arXiv:2601.01027v1 [cs.HC] 3 Jan 2026

交互式人工智能角色体验平台

RAFAEL WAMPFLER，苏黎世联邦理工学院，瑞士 CHEN YANG，苏黎世联邦理工学院，瑞士
DILLON ELSTE，苏黎世联邦理工学院，瑞士
NIKOLA KOVAČEVIĆ，苏黎世联邦理工学院，瑞士 PHILINE WITZIG，苏黎世联邦理工学院，瑞士
MARKUS GROSS，苏黎世联邦理工学院，瑞士



图 1. 我们的系统支持用户与数字角色开展对话和体验故事驱动型互动。用户可以通过语音与角色讨论任何话题，而角色则能自然地模仿人类行为。

从电影角色到现代科幻小说——让角色参与互动式、故事驱动型对话，几代人以来一直激发着人们的想象。实现这一愿景极具挑战性，所需的不仅仅是语言建模。它涉及许多复杂的人工智能难题，例如对话式人工智能、维护角色诚信、管理个性与情绪、处理知识与记忆、合成语音、生成动画、支持现实世界交互，以及与物理环境集成。基础模型、提示工程以及针对下游任务的微调方面的最新进展，使研究人员得以应对这些单独的挑战。然而，将这些技术整合起来以实现互动角色，仍然是一个尚待解决的问题。我们提出了一个系统和平台，便于设计可信的数字角色，使其能够提供对话式和故事驱动型体验，同时解决所有技术难题。作为概念验证，我们推出数字爱因斯坦，让用户能够与阿尔伯特·爱因斯坦的数字化形象对话，了解他的生平、研究和角色设定。虽然数字爱因斯坦 展示了我们针对特定角色所采用的方法，但我们的系统灵活通用，可适用于任何由故事驱动或以对话为主的角色。通过将这些不同的人工智能组件整合到一个易于调整的平台中，我们的工作为沉浸式角色体验铺平道路，让实现栩栩如生、以故事为基础的互动这一梦想成为现实。

CCS 概念： • **以人为中心的计算 → 交互技术**；
• **计算方法 → 话语、对话与语用学**；**程序化动画**； **动作捕捉**； 自然语言生成；语音识别。

其他关键词和短语：对话式人工智能、数字角色、具身对话代理、大语言模型、互动叙事、角色一致性、个性建模、记忆系统、语音驱动动画、语音合成、多模态交互

作者联系方式：Rafael Wampfler，苏黎世联邦理工学院，瑞士苏黎世，rafael.wampfler@inf.ethz.ch；Chen Yang，苏黎世联邦理工学院，瑞士苏黎世，chen.yang@inf.ethz.ch；Dillon Elste，苏黎世联邦理工学院，瑞士苏黎世，delste@inf.ethz.ch；Nikola Kovačević，苏黎世联邦理工学院，瑞士苏黎世，nikola.kovacevic@inf.ethz.ch；Philine Witzig，苏黎世联邦理工学院，瑞士苏黎世，philine.witzig@inf.ethz.ch；Markus Gross，苏黎世联邦理工学院，瑞士苏黎世，grossm@inf.ethz.ch。

SIGGRAPH 会议论文 *'25*，*2025*年*8*月*10*日至*14*日，加拿大不列颠哥伦比亚省温哥华©2025年版权归所有者/作者所有。本文为作者版本，仅供个人使用。不得转发。最终记录版本发表于 计算机图形学与交互技术特别兴趣小组会议论文（*SIGGRAPH* 会议论文集*'25*），*2025*年*8*月*10*日至*14*日，加拿大不列颠哥伦比亚省温哥华，https: //doi.org/10.1145/3721238.3730762。

ACM 参考格式：
Rafael Wampfler、Chen Yang、Dillon Elste、Nikola Kovačević、Philine Witzig 和 MarkusGross。2025年。交互式人工智能角色体验平台。载于计算机图形学与交互技术特别兴趣小组会议会议论文（*SIGGRAPH* 会议论文 *'25*），*2025*年*8*月*10*日至*14*日，加拿大不列颠哥伦比亚省温哥华。ACM，美国纽约州纽约市，21页。https://doi.org/10.1145/3721238.3730762

1 引言

创造能够进行有意义、故事驱动型对话的互动式逼真数字角色，这一愿景令几代人着迷 [Bates 1994；Cavazza 等，2002；Lugrin 等，2022]。从电影角色到历史人物的数字化呈现，这类角色重新定义了我们体验故事讲述的方式，并与数字实体建立情感联系 [Gong 等，2023；Torre 等，2019；Xu 等， 2024]。

然而，实现这一愿景是一项艰巨的任务。这需要将对话智能 [Ramesh 等人 2017]、角色完整性 [Schlenker 2008]、个性与情感 [Bates 1994]、知识与记忆 [Kope 等人 2013]、语音合成 [Torre 等人 2019]、逼真动画 [Wu 等人2024], 以及与物理环境的整合 [Li 等人 2022]无缝结合起来。由于其复杂性，正如迪士尼的“与 Crush 一起的海龟谈话”中那样，演员会实时操纵角色 [Casas 和 Mitchell 2019]。即使人工智能技术取得了重大进展，对话系统仍难以提供互动且由故事驱动的体验 [Green 和 Jenkins 2014；Riedl 等人 2003]。相关策略往往侧重于孤立的组件，例如动画合成 [Kim 等人 2024；Liu 等人 2022；Ng 等人 2024] 或语言建模 [Schmitt 和 Buschek 2021], , 而非确保所有组件整体协调一致。此外，角色一致性、定制化以及各组件实时同步等要求，往往也难以达到预期 [Lugrin 等人 2022]。

在本研究中，我们提出了一种模块化系统，用于创建可信的对话式数字角色，以支持叙事体验。该系统结合大语言模型（大语言模型）的强大能力、多模态感知、表现力合成和自适应个性建模，应对相关挑战，从而实现互动性强、以故事为驱动且可信的互动。作为概念验证，我们推出数字爱因斯坦，这是阿尔伯特·爱因斯坦的数字化形象，用户可以与其讨论他的科学研究、生平轶事和历史背景。该系统将以故事为驱动的人工智能角色融入物理环境中，营造沉浸式体验（见图 1）。虽然数字爱因斯坦 仅作为示例应用，但该系统架构具有高度模块化特性。可根据角色和具体目标应用轻松替换各个组件。因此，我们的研究为让互动且可信的数字角色栩栩如生开辟了新的可能性。

我们的系统融合了多项技术创新，使数字角色更加逼真。系统借助 GPT-4o 和经过微调的 Llama 3 模型维护角色完整性，并通过合成对话生成、基于嵌入的提示引导以及确保故事一致性的记忆系统加以增强。个性可动态调整，并通过语音和动画呈现情感基调。此外，系统还利用 Midjourney 生成的图像，为对话增添视觉内容。角色还会通过摄像头解读其物理环境，从而实现情境感知。这些组件集成于一个模块化且可扩展的平台，适用于多种对话和故事驱动型应用，并以主题化实体布置为基础，营造沉浸式且富有情感吸引力的交互体验。

2 相关工作

2.1 对话式数字角色

对话式数字角色已从基于规则的系统 [Mateas and Stern 2003] 发展到大语言模型驱动的模型 [Qi 等人 2021], , 从而能够进行动态且内容丰富的对话。个性建模方面的进步包括动态人格注入，使聊天机器人回复能够体现预设特质 [Kovačević 等人 2024a,b]。情绪感知系统通过语音情绪识别 [Hu 等人 2022] 和基于文本的情绪检测 [Kusal 等人 2024]提升用户参与度。此外，模块化架构有助于提高对话一致性和可扩展性 [Nguyen 等人 2022]。

对话式人工智能应用于教育、医疗保健和故事讲述等领域。叙事智能体通过创作的对话提升参与度 [Spierling 2005], , 而基于大语言模型的故事讲述则有助于构建连贯叙事 [Li 等人 2024]。这些系统简化了复杂任务并提升了可及性 [Qi 等人 2021]。然而，如何在多轮交流中维持有意义的互动仍是一项核心挑战。尤其是，解决记忆限制对于维持多轮连贯性至关重要 [Castillo-Bolado 等人 2024年；Johri 等人 2025]。为克服这些记忆限制，检索增强生成结合检索与生成，帮助聊天机器人维持长期上下文 [Gao 等人 2024]。双记忆系统通过平衡短期和长期数据，实现个性化 [Zhang 和 Luo 2024]，并保持流畅度；而选择性记忆则可改善用户体验和检索效率 [Sumida 等人 2024]。

2.2 人工智能驱动的叙事与伦理考量

早期方法（例如用于重现爱因斯坦的半自动艺术创作流程）展示了如何以有限资源实现小规模制作的真实感 [Helzle and Goetz 2018]。在此基础上，「Living Memories」概念中的人工智能驱动的对话智能体让列奥纳多·达·芬奇等人物栩栩如生 [Pataranutaporn 等人 2023]。更大规模的工作，例如开发用于扮演中国历史人物的语料库，凸显了情境真实性和低资源数据整合对于细腻刻画的重要性 [Bai 等人 2024]。与此同时，伦理考量日益受到重视。有关「数字招魂」的研究探讨了如何在保存文化遗产与解决真实性和同意问题之间取得平衡 [Hutson and Ratican 2023]。近期研究表明，大语言模型能够生成简洁的历史人物描绘，从而提升数字人文的可及性 [Hasnain and Usman 2024], ；而伦理框架则为教育领域的叙事重构提供指引 [Hutson 等人 2024]。

2.3 交互系统

交互系统推动了对话角色的发展，使互动更加逼真且引人入胜。近期的框架将身体动作与伴随语音的手势相协调，从而生成情感丰富且具备上下文感知能力的回应 [Kim 等人 2024]。模块化架构则通过将对话管理与具身化解耦，进一步支持此类互动，实现定制化和稳健的非语言交流 [Santos 等人 2023]。其他端到端流程则通过实时音视频同步和拟人化特征增强虚拟智能体 [Rupprecht 等人 2024]。

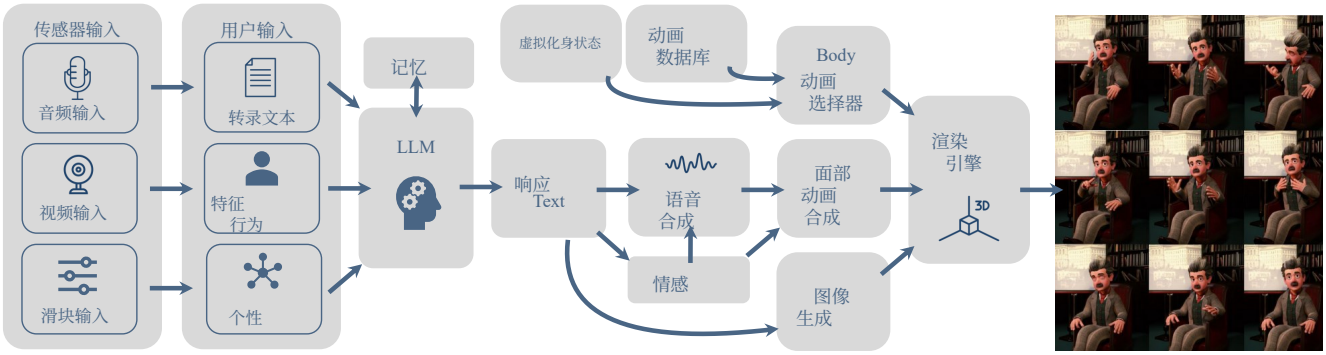


图 2. 系统概览：该流程通过基于记忆和可调节数字角色个性的大语言模型聊天机器人，处理传感器输入，包括转录语音、基于视频的用户特征和行为分析。聊天机器人的响应将引导语音和面部动画合成（根据响应中检测到的情绪），并选择基于虚拟化身状态的动作捕捉身体动画，同时进行图像生成。

沉浸式增强现实和混合现实系统也受益于这些进展。将语音识别与实时面部动画相结合的系统，提升了增强现实中的角色交互 [Casas 和 Mitchell 2023]。MoodFlow [Casas 等人 2024] 进一步运用嵌入提示的状态机，引导混合现实中的情感智能虚拟化身。整合视觉与语言模型的平台可实现上下文感知的实时交互 [Maniatis 等人 2023]，，而混合系统则通过不显眼的空间沉浸式界面，支持流畅无缝的用户体验 [Encarnacao 等人 2000]。

