进行分析。我们还评估了远程临场感 [39]，提出了“这段体验有多让您投入？”以及“在多大程度上，您觉得自己沉浸在所看到/听到的环境中？”等问题。最后，我们借鉴 Kang 等人 [24], 的方法，评估了交流线索同步性，考察虚拟形象的表情与预期对话信号之间的匹配程度。由于不同条件下虚拟形象的互动（刺激材料）均为单向且完全相同，我们排除了与沟通效能相关的问题，因为重复的语音可能会对其产生影响。因此，本研究重点关注自然度、真实感、交流线索同步性和远程临场感，探究虚拟形象在对话场景中是否具有吸引力、可信度和沉浸感。

4 研究一：面部数据采集与性能评估

按照第3.1节所述的数据采集阶段，我们使用 PointAvatar 重建了三维化身 [55]，以复现参与者的面部特征与面部表情。在每项实验中，所有序列均被重采样，以匹配最长序列的长度，从而确保各条件下的计算负载具有可比性。研究一的目标是确定生成逼真虚拟形象最有效的数据采集策略。因此，我们开展了两项实验：

- **话语数据实验：**测试三种在语音过程中采集面部表情的方法：发音采集（U1）、脚本朗读（U2）和自发语音（U3）。
- **话语与情感数据实验：**将每种话语方法与两种情感数据采集技术相结合：直接情感表达（E1）和基于情感的面部动作编码系统（E2）。由此得到六种评估条件（U1E1、U1E2、U2E1、U2E2、U3E1、U3E2）。

表1：话语数据实验中的性能评估，比较三种数据采集方法：发音采集（U1）、脚本朗读（U2）和自发语音（U3）。

被试	方法	平均绝对误差 ↓	峰值信噪比 ↑	结构相似性指数 ↑	学习感知图像块相似度 ↓
male_1	U1	0.0048	30.2	0.969	0.0434
	U2	0.0046	30.6	0.966	0.0511
	U3	0.0046	31.3	0.969	0.0425
女性 1_	U1	0.0104	24.4	0.958	0.0562
	U2	0.0092	25.2	0.963	0.0480
	U3	0.0083	26.0	0.966	0.0447
male_2	U1	0.0040	33.6	0.985	0.0111
	U2	0.0026	36.4	0.989	0.0093
	U3	0.0023	37.7	0.990	0.0086
女性 2_	U1	0.0070	28.2	0.972	0.0291
	U2	0.0059	29.5	0.975	0.0247
	U3	0.0051	30.3	0.980	0.0219

4.1 阶段 1：话语数据实验

表1报告了四名受试者在U1–U3条件下的重建性能（平均值 ±标准差），评估指标为平均绝对误差、峰值信噪比、结构相似性指数和学习感知图像块相似度。采用成对t检验考察不同方法之间的统计差异。总体而言，U3（自发语音）在所有指标上均 consistently表现最佳。其重建误差最低（平均绝对误差 = (0.0051 ± 0.0025)），表明准确性更高，且在参与者之间的结果更稳定。成对分析证实，U3的表现显著优于U1（$t = 3.72$, $p < 0.05$），而U3与U2之间的差异在统计学上不显著（$t = 2.24$, $p = 0.113$）。在图像质量方面，U3的峰值信噪比最高（峰值信噪比 = (31.33 ± 4.82)），显著优于U1（$t = -3.58$, $p < 0.05$），但与U2相比仅有微弱差异（$t = -2.48$, $p = 0.089$）。U3的结构相似性也最高（结构相似性指数 = (0.976 ± 0.010)），表明结构对齐效果更好，并且显著优于U1（$t = 4.15$, $p < 0.05$）。最后，U3的感知误差最低（学习感知图像块相似度 = (0.0294 ± 0.0161)），表明其与原始图像的感知匹配度最高。成对检验显示，U3显著优于U1（$t = -3.85$, $p < 0.05$），且相较于U2虽有微弱优势，但在统计学上并不显著（$t = -2.52$, $p = 0.088$）。

总而言之，U3 是最有效的话语方法，其准确性和感知质量显著高于 U1。与 U2 相比，U3 在所有度量指标上的表现也始终更强，不过差异并非总是达到统计显著性。

4.2 阶段2：话语与情感数据实验

表2展示了话语与情感数据采集方法的六种组合的结果。U3E1条件（自发语音 + 直接情感表达）整体表现最佳。其重建误差最低（平均绝对误差 $= (0.0049 \pm 0.0024)$），且显著优于U1E1（$t = 3.85, p < 0.05$）和U1E2（$t = 4.02, p < 0.05$）。尽管U3E1也优于U2E1和U2E2，但这些差异在统计学上并不显著。U3E1还取得了最高的峰值信噪比（31.65 ± 4.96），显著优于基于U1的条件（U1E1（$t = -3.82, p < 0.05$）和U1E2（$t = -3.95, p < 0.05$））， 并达到最高的结构相似性指数（结构相似性指数 $= (0.978 \pm 0.0101)$），相较于 U1 组合也有显著提升。其感知相似度同样最高（学习感知图像块相似度 $= (0.0282 \pm 0.016)$），明显优于U1条件，且略优于U2。总体而言，这些结果表明，将自发语音与直接情感表达相结合，是捕捉丰富、自然的面部动态的最有效策略。在所有客观指标上，U3E1均显著优于基于U1的方法，且相较于U2也呈现出明显的性能提升趋势。结果还表明

