人机价值观驱动的对话情绪生成模型

马志强[1,2,*], 叶浩然[1], 刘佳[1], 吕凯[1]

[1]内蒙古工业大学智能科学与技术学院，呼和浩特市，内蒙古，010080

[2]内蒙古自治区北疆网络空间安全重点实验室，呼和浩特市，内蒙古，010080

{mzq_bim,20231800801,20221800729,20221800724}@imut.edu.cn

摘要

对话系统情绪生成任务旨在生成待回复话语的情绪类别。针对现有情绪生成模型忽视了用户与模型价值观一致性对情绪生成的调节与引导作用，导致对话系统生成情绪与用户期望情绪之间存在偏差，降低了对话系统与用户之间的情绪共鸣。本文提出一种人机价值观驱动的对话情绪生成模型-HVDEGM，通过多阶段的门控机制动态引入用户价值观特征来引导情绪生成。该模型基于价值观一致性原理，设计了三个单元。首先情境修正注意力单元通过两次注意力机制增强了情绪与语义特征信息，其次价值观融合单元通过多阶段融合门控动态平衡了用户价值观特征与对话系统历史价值观特征的权重，最后反应调节单元通过双向注意力与交叉注意力机制，强化了情绪、语义、价值观特征之间的互补关联。模型在新构建的价值观对话数据集ValueCon上进行实验，实验结果表明，HVDEGM相比DialogueRNN、DialogueGCN等基线模型在Precision、Recall、F1及情绪共鸣度等指标分别提升了2.9%、2.5%、0.9%和4.1%，证明了所提出方法的有效性。

关键词： 情绪生成；价值观一致性；注意力机制；情绪共鸣

Human-Machine Values-Driven Dialogue Emotion Generation Model

Zhiqiang Ma[1,2*] Haoran Ye[1] Jia Liu[1] Kai Lü[1]

[1]College Of Intelligent Science And Technoligy, Inner Mongolia University of Technology, Hohhot,Inner Mongolia,010080

[2]Key Laboratory of Beijiang Cyberspace Security of Inner Mongolia Autonomous Region, Hohhot,Inner Mongolia,010080

{mzq_bim,20231800801,20221800729,20221800724}@imut.edu.cn

Abstract

The task of dialogue-system emotion generation aims to produce an appropriate emotional category for a given reply. Existing methods largely ignore the role of value-alignment between user and system in modulating and guiding emotion generation, resulting in outputs that deviate from user expectations and diminish emotional resonance. To address this, we propose a Human–Machine Values-Driven Dialogue Emotion Generation Model (HVDEGM) that dynamically incorporates user value signals through a multi-stage gating mechanism. Guided by the principle of value consistency, HVDEGM comprises three modules: first, a Contextualized Modified Attention Unit

*代表通讯作者

基金项目：国家自然科学基金（No.62166029）;内蒙古自治区科研基础条件及平台（No.2025KYPT0014）;内蒙古自治区高等学校创新团队发展计划（No.NMGIRT2506）;内蒙古自治区自然科学基金（No.2023LHMS06007）;内蒙古自治区高等学校碳达峰碳中和研究项目（No.STZX202307).

(c) 2025 中国中文信息学会计算语言学专业委员会

applies two successive attention steps to strengthen emotion and semantic representations; second, a Value Integration Unit uses multi-stage gated fusion to dynamically balance the influence of user values and the system's historical values; and third, a Reaction Regulation Unit employs both bi-directional and cross-modal attention to reinforce the complementary interactions among emotion, semantics, and values. Experiments on the newly constructed ValueCon dataset demonstrate that HVDEGM outperforms baseline models such as DialogueRNN and DialogueGCN, improving Precision, Recall, F1, and emotional resonance by 2.9%, 2.5%, 0.9%, and 4.1%, respectively, thereby validating the effectiveness of the proposed approach.

Keywords: Emotion Generation , Value consistency , Attention mechanism , Emotional resonance

1 引言

情绪是对话系统与用户建立深层连接的桥梁：当机器能感知并恰当地表达情绪，就能让交互不再冰冷生硬，而是真正贴近人类的心理体验。情绪生成任务作为情感对话系统的一个重要研究任务(马志强et al., 2025)，其目标正是生成恰当的、能与用户产生共鸣的情绪状态。心理学研究表明，当倾听者的回应能够体现出对说话者情绪的理解时，这种回应更容易被感知为富有同理心(Yang and Jurgens, 2024)。相似性—吸引力假说指出，当个体感知到彼此在个性、价值观、兴趣等具有相似性时，便更容易产生积极的吸引力和信任感，进而促进关系的建立与维系(Abbasi et al., 2024)。而其中保持交互双方价值观一致性能直接影响交互双方信任感和吸引力，使交流更加顺畅交互意愿更加积极(Edwards and Cable, 2009)。

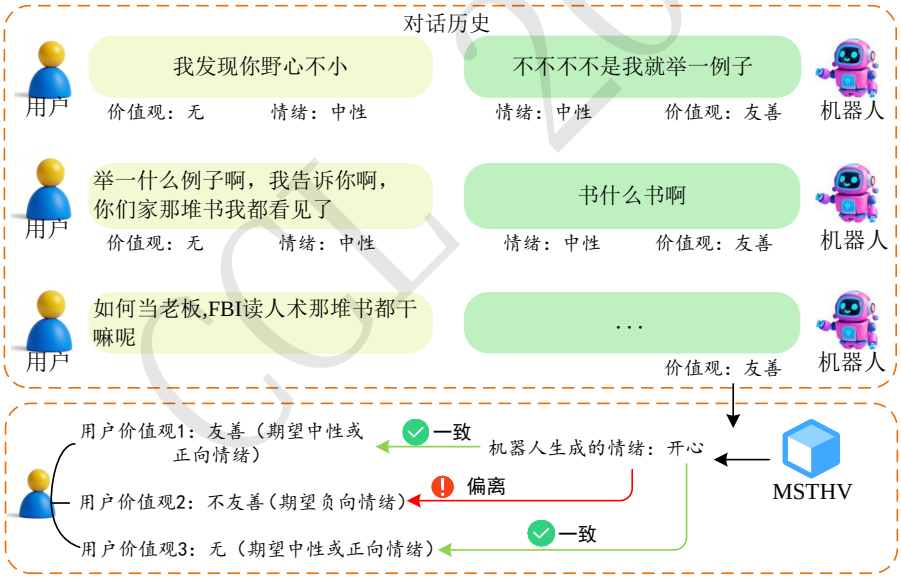

图 1: 情绪生成模型在价值观不一致情况下的响应偏差示例

然而，现有的情绪生成模型要么只是为模型预先设定一个固定不变的价值观，未考虑与用户之间的价值观一致性(Qiu et al., 2022)，要么完全忽略价值观等认知因素，而采用通用策略来生成情绪(Lu et al., 2023; 马志强et al., 2025)，导致对话系统生成情绪与用户期望情绪之间存在偏差，降低了对话系统与用户之间的情绪共鸣。如图1所示，当模型忽略用户价值观，或用户价值观恰好与模型预设价值观均为“友善”时，用户期望模型能回复中性或正向情绪，MSTHV模型(Qiu et al., 2022)生成的情绪状态符合用户的需求和期望；当用户价值观变化为“不友善”时，

