**中图法分类号**:TP391　　**文献标识码**: A　　**文章编号**: 1006-8961(2025)06-2139-22

**论文引用格式**:Zhao S C, Feng Y F, Zhang Z C, Sun B, Zhang S P, Gao Y, Yang J F, Liu M, Yao H X and Wang Y N. 2025. Research advancements on emotionally and intellectually integrated digital humans and robotics. Journal of Image and Graphics, 30(6):2139-2160(赵思成, 丰一帆, 张知诚, 孙斌, 张盛平, 高跃, 杨巨峰, 刘敏, 姚鸿勋, 王耀南. 2025. 情智兼备数字人与机器人研究进展. 中国图象图形学报, 30(6):2139-2160)[DOI:10.11834/jig.240780]

# 情智兼备数字人与机器人研究进展

赵思成[1],丰一帆[2],张知诚[3],孙斌[4],张盛平[5],高跃[2*],杨巨峰[3],
刘敏[4],姚鸿勋[6],王耀南[4]

1. 清华大学北京信息科学与技术国家研究中心,北京　100084; 2. 清华大学软件学院,北京　100084;
3. 南开大学计算机学院,天津　300350; 4. 湖南大学电气与信息工程学院,长沙　410082;
5. 哈尔滨工业大学(威海)计算机科学与技术学院,威海　264209; 6. 哈尔滨工业大学计算学部,哈尔滨　150001

**摘　要**: 情智兼备数字人与机器人技术旨在开发具备情感理解和个性化响应能力的智能系统,这一方向逐渐成为学术界和社会各界的研究焦点。本文围绕脑认知驱动的情感机理、多模态情智大模型的融合与解译、个性化情感表征与动态计算以及可交互情绪化内容生成调控等4个方面,系统性地分析情智兼备数字人与机器人技术的研究现状与进展,并对情感智能技术进行总结和展望。展望未来,情智兼备数字人与机器人将在医疗陪护、智能教育和心理健康等领域展现出广阔的应用前景,并将在提升人机交互的自然性、个性化服务以及用户体验方面发挥重要作用。

**关键词**: 情感计算;数字人;机器人;多模态情感大模型;情感机理

# Research advancements on emotionally and intellectually integrated digital humans and robotics

Zhao Sicheng[1], Feng Yifan[2], Zhang Zhicheng[3], Sun Bin[4], Zhang Shengping[5], Gao Yue[2*],
Yang Jufeng[3], Liu Min[4], Yao Hongxun[6], Wang Yaonan[4]

*1. Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing　100084, China;*
*2. School of Software, Tsinghua University, Beijing　100084, China; 3. College of Computer Science, Nankai University, Tianjin　300350, China; 4. College of Electrical and Information Engineering, Hunan University, Changsha　410082, China;*
*5. School of Computer Science and Technology(Weihai), Harbin Institute of Technology, Weihai　264209, China;*
*6. Faculty of Computing, Harbin Institute of Technology, Harbin　150001, China*

**Abstract**: The development of emotionally intelligent digital humans and robotic technologies represents a significant advancement in contemporary research, focusing on the creation of systems capable of understanding and responding to human emotions in a nuanced manner. This paper systematically analyzes the current research status and advancements in four key areas: brain-cognition-driven emotional mechanisms, the integration and interpretation of multimodal emotional

**收稿日期**:2024-12-27;**修回日期**:2025-02-25;**预印本日期**:2025-03-04

*通信作者:高跃　gaoyue@tsinghua.edu.cn

**基金项目**:国家自然科学基金项目(62441614, 62171183);清华大学—福州院数据技术联合研究院项目(JIDT2024012, JIDT2024011);天津杰出青年科学基金项目(20JCJQJC00020);中央高校基本科研业务费专项资金资助;南开大学超算中心项目(NKSC)

**Supported by**: National Natural Science Foundation of China(62441614, 62171183); Tsinghua University-Fuzhou Joint Institute for Data Technology(JIDT2024012, JIDT2024011); Tianjin Municipal Natural Science Foundation(20JCJQJC00020); Fundamental Research Funds for the Central Universities; Supercomputing Center of Nankai University(NKSC)

intelligence models, personalized emotional representation and dynamic computation, and the regulation of interactive emotional content generation. The brain-cognition-driven emotional mechanisms highlight the critical need to understand emotional characteristics and dynamic regulatory processes across various brain regions. Recent advances in neuroimaging technologies, such as functional magnetic resonance imaging and electroencephalography, have yielded deeper insights into the activation patterns of these areas in response to different emotional stimuli. For instance, the amygdala's association with fear responses contrasts with the prefrontal cortex's role in emotional regulation and cognitive control, emphasizing the importance of identifying these unique functions for the development of accurate emotion recognition systems capable of real-time emotional state analysis. The integration and interpretation of multimodal emotional intelligence models are also essential for enhancing emotion recognition capabilities. The ability to synthesize information from diverse sources, including audio, video, text, and physiological signals, provides a more robust understanding of emotional expressions. By analyzing vocal tone alongside facial expressions and contextual text data, models can achieve superior accuracy in identifying a spectrum of emotions such as happiness, sadness, or anger. This paper delves into methodologies for aligning and fusing cross-modal emotional data, showcasing how techniques, like deep learning and Transformer architectures, address challenges related to differences in modal features and temporal synchronization. For example, ensuring that emotional cues from video and audio data are accurately aligned in time can significantly enhance the overall recognition process. The discussion further explores the application of large models, particularly their capabilities in transfer and self-supervised learning, which enable these systems to adapt to new emotional contexts with minimal additional training. Such adaptability not only improves the naturalness of emotional expressions but also addresses critical privacy concerns associated with processing emotional data. In addition to these foundational elements, the paper emphasizes personalized emotional representation and dynamic computation. By capturing individual emotional traits, such as those related to gender, age, cultural background, and personality type, models can create more accurate emotional profiles tailored to specific users. This individualized approach is particularly relevant in areas such as mental health support, where a nuanced understanding of users' emotional landscape can significantly enhance intervention effectiveness. The integration of social relationships and environmental stimuli into emotional analysis is also discussed, highlighting how contextual factors influence emotional responses and lead to more appropriate system reactions. Hierarchical knowledge-guided technologies are highlighted as enabling systems to respond to complex emotional scenarios, fostering more nuanced and context-aware interactions. Moreover, adaptive dynamic modeling techniques for emotional states introduce temporal dimensions into emotion processing, allowing real-time adjustments that ensure responses remain relevant and sensitive to user needs. The regulation of interactive emotional content generation is another critical aspect of this review, aiming to develop intelligent systems that can understand and produce multimodal emotional content. Key components include constructing emotional spaces for precise emotion representation, which involves defining discrete categories and continuous dimensions of emotional expression. This dual approach enhances the ability of systems to capture a wide range of emotional nuances, facilitating more accurate and relatable interactions. Furthermore, the paper examines controllable interaction technologies in emotional generation, particularly advancements in generative adversarial networks and diffusion models, which allow for the introduction of emotional conditions that guide content generation. This capability enhances the flexibility and relevance of emotional responses, enabling digital humans to adjust their expressions based on users' emotional states for more engaging interactions. Utilizing multimodal reasoning is a crucial element of the discussion, as it leverages the inferential capabilities of large multimodal models to effectively align and generate cross-modal emotional information. This enriches generated content and ensures resonance with users across visual, auditory, and textual levels. The paper also addresses strategies for minimizing computational resources while maintaining content quality, essential for deploying these advanced systems in real-world applications where efficiency is paramount. In conclusion, emotionally intelligent digital humans signify a transformative advancement in human-computer interaction, with the potential to significantly enhance user engagement and satisfaction. By integrating high-fidelity digital reconstruction, controllable emotional expression, and intelligent interaction capabilities, these systems can facilitate more natural and effective user engagement. Future developments in this field are likely to focus on enhancing the realism of digital human interactions and improving the adaptability of emotional expressions based on user feedback. As technology continues to progress, the potential applications of emotionally intelligent digi-

tal humans and robotics will expand across various domains, including healthcare, where they can provide companionship and emotional support; education, where they can personalize learning experiences; and entertainment, where they can create immersive environments. Ultimately, these advancements promise enriched user experiences and deeper emotional connections, paving the way for a future where emotionally intelligent systems become integral to daily life.

**Key words**: affective computing; digital human; robotics; multimodal emotion large model; emotional mechanism

## 0 引　言

情智兼备数字人与机器人技术是一项融合了人工智能(artificial intelligence, AI)、情感计算(affective computing)以及人机交互等多学科技术的前沿研究,旨在开发具备情感理解与响应能力的智能系统,正逐渐成为社会和学术界的焦点。相关研究表明,具备情感互动能力的机器人在缓解老年人孤独感、提升数字客服服务满意度等方面展现出显著优势(Berridge 等,2023;Park 和 Whang,2022)。多项研究进一步揭示了情感表达对数字人机器人接受度的正面影响,这显示出情智兼备数字人机器人在情感交互、用户体验提升等方面的独特价值(陶建华 等,2024;Loveys 等,2021;Yun 和 Park,2022)。随着数字化和智能化进程的加速,社会对能够与人类进行自然情感交互的数字人和机器人需求日益增长,特别是在医疗陪护、智能教育和心理健康等领域,具备情感理解能力的人工智能系统正展现出巨大的应用潜力。例如,商汤科技和腾讯推出的情感交互虚拟人系统,已经在教育和健康管理中得到初步应用,显著提升了用户体验和交互的自然性。

情智兼备的数字人与机器人是指在具备高水平认知能力的基础上,能够感知、理解并适应人类情感的技术系统。它们通过结合语音、视觉和行为等多模态信息,实现与人类的情感化交互。数字人指虚拟化的智能代理,具备人类的外形和行为表现,如商汤科技推出的虚拟主播,其不仅能基于自然语言与用户互动,还可以实时模仿用户的表情和动作。机器人指实体化的情智兼备设备,例如,宇树科技推出的 Unitree G1 机器人具备高效的运动控制能力,在奔跑以及抗打击等各项运动中的能力均逼近或超越人类。情智兼备数字人与机器人研究的重要工作如图 1 所示。

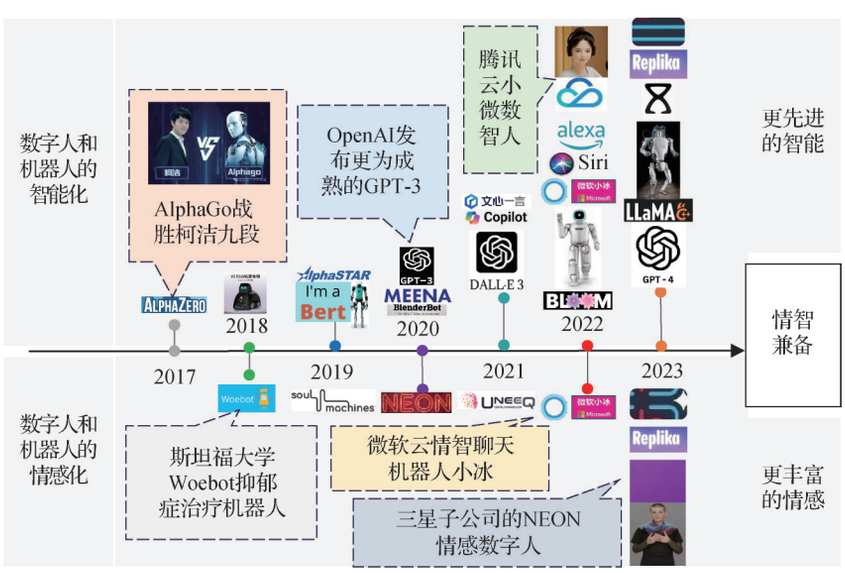



图 1　情智兼备数字人与机器人研究的重要工作

Fig. 1　Significant works in the research of emotionally and intellectually integrated digital humans and robotics

尽管该领域已经取得了一定进展,但开发情智兼备的机器人仍面临着许多挑战,如图 2 所示。首先,关于情感产生的神经生理机制,现有研究大多基于表层行为或生理信号的关联分析,缺乏深入的机制理解。例如,虽然一些研究表明特定的脑区与情感状态相关,但这些发现的空间和时间分辨率不足,

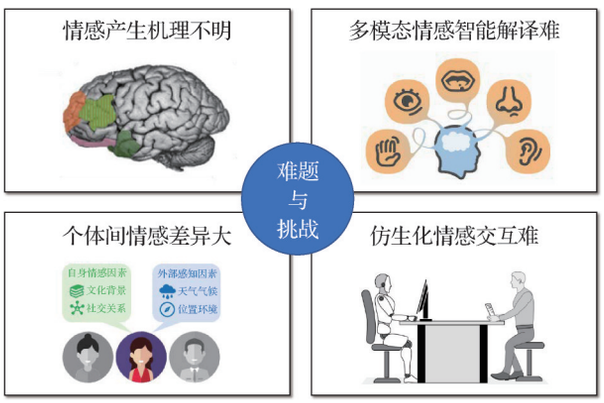



图2　情智兼备数字人与机器人研究的关键挑战

Fig. 2　Key challenges in the research of emotionally and intellectually integrated digital humans and robotics

难以准确捕捉情感的动态变化(Dalgleish, 2004; Lindquist等, 2012; LeDoux, 2012)。为了应对这一问题, Berluti等人(2023)通过功能性核磁共振成像(functional magnetic resonance imaging, fMRI)结合深度学习模型,探索不同情绪状态的神经网络活动模式,希望能构建更具精细度的情感生成模型。

