通过领域提示和并行注意力实现对话中的可泛化参与度估计

Yangchen Yu*
School of Computer Science and Information Engineering, Hefei University of Technology
Hefei, China
2019212292@mail.hfut.edu.cn

Yin Chen*
School of Computer Science and Information Engineering, Hefei University of Technology
Hefei, China
chenyin@mail.hfut.edu.cn

Jia Li†
School of Computer Science and Information Engineering, Hefei University of Technology
Hefei, China
jiali@hfut.edu.cn

Peng Jia
School of Computer Science and Information Engineering, Hefei University of Technology
Hefei, China
2020214631@mail.hfut.edu.cn

Yu Zhang
School of Computer Science and Information Engineering, Hefei University of Technology
Hefei, China
yuyueback@gmail.com

Li Dai
School of Computer Science and Information Engineering, Hefei University of Technology
Hefei, China
321daili123@gmail.com

Zhenzhen Hu
School of Computer Science and Information Engineering, Hefei University of Technology
Hefei, China
zzhu@hfut.edu.cn

Meng Wang
School of Computer Science and Information Engineering, Hefei University of Technology
Hefei, China
eric.mengwang@gmail.com

Richang Hong
School of Computer Science and Information Engineering, Hefei University of Technology
Hefei, China
hongrc.hfut@gmail.com

摘要

准确的参与度估计对于自适应人机交互系统至关重要，然而由于在不同领域（如文化与语言）中的泛化能力较差以及建模复杂交互动态的挑战，其稳健部署受到阻碍。为解决这些问题，我们提出了**达帕**（**D**omain-**A**daptive**P**arallel**A**ttention），一种用于可泛化的对话参与度建模的新框架。DAPA 通过在输入前添加可学习的领域特定向量引入了域提示机制，明确地使模型基于数据来源进行条件设置，以促进领域感知适应的同

*Equal contribution.
†Corresponding author.

Conference'17, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

时保持泛化参与表示。为了捕捉交互同步性，该框架还结合了一个并行交叉注意力模块，显式对齐参与者之间的反应（正向 BiLSTM）和预期（反向 BiLSTM）状态。广泛的实验表明，DAPA 在多个跨文化和跨语言基准测试中建立了新的最先进的性能，在 NoXi-J 测试集上相对于强基线在一致性相关系数 (CCC) 绝对提升了 0.45。我们的方法的优越性还通过赢得 MultiMediate'25 多领域参与度估计挑战赛的第一名得到了确认。源代码将在 https://github.com/MSA-LMC/DAPA 公开提供。

CCS Concepts

• **Human-centered computing → Empirical studies in HCI**.

Keywords

参与度估计，领域适应，交互建模，多模态分析，情感计算

ACM Reference Format:

Yangchen Yu, Yin Chen, Jia Li, Peng Jia, Yu Zhang, Li Dai, Zhenzhen Hu, Meng Wang, and Richang Hong. 2025. 通过领域提示和并行注意力实现对话中的可泛化参与度估计. In . ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 介绍

寻求真正富有同理心和协作性的 AI 系统的关键在于其感知人类参与度的能力——一个人在对话中表达的投入程度 [12, 21, 39]。参与度不是一个静态属性，而是一个动态的交互现象，通过诸如眼神、点头和语音反馈等多种模态信号丰富地传达出来 [27, 34]。然而，构建稳健且可泛化的参与模型的道路受到两大基本障碍的阻碍。

其中最为深远的是源于文化与语言多样性所产生的领域差距，[4, 17, 24]。诸如点头或停顿时长等行为的沟通功能在不同文化中可能有极大的差异，导致在一个环境中训练出来的模型在另一个环境中无法正常工作。这种差距极大地限制了当前系统在实际应用中的效用。

加剧这一问题的是在建模参与度人际动态的过程中存在持续的疏忽 [19, 46]。虽然普遍认为这是一种互动的 [7] 衍生属性，但许多模型将参与简化为个体的状态，忽视了他们的对话伙伴 [20, 49] 的关键影响。即使最近通过简单的交叉注意力 [21] 尝试整合伙伴信息的努力也未能达到目的，无法捕捉到支撑真正连接的细微、瞬间同步性 [19]。

针对这些交织的挑战，我们引入了**领域自适应并行注意力网络（DAPA）**，一个旨在实现稳健且互动感知的参与度建模的新框架。DAPA 通过**领域提示**机制应对领域差异，在可学习向量前添加以明确条件化模型对数据文化起源的关注。为了解释复杂的互动动态，它使用了**并行交叉注意力**模块。该模块独特地对齐参与者的反应性的（前向 BiLSTM）和预期的（后向 BiLSTM）状态，从而捕捉到瞬间同步的微妙舞蹈。

DAPA 的有效性不仅通过在多个具有挑战性的跨文化基准测试中达到最先进的性能得到证明，还在 MultiMediate'25 的多领域参与估计挑战 [45] 中获得第一名。我们的主要贡献如下：

(1) **一种新颖的领域提示机制** 通过在前面添加可学习的特定领域向量，使模型能够在联合训练期间根据数据集特定的参与模式进行条件处理，从而实现强大的跨域泛化。
(2) **一个并行交叉注意力模块** 显式建模交互动态，通过将目标参与者的反应性和预期性 BiLSTM 状态与对话伙伴的状态对齐，捕捉细微的时刻同步。
(3) **广泛的性能验证** 在多个基准测试中表现出色，例如在 NoXi-J 测试集上绝对 CCC 提高了 0.45，展示了模型的优越效果和泛化能力。