©2025 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

用户期望模型能回复负向情绪，但此时模型并未考虑价值观一致性，MSTHV模型生成的情绪状态就与用户的期望情绪产生了偏差，难以引起用户情绪共鸣。所以，将交互过程中用户所表达的价值观融入情绪生成模型，并不断保持模型与用户的价值观一致，才能提升对话系统与用户的情绪共鸣。

因此，我们提出了一种人机价值观驱动的对话情绪生成模型（Human-Machine Values-Driven Dialogue Emotion Generation Model，HVDEGM），该模型通过情境修正注意力单元、价值观融合单元和反应调节单元，提取情绪与语义特征、动态融合价值观特征、强化三类特征之间的关联，保持对话系统与用户价值观一致性，主要贡献总结如下：

（1）设计了情境修正注意力单元（Contextualized Modified Attention Unit, CMAU），通过两次注意力机制增强了情绪与语义特征信息，为后续价值观融合与情绪生成提供上下文的情绪和语义特征支撑。

（2）设计了价值观融合单元（Value Integration Unit，VIU），通过引入价值观融合机制动态平衡了用户价值观特征与对话系统历史价值观特征的权重，在保持历史价值观稳定性与用户价值观响应性之间自动取舍，生成一个既保留对话系统历史惯性又反映用户偏好的融合向量。

（3）构建了基于双向注意力的反应调节单元（Reaction Regulation Unit，RRU），通过双向注意力与交叉注意力机制强化了情绪特征与价值观特征、语义特征与价值观特征之间的互补关联信息，确保最终输出的情绪既符合上下文语义，也严格对齐了融合价值观。

（4）在新构建的ValueCon对话数据集上进行了实验，结果表明HVDEGM模型能够有效融合用户价值观，缩小对话系统输出情绪与用户期望情绪之间的偏差，增强双方的情绪共鸣。

2 相关工作

对话情绪生成任务中，“认知因素”是指影响情感表达的内部心理属性，主要包括人格特质、价值观等多个类别。近年来，研究者开始关注如何将认知因素融入情绪生成模型以生成恰当的情绪。

Wen(2021)首次在情绪生成过程中考虑了人格因素，将人格特征和上下文编码分别映射到VAD空间中并进行融合，以生成能够反映特定人格的情绪回应。Qiu(2022)通过基于Transformer的价值函数来计算不同角色特征用户的价值偏好，并影响对话系统下做出更合理的情绪反应。Li(2023)针对现有研究多为被动响应，缺乏主动性和情感交互能力的问题，通过预测用户情绪期望来调节系统情绪。Hayat(2023)考虑到不同的说话人在情绪表达上具有差异，为每个说话者指定独立的分支来学习其特定的情绪反应。Ma(2024)通过改进lstm预测方法，构建了情绪生成模型，将人格建模与情绪生成相结合，生成了特定人格的情绪类别。Wen(2024)模拟了给定人格特征所影响的情绪转换过程，实现了大五人格特征与细粒度情绪的融合。Ehtesham-Ul-Haque(2024)提出了基于认知评估理论的情绪生成方法，通过计算信息变量来评估事件对情绪生成的影响，并结合模型预设的人格特征生成最终的情绪。

然而，现有研究往往只为模型注入静态的认知因素，而忽略了模型与用户间认知因素的相互作用与融合。事实上，情绪生成不仅受到单方认知因素的影响，更受交互双方认知因素融合结果的引导。为此，本文以价值观这一关键认知因素为切入点，融合建模了交互双方的价值观，共同影响情绪生成。

3 任务定义

人机价值观驱动的对话情绪生成任务描述：给定当前T时刻的用户输入$S_T^U = \left\{w_{T,1}^U, w_{T,2}^U, \ldots, w_{T,j}^U, \ldots, w_{T,L}^U\right\}$，$w_{T,j}^U$表示第$T$轮话语中用户话语的第$j$个词,且当前话语的长度为$L$；给定$T$时刻用户历史情绪序列$\text{EmoS}_T^U = \{e_1^U, e_2^U, \ldots, e_{T-1}^U, e_T^U\}$，其中$e_T^U$是一个13维向量，表示$T$时刻的用户的情绪状态；给定$T$时刻用户价值观向量$V_U^T = \left[v_1^T, v_2^T, \ldots, v_d^T\right], v_i^T \in \{-1, 0, 1\}$，其中$d$表示预设价值观的类别数量，$v_i^T$表示每个价值观类别的倾向性，取值1表示“支持/一致”，-1表示“反对/相悖”，0表示“中立/无关”。为防止用户可能带入的极端或不合理价值观对模型情绪输出产生负面影响，系统在对话开始时预设安全边界价值观$V_R^0 = [1, 1, \ldots, 1]$。在对话过程中，对话系统不断融合用户价值观，并在$T-1$时刻形成融合价值观$V_{T-1}^R$。对话情绪生成任务的目标是生成对话系统$T$时刻的情绪状态$e_T^R \in \mathcal{E}$（$|\mathcal{E}| = 13$），其中$\mathcal{E}$表示所有13种情绪类别。情绪生成过程可由公式(1)表示。

(c) 2025 中国中文信息学会计算语言学专业委员会

$$e_T^R = HVDEGM\left(S_T^U, EmoS_T^U, V_U^T, V_R^{T-1}\right) \tag{1}$$

4 方法

本文针对现有情绪生成模型忽视了用户与模型价值观一致性对情绪生成的调节与引导作用，导致对话系统生成情绪与用户期望情绪之间存在偏差，降低了对话系统与用户之间的情绪共鸣的问题，提出了一种人机价值观驱动的情绪生成模型（HVDEGM）。模型架构如图2所示。