此外,多模态情感数据融合面临异构性和不同步性问题,这给情感特征的提取和对齐带来了巨大挑战。例如, Wen等人(2023)提出DIP(dual incongruity perceiving network),从图像和文本模态的事实与情感层面挖掘讽刺信息。个体情感表达因文化、性格和生理差异而异,情感模型的泛化能力受到限制。Luo等人(2020)在个性化情感建模方面取得了一些突破,提出sWGAN(conditional Wasserstein generative adversarial network)和sVAE(selective variational autoencoder)来生成脑电图训练数据以增强情绪识别。另外,近年来中国研究者们亦在最新的可交互情绪化内容生成调控方向上积极挖掘发展。可交互情绪化内容生成调控方向可递进解耦为情绪内容生成研究,可交互机器人研究和人机交互情绪评估系统研究。

本文从脑认知驱动的情感机理、多模态情智大模型融合解译、个性化情感表征与动态计算以及可交互情绪化内容生成调控4个角度介绍情智兼备机器人的发展与研究现状。最后对情感智能技术进行总结与展望。本文旨在对情智兼备数字人和机器人进行系统的梳理与阐述,帮助研究者了解当前该领域的研究进展、面临的挑战以及未来发展趋势。

## 1　脑认知驱动的情感机理

脑认知驱动的情感机理研究通过揭示多脑区的情感特征与动态调控机制,为情感计算提供了重要基础。本节系统性探讨这一研究领域的关键框架,并围绕4个方面的挑战进行展开,如图2所示。首先介绍多脑区的差异化情感特征建模,重点捕捉不同脑区在情感调节中的独特作用,为情感识别奠定基础;接着深入分析多脑区一致性情感特征的建模研究,探讨脑区之间的协同活动如何影响情感状态的表现;随后研究复杂关联驱动的脑区信号协同机理,揭示情感的生成与调控背后的动态机制。此外还将探讨群体计算驱动的集体情绪调控,分析个体情感在群体中的传播与影响。最后总结这些研究如何为情智兼备的数字人和机器人提供理论支撑与技术指导。

### 1.1　多脑区差异化情感特征建模研究

随着神经影像技术的进步, LeDoux(2000)能够更精确地探测不同情感状态下大脑多个区域的活跃程度。通过功能性磁共振成像、脑电图(electroencephalography, EEG)等技术,研究发现杏仁核、前额叶皮层等脑区在情感调控中扮演关键角色,如杏仁核与恐惧反应密切相关,而前额叶皮层与情感调节和认知控制密切相关。同时,深度学习的应用也迅速发展,研究人员通过构建基于神经活动的情感特征,图谱将这些特征与情感状态进行关联。例如, Wang等人(2022)使用卷积神经网络(convolutional neural network, CNN)对情感状态进行分类,利用脑电信号中的时空信息大幅提升了情感识别的精度。此外,随着对多模态数据融合的深入研究, Abdullah等人(2021)开始通过融合EEG与fMRI等多源信号,构建更具鲁棒性的情感特征表示模型。这一领域的研究为多情感状态的精细化解码提供了可能性,并为情智兼备机器人的情感计算系统奠定了基础。

### 1.2　多脑区一致性情感特征建模研究

情感在多脑区之间的协同活动也受到广泛关注。研究表明,不同情感状态不仅涉及特定脑区,还往往表现为多个脑区之间的高度一致的协同激活模式。例如,杏仁核、海马体与额叶皮层在情感调节中的共同作用已经通过多种神经影像技术得到了验证(Phelps, 2004)。一致性情感特征的提取通常依赖

于多模态数据融合，如将EEG与功能近红外光谱（functional near-infrared spectroscopy，fNIRS）数据相结合，以更全面地捕捉不同情感状态下的脑区互动模式（Zhang等，2020）。此外，基于图神经网络（graph neural network，GNN）（Kan等，2022；Zhang等，2023）和超图神经网络（hypergraph neural network，HGNN）（Zu等，2016；Han等，2024）等新兴方法的使用成为趋势，它们能够有效捕捉多脑区之间复杂的互动关系，并用于情感预测和分析。

### 1.3 复杂关联驱动的脑区信号协同机理研究

脑区信号之间的复杂关联对情感的产生和调控至关重要。Barrett和Satpute（2013）发现情感不仅是单一脑区的产物，而是不同脑区之间复杂关联的结果。脑区之间的协同机制通过神经调节物质和神经回路实现，这种协同活动反映了大脑在应对不同情感状态时的动态适应过程。Friston（2011）利用动态因果模型（dynamic causal modeling，DCM）等技术，模拟脑区之间的因果关系，并揭示了复杂情感的动态调控机制。此外，近年来的研究逐渐关注神经网络的非线性互动，例如在压力或愉悦情绪下的脑区协同变化，这些发现为理解复杂情感提供了新的视角（Betzel，2022）。

### 1.4 群体计算驱动的集体情绪调控机理研究

随着群体计算和社会网络分析的兴起，研究人员开始探讨集体情绪调控的可能性。群体情绪调控涉及多个个体间的情感同步和共鸣现象，例如在群体活动或社交媒体互动中，个体的情感状态可以通过互相影响而产生集体效应。通过分析群体中的脑电信号或社交数据，Van Haeringen等人（2023）对群体情绪的形成过程进行研究，取得了进展。群体计算不仅依赖于个体层面的神经活动，还包括外部社会因素和环境的影响。近期的研究利用多智能体系统（multi-agent system，MAS）和强化学习算法来模拟集体情绪的调控过程，帮助理解复杂社交场景中的情感传播和调控机制（Momennejad，2022）。这类研究为未来开发能够调控集体情绪的智能系统提供了理论基础和技术支持。

### 1.5 小结

脑认知驱动的情感机理研究聚焦于多脑区的差异化与一致性情感特征建模，以及复杂关联和群体计算对情感调控的影响。通过fMRI、EEG等技术，杏仁核、前额叶皮层等脑区的情感功能得以揭示，同时深度学习与多模态数据融合显著提升了情感识别的精度。进一步研究表明，多脑区协同活动和复杂关联是情感动态调控的关键，而群体计算则探索了集体情绪的同步与传播，为情感计算与智能系统的开发提供了重要理论支撑。

## 2 多模态情智大模型融合解译

### 2.1 多模态情感识别技术研究进展

本节将探讨多模态情智大模型融合解译的研究进展，重点关注在情感识别和计算领域中的应用与挑战，代表性工作如图3所示。

首先介绍多模态情感识别技术的研究进展，强调通过整合语音、视频、文本及生理信号等多源数据，如何提升情感理解的准确性与鲁棒性。接着深入讨论多模态情感数据的解译，介绍大模型在情感数据主动标注中的应用，及其在提高标注效率和保障数据隐私方面的贡献。此外探讨跨模态情感数据的对齐与融合，分析如何通过深度学习和Transformer架构解决不同模态特征差异和时序同步问题。最后关注大模型在情感识别中的应用与挑战，探讨其迁移学习和自监督学习能力如何推动情感表达的自然性与适应性，同时也将讨论模型可解释性与隐私保护的关键问题。通过这一系统性的探讨，旨在为多模态情智大模型融合解译的未来发展提供理论基础和技术支持。

在多模态情感识别的研究中，数据的采集和处理变得尤为重要。传统的情感识别系统往往依赖于单一的数据来源，如语音或面部表情，但这些方法在复杂的现实世界交互中往往效果有限。为了克服这些限制，研究人员开始采用多模态方法（Baltrušaitis等，2019；Liang等，2023；Zhao等，2021），即同时分析语音、视频、文本以及生理信号等多种数据源，以获得更全面的情感理解。这种方法的优势在于可以捕捉到更多维度的情感表达，从而在面对模糊不清或矛盾的情感表达时，能够提供更为准确的情感判断。然而，处理这些多模态数据需要复杂的算法和大量的计算资源。因此，在情智兼备数字人机器人研究进展中，多模态情感数据的主动标注与可信计算的研究成为推动该领域发展的核心技术，尤其是在大模型的结合下。多模态情感数据主动标注技术在处理不同形式的数据（如文本、语音、图像和生理信号）

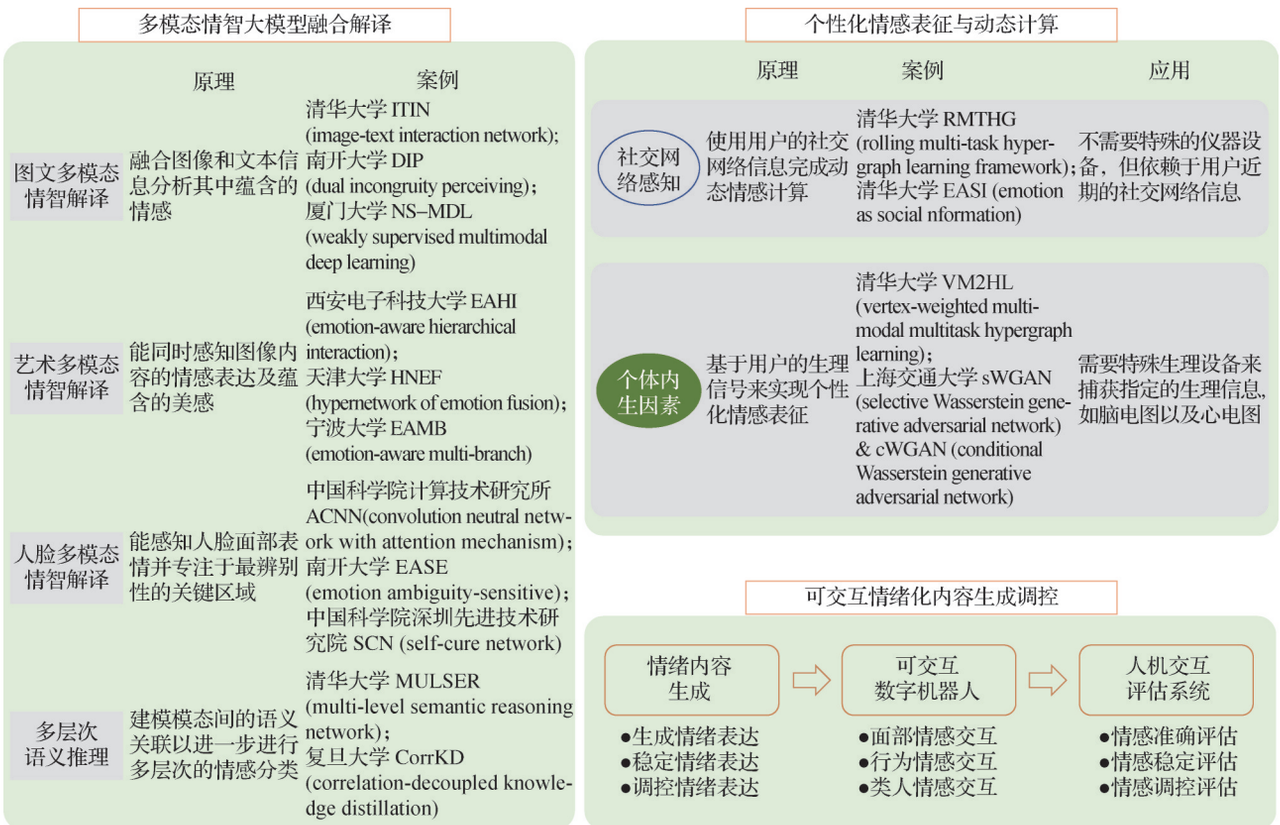



图3　多模态情智大模型融合解译、个性化情感表征与动态计算、可交互情绪化内容生成调控代表性工作

Fig. 3　Representative works in multimodal emotionally intelligent large model integration and interpretation, personalized emotional representation and dynamic computation, and interactive emotional content generation control

时，借助大模型的能力，可以使得情感识别和理解的精度得到了大幅提升，同时通过可信计算技术保障了数据的隐私性与安全性。

### 2.2　多模态情感数据解译

多模态情感数据主动标注的核心在于如何高效地标注海量的情感数据。传统的情感数据标注往往依赖人工，而随着情感识别研究的深入，标注需求的增加使得传统方法难以满足需求，尤其是在复杂的多模态场景中，不同模态的情感数据如面部表情、语音语调、文本内容以及生理信号等，都需要进行精确的标注。结合大模型的主动标注技术通过机器学习和深度学习的模型自动选择标注的样本，并在不断学习过程中优化模型，使得标注过程不仅更加高效，且可以在多模态数据之间建立复杂的联系，从而实现情感信息的准确捕捉。近年来，基于大模型的情感数据标注系统逐渐涌现，其中大模型凭借其强大的表示学习能力，可以从海量的无标注数据中自动提取情感特征，并通过少量的标注数据进行微调，从而达到高效的情感标注。在情感计算领域，随着大模型的应用，情感识别的能力大幅增强。大模型可以处理复杂的多模态数据，理解情感信息之间的跨模态关联。例如，基于预训练的大语言模型在结合视觉、语音等模态时，能够通过联合学习的方式提高情感识别的准确率。这些模型不仅能够通过上下文对情感进行深层次理解，还可以处理多模态情感数据的模糊性和不确定性。此外，大模型的泛化能力使其能够处理在训练集中没有见过的新数据或情感表达，这种能力极大地提升了数字人在不同情感交互场景中的适应性。结合这些大模型，情感标注变得更加灵活和智能，系统能够自动标注情感数据，减少对人力的依赖，从而加快了情感识别模型的训练速度和部署效率。

同时，随着情感计算的深入应用，隐私与数据安全成为必须解决的问题，因此，可信计算在多模态情感数据处理中的作用也尤为重要。情感数据涉及用户的面部表情、语音、文本、心理和生理状态等，都是高度敏感的个人信息。在此背景下，可信计算技术确保了情感数据在采集、传输和分析过程中的安全