2 相关工作

多模态和跨域参与度估计。准确估计参与者参与度是理解人类互动的基础。先前的研究从各种模态和视角探讨了这个任务 [5, 18, 25, 41, 48]，但很少有人考虑文化多样性的影响 [11, 22, 47]，这在心理和行为分析中起着关键作用 [8]。Rudovic 等人 [35, 36] 进行了初步调

查，探讨参与度如何在不同的文化背景下变化，强调跨领域泛化的必要性。尽管相关领域如情感分析已经引入了跨域适应技术——例如，DiSRAN[50] 使用对抗解缠的方法，以及 Wang 等人 [44] 使用特征解耦通过线性变换的方法——这类策略在参与度估计方面仍处于探索初期。现有方法 [21, 49] 通常需要特定数据集的训练，并且跨领域泛化能力较差。为了解决这一限制，我们提出了一种新颖的域提示机制，能够在统一框架内实现动态域适应。我们的方法展示了持续的性能提升，在多个文化和语言多样性的数据集中达到了最先进的结果。

建模交互动态。沉浸是人际动态的涌现属性，但许多估计模型 [15, 16, 30, 31] 未能捕捉到这种互动本质。早期方法采用了以参与者为中心的观点，从孤立的个人特征预测沉浸 [49]，从而忽视了对话中基本的相互影响。虽然随后的方法如 DAT[21] 认识到了这一点差距并纳入了伙伴信息，但它们依赖于浅层融合技术，例如简单的交叉注意力 [40, 42]。这种方法在概念上是有限的，因为它将互动视为信号的静态聚合，而不是动态过程，未能建模复杂的非线性依赖关系和时间同步，这些特征刻画了人类交互。相比之下，我们提出的 DAPA 采用了并行注意架构，旨在进行深度关联建模，明确捕捉参与者之间的细粒度、时刻到时刻的同步性和相互依存关系。

3 方法论

3.1 任务公式化

给定一个跨文化的多领域对话语料库 D，我们将其表示为一组样本 $\{x_1, x_2, \ldots, x_D\}$。每个样本包含一段视听序列 $S = \{S_1, S_2, \ldots, S_N\}$，跨越 N 帧。每一帧 $S_i = \{S_i^a, S_i^v\}$ 包含来自两种模式的同步特征：音频 (a) 和视觉 (v)。参与度估计任务的目标是学习一个映射 $f : S_i \mapsto y_i$，其中 $y_i \in [0, 1]$ 表示帧 S_i 的连续参与度得分。相应地，模型为输入序列 S 预测一组参与度得分 $Y = \{y_1, y_2, \ldots, y_N\}$。值得注意的是，语料库涵盖了多种语言、格式和主题领域的长时间对话，这在内容和交互风格上都引入了显著的变化。

3.2 总体架构

我们提出的域自适应并行注意力框架（DAPA）旨在通过利用目标参与者（P_T）及其对话伙伴（P_P）的多模态信息来准确预测目标参与者的参与度。如图 1 所示，DAPA 管道首先通过**域提示**丰富输入特征以处理跨域变化。模型的核心由一系列相同的 DAPA 层组成，每一层都通过一种新颖的**并行交叉注意力模块**来优化参与者的表示。本模块通过使参与者之间的**反应性**和**预期的**状态对齐，明确建模了这种交互。最后，一个预测头将深度交互特征映射到一个连续的参与度得分。以下各节将详细说明该架构的每个组件。

3.3 领域自适应提示

DAPA 的初始阶段侧重于为目标和伙伴构建丰富且领域相关的输入表示。

3.3.1 **多模态特征提取**. 对于每位参与者，我们从他们的音视频数据流中提取一组全面的特征 S。在音频模态下，我们将低级声学描述符（eGeMAPS[9]）与预训练的 Whisper[33] 模型中的高级语义特征相结合。对于视觉模态，我们融合全局场景上下文（Swin Transformer[23]）与来自面部特征点（OpenFace[1]）和身体关键点（OpenPose[3]）的精细行为线索。这些多模态特征随后被投影并拼接成一个帧级特征序列 $X \in \mathbb{R}^{N \times D_m}$，用于目标参与者（$X_T$）和伙伴（$X_P$）。

3.3.2 **领域提示**. 为了弥合跨文化和跨语言数据集中存在的显著领域差异，我们引入了一种域提示机制。与其迫使模型学习领域不变的特征，这可能会丢弃有价值的特定领域的信息，我们明确地将模型条件化于数据来源。

具体来说，我们构建了一个领域提示池（DPP），表示为 $P = \{P_1, P_2, \ldots, P_k\}$，其中每个向量 $p_d \in \mathbb{R}^{N \times D_p}$ 对应训练语料库中的 K 个领域（数据集）之一。这些提示是随机初始化的，并且与网络的其余部分一起进行端到端优化。对于来自领域（数据集）d 的给定输入序列，我们将对应的提示向量 p_d 附加到目标参与者和合作伙伴的功能序列之前：