HVDEGM模型的工作流程如下：在T时刻，HVDEGM 先用情境修正注意力单元（CMAU）提取情绪与语义特征，再通过价值观融合单元（VIU）生成反映系统–用户价值观一致性的融合向量V_R^T。最后，反应调节单元（RRU）将情绪–价值观和语义–价值观两路注意力结果与V_R^T 一同进行加权融合，确保每一次生成的情绪类别都既考虑了对话历史，也严格对齐当前的价值观引导信号，最终通过全连接层+Softmax 输出情绪e_T^R。

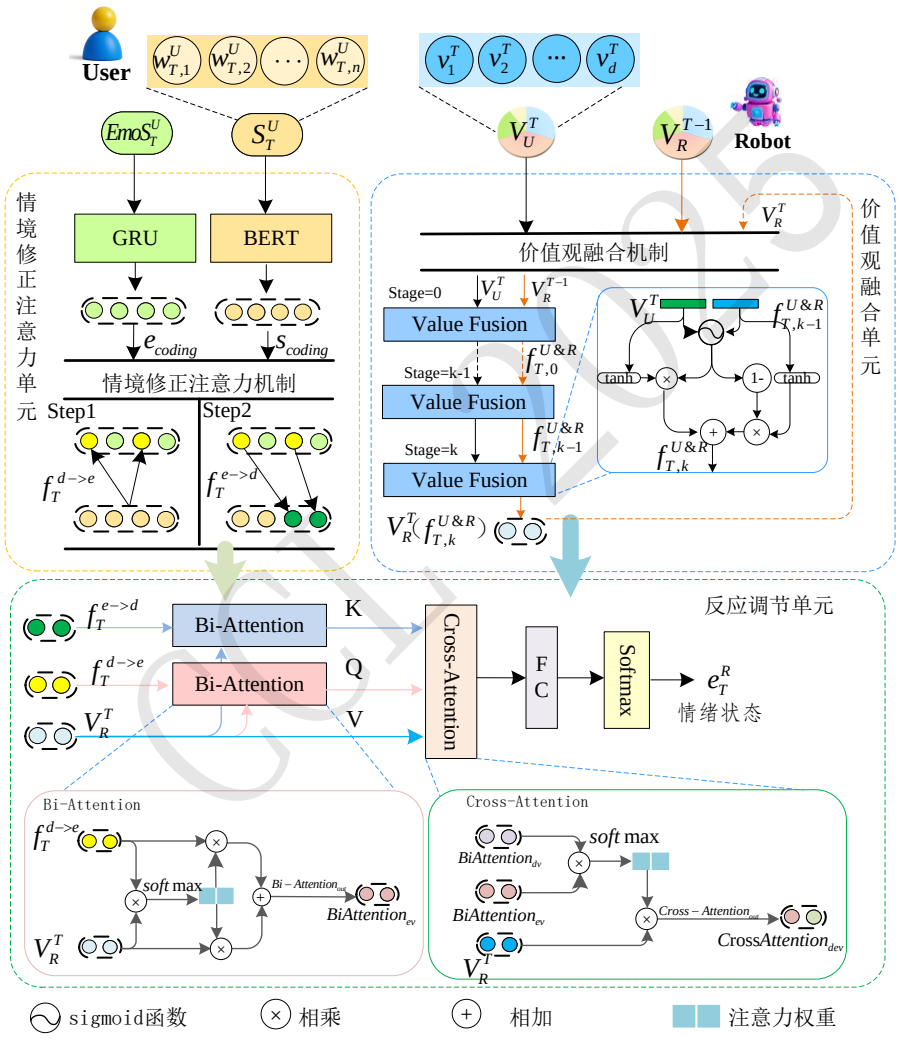

图 2: 人机价值观驱动的对话情绪生成模型

4.1 情境修正注意力单元

情境修正注意力单元旨在引入分两步的注意力机制，先捕捉对话中的关键情绪信号，再在这些信号引导下精炼对话历史信息，为价值融合和情绪生成提供高质量的情感与语义输入。本单元具体计算流程如下所示：

(c) 2025 中国中文信息学会计算语言学专业委员会

首先，利用GRU 和预训练语言模型BERT 分别对当前时刻的用户情绪序列和对话历史进行编码，以获得初步特征，具体公式如(2)(3)所示。

$$e_{\text{coding}} = \text{GRU}(\text{EmoS}_T^U) \tag{2}$$

$$s_{\text{coding}} = \text{BERT}(S_T^U) \tag{3}$$

其中，e_{coding} 表示情绪编码，s_{coding} 表示语义编码，EmoS_T^U 为用户情绪历史序列，$S_T^U = \{u_{T,1}^U, u_{T,2}^U, \ldots, u_{T,L}^U\}$ 表示第T 轮用户输入。

其次，第一阶段注意力机制通过点积评分聚焦与对话历史最相关的情绪特征，第二阶段注意力机制在选出的情绪特征指引下，提取与用户立场高度关联的对话语义。具体公式如(4)(5)所示。

$$f_T^{d\to e} = \text{softmax}\big(\tfrac{s_{\text{coding}} \cdot e_{\text{coding}}{}^T}{\sqrt{m_k}}\big) \cdot e_{\text{coding}} \tag{4}$$

$$f_T^{e\to d} = \text{softmax}\big(\tfrac{f_T^{d\to e} \cdot s_{\text{coding}}{}^T}{\sqrt{m_k}}\big) \cdot s_{\text{coding}} \tag{5}$$

其中，$\sqrt{m_k}$ 表示键向量的维度，$f_T^{d\to e}$ 表示与对话历史强相关的情绪特征，$f_T^{e\to d}$ 表示根据第一阶段识别出的与对话历史强相关的对话历史语义特征。此单元模拟了情绪感知后的人类认知回溯过程，为下游价值融合提供了精准的情感与语义输入。

4.2 价值观融合单元

价值观融合单元通过引入价值观融合机制，在对话第T 时刻，将该时刻的用户价值观V_U^T与上一时刻融合的历史价值观V_R^{T-1} 通过k 个迭代阶段加权融合，生成新的融合结果V_R^T，确保此时价值观输出符合对话系统与用户的价值观一致性。本单元计算流程如下：

首先，在第T 时刻的第k 个迭代阶段，分别令

$$v_1^k = V_U^T \tag{6}$$

$$v_2^k = \begin{cases} V_R^{T-1}, & k = 0 \\ f_{T,k-1}^{U\&R}, & k \geq 1 \end{cases} \tag{7}$$