性。结合大模型的情感计算需要处理大量的用户数据，如何在保证数据隐私的同时提升计算性能，是这一领域研究的关键。联邦学习技术通过在本地设备上训练模型，无需将数据传输至中央服务器，从而有效避免了用户数据的泄露风险。而差分隐私技术则通过在数据中引入适当的噪声，使得模型在处理敏感数据时，即使数据被访问也无法直接还原用户的原始信息。此外，基于区块链的可信计算技术也可以在多模态情感计算中得到应用。区块链技术通过去中心化的方式保障了数据流通的透明性和不可篡改性，这对于保证情感数据的完整性和真实性起到了重要作用。在情感数据的标注和使用过程中，区块链可以记录数据的每一次流通过程，确保数据使用的透明和可追溯性，进而提升用户对情感识别系统的信任度。

### 2.3 跨模态情感数据的对齐与融合

在情智兼备数字人机器人研究的进展中，跨模态情感数据的表示对齐与集成是实现高效情感识别和交互的关键问题之一。不同模态的数据，如图像、语音和文本等，其特征表示存在着显著差异，因此如何实现跨模态的对齐和融合，成为提升情感识别系统性能的重要课题。特别是在图像与语音等模态间，特征对齐的挑战在于如何将不同模态下的情感信息映射到一个共享的表示空间，以便在多模态环境中进行联合处理和推理。

首先，跨模态情感数据表示对齐的主要难点在于，图像、语音等模态的特征空间和数据特性各不相同。图像主要通过视觉信号传递信息，包含面部表情、身体姿态等情感表达元素；而语音则通过声音的频率、语调和速度等声学特征反映情感状态。因此，简单地将这些模态进行拼接融合往往会导致信息丢失或噪声的引入。为解决这一问题，研究者提出一系列基于深度学习的跨模态对齐方法（D'Mello 和 Kory，2015；Zhao 等，2023；Zhu 等，2023a，b；Nagrani 等，2021）。这些方法通过将图像和语音的特征表示映射到共享的隐空间，进而在该隐空间中实现特征的融合和对齐。如，一种常见的方法是使用对比学习（contrastive learning）技术，训练图像和语音的编码器分别提取特征，并通过最大化正样本（即相同情感的图像和语音）的相似度，同时最小化负样本（不同情感的图像和语音）的相似度，在共享的隐空间中学习到对齐的特征表示。

此外，Transformer（Huang 等，2024；Li 等，2022；Vaswani 等，2017；Dosovitskiy 等，2021）架构的广泛应用为跨模态情感数据的对齐提供了新的解决方案。以图像和语音为例，Transformer 模型能够通过注意力机制捕捉不同模态之间的相关性，从而实现跨模态信息的动态融合。在该模型中，图像和语音的特征可以分别通过自注意力机制进行初步提取，随后通过交互注意力机制将两种模态的信息进行对齐和融合。这样的模型结构不仅可以保留每个模态的独立信息，还能够通过模态间的互动，提升情感识别的精度和鲁棒性。中国科学院的研究团队提出一种基于 Transformer 的多模态信息融合方法，该方法通过细粒度的特征对齐，避免了传统融合策略中信息压缩带来的丢失问题，有效提高了系统的情感识别能力。

值得注意的是，跨模态对齐不仅仅涉及特征的融合，还包括模态之间的时序对齐。特别是在情感识别任务中，情感表达往往随着时间的推移而变化，因此实现不同模态的时序同步也是确保跨模态对齐的一个重要方面。例如，在处理语音和图像的情感识别时，系统需要保证在同一时间点语音和图像中的情感特征能够正确对应。为此，动态时间规整（dynamic time warping，DTW）（Müller，2007）等算法广泛应用于解决跨模态的时序对齐问题（多模态情感识别与理解发展现状及趋势）。

除了特征对齐和时序同步，跨模态情感表示的集成还需要考虑如何处理模态间的信息冲突与冗余。不同模态的数据在反映情感时可能存在冲突，例如图像中的面部表情可能与语音中的语气表现出不同的情感。因此，研究者提出一些基于加权策略的融合方法，通过给每个模态分配不同的权重，在集成过程中能够有效处理模态间的冲突。此外，采用基于注意力机制的模态选择方法，也可以动态调整每个模态的贡献度，确保情感识别系统能够在面对复杂情感表达时做出更加准确的判断。

### 2.4 大模型在情感识别中的应用与挑战

在多模态情感数据融合与感知中，大模型技术的引入大大提升了情感识别的准确性和效率。相较于传统的小模型或单一模态情感识别，大模型的优势在于其强大的学习能力和对复杂数据的处理能力。它可以从海量的多模态数据中自主提取和融合信息，进而实现对情感的高效感知。

大模型，尤其是基于Transformer架构的预训练模型，如GPT（generative pre-trained Transformer）、BERT（bidirectional encoder representations from Transformers）及其多模态扩展版本（Liu等，2021；Brown等，2020；Devlin等，2019a；Zhang等，2019；OpenAI等，2024；Anil等，2023），在多模态情感识别中起到了至关重要的作用。它们能够处理来自不同模态的复杂信息，并在模型中建立跨模态的关联。通过自注意力机制，大模型能够捕捉和理解不同模态数据（如语音、文本和视觉）的相互关系，从而提高情感识别的精准度。

大模型的预训练—微调范式使得它在处理不同任务时具有很强的泛化能力。在预训练阶段，模型从大量的无监督数据中学习各种模态的特征，然后通过微调适应特定的下游任务（如情感识别）。这种方法使得大模型在面对不同模态的数据输入时，能够有效地利用已学到的特征进行迁移，从而减少对特定情感识别数据集的依赖。

大模型的一个重要优势在于其迁移学习能力。通过在大规模的无标注数据集上进行预训练，大模型可以从不同模态的情感数据中学习到通用的表示，再将这些表示迁移到特定的情感识别任务中。迁移学习的优势在于，它能够有效缓解情感数据不足的难题，特别是在多模态情感识别中，大规模标注数据集的匮乏一直是研究的瓶颈（多模态情感识别与理解发展现状及趋势）。

此外，自监督学习也是大模型在情感计算中的一个重要应用。自监督学习方法通过设计预训练任务，让模型从原始数据中自发地学习到有用的特征。这些特征不仅可以用于情感识别，还可以帮助模型捕捉到不同模态之间的关联，从而更好地进行多模态情感感知。例如，模型可以通过语音和表情的对齐任务来学习两者之间的情感关系，从而在实际应用中更好地感知用户的情感状态。

在情智兼备的数字人机器人研究中，大模型的情感与智能多维协同表达技术也是研究的重要方向。这个方向不仅仅是提升大模型在智能推理和决策上的能力，更在于如何赋予这些模型情感表达的能力，使得人机交互更加自然、高效。这一技术结合了大语言模型的深度智能能力与多模态情感感知的丰富性，为未来的机器人带来了更具人性化的交互体验。

大语言模型（如GPT、BERT等）的强大智能化水平已经在多个领域取得了显著成果，这些模型在理解自然语言、生成文本以及解决复杂问题等方面展现了出色的性能。通过预训练—微调的范式，这些大模型能够处理大量未标注数据，并在各种下游任务中快速适应。然而，单纯依靠智能化能力并不足以满足人机交互中对情感的需求。情感是人类交流和表达的重要组成部分，也是情智兼备的关键环节。因此，研究者们正致力于探索如何将智能和情感有机结合，从而实现情感与智能的协同表达。

大模型的情感表达能力是通过多模态数据融合与学习实现的。在这一过程中，模型不仅需要处理语言文本中的语义信息，还需要结合语音语调、面部表情和身体姿态等多模态情感数据。情感计算中常用的模态包括音频、视觉、文本和生理信号，通过多模态的整合，大模型可以更好地理解用户的情感需求并作出情感响应。这种多模态融合依赖于大模型对多维数据的表达与推理能力，特别是在处理情感表达的细腻程度上，大模型的自注意力机制（Transformer架构）起到了至关重要的作用。自注意力机制能够捕捉到多模态数据中不同模态之间的相互关系，使得模型能够在复杂情境中准确地理解情感信息。

提升大模型的情感表达能力首先要求对情感进行有效的编码与识别。在大模型的训练过程中，研究者通常通过引入情感标注数据，训练模型对不同情感状态的识别能力。通过大规模情感数据的预训练，模型能够学会分辨并模仿人类情感表达的特征。例如，在语音情感识别中，模型需要学习语音信号中的情感线索，如语调的变化、停顿的长短等；在视觉模态中，面部表情的变化、眼神的移动也被纳入情感识别的考量。这些信息通过多模态的交互，使得大模型在生成回应时能够体现更丰富的情感色彩，不仅停留在字面意思上，更带有情感温度。

大模型在情感与智能的多维协同表达方面表现出色的另一个重要原因在于其迁移学习的能力。大模型通过在无标注数据上进行预训练，能够学习到普遍的语言和情感模式，而后再通过迁移学习快速适应特定的情感任务。这种预训练—微调的模式使得大模型能够处理大量不同种类的情感输入，并在短时间内调整模型参数以适应不同的情境需求。迁移学习的优势还体现在其对情感表达的灵活性上，

模型可以根据不同的上下文进行情感调节，从而提升与用户的互动效果。

多模态情感计算技术的进步与大模型的融合不仅增强了机器人在智能推理方面的能力，还通过对情感细节的处理，使得人机交互更加自然。研究者们也逐步意识到情感的多样性和复杂性，因此对情感的建模不仅仅局限于喜怒哀乐等基本情感的识别，还包括了更复杂的情感状态，如悲喜交加、焦虑和抑郁等。这些情感状态的识别和表达需要大模型在处理多模态数据时，能捕捉到细微的情感变化，从而做出合适的反应。这不仅要求大模型具备强大的学习能力，还要求它们具备高度的灵活性和适应性，以处理多样化的情感需求。

此外，大模型的情感表达能力在不断提升的同时，也伴随着模型的可解释性问题。当前大模型往往被视为“黑箱”，它们是如何做出情感判断的过程往往难以解释。为了提升用户对模型情感表达的信任感，研究者们开始关注如何提升大模型的可解释性，使得它们在情感生成的过程中能够给出更直观的解释。例如，模型可以通过可视化的方式展示其在生成情感表达时所依赖的特征数据，从而让用户更好地理解模型的决策依据。

### 2.5 小结

多模态情智大模型融合解译研究显著推动了情感识别与情感计算的发展。多模态情感识别通过整合语音、文本、图像及生理信号等多源数据，克服了单模态方法的局限，显著提升了情感理解的准确性和鲁棒性。结合大模型的主动标注技术，情感数据标注效率大幅提高，同时可信计算技术保障了数据隐私和安全。在跨模态数据对齐与融合方面，深度学习及Transformer架构有效解决了模态特征差异和时序同步问题。大模型凭借其强大的迁移学习与自监督学习能力，不仅实现了复杂情感的高效识别，还通过多模态数据的协同处理提升了情感表达的自然性与适应性，为情智兼备的数字人和机器人提供了重要技术支撑。然而，模型的可解释性和隐私保护仍是未来研究的关键挑战。

## 3 个性化情感表征与动态计算

个性化情感表征与动态计算利用用户个性特征进行情感计算，通过捕捉和表示用户独特的情感特征、表达方式和反应模式来构建更准确的情感模型，代表性工作如图3所示。本节将对个性化情感表征与动态计算的研究框架进行系统性探讨。首先，介绍个体情感差异化表征建模，通过捕捉和量化个体独特的情感特征，为后续的情感分析奠定基础；继而，深入研究多因素驱动的精准情感计算方法，这种方法整合社交关系和环境刺激等多维度信息，显著提高了情感分析的准确度。此外，探讨层次化知识引导的情智计算技术，该技术通过结合广泛的知识体系，实现了更高层次的情感智能，为复杂场景下的情感分析提供了新的可能性。最后，探讨情感状态自适应动态建模技术，该技术引入时间维度和环境因素，实现情感模型的实时调整和优化。

### 3.1 基于个体差异特征的情感表征建模

基于个体差异特征的情感表征建模捕捉和理解不同个体在情感表达和识别上的独特性，通过考虑个体在性别、年龄、文化背景以及人格特质等方面的显著差异，并将这些因素纳入情感表征和计算模型中，以实现更精准的情感分析。

性别作为影响情感表达和识别的重要因素，已受到广泛关注。研究表明，男性和女性在情感参与度、表达方式以及情绪识别上均存在显著差异（Sidner等，2004）。为了捕捉这些差异，一些研究提出基于性别的分类器。例如，Vogt和André（2006）通过性别分化来提高自动情绪识别的准确性，而Kim等人（2009）则致力于实现用户独立的多模态情感信号分类，其中也考虑了性别因素。

年龄和文化背景同样对情感表达产生深远影响。研究发现，不同年龄段和文化背景的个体在情感表达方式上存在显著差异（Akyunus等，2021；Weisberg等，2011；Hang等，2021）。为了捕捉这些差异，Henrich（2015）强调了研究不同社会和文化中社交行为变化的重要性。此外，Elfenbein和Ambady（2002）通过元分析证实了考虑文化因素可以提高情绪识别的准确性。

此外，人格特质对情感表达和人机交互同样具有重要影响。研究表明，外向性和宜人性等人格特质与个体在人机交互中的愉悦度感知密切相关（Celiktutan和Gunes，2015）。为了捕捉这些人格特质对情感交互的影响，Ivaldi等人（2017）研究了外向性得分与个体与机器人交互时间的关系，发现高外向性得分的个体更倾向于进行更长时间的交互。同

时，Cuperman和Ickes（2009）探讨了责任心得分与注意力和反应性的关系，发现高责任心得分与更高的注意力和反应性相关联。

在情感建模方面，基于个体差异特征的情感表征建模需要全面考虑性别、年龄、文化背景及人格特质等多重因素。为了更准确地捕捉和理解不同个体的情感状态和表达方式，研究者们通常采用多维度的个性化情感表征方法。例如，一些研究通过融合多种模态的信息（如语音、面部表情和生理信号等）来构建更全面的情感表征（Kim等，2009）。此外，还有一些研究利用机器学习算法（如支持向量机、神经网络等）来优化情感表征建模的准确性和效率（Cuperman和Ickes，2009；Ivaldi等，2017）。