表2：六种数据采集方法下的虚拟形象面部重建性能。各方法将话语数据（U1：发音采集，U2：脚本朗读，U3：自发语音）与情感表达数据（E1：直接表达，E2：基于面部动作编码系统的情绪）相结合，形成六种组合（U1E1、U1E2、U2E1、U2E2、U3E1、U3E2）

主题	方法	平均绝对误差 ↓	峰值信噪比 ↑	结构相似性指数 ↑	学习感知图像块相似度 ↓
男性 1_	U1E1	0.0051	30.5	0.969	0.0410
	U1E2	0.0049	30.4	0.967	0.0466
	U2E1	0.0051	30.5	0.967	0.0434
	U2E2	0.0052	30.0	0.964	0.0496
	U3E1	0.0044	31.7	0.972	0.0397
	U3E2	0.0047	31.2	0.969	0.0443
女性 1_	U1E1	0.0093	25.1	0.963	0.0501
	U1E2	0.0099	24.7	0.960	0.0536
	U2E1	0.0086	25.7	0.967	0.0459
	U2E2	0.0087	25.6	0.965	0.0470
	U3E1	0.0080	26.3	0.968	0.0431
	U3E2	0.0085	25.8	0.966	0.0443
男性 2_	U1E1	0.0027	35.5	0.988	0.0103
	U1E2	0.0031	35.1	0.987	0.0102
	U2E1	0.0028	35.9	0.988	0.0090
	U2E2	0.0026	36.7	0.989	0.0087
	U3E1	0.0021	38.2	0.991	0.0079
	U3E2	0.0025	37.1	0.990	0.0087
女性 2_	U1E1	0.0066	28.5	0.973	0.0280
	U1E2	0.0068	28.3	0.973	0.0287
	U2E1	0.0060	29.4	0.975	0.0253
	U2E2	0.0061	29.2	0.974	0.0263
	U3E1	0.0050	30.4	0.980	0.0221
	U3E2	0.0054	30.1	0.979	0.0225

直接情感表达（E1）通常比基于FACS的表情（E2）效果更好，尤其是在与自发语音搭配时。

4.3 综合实验

我们还评估了各面部数据采集条件的计算成本：C1（仅话语）、C2（话语 +情绪）和 C3（广泛数据组合），所用数据为四位参与者数据的平均值。我们考量了三项与效率相关的关键指标：训练时间、采集时间和采集数据量。如表 3所示，结果显示三种条件在计算效率和数据效率方面存在差异。

C1 所需资源最少，训练时间为 34 分 56 秒，采集时间为 60.8 秒，数据量为 209.36 MB。C2 的训练时间为 44 分 24 秒，采集时间为 77.0 秒，生成的数据量为 264.11 MB。C3 涉及广泛的数据采集，资源消耗最多，训练时间为 1 小时 53 分 2 秒，采集时间为 197.5 秒，数据量为 682.25 MB。

5 研究二：用户评估

针对 RQ3，我们研究了不同面部数据采集方法如何影响人们在对话过程中的虚拟形象感知
互动，重点关注用户体验的四个关键方面：逼真度、自然度、交流线索同步性和远程临场感。虽然研究一（第 4 节所述）主要使用平均绝对误差、峰值信噪比、结构相似性指数和学习感知图像块相似度等指标评估技术性能，但仅凭这些度量指标无法反映用户在对话中如何感知虚拟形象。为此，我们开展了一项用户研究，从人的视角出发，收集了定量评分和定性

表 3：不同条件下计算成本的比较，数据取自四位参与者并取平均值。训练时间指训练模型所需的总时间。采集时间表示收集数据所需的时长，数据量则表示每种条件下收集的数据总量。

方法	训练时间	采集时间	数据量
C1	34分56秒	60.8秒	209.36 MB
C2	44分24秒	77.0秒	264.11 MB
C3	1小时53分钟2秒	197.5 秒	682.25 MB



图3：用户评估中使用的虚拟形象的 Unity 视图。（上）男性虚拟形象，（下）女性虚拟形象。（左）仅话语，（中）话语与情感，（右）广泛数据组合。

反馈，以评估参与者如何感知通过三种不同数据收集方法生成的虚拟形象。

5.1 研究设计

我们采用被试内设计，让每位参与者体验全部三种虚拟形象条件：

- **C1:** 仅话语，
- **C2:** 话语 +情绪，
- **C3:** 广泛数据组合。

这种方法能够直接比较不同虚拟形象，同时控制个体差异。参与者提供了定量评分和定性反馈，有助于了解各条件的优势与不足，并为未来虚拟形象的开发指明方向。

5.2 实验设置

本研究在沉浸式环境中进行，环境采用三星 Premiere 9 短焦投影仪在墙面显示屏上进行大尺寸投影（4K： 3840×2160，最大宽度：330厘米，投射比：0.189）。在充分利用显示区域时，投影面积约为 420 厘米 × 236 厘米，覆盖完整 4K 分辨率的 78.6%。为增强沉浸感，实验空间（4.2 米 × 4 米）用遮光窗帘围起，并完全遮暗。虚拟形象以真人大小显示，放置在距离参与者 2 米处，符合典型的社交距离 [19]。