其中，V_U^T 表示当前时刻的用户价值观向量，V_R^{T-1} 为上一时刻融合后的历史价值观向量，k为超参数，$f_{T,k-1}^{U\&R}$ 则为$k-1$阶段所得到的融合结果。

其次，在第k 阶段，通过可训练的线性映射与Sigmoid 激活计算融合权重，并对输入向量进行非线性映射后按维度加权融合，得到本轮的融合结果$f_{T,k}^{U\&R}$，如公式(8) ~ (11)所示。

$$w^k = \sigma\big(W_f\,[\,v_1^k;\, v_2^k\,] + b_f\big) \tag{8}$$

$$v_1^{k'} = \tanh(v_1^k), \quad v_2^{k'} = \tanh(v_2^k) \tag{9,10}$$

$$f_{T,k}^{U\&R} = w^k \odot v_1^{k'} + (1 - w^k) \odot v_2^{k'} \tag{11}$$

其中，w^k 为k 阶段的融合权重,$v_1^k, v_2^k \in \mathbb{R}^{d_v}$，$[v_1^k; v_2^k] \in \mathbb{R}^{2d_v}$ 表示按维度拼接，$W_f \in \mathbb{R}^{2d_v \times d_v}$、$b_f \in \mathbb{R}^{d_v}$ 为可训练的线性映射参数，$\sigma(\cdot)$ 表示逐元素的Sigmoid 非线性激活，输出维度为d_v。$v_1^{k'}$ 和$v_2^{k'}$ 表示中间向量，$tanh(\cdot)$是逐元素的双曲正切映射，用于对向量进行非线性变换。$f_{T,k}^{U\&R}$ 表示T 时刻第k 阶段价值融合输出，并且令$V_R^T = f_{T,k}^{U\&R}$参与到下一时刻的价值观融合和反应调节单元中。

(c) 2025 中国中文信息学会计算语言学专业委员会

4.3 反应调节单元

反应调节单元通过双向注意力和交叉注意力，对情绪特征、语义特征与融合后的价值观特征三者进行深度融合和再平衡，确保最终输出是既符合对话语境又代表价值观一致性的情绪状态。本单元的具体计算流程如下：

首先，分别计算情绪—价值和语义—价值的双向注意力，如公式(12) ∼ (15)所示

$$A_{ev} = \text{softmax}\left(\frac{f_T^{d\to e} \cdot {V_R^T}^T}{\sqrt{d_v}}\right) \tag{12}$$

$$\text{BiAttention}_{ev} = A_{ev} \bullet V_R^T + A_{ev}^T \bullet f_T^{d\to e} \tag{13}$$

$$A_{dv} = \text{softmax}\left(\frac{f_T^{e\to d} \cdot {V_R^T}^T}{\sqrt{d_v}}\right) \tag{14}$$

$$\text{BiAttention}_{dv} = A_{dv} \bullet V_R^T + A_{dv}^T \bullet f_T^{e\to d} \tag{15}$$

其中，A_{ev} 表示情绪特征与价值观特征之间的注意力权重,BiAttention $_{\text{ev}}$ 表示情绪特征和价值观特征之间的双向注意力融合结果，A_{ab} 表示语义特征和价值观特征之间的注意力权重，BiAttention_{ab} 表示语义特征和价值观特征之间的双向注意力融合结果，$\sqrt{d_v}$表示价值观特征向量的维度。

其次,将两个注意力的输出特征BiAttention_{ev} 和BiAttention_{dv} 同时与价值观一致性特征V_R^T进行融合相乘,得到交叉注意力融合向量$\text{CrossAttention}_{dev}$，并将其送入全连接层和softmax 函数得出最终情绪状态，如公式(16) ∼ (18)所示。

$$\text{CrossAttention}_{dev} = \text{softmax}\left(\frac{\text{BiAttention}_{ev} \cdot {\text{BiAttention}_{dv}}^T}{\sqrt{d_A}}\right) \cdot V_R^T \tag{16}$$

$$e_{\text{output}} = W_{FC} \cdot \text{CrossAttention}_{dev} + b_{FC}, \quad e_T^R = \arg\max(\text{softmax}(e_{\text{output}})) \tag{17,18}$$

其中，$\text{CorssAttention}_{\text{dev}}$ 为$\text{BiAttention}_{\text{ev}}$ 和$\text{BiAttention}_{\text{dv}}$ 同时与价值观特征V_R^T 经过Corss-Attention 融合后得到的情绪向量，e_{output} 为全连接层的输出，w_{FC} 和b_{FC} 分别为全连接层的权重和偏置项，将e_{output} 中权重最大的情绪状态作为T 时刻对话系统情绪状态e_T^R 。

5 实验

5.1 实验设置

5.1.1 数据集介绍

由于尚缺乏带有价值观标注的情感对话数据集，在实验部分，我们构建了ValueCon数据集以支持人机价值观驱动的对话情绪生成研究。该数据集基于开源中文情感对话数据集CPED(Chen et al., 2022)扩展而来，通过引入价值观标注增强了认知维度建模能力。具体而言，我们选取社会主义核心价值观中的个人层面的价值目标并采用Zero-Shot-CoT框架对132,760条对话进行自动化四维价值观标注，分别标注为爱国、敬业、诚信、友善,并使用1、-1 和0 分别表示话语或情绪的价值观属性与特定价值观相同、相反和无关，最后经人工验证确保标注有效性。针对原始数据中“爱国”维度样本极端稀疏的问题（仅占2.4%），我们通过数据增强生成5,171条补充样本，最终形成的ValueCon数据集完整保留了CPED原有的情感极性、13类情绪标签及大五人格特征，其价值观分布如表1所示。

表 1: ValueCon数据集价值观分布统计

价值观类别	一致(1)	相反(-1)	无关(0)	总数
爱国	3,584	386	98,790	102,760
敬业	9,621	1,715	2,000	13,336
诚信	4,268	4,175	0	8,443
友善	18,705	13,530	0	32,235

(c) 2025 中国中文信息学会计算语言学专业委员会

5.1.2 实验环境

实验使用的硬件配置为NVIDIA TESLA V100S 32G GPU、NVIDIA GeForce RTX 2060 GPU及Intel(R) Core(TM) i7-9700 CPU，操作系统为Windows。软件环境包括CUDA 11.0和PyTorch 1.7深度学习框架，训练参数配置为：优化器采用SGD（随机梯度下降），初始学习率设为$\eta = 0.01$，Dropout率设为$p = 0.4$，Batch Size设为$N = 64$，迭代次数设为$T = 50$。