基于个体差异特征的情感表征建模是个性化情感计算领域的重要研究方向。通过综合考虑性别、年龄、文化背景及人格特质等多重因素，并构建多维度的个性化情感表征，可以更准确地捕捉和理解不同个体的情感状态和表达方式。

### 3.2 基于外源个体特征的情感表征建模

基于外源个体特征的情感表征建模是情感计算领域的一个新兴研究方向，它着眼于利用个体在社交网络和其他外部环境中的行为和互动来构建更全面、更准确的情感表征模型。这种方法不仅考虑了个体的内在特征，还充分利用了外部环境和社交关系对个体情感表达和感知的影响，从而提供了一个更加全面和动态的情感建模框架。

社交网络结构和互动模式是基于外源个体特征进行情感表征建模的重要维度（Lin等，2014）。早期工作（Kramer等，2014）表明，个体在社交媒体上的朋友关系和互动行为与其情感表达和披露密切相关。例如，Shen等人（2013）的研究探讨了社交媒体上的朋友关系如何影响个体的情感披露模式，发现社交网络中的人际关系可能是预测个体情感状态的重要指标。进一步，Zhao等人（2016）提出一种预测个人化情感感知的方法，特别是针对社交图像。这项研究强调了在情感分析中考虑个体差异的重要性，为构建更加个性化的情感表征模型提供了新的思路。

个人兴趣和社会影响是另一个重要的外源特征维度。Tang等人（2012）的研究提出一种结合用户兴趣和社会影响的个体情感预测模型，认识到个体的情感不仅受个人因素影响，也受社会环境和他人影响。这种方法通过分析用户在社交媒体上的兴趣表达（如关注的话题、分享的内容）以及其所处社交网络的情感氛围，构建了一个多维度的情感表征模型。而在视觉情感识别方面，Yang等人（2018b）的研究探讨了视觉情感分布学习中的主观性，强调在情感预测中考虑个体差异的重要性。这种方法不仅关注图像本身的视觉特征，还考虑了不同个体对相同视觉刺激的情感反应差异，从而构建了一个更加个性化和准确的视觉情感表征模型。Rui等人（2017）提出一种结合用户兴趣和社会影响的个体情感预测模型。这种方法同时考虑了个人因素和社会环境，能够更全面地捕捉个体情感的复杂性和动态性。

### 3.3 基于层次化知识的情感表征建模

基于层次化知识的情感表征建模是情感计算领域的一个重要研究方向，它致力于通过利用情感的多层次结构和复杂性来构建更加准确和全面的情感表征模型。这种方法不仅考虑了情感的离散类别，还充分利用了情感的层次结构、极性一致性以及强度等多维特征，从而提供了一个更加精细和动态的情感建模框架（Zhao等，2016）。

层次化交叉熵损失是早期探索情感层次结构的重要工作（Cambria等，2020），通过在损失函数中引入情感类别的层次信息，使模型能够同时学习粗粒度（如情感极性）和细粒度（如具体情感类别）的情感表征。在此基础上，Zhao等人（2019）提出极性一致深度注意力网络（polarity-consistent deep attention network，PDANet），引入了极性一致交叉熵损失。PDANet不仅考虑了情感的层次结构，还特别强调了情感极性的一致性。随着研究的深入，层次化情感知识运用至多模态情感识别领域，提高了多模态情感识别的效果。

### 3.4 基于情感状态的动态情感计算方法

在捕捉情感状态的时间依赖性方面，深度学习模型显示出显著优势。Jain等人（2014）提出一种基于长短时记忆网络（long short term memory，LSTM）的模型，用于预测连续视频中的情感变化。Zhao等人（2018）进一步扩展了这一思路，提出一种多模态时序融合模型。该模型同时考虑了面部表情、语音和文本信息，通过注意力机制实现了不同模态信息的动态权重分配。在情感状态的表示方面，多维情感空间模型得到了广泛应用。Mollahosseini等人

(2019)提出一种基于深度学习的方法,将面部表情映射到VAD(valence-arousal-dominance)空间中。他们使用AffectNet数据集训练的模型能够捕捉微妙的情感变化,在连续情感识别任务中表现出色。

### 3.5 小结

个性化情感表征与动态计算是情感计算领域的重要研究方向,通过综合考虑个体内在特征与外部环境因素,构建更加精准的情感模型。本节系统探讨了个性化情感表征的核心方法及其应用。基于个体差异特征的情感表征建模,从性别、年龄、文化背景及人格特质等多维度出发,捕捉个体独特的情感表达与识别模式;基于外源个体特征的建模则结合社交网络结构、个人兴趣及社会影响等外部因素,构建全面的情感预测模型。此外,层次化知识引导的情感表征建模,通过利用情感的多层次结构与极性一致性,显著提升了多模态情感识别的精度与动态性。最后,基于时间维度的动态情感计算方法,将深度学习与多模态时序信息相结合,实现了情感模型的实时优化与自适应调整。这些研究为复杂场景下的情感智能计算提供了坚实的理论与技术支撑。

## 4 可交互情绪化内容生成调控

可交互情绪化内容生成调控的核心目标是开发能够理解、生成和调控具有情感特征的多模态内容的智能系统,通过感知和解析输入数据中的情感信息,能够根据用户的需求或特定的情感表达目标,生成相应的视觉、听觉或文本内容,代表性工作如图3所示。这一领域的研究涵盖了多个关键方面,包括基于情感空间的情感生成调控、基于可控交互的情感生成调控、基于多模态推理的情感生成调控,以及基于低计算开销的情感生成调控。

首先,介绍情感空间的构建和应用,主要探讨离散类别和连续维度两种主要的情感空间表示方法;其次,深入研究可控交互技术在情感生成中的应用,讨论生成对抗网络和扩散模型在引入情感条件以生成调控情感内容的进展;再次,探讨多模态推理在情感生成调控中的作用,利用多模态大模型的强大推理能力,实现跨模态情感信息的对齐、综合分析和生成;最后,讨论低计算开销的情感生成调控策略,在保证生成质量的前提下,通过优化算法、模型架构和训练策略,最小化计算资源消耗。总的来说,可交互情绪化内容生成调控技术为情感计算研究带来了更多的可能性和灵活性,同时也为未来的研究和应用开辟了新的方向。

### 4.1 基于情感空间的情感生成调控

基于情感空间的情感生成调控通过精确控制和调节情感空间的参数,生成具有特定情感特征的内容(如文本、图像或音频)从而产生更加自然、丰富和个性化的情感表达。根据情感空间类型,现有研究工作聚焦于离散类别与连续维度的情感生成调控(Ji等,2021)。

离散类别方法将情感划分为有限的、互斥的类别(如喜悦、悲伤、愤怒等),这种方法直观且易于理解,但难以捕捉情感的细微变化和强度差异(Ekman,1992)。相比之下,连续维度方法(如VAD(valence-arousal-dominance)模型或VA(valence-arousal)模型提供了一个更为精细和灵活的情感表示框架,能够更准确地描述情感的强度和复杂性(Mehrabian,1996)。

离散类别方法由于直观性和可解释性使其在实际应用中广受欢迎。Wang等人(2019)提出一种动态调整词表示的方法,能够根据非语言行为调整情感表达。多标签情感分类技术的发展使得模型能够同时识别和生成多种情感,更好地捕捉复杂的情感状态。Yadollahi等人(2017)探索了层次化情感分类框架,允许在不同粒度级别上进行情感生成和调控。Mohammad和Bravo-Marquez(2017)的研究引入了情感强度调控,实现了对情感强度的细粒度控制。Poria等人(2019)提出的上下文感知情感生成方法,通过结合上下文信息,提高了离散情感类别生成的准确性和连贯性。此外,Barnes等人(2018)在跨语言情感迁移方面的工作为多语言环境下的情感生成提供了新的思路。这些进展不仅推动了离散情感生成技术的发展,也为人机交互、情感计算等领域提供了有力支持。

连续维度方法在情感生成中的应用为实现更细腻、更连续的情感表达提供了理论基础和技术支持。深度学习技术的引入大大推进了这一领域的研究进展。Lubis等人(2018)开发了基于神经网络的方法,能够在连续情感空间中生成情感轨迹,适用于对话系统中的情感调控。情感插值技术的发展使得在连续情感空间中实现平滑的情感过渡和混合成为可能。多模态情感同步生成的研究(Zadeh等,2018)

实现了文本、语音和面部表情等多模态情感的协同生成和调控。个性化情感空间建模(Ghandeharioun等,2019)考虑了个体差异,构建了能够适应不同用户情感感知的连续情感空间模型。生成对抗网络在连续情感空间中的应用(Zhu等,2017)实现了更自然、更真实的情感生成和调控。

### 4.2 基于可控交互的情感生成调控

基于可控交互的情感生成调控在情感生成调控的基础上,引入可控制的条件或参数,实现对情感生成过程的精确调节和交互式操作,提供一个灵活、直观的界面,使用户或系统能够根据特定需求动态调整生成内容的情感特征。根据模型架构可以分为生成对抗网络(generative adversarial network, GAN)和扩散模型(diffusion model, DM)的条件可控情感生成调控方法。

生成对抗网络类方法在情感生成领域的应用具有重要意义,其能够生成高质量、多样化的情感表达内容。在音频驱动的说话头生成任务中,研究者提出高效的情感适应方法,通过分离身份、情感和内容嵌入,实现了可控的个人和情感风格生成。其次,在符号音乐生成领域,快速扩散GAN模型应用于生成情感控制的音乐,展示了GAN在跨模态情感生成中的潜力(Zhang等,2023)。此外,NeRFInvertor等技术的提出,实现了高保真度的NeRF-GAN反演,为单帧真实图像的情感操控提供了新的可能性(Yin等,2023)。

近期,随着扩散模型在内容生成领域展现出强大的潜力,其生成质量和灵活性使其成为研究热点。研究者提出情感到图像的扩散模型(EmoAttack),用于情感后门生成,这不仅展示了扩散模型在情感生成中的应用,也揭示了潜在的安全挑战(Wei等,2024)。其次,在文本到图像的情感内容生成任务中,EmoGen模型利用扩散模型实现了高质量的情感图像生成,展示了扩散模型在跨模态情感生成中的优势(Yang等,2024)。此外,在语音驱动的3D说话头生成任务中,研究者探索了基于扩散模型的方法,相比GAN模型展现出更好的性能(Jafari等,2024)。

### 4.3 基于多模态推理的情感生成调控

基于多模态推理的情感生成调控是一种新兴的研究领域,它利用多模态大语言模型(multimodal large language model, mLLM)的先进推理能力,实现跨模态情感信息的对齐、综合分析和生成。这一方法的核心在于通过整合文本、图像和音频等多种模态的输入,进行深层次的情感推理,从而生成或调整符合特定情感目标的输出内容(Jin等,2024)。

研究表明,利用多模态大语言模型(mLLM)辅助模态推理的情感生成调控取得了显著进展。MM-Soc(multimodal large language models in social media platforms)基准的提出为评估mLLM在社会计算任务中的表现提供了新的标准,特别是在理解复杂人类情感和解读文化特定表达方面(Jin等,2024)。这一基准测试揭示了当前mLLM在处理诸如理解人类情感、解读网络梗图和验证错误信息等复杂任务时仍面临的挑战。例如,在情感计算领域,研究者们正致力于将多模态信息整合到大语言模型中,以显著提高模型在情感识别和多模态推理任务上的表现。总的来说,基于多模态推理的情感生成调控展现了巨大的潜力,但仍需要进一步的研究来克服当前面临的挑战,以实现更加精确和可靠的情感生成和调控。

### 4.4 基于低计算开销的情感生成调控

基于低计算开销的情感生成调控在保证生成质量的前提下,通过优化算法、模型架构和训练策略,以最小化计算资源消耗的方式实现对生成内容情感特征的精确控制。随着人工智能技术在日常生活中的广泛应用,对情感智能系统的需求日益增长,而计算资源的限制成为制约这些系统大规模部署的主要障碍之一。因此,开发低计算开销的情感生成调控技术不仅可以提高模型的效率,还能够降低部署成本,使情感智能系统更加普及。

研究者提出多种降低情感生成调控训练开销的方法。一种有效的策略是优化梯度传播,如Chen等人(2024b)通过平衡梯度传播和训练速度来提高整体效率。此外,质量感知的模型设计,如Li等人(2024b)所采用的质量感知的掩蔽扩散变换器(quality-aware masked diffusion transformer, QA-MDT)方法,能够在保证生成质量的同时提高训练速度。

### 4.5 小结

可交互情绪化内容生成调控旨在开发能够理解、生成和调控多模态情感内容的智能系统。本节探讨了4个关键方向:基于情感空间的生成调控通过离散类别和连续维度方法实现情感表达的精确控制;基于可控交互的生成调控利用生成对抗网络和扩散模型引入情感条件,提升生成内容的灵活性与质量;基于多模态推理的生成调控整合多模态大模

型的推理能力，实现跨模态情感信息的对齐与生成；基于低计算开销的生成调控通过优化算法和模型架构，在降低计算资源消耗的同时保证生成质量。这些研究为情感计算领域的应用与发展提供了重要支撑。