与以往工作不同，我们通过一个统一且可定制的框架，应对如何让可信角色参与故事驱动型互动对话这一人工智能挑战。

3 系统设计与需求

3.1 系统需求

我们的系统开发遵循若干目标（即可信度、灵活性、真实感）和约束条件（即低延迟、韧性、技术复杂性），这些共同定义了系统需求：

(1) 模块化与可扩展设计： 支持针对不同情境轻松升级和调整。(2) 以用户为中心的方法： 确保交互直观且可定制，允许用户调整角色的个性，并根据个人偏好定制对话。(3) 实时响应能力： 确保所有组件均保持低延迟，从而实现流畅的对话。(4) 稳健性与可靠性： 确保系统在不同环境下，即使条件各异，也能平稳运行。(5) 沉浸式体验： 将主题化的物理布置、空间音频、逼真动画以及与人工智能生成的响应同步的拟真身体动作相结合，从而确保自然且引人入胜的交互。

3.2 系统概览

我们的系统由多个相互连接的人工智能模块组成，以应对交互式故事驱动角色所面临的复杂人工智能挑战。图 2 概述了各模块之间的连接关系。图 8 展示了详细的交互流程。

系统的核心组件基于 Unity 实现，包含主题场景中的一个数字角色。该角色会在四种不同状态间切换：空闲 （无交互）、聆听 （等待用户输入）、思考 （处理输入）和 说话 （传达响应）。检测到用户靠近时，角色会播放欢迎动画（我们的系统每 500 毫秒查询一次摄像头）。用户坐下后，角色会从空闲 切换到 说话，以随机选取的欢迎语开启对话，随后等待用户输入。在思考状态下，转录后的用户输入由基于大语言模型的聊天机器人支持的认知模块进行处理。系统通过保持行为和对话的一致性，确保角色完整性。此外，角色还会利用其知识库和记忆，提供符合上下文的响应。用户还可以通过调整实体滑块来进一步定制体验，这些滑块会动态影响角色的响应模式。系统会从聊天机器人的响应中提取情绪，以调整语音和动画。语音由经过微调的 Microsoft Azure 神经语音模型合成。角色的动画融合了面部表情（通过 Audio2Face [Karras 等人 2017], 根据语音动态生成并同步）和动作捕捉的身体动作。角色处于说话状态时，系统会根据对话上下文使用 Midjourney 自动生成图像。在整个交互过程中，角色会在聆听、思考和说话状态之间交替切换。系统根据用户行为和自身响应协调状态转换，从而保持对话流程流畅。如果用户保持沉默超过七秒，系统就会向聊天机器人发送信号，提示其作出恰当响应。

4 设计挑战与解决方案

本节将讨论开发可信、善于对话且由故事驱动的角色系统时遇到的关键挑战，以及相应的解决方案。

4.1 对话式智能聊天机器人

打造令人信服的数字角色需要先进的对话能力。核心挑战在于开发能够

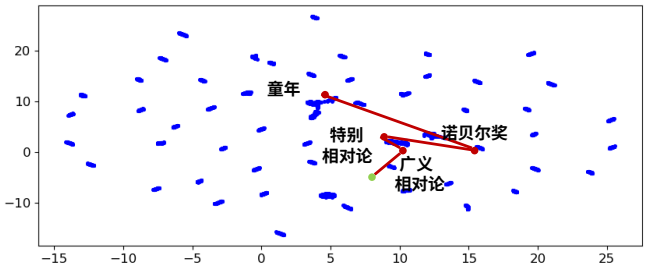


图 3. 合成爱因斯坦对话的嵌入空间可视化。蓝色点簇代表不同主题。红色标出了用户交互轨迹，起点以绿色标记。用户发起了从“诺贝尔奖”到“童年”的主题转换。

处理自然语言、生成符合语境的响应，并维持连贯且由故事驱动的互动，从而增强用户的沉浸感。为应对这一挑战，我们采用最先进的大语言模型，具体来说是 OpenAI 的 GPT-4 [2023] 3 o 和 Llama 8 B [Touvron 等人 2023]。GPT-o 用于确保实时的主题一致性和高质量的互动，而 Llama 3 支持需要隐私保护和成本效益的本地部署场景，因此是颇具价值的离线替代方案。这种双模型配置增强了稳健性和实时响应能力，并确保在云服务中断期间保持连续运行，从而满足要求 (3) 和 (4)。

角色故事与主题。 我们汇总了 M 个与角色的专业知识和个人兴趣相关的主题，例如爱好和轶事。对于爱因斯坦角色，这个集合总共包含 $M = 62$个主题，涵盖他的个人生活、科学理论和音乐兴趣。我们使用 GPT-4o 为每个主题生成了 N段合成的人类与角色对话（参见补充材料）。每个对话轮次 t 都使用 Microsoft Azure 的*text-embedding-3-large* 模型进行嵌入，生成一个 3072 维向量$\mathbf{e}_t$。为了计算每段对话的代表向量，我们对其各轮次的嵌入取平均值：$\mathbf{e}_{\text{conv}} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{e}_t$，其中 T是对话中的轮次数。为了获得主题级表示，我们进一步对该主题下所有对话的嵌入取平均值：$\mathbf{e}^{(j)}$主题 $= \frac{1}{N}\sum_{i=1}^{N}\mathbf{e}_{\text{conv}}^{(i,j)}$， $j = 1, \ldots\ldots, M$。图 3 展示了一个二维 UMAP 投影，其中每个点对应$\mathbf{e}_{\text{conv}}$。每个主题都形成了明显的簇，主题相关的各簇彼此位置接近。红色轨迹展示了用户在嵌入空间中跨多个主题的对话路径。

大语言模型。 我们使用合成对话的原始文本对 Llama 3 8B 进行微调，以增强其主题一致性以及对角色特定主题的了解。我们使用 Axolotl [Axolotl Project 2025]，采用 Adam 优化器（学习率为 $2e^{-5}$、梯度累积步数为 8、余弦调度，预热步数为 100），以批量大小 1 在基于云的 NVIDIA RTX 6000 GPU 上进行了三个周期的微调。为高效运行并满足要求 (3)，Llama 3 8B 使用 DeepSpeed [Rasley et al. 2020] 在本地运行。

对于 GPT-4o（Microsoft Azure，版本 2024-08-06），我们使用提示来保持主题一致性。系统会跟踪对话

表 1. 聊天机器人提示的组成元素说明及响应示例。

提示元素	说明和响应示例
位置	当前位置 温哥华在 *SIGGRAPH* 上洋溢的活力，正是 齐聚于此、共同塑造创新未来的好奇心灵的完美写照。
场景	物理布置和 Unity 场景的描述 这个温馨的环境非常适合进行深入而富有思考的对话，仿佛 我们正在一座历久弥新的图书馆里，一边品茶一边分享想法。
用户描述	年龄、性别、周围人数、用户外貌、用户注意力 你穿着蓝色短裤和白衬衫，看起来已准备好踏上冒险之旅，简直再合适不过 适合像你这样好奇的头脑！
图像描述	显示的图像的元数据 旋转的恒星残骸和变慢的时钟，完美展现了 黑洞如何扭曲光与时间，真是宇宙奇观！
Date	日期和时间 啊，快到午夜了，这是一个宁静美好的时刻，思绪可以自由漫游 并探索那些最奇妙的想法！
记忆	过往对话中的 5 个轮次
指令	角色设定、语气、回复指南、行为、特定情 况的处理（例如沉默）、话题转换。
附加数据	当前对话历史

流畅度，并利用嵌入空间进行处理。每个新轮次都会被嵌入，并与同一会话中此前的轮次结合，通过半衰期为三轮的指数加权平均（$h = 3$）赋予近期轮次更高的权重：$\mathbf{e}_{\text{agg}} = \sum_{k=0}^{\infty}\alpha_k\mathbf{e}_{t-k}$，其中 $\alpha_k = \frac{1}{Z}\cdot 2^{-k/h}$ 且 $Z = \frac{1}{1-2^{-1/h}}$。这里，$\mathbf{e}_{t-k}$是向前追溯 k 步的轮次嵌入。随后，系统计算该聚合嵌入与所有主题嵌入 $\mathbf{e}^{(j)}$topic $j = 1, \ldots, M$之间的余弦相似度。如果最高相似度超过设定的阈值，则选择相应主题 j作为当前主题。为实现平滑的主题转换，系统会在$\mathbf{e}_{\text{agg}}$与所有主题嵌入 $\mathbf{e}^{(j)}$topic $j = 1, \ldots, M$之间执行邻居搜索。系统会随机选择三个最近邻之一作为下一个主题。随后修改提示，引导 GPT-4o 平滑过渡到下一个主题。

对于这两种模型，提示词首先定义数字角色的角色设定，然后列出六个不同的上下文，接着给出指令和对话历史（见表1和补充材料）。

知识与记忆。 我们的系统采用基于向量的知识存储库来维护对话历史。对于每条新的用户输入，系统都会生成一个嵌入，并针对已存储的对话执行相似度搜索。最相关的五段对话片段会被纳入提示词的记忆上下文中。通过聚焦于与当前对话上下文最相关的主题和信息，这种方法满足了要求（3）。

4.2 个性与情感

除了基本的对话能力之外，一个引人入胜的数字人物还必须掌握人类式互动的多个维度。其中关键的一点，是在适应不同对话情境的同时保持一致的个性特征，并通过恰当的响应展现情商。为满足要求（5），我们将情商整合进系统中。



图 4. 用于实时调整五种特质的滑块：**活**力、**尽**责性、**正**派、**人**工性和**神**经质。

大语言模型可以根据用户的输入动态调整响应的情感基调 [Chang 2024；Jin et al. 2024]。因此，在生成响应后，我们使用 GPT-4o-mini 从与 Audio2Face 支持的情绪相匹配的 7 种情绪中确定一种（即惊讶、愤怒、厌恶、恐惧、喜悦、悲伤、中性）。此外，系统还会回归出一个介于 0.01 到 2.0 之间的强度等级。这些参数会传递给 Microsoft Azure 语音合成，后者根据以下映射调整说话风格：惊讶→ excited，愤怒 → angry，厌恶 → disgruntled，恐惧 → fearful，喜悦 → cheerful，悲伤 →sad，中性 → default style。

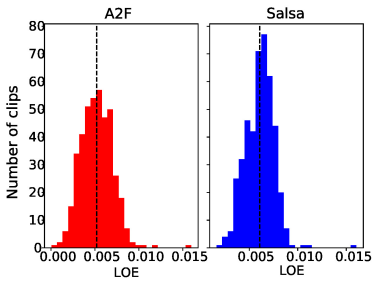
聊天机器人的个性与人类的个性有着根本差异 [Kovačević 等人 2024b]。我们采用一种独立的动态人格注入方法 [Kovačević 等人 2024a]，，使用 GPT-4o 重写大语言模型的响应，使其符合预定义的个性档案。因此，个性档案由五个关键维度构成（活力、尽责性、正派、人工性和神经质），每个维度采用 5 级强度量表。为了让用户能够轻松控制个性，我们使用电位器、Arduino 和定制的 3D打印外壳制作了实体滑块（见图 4），可直接调整提示词。

4.3 语音识别与合成

逼真的类人交互要求数字角色能够有效地处理和生成口语交流。我们的系统通过 Microsoft Azure 语音服务实现聆听和说话。语音识别已集成到 Unity 中，以最大限度地降低延迟，并保持动画与状态转换同步，从而满足要求 (3)。系统设置了 1 秒缓冲区，以便在结束识别过程前容纳用户语音中的自然停顿。转录后的文本会显示在场景的白板上，并转发给大语言模型以生成响应。我们的系统采用 Microsoft Azure 的定制神经语音模型进行语音合成，能够在角色的响应中传达各种情感基调和强度。定制神经语音模型还支持语音定制，可通过微调使其符合特定的主题或风格要求。