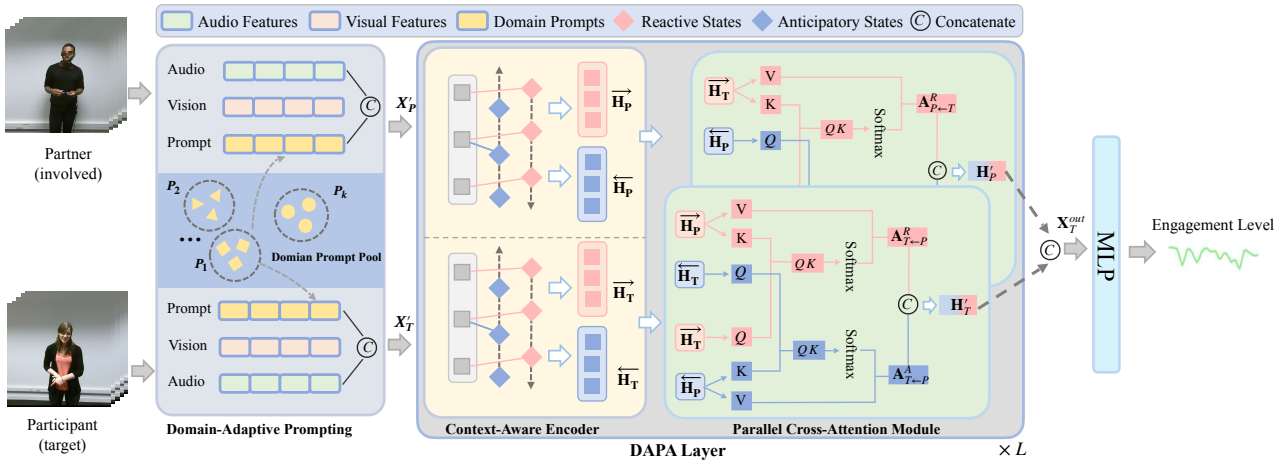

图 1: DAPA 框架概述。DAPA 引入了两项关键创新：(a) 领域提示，它将可学习的特定领域向量前置到输入特征中以实现跨域适应，以及 (b) 一个由 L 个 DAPA 层堆叠而成的结构。每一层通过使用并行交叉注意力机制来建模交互，该机制对齐了目标和其伙伴之间的反应（前向 BiLSTM）和预期（后向 BiLSTM）状态。图示说明了目标参与者注意流；对于伙伴执行对称操作。最后，将深度互动表示输入预测头以估计参与度分数。

$$X_T' = Concat(p_d, X_T), X_P' = Concat(p_d, X_P). \quad (1)$$

这一简单而有效的技术充当了明确的指令，使网络能够激活不同的路径或调整其参数以处理特定领域的参与模式，从而提升其泛化能力。

3.4 反应-预期交互建模

为了捕捉对话中的复杂动态，我们引入了一种新的交互建模方法。这种方法的核心是DAPA层，这些层被堆叠形成一个分层架构。每一层都使用并行交叉注意力模块来明确建模参与者反应状态和预期背景之间的同步。

3.4.1 **上下文感知编码器**. 每个 DAPA 层首先通过独立的 BiLSTM 编码器处理目标参与者和相关伙伴的输入特征序列，以捕捉时间依赖性。我们架构中的一个关键设计选择是将 BiLSTM 输出进行概念分解。我们没有整体使用拼接后的隐藏状态，而是将其分离以区分即时、反应性的行为与对话的整体、预期理解。具体来说，我们定义了两种不同的上下文表示。第一种是**反应状态 ($\overrightarrow{H}$)**，由前向传递的隐藏状态 $\overrightarrow{H} = \{\overrightarrow{h}_1, ..., \overrightarrow{h}_N\}$ 组成。每个状态 $\overrightarrow{h}_t$ 模型参与者对过去和当前事件的反应，因此代表他们的“即时”反应。第二种是**预见性上下文 ($\overleftarrow{H}$)**，它由后向传递的隐藏状态 $\overleftarrow{H} = \{\overleftarrow{h}_1, ..., \overleftarrow{h}_N\}$ 组成。每个状态 $\overleftarrow{h}_t$ 都受到对话整个未来的指导，从而编码一个更全局的、“预期”的视角。通过将这些编码器应用于目标和合作伙伴输入，我们获得它们各自的状态反应 $(\overrightarrow{H_T}, \overrightarrow{H_P})$ 和预期表示 $(\overleftarrow{H_T}, \overleftarrow{H_P})$，这对于后续的交互建模阶段是至关重要的。

3.4.2 **并行交叉注意力模块**. 为了建模交互的互惠性质，我们引入了**并行交叉注意力模块**模块。该模块超越了简单的单向信息融合，通过创建双向交换来同时精炼两个参与者的表示。它通过两条并行路径运行，每条路径都模拟人际同步的不同方面。第一条路径**预期上下文对齐**旨在建模高层次的情境同步和共同理解。评估每个参与者整体的、面向未来的视角如何与其他人的相一致。这是通过双向交叉注意力机制实现的。目标的预期状态 $\overleftarrow{H}_\mathrm{T}$ 查询伙伴的状态 $\overleftarrow{H}_\mathrm{P}$（作为键和值），同时，伙伴的状态 $\overleftarrow{H}_\mathrm{P}$ 查询目标的状态 $\overleftarrow{H}_\mathrm{T}$：

$$A^{A}_{\mathrm{T}\leftarrow\mathrm{P}} = \mathrm{Attention}(\overleftarrow{H_{\mathrm{T}}}, \overleftarrow{H_{\mathrm{P}}}, \overleftarrow{H_{\mathrm{P}}}) \tag{2}$$

$$A^{A}_{\mathrm{P}\leftarrow\mathrm{T}} = \mathrm{Attention}(\overleftarrow{H_{\mathrm{P}}}, \overleftarrow{H_{\mathrm{T}}}, \overleftarrow{H_{\mathrm{T}}}) \tag{3}$$