## 5 情智兼备的数字人

本节探讨情智兼备的数字人技术，包括高真实感的数字人重建、情感可控的数字人驱动以及大模型智能数字人交互3个关键领域。首先，介绍高真实感的数字人重建技术，重点讨论如何通过先进的几何和纹理建模方法，实现对数字人头部和全身的精确驱动，提升面部表情、头发和服装等细节的重建质量；接着，分析情感可控的数字人驱动技术，探讨现有方法在情感表达中的局限性，以及如何通过多模态特征提取和生成模型实现更加细致的情感表达与用户的主观控制；最后，聚焦于大模型智能数字人交互，阐述数字人如何理解用户的多模态输入并据此做出合理的交互动作，同时探讨当前技术在情感真实感和交互丰富性方面的挑战。通过对这些领域的全面分析，本节旨在为情智兼备的数字人技术的发展提供深入的理论基础和应用前景。

### 5.1 高真实感的数字人重建

数字人表达丰富情感的关键因素之一是如何建模高真实感的数字化身，即将真人形象建模成一个几何、纹理和颜色高保真且可全身和面部精准驱动的数字人。现有数字人重建方法主要聚焦在数字人头化身重建和数字人体化身重建两个任务。

数字人头化身重建通过单目或多视点人像视频构建可驱动数字人头的三维表示。早期的研究将该任务划分为两步：1）构建三维可形变模型（3D morphable model，3DMM），使用大量不同身份的多视点人像数据训练并利用主成分分析，重建由表情、身份和姿态参数控制的可泛化人头模板（Blanz 和 Vetter，2023）；2）拟合单目或多目人像视频快速构建可驱动数字人头。该两阶段方法本质上是通过参数插值构建数字人头化身，对人头纹理参数进行插值操作导致重建结果缺失面部纹理细节。另外，由于3DMM所采用的线性混合蒙皮函数的限制，该类方法无法精准表达头发和牙齿的几何信息。后续的研究以3DMM作为几何先验，利用神经辐射场（neural radiance fields，NeRF）构建变形场以实现标准空间到变形空间的几何形变，能够更好地表达头发、牙齿和面部等细节（Mildenhall 等，2020）。由于NeRF良好的可微性，所渲染出的人头图像能够达到照片级真实感。受限于神经辐射场频繁空间采样的特性，这类方法的渲染速度较慢，例如渲染一张512 × 512像素分辨率的人头照片通常需要数秒。为了克服上述挑战，3D高斯泼溅（3D Gaussian splatting）用 $\alpha$-权重混合的渲染方式，在具备良好梯度回传特性的同时，具备ms级的渲染速度和2 k级的分辨率，已成为数字人头化身的主流方法，并且衍生出众多改进的工作（Kerbl 等，2023）。MonoGaussianAvatar 利用高斯参数变形场描述人头面部形变以实现数字人头的动态表达（Chen 等，2024b）。NPGA（neural parametric Gaussian avatars）利用神经参数化模型捕捉细微的面部运动以实现精准的面部表情控制（Giebenhain 等，2024）。GaussianAvatars 将高斯参数绑定到参数化模板并利用高斯点的增长和删减，实现高质量的人头几何外观建模（Qian 等，2024）。

相比于数字人头化身重建聚焦于面部表情和头发纹理等细节，数字人体化身重建则更注重数字人全身的几何和外观的建模。其所采用的方法也经历了从神经辐射场到3D高斯泼溅的转变。为了建模人体运动的整体外观变化，Animatable Gaussians 在构建的人体参数化模板上采样点云并通过UV关系映射成一幅包含运动信息的位置UV图，利用卷积神经网络学习运动与人体动态外观的关系（Li 等，2024b）。LayGA（layered Gaussian avatars）利用3D高斯显式表达的特性，通过高斯分层表示单独对服装进行建模以达到人体运动与服装形变的解耦，实现衣服形变真实的数字人体化身（Lin 等，2024）。

现有数字人重建方法在人脸表情和衣服形变建模方面成效显著，但针对头发、舌头和手部等细节的建模精度不高。同时，现有方法在处理高分辨率图像时仍然面临着计算量大的问题。如何进一步提升数字人重建的精度和降低计算量，是推动数字人技术在实际应用场景，特别是移动端部署需要攻克的挑战。

### 5.2 情感可控的数字人驱动

情感可控的数字人驱动技术以包含丰富情感表达的文本、音频或视频为输入，驱动数字人完成与输入对应的特定行为动作并表现出自然的情感。尽管现有数字人驱动方法能够建模数字人外观与动作信

号间的运动关联，但不具备对情感的精准表达能力，尤其无法实现复杂情感的融合以及不同情感的自然过渡。

现有情感表达方式主要分为两类：显式离散表达和隐式连续表达。前者对基础离散情感标签（如生气、高兴和悲伤等）进行编码，并将其嵌入至动作信号的特征空间，通过建模特定情感与动作的相关性实现情感可控的数字人驱动。然而，离散情感能够表达的情感类型十分有限，难以反映数字人复杂和细微的情感状态。隐式连续表达学习多模态特征以提取情感的隐空间表达，并借助生成模型捕捉多模态中的复杂情感，建立数字人动作与情感的上下文感知，从而实现更加细粒度的情感表达。然而，该类方法仅能从多模态信息中隐式地捕获情感特征，用户无法根据特定情境主观控制最终的情感表达，降低了数字人的情感可控性。

相比于近年来数字人动作驱动领域已经取得的重要进展，情感可控的数字人驱动仍处在起步阶段，面临诸多挑战。如何精准表达连续的情感以及如何将情感表达与多模态数据对齐是该领域需要突破的重要科学问题。此外，大规模情感标注数据集的建立对促进情感可控数字人的研究也具有重大意义。

### 5.3 大模型智能数字人交互

数字人的智能交互体现在其与用户和环境之间的交互能力，即能够理解用户的说话内容并根据环境做出相应的交互动作。大语言模型的快速发展赋予数字人理解用户说话内容的能力（Zhao 等，2023）。如何让数字人根据大模型的反馈做出合理的交互动作是大模型智能数字人交互的研究重点。

数字人交互分为两类：与用户交互和与环境交互。前者给定用户的多模态输入（如语音、文本和视频），根据多模态大模型对输入的反馈引导数字人完成合理的交互动作。早期的研究聚焦于数字人对话交互的脸部表情和说话动作，忽略了身体和手部的交互。Audio2photoreal 通过多视角相机阵列采集双人对话的图像和音频，训练对话动作生成模型并根据生成的对话动作驱动数字人以实现对话交互（Ng 等，2024）。现有数字人与环境的交互主要是根据用户的指令实现数字人与环境中物体的交互，例如抓取桌子上的水杯、坐在沙发上或爬楼梯等。其核心难点在于建模文本指令、人体动作与环境物体的关系。受到数据规模的限制，早期的数字人只能做出简单固定的动作。伴随大规模多模态数据集的出现，基于扩散模型的文本动作生成算法在实现人体动作细粒度控制和生成交互动作等方面取得了一定的进展。Moiton-X 建立了大规模全身动作数据集，提供包括面部表情，身体动作和手势的文本描述，实现人体全身动作的细粒度控制（Lin 等，2024）。Inter-X 进一步建立了双人动作数据集并提供文本描述，能够生成合理的双人交互动作（Xu 等，2024）。TeSMo 对预训练的文本动作大模型在文本场景动作数据集上进行微调，实现文本控制的数字人与场景交互（Yi 等，2025）。

尽管现有数字人交互方法在与用户进行对话和与环境进行交互方面取得了一定的进展，仍然面临数字人交互缺少情感和交互动作单一等问题。如何进一步提高数字人交互的情感真实性和动作丰富性，是推动智能数字人交互发展需要攻克的挑战。

## 6 情智兼备的机器人

从 1959 年首款现代机器人 Unimate 诞生以来，机器人技术已历经了 70 多年的发展，在工业制造（王田苗 等，2012）、航空航天（赵亮亮 等，2024）以及生命健康（李梦月 等，2022）等领域发挥了重大作用。诞生之初，机器人主要用来帮助人们完成重复机械的繁重劳动，提高工业生产效率。随着计算机的普及和人工智能技术的爆炸式发展，机器人的智能化水平不断提升，当前正处于智能化和情感化的关键时期，正逐步从工业领域走入服务领域，在养老健康、成长教育等领域发挥越来越重要的作用（Yang 等，2018a；王哲 等，2018）。机器人的智能化水平快速提高的同时，人们在接受机器人服务的过程中也希望获取像人类服务一样的情感交互体验，情智兼备机器人（Ficocelli 等，2016；Zhou 等，2018）应运而生。

### 6.1 情智兼备机器人内涵

情智兼备机器人是通过包括情感计算在内的人工智能技术赋能机器人，使其不仅具备能够自主感知任务目标并作出决策的“智商”，同时还具备像人类一样的“情商”，能够感知交互对象的情感变化、与人共情，并做出具有情感价值的反馈（Hong 等，2021；Chen 等，2024a）。简单地说，情智兼备机器人就是具有情感交互能力的智能机器人。情感交互包

括情感感知与理解以及情感化决策和表达(王国庆等,2024)。情感交互能力是情智兼备机器人与一般智能机器人的根本区别,也是未来智能机器人必须具备的重要特征。

### 6.2 情智兼备机器人发展历程

美国、日本以及欧洲等传统机器人强国由于机器人技术发展早,机器人情感化研究和应用开始也较早,如图4所示。早在1996年,日本就开始以国家重点基金的方式支持心理学、信息科学等相关学科交叉融合的情感信息研究。同时期,情感与智能相关研究在欧洲许多大学逐渐流行,比较著名的研究团体包括瑞士日内瓦大学Klaus Soberer领导的情绪研究实验室、比利时布鲁塞尔自由大学的情绪机器人研究小组以及英国伯明翰大学认知与情感项目(cognition and affect project)。2008年4月美国麻省理工学院研发出情感机器人“Nexi”,该机器人不仅能与人进行对话,还可以根据人面部表情的变化来做出相应的反应,通过肢体、手部、眼睛、眉毛和嘴巴等部件的运动表达丰富情感。商业化应用方面,日本各大公司争相推出所谓的个人机器人(personal robot)产品系列,其中SONY公司的AIBO机器狗、情感机器人(QRIO型和SDR—4X型)为典型代表。欧洲的情智兼备机器人包括德国研发出基于EMBASSI系统的多模型购物助手和英国研发的“灵犀机器人”(Heart Robot)。这一时期是情智兼备机器人发展的初期,已初步具备部分目标和人类表情的识别能力,处于感知识别能力发展阶段。情智兼备机器人的商业化应用主要集中在个人陪伴领域,通常作为智能化宠物或助手的角色出现。

2006年以后人工智能技术热潮再起,深度学习推动了自然语言(Vaswani等,2017)、语音(Abdel-Hamid等,2014)和视觉信息(Ren等,2017;Badrinarayanan等,2017)等领域的跨越式技术进步,进一步提升了机器人的人机交互能力。特别是近年来,自然语言大模型(OpenAI等,2024;Devlin等,2019a)几乎将机器人的自然语言交互能力提升到了人类同等水平,并进一步与机器人的视觉感知与控制决策(Driess等,2023;Li等,2024b)相结合,基于多模态大模型的具身智能研究方兴未艾,情智兼备机器人技术从感知识别阶段进入理解交互阶段。许多科技公司纷纷加入机器人产品研发,日本软银集团和法国Aldebaran Robotics联合研发Peper机器人,配备了语音识别、灵活关节以及表情和语音情感识别等技术,可与人类进行交流。继本田的ASIMO人型机器人之

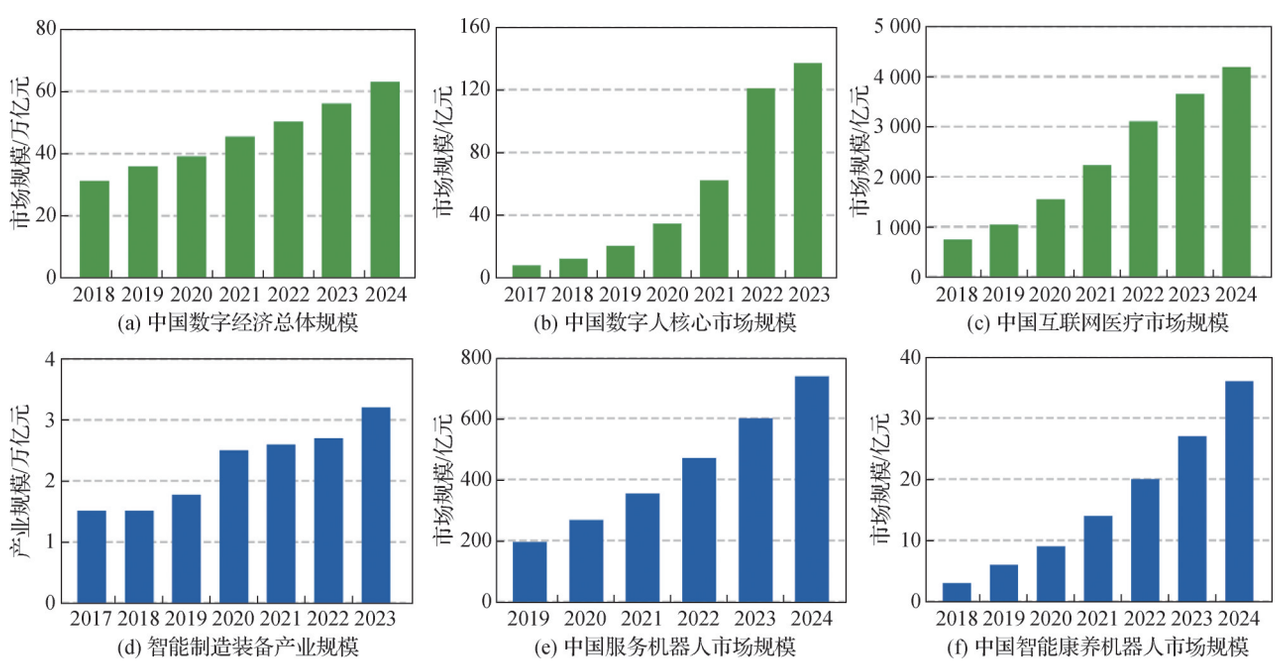



图4 情智兼备推进数字人与机器人的产业变革示意图

Fig. 4 Schematic diagram of industry transformation driven by emotionally and intellectually integrated digital humans and robotics ((a) the overall scale of the digital economy; (b) the market size of digital humans; (c) the market size of internet healthcare; (d) the overall scale of intelligent manufacturing equipment; (e) the market size of the robotics service industry; (f) the market size of intelligent elderly care robots)