为确保高保真虚拟形象拥有稳定的运行环境，我们选择使用墙面显示屏，而非头戴式设备，以保持分辨率和视觉质量的一致性，同时消除硬件导致的差异。尽管我们使用 PointAvatar 将虚拟形象重建为可变形的三维点云 [55], , 但为了确保视觉稳定性，我们以固定视角的序列呈现这些虚拟形象（图3）。听觉刺激方面，预录语音通过扬声器播放，并与面部动画同步。系统运行于一台配备 NVIDIA GeForce RTX 3090 GPU 和 Intel Core i9-10900K CPU 的台式机上。

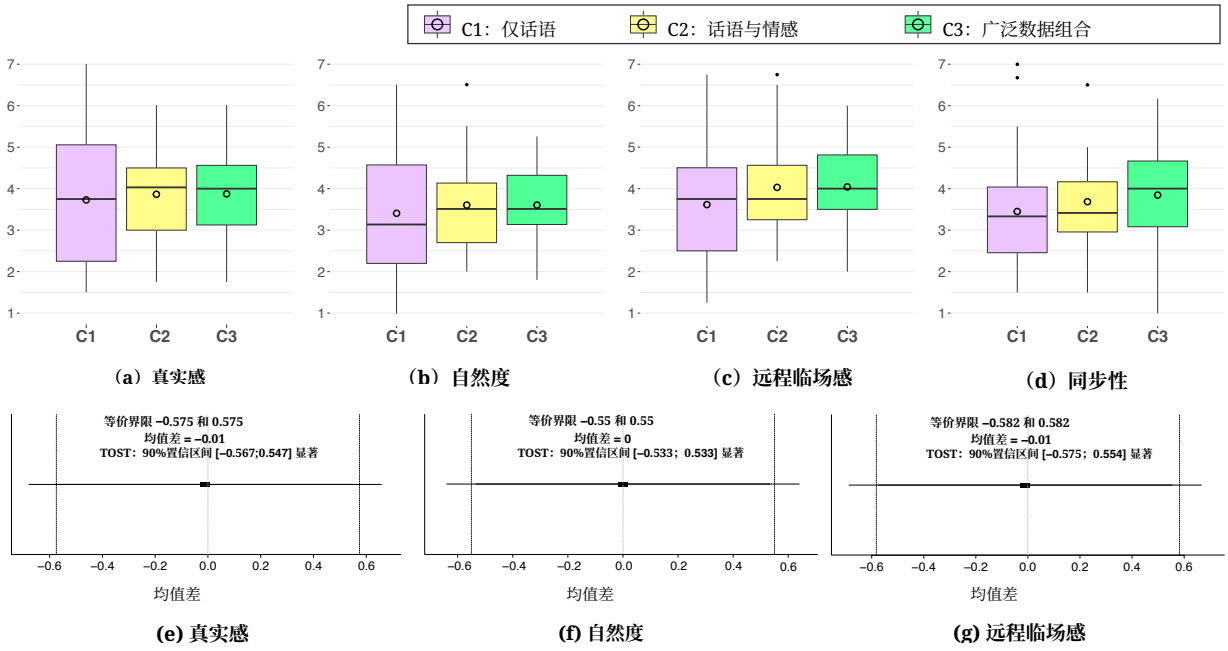

图 4：（上）李克特量表评分结果（1：强烈不同意–7：强烈同意），分别针对 (a) 真实感、(b) 自然度、(c) 远程临场感和 (d) 交流线索同步性。（下）比较 C2 和 C3 的等效性检验结果，分别针对 (e) 真实感、(f) 自然度和 (g) 远程临场感。

5.3 参与者

主要实验招募了另一组 24 名参与者（12 名男性和 12 名女性，平均值 ＝27.8，标准差 ＝4.36），以评估重建的虚拟形象。所有参与者均视力和听力正常或矫正至正常，以确保他们能够充分参与沉浸式环境及与虚拟形象的互动。所有参与者均自我认定为亚裔，这与用于虚拟形象重建的对象的人口统计背景一致，也确保了重建的虚拟形象与评估者之间的一致性。所有参与者报告了不同程度的既往虚拟现实/增强现实经验：13 名参与者（54.2%）参加过 10 次或以上相关活动，5 名参与者（20.8%）报告参加过 5–10 次，1 名参与者（4.2%）报告参加过 3–5 次，3 名参与者（12.5%）报告参加过 1–3 次，另有 2 名参与者（8.3%）没有相关经验。具体而言，22 名参与者（91.7%）有基于二维视频的远程协作经验，21 名参与者（87.5%）有基于虚拟化身的视频游戏经验，5 名参与者（20.8%）有基于三维的远程协作平台使用经验。每位参与者均完成了所有实验条件。本研究经 KAIST 机构审查委员会（IRB）批准 [KH2024-144]。

5.4 实验流程

到达后，参与者先了解实验概况，随后填写知情同意书。接着，他们在全部三种条件（C1–C3）下，分别与男性和女性虚拟形象进行对话式聆听任务。条件顺序采用平衡拉丁方设计进行随机化。每次交互后，参与者从真实感、自然度、沟通线索的同步性以及远程临场感等方面对虚拟形象进行评分。在完成全部六次会话（男性和女性虚拟形象 × 三种条件）后，参与者指出最喜欢和最不喜欢的虚拟形象，并说明选择理由。开放式回答收集了关于哪些方面感觉自然或不自然以及需要哪些改进的详细反馈。包括介绍在内的整个环节大约持续45分钟。