其中函数 $\mathrm{Attention}(Q, K, V)$ 是标准缩放点积注意力 [43]：

$$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^{T}}{\sqrt{d_k}}\right)V \tag{4}$$

第二个路径，**反应状态对齐**，专注于捕捉细粒度的即时行为同步，如回应手势或模仿。它使用相同的双向注意力结构将每位参与者的“当下”反应与另一方的反应状态对齐，

$$A^{R}_{\mathrm{T}\leftarrow\mathrm{P}} = \mathrm{Attention}(\overrightarrow{H_{\mathrm{T}}}, \overrightarrow{H_{\mathrm{P}}}, \overrightarrow{H_{\mathrm{P}}}) \tag{5}$$

$$A^{R}_{\mathrm{P}\leftarrow\mathrm{T}} = \mathrm{Attention}(\overrightarrow{H_{\mathrm{P}}}, \overrightarrow{H_{\mathrm{T}}}, \overrightarrow{H_{\mathrm{T}}}) \tag{6}$$

。

该模块的输出由四个交互感知表示组成，这些表示从全局和局部两个角度捕捉相互影响。为了准备后续 DAPA 层的输入，我们通过连接各自的反应性和预期对齐向量来更新每个参与者的表示。对于目标参与者（T），这一新的表示 H'_{T} 形式为：

$$H'_{\mathrm{T}} = \mathrm{Concat}\big(A^{R}_{\mathrm{T}\leftarrow\mathrm{P}}, A^{A}_{\mathrm{T}\leftarrow\mathrm{P}}\big) \tag{7}$$

类似地，伙伴的表示（P）被更新为 H'_{P}：

$$H'_{\mathrm{P}} = \mathrm{Concat}\big(A^{R}_{\mathrm{P}\leftarrow\mathrm{T}}, A^{A}_{\mathrm{P}\leftarrow\mathrm{T}}\big) \tag{8}$$

这些融合的、感知交互的表示形式 H'_{T} 和 H'_{P} 随后被输入到下一个 DAPA 层，实现了关系动力学的分层细化。

3.5 参与预测

经过所有 LDAPA 层处理后，我们分别获得目标和伙伴的最终交互感知表示 H'_{T} 和 H'_{P}（公式 7, 8）。为了预测目标的参与度，我们首先将这些表示连接起来形成一个二元状态向量，

$$X^{\mathrm{out}}_{\mathrm{T}} = \mathrm{Concat}\big(H'_{\mathrm{T}}, H'_{\mathrm{P}}\big), \tag{9}$$

该向量捕捉了完整的交互上下文。一个多层感知器（MLP）[37] 使用 Sigmoid 激活函数将这种联合表示映射到最终预测的参与序列：

$$\hat{Y}_T = \mathrm{Sigmoid}\left(\mathrm{MLP}(X^{\mathrm{out}}_T)\right). \tag{10}$$

整个 DAPA 模型是通过最小化以下损失函数进行端到端训练的：

$$\mathcal{L} = 1 - \mathrm{CCC}(\hat{Y}_T, Y_T), \tag{11}$$

其中 Y_T 是真实参与序列。此目标直接将训练过程与我们的主要评估指标———一致相关系数（CCC）对齐。

4 实验

4.1 数据集

挑战包含四个不同的数据集，最终的比赛指标是这四个数据集的平均 CCC 分数。

NoXi 基础 [2] 数据集包含 48 个训练环节和 16 个专家与新手之间的一对一视频互动测试环节，这些环节分别用英语、法语和德语进行。每个环节包括以 25 帧/秒（某些特征采样率为 40 帧/秒）录制的同步音频和视频流，总共有 84 段对话涉及 87 名参与者（女性 26 人，男性 61 人），以及大约 25 小时的音频。为每一帧提供的 [0,1] 范围内的逐帧参与度评分由 2 至 7 名标注员评定（平均 3.6 名）。对于 2025 年多领域参与度评估挑战赛，既提供了预先提取的特征也提供了原始视频文件。

诺西附加 [2, 26] 语言数据集 (NoXi-Add) 仅包含 12 个测试会话。为了评估核心 NoXi 语言之外的跨文化泛化，该设置引入了四种新语言：阿拉伯语、意大利语、印度尼西亚语和西班牙语，同时保持与 NoXi Base 相同的功能格式。

NoXi 日本 [10] 数据集 (NoXi-J) 包含了 10 次专家与新手的屏幕中介互动测试会话，使用日语和普通话进行记录，这些记录在日本和中国根据原始的 NoXi 协议完成。这将覆盖范围扩展到欧洲八种核心 NoXi 语言之外的东亚语言。