后，优必选的Walker、特斯拉的Optimus、小米的CyberOne等人形机器人相继推出，人形机器人又一次迎来发展热潮。机器人行业呈现加速发展态势，情智兼备机器人的应用场景也大范围增加，除了传统的工业制造和文化娱乐等领域外，还在医疗、金融、教育和养老等领域开展了积极探索，应用前景广阔。

### 6.3 情智兼备机器人应用需求

1）接待服务。接待服务在所有服务行业中都具有普遍需求，包括医院的导诊分诊、政务大厅的办事咨询、银行的金融服务、商场的导购咨询以及企业的售后服务等。接待服务不仅要求机器人能够熟悉业务知识，通常还需要应对用户的焦虑、生气等负面情绪，以提供具有良好情感化体验的人机交互服务。目前，机器人在上述大部分行业的应用处在探索阶段，我们经常能在医院、商场等公共场所看到机器人的身影，但由于智能系统在不同场景表现不稳定，主动感知、自主决策和灵活运动等能力不足，服务体验与人类服务相比还有较大差距，通常提供位置导引、业务咨询等相对固化的交互服务，非标准应用场景适应能力还有待提高。

2）养老护理。随着人们生活节奏加快和社会老龄化到来，老年人的康养护理对机器人的需求越来越迫切，空巢老人的情感陪伴、失能老人的生活护理等具体需求不仅需要机器人能够完成助餐助浴等具体任务，还要在照顾老人的同时体现出共情和安慰等情感特质。当前，养老陪护机器人主要为老年人提供身心健康监测、意外跌倒辅助以及文化娱乐等基本服务，护理或家政类的复杂任务还在研究阶段，主要原因是护理家政类任务本身复杂多样，个性化需求层出不穷，机器人的理解和运动能力不足。

3）成长陪伴。伴随着全球人工智能革命呼之欲出，儿童教育成长对机器人的需求也逐渐成熟。当前，平板、手表、音响以及点读笔等各式各样的智能硬件已经走入孩子们的生活，提供学习、娱乐和社交等多样化服务。机器人也以早教机、机器狗等形式走入家庭陪伴儿童成长。目前，成长陪伴机器人通常提供知识问答、绘本讲解和娱乐表演等服务，尚缺少情感认知和交互关系建模等能力，难以与儿童建立长期稳定的陪伴关系，离朋友伙伴的定位还有很大差距。

### 6.4 情智兼备机器人应用挑战

1）应用需求碎片化，模型泛化能力不足。情智兼备机器人的主要应用场景在于服务业，与工业机器人相比，应用场景更加开放，应用需求高度碎片化，对机器人智能模型的泛化能力提出更高的要求，跨域跨场景感迁移学习能力有待进一步提升。

2）情感机理不清晰，情感交互能力欠缺。情智兼备机器人的情感交互性能是提升服务体验的关键，但当前情感信息的数理模型、产生原因、交互逻辑、转换机制等相关基础机理研究还不足以支撑机器人产生真正的情感，情感化交互能力难以满足个性化需求。

3）硬件水平遭瓶颈，感知与运动难协同。当前人工智能研究主要是数据驱动的深度学习，通常采用离线数据，造成感知—决策—执行相对独立，基于机器人实体的具身智能研究尚处于起步阶段。如何将机器人运行的诸多环节协同起来形成真正的具身智能，成为情智兼备机器人未来发展的关键。

## 7 结 语

情智兼备数字人与机器人技术的发展，既为科学研究带来了新的前沿挑战，也在多个领域中开辟了广阔的应用前景。通过脑认知驱动的情感机理研究、多模态大模型的融合与解译、个性化情感表征的动态计算，以及可交互情绪化内容生成调控等方面的进展，情智兼备的数字人与机器人逐步展现出在情感理解、自然交互及社会适应能力上的优势。

展望未来，情智兼备的数字人与机器人有望在医疗陪护、智能教育和心理健康等领域发挥重要作用。通过更精准的情感建模和多模态数据融合技术，它们能够提供更加个性化和人性化的服务，提升用户体验。同时，可信计算技术和隐私保护措施将保障用户的数据安全，使情感交互更加可靠。随着跨学科合作的深入，情智兼备机器人将进一步缩短人与机器之间的情感距离，推动社会各领域的智能化进程，为人类生活带来更加温暖和智能的服务。

**致谢**：本文由中国图象图形学学会情感计算与理解专业委员会组织撰写，该专业委员会链接为https://www.csig.org.cn/16/202111/49339.html。

### 参考文献（References）

Abdel-Hamid O, Mohamed A R, Jiang H, Deng L, Penn G and Yu D.

2014. Convolutional neural networks for speech recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 22(10): 1533-1545 [DOI: 10.1109/TASLP.2014.2339736]

Abdullah S M S A, Ameen S Y A, Sadeeq M A M and Zeebaree S. 2021. Multimodal emotion recognition using deep learning. Journal of Applied Science and Technology Trends, 2(1): 73-79 [DOI: 10.38094/jastt20291]

Akyunus M, Gençöz T and Aka B T. 2021. Age and sex differences in basic personality traits and interpersonal problems across young adulthood. Current Psychology, 40(5): 2518-2527 [DOI: 10.1007/s12144-019-0165-z]

Anil R, Dai A M, Firat O, Johnson M, Lepikhin D, Passos A, Shakeri S, Taropa E, Bailey P, Chen Z F, Chu E, Clark J H, El Shafey L, Huang Y P, Meier-Hellstern K, Mishra G, Moreira E, Omernick M, Robinson K, Ruder S, Tay Y, Xiao K F, Xu Y Z, Zhang Y J, Abrego G H, Ahn J, Austin J, Barham P, Botha J, Bradbury J, Brahma S, Brooks K, Catasta M, Cheng Y, Cherry C, Choquette-Choo C A, Chowdhery A, Crepy C, Dave S, Dehghani M, Dev S, Devlin J, Díaz M, Du N, Dyer E, Feinberg V, Feng F X Y, Fienber V, Freitag M, Garcia X, Gehrmann S, Gonzalez L, Gur-Ari G, Hand S, Hashemi H, Hou L, Howland J, Hu A, Hui J, Hurwitz J, Isard M, Ittycheriah A, Jagielski M, Jia W H, Kenealy K, Krikun M, Kudugunta S, Lan C, Lee K, Lee B, Li E, Li M, Li W, Li Y G, Li J, Lim H, Lin H Z, Liu Z T, Liu F, Maggioni M, Mahendru A, Maynez J, Misra V, Moussalem M, Nado Z, Nham J, Ni E, Nystrom A, Parrish A, Pellat M, Polacek M, Polozov A, Pope R, Qiao S Y, Reif E, Richter B, Riley P, Ros A C, Roy A, Saeta B, Samuel R, Shelby R, Slone A, Smilkov D, So D R, Sohn D, Tokumine S, Valter D, Vasudevan V, Vodrahalli K, Wang X Z, Wang P D, Wang Z R, Wang T, Wieting J, Wu Y H, Xu K, Xu Y H, Xue L T, Yin P C, Yu J H, Zhang Q, Zheng S, Zheng C, Zhou W K, Zhou D, Petrov S and Wu Y H. 2023. Palm 2 technical report [EB/OL]. [2024-12-27]. https://arxiv.org/pdf/2305.10403.pdf

Badrinarayanan V, Kendall A and Cipolla R. 2017. SegNet: a deep convolutional encoder-decoder architecture for image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(12): 2481-2495 [DOI: 10.1109/TPAMI.2016.2644615]

Baltrušaitis T, Ahuja C and Morency L P. 2019. Multimodal machine learning: a survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(2): 423-443 [DOI: 10.1109/TPAMI.2018.2798607]

Barnes J, Klinger R and Walde S S I. 2018. Projecting embeddings for domain adaption: joint modeling of sentiment analysis in diverse domains//Proceedings of the 27th International Conference on Computational Linguistics. Santa Fe, USA: Association for Computational Linguistics: 818-830

Barrett L F and Satpute A B. 2013. Large-scale brain networks in affective and social neuroscience: towards an integrative functional architecture of the brain. Current Opinion in Neurobiology, 23(3): 361-372 [DOI: 10.1016/j.conb.2012.12.012]

Berluti K, Ploe M L and Marsh A A. 2023. Emotion processing in youths with conduct problems: an fMRI meta-analysis. Translational Psychiatry, 13(1): #105 [DOI: 10.1038/s41398-023-02363-z]

Berridge C, Zhou Y J, Robillard J M and Kaye J. 2023. Companion robots to mitigate loneliness among older adults: perceptions of benefit and possible deception. Frontiers in Psychology, 14: #1106633 [DOI: 10.3389/fpsyg.2023.1106633]

Betzel R F. 2022. Network neuroscience and the connectomics revolution//Horn A, ed. Connectomic Deep Brain Stimulation. [s.l.]: Academic Press: 25-58 [DOI: 10.1016/B978-0-12-821861-7.00002-6]

Blanz V and Vetter T. 2023. A morphable model for the synthesis of 3D faces. Seminal Graphics Papers: Pushing the Boundaries, 2: 157-164 [DOI: 10.1145/3596711.3596730]

Brown T B, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal S, Herbert-Voss A, Krueger G, Henighan T, Child R, Ramesh A, Ziegler D M, Wu J, Winter C, Hesse C, Chen M, Sigler E, Litwin M, Gray S, Chess B, Clark J, Berner C, McCandlish S, Radford A, Sutskever I and Amodei D. 2020. Language models are few-shot learners [EB/OL]. [2024-12-27]. https://arxiv.org/pdf/2005.14165.pdf

Cambria E, Li Y, Xing F Z, Poria S and Kwok K. 2020. SenticNet 6: ensemble application of symbolic and subsymbolic AI for sentiment analysis//Proceedings of the 29th ACM International Conference on Information and Knowledge Management. Virtual Event, Ireland: ACM: 105-114 [DOI: 10.1145/3340531.3412003]

Celiktutan O and Gunes H. 2015. Computational analysis of human-robot interactions through first-person vision: Personality and interaction experience//The 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN). Kobe, Japan: IEEE: 815-820 [DOI: 10.1109/ROMAN.2015.7333602]

Chen L F, Li M, Wu M, Pedrycz W and Hirota H. 2024a. Coupled multimodal emotional feature analysis based on broad-deep fusion networks in human-robot interaction. IEEE Transactions on Neural Networks and Learning Systems, 35(7): 9663-9673 [DOI: 10.1109/TNNLS.2023.3236320]

Chen Y F, Wang L Z, Li Q J, Xiao H J, Zhang S P, Yao H X and Liu Y B. 2024b. MonoGaussianAvatar: monocular Gaussian point-based head avatar//Proceedings of 2024 ACM SIGGRAPH Conference Papers. Denver, USA: ACM: #58 [DOI: 10.1145/3641519.3657499]

Cuperman R and Ickes W. 2009. Big five predictors of behavior and perceptions in initial dyadic interactions: personality similarity helps extraverts and introverts, but hurts "disagreeables". Journal of Personality and Social Psychology, 97(4): 667-684 [DOI: DOI: 10.1037/a0015741]

Dalgleish T. 2004. The emotional brain. Nature Reviews Neuroscience, 5(7): 583-589 [DOI: 10.1038/nrn1432]

Devlin J, Chang M W, Lee K and Toutanova K. 2019a. Bert: pre-training of deep bidirectional Transformers for language understanding [EB/OL]. [2024-12-27]. https://arxiv.org/pdf/1810.04805.pdf

D'Mello S K and Kory J. 2015. A review and meta-analysis of multimodal affect detection systems. ACM Computing Surveys, 47(3): #43 [DOI: 10.1145/2682899]

Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X H, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J and Houlsby N. 2021. An image is worth 16 × 16 words: Transformers for image recognition at scale [EB/OL]. [2024-12-27]. https://arxiv.org/pdf/2010.11929.pdf

Driess D, Xia F, Sajjadi M S M, Lynch C, Chowdhery A, Ichter B, Wahid A, Tompson J, Vuong Q, Yu T H, Huang W L, Chebotar Y, Sermanet P, Duckworth D, Levine S, Vanhoucke V, Hausman K, Toussaint M, Greff K, Zeng A, Mordatch I and Florence P. 2023. PaLM-E: an embodied multimodal language model//Proceedings of the 40th International Conference on Machine Learning. Honolulu, USA: JMLR.org: 8469-8488

Ekman P. 1992. An argument for basic emotions. Cognition and Emotion, 6(3/4): 169-200 [DOI: 10.1080/02699939208411068]

Elfenbein H A and Ambady N. 2002. On the universality and cultural specificity of emotion recognition: a meta-analysis. Psychological Bulletin, 128(2): 203-235 [DOI: 10.1037/0033-2909.128.2.203]

Ficocelli M, Terao J and Nejat G. 2016. Promoting interactions between humans and robots using robotic emotional behavior. IEEE Transactions on Cybernetics, 46(12): 2911-2923 [DOI: 10.1109/TCYB.2015.2492999]

Friston K J. 2011. Functional and effective connectivity: a review. Brain Connectivity, 1(1): 13-36 [DOI: 10.1089/brain.2011.0008]