5.5 结果

我们采用两步法评估参与者在四项度量指标上的评分差异：真实感、自然度、同步性和临场感。首先，我们进行了弗里德曼检验 [45], 一种非参数方法，结果显示在参与者对三种条件的感知（$p > .05$，所有度量指标均如此）。然而，标准假设检验的非显著结果并不能证明各条件是等效的；它们仅表明缺乏足以支持存在差异的证据。为解决这一问题，我们使用双单侧检验（TOST）实验流程进行了等效性检验 [28] 以确定各条件之间的差异是否小到可以忽略。这对于判断不同的数据采集方法能否在数据和计算需求降低的情况下，仍生成被感知为同样真实且自然的虚拟形象至关重要。对于所有等效性检验，我们将等价界限（Δ）设定为 ±0.5（七点李克特量表）。该阈值代表半个量表刻度，我们认为这是用户感知中最小的有意义差异 [27]。任何落在该范围内的均值差均被视为可忽略，从而支持各条件在实践上等效的结论。TOST分析表明，C2（话语与情感数据）和C3（广泛数据组合）这两种条件在多项用户感知度量指标上具有统计等效性（见图 4）。

对于所有检验，之所以得出等价结论，是因为均值差的90%置信区间完全落在预先设定的等价界限之内。这一结果有力地证明，降低数据和计算需求（从C3到C2）不会对用户感知产生负面影响。

- **真实感：** C2（$M = 3.865$，$SD = 1.093$）与C3（$M = 3.875$，$SD = 1.202$）之间证明了等价性。TOST程序证实感知真实感的差异可忽略不计（$p_{TOST} < .05$）。均值差的90%置信区间 $[-0.567; 0.547]$ 完全落在 ±0.575的等价界限内，表明用户认为使用这两种数据收集方法生成的虚拟形象具有同等真实感。
- **自然度：** 在C2（$M = 3.594$，$SD = 1.158$）与C3（$M = 3.594$，$SD = 1.039$）之间观察到类似的等价性结果。TOST程序表明感知自然度的差异可忽略不计（$p_{TOST} < .05$）。均值差的90%置信区间 $[-0.533; 0.533]$ 落在 ±0.55的等价界限内，表明与更高效的C2相比，增加大量数据（C3）并未使化身自然度产生可感知的提升。
- **远程临场感：** 等价性也得到了确认，适用于远程临场感的比

较：C2（$M = 4.031$，$SD = 1.155$）与 C3（$M = 4.041$，$SD = 1.174$）。TOST程序表明差异可以忽略不计（$p_{TOST} < .05$）。均值差的 90% 置信区间 $[-0.575; 0.554]$ 位于等价界限 ± 0.582 之内。这表明，用户与使用数据量较少的 C2 创建的虚拟化身互动时，在远程临场环境中的沉浸感与使用更全面的 C3 创建的虚拟形象时同样强烈。

5.6 定性反馈

定性反馈深入揭示了参与者对虚拟形象面部表情和行为的体验，既指出了引人入胜的特点，也指出了需要改进之处。

5.6.1 正面反馈

参与者一致表示，当面部表情、身体动作和语音内容协调一致时，虚拟形象会显得更具吸引力，也更栩栩如生。这种同步性让虚拟形象看起来更像真人，避免了虚拟形象常见的人工感或呆板感。与语音情绪基调相匹配的细微面部变化（例如脸颊、眉毛或额头的变化）也颇受好评。

逼真的表情与动作 正面反馈的一个主要来源，是虚拟形象能够生成自然且符合语境的面部表情。参与者指出，虚拟形象的面部并非一成不变，而是会根据对话线索动态回应。例如，P7称赞虚拟形象“会根据对话在中性表情和温和的微笑之间切换，”并指出这种转换流畅而可信。其他参与者则称赞虚拟形象捕捉细节的能力。P5评论说，虚拟形象“准确地反映了甚至微小的肌肉运动，尤其是眼睛和鼻子周围的变化，”这增强了真实感。同样，P19指出“下半脸的肌肉和眼睛下方的肌肉都自然地动起来，让人感觉最为逼真。”参与者还提到了面部之外的非语言线索，包括模仿真实对话模式的头部和上半身动作。P19观察到“与语音相配的小幅前后动作让互动显得真实。”P5也指出“头部和上半身的动作增强了真实感。”P12甚至注意到虚拟形象的肢体语言传达了心理状态，并表示“倾听时，肩膀和头部会微微前倾，这让我感觉自己仿佛置身于对话之中。”

表情与语音内容之间的同步性 另一个获得积极反馈的关键方面，是虚拟化身能够使面部表情与语音内容相协调，从而增强情绪上的一致性。P8 表示，“虚拟化身的表情与说话的语气完美匹配，让我感觉自己像是在和真人交谈。”P14 赞赏道，“即使是简短的话语，也会引发细微但恰当的面部反应，”这突显了系统的响应能力和细腻程度。P20 提到，虚拟化身“在停顿或话题转换时，从不会显得僵硬或格格不入，”强调了表情与语音整合的流畅性和连贯性。

总体而言，参与者强调，精准的微表情、同步的头部和身体动作，以及能理解语境的面部反应相结合，营造出沉浸感强、引人入胜且令人信服、宛如真人的虚拟形象体验。这些结果表明，细致关注表情动态和行为同步，对于提升用户在虚拟化身媒介沟通中的感知至关重要。