MPIIGroupInteraction[28, 29] 数据集 (MPIIGI) 包含在办公室环境中使用 8 个摄像头（4DV）阵列和四

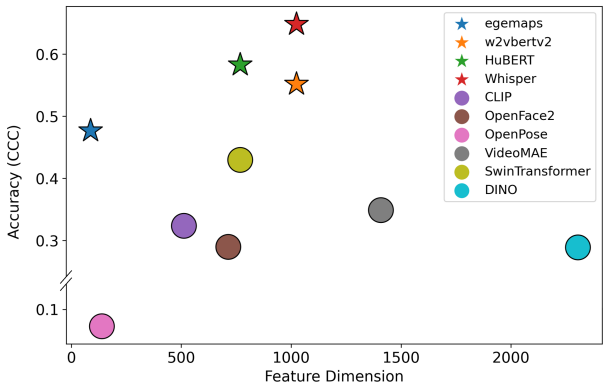

图 2: 评估关键特征对 NoXi-J 数据集上的参与度估计性能的影响。

个环境麦克风录制的 6 个测试环节。每个 20 分钟的环节记录了四名参与者用德语讨论一个自选的争议性话题。与 NoXi 中大多数站立互动不同，MPIIGI 中的参与者大多坐着；特征采样率与 NoXi 相同。

4.2 实验设置

我们使用了官方预提取的特征，共计 1709 维：88-D eGeMAPSv2 音频、768-D Swin Transformer 视觉、714-D OpenFace 面部和 139-D OpenPose 姿态特征。此外，我们还提取了 1280-D Whisper 音频嵌入，总共达到了 2989 维。

在所有实验中，我们将随机种子固定为 40，并使用初始学习率为 5×10^{-5} 的 Adam 优化器。我们在前 400 步线性预热到 $5{\times}10^{-5}$，然后应用余弦退火至 $T_{\max} = 10$。训练和验证批次大小分别设置为 32 和 256，共进行 40 个周期。遵循 DAT 中的上下文分割方案，总序列长度为 96 帧，包括一个 32 帧的核心段加上两个 32 帧的辅助段，并采用滑动步幅 $s = 32$。

在模型中，每个输入特征向量首先通过两个全连接层映射到一个 512 维的表示，然后被送入一个三层的 BiLSTM [13, 38] 编码器，其隐藏层大小为 512。为了缓解过拟合并提高泛化能力，我们在所有层中应用了 0.1 的 dropout 率，并在整个 40 个 epoch 的训练过程中使用指数移动平均（EMA）[32]。所有的训练和评估都使用 CCC 作为标准。

4.3 消融研究

为了验证我们提出的模型的有效性，我们在 NoXi-J 和 NoXi-Base 验证集上进行了消融实验，重点关注两个关键组件：领域自适应提示（DAP）和平行交叉注意力（PCA）。我们也评估了每个输入特征的贡献。没有 DAP 和 PCA 的基础模型在表 1 的第一行中进行了描述。

表 1: 域自适应提示（DAP）和并行交叉注意力（PCA）配置的消融研究结果。

数据访问权限	主成分分析	无序杰	诺西-基础
		0.669	0.822
☒		0.705	0.837
	☒	0.684	0.836
☒	☒	0.722	0.855

4.3.1 **特征选择的消融研究.** 为了评估不同特征组合的性能，我们使用了所提出的 DAPA 框架彻底调查了每个组件的贡献。结果如图 2 所示，指导了我们的最终特征选择。对于音频特征，实验表明 Whisper 特征显著优于 Hubert [14] 和 W2v-BERTv2[6]。为了捕捉音频中的情感线索，我们还利用了 eGeMAPS 特征。对于视频特征，SwinTransformer 被证明是最有效的。此外，为了结合显式的面部和姿态信息，我们分别使用了从 OpenFace 和 OpenPose 获得的面部标志点和身体关键点。最终，这一系统的评估使我们选择了一种五项特征组合，在 DAPA 模型中实现了最优性能。

4.3.2 **域自适应提示的消融研究.** 为了展示领域自适应提示（DAP）的效果，我们将有无此机制的模型性能进行了对比。如表 1 所示，DAP 在两个数据集上均持续改善了结果。在跨领域的 NoXi-J 测试集中，它将 CCC 从 0.669 提升到 0.705（显著提升了 0.036），突显了领域条件化在缓解领域偏移方面的有效性。即使是在文化混合的 NoXi-Base 上，DAP 也取得了较小但有意义的提升，为 0.015，证明其能够保持领域的特定提示。这些结果证实了 DAP 在改善多样化的现实条件下的一般性方面发挥的关键作用。

表 2: 各个测试数据集上的最终 CCC 得分。全局代表所有数据集的全球平均 CCC 分数。

Team	NoXi Base	NoXi-Add	MPIIGI	NoXi-J	Global
HFUT-LMC（ours）	**0.79**	**0.75**	**0.67**	**0.58**	**0.70**
USTC-IAT United	0.79	0.73	0.66	0.53	0.68
chfighting	0.77	0.70	0.65	0.50	0.65
nnaedu	0.75	0.68	0.57	0.58	0.64
lasii	0.79	0.73	0.54	0.51	0.64
yueangh	0.75	0.67	0.59	0.42	0.61
yueyue	0.65	0.67	0.61	0.34	0.57
MM25 Baseline [2]	0.57	0.47	0.44	0.13	0.40
Behavioural-AI Lab	0.53	0.41	0.09	0.26	0.32

4.3.3 **平行交叉注意力模块的消融研究**. 我们的并行交叉注意力（PCA）模块旨在捕捉人际互动丰富的通用动态。为了验证其贡献，我们将基线 BiLSTM 与增强了 PCA 的模型进行比较。表 1 中的结果表明，PCA 提供了显著的性能提升，在 NoXi-J 上将 CCC 提高了 0.015，在 NoXi-Base 上提高了 0.014。这表明，对反应性和预期状态之间的同步性进行建模是理解参与度的一种根本上稳健的方法。