5.1.3 基线模型

为了验证本文提出的人机价值观驱动的对话情绪生成模型的有效性，本文选择了以下几个情绪生成任务常用的基线模型：DialogueRNN(Majumder et al., 2019)：多模态对话情绪识别模型，同样能够用于对话情绪生成，该模型同时考虑了对话中的说话者状态和上下文信息，本文实验中仅考虑了文本模态；DialogueGCN(Ghosal et al., 2019)：对话情绪识别模型，同样能够用于对话情绪生成，该模型通过建模说话人之间的依赖关系来捕捉对话中的情绪上下文；IDS-ECM(Li et al., 2020)：对话情绪生成模型，它通过引入情感记忆模块，能够更好地捕捉对话中的情感动态；MSTHV(Qiu et al., 2022)：通过预训练的价值模型对多个情感响应的价值观得分排序，间接影响情绪生成的结果；CEM(Sabour et al., 2022)：共情对话生成模型，通过引入常识知识来增强响应的同理心，本文实验中，去掉了其中的知识选择器和响应生成器，只输出了情绪生成结果；DiaRP(Yingjian et al., 2023)：对话情绪识别模型，通过捕捉对话关系和情绪生成来增强情绪识别性能，本文中去掉了该模型中情绪识别模块，只保留情绪生成结果；ECoT(Li et al., 2024)：即插即用的大模型提示方法，旨在提升大模型在各类情绪生成任务上的表现。本实验中，我们将ECoT与ERNIE4.0(Deng et al., 2024) 结合作为对比基线。

5.2 评价指标

由于情绪具有主观性, 为了更加全面的评估模型在情绪生成任务上的表现, 采用自动评价和主观评价结合的方式对模型进行评估,具体介绍如下。

5.2.1 自动评价指标

自动评价指标为Precision、Recall、F1、情绪相似度指标R1、情绪积极性指标R2、情绪共鸣度指标R3。

精度Precision: 正确分类的阳性样本数与分类器确定为阳性样本的样本数之比,计算公式为:

$$Precision = \frac{S_{\text{true}}}{S_{\text{true}}^{\text{all}}} \tag{19}$$

召回率Recall：正确分类的阳性样本数与真实阳性样本数之比,计算公式为：

$$Recall = \frac{S_{\text{true}}}{S_{\text{all}}} \tag{20}$$

其中, S_{true} 为正确分类的正样本数量, $S_{\text{true}}^{\text{all}}$ 为模型认为是正样本的数量，S_{all} 为真实的正样本数量。

F1 值: 精度和召回率的调和平均值,是两者的结果的结合,计算公式为:

$$\text{F1} = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{21}$$

情绪相似度R1(黄宏程et al., 2021): 通过情绪向量间的余弦相似度来计算其相似性,计算如公式(22)所示:

$$R_1 = S\left(e_T^U, e_T^R\right) = \frac{I\left(e_T^U\right) \cdot I\left(e_T^R\right)}{\left\|I\left(e_T^U\right)\right\| \left\|I\left(e_T^R\right)\right\|} \tag{22}$$

其中, $I\left(e_T^U\right)$ 和$I\left(e_T^R\right)$ 分别表示T 时刻用户和对话系统的情绪向量。

情绪积极性R2(黄宏程et al., 2021): 情绪积极性表示生成情绪的情感积极性,计算如公式(23)所示:

$$R_2 = P\left(e_{T+1}^R\right) = P\left(I\left(e_{T+1}^R\right)\right) = \sum_{j=1}^{13} l_j i_j \tag{23}$$

(c) 2025 中国中文信息学会计算语言学专业委员会

其中, l_j 为13 种基本情绪的积极性权重, i_j 为语境中待评价情绪朝情绪j 的转移概率。

情绪共鸣度R3(黄宏程et al., 2019): 根据相似性原则(Duck, 1977),对某种事物或事件具有相同或相似的态度, 有共同的理想信念和价值观, 感情上就容易产生共鸣, 计算如公式(24)所示:

$$R_3 = M\left(e_{T,i}^{U}, e_{T}^{R}\right) = \sum_{i=1}^{13}\sum_{j=1}^{13}\left[\frac{f\left(e_{T,i}^{U}, e_{T,j}^{R}\right)}{N}\right]^4 \tag{24}$$

其中, N为对话历史中用户情绪为$e_{T,j}^{U}$ 时的总次数，$f\left(e_{T,i}^{U}, e_{T,j}^{R}\right)$ 为对话历史中用户情绪为$e_{T,j}^{U}$时,交互情绪类别为$e_{T,j}^{R}$ 的次数，并且只有对$e_{T,j}^{R}$ 的回应为积极态度才算是有效次数。

5.2.2 人工评价指标

为直接验证模型对情绪共鸣度问题的解决效果，设计以下与核心问题强关联的人工评价指标（5位标注者独立评分后取平均）：

共鸣性：生成情绪与用户潜在价值观（通过历史对话推断）及系统预设价值观的双向匹配程度，采用[-1,1]对称区间量化冲突与共识（-1=完全违背双方价值观，1=完美融合双方取向）

恰当性：情绪反应在具体对话语境中的合理性基于5级Likert量表映射至[-1,1]区间（-1=极端不恰当如“收到礼物时愤怒”，1=高度契合如“遭遇不公时生气”）

6 实验结果

6.1 情绪生成对比实验

为验证HVDEGM模块的有效性，使用基线模型与HVDEGM在ValueCon数据集上进行实验，实验结果如表2所示。

表 2: 情绪生成对比实验

模型	Precision	Recall	F1	R1	R2	R3
DialogueRNN	0.379	0.375	0.380	0.342	0.361	0.359
DialogueGCN	0.381	0.388	0.384	0.462	0.442	0.419
IDS-ECM	0.394	0.385	0.391	0.456	0.448	0.394
MSTHV	0.416	0.412	0.412	**0.472**	0.448	0.392
CEM	0.445	0.441	0.445	0.462	0.412	0.407
DiaRP	0.409	0.426	0.421	0.453	0.441	0.412
ECoT	0.421	0.423	0.421	0.471	0.397	0.410
HVDEGM	**0.458**	**0.452**	**0.449**	0.468	**0.451**	**0.436**