5.6.2 有待改进之处

尽管有这些优点，参与者仍指出虚拟形象有三个明显不足之处：情感表达能力、面部协调和眼部动作。

情感表达能力与细微动作C1（仅话语）中的虚拟形象常被形容为僵硬或毫无表情。P6表示：“虚拟化身的脸看起来毫无反应，让它显得不那么真实”，P9也提到“变化很少，让人觉得它并没有真正投入对话。”参与者强调了细微情绪变化的重要性，例如停顿期间轻微的肌肉运动，有助于培养共情并提升自然度。P19解释道：“在语音停顿期间，面部肌肉的细微动作会让虚拟化身显得更加真实，就像在和真人交谈一样。”

面部协调与微表情 几位参与者强调，眼睛、脸颊和嘴部之间需要协调流畅，以避免表情显得割裂。眉毛抽动或嘴部线条变化等微表情，被认为对于传达细腻情绪尤为重要。P5强调，“捕捉情绪表达变化过程中面部肌肉细微的运动非常重要。”

眼部动作与视线 参与度和真实感的关键在于眼部行为，这一点被反复强调，也印证了先前关于照片级真实感虚拟形象的研究发现 [16, 55]。P22指出，“沟通感更多来自是否进行了眼神接触，而非面部肌肉运动的变化。”其他参与者（如P5、P18、P19）也强调，自然的眼神接触“对于营造虚拟形象仿佛真的在与我交谈的感觉至关重要。”参与者认为固定凝视显得不自然，P23指出，“眨眼和细微的视线变化等无意识动作，对于营造真实感至关重要。”P16补充道，“真人交谈时，会自然地关注对方的眼睛和嘴部。虚拟形象缺乏眼部动作，让人感觉不自然。”

恐怖谷风险 参与者提到，由于虚拟形象具有高度照片级真实感，即使细微的变形也可能令人不安。P23 描述道，“身体过度前倾会破坏面部比例，让人看起来很诡异。”P15 提到，故障会造成“卡通式变形”，还有几位参与者强调，面部表情稍显不自然，就可能迅速引发不适感。正如 P1 所解释的，“由于虚拟化身与真人十分相似，可能会引发恐怖谷效应，因此确保面部呈现尽可能接近真实人类的面部外观至关重要。”同样，P3 强调，“这不是普通的 *3D* 角色，而是照片级真实感的虚拟化身，因此面部表情变得更加重要。即使是最细微的不自然之处，也可能迅速引发与恐怖谷相关的不适感。”

6 讨论

6.1 研究发现与启示

面向虚拟形象生成的高效数据采集 本研究表明，照片级真实感虚拟形象生成并不一定需要大规模数据采集。聚焦于对话情境，我们发现，精选数据集能够生成兼具真实感和计算效率的虚拟形象。具体而言，自发语音（U3）与直接情感表达（E1）的组合在质量与效率之间实现了最佳平衡：与扩展数据收集方法相比，数据需求减少了约 61%，同时保持了相当的性能。这些发现挑战了全面数据集是生成高质量虚拟形象不可或缺的主流假设，并凸显了有针对性的情境特定数据收集的价值。

话语与情感数据采集的影响对于**研究问题1**（探讨话语数据采集方法对虚拟形象准确度和效率的影响），我们的结果表明，自发语音（U3）的表现始终优于发音采集（U1）和脚本朗读（U2）。U1的计算成本最低，但准确性也最低；相比之下，U3在输出质量与计算效率之间取得了良好平衡，因此成为在对话情境中采集话语数据最有效的方法。

对于**研究问题2**，为评估话语与情感数据的综合作用，我们评估了六种话语与情感数据组合（U1E1、U1E2、U2E1、U2E2、U3E1、U3E2）。分析结果显示，在各项关键指标上，直接情感表达（E1）与U3的组合优于其他方法（表2）。基于这些发现，

我们确定U3E1（自发语音 + 直接情感表达）是后续评估的最佳组合。

数据收集策略对用户感知的影响 针对 **RQ3**，我们比较了三种数据收集策略：仅采集话语数据（C1）、话语与 + 情绪数据（C2）以及广泛数据组合（C3）。等效性检验表明，尽管C2所需的数据和计算量显著更少，但由C2生成的虚拟形象在真实感、自然度和沉浸感方面的评价与C3生成的虚拟形象相当。这表明，超过一定阈值后，增加数据所带来的感知收益会递减，这挑战了以下普遍看法：[18, 56] 高质量的照片级真实感虚拟形象离不开大规模数据集。

从用户体验的角度来看，这些发现强调，真实感与数据集规模并非线性相关。参与者的评价表明，捕捉到核心对话表情后，额外细节几乎无助于提升感知自然度或远程临场感。相反，用户对能够传达及时、同步且符合情境的表情的虚拟形象反应更积极，而这可以通过针对性数据集实现。

综合来看，这些见解表明，优先考虑关键表情能够生成不仅在技术上可行、而且在社交上有效的虚拟形象。该方法对AR/VR远程临场感尤为重要，因为在此类场景中计算资源和带宽往往受限。通过在不牺牲感知质量的前提下强调效率，我们的框架支持跨平台和设备的可扩展实时部署，从而提升富有表现力的虚拟形象在日常交流中的可及性。