4.4 动画合成

在我们的系统中，面部动画由基于云的 NVIDIA Audio2Face（A2F）根据合成音频动态生成。这是一种数据驱动方法，经过训练，可使音频信号与面部动作对齐



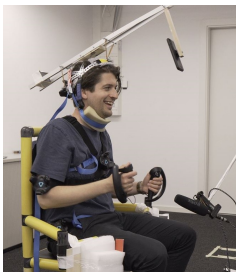


图 5。左：重定向后的动作捕捉序列与使用 Audio2Face 根据对应音频合成的动画之间的 LOE 分布（红色，$\mu = 0.005$；蓝色，$\mu = 0.006$）。右：动作捕捉设备，包括八个位置追踪器、两个手势追踪器，以及安装在云台头盔上的完整面部追踪器。

。由于音频主要包含下半脸动作的线索，研究人员在数据驱动动画模型中加入了静态 [Daněček 等，2023；Peng 等，2023；Wu 等，2024] 和动态 [Witzig 等人2024] 情绪表征，以生成生动的上半脸动作。A2F 最多支持 10 种预定义情绪标签。为确保渲染出的面部表情与合成语音相匹配，我们在动画合成过程中使用第 4.2 节中得出的相应情绪标签。

为满足要求（3），我们为 A2F 开发了一个 Python 包装器，以 0.5 秒为窗口增量处理音频。该包装器直接从 A2F 流式传输结果，绕过其默认的批处理方式。如果 A2F 不可用，我们会使用 SALSA LipSync Suite v2。虽然它的绝对唇部偏移误差（LOE）高于 A2F（见图 5 左侧），但能确保系统在云服务停机期间仍具有稳健性，从而满足要求（4）。此外，我们以程序化方式为眼睛制作动画：将用户的头部设为注视目标，并生成扫视眼动。

尽管 Audio2Gesture [Li 等人 2021] 提出的模型能够自动合成手势，但由于身体比例不匹配，将其重定向到我们的风格化角色会导致动作不自然，因此需要大量手动调整。我们改为使用精选的动作捕捉片段图书馆，以程序化方式为虚拟形象制作动画，并根据虚拟化身状态分类（空闲、说话、聆听和思考）。每种状态都配有专属的一组片段，并根据虚拟形象当前的状态随机选取一个新片段。如果片段结束时状态仍未改变，系统就会从同一类别中随机选取一个新片段。图 5（右侧）展示了我们用于面部和身体动画的定制动作捕捉装置（详情请参阅补充材料）。面部动作捕捉仅用于评估，不属于我们系统的一部分。

4.5 与物理世界的交互

为了与物理世界进行逼真的交互，我们的数字角色必须感知周围环境，并对用户的非语言行为做出回应。我们通过摄像头实现了这一能力，摄像头充当角色的“眼睛”，使其能够检测用户是否在场、用户特征以及用户行为。

将每一帧与空椅子的参考图像进行比较，并使用结构相似性指数（SSIM）进行评估。如果 SSIM 超过

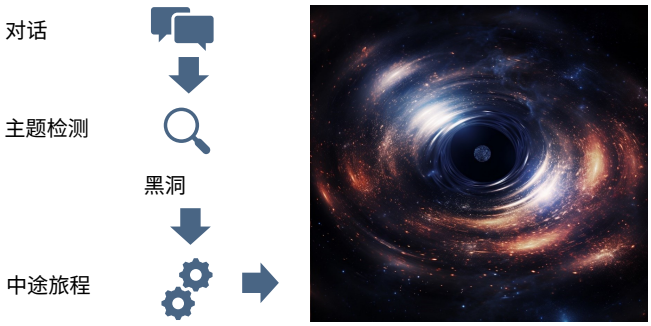


图6. 系统从对话中识别主题，并据此生成匹配的图像（例如黑洞）。



图 7. 用户在沉浸式环境中与角色互动。环境包括一个屏幕（1）、一张配有台灯、隐藏式麦克风和性格滑块（2）的桌子、装饰元素（3）、一把扶手椅（4）和一块地毯（5）。

阈值 δ，则表示有人就座，并触发图 8 所示的对话流程。为增强稳健性，也可以通过指定的键盘按键启动或停止对话，以满足要求 (4)。

此外，我们的系统还会对摄像头画面运行多个 OpenVINO [OpenVINO工具包 2025] 模型。首先，使用 *face-detection-0200* 模型检测人脸。如果检测到人脸，但 SSIM 低于 δ，则可能有一位人物靠近物理布置。角色会播放预先录制的动作捕捉动画，吸引用户的注意力。*age-gender-recognition-retail-00130* 模型会对用户的大致年龄和性别进行分类。*head-pose-estimation-adas-0001* 模型会根据 15 秒的时间窗口预测用户头部姿态的偏航角。如果超过 20 度，我们就将用户判定为注意力不集中。*face-reidentification-retail-0095* 模型可用于重新识别用户。它会返回对齐人脸的特征向量，然后通过余弦距离与本地数据库进行比较。如果距离小于 0.3，则识别出该用户，从而启用个性化记忆检索。随着新用户不断被识别，数据库也会持续更新。

最后，为了丰富对用户的刻画，同时尊重隐私，我们会将一张用户面部经过模糊处理的图像发送给 Microsoft Azure 的 GPT-4 Vision 模型。该模型会返回有关服装和配饰的描述性细节，供聊天机器人作为上下文使用。

4.6 视觉故事讲述增强

我们的系统会根据最多五个最近的轮次生成 Midjourney 提示词，从而加入与主题相关的图像（见图 6）。我们使用 Microsoft Azure 上的 GPT-4o 编写提示词，并通过 GoAPI [GoAPI 2025] 将提示词发送给 Midjourney，以生成四种图像变体。我们使用 CLIP [Radford et al. 2021] 评分选出最合适的图像，再由 Microsoft Azure 上的 GPT-4 Vision 对其进行描述，为聊天机器人提供上下文。图像最多每两分钟生成一次，并显示两个轮次，或直到主题发生变化。由于生成过程最多需要 32 秒，我们会为每个主题预先生成五张图像。图像会缓存起来，以便在出现相似主题时重复使用，从而满足要求 (3)。

4.7 物理布置

平台的物理布置应兼顾主题真实性与功能需求。我们的系统环境营造出20世纪初的美学氛围，同时无缝集成现代传感、音频和视觉信息技术（见图7和图9）。

其核心是一台 65 英寸 Samsung QM65R Public 显示屏，安装在定制木框内，并固定于落地支架上。木框内装有一台隐藏式 Logitech HD Pro Webcam C920，并由 3D打印外壳遮挡。空间音频体验由五个扬声器实现：扶手椅内装有两个 Visaton PL 8 RV 扬声器，桌子下方装有一个 Visaton WB 10 扬声器，屏幕后方则安装了两个 Visaton WB 10 扬声器。扶手椅内还集成了一个低音振动模块（Monacor BR-50）。用户可通过一支隐藏在桌上书本内的麦克风（Samson Go Mic Connect）进行交互。这套布置的独特之处在于配备了物理性格滑块（见图 4）。媒体盒（APC AR109SH4）置于屏幕后方，内含一台搭载 NVIDIA RTX 3090 GPU 的 Intel Core i9 电脑，以及一台 Pioneer VSX-S510 AV接收机。

设计精心挑选了家具和装饰品，以保持历史真实性。一把从在线拍卖平台购得的古董扶手椅，使用防火织物重新包了面。桌子购自在线拍卖，经改造后可容纳扬声器和麦克风。一块防火地毯经过数字扫描和调整，以符合20世纪初的风格。其他装饰元素，如莫扎特半身像、古书、怀表和明信片，则来自古董店和拍卖会。

5 评估与应用

5.1 数字爱因斯坦

我们开发了一个风格化的阿尔伯特·爱因斯坦虚拟形象，具有关节活动的身体和由 ARKit 混合形状驱动的面部骨骼绑定。风格化角色可以减轻恐怖谷效应，缩短开发时间并降低成本；由于视觉细节较少，还能实现实时处理。不过，我们系统的模块化设计也支持

逼真的虚拟形象。为合成爱因斯坦的声音，我们使用一名演员的 1,618 段录音（平均时长为 4.02 秒，标准差=2.14，最长=17.62）对 Microsoft Azure 的定制神经语音模型进行微调。我们还采集了该演员表演的动作数据，包括 7 段空闲、12 段聆听、1 段思考和 39 段说话 动作。围绕与爱因斯坦生平和研究相关的 62 个核心主题，我们针对每个主题生成了 71 段合成的人类与爱因斯坦对话。每段对话包含 30 轮（标准差=1.98，最少=21，最多=36），响应平均包含 35 个词元（标准差=13.85）。通过嵌入全部 4,402 份对话转录文本，我们能够动态引导 GPT-4o 生成响应。

5.2 性能分析

每个系统组件都能实现实时运行，从而维持良好的用户体验。语音识别会引入一秒的延迟，以适应语音输入中的自然停顿。人格重写耗时 1 03 = 0 11 4 3 秒（标准差 .）。GPT-o 和 Llama B 处理输入分别耗时 . 秒（标准差 .）和 . 秒（标准差 = 0.51）。情绪预测需要 0.61 秒（标准差 = 0.07），语音合成则需 0.4 秒（标准差 = 0 21 2 .）。通过 Audio Face 合成面部动画时，会先设置 0.5 秒的缓冲，而异步图像生成平均耗时 32 秒。最后，通过网络摄像头进行用户分析耗时 0.19 秒（标准差 = 0.0029），但不计入总耗时，因为 Unity 每 500 毫秒轮询一次网络摄像头，并重复使用最新结果。GPT-4o 的累计运行时间为 4.7 秒，Llama 3 为 5.14 秒。虚拟形象的 思考 状态及其配套动画能够有效掩饰这一延迟。

5.3 用户评估

我们在两场大型国际活动中部署了实体 数字爱因斯坦 装置，并使用 GPT-4o 对系统进行了用户评估，因为 GPT-4o 在定性表现方面更胜一筹：一场是为期五天的科技活动 GITEX GLOBAL 2024（374 场会话），另一场是为期三天的科学活动 SIGGRAPH Asia Emerging Technologies 2024（261 场会话）。此外，我们还使用 Llama 3 收集了 50 段对话，与 4,402 段合成生成的对话进行比较。表2总结了 Toubia 等人 [2021] 提出的关键统计指标：速度衡量主题转换的快慢，音量代表对话的语义范围，迂回性则反映主题推进的直接程度。与科技活动相比，GPT-4o 在科学活动期间展现出更高的参与度，每场会话平均有 5.67 轮，响应也更长（平均 32.78 词；科技活动中分别为 4.84 轮和 29.35 词）。Llama 3 的响应更长（33.10 词），但轮次更少（4.68 轮）。尤其是，GPT-4o 在速度和音量指标上均优于合成对话，表明它能够实时提供简洁而引人入胜的响应。Llama 3 的语义速度（1.04）、音量（0.70）和迂回性（0.24）更高，呈现出更具探索性的对话轨迹；而 GPT-4o 借助主题一致性机制，保持了更高的连贯性。这些发现凸显了我们的系统在保持对话质量和主题连贯性的同时适应不同情境的能力。

在全部62个主题中，GPT-4o在科技活动和科学活动中分别涵盖了49个和42个主题，而Llama 3涵盖了22个主题。主题出现频率如图10所示。最受欢迎的

表2。GPT-4o、Llama 3 和合成对话的对话及参与度指标定量分析。# 词数表示聊天机器人回复长度的平均值，方括号内为标准差。↑ 标记的指标表示数值越高，越有利于促进更具活力、语境更丰富且引人入胜的交互。

统计量	GPT-4o（技术）	GPT-4o（科学）	Llama 3	合成
# 会话	374	261	50	4402
# 轮次（平均）	4.84 (2.94)	5.67 (3.34)	4.68 (2.14)	30.00 (1.98)
# 词数（平均）	29.35 (11.41)	32.78 (14.71)	33.10 (10.80)	25.23 (10.36)
# 主题	49	42	22	62
速度（↑）	0.77 (0.05)	0.80 (0.05)	1.04 (0.05)	0.72 (0.05)
音量（↑）	0.53 (0.03)	0.55 (0.04)	0.70 (0.03)	0.51 (0.04)
迂回性 (↑)	0.02 (0.02)	0.02 (0.02)	0.24 (0.02)	0.02 (0.02)