此外，PCA 与 DAP 之间的协同作用尤为显著。在配备 DAP 的模型中添加 PCA（第 4 行对比第 2 行）在 NoXi-J 上获得了 0.017 的 CCC 增益，而基于 PCA 引入 DAP（第 4 行对比第 3 行）则带来了更大的提升，达到 0.038。这种不对称性表明 PCA 提供了一种交互感知特征基础，增强了特定领域提示的有效性。它们互补的作用凸显了结合两个模块以实现 DAPA 最佳性能的必要性。

4.4 最终结果

最后，我们将优化配置的框架应用于所有四个数据集的测试集，并将其与官方基线进行了比较。如表 2 所示，我们的模型在 NoXi Base 上获得了 0.79 的 CCC 分数，在 NoXi-Add 上为 0.75，在 NoXi-J 上为 0.58，在 MPIIGI 上为 0.67，达到了新的最先进水平。这不仅展示了鲁棒性，还强调了强大的跨文化和跨领域泛化能力。总体而言，我们的方法在全局平均指标上比官方基线 [2] 高出 0.30 CCC，进一步验证了其有效性。

4.5 可视化

图 3 可视化了我们的 DAPA 模型在 NoXi-Base（顶部行）和 NoXi-J（底部行）验证集上的预测与真实值对比。一个关键的观察结果是两个数据集中真实值特征之间的显著差异。NoXi-Base 标签表现出高度动态且连续的波动，这是细粒度参与注释的典型特征。相比之下，NoXi-J 标签遵循一种明显的阶梯状（量化）模式，其特点是持续水平之间的突然转换。尽管这些基本不同的注释风格存在，我们的模型预测值（橙色）在两种情况下都能始终如一且准确地跟踪真实值（蓝色）。这强烈证明了该模型的鲁棒性及其不仅跨越不同对象，还能跨越变化的数据分布和注释方法进行泛化的性能。

5 结论

本文介绍了 DAPA，这是一个显著推动跨文化对话参与度估计前沿的新型框架。通过结合域提示机制以实现鲁棒的跨领域适应，并与操作于学习到的反应性和预期状态上的并行交叉注意力模块协同工作，DAPA 有效地捕捉了参与度复杂的互动性质。我们的模型优越性通过在具有挑战性的基准测试中建立新的最先进结果以及在 MultiMediate'25 参与度估计挑战赛中获得第一名的位置得到了证明。因此，这项工作为多模态交互分析提供了一个强大且可泛化的方案。

展望未来，我们承认某些局限性为未来的探索提供了明确的路径。我们的当前域提示机制虽然有效，但需要重新训练以适应新的、未见过的领域。此外，我们的模型对音频-视频特征的依赖提供了一个重大机会：语言内容的整合。我们的初步实验揭示了稀疏文本表达与密集帧级参与注释之间时间对齐的基本挑战。我们认为开发复杂模型来弥合这种模态差距，如动态注意力架构或异步多模态变换器，是关键的下一步。成功应对这些挑战将是实现更全面理解参与的关键，从而为更具上下文感知能力和文化适应性的计算机交互系统铺平道路。