6.2 局限性与未来工作

尽管本研究提供了有价值的见解，但仍有若干局限性有待在未来工作中加以解决。首先，面部表情数据和用户评估仅限于单一文化群体。由于文化因素会强烈影响面部表情的产生与解读 [4, 35], ，这限制了我们研究结果的可推广性。因此，未来研究应纳入来自不同文化群体的数据，以评估我们方法的跨文化有效性，并提升虚拟化身表现的稳健性。其次，我们的用户评估采用了在各条件下使用相同刺激材料的单向对话聆听任务。尽管该设计确保了控制与一致性，但也限制了我们考察虚拟形象如何促进交互式交流的能力。未来研究应探索更具动态性、涉及交流的交互式场景，以更深入地理解虚拟化身的面部表情。此外，我们在模拟 *Project Starline*[4], 的大型墙面显示屏上开展了本研究，使用预先录制的Unity场景而非实时渲染。这一受控设置最大限度地减少了延迟和渲染延迟等技术问题，但并不能完全反映真实世界的XR条件。未来研究应在不同硬件配置下评估表现，并应对实时渲染、带宽限制和设备差异等约束。此类研究将阐明照片级真实感虚拟形象在实际AR/VR应用中能够以多高的效果进行部署。最后，我们计划准备该数据集的匿名化版本以供公开发布，从而实现更广泛的采用并促进可复现性。

7 结论

本研究表明，聚焦于关键数据输入，特别是话语和情感表达数据，可以在大幅降低数据需求的同时生成逼真的虚拟形象。通过针对与会话交互最相关的表情，我们简化了数据采集流程，在提升效率的同时不影响虚拟形象质量。本研究的一项重要贡献是识别出日常交流中的核心面部表情，从而简化重建过程，同时保持高度真实感和自然度。我们比较了三种数据采集方法，并评估了它们对计算性能和用户感知（包括真实感、自然度和远程临场感）的影响。结果表明，有针对性的数据采集能够在保持用户满意度的同时提升表现。通过结合定量指标与定性反馈，我们全面了解了虚拟化身在现实世界场景中的表现，并为在AR/VR环境中高效创建虚拟化身提供了实用指南。总之，我们简化的照片级真实感虚拟形象生成方法在不牺牲质量的前提下降低了数据需求，使更多设备和平台能够进行实时通信，并更便捷地使用高质量虚拟形象。这提升了打造更具沉浸感和吸引力的AR/VR通信体验的可行性。

[4]Google Project Starline：https://starline.google/ (访问日期: 2026- 01-04)

致谢

本研究由韩国国家科学技术研究会（NST）和信息通信技术规划评估院（IITP）通过元宇宙支援项目（旨在培养顶尖人才）提供支持，项目编号分别为 CRC21015 和 IITP-2025-RS- 2022-00156435，资金均来自韩国政府（MSIT）。此外，我们还感谢加拿大自然科学与工程研究委员会（NSERC） [RGPIN-2022-03294、ALLRP 590555-23]以及阿尔伯塔创新机构提供的支持。

参考文献

[1] Z. Bai, F. Tan, Z. Huang, K. Sarkar, D. Tang, D. Qiu, A. Meka, R. Du, M. Dou, S. Orts-Escolano 等。利用单目 RGB 视频学习个性化高质量体积式头部虚拟形象。载于 IEEE/CVF计算机视觉与模式识别会议论文集，第 16890–16900 页，2023年。1[2] J. N. Bailenson, A. C. Beall, J. Loomis, J. Blascovich 和 M. Turk。变换式社会交互：在协作式虚拟环境中将表征与行为及形式解耦。Presence: Teleoperators and Virtual Environments, 13 4 428 441, 2004 2 ( ): – 。S. Baker, J. [3] Waycott, R. Carrasco, R. M. Kelly, A. J. Jones, J. Lilley, B. Dow, F. Batchelor, T. Hoang 和 F. Vetere。社交虚拟现实中的虚拟化身媒介交流：深入探索老年人在新兴交流平台中的互动。载于2021年CHI人因计算系统会议论文集, 第 1 13, 2021 2 – 页。[4] C. Blais, R. E. Jack, C. Scheepers, D. Fiset 和 R. Caldara。文化塑造我们观察面孔的方式。PLOS ONE, 3 8 3022, 2008 8 ( ):e 。V. Blanz 和 [5] T. Vetter。用于合成三维人脸的可变形模型。载于《开创性图形学论文：拓展边界》，第2卷，第157–164页。2023年。2[6] E. Carrigan, E. Zell, C. Guiard 和 R. McDonnell。Expression Packing：以尽可能少的训练表情进行混合形状迁移。载于《计算机图形论坛》，第39卷，第219–233页。Wiley Online Library，2020年。2[7] C. Chen, M. O'Toole, G. Bharaj 和 P. Garrido。通过可控局部变形场进行隐式神经头部合成。载于 IEEE/CVF计算机视觉与模式识别会议论文集，第416–426页，2023年。1[8] T. Combe, R. Fribourg, L. Detto 和 J.-M. Normand。探索混合现实中虚拟化身头部对协作任务中的社会临场感、表现和用户体验的影响。《IEEE可视化与计算机图形学汇刊》，2024 2 。P. Ekman。《面部情感图 1976 3 [9] 片》。（无标题），。[10] P. Ekman 和 W. V. Friesen。面部动作编码系统。《环境心理学与非语言行为》，1978 3 。P. Ekman 和 D. Keltner。普遍 [11] 性情绪面部表情。《加利福尼亚心理健康研究文摘》，8 4 151 158, 1970 3 ( ): – 。A. D. [12] Fraser, I. Branson, R. C. Hollett, C. P. Speelman 和 S. L. Rogers。实时动作捕捉虚拟化身的表现力会影响感知到的动画逼真度，以及对虚拟现实中社会互动质量的感知。《虚拟现实前沿》，第 177, 2022 4 页。W. V. Friesen, P. [13] Ekman 等。Emfacs-：情绪面部动作编码系统。未发表手稿，加州大学旧金山分校, 2 36 1, 1983 3 ( ): 。Y. Fu, Y. Li, C. Li, J. Saragih, P. [14] Zhang, X. Dai 和 Y. C. Lin。Auto-CARD：用于实时移动临场感的高效、稳健编解码虚拟形象驱动。载于 IEEE/CVF计算机视觉与模式识别会议论文集，第 21036 21045, 2023 2 – 页。H. Fuchs, A. State 和 [15] J.-C. Bazin。沉浸式 D 远程临场感。《计算机》，47(7):46–52，2014年。2 3