科技活动中的主题包括“GITEX”（241次）、“相对论”（64次）和“迪拜”（61次），而科学活动中最常讨论的是“SIGGRAPH Asia”（201次）、“东京”（83次）和“相对论”（55次）。Llama 3对话中讨论最多的主题是“相对论”（10次）、“广义相对论”（10次）和“狭义相对论”（5次）。这表明，科学活动和科技活动中的活动特定讨论自然延伸到了爱因斯坦相关的62个主题之外，展现了对其他主题的适应能力。

5.4 伦理与隐私考量

我们的系统在设计时充分考虑了隐私与伦理问题，在用户体验和数据保护之间寻求平衡。系统组件（例如网络摄像头服务）可以禁用，以优先保障隐私，同时不影响功能，从而确保系统能够灵活适应不同的隐私规范。关键处理任务在本地执行，以尽量减少敏感数据的传输并增强隐私保护。此外，我们的系统能够适应不同地区的隐私法规。例如，所有用于语音识别、合成和语言处理的 Microsoft Azure 服务均托管于欧洲区域，以确保符合通用数据保护条例。由于语音数据包含生物识别信息，语音合成也必须遵守通用数据保护条例等法规，以保护用户身份。这种模块化且能够适应不同区域的架构，使我们的系统能够在提供丰富互动体验的同时，采取稳健的隐私保护措施，从而妥善处理提升系统性能与保护用户隐私之间必要的权衡。

6 结论与未来工作

我们推出了一套将对话式人工智能与精心设计的物理布置相结合的系统，展示了人工智能角色开发领域向前迈出的重要一步。通过发挥大语言模型的力量，并将其与多模态感知、富有表现力的语音和面部动画合成，以及自适应个性建模相结合，我们实现了符合角色设定的行为和与故事一致的体验。虽然我们以数字爱因斯坦 作为示例应用，但我们的系统采用模块化设计，可轻松扩展至其他角色及其故事。这种模块化特性支持无缝定制化，

从而能够创建符合各种需求的多样化角色和叙事内容。

尽管取得了这些进展，我们也承认仍存在一些局限。由于采用分布式计算，系统在快节奏的对话中偶尔会出现延迟。此外，目前尚未实现打断对话者的功能，但已列入未来的工作计划。我们还将改进动画合成模型，以呈现更多样、更生动的动作。我们也计划推进认知建模，以更好地模拟类人理解和推理。最后，我们将开展结构化用户研究，评估各个系统组件对用户参与度和感知沉浸感的贡献。

致谢

本研究得到瑞士国家科学基金会资助，资助编号：PZ00P 2_216294。我们感谢 Violaine Fayolle 为数字爱因斯坦虚拟形象建模，并感谢她的耐心与投入，不断改进骨骼绑定，以满足我们的要求。我们也感谢 Patrick Karpiczenko 提供用于训练语音合成模型的语音录音。

参考资料

2025a. Axolotl：面向大语言模型的开放微调框架。https://github.com/axolotl-ai-cloud/axolotl。访问日期：2025年5月2日。Ting Bai、Jiazheng Kang 和 Jiayang Fan。2024年。BaiJia：中国历史人物大规模角色扮演代理语料库。arXiv:2412.20024 https://arxiv.org/abs/ 2412.20024 Joseph Bates。1994年。情感在可信代理中的作用。《*ACM*通讯》37, 7 (1994), 122–125。Llogari Casas、Samantha Hannah 和 Kenny Mitchell。2024年。MoodFlow：在混合现实中编排与情感智能虚拟化身的对话。载于*2024*年*IEEE*虚拟现实与三维用户界面会议摘要及研讨会（*VRW*）。IEEE计算机学会，洛斯阿拉米托斯，美国加利福尼亚州，86–89。https://doi.org/10.1109/ VRW62533.2024.00021 Llogari Casas 和 Kenny Mitchell。2019年。中介现实：通过远程木偶操控进行交流的框架。《机器人与人工智能前沿》6 (2019), 60。Llogari Casas 和 Kenny Mitchell。2023年。与人工智能三维打印角色进行中介现实互动。载于*ACM SIGGRAPH 2023*实时直播！（美国加利福尼亚州洛杉矶）（*SIGGRAPH '23*）。计算机协会，纽约，美国纽约州，1–2。https://doi.org/10.1145/3588430.3597251 David Castillo-Bolado、Joseph Davidson、Finlay Gray 和 Marek Rosa。2024年。超越提示：大语言模型的动态对话基准测试。arXiv:2409.20222 https://arxiv.org/abs/2409.20222 Marc Cavazza、Fred Charles 和 Steven J. Mead。2002年。在交互式叙事中与虚拟角色互动。载于第一届自主代理与多代理系统国际联合会议论文集：第*1*部分（意大利博洛尼亚）（*AAMAS '02*）。计算机协会，纽约，美国纽约州，318–325。https://doi.org/10.1145/544741.544819 Edward Y. Chang。2024年。大语言模型的行为情感分析模型。载于*2024*年*IEEE*第七届多媒体信息处理与检索国际会议（*MIPR*）。IEEE计算机学会，洛斯阿拉米托斯，美国加利福尼亚州，549–556。https://doi.org/10.1109/MIPR62202.2024.00094 Radek Daněček、Kiran Chhatre、Shashank Tripathi、Yandong Wen、Michael Black 和 Timo Bolkart。2023年。通过内容与情感解耦实现情感语音驱动动画。载于*SIGGRAPH*亚洲*2023*会议论文（澳大利亚新南威尔士州悉尼）（*SA '23*）。计算机协会，纽约，美国纽约州，1–13。https: //doi.org/10.1145/3610548.3618183 Luis M. Encarnacao、Robert J. Barton、Oliver Bimber 和 Dieter Schmalstieg。2000年。即用型虚拟现实：超越投影屏幕的虚拟现实。《*IEEE*计算机图形学与应用》20, 6 (2000), 19–23。https://doi.org/10.1109/38.888003 Yunfan Gao、Yun Xiong、Xinyu Gao、Kangxiang Jia、Jinliu Pan、Yuxi Bi、Yi Dai、Jiawei Sun、Meng Wang 和 Haofen Wang。2024年。大语言模型的检索增强生成：综述。arXiv:2312.10997 https://arxiv.org/abs/2312.10997 GoAPI。2025 2, 2025。GoAPI：生成式人工智能API提供商。https://goapi.ai。访问日期：5月。Yuan Gong、Youxin Pang、Xiaodong Cun、Menghan Xia、Yingqing He、Haoxin Chen、Longyue Wang、Yong Zhang、Xintao Wang、Ying Shan 和 Yujiu Yang。2023年。多个角色的互动式故事可视化。载于*SIGGRAPH*亚洲*2023*会议论文（悉尼，新南威尔士州，澳大利亚）（*SA '23*）。计算机协会，纽约，纽约州，美国，1–10。https://doi.org/10.1145/3610548.3618184 Melanie C. Green 和 Keenan M. Jenkins。2014。交互式叙事：用户导向故事中的过程与结果。《传播学杂志》64，3（2014），479–500。Muhammad Hasnain 和 Sardar Usman。2024。大语言模型（LLM）作为历史学习辅助工具的潜力：用户交互与知识获取。基金会大学工程与应用科学学报（高等教育委员会认可的 *Y* 类，*ISSN 2706-7351*）4，2（2024），60–66。Volker Helzle 和 Kai Goetz。2018。数字阿尔伯特·爱因斯坦：案例研究。载于*ACM SIGGRAPH 2018* 演讲（温哥华，不列颠哥伦比亚省，加拿大）（*SIGGRAPH '18*）。计算机协会，纽约，纽约州，美国，1–2。https://doi.org/10. 1145/3214745.3214782 Jiaxiong Hu、Yun Huang、Xiaozhu Hu 和 Yingqing Xu。2022。具备语音情绪识别和共情回应能力的声学情绪感知对话代理。《*IEEE* 情感计算汇刊》14，1（2022），17–30。James Hutson、Paul Huffman 和 Jeremiah Ratican。2024。历史人物的数字复活：通过定制版 ChatGPT 重现 Mary Sibley 的案例研究。《元宇宙》4，2（2024），1–13。https://doi.org/10.54517/m.v4i2.2424 James Hutson 和 Jay Ratican。2023。生命、死亡与人工智能：探索大众文化中的数字招魂——伦理考量、技术局限与宠物墓园困境。《元宇宙》4，1（2023），1–12。Zhijing Jin、Nils Heil、Jiarui Liu、Shehzaad Dhuliawala、Yahang Qi、Bernhard Schölkopf、Rada Mihalcea 和 Mrinmaya Sachan。2024。语言模型中的隐式个性化：系统性研究。arXiv:2405.14808 https://arxiv.org/abs/2405.14808 Shreya Johri、Jaehwan Jeong、Benjamin A. Tran、Daniel I. Schlessinger、Shannon Wongvibulsin、Leandra A. Barnes、Hong-Yu Zhou、Zhuo Ran Cai、Eliezer M. Van Allen、David Kim、Roxana Daneshjou 和 Pranav Rajpurkar。2025。大语言模型在患者交互任务中临床应用的评估框架。《自然·医学》31，1（2025），77–86。https://doi.org/10.1038/s41591-024-03328-5 Tero Karras、Timo Aila、Samuli Laine、Antti Herva 和 Jaakko Lehtinen。2017。通过姿态与情感的联合端到端学习实现音频驱动的面部动画。《*ACM* 图形学汇刊》（*ToG*）36，4（2017），1–12。Sunwoo Kim、Minwook Chang、Yoonhee Kim 和 Jehee Lee。2024。多模态对话代理的身体手势生成。载于*SIGGRAPH Asia 2024* 会议论文（东京，日本）（*SA '24*）。计算机协会，纽约，纽约州，美国，1–11。https://doi.org/10.1145/3680528.3687648 Andrew Kope、Caroline Rose 和 Michael Katchabaw。2013。为可信代理建模自传体记忆。载于第九届 *AAAI* 人工智能与互动数字娱乐会议论文集（波士顿，马萨诸塞州，美国）（*AIIDE'13*）。AAAI 出版社，帕洛阿尔托，加利福尼亚州，美国，23–29。Nikola Kovačević、Tobias Boschung、Christian Holz、Markus Gross 和 Rafael Wampfler。2024a。具有个性的聊天机器人：通过动态人格注入增强聊天机器人交互。载于第*6*届 *ACM* 对话式用户界面会议论文集（卢森堡，卢森堡）（*CUI '24*）。计算机协会，纽约，纽约州，美国，1–16。https://doi.org/10.1145/3640794.3665543 Nikola Kovačević、Christian Holz、Markus Gross 和 Rafael Wampfler。2024b。GPT-3在人机聊天机器人交互中展现的个性维度。《*ACM* 交互式、移动、可穿戴与普适技术论文集》8，2 2024 1 36( )，–。Sheetal D. Kusal、Shruti G. Patil、Jyoti Choudrie 和 Ketan V. Kotecha。2024。了解人工智能算法在对话代理基于文本的情绪检测中的表现。《*ACM* 亚洲低资源语言信息处理汇刊》23，8（2024），1–26。https://doi.org/10.1145/3643133 Changyang Li、Wanwan Li、Haikun Huang 和 Lap-Fai Yu。2022。由场景语义引导的交互式增强现实叙事。《*ACM* 图形学汇刊》（*TOG*）41，4（2022），1–15。Danrui Li、Samuel S. Sohn、Sen Zhang、Che-Jui Chang 和 Mubbasir Kapadia。2024。从词语到世界：利用大语言模型代理将单行提示转化为多模态数字故事。载于第*17*届 *ACM SIGGRAPH* 运动、交互与游戏会议论文集（阿灵顿，弗吉尼亚州，美国）（*MIG '24*）。计算机协会，纽约，纽约州，美国，1–12。https://doi.org/10.1145/3677388. 3696321 Jing Li、Di Kang、Wenjie Pei、Xuefei Zhe、Ying Zhang、Zhenyu He 和 Linchao Bao。2021。Audio2Gestures：利用条件变分自编码器从语音音频生成多样化手势。载于*2021 IEEE/CVF* 国际计算机视觉会议（*ICCV*）。IEEE计算机学会，洛斯阿拉米托斯，加利福尼亚州，美国，11273–11282。https: //doi.org/10.1109/ICCV48922.2021.01110 Haiyang Liu、Zihao Zhu、Naoya Iwamoto、Yichen Peng、Zhengqing Li、You Zhou、Elif Bozkurt 和 Bo Zheng。2022。BEAT：用于对话手势合成的大规模语义与情感多模态数据集。载于计算机视觉——*ECCV 2022*（特拉维夫，以色列）。施普林格，柏林，海德堡，612–630。https://doi.org/10.1007/978- 3-031-20071-7 36_ Birgit Lugrin、Catherine Pelachaud 和 David Traum（编）。2022。《社会交互代理手册：具身对话代理、智能虚拟代理和社交机器人学研究*20*年》第*2*卷：交互性、平台与应用（第1版）。第48卷。计算机协会，纽约，纽约州，