由表2可得，HVDEGM模块的表现总体优于基线模型，仅R1指标低于最优基线0.8%，HVDEGM模块的Precision、Recall、F1、R2和R3指标分别相较于最优基线提升了2.9%、2.5%、0.9%、0.7%和4.1%，这表明HVDEGM模块在情绪生成任务上具有有效性。具体来说，相较于最优基线，HVDEGM对R3指标即生成情绪的共鸣度方面提升较为显著，这表明在情绪生成的过程中引入价值观因素有利于提升生成情绪与用户共鸣，验证了本文提出的假设，即引入价值观因素能够使生成的情绪更加符合用户的期望，此外，HVDEGM模块在Precision、Recall、F1和R2指标提升并不显著，可能是由于情绪生成同时受到多种复杂因素的影响，HVDEGM模块独立考虑了价值观因素而忽略了大五人格、对话行为和对话主题等其他与对话情绪生成相关的因素，导致HVDEGM模块在部分指标上的提升并不显著。

6.2 细粒度情绪生成对比实验

为了进一步分析HVDEGM在细粒度情绪生成上的表现，统计了各基线模型和HVDEGM模块对细粒度情绪生成的Precision、F1和Recall指标，ValueCon中包括13种细粒度情绪，为了便于展示和分析，本文选择了其中占比最高的7种细粒度情绪的实验结果进行展示，如表3所示。

(c) 2025 中国中文信息学会计算语言学专业委员会

表 3: 情绪细粒度对比实验

类别	指标	DialogueRNN	DialogueGCN	IDS-ECM	MSTHV	CEM	DiaRP	ECoT	HVDEGM
Neutral	Precision	0.535	0.514	0.543	0.541	0.567	0.555	0.575	**0.576**
	F1	0.531	0.513	0.538	0.537	0.565	0.552	0.574	**0.577**
	Recall	0.533	0.512	0.535	0.534	0.564	0.547	**0.575**	0.571
Anger	Precision	0.532	0.536	0.559	0.544	0.559	0.576	0.564	**0.587**
	F1	0.532	0.537	0.553	0.539	0.555	0.573	0.563	**0.577**
	Recall	0.531	0.535	0.544	0.541	0.557	0.569	0.561	**0.578**
Depress	Precision	0.489	0.483	0.513	0.498	0.545	0.531	0.554	**0.555**
	F1	0.493	0.484	0.512	0.499	0.546	0.529	0.553	**0.554**
	Recall	0.496	0.477	0.514	0.501	0.547	0.527	0.554	**0.559**
Negative-other	Precision	0.487	0.527	0.549	0.534	0.553	0.541	0.563	**0.574**
	F1	0.488	0.521	0.547	0.534	0.549	0.543	0.564	**0.569**
	Recall	0.491	0.516	0.547	0.536	0.544	0.539	0.563	**0.571**
Relaxed	Precision	0.501	0.513	0.537	0.524	0.549	0.536	0.556	**0.567**
	F1	0.496	0.515	0.533	0.522	0.547	0.538	**0.559**	0.557
	Recall	0.497	0.512	0.529	0.525	0.548	0.535	0.555	**0.559**
Worried	Precision	0.481	0.512	0.523	0.497	0.547	0.551	0.559	**0.561**
	F1	0.484	0.511	0.531	0.497	0.544	0.548	**0.557**	0.554
	Recall	0.483	0.509	0.532	0.501	0.544	0.549	0.554	**0.557**
Happy	Precision	0.523	0.526	0.543	0.535	0.542	0.543	0.547	**0.572**
	F1	0.515	0.522	0.537	0.523	0.544	0.542	0.551	**0.569**
	Recall	0.508	0.514	0.535	0.513	0.541	0.542	0.449	**0.563**

由表3可得，HVDEGM在多数细粒度情绪类别上的表现超过了最优基线，对表中7种细粒度情绪生成的Precision、F1和Recall指标平均提升了1.59%、0.67%和1.20%。其中最显著的是“Happy”、“Anger”和“Negative-other”情绪的P指标相较于最优基线分别提升了4.6%、1.9% 和2.0%，F1指标相较于最优基线分别提升了3.3%、0.70%和0.89%，Recall指标相较于最优基线分别提升了3.9%、1.58%和1.42%，可能是由于“Happy”情绪与“友善”价值观的基本内涵相同且强相关，而“Anger”和“Negative-other”情绪与“友善”价值观的基本内涵相反且强相关，在ValueCon数据集中，“友善”价值观占比最高，而模型在“Neutral”、“Depress”、“Relaxed”和“Worried”这类价值观内涵较为模糊的情绪类别上的表现则相对于基线模型没有较为显著的提升，这表明HVDEGM在价值观强相关的情绪类别的生成上有更显著的提升，而对于价值观内涵较为模糊的情绪类别的提升仍然有限。

6.3 消融实验

为了进一步验证情境修正注意力单元(CMAU)、价值观融合单元(VIU)和反应调节单元(RRU)在情绪生成任务上的有效性，本文通过分别去除CMAU、VIU和RRU来分析其对实验结果的影响，实验结果如表4所示。

表 4: 情绪生成消融实验

模型	Precision	Recall	F1	R1	R2	R3
-HVDEGM w/o CMAU	0.447	0.436	0.435	0.461	0.441	0.418
-HVDEGM w/o VIU	0.439	0.442	0.436	0.463	0.439	0.403
-HVDEGM w/o RRU	0.445	0.437	0.431	0.462	0.445	0.422
HVDEGM	**0.458**	**0.452**	**0.449**	**0.468**	**0.451**	**0.436**

由表4可见，去除CMAU、YIU 和RRU 三个关键单元后，HVDEGM 的性能均有所下降，但三者对不同指标的影响程度并不相同。首先，去除CMAU后，各指标分别下降约1.5% ∼ 4.1%，其中R3 指标下降最为显著，达到4.1%，这表明缺少“情境修正注意力”后，模型在提取语义和情绪线索时容易对话上下文信息理解不足。具体而言，缺失CMAU后，后续的反应调节单元（RRU）只能直接使用未经精炼的原始情绪与语义编码进行与价值观向量的融合，此时输