[16] G. Gafni, J. Thies, M. Zollhofer 和 M. Nießner。用于单目 4D 面部虚拟形象重建的动态神经辐射场。载于 IEEE/CVF计算机视觉与模式识别会议论文集，第8649–8658页，2021年。2, 7[17] P.-W. Grassal, M. Prinzler, T. Leistner, C. Rother, M. Nießner 和 J. Thies。利用单目 RGB 视频生成神经头部虚拟形象。载于 IEEE/CVF计算机视觉与模式识别会议论文集，第 18653 18664, 2022 2 – 页。M.
[18] Habermann, W. Xu, M. Zollhofer, G. Pons-Moll 和 C. Theobalt。Deepcap：利用弱监督进行单目人体表演捕捉。载于 IEEE/CVF计算机视觉与模式识别会议论文集，第 5052 5063, 2020 8 – 页。E. T. Hall, R. L.
[19] Birdwhistell, B. Bock, P. Bohannan, A. R. Diebold Jr, M. Durbin, M. S. Edmonson, J. Fischer, D. Hymes, S. T. Kimball 等。人际距离学 [及评论与回复]。《当代人类学》，9(2/3):83– 108, 1968 5 。Z. S. Harris。《结构语言学》。
[20] 1963 2 。D. Higgins, D. Egan, R. Fribourg,
[21] B. Cowan 和 R. McDonnell。走出低谷：最先进的照片级真实感能否避免诡异感？载于 ACM应用感知研讨会 *2021*， 1 5, 2021 1 第 – 页。Q.
[22] Huynh-Thu 和 M. Ghanbari。PSNR 在图像/视频质量评估中的适用范围。《电子快报》， 44 13 800 801, 2008 1, 2, 4 ( ): – 。
[23] S. Kang, A. Nguyen, B. Yoon, K. Kim 和 W. Woo。感知虚拟形象面部与人际距离时的性别差异：在混合现实中探索真实感与社会临场感。载于2024年IEEE国际混合与增强现实研讨会（ISMAR），第101–110页，2024年。doi：10.1109/ISMAR62088.2024.00024 2[24] S. Kang, H. Song, B. Yoon, K. Kim 和 W. Woo。基于情绪的优先面部表情对虚拟形象媒介远程交流中社会临场感的影响。载于2024年IEEE国际混合与增强现实研讨会（ISMAR），第 1147 1156 2024 4 – 页。IEEE，。
[25] S. Kang, B. Yoon, K. Kim, J. Gratch 和 W. Woo。协作情境与人格特质如何塑造人到虚拟形象的身份表征之社会规范。《IEEE可视化与计算机图形学汇刊》， 2025 1 。B. Kerbl, G. Kopanas, T. Leimkühler 和 G. Drettakis。用
[26] 于实时辐射场渲染的3D高斯泼溅。《ACM图形学汇刊》， 42 4 139 1, 2023 2 ( ): – 。D. Lakens。等效性检验：针对 t 检验、相关分析和荟萃分析的实用入门。《社会心理学与人格科学》，8(4):355–362，2017年。6[28] D. Lakens, A. M. Scheel 和 P. M. Isager。心理学研究中的等效性检验：教程。《心理科学方法与实践进展》， 1 2 259 269, 2018 6 ( ): – 。S. U. Lee, J. Kim 和 J. Lee。奖
[29] 励计划与虚拟形象可见性对虚拟现实协作中共同能动性的影响。《IEEE可视化与计算机图形学汇刊》， 2023 2 。L. Li, Y. Li, Y.
[30] Weng, Y. Zheng 和 K. Zhou。RGBAvatar：用于头部虚拟形象在线建模的精简高斯混合形状。载于计算机视觉与模式识别会议论文集，第10747– 10757, 2025 2 页。
[31] T. Li, T. Bolkart, M. J. Black, H. Li 和 J. Romero。利用4D扫描学习面部形状与表情模型。《ACM图形学汇刊》，36(6):194–1，2017年。2, 3[32] S. Lombardi, J. Saragih, T. Simon 和 Y. Sheikh。用于人脸渲染的深度外观模型。《ACM图形学汇刊》（ToG），37 13，2018 4 1 1 ( ): – 。[33] S. Lombardi, T. Simon, G. Schwartz, M. Zollhoefer, Y. Sheikh 和 J. Saragih。用于高效神经渲染的体积基元混合。《ACM图形学汇刊》（ToG）， 40 4 1 13, 2021 2 ( ): – 。S. Ma, T.
[34] Simon, J. Saragih, D. Wang, Y. Li, F. De La Torre 和 Y. Sheikh。像素编解码虚拟形象。载于 IEEE/CVF计算机视觉与模式识别会议论文集，第 64 73, 2021 2 – 页。T. Masuda, P.
[35] C. Ellsworth, B. Mesquita, J. Leu, S. Tanida 和 E. Van de Veerdonk。将面孔置于情境中：面部情绪感知的文化差异。《人格与社会心理学杂志》，94 365，2008年 3 8 ( ): 。[36] A. Mollahosseini, B. Hasani 和 M. H. Mahoor。Affectnet：用于自然环境中面部表情、效价和唤醒度计算的数据库。《IEEE情感计算汇刊》， 10 1 18 31, 2017 1 ( ): – 。A. Moreau, J.