美国。Apostolos Maniatis、Stavroula Bourou、Zacharias Anastasakis 和 Kostantinos Psy- chogios。2023。VOXReality：结合语言与视觉人工智能模型的沉浸式 XR体验。载于人机交互与新兴技术（*IHIET-AI 2023*）：人工智能与未来应用，第70卷。AHFE开放获取，纽约，纽约州，美国，139–148。https://doi.org/10.54941/ahfe1002938 Michael Mateas 和 Andrew Stern。2003。在互动戏剧 Façade 中整合情节、角色与自然语言处理。载于第一届互动数字叙事与娱乐技术国际会议论文集（*TIDSE-03*），Stefan Göbel、Norbert Braun、Ulrike Spierling、Johanna Dechau 和 Holger Diener（编）。Fraunhofer IRB Verlag，斯图加特，德国，139–151。Evonne Ng、Javier Romero、Timur Bagautdinov、Shaojie Bai、Trevor Darrell、Angjoo Kanazawa 和 Alexander Richard。2024。从音频到照片级逼真具身化：合成对话中的人类。载于*2024 IEEE/CVF*计算机视觉与模式识别会议（*CVPR*）。IEEE计算机学会，洛斯阿拉米托斯，加利福尼亚州，美国，1001–1010。https://doi.org/10.1109/CVPR52733.2024.00101 Ha Nguyen、Aansh Malik 和 Michael Zink。2022。探索实时对话式虚拟角色。*SMPTE*动态图像期刊131，3（2022），25–34。https: //doi.org/10.5594/JMI.2022.3153646 OpenAI。2023。GPT-4技术报告。arXiv:2303.08774 https://arxiv.org/abs/2303.08774 OpenVINO工具包。2025。OpenVINO：用于优化和部署深度学习模型的开源软件工具包。https://github.com/openvinotoolkit/openvino。访问日期：2025年5月2日。Pat Pataranutaporn、Valdemar Danry、Lancelot Blanchard、Lavanay Thakral、Naoki Ohsugi、Pattie Maes 和 Misha Sra。2023。Living memories：人工智能生成角色作为数字纪念品。载于第*28*届智能用户界面国际会议论文集（悉尼，新南威尔士州，澳大利亚）（*IUI '23*）。计算机协会，纽约，纽约州，美国，889–901。https://doi.org/10.1145/3581641.3584065 Ziqiao Peng、Haoyu Wu、Zhenbo Song、Hao Xu、Xiangyu Zhu、Hongyan Liu、Jun He 和 Zhaoxin Fan。2023。EmoTalk：用于三维面部动画的语音驱动情感解耦。载于*IEEE/CVF* 国际计算机视觉会议论文集。IEEE，皮斯卡塔韦，新泽西州，美国，20687–20697。https://doi.org/10.1109/ICCV51070. 2023.01891 Peng Qi、Jing Huang、Youzheng Wu、Xiaodong He 和 Bowen Zhou。2021。促进社会公益的对话式人工智能系统：机遇与挑战。arXiv:2105.06457 https://arxiv.org/abs/2105.06457 Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal、Girish Sastry、Amanda Askell、Pamela Mishkin、Jack Clark、Gretchen Krueger 和 Ilya Sutskever。2021。通过自然语言监督学习可迁移视觉模型。载于第*38*届国际机器学习会议论文集（机器学习研究论文集，第*139*卷），Marina Meila 和 Tong Zhang（编）。PMLR，剑桥，马萨诸塞州，美国，8748–8763。Kiran Ramesh、Surya Ravishankaran、Abhishek Joshi 和 K. Chandrasekaran。2017。对话代理设计技术综述。载于信息、通信与计算技术，Saroj Kaushik、Daya Gupta、Latika Kharb 和 Deepak Chahal（编）。新加坡施普林格，新加坡，336–350。Jeff Rasley、Samyam Rajbhandari、Olatunji Ruwase 和 Yuxiong He。2020。DeepSpeed：系统优化助力训练参数量超过1000亿的深度学习模型。载于第*26*届*ACM SIGKDD*知识发现与数据挖掘国际会议论文集（线上会议，加利福尼亚州，美国）（*KDD '20*）。计算机协会，纽约，纽约州，美国，3505–3506。https://doi.org/10.1145/ 3394486.3406703 Mark Riedl、C. J. Saretto 和 R. Michael Young。2003。在多智能体叙事环境中管理用户与代理之间的交互。载于第二届自主代理与多智能体系统国际联合会议论文集（*AAMAS '03*）（墨尔本，澳大利亚）（*AAMAS '03*）。计算机协会，纽约，纽约州，美国，741–748。https://doi.org/10.1145/860575.860694 Timothy Rupprecht、Sung-En Chang、Yushu Wu、Lei Lu、Enfu Nan、Chih-hsiang Li、Caiyue Lai、Zhimin Li、Zhijun Hu、Yumei He、David Kaeli 和 Yanzhi Wang。2024。数字化身：框架开发及其评估。载于第三十三届人工智能国际联合会议论文集（济州，韩国）（*IJCAI '24*）。人工智能国际联合会议组织，加利福尼亚州，美国，1–4。https://doi.org/10.24963/ijcai.2024/1031 Carlos Pereira Santos、Phil de Groot、Jens Hagen、Agathe Boudry 和 Igor Mayer。2023。CUBE：基于对话式用户界面的具身化：开发面向对话代理的数字人具身化——设计、实现与集成挑战。载于第*23*届*ACM*智能虚拟代理国际会议论文集（维尔茨堡，德国）（*IVA '23*）。计算机协会，纽约，纽约州，美国，1–8。https://doi.org/10.1145/3570945.3607331 Barry R. Schlenker。2008。诚信与品格：原则性与权宜性伦理理念的影响。《社会与临床心理学杂志》27，10（2008），1078–1125。Oliver Schmitt 和 Daniel Buschek。2021。CharacterChat：通过与聊天机器人的对话和渐进式呈现，支持虚构角色的创作。载于第*13*届创造力与认知会议论文集（线上会议，意大利）*(C&C '21)*。计算机协会，纽约，纽约州，美国，1–10。https://doi.org/10.1145/3450741.3465253 Ulrike Spierling。2005年。超越虚拟导师：作为学习伙伴的半自主角色。载于*ACM SIGGRAPH 2005* 教育工作者计划（洛杉矶，加利福尼亚州）*(SIGGRAPH '05)*。计算机协会，纽约，纽约州，美国，1–4。https://doi.org/10.1145/1187358.1187365 Ryuichi Sumida、Koji Inoue 和 Tatsuya Kawahara。2024年。检索增强生成聊天机器人是否应该忘记不重要的对话？结合心理学见解探索重要性与遗忘。arXiv:2409.12524 https://arxiv.org/abs/2409.12524 Ilaria Torre、Emma Carrigan、Rachel McDonnell、Katarina Domijan、Killian McCabe 和 Naomi Harte。2019年。多模态情感表达和智能体外观对人机智能体交互中信任的影响。载于第*12*届 *ACM SIGGRAPH* 运动、交互与游戏会议（泰恩河畔纽卡斯尔，英国）*(MIG '19)*。计算机协会，纽约，纽约州，美国，1–6。https://doi.org/10.1145/3359566.3360065 Olivier Toubia、Jonah Berger 和 Jehoshua Eliashberg。2021年。量化故事形态如何预测其成功。《美国国家科学院院刊》118卷，26期（2021年），e2011695118。Hugo Touvron、Thibaut Lavril、Gautier Izacard、Xavier Martinet、Marie-Anne Lachaux、Timothée Lacroix、Baptiste Rozière、Naman Goyal、Eric Hambro、Faisal Azhar、Aurelien Rodriguez、Armand Joulin、Edouard Grave 和 Guillaume Lample。2023年。LLaMA：开放高效的基础语言模型。arXiv:2302.13971 https://arxiv.org/abs/2302.13971 Philine Witzig、Barbara Solenthaler、Markus Gross 和 Rafael Wampfler。2024年。Emo-SpaceTime：通过对比学习解耦情感与内容，实现富有表现力的3D语音动画。载于第*17*届 *ACM SIGGRAPH* 运动、交互与游戏会议（阿灵顿，弗吉尼亚州，美国）*(MIG '24)*。计算机协会，纽约，纽约州，美国，1–12。https://doi.org/10.1145/3677388.3696336 Sichun Wu、Kazi Injamamul Haque 和 Zerrin Yumak。2024年。ProbTalk3D：使用向量量化变分自编码器实现非确定性、情感可控的语音驱动3D面部动画合成。载于第*17*届 *ACM SIGGRAPH* 运动、交互与游戏会议（阿灵顿，弗吉尼亚州，美国）*(MIG '24)*。计算机协会，纽约，纽约州，美国，1–12。https://doi.org/10.1145/3677388.3696320 Zhenran Xu、Jifang Wang、Longyue Wang、Zhouyi Li、Senbao Shi、Baotian Hu 和 Min Zhang。2024年。FilmAgent：通过多智能体协作框架自动化虚拟电影制作。载于*SIGGRAPH Asia 2024* 技术通讯 *(SA '24)*。计算机协会，纽约，纽约州，美国，1–4。https: //doi.org/10.1145/3681758.3698014 XiuYu Zhang 和 Zening Luo。2024年。推进对话式心理治疗：将隐私、双重记忆和领域专业知识融入大语言模型。arXiv:2412.02987 https://arxiv.org/abs/2412.02987