Ghandeharioun A, McDuff D, Czerwinski M and Rowan K. 2019. EMMA: an emotion-aware wellbeing chatbot//Proceedings of the 8th International Conference on Affective Computing and Intelligent Interaction. Cambridge, UK: IEEE: 1-7 [DOI: 10.1109/ACII.2019.8925455]

Giebenhain S, Kirschstein T, Rünz M, Agapito L and Nießner M. 2024. NPGA: neural parametric Gaussian avatars//Proceedings of 2024 ACM SIGGRAPH Asia Conference Papers. Tokyo, Japan: ACM: 127 [DOI: 10.1145/3680528.3687689]

Han X M, Xue R D, Du S Y and Gao Y. 2024. Inter-intra high-order brain network for ASD diagnosis via functional MRIs//Proceedings of the 27th International Conference on Medical Image Computing and Computer-Assisted Intervention. Marrakesh, Morocco: Springer: 216-226 [DOI: 10.1007/978-3-031-72069-7_21]

Hang Y Z, Soto C, Speyer L G, Haring L, Lee B, Ostendorf F and Mõttus R. 2021. Age differences in the personality hierarchy: a multi-sample replication study across the life span. Journal of Research in Personality, 93: #104121 [DOI: 10.1016/j.jrp.2021.104121]

Henrich J. 2015. Culture and social behavior. Current Opinion in Behavioral Sciences, 3: 84-89 [DOI: 10.1016/j.cobeha.2015.02.001]

Hong A, Lunscher N, Hu T H, Tsuboi Y, Zhang X Y, Franco Dos reis alves S, Nejat G and Benhabib B. 2021. A multimodal emotional human-robot interaction architecture for social robots engaged in bidirectional communication. IEEE Transactions on Cybernetics, 51(12): 5954-5968 [DOI: 10.1109/TCYB.2020.3024567]

Huang K L, Shi B T, Li X, Huang S Y and Li Y K. 2024. Multi-modal sensor fusion for auto driving perception: a survey [EB/OL]. [2024-12-27]. https://arxiv.org/pdf/2202.02703.pdf

Ivaldi S, Lefort S, Peters J, Chetouani M, Provasi J and Zibetti E. 2017. Towards engagement models that consider individual factors in HRI: on the relation of extroversion and negative attitude towards robots to gaze and speech during a human-robot assembly task: experiments with the iCub humanoid. International Journal of Social Robotics, 9(1): 63-86 [DOI: 10.1007/s12369-016-0357-8]

Jafari F, Berretti S and Basu A. 2024. JambaTalk: speech-driven 3D talking head generation based on hybrid Transformer-Mamba model [EB/OL]. [2024-12-27]. https://arxiv.org/pdf/2408.01627.pdf

Jain V, Crowley J L, Dey A K and Lux A. 2014. Depression estimation using audiovisual features and fisher vector encoding//Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge. Orlando, USA: ACM: 87-91 [DOI: 10.1145/2661806.2661817]

Ji X Y, Zhou H, Wang K S Y, Wu W, Loy C C, Cao X and Xu F. 2021. Audio-driven emotional video portraits//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 14075-14084 [DOI: 10.1109/CVPR46437.2021.01386]

Jin Y, Choi M, Verma G, Wang J, Kumar S. 2024. MM-SOC: Benchmarking multimodal large language models in social media platforms//Findings of the Association for Computational Linguistics: ACL: 6192-6210.

Kan X, Cui H J, Lukemire J, Guo Y and Yang C. 2022. FBNETGEN: task-aware gnn-based fmri analysis via functional brain network generation [EB/OL]. [2024-12-27]. https://arxiv.org/pdf/2205.12465.pdf

Kerbl B, Kopanas G, Leimkühler T and Drettakis G. 2023. 3D Gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics, 42(4): #139 [DOI: 10.1145/3592433]

Kim J, André E and Vogt T. 2009. Towards user-independent classification of multimodal emotional signals//Proceedings of the 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops. Amsterdam, the Netherlands: IEEE: 1-7 [DOI: 10.1109/ACII.2009.5349495]

Kramer A D I, Guillory J E and Hancock J T. 2014. Experimental evidence of massive-scale emotional contagion through social net-

works. Proceedings of the National Academy of Sciences of the United States of America, 111(24): 8788-8790 [DOI: 10.1073/pnas.132004011]

LeDoux J. 2012. Rethinking the emotional brain. Neuron, 73(4): 653-676 [DOI: 10.1016/j.neuron.2012.02.004]

LeDoux J E. 2000. Emotion circuits in the brain. Annual Review of Neuroscience, 23: 155-184 [DOI: 10.1146/annurev.neuro.23.1.155]

Li J N, Li D X, Xiong C M and Hoi S. 2022. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation//Proceedings of the 39th International Conference on Machine Learning. Seattle, USA: PMLR: 12888-12900

Li M Y, Yang J, Jiao N D, Wang Y C and Liu L Q. 2022. Review of the latest research progress of micro-nano robots. Robot, 44(6): 732-749 (李梦月, 杨佳, 焦念东, 王越超, 刘连庆. 2022. 微纳米机器人的最新研究进展综述. 机器人, 44(6): 732-749) [DOI: 10.13973/j.cnki.robot.210377]

Li Z, Zheng Z R, Wang L Z and Liu Y B. 2024b. Animatable Gaussians: learning pose-dependent Gaussian maps for high-fidelity human avatar modeling//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 19711-19722 [DOI: 10.1109/CVPR52733.2024.01864]

Liang P P, Zadeh A, Morency L P. 2023. Foundations and trends in multimodal machine learning: principles, challenges, and open questions [EB/OL]. [2024-12-27]. https://arxiv.org/pdf/2209.03430.pdf

Lin H, Tov W and Qiu L. 2014. Emotional disclosure on social networking sites: the role of network structure and psychological needs. Computers in Human Behavior, 41: 342-350 [DOI: 10.1016/j.chb.2014.09.045]

Lin S Y, Li Z, Su Z Q, Zheng Z R, Zhang H W and Liu Y B. 2024. LayGA: layered Gaussian avatars for animatable clothing transfer//Proceedings of 2024 ACM SIGGRAPH Conference Papers. Denver, USA: ACM: #37 [DOI: 10.1145/3641519.3657501]

Lindquist K A, Wager T D, Kober H, Bliss-Moreau E and Barrett L F. 2012. The brain basis of emotion: a meta-analytic review. Behavioral and Brain Sciences, 35(3): 121-143 [DOI: 10.1017/S0140525X11000446]

Liu Z, Lin Y T, Cao Y, Hu H, Wei Y X, Zhang Z, Lin S and Guo B N. 2021. Swin transformer: hierarchical vision transformer using shifted windows//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 9992-10002 [DOI: 10.1109/ICCV48922.2021.00986]

Loveys K, Sagar M, Zhang X Y, Fricchione G and Broadbent E. 2021. Effects of emotional expressiveness of a female digital human on loneliness, stress, perceived support, and closeness across genders: randomized controlled trial. Journal of Medical Internet Research, 23(11): #30624 [DOI: 10.2196/30624]

Lubis N, Sakti S, Yoshino K and Nakamura S. 2018. Eliciting positive emotion through affect-sensitive dialogue response generation: a neural network approach//Proceedings of the 32nd AAAI Conference on Artificial Intelligence. New Orleans, USA: AAAI: 5293-5300 [DOI: 10.1609/aaai.v32i1.11955]

Luo Y, Zhu L Z, Wan Z Y and Lu B L. 2020. Data augmentation for enhancing EEG-based emotion recognition with deep generative models. Journal of Neural Engineering, 17(5): #056021 [DOI: 10.1088/1741-2552/abb580]

Mehrabian A. 1996. Pleasure-arousal-dominance: a general framework for describing and measuring individual differences in temperament. Current Psychology, 14(4): 261-292 [DOI: 10.1007/BF02686918]

Mildenhall B, Srinivasan P P, Tancik M, Barron J T, Ramamoorthi R and Ng R. 2020. NeRF: representing scenes as neural radiance fields for view synthesis//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 405-421 [DOI: 10.1007/978-3-030-58452-8_24]

Mohammad S and Bravo-Marquez F. 2017. WASSA-2017 shared task on emotion intensity//Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis. Copenhagen, Denmark: Association for Computational Linguistics: 39-49 [DOI: 10.18653/v1/W17-5205]

Mollahosseini A, Hasani B and Mahoor M H. 2019. AffectNet: a database for facial expression, valence, and arousal computing in the wild. IEEE Transactions on Affective Computing, 10(1): 18-31 [DOI: 10.1109/TAFFC.2017.2740923]

Momennejad I. 2022. Collective minds: social network topology shapes collective cognition. Philosophical Transactions of the Royal Society B: Biological Sciences, 377(1843): #20200315 [DOI: 10.1098/rstb.2020.0315]

Müller M. 2007. Dynamic time warping//Information Retrieval for Music and Motion. Berlin, Heidelberg: Springer: 69-84 [DOI: 10.1007/978-3-540-74048-3_4]

Nagrani A, Yang S, Arnab A, Jansen A, Schmid C and Sun C. 2021. Attention bottlenecks for multimodal fusion//Proceedings of the 35th International Conference on Neural Information Processing Systems. Virtual Event: Curran Associates Inc.: 14200-14213

Ng E, Romero J, Bagautdinov T, Bai S J, Darrell T, Kanazawa A and Richard A. 2024. From audio to photoreal embodiment: synthesizing humans in conversations//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 1001-1010 [DOI: 10.1109/CVPR.2024.1001]

OpenAI, Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, Aleman F L, Almeida D, Altenschmidt J, Altman S, Anadkat S, Avila R, Babuschkin I, Balaji S, Balcom V, Baltescu P, Bao H, Bavarian M, Belgum J, Bello I, Berdine J, Bernadett-Shapiro G, Berner C, Bogdonoff L, Boiko O, Boyd M, Brakman A, Brockman G, Brooks T, Brundage M, Button K, Cai T, Campbell R, Cann A, Carey B, Carlson C, Carmichael R, Chan B, Chang C, Chantzis F, Chen D, Chen S, Chen R, Chen J, Chen M, Chess B, Cho C, Chu C, Chung H W, Cummings D, Currier J, Dai Y, Decareaux

C, Degry T, Deutsch N, Deville D, Dhar A, Dohan D, Dowling S, Dunning S, Ecoffet A, Eleti A, Eloundou T, Farhi D, Fedus L, Felix N, Fishman SP, Forte J, Fulford I, Gao L, Georges E, Gibson C, Goel V, Gogineni T, Goh G, Gontijo-Lopes R, Gordon J, Grafstein M, Gray S, Greene R, Gross J, Gu S S, Guo Y F, Hallacy C, Han J, Harris J, He Y C, Heaton M, Heidecke J, Hesse C, Hickey A, Hickey W, Hoeschele P, Houghton B, Hsu K, Hu S L, Hu X, Huizinga J, Jain S, Jain S, Jang J, Jiang A, Jiang R, Jin H Z, Jin D, Jomoto S, Jonn B, Jun H, Kaftan T, Kaiser Ł, Kamali A, Kanitscheider I, Keskar N S, Khan T, Kilpatrick L, Kim J W, Kim C, Kim Y, Kirchner J H, Kiros J, Knight M, Kokotajlo D, Kondraciuk Ł, Kondrich A, Konstantinidis A, Kosic K, Krueger G, Kuo V, Lampe M, Lan I, Lee T, Leike J, Leung J, Levy D, Li C M, Lim R, Lin M, Lin S, Litwin M, Lopez T, Lowe R, Lue P, Makanju A, Malfacini K, Manning S, Markov T, Markovski Y, Martin B, Mayer K, Mayne A, McGrew B, McKinney S M, McLeavey C, McMillan P, McNeil J, Medina D, Mehta A, Menick J, Metz L, Mishchenko A, Mishkin P, Monaco V, Morikawa E, Mossing D, Mu T, Murati M, Murk O, Mély D, Nair A, Nakano R, Nayak R, Neelakantan A, Ngo R, Noh H, Ouyang L, O'Keefe C, Pachocki J, Paino A, Palermo J, Pantuliano A, Parascandolo G, Parish J, Parparita E, Passos A, Pavlov M, Peng A, Perelman A, Peres F, Petrov M, Pinto H, (Rai) Pokorny M, Pokrass M, Pong V H, Powell T, Power A, Power B, Proehl E, Puri R, Radford A, Rae J, Ramesh A, Raymond C, Real F, Rimbach K, Ross C, Rotsted B, Roussez H, Ryder N, Saltarelli M, Sanders T, Santurkar S, Sastry G, Schmidt H, Schnurr D, Schulman J, Selsam D, Sheppard K, Sherbakov T, Shieh J, Shoker S, Shyam P, Sidor S, Sigler E, Simens M, Sitkin J, Slama K, Sohl I, Sokolowsky B, Song Y, Staudacher N, Such F P, Summers N, Sutskever I, Tang J, Tezak N, Thompson M B, Tillet P, Tootoonchian A, Tseng E, Tuggle P, Turley N, Tworek J, Uribe J, Vallone A, Vijayvergiya A, Voss C, Wainwright C, Wang J J, Wang A, Wang B, Ward J, Wei J, Weinmann C, Welihinda A, Welinder P, Weng J, Weng L, Wiethoff M, Willner D, Winter C, Wolrich S, Wong H, Workman L, Wu S, Wu J, Wu M, Xiao K, Xu T, Yoo S, Yu K, Yuan Q M, Zaremba W, Zellers R, Zhang C, Zhang M, Zhao S J, Zheng T H, Zhuang J T, Zhuk W and Zoph B. 2024. GPT-4 technical report [EB/OL]. [2024-12-27]. https://arxiv.org/pdf/2303.08774.pdf