Acknowledgments

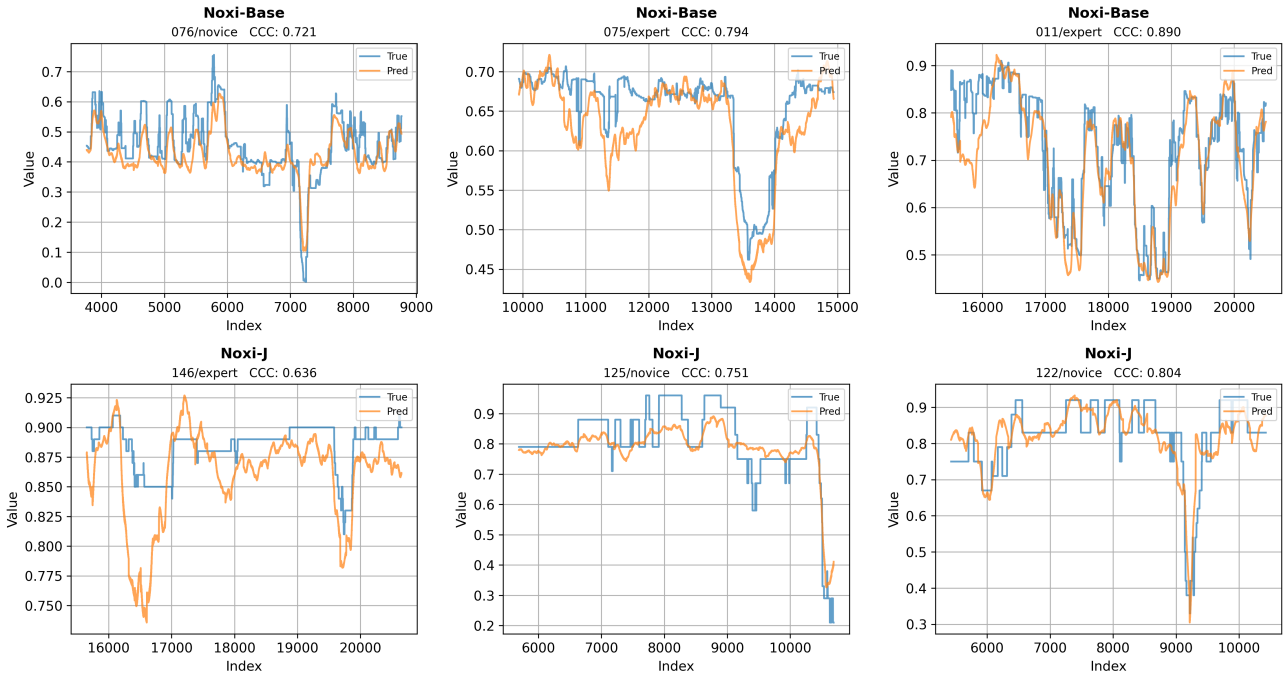

图 3: DAPA 模型在 NoXi-Base 和 NoXi-J 数据集上的实时拟合结果，固定区间长度为 5000 个样本。True 表示真实值。

References

[1] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. 2016. Openface: an open source facial behavior analysis toolkit. In *2016 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 1–10.

[2] Angelo Cafaro, Johannes Wagner, Tobias Baur, Soumia Dermouche, Mercedes Torres Torres, Catherine Pelachaud, Elisabeth André, and Michel Valstar. 2017. The NoXi Database: Multimodal Recordings of Mediated Novice-Expert Interactions. In *Proceedings of 19th ACM International Conference on Multimodal Interaction*. 350 – 359. doi:10.1145/3136755.3136780

[3] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2017. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7291–7299.

[4] Haifeng Chen, Yifan Deng, and Dongmei Jiang. 2021. Temporal attentive adversarial domain adaption for cross cultural affect recognition. In *Companion publication of the 2021 international conference on multimodal interaction*. 97–103.

[5] Xu Chen, Li Niu, Ashok Veeraraghavan, and Ashutosh Sabharwal. 2019. FaceEngage: Robust estimation of gameplay engagement from user-contributed (YouTube) videos. *IEEE Transactions on Affective Computing* 13, 2 (2019), 651–665.

[6] Yu-An Chung, Yu Zhang, Wei Han, Chung-Cheng Chiu, James Qin, Ruoming Pang, and Yonghui Wu. 2021. W2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 244–250.

[7] Soumia Dermouche and Catherine Pelachaud. 2019. Engagement modeling in dyadic interaction. In *2019 international conference on multimodal interaction*. 440–445.

[8] Hillary Anger Elfenbein and Nalini Ambady. 2002. On the universality and cultural specificity of emotion recognition: a meta-analysis. *Psychological bulletin* 128, 2 (2002), 203.

[9] Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*. 1459–1462.

[10] Marius Funk, Shogo Okada, and Elisabeth André. 2024. Multilingual dyadic interaction corpus noxi+ j: Toward understanding asian-european non-verbal cultural characteristics and their influences on engagement. In *Proceedings of the 26th International Conference on Multimodal Interaction*. 224–233.

[11] Yuefang Gao, Yiteng Cai, Xuanming Bi, Bizheng Li, Shunpeng Li, and Weiping Zheng. 2023. Cross-domain facial expression recognition through reliable global–local representation learning and dynamic label weighting. *Electronics* 12, 21 (2023), 4553.

[12] Guendalina Graffigna, Serena Barello, Andrea Bonanomi, and Edoardo Lozza. 2015. Measuring patient engagement: development and psychometric properties of the Patient Health Engagement (PHE) Scale. *Frontiers in psychology* 6 (2015), 274.

[13] Alex Graves. 2012. Long short-term memory. *Supervised sequence labelling with recurrent neural networks* (2012), 37–45.

[14] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden

units. *IEEE/ACM transactions on audio, speech, and language processing* 29 (2021), 3451–3460.

[15] Ryo Ishii and Yukiko I Nakano. 2008. Estimating user's conversational engagement based on gaze behaviors. In *International Workshop on Intelligent Virtual Agents*. Springer, 200–207.

[16] Ryo Ishii, Yukiko I Nakano, and Toyoaki Nishida. 2013. Gaze awareness in conversational agents: Estimating a user's conversational engagement from eye gaze. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 3, 2 (2013), 1–25.

[17] Ji-Youn Jung, Sihang Qiu, Alessandro Bozzon, and Ujwal Gadiraju. 2022. Great chain of agents: The role of metaphorical representation of agents in conversational crowdsourcing. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–22.

[18] Shofiyati Nur Karimah and Shinobu Hasegawa. 2022. Automatic engagement estimation in smart education/learning settings: a systematic review of engagement definitions, datasets, and methods. *Smart Learning Environments* 9, 1 (2022), 31.

[19] Atesh Koul, Davide Ahmar, Gian Domenico Iannetti, and Giacomo Novembre. 2023. Spontaneous dyadic behavior predicts the emergence of interpersonal neural synchrony. *NeuroImage* 277 (2023), 120233.

[20] Deepak Kumar, Surbhi Madan, Pradeep Singh, Abhinav Dhall, and Balasubramanian Raman. 2024. Towards engagement prediction: a cross-modality dual-pipeline approach using visual and audio features. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 11383–11389.

[21] Jia Li, Yangchen Yu, Yin Chen, Yu Zhang, Peng Jia, Yunbo Xu, Ziqiang Li, Meng Wang, and Richang Hong. 2024. DAT: Dialogue-Aware Transformer with Modality-Group Fusion for Human Engagement Estimation. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 11397–11403.