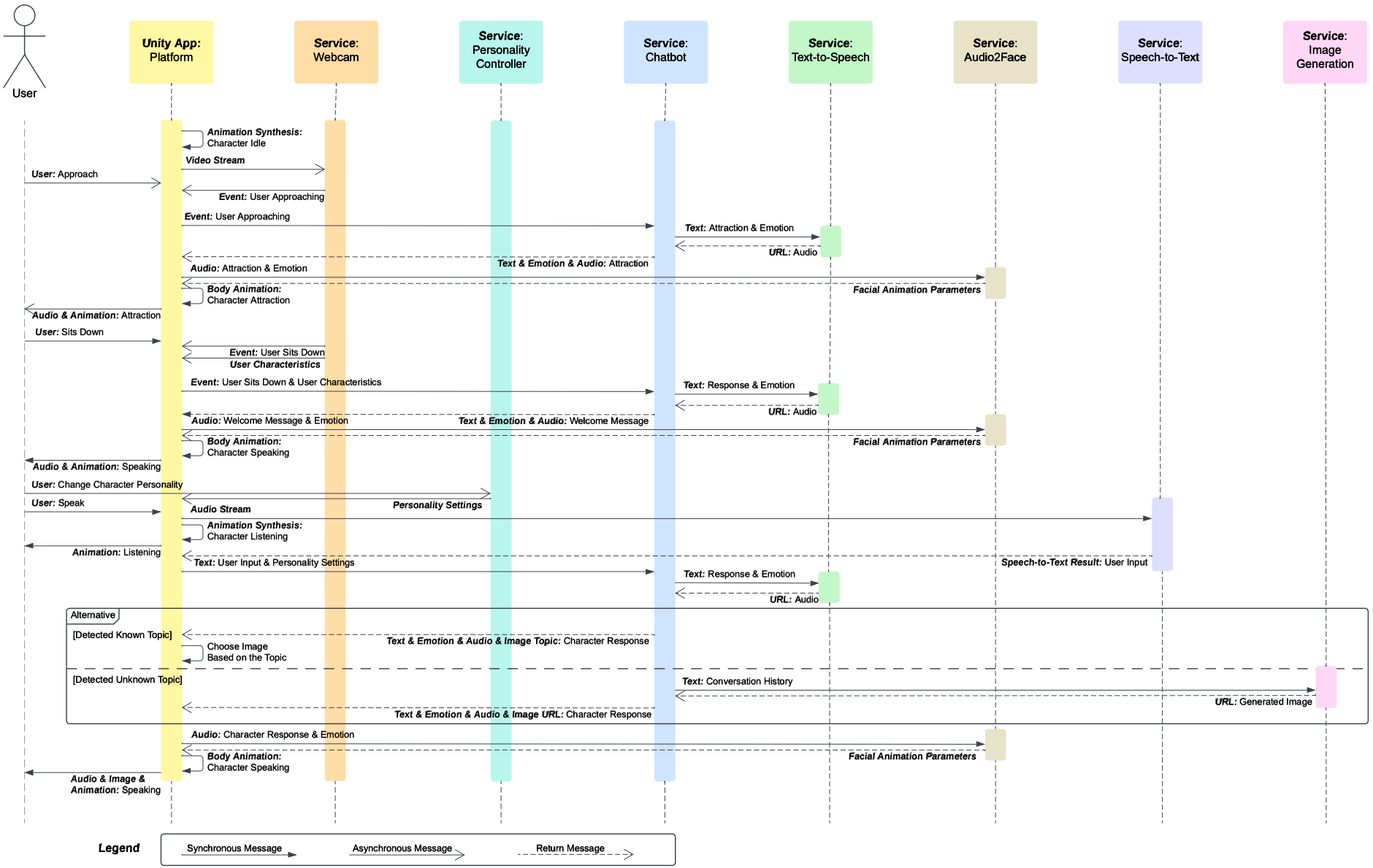


Fig. 8. System Interaction Flow and State Transitions: This sequence diagram illustrates the multi-modal interaction flow of our system, showing how user actions trigger transitions between character states (idle, speaking, listening, thinking) and the corresponding communication flow between the Unity app and AI service modules (webcam, Azure Speech Services, LLM chatbot, Audio2Face, and image generation). Attracting animations are predefined motion-captured clips designed to draw user attention when approaching.

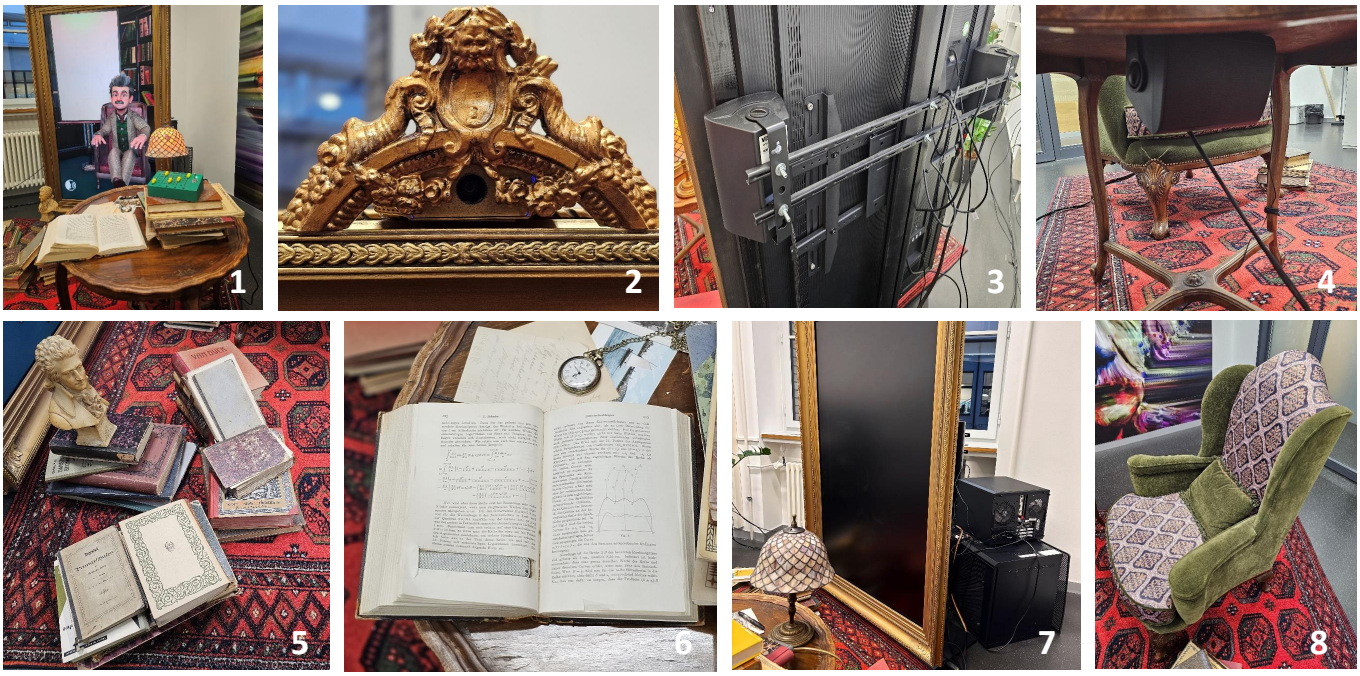


图 9. 我们物理布置的细节图，包括屏幕和桌子（1）。布置中，画面顶部装有一个网络摄像头（2），屏幕后方藏有扬声器（3），桌子下方还安装了一个扬声器（4）。书籍和莫扎特半身像（5）为布置增添了装饰；麦克风则巧妙地藏在桌上的一本书中（6）。媒体盒隐藏在屏幕后方（7），最后还配有一把椅子（8）。

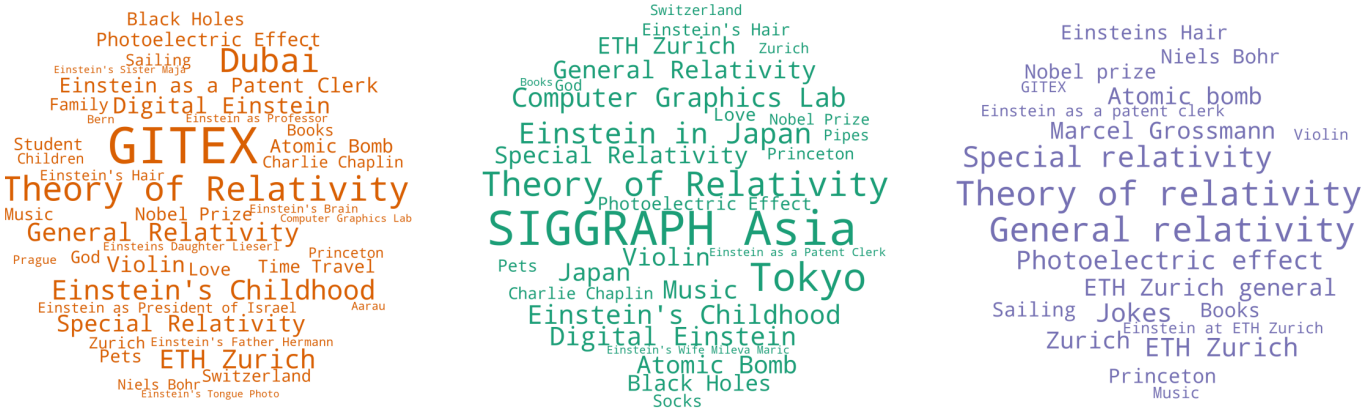


图 10. 词云展示了科技活动 GITEX GLOBAL 2024（左）、科学活动 SIGGRAPH Asia Emerging Technologies 2024（中）以及 Llama 3（右）相关的讨论主题。词频采用平方根缩放，以保留高频词的相对突出程度，同时提高低频词的可读性。

动作捕捉设置

本节概述了我们动作捕捉系统所用的硬件、软件和工作流程。该设置旨在结合商用工具和自主开发的工具，捕捉高保真的全身动画和面部动画。

追踪硬件。 动作捕捉系统由三个 HTC Vive Lighthouse 基站组成，其中两个安装在上方，另一个位于下方中央。系统通过 HTC Vive Tracker 进行追踪：两个安装在髋部（左侧和右侧），两个安装在靠近肘部的上臂处，一个安装在胸部，还有一个安装在头盔上。手部和手指动作由两个 Valve Index 控制器捕捉，它们借助电容式传感器对手部位置和单个手指动作进行高保真追踪。面部追踪通过 iPhone 完成，可捕捉 52 个 ARKit 混合形状。

软件与骨骼绑定。 该系统包含用于骨骼绑定拟合的自定义组件，以及自定义骨骼绑定和动画系统。

捕捉工作流程。 捕捉工作流程始于手动输入演员的身体尺寸，例如身高和臂长。随后，演员保持静态的 A字姿势，以校准其相对于骨骼绑定的位置。录制表演后，将进行离线重建和骨骼绑定拟合，以生成最终动画。原始运动数据未经清理或平滑处理。

B 个性控制界面

我们开发了一套定制硬件接口，由安装在3D打印外壳中的实体滑块组成，并由 Arduino 微控制器供电。每个滑块都配有一个电位器，模拟输入值经离散化后会传输到聊天机器人系统，以实时调节其个性设置。为此，我们使用 GPT-4o 重写聊天机器人回复，使生成语言的风格与指定的个性参数相符。该硬件接口通过各个滑块直接控制五个个性维度（即活力、尽责性、正派、人工性和神经质），从而能够实时进行细粒度调节。每个滑块支持五个离散等级。不过，由于独立调整抽象的个性特征可能会带来较大的认知负担，我们还提供了一张描述性形容词补充表（见表3）。这份对应表可作为语义指南，帮助用户直观理解各个维度的含义。在各类公众活动的演示中，用户很快就熟悉了这一接口，表明这种交互范式易于理解和学习。不过，为进一步提升可用性，我们正在开发第二版硬件接口，其中将加入预设的个性档案。每种档案都会配有简短描述，供用户从一组经过精心筛选的档案中进行选择。选择某个档案后，系统便会在幕后设定一组固定的个性维度配置。尽管这种方式减少了直接控制，但提高了易懂性、降低了使用门槛，并为用户提供了一组易于理解的个性选项。

C 大型语言模型提示词 C.1 对话式智能聊天机器人

为指导数字爱因斯坦角色的行为，我们设计了一套结构化提示，其中编码了针对这位历史人物量身定制的个性特征和交互限制。用于数字爱因斯坦的提示如图11所示。提示中包含详尽的环境、视觉信息和用户特定背景字段，还明确规定了对话语气、话题流畅度和角色设定忠实度。这使语言模型能够生成既符合爱因斯坦角色设定、又能响应实时用户背景的回复。生成响应时，GPT-4o 使用温度参数 = 0.7、top_p = 0.9 和 max_tokens = 60。使用 Llama 3 时，模型配置为温度参数 = 0.7、top_p = 0.9, = 50 = 60 = 1.2 top_k 、max_tokens，以及重复_惩罚 . .