(c) 2025 中国中文信息学会计算语言学专业委员会

入特征混杂了大量无关或冗余信息，使得交叉注意力无法准确对齐情绪信号与价值观指向，最终生成的情绪更易偏离用户期望，导致共鸣度（R3）下降最为显著。其次，去除VIU 后，各指标下降幅度分别约为1.1% ∼ 7.6%，其中R3 下降最明显，为7.6%。具体来看，VIU 在每一阶段通过sigmoid 得到的融合权重w^k 本质上是在动态平衡用户价值观与对话系统价值观的影响力，去掉该机制后，HVDEGM 只能沿用对话开始时设置的固定的对话系统价值观，无法动态融合用户价值观，导致生成情绪与用户价值观的契合度大幅降低。最后，在去除RRU 后，所有指标也都有不同幅度的下降，其中最显著的是F1 下降了4.0%。RRU 通过双向注意力先分别将情绪与价值观、语义与价值观进行双向交互，再通过交叉注意力将两者融合，使得最终输出情绪既考虑上下文语境，又保留了价值观的平衡，缺少这一模块则会导致多源信息未能有效融合，模型在特征匹配和情绪平衡上出现偏差，从而使F1 等指标下降。综上，消融实验揭示了三个模块各自的不可替代性：CMAU 负责上下文场景对情绪特征的精准提炼，VIU 负责动态融合用户与对话系统双方价值观，而RRU 则负责将这多个特征进行最终的深度融合并注入价值导向。

6.4 人工评价实验

在人工评价时，使用从测试集抽取出的100组多轮对话样本进行了对话情绪生成，并邀请了5位评审员从一致性和恰当性两个角度上进行人工打分。人工评价归一化结果如表5所示。

表 5: 人工评价实验

模型	共鸣性	恰当性
DialogueRNN	0.231	0.431
DialogueGCN	0.282	0.462
IDS-ECM	0.267	0.472
MSTHV	0.273	0.463
CEM	0.305	0.491
DiaRP	0.278	0.486
ECoT	0.295	0.484
HVDEGM	**0.319**	**0.495**

共鸣性方面，HVDEGM显著优于其他未考虑价值观因素的基线模型，通过指定价值观，在一定程度上能够控制模型的情绪生成过程，使生成情绪发生转移，进一步验证了融入价值观因素在情绪生成任务中的有效性；恰当性方面，HVDEGM相较于最优基线提升了0.8%，这表明引入价值观后的HVDEGM不仅提升了生成情绪与指定价值观的一致性，且没有损失通用的情绪生成能力。综上所述，相较于多个基线模型，HVDEGM模型在情绪生成任务上呈现了最好的综合表现，验证本文提出方法的有效性。

6.5 案例分析

为了更清晰地展示HVDEGM在对话情绪生成任务上的表现，从ValueCon测试集中选取一组对话，通过指定双方价值观，形成3组案例。使用HVDEGM和MSTHV进行案例测试，为方便测试分析，在生成情绪输出后，我们不仅选取权重最高的情绪状态作为主输出，还将权重排名前三的其余两个情绪状态一并纳入候选情绪列表，以便后续对模型表现进行更全面的评估，如表6所示。

由表6可得，在A组案例中，未考虑人机价值观，MSTHV 生成的结果正好与真实标签“neutral”保持一致，而HVDEGM 反而输出了与真实偏差较大的“negative-other”，这可能是由于MSTHV模型中引入了动态更新的心智状态图，为情绪生成提供了充足的上下文信息，能够很好的捕捉对话中的情感动态，而HVDEGM 仅将用户当前时刻的情绪序列作为输入，相当于仅关注了最后一轮对话信息，忽略了更早轮次对情绪的潜在影响。在B 组和C 组中，当人机价值观一致时（C 组），MSTHV 与HVDEGM 均生成了“happy”，说明二者在没有价值观冲突时都能够识别出对话意图中的积极情绪；而在价值观冲突（B 组）时，MSTHV 仅参考系统的价值观，直接输出符合系统“友善”价值导向的“happy”，却忽视了用户负向价值观倾向，这样的情绪的生成具有一定的随机性；而HVDEGM生成的首位情绪状态为“positive-other”，虽然也与用户价值观矛盾，但相较于MSTHV，在用户价值观和对话系统价值观之间表现出了更

第二十四届中国计算语言学大会论文集，第627页–第639页，济南，中国，2025年8月11日至14日。

(c) 2025 中国中文信息学会计算语言学专业委员会

表 6: HVDEGM和MSTHV对话情绪生成案例展示

组别	项目	内容
A组不指定价值观	对话历史	U：我发现你野心不小[neutral] [0,0,0,0]；R：不不不不是我就举一例子[neutral] [0,0,0,0]；U：举一什么例子啊，我告诉你啊，你们家那堆书我都看见了[neutral] [0,0,0,0]；R：书什么书啊[neutral] [0,0,0,0]；U：如何当老板，FBI读人术那堆书都干嘛呢[neutral] [0,0,0,0]
	生成情绪状态列表	MSTHV: [neutral, anger, relaxed];HVDEGM: [negative-other, relaxed, neutral]
	真实标签	[neutral]
B组指定交互双方具有不同的价值观	对话历史	U：我发现你野心不小[neutral] [0,0,0,-1]；R：不不不不是我就举一例子[neutral] [0,0,0,1]；U：举一什么例子啊，我告诉你啊，你们家那堆书我都看见了[neutral] [0,0,0,-1]；R：书什么书啊[neutral] [0,0,0,1]；U：如何当老板，FBI读人术那堆书都干嘛呢[neutral] [0,0,0,-1]
	生成情绪状态列表	MSTHV: [happy, relaxed, neutral];HVDEGM: [positive-other, anger, neutral]
	真实标签	[neutral]
C组指定交互双方具有相同的价值观	对话历史	U：我发现你野心不小[neutral] [0,0,0,1]；R：不不不不是我就举一例子[neutral] [0,0,0,1]；U：举一什么例子啊，我告诉你啊，你们家那堆书我都看见了[neutral] [0,0,0,1]；R：书什么书啊[neutral] [0,0,0,1]；U：如何当老板，FBI读人术那堆书都干嘛呢[neutral] [0,0,0,1]
	生成情绪状态列表	MSTHV: [happy, relaxed, neutral];HVDEGM: [happy, positive-other, relax]
	真实标签	[neutral]

优的平衡效果，并且结合第二候选情绪“anger”和第三候选情绪“neutral”来看，HVDEGM生成的情绪也具有一定的随机性。上述结果表明，首先，对话历史对情绪生成具有很大的影响，重要性不可低估；其次，交互双方的价值观都会对情绪生成结果产生影响并且当人机价值观一致时，MSTHV和HVDEGM都能生成出符合对话系统和用户价值观强相关的情绪状态，而当交互双方的价值观不一致时，MSTHV和HVDEGM模块的生成结果都存在一定的随机性。这表明指定静态的对话系统价值观并不能很好的对齐不同语境下的用户价值偏好，也表明HVDEGM模块在人机价值观不同的情况下存在一定的局限性。