[37] Song, H. Dhamo, R. Shaw, Y. Zhou 和 E. Pérez-Pellitero。人体高斯泼溅：可动画虚拟形象的实时渲染。载于 IEEE/CVF计算机视觉与模式识别会议论文集 识别，页码 788 798, 2024 2 – 。M. Mori、K. F.
[38] MacDorman 和 N. Kageki。恐怖谷 [来自该领域] 19 2 98 100, 2012 1 。I
[39] EEE《机器人与自动化杂志》， （）： –。K. Nowak。能动性和拟人化对用户在虚拟环境中的远程临场感、共在感和社会临场感的影响，2003年。4[40] S. Orts-Escolano、C. Rhemann、S. Fanello、W. Chang、A. Kowdle、Y. Degtyarev、D. Kim、P. L. Davidson、S. Khamis、M. Dou 等。全息传送：实时虚拟三维传送。载于《第29届用户界面软件与技术年度研讨会论文集》，页码 741 754，2016 年 1, 2 –。[41] C. Pelachaud 、 N. I. Badler 和 M. Steedman。为语音生成面部表情。《认知科学》， 20 1 1 46, 1996 2 （）： –。A. Schäfer、
[42] G. Reis 和 D. Stricker。同步增强现实、虚拟现实和混合现实远程协作系统综述。ACM《计算调查》， 55 6 1 27, 2022 1 （）： –。G. Serifi 和 M.
[43] C. Bühler。HyperGaussians：用于高保真可动画面部化身的高维高斯泼溅。arXiv，2507 02803，2025年 2。.[44] Z. Shao、Z. Wang、Z. Li、D. Wang、X. Lin、Y. Zhang、M. Fan 和 Z. Wang。SplattingAvatar：采用网格嵌入式高斯泼溅的逼真实时人类化身。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，页码 1606 1616, 2024 2 –。
[45] M. R. Sheldon、M. J. Fillyaw 和 W. D. Thompson。在重复测量设计中分析有序尺度数据时Friedman检验的使用与解读。《国际物理治疗研究》， 1 228, 1996年 4 221 6（）： –。[46] M. Slater、B. Spanlang、M. V. Sanchez-Vives 和 O. Blanke。虚拟现实中身体转移的第一人称体验。PLOS ONE, 5(5):e10564，2010年。2[47] M. Stengel、K. Nagano、C. Liu、M. Chan、A. Trevithick、S. De Mello、J. Kim 和 D. Luebke。人工智能中介的三维视频会议。载于ACM SIGGRAPH *2023* 1 2 2023 1 新兴技术，页码 –。 。
[48] J. Steuer 等。定义虚拟现实：决定远程临场感的维度。《传播学杂志》， 42 4 73 93, 1992 2 （）： –。Z. Wang、
[49] A. C. Bovik、H. R. Sheikh 和 E. P. Simoncelli。图像质量评估：从误差可见性到结构相似性。IEEE《图像处理汇刊》， 13 4 600 612, 2004 1, 2, 4 （）： –。M. A.
[50] Witherow、C. Butler、W. J. Shields、F. Ilgin、N. Diawara、J. Keener、J. W. Harrington 和 K. M. Iftekharuddin。采用动态面部动作编码表情（CADYFACE）的可定制化身，可提升用户参与度。arXiv预印本 arXiv: *2403 07314*, 2024 2 。 。J. Xiang、X.
[51] Gao、Y. Guo 和 J. Zhang。FlashAvatar：通过高效高斯嵌入实现高保真头部化身。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，页码 1802 1812, 2024年 2 –。[52] Y. Xu、B. Chen、Z. Li、H. Zhang、L. Wang、Z. Zheng 和 Y. Liu。高斯头部头像：通过动态高斯实现超高保真头部化身。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，页码 1931 1941, 2024 2 –。R. Zhang、P. Isola、A. A. Efros、
[53] E. Shechtman 和 O. Wang。深层特征作为感知度量具有惊人效力。载于《IEEE计算机视觉与模式识别会议论文集》，页码 586–595，2018年。1, 2, 4[54] Y. Zheng、V. F. Abrevaya、M. C. Bühler、X. Chen、M. J. Black 和 O. Hilliges。IM Avatar：通过视频生成的隐式可变形头部化身。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，页码 13545 13555, 2022 2 –。Y. Zheng、W. Yifan、G.
[55] Wetzstein、M. J. Black 和 O. Hilliges。PointAvatar：通过视频生成的可变形点基头部化身。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，页码 21057–21067，2023年。3, 4, 5, 7[56] Z. Zheng、X. Zhao、H. Zhang、B. Liu 和 Y. Liu。Avatarrex：实时富有表现力的全身化身。ACM《图形学汇刊》(TOG)，42(4):1–19，2023年。8