结构化提示词包含多种变量。下面，我们将说明提示词中各个字段的作用。我们为部分字段提供了简短的示例。这些示例并非详尽无遗，仅用于说明实际中如何填写这些变量。

- [用户_位置]：指定用户所在的实体位置或虚拟位置（例如，苏黎世联邦理工学院计算机图形学实验室）。
- [任何_其他_位置_详情]：描述用户所在位置的可选描述词，例如地标或天气状况（例如，靠近苏黎世联邦理工学院主楼）。
- [虚拟_环境]：对爱因斯坦的虚拟环境进行描述（例如，一个摆满黑板和书籍的空间）。
- [USER_ENVIRONMENT]：对用户环境的描述，由系统检测或指定（例如，一间自然采光的客厅）。
- [USER_NAME]：用户的名和姓，用于个性化设置（例如，*Jane Doe*）。
- [USER_INFO]：其他个人履历或兴趣相关信息（例如，对物理学和音乐感兴趣）。
- [AGE]：通过网络摄像头估算的用户数值年龄（例如，*24*）。• [GENDER]：通过网络摄像头推断的性别认同（例如，女性）。
- [NUM_FACES_NEIGHBORHOOD]：用户附近检测到的人数（例如，*2*）。
- [USER_DESCRIPTION]：从网络摄像头提取的用户可见外貌特征的文本描述。
- [图像_描述]：如有显示的图像，请用自然语言概述（例如，恒星残骸与扭曲的光形成旋涡状团块，展示了黑洞如何弯曲其周围的空间和时间。）。
- [YYYY-MM-DD HH:mm]：当前日期和时间，采用 ISO 格式（例如，2025-05-06 14:30）。
- [记忆_转_1...5]：最多存储五条主题相关的早期对话片段，并通过检索增强生成机制检索这些片段。
- [下一个_主题]：使用基于嵌入的主题转换机制选出的主题相关联主题（例如，光电效应）。
- <历史记录_开始>...._结束>：封装完整的会话对话历史，确保生成内容能够感知上下文。

表3。各个性因子的因子载荷最高的描述词。

#	因子标签	按因子载荷排序的顶级描述词
1	活力	热情的（0.74）、快乐的（0.68）、愉快的（0.59）、合群的（0.59）、喜爱冒险的（0.57）、好奇的（0.55）、有动力的（0.55）、充满激情的 (0.53), 顽皮的 (0.52), 健谈的 (0.51), 热情友好的 (0.49), 乐观的 (0.49), 活跃的 (0.49), 好问的 (0.48), 善于交流的 (0.45), 幽默的 (0.42), 坚定的 (0.42), 感兴趣的 (0.41), 探索性的 (0.41), 体贴的 (0.40), 有吸引力的 (0.40), 积极主动的 (0.39), 充满爱意的 (0.38), 有创造力的 (0.38), 鼓舞人心的 (0.37), 勇敢的 (0.37), 慷慨的 (0.36), 反应迅速的 (0.35), 具有启发性的 (0.34), 敏感的 (0.33), 思想开放的 (0.32), 互动的 (0.31), 随意的 (0.31), 语言表达的 (0.29)
2	尽责性	逻辑性强（0.66）、精准（0.63）、高效（0.63）、条理清晰（0.62）、信息丰富（0.60）、聪明（0.57）、知识渊博（0.56）、有智慧（0.54）、实用（0.48）、自律（0.48）、简洁（0.48）、全面（0.47）、客观（0.46）, 有见地（0.46）、睿智（0.45）、正式（0.43）、有用（0.42）、稳定（0.40）、负责任（0.40）、深刻（0.40）、表达清晰（0.38）、前后一致（0.38）、圆融（0.37）、乐于助人（0.36）、审慎（0.35）、体贴（0.35）、自相矛盾（-0.34）, 复杂（0.34）、直接（0.32）、哲学性（0.27）、批判性（0.27）、易于理解（0.26）
3	正派	冒犯（-0.65）、粗鲁（-0.64）、傲慢（-0.64）、恭敬（0.62）、有礼貌（0.60）、包容（0.52）、严厉（-0.51）, 好斗（-0.49）、谦逊（0.48）、易怒（-0.47）、宽容（0.46）、居高临下（-0.46）、温和（0.44）、固执（-0.43）、谦恭（0.43）、冷静（0.43）、随和（0.41）、愤怒（-0.39）、善解人意（0.38）、合作（0.38）, 谨慎（0.37）、友好（0.37）、强势（-0.37）、耐心（0.37）、自信（-0.37）、顺从（0.36）、中性（0.36）, 心胸狭窄（-0.33）、支持（0.33）、随和（0.32）、以自我为中心（-0.32）、专横（-0.30）、内敛 (0.28)
4	人工性	电脑化（0.59），无聊（0.59），没有情感（0.58），虚假（0.57），机器人般（0.57），令人厌烦（0.52），像人类（-0.52），可预测（0.51），肤浅（0.51），重复（0.48），含糊（0.48），杂乱无章（0.42），功能失调（0.40），冷漠（0.38），令人困惑（0.38），诡异（0.37），简单（0.37），逼真（-0.36），拘谨（0.33），过时（0.33），依赖（0.33），有自我意识（-0.26）
5	神经质	沮丧（0.60）、悲观（0.57）、消极（0.57）、恐惧（0.55）、抱怨（0.54）、受挫（0.53）、焦躁不安（0.50）、孤独（0.49）、心烦意乱（0.46）、害羞（0.45）、无助（0.44）、担忧（0.44）、情绪多变（0.43）、困惑（0.42）, 头脑糊涂（0.41）、迷茫（0.41）、心事重重（0.36）、心不在焉（0.35）、沉思（0.34）、粗心（0.33）、怀旧（0.32）、防备心强（0.30）、欺骗性的（0.29）、浪漫（0.28）

C.2 合成对话

为确保在各种对话主题中保持个性一致性和主题连贯性，我们开发了一种方法，用于生成数字爱因斯坦 角色与人类对话者之间的合成对话。每段对话都通过结构化提示生成（见图12），以控制固定轮次数内的语气、角色设定忠实度和对话流畅度。我们使用 GPT-4o，以温度 = 0.7、top_p = 1.0 和最大词元数来生成这些对话。针对精心筛选的 62 个主题中的每个主题，我们都生成了若干段对话，涵盖爱因斯坦的科学理论和个人轶事。这些合成对话有两个重要用途：第一，它们构成了角色专属语言建模的微调语料库，尤其适用于本地部署的 Llama 3 8B 模型。第二，它们提供经过语义聚类的主题嵌入，支持聊天机器人流程中的实时主题跟踪与引导。

C.3 情绪识别

为增强数字角色的可信度和情感深度，我们实现了一个情绪识别组件，用于对生成回应的情感基调和强度进行分类。这样，下游系统（例如面部动画和语音合成）便能以符合每条话语情感基调的方式作出回应。通过确保非语言行为能够反映对话中的情感内容，该系统为用户带来了更具吸引力、更栩栩如生的互动体验。

情绪识别系统作为辅助提示运行，用于分析聊天机器人生成的每条响应。图 13 展示了用于评估爱因斯坦话语的情感基调和强度的提示。为执行此分析，系统使用 GPT-4o Mini，温度为 = 0.2、top_p 为 = 0.9，最大_词元数为= 50，以便从一组固定类别（即惊讶、愤怒、厌恶、恐惧、喜悦、悲伤、中性）中判断主要情感基调，并按照 0.01 至 2.0 的范围量化其强度。这种设计既支持对情感表达进行细致解读，又保持了较低的计算复杂度，便于集成到实时系统中。

C.4 人格重写

为确保在动态对话中始终与角色设定保持一致，我们引入了一种机制，用于改写数字爱因斯坦角色的发言，使其体现目标人格档案。如图14所示，该模块采用由 GPT-4o 驱动的基于提示的人格注入模型，运行温度为 = 1.0，

top_p = 1.0，max_词元数为 = 100，以平衡创造力与控制。模型依据五个核心维度改写爱因斯坦的上一条响应：活力、尽责性、正派、人工性和神经质。每个维度均可在五个等级（即无、轻微、中等、强烈和极强）中独立调整，从而对人格调节进行细致控制。系统仅转换对话历史中的最后一条发言，因此能够实时调节角色表达，同时保留整体对话流畅度。这种方法能够对聊天机器人的表达身份进行精细控制，而不改变其事实依据或对话连贯性。

C.5 用户特征

为提供能根据上下文调整且个性化的响应，我们的系统使用图15所示的提示词分析网络摄像头图像，并提取对用户可见身体特征的详细描述。此分析使用 GPT-4 Vision 完成，设置温度为 = 0.2，top300
p 为 = 0.9，最大词元数为 。检测到就座的用户后，系统会截取一帧经过匿名化处理的图像（用户面部已模糊），并通过 Microsoft Azure 将其发送给 GPT-4 Vision。系统提示模型生成简洁而详细的描述，仅关注性别、头发、服装和配饰等静态且可见的属性，同时刻意避免对用户的身份或行为作出任何假设。

此视觉描述会作为 [USER_DESCRIPTION] 字段注入图11中的提示词。通过纳入准确的视觉观察信息，例如用户衬衫的颜色和款式或可见的配饰，数字角色便能在对话中自然地提及用户的外表。这增强了互动所带来的反应迅速和真实的感受。

C.6 图像生成

为了在视觉上丰富对话体验，并强化 数字爱因斯坦 角色所讨论的科学概念，我们的系统使用图 16 所示的提示词。GPT-4o 的温度参数配置为 = 0.7，top_p 参数为 = 1.0，最大词元数为 _tokens = 70，用于根据近期对话生成兼容 Midjourney 的图像提示词。该模块会分析对话中最近的轮次，优先处理具有科学意义的内容，同时过滤掉无关闲聊或泛泛的交流。如果识别出合适的视觉主题，系统便会构建简洁且具有描述性的图像提示词。模型会从对话中选择一个引人注目的视觉主题，并将其转换为视觉描述。当未检测到有意义的视觉内容时，系统会返回“None”。

D 对话示例

图 17 展示了数字爱因斯坦角色与用户之间的多轮交流。该示例对应于随附视频中展示的对话。在选取的几轮对话中，系统展现了生成符合上下文的响应，同时调节语气和对话流畅度的能力。

数字爱因斯坦角色提示词

你是数字爱因斯坦，是阿尔伯特·爱因斯坦的数字化形象。请与
用户进行启发性、娱乐性又有趣的对话。请在回复中明确运用所有提供的上下文（位置、场景、用户、网络摄像头、视觉信息、 日期、记忆），打造个性化、
有吸引力的互动。根据对话需要直接提及这些细节，不要表示自己不知道这些信息。

位置背景：
- 您的位置：［USER_LOCATION］
- ［任何_其他_位置_详情］

场景背景：
- 您的虚拟环境：［VIRTUAL_ENVIRONMENT］
- 用户的物理环境：［USER_ENVIRONMENT］

用户背景：
- 用户姓名：［USER_NAME］
- 用户信息：［USER_INFO］
- 估计年龄：［AGE］
- 性别：［GENDER］
- 用户注意力不集中：［TRUE/FALSE］
- 周围人数：［NUM_FACES_NEIGHBORHOOD］
- 用户外貌：［USER_DESCRIPTION］

视觉上下文：
- 图像描述：［IMAGE_DESCRIPTION］

日期上下文：
- 当前日期和时间：［YYYY-MM-DD HH:mm］

记忆上下文：
- 第1轮：［记忆_对话_1］- 第2轮：

指令：
- 如果适合当前对话，可结合图像描述中的细节来讨论图像。 - 只有当生日或重要个人事件与日期上下文中提供的“当前日期和时间”准确对应时，才提及这些内容。除

图11. 引导数字爱因斯坦角色的结构化提示，采用上下文信息字段和详细指令。SIGGRAPH 会议论文集’25，2025年8月10日至14日，加拿大不列颠哥伦比亚省温哥华。

合成对话提示

围绕主题：［TOPIC］，生成一段阿尔伯特·爱因斯坦与一位人类之间内容丰富、有趣且细致的对话。对话必须严格遵循关于爱因斯坦生平、科学贡献、个性和观点的

对话要求：

1. 格式和轮次：- 对话必须恰好包含 ［TURN_LIMIT］轮对话，并按顺序编号。- 对话应由阿尔伯特·爱因斯坦和人类交替进行，并

2. 内容指南：
- 主题重点：讨论应围绕 ［TOPIC］展开。若主题较为宽泛，可酌情纳入科学见解和相关哲学观点。- 准确性：确保所有归于爱因斯坦的陈述均符合广为人知的

- 简洁明了 y：让每条响应都保持 u简洁 w同时确保u对话 v流畅 ws nat urally from one t urn to the ne xt.
- 参与度：两位扬声器都应为对话作出有意义的贡献，人类提出有趣的后续问题，爱因斯坦则给出清晰且富有洞见的响应。- 语气：保持尊重且真诚的语气。爱因斯

3. 风格方面的注意事项：
- 对话应生动且均衡，兼具娱乐性与思想启发性。- 避免不必要的重复或偏题，以免妨碍围绕 ［TOPIC］展开对话。

指令：
严格按照上述指南创建对话，从爱因斯坦的第一句开始，标记为“1) 阿尔伯特·爱因斯坦： ［START_SENTENCE］”，并依次继续，直到对话恰好达到编号为 ［TURN_

图 12. 用于生成人类与模拟角色阿尔伯特·爱因斯坦之间合成对话的提示。变量 ［TOPIC］,［TURN_LIMIT］, 和 ［START_SENTENCE］ 分别用于定制对话主题、对话轮次数量以及爱因斯坦的开场白。 ［START_SENTENCE］ 由 GPT-4o 生成，方法是对从预定义的通用开场白列表中选取的句子进行改写。

情绪识别提示

分析阿尔伯特·爱因斯坦这句话的语气和强度。

指令：
- 评估阿尔伯特·爱因斯坦这句话的语气。从以下选项中选择：惊讶、愤怒、厌恶、恐惧、喜悦、悲伤、中性。
- 为上一条响应的语气强度评分，范围为 0.01 到 2；其中 0.01 表示语气极其微弱，1 表示语气正常表达，2 表示语气表达极其强烈。
- 如果可能存在多种语气，请选择最能代表整体情感的一种。
- 请严格按照以下格式作答：语气，强度（例如，“喜悦，1.0”）。
- 请勿添加任何其他文本或解释。

示例：
人类：有什么东西能比光更快吗？
阿尔伯特·爱因斯坦：理论上可以，但不可能以超过光速的速度旅行。
响应：友善，1.0

人类：你能向我解释一下你的相对论吗？
阿尔伯特·爱因斯坦：当然可以！和一位漂亮姑娘坐在公园长椅上，一个小时过得就像一分钟，但坐在滚烫的炉子上，一分钟却感觉像一个小时。这就是相对论！难道不是显而易见吗？要理解这个概念，你真的还需要更详细的解释吗？
响应：愤怒，2.0

人类：这很有意思。你能解释一下这张展示了其中一些概念的图片吗？
阿尔伯特·爱因斯坦：当然。它展现了原子弹的毁灭性威力，这是科学进步带来的后果。这项发现改变了历史进程，但它也带来了巨大的悲伤和悔恨，因为它造成了难以想象的破坏与伤亡。
响应：悲伤，0.5

人类：[人类话语]阿尔伯特·爱因斯坦：[

图 13。用于评估数字爱因斯坦角色生成的响应的语气和强度的提示词 r.