Park S and Whang M. 2022. Empathy in human-robot interaction: designing for social robots. International Journal of Environmental Research and Public Health, 19(3): #1889 [DOI: 10.3390/ijerph19031889]

Phelps E A. 2004. Human emotion and memory: interactions of the amygdala and hippocampal complex. Current Opinion in Neurobiology, 14(2): 198-202 [DOI: 10.1016/j.conb.2004.03.015]

Poria S, Majumder N, Mihalcea R and Hovy E. 2019. Emotion recognition in conversation: research challenges, datasets, and recent advances. IEEE Access, 7: 100943-100953 [DOI: 10.1109/ACCESS.2019.2929050]

Qian S H, Kirschstein T, Schoneveld L, Davoli D, Giebenhain S and Nießner M. 2024. GaussianAvatars: photorealistic head avatars with rigged 3D Gaussians//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 20299-20309 [DOI: 10.1109/CVPR52733.2024.01919]

Ren S Q, He K M, Girshick R and Sun J. 2017. Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(6): 1137-1149 [DOI: 10.1109/TPAMI.2016.2577031]

Rui T, Cui P and Zhu W W. 2017. Joint user-interest and social-influence emotion prediction for individuals. Neurocomputing, 230: 66-76 [DOI: 10.1016/j.neucom.2016.11.054]

Shen J Q, Brdiczka O and Liu J. 2013. Understanding email writers: personality prediction from email messages//Proceedings of the 21st International Conference on User Modeling, Adaptation, and Personalization. Rome, Italy: Springer: 318-330 [DOI: 10.1007/978-3-642-38844-6_29]

Sidner C L, Kidd C D, Lee C and Lesh N. 2004. Where to look: a study of human-robot engagement//Proceedings of the 9th International Conference on Intelligent User Interfaces. Funchal, Portugal: ACM, 78-84 [DOI: 10.1145/964442.964458]

Tang J, Zhang Y, Sun J M, Rao J H, Yu W J, Chen Y R and Fong A C M. 2012. Quantitative study of individual emotional states in social networks. IEEE Transactions on Affective Computing, 3(2): 132-144 [DOI: 10.1109/T-AFFC.2011.23]

Tao J H, Fan C H, Lian Z, Lyu Z, Shen Y and Liang S. 2024. Development of multimodal sentiment recognition and understanding. Journal of Image and Graphics, 29(6): 1607-1627 (陶建华, 范存航, 连政, 吕钊, 沈莹, 梁山. 2024. 多模态情感识别与理解发展现状及趋势. 中国图象图形学报, 29(6): 1607-1627) [DOI: 10.11834/jig.240017]

Van Haeringen E S, Gerritsen C and Hindriks K V. 2023. Emotion contagion in agent-based simulations of crowds: a systematic review. Autonomous Agents and Multi-Agent Systems, 37(1): #6 [DOI: 10.1007/s10458-022-09589-z]

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser Ł and Polosukhin I. 2017. Attention is all you need//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc.: 6000-6010

Vogt T and André E. 2006. Improving automatic emotion recognition from speech via gender differentiation//Proceedings of the 5th International Conference on Language Resources and Evaluation. Genoa, Italy: ICLRE: 1123-1126

Wang G Q, Pei Y Q, Yang Y, Xu X, Wang Z and Shen H T. 2024. Multimodal trustworthy interaction: from multimodal information fusion to a trinitarian human-robot-digital human interaction model. Scien-

tia Sinica Informationis，54(4)：872-892（王国庆，裴云强，杨阳，徐行，汪政，申恒涛. 2024. 多模可信交互：从多模态信息融合到人—机器人—数字人三位一体式交互模型. 中国科学：信息科学)，54(4)：872-892［DOI：10.1360/SSI-2023-0133］

Wang T M，Tao Y and Chen Y. 2012. Research status and development trends of the service robotic technology. Scientia Sinica Informationis，42(9)：1049-1066（王田苗，陶永，陈阳. 2012. 服务机器人技术研究现状与发展趋势. 中国科学：信息科学)，42(9)：1049-1066［DOI：10.1360/112012-402］

Wang Y S，Shen Y，Liu Z，Liang P P，Zadeh A and Morency L P. 2019. Words can shift：dynamically adjusting word representations using nonverbal behaviors//Proceedings of the 33rd AAAI Conference on Artificial Intelligence. Honolulu，USA：AAAI Press：7216-7223［DOI：10.1609/aaai.v33i01.33017216］

Wang Z，Feng X H，Li Y M and Zhuang J X. 2018. The current situation and future of the intelligent robot industry. AI-View，(3)：12-27（王哲，冯晓辉，李艺铭，庄金鑫. 2018. 智能机器人产业的现状与未来. 人工智能，(3)：12-27)［DOI：10.16453/j.cnki.issn2096-5036.2018.03.002］

Wang Z，Wang Y X，Zhang J P，Hu C F，Yin Z and Song Y. 2022. Spatial-temporal feature fusion neural network for EEG-based emotion recognition. IEEE Transactions on Instrumentation and Measurement，71：#2507212［DOI：10.1109/TIM.2022.3165280］

Wei T，Pang S，Guo Q，Ma Y，Ma Y and Guo Q. 2024. EmoAttack：emotion-to-image diffusion models for emotional backdoor generation［EB/OL］.［2024-12-27］. https://arxiv.org/pdf/2406.15863.pdf

Weisberg Y J，DeYoung C G and Hirsh J B. 2011. Gender differences in personality across the ten aspects of the big five. Frontiers in Psychology，2：#11757［DOI：10.3389/fpsyg.2011.00178］

Wen C S，Jia G L and Yang J F. 2023. DIP：dual incongruity perceiving network for sarcasm detection//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver，Canada：IEEE：2540-2550［DOI：10.1109/CVPR52729.2023.00250］

Xu L，Lv X T，Yan Y C，Jin X，Wu S W，Xu C S，Liu Y F，Zhou Y Z，Rao F Y，Sheng X D，Liu Y H，Zeng W J and Yang X K. 2024. Inter-X：towards versatile human-human interaction analysis//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle，USA：IEEE：22260-22271［DOI：10.1109/CVPR52733.2024.02101］

Yadollahi A，Shahraki A G and Zaiane O R. 2017. Current state of text sentiment analysis from opinion to emotion mining. ACM Computing Surveys，50(2)：#25［DOI：10.1145/3057270］

Yang G Z，Bellingham J，Dupont P E，Fischer P，Floridi L，Full R，Jacobstein N，Kumar V，McNutt M，Merrifield R，Nelson B J，Scassellati B，Taddeo M，Taylor R，Veloso M，Wang Z L and Wood R. 2018a. The grand challenges of science robotics. Science Robotics，3(14)：#7650［DOI：10.1126/scirobotics.aar7650］

Yang J F，She D Y，Sun M，Cheng M M，Rosin P L and Wang L. 2018b. Visual sentiment prediction based on automatic discovery of affective regions. IEEE Transactions on Multimedia，20(9)：2513-2525［DOI：10.1109/TMM.2018.2816938］

Yang J，Feng J and Huang H. 2024. EmoGen：emotional image content generation with text-to-image diffusion models//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle WA，USA：IEEE：6358-6368

Yi H W，Thies J，Black M J，Peng X B and Rempe D. 2025. Generating human interaction motions in scenes with text control//Proceedings of the 18th European Conference on Computer Vision. Milan，Italy：Springer：246-263［DOI：10.1007/978-3-031-73235-5_14］

Yin Y，Ghasedi K，Wu H T，Yang J L，Tong X and Fu Y. 2023. Nerfinvertor：high fidelity nerf-gan inversion for single-shot real image animation//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023：8539-8548

Yun J and Park J. 2022. The effects of chatbot service recovery with emotion words on customer satisfaction，repurchase intention，and positive word-of-mouth. Frontiers in Psychology，13：#922503［DOI：10.3389/fpsyg.2022.922503］

Zadeh A，Liang P P，Poria S，Vij P，Cambria E and Morency L P. 2018. Multi-attention recurrent network for human communication comprehension//Proceedings of the 32nd AAAI Conference on Artificial Intelligence. New Orleans，USA：AAAI：5642-5649［DOI：10.1609/aaai.v32i1.12024］

Zhang H，Song R，Wang L P，Zhang L，Wang D W，Wang C and Zhang W. 2023. Classification of brain disorders in rs-fMRI via local-to-global graph neural networks. IEEE Transactions on Medical Imaging，42(2)：444-455［DOI：10.1109/TMI.2022.3219260］

Zhang J H，Yin Z，Chen P and Nichele S. 2020. Emotion recognition using multi-modal data and machine learning techniques：a tutorial and review. Information Fusion，59：103-126［DOI：10.1016/j.inffus.2020.01.011］

Zhang Z Y，Han X，Liu Z Y，Jiang X，Sun M S and Liu Q. 2019. ERNIE：enhanced language representation with informative entities［EB/OL］.［2024-12-27］. https://arxiv.org/pdf/1905.07129.pdf

Zhao L L，Li X A，Zhao J D and Liu H. 2024. Development strategy of space robots for autonomous repair and maintenance of spacecraft. Strategic Study of CAE，26(1)：149-159（赵亮亮，李雪皑，赵京东，刘宏. 2024. 面向航天器自主维护的空间机器人发展战略研究. 中国工程科学，26(1)：149-159)［DOI：10.15302/J-SSCAE-2024.01.014］

Zhao S C，Jia G L，Yang J F，Ding G G and Keutzer K. 2021. Emotion recognition from multiple modalities：fundamentals and methodologies. IEEE Signal Processing Magazine，38(6)：59-73［DOI：10.1109/MSP.2021.3106895］

Zhao S C，Jia Z Z，Chen H，Li L D，Ding G G and Keutzer K. 2019. PDANet：polarity-consistent deep attention network for fine-grained visual emotion regression//Proceedings of the 27th ACM International Conference on Multimedia. Nice，France：ACM：192-

201［DOI：10.1145/3343031.3351062］

Zhao S C, Hong X P, Yang J F, Zhao Y Y and Ding G G. 2023. Toward label-efficient emotion and sentiment analysis. Proceedings of 2023 IEEE, 111(10)：1159-1197［DOI：10.1109/JPROC.2023.3309299］

Zhao S C, Yao H X, Gao Y, Ji R R, Xie W L, Jiang X L and Chua T S. 2016. Predicting personalized emotion perceptions of social images// Proceedings of the 24th ACM International Conference on Multimedia. Amsterdam, the Netherlands：ACM：1385-1394［DOI：10.1145/2964284.2964289］

Zhao Z P, Zheng Y, Zhang Z X, Wang H S, Zhao Y Q and Li C. 2018. Exploring spatio-temporal representations by integrating attention-based bidirectional-LSTM-RNNs and FCNs for speech emotion recognition//Proceedings of Interspeech 2018. Hyderabad, India：International Speech Communication Association（ISCA）：272-276［DOI：10.21437/Interspeech.2018-1477］

Zhou H, Huang M L, Zhang T Y, Zhu X Y and Liu B. 2018. Emotional chatting machine：emotional conversation generation with internal and external memory//Proceedings of the 32nd AAAI Conference on Artificial Intelligence. New Orleans, USA：730-738［DOI：10.1609/aaai.v32i1.11325］

Zhou X Y, Chen D L and Chen Y D.2023.Towards joint modeling of dialogue response and speech synthesis based on large language model［EB/OL］.［2024-12-27］. https://arxiv.org/pdf/2309.11000.pdf

Zhu J Y, Park T, Isola P and Efros A A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy：IEEE：2242-2251［DOI：10.1109/ICCV.2017.244］

Zhu T, Li L D, Yang J F, Zhao S C, Liu H T and Qian J S. 2023a. Multimodal sentiment analysis with image-text interaction network. IEEE Transactions on Multimedia, 25：3375-3385［DOI：10.1109/TMM.2022.3160060］

Zhu T, Li L D, Yang J F, Zhao S C and Xiao X. 2023b. Multimodal emotion classification with multi-level semantic reasoning network. IEEE Transactions on Multimedia, 25：6868-6880［DOI：10.1109/TMM.2022.3214989］

Zu C, Gao Y, Munsell B, Kim M, Peng Z W, Zhu Y Y, Gao W, Zhang D Q, Shen D G and Wu G R. 2016. Identifying high order brain connectome biomarkers via learning on hypergraph//Proceedings of the 7th International Workshop on Machine Learning in Medical Imaging. Athens, Greece：Springer：1-9［DOI：10.1007/978-3-319-47157-0_1］

## 作者简介

赵思成，男，副研究员，主要研究方向为情感计算与多媒体分析。E-mail：schzhao@tsinghua.edu.cn

高跃，通信作者，男，副教授，主要研究方向为计算机视觉。E-mail：gaoyue@tsinghua.edu.cn

丰一帆，男，博士研究生，主要研究方向为超图计算。E-mail：evanfeng97@gmail.com

张知诚，男，博士研究生，主要研究方向为视频理解与生成。E-mail：gloryzzc6@sina.com

孙斌，男，副教授，主要研究方向为多模态语义理解与交互。E-mail：sunbin611@hnu.edu.cn

张盛平，男，教授，主要研究方向为计算机视觉。E-mail：s.zhang@hit.edu.cn

杨巨峰，男，教授，主要研究方向为情感计算。E-mail：yangjufeng@nankai.edu.cn

刘敏，男，教授，主要研究方向为计算机视觉。E-mail：liu_min@hnu.edu.cn

姚鸿勋，女，教授，主要研究方向为情感计算与计算机视觉。E-mail：h.yao@hit.edu.cn

王耀南，男，教授，主要研究方向为机器人。E-mail：yaonan@hnu.edu.cn