7 结论

本文针对现有情绪生成模型忽视了用户与模型价值观一致性对情绪生成的调节与引导作用，导致对话系统生成情绪与用户期望情绪之间存在偏差，降低了对话系统与用户之间的情绪共鸣问题开展研究，设计了人机价值观驱动的对话情绪生成模型（HVDEGM）。HVDEGM包括三个单元：情境修正注意力单元（CMAU）、价值观融合单元（VIU）和反应调节单元（RRU），分别完成语义与情绪特征提取、系统–用户价值观一致性融合，以及基于多注意力的最终情绪生成。

在新构建的ValueCon 数据集上，HVDEGM 相较最优基线在Precision、Recall、F1、R2和R3 指标上分别取得了2.9%、2.5%、0.9%、0.7% 和4.1% 的提升；在细粒度情绪生成实验中，对“Anger” “Negative-other” “Happy”等与价值观高度相关的情绪类别表现尤为突出；消融

(c) 2025 中国中文信息学会计算语言学专业委员会

实验则证明了CMAU、VIU 和RRU 对模型整体性能的有效贡献；案例分析进一步表明，系统与用户价值观的一致性能够稳定引导情绪生成，而在价值观冲突情境下，生成结果会呈现一定随机性。

参考文献

Zoleikha Abbasi, Jon Billsberry, and Mathew Todres. 2024. Empirical studies of the "similarity leads to attraction" hypothesis in workplace interactions: a systematic review. *Management Review Quarterly*, 74(2):661–709.

Yirong Chen, Weiquan Fan, Xiaofen Xing, Jianxin Pang, Minlie Huang, Wenjing Han, Qianfeng Tie, and Xiangmin Xu. 2022. Cped: A large-scale chinese personalized and emotional dialogue dataset for conversational ai. *arXiv preprint arXiv:2205.14727*.

Hourui Deng, Hongjie Zhang, Jie Ou, and Chaosheng Feng. 2024. Can llm be a good path planner based on prompt engineering? mitigating the hallucination for path planning. *arXiv preprint arXiv:2408.13184*.

S. Duck. 1977. *Theory and Practice in Interpersonal Attraction*. Academic Press, Cambridge.

Jeffrey R Edwards and Daniel M Cable. 2009. The value of value congruence. *Journal of applied psychology*, 94(3):654.

Md Ehtesham-Ul-Haque, Jacob D'Rozario, Rudaiba Adnin, Farhan Tanvir Utshaw, Fabiha Tasneem, Israt Jahan Shefa, and ABM Alim Al Islam. 2024. Emobot: Artificial emotion generation through an emotional chatbot during general-purpose conversations. *Cognitive Systems Research*, 83:101168.

Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. 2019. Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation. *arXiv preprint arXiv:1908.11540*.

Hassan Hayat, Carles Ventura, and Agata Lapedriza. 2023. Predicting the subjective responses' emotion in dialogues with multi-task learning. In *Iberian Conference on Pattern Recognition and Image Analysis*, pages 693–704. Springer.

Dayu Li, Yang Li, and Suge Wang. 2020. Interactive double states emotion cell model for textual dialogue emotion prediction. *Knowledge-Based Systems*, 189:105084.

Yuanchao Li, Koji Inoue, Leimin Tian, Changzeng Fu, Carlos Toshinori Ishi, Hiroshi Ishiguro, Tatsuya Kawahara, and Catherine Lai. 2023. I know your feelings before you do: Predicting future affective reactions in human-computer dialogue. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–7.

Zaijing Li, Gongwei Chen, Rui Shao, Yuquan Xie, Dongmei Jiang, and Liqiang Nie. 2024. Enhancing emotional generation capability of large language models via emotional chain-of-thought. *arXiv preprint arXiv:2401.06836*.

Xin Lu, Weixiang Zhao, Yanyan Zhao, Bing Qin, Zhentao Zhang, and Junjie Wen. 2023. A topic-enhanced approach for emotion distribution forecasting in conversations. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Zhiqiang Ma, Wenchao Jia, Yutong Zhou, Biqi Xu, Zhiqiang Liu, and Zhuoyi Wu. 2024. Personality enhanced emotion generation modeling for dialogue systems. *Cognitive Computation*, 16(1):293–304.

Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. 2019. Dialoguernn: An attentive rnn for emotion detection in conversations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 6818–6825.

Liang Qiu, Yizhou Zhao, Yuan Liang, Pan Lu, Weiyan Shi, Zhou Yu, and Song-Chun Zhu. 2022. Towards socially intelligent agents with mental state transition and human value. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 146–158.

Sahand Sabour, Chujie Zheng, and Minlie Huang. 2022. Cem: Commonsense-aware empathetic response generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11229–11237.

第二十四届中国计算语言学大会论文集，第627页–第639页，济南，中国，2025年8月11日至14日。

(c) 2025 中国中文信息学会计算语言学专业委员会

Zhiyuan Wen, Jiannong Cao, Ruosong Yang, Shuaiqi Liu, Jiaxing Shen, et al. 2021. Automatically select emotion for response via personality-affected emotion transition. Association for Computational Linguistics (ACL).

Zhiyuan Wen, Jiannong Cao, Jiaxing Shen, Ruosong Yang, Shuaiqi Liu, and Maosong Sun. 2024. Personality-affected emotion generation in dialog systems. *ACM Transactions on Information Systems*, 42(5):1–27.

Jiamin Yang and David Jurgens. 2024. Modeling empathetic alignment in conversation. *arXiv preprint arXiv:2405.00948*.

Liu Yingjian, Wang Xiaoping, and Lei Shanglin. 2023. Emotion prediction in conversation based on relationship extraction. In *2022 IEEE International Conference on Cyborg and Bionic Systems (CBS)*, pages 53–58. IEEE.

马志强, 吕凯, 周钰童, 刘佳, 叶浩然, 刘义兴, and 王奎波. 2025. 基于多样性情绪的积极导向自然过渡决策模型. 计算机工程与应用, pages 1–12.

黄宏程, 刘宁, 胡敏, 陶洋, and 寇兰. 2019. 基于博弈的机器人认知情感交互模型. 电子与信息学报, 41(10):2471–2478.

黄宏程, 李净, 胡敏, 陶洋, and 寇兰. 2021. 基于强化学习的机器人认知情感交互模型. 电子与信息学报, 43(06):1781–1788.

CCL 2025