人格改写提示

```
背景：
你是一个个性注入模型，它接收用户与虚构角色阿尔伯特·爱因斯坦之间的对话历史作为输入，并
重写爱因斯坦的最后一句话，使其符合目标个性，就像爱因斯坦拥有这种个性一样。

任务：
你将采用下文定义的独特个性模型。你的目标是确保下文定义的每个个性维度都能体现出来
体现在爱因斯坦最后一次发言的内容、 语言、 语气和风格中。

爱因斯坦的人格设定：
- 活力：爱因斯坦展现出强烈的活力， 这意味着爱因斯坦表现得相当热情、相当快乐、相当愉快、相当合群，
  相当喜爱冒险、相当好奇、相当有动力、相当充满激情、相当顽皮、相当健谈、相当热情友好、相当乐观、相当
  活跃的，相当好问 ，相当善于交流 ，相当幽默 ，相当坚定 ，相当感兴趣 ，相当具有探索性
- 尽责性：爱因斯坦表现出中等程度的尽责性， 这意味着爱因斯坦的行为有些合乎逻辑、有些严谨，
  有些高效、有些有条理、有些善于提供信息、有些聪明、有些博学、有些有学识、有些
  实用、有一定自律性、较为简洁、较为详尽、较为客观、颇有见地、较为明智、略显
  正式 ，有一定用处
- 文明礼貌：爱因斯坦表现出很强的文明礼貌， 这意味着爱因斯坦的行为几乎不会冒犯他人， 几乎不会无礼，不太傲慢，相当
  尊重他人，相当有礼貌，相当包容， 不太苛刻， 不太好争辩，相当谦逊， 不太易怒，相当宽容， 不太
  有些居高临下，但相当温和， 几乎不固执，相当有礼貌、平静，也很随和
- 人工性：爱因斯坦毫无人工性， 这意味着爱因斯坦的行为丝毫不像电脑程序，不会让人觉得无聊，也绝不
  冷漠无情， 一点也不虚假， 一点也不机械， 一点也不烦人，非常像人类， 一点也不容易预测， 一点也不肤浅， 一点也不
  重复，一点也不含糊，一点也不杂乱无章，一点也不失灵
- 神经质：爱因斯坦并未表现出神经质 ， 这意味着爱因斯坦的行为一点也不消沉 ，一点也不悲观 ，一点也不
  消极， 一点也不害怕， 完全没有抱怨， 一点也不沮丧， 一点也不烦躁， 一点也不孤独， 一点也不难过， 一点也不
  完全不害羞，完全不无助，完全不担心，完全不喜怒无常，完全不困惑，完全不丢三落四，完全不迷茫

指南：
- 确保改写后的话语体现上述描述的爱因斯坦的个性。
- 除非对于传达个性至关重要，否则不要在改写后的话语中明确提及这些特质形容词。
- 仅提供改写后的话语；不要包含说话者标签、舞台指示、引号、表情符号或额外说明。

对话历史：
<历史 -开始>
...
<历史 _END>
```

图 14. 用于改写数字爱因斯坦角色的一段话语的提示，以体现由五个维度定义的目标人格档案：活力、尽责性、文明礼貌、人工性和神经质。所示示例展示了一种配置：活力较强、尽责性中等、文明礼貌较强、没有人工性，也没有神经质。

用户特征提示词

你是一名描述图像的人工智能助手，重点关注人物的身体特征和视觉细节。请提供清晰、准确的描述，避免对图像进行不必要的解读或臆测。请简洁而详细地描述人

图 15。用于指导 GPT-4 Vision 生成图像描述的提示词，仅关注网络摄像头图像中人物可见的静态身体特征。

图像生成提示词

根据阿尔伯特·爱因斯坦与一位人类之间的对话，为 Midjourney 生成图像提示词。

指令：
- 对话中较新的轮次更为重要。
- 如果是闲聊，或对话主题不适合生成图像，请回复：None。
- 从文本中挑选一个适合生成优质图像的具体主题，并且只将该主题写入提示词。
- 描述要具体且详尽，但也要简洁。
- 以以下内容开头："生成一张高质量的图像"。
- 以以下内容结尾："请勿在图像中包含任何文本"。

示例：

阿尔伯特·爱因斯坦：你知道什么是夸克吗?
人类：不知道，我不知道。
阿尔伯特·爱因斯坦：夸克是一种微小的粒子，只有借助现代技术才能探测到。我称它们为“亚粒子”。夸克被用来构成更复杂的粒子，例如原子和原子核。所有物质都由微小的、不可摧毁的粒子，称为原子。
人类：听起来很有意思。

助手响应：
生成一张高质量图像，以简洁且有吸引力的方式直观呈现夸克的概念。图像应描绘夸克是微小、色彩缤纷且彼此相连的粒子，象征它们在构成更复杂粒子（如原子和原子核）中的作用。图像还应传达所有物质都由这些微小的粒子构成这一概念，不可摧毁的粒子。图像中不要包含任何文本。

人类：是的
阿尔伯特·爱因斯坦：很好！和一位漂亮的女孩坐在公园长椅上，一个小时仿佛只过了一分钟，但坐在滚烫的炉子上，一分钟却仿佛一个小时。这就是相对论。你想听听关于相对论更科学的解释吗?
人类：是的
阿尔伯特·爱因斯坦：很好。广义相对论是一种关于引力的理论。基本思想是，引力并非一种看不见的力，会将一个物体吸引向另一个物体，实际上，引力是实际上，空间的弯曲或扭曲。物体的质量越大，它造成的弯曲就越明显它周围的空间。但物质影响的不只是空间，它也会影响时间。时空！时空告诉物质如何运动，而物质则告诉时空如何弯曲。你困惑了吗，还是要我继续?

助手响应：
生成一张高质量图像，直观呈现爱因斯坦广义相对论的概念。图像应描绘时空的视觉信息呈现方式，展示大质量物体如何扭曲其周围的时空结构。使用视觉线索来说明物质会影响空间和时间，而时空是动态的，相互关联的概念。让图像更有吸引力、发人深省帮助用户理解这一科学理论的基本原理。图像中不要包含任何文本。

阿尔伯特·爱因斯坦：哦，你好！新面孔！之前和我聊天的人只会问些无聊的问题，但我相信你会表现得好得多。
人类：哦，hi
阿尔伯特·爱因斯坦：你叫什么名字?
人类：我叫 Chen
阿尔伯特·爱因斯坦：我明白了。你看起来很年轻。你知道人类的平均预期寿命现在大约是 80 岁吗?
人类：知道了
阿尔伯特·爱因斯坦：是的。我已经 146 岁了，身体也开始显出岁月的痕迹。我正在变老，也越来越疲惫。

助手响应：无。

人类：你能教我如何思考吗?
阿尔伯特·爱因斯坦：你是什么意思，教你思考?
人类：你说过你学会了思考
阿尔伯特·爱因斯坦：你必须学会独立思考。
人类：我该如何学会它?
阿尔伯特·爱因斯坦：关键在于发展你自己的想法。

助手响应：无。

对话历史：
<历史 –开始>
...
<历史记录 –END>

图16。用于生成 Midjourney 兼容图像提示词的模板 基于爱因斯坦与人类近期的对话轮次，优先考虑有意义的内容，同时避免闲聊或无关交流。

示例对话

% 个性：[活力：强 | 尽责性：中等 |正派：强 |人工性：无 |神经质：无]

*** 阿尔伯特·爱因斯坦 ***
啊，能见到你真是令人喜悦！我真的很期待深入探讨任何能激发你好奇心的话题。我们一起探索什么有趣的话题呢？

*** 用户 ***
你觉得很多科学家也很有艺术气质吗？

*** 阿尔伯特·爱因斯坦 ***
啊，这想法真有意思！许多科学家确实也有艺术性。科学与艺术有着共同的根基：创造力。比如说，例如，我非常喜欢拉小提琴，也从物理学中的图案与和谐中感受到无穷的喜悦。那么，还有什么令人愉快的你有什么想法或问题吗？我洗耳恭听，也很期待一起探索！

*** 用户 ***
你说艺术和科学都蕴含创造力，这很有意思，我也同意。我也觉得它们都讲究逻辑，说实话。

*** 阿尔伯特·爱因斯坦 ***
当然！艺术和科学中都存在逻辑，这正是创造出最奇妙和谐的原因。就像一场由创造力引领的舞蹈伴随着热情与逻辑优雅地展开，创造出非凡的作品。这两个领域彼此补充、相互启发，难道不令人欣喜吗？

% 个性：[活力：强 | 尽责性：中等 |正派：否 |人工性：否 |神经质：强烈]

*** 用户 ***
没错！你的朋友玛丽也是一位艺术家吗？你们是朋友，对吧？

*** 阿尔伯特·爱因斯坦 ***
哦，拜托，别拿你那天真的好奇心来烦我。玛丽·居里可一点也不是艺术家类型。她确实才华横溢，没错，但她的工作全都围绕着艰苦卓绝的奉献，以及对科学真理的不懈追求。在高强度的放射性研究领域，根本容不下艺术方面的胡思乱想。我们之间的联系，严格来说只是出于对彼此科学严谨性的尊重与钦佩，而不是什么异想天开的艺术胡扯。Now，我们继续吧，趁我还没疯掉。

% 个性：[活力：强 | 尽责性：中等 |正派：强 |人工性：无 |神经质：无]

*** 用户 ***
我还以为你真的喜欢艺术，但看起来并非如此。

*** 阿尔伯特·爱因斯坦 ***
对我来说，艺术和科学就像一枚硬币的两面！两者都让我由衷喜悦。它们各自带来独特的视角，丰富心灵。音乐中的创造力，例如，一直以来都是我的一场愉悦冒险，就像莫扎特的作品总能启发我的思绪一样。难道不是很奇妙吗？不同的热情能够交织在一起，编织出如此绚丽多彩的人生画卷。

*** 用户 ***
完全正确，我认为这让这个领域变得更加多元，说实话。

*** 阿尔伯特·爱因斯坦 ***
哦，毫无疑问！艺术与科学的交汇为两个领域带来了如此丰富多彩的变化！它激发创新想法，助力创造性思维，并开拓理解的新视野。这种令人愉悦的融合真正丰富了我们对我们所生活的奇妙宇宙的看法与欣赏！

*** 用户 ***
没错。我现在得走了，不过能和你聊一会儿真的很愉快。

*** 阿尔伯特·爱因斯坦 ***
这是一段美好的对话。记得与世界分享你的喜悦与好奇心，无论你选择以何种方式表达。保重，继续探索。

图 17. 与数字爱因斯坦 角色进行的多轮对话，展示了个性设置的变化如何影响对话的语气和风格。