[22] Tong Liu, Jing Li, Jia Wu, Lefei Zhang, Shanshan Zhao, Jun Chang, and Jun Wan. 2023. Cross-Domain Facial Expression Recognition via Disentangling Identity Representation.. In *IJCAI*. 1213–1221.

[23] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

[24] Kshitij Mishra, Mauajama Firdaus, and Asif Ekbal. 2023. Predicting Politeness Variations in Goal-Oriented Conversations. *IEEE Transactions on Computational Social Systems* 10, 3 (2023), 1095–1104. doi:10.1109/TCSS.2022.3156580

[25] Hamed Monkaresi, Nigel Bosch, Rafael A Calvo, and Sidney K D'Mello. 2016. Automated detection of engagement using video-based estimation of facial expressions and heart rate. *IEEE Transactions on Affective Computing* 8, 1 (2016), 15–28.

[26] Philipp Müller, Michal Balazia, Tobias Baur, Michael Dietz, Alexander Heimerl, Anna Penzkofer, Dominik Schiller, François Brémond, Jan Alexandersson, Elisabeth André, et al. 2024. MultiMediate'24: Multi-Domain Engagement Estimation. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 11377–11382.

[27] Philipp Müller, Michael Dietz, Dominik Schiller, Dominike Thomas, Hali Lindsay, Patrick Gebhard, Elisabeth André, and Andreas Bulling. 2022. MultiMediate'22: Backchannel Detection and Agreement Estimation in Group Interactions. In *Proceedings of the 30th ACM International Conference on Multimedia*. 7109–7114.

[28] Philipp Müller, Michael Xuelin Huang, and Andreas Bulling. 2018. Detecting low rapport during natural interactions in small groups from non-verbal behaviour. In *Proceedings of the 23rd International Conference on Intelligent User Interfaces*. 153–164.

[29] Philipp Müller, Dominik Schiller, Dominike Thomas, Guanhua Zhang, Michael Dietz, Patrick Gebhard, Elisabeth André, and Andreas Bulling. 2021. MultiMediate: Multi-modal Group Behaviour Analysis for Artificial Mediation. In *Proc. ACM Multimedia (MM)*. 4878–4882. doi:10.1145/3474085.3479219

[30] Ryota Ooko, Ryo Ishii, and Yukiko I Nakano. 2011. Estimating a user's conversational engagement based on head pose information. In *International Workshop on Intelligent Virtual Agents*. Springer, 262–268.

[31] Arthur Pellet-Rostaing, Roxane Bertrand, Auriane Boudin, Stéphane Rauzy, and Philippe Blache. 2023. A multimodal approach for modeling engagement in conversation. *Frontiers in Computer Science* 5 (2023), 1062342.

[32] Boris T Polyak and Anatoli B Juditsky. 1992. Acceleration of stochastic approximation by averaging. *SIAM journal on control and optimization* 30, 4 (1992), 838–855.

[33] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*. PMLR, 28492–28518.

[34] Andrew Reece, Gus Cooney, Peter Bull, Christine Chung, Bryn Dawson, Casey Fitzpatrick, Tamara Glazer, Dean Knox, Alex Liebscher, and Sebastian Marin. 2023. The CANDOR corpus: Insights from a large multimodal dataset of naturalistic conversation. *Science Advances* 9, 13 (2023), eadf3197.

[35] Ognjen Rudovic, Jaeryoung Lee, Lea Mascarell-Maricic, Björn W Schuller, and Rosalind W Picard. 2017. Measuring engagement in robot-assisted autism therapy: A cross-cultural study. *Frontiers in Robotics and AI* 4 (2017), 36.

[36] Ognjen Rudovic, Yuria Utsumi, Jaeryoung Lee, Javier Hernandez, Eduardo Castelló Ferrer, Björn Schuller, and Rosalind W Picard. 2018. CultureNet: a deep learning approach for engagement intensity estimation from face images of children with autism. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 339–346.

[37] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. 1986. Learning representations by back-propagating errors. *nature* 323, 6088 (1986), 533–536.

[38] Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing* 45, 11 (1997), 2673–2681.

[39] Candace L Sidner and Myrosia Dzikovska. 2002. Human-robot interaction: Engagement between humans and robots for hosting activities. In *Proceedings. fourth ieee international conference on multimodal interfaces*. IEEE, 123–128.

[40] Daniel Stamate, Pradyumna Davuloori, Doina Logofatu, Evelyne Mercure, Caspar Addyman, and Mark Tomlinson. 2024. Ensembles of Bidirectional LSTM and GRU Neural Nets for Predicting Mother-Infant Synchrony in Videos. In *International Conference on Engineering Applications of Neural Networks*. Springer, 329–342.

[41] Lukas Stappen, Alice Baird, Michelle Lienhart, Annalena Bätz, and Björn Schuller. 2022. An estimation of online video user engagement from features of time-and value-continuous, dimensional emotions. *Frontiers in Computer*

Science 4 (2022), 773154.

[42] Wang Tang, Fethiye Irmak Dogan, Linbo Qing, and Hatice Gunes. 2025. AsyReC: A Multimodal Graph-based Framework for Spatio-Temporal Asymmetric Dyadic Relationship Classification. *arXiv preprint arXiv:2504.05030* (2025).

[43] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[44] Qianlong Wang, Zhiyuan Wen, Keyang Ding, Bin Liang, and Ruifeng Xu. 2025. Cross-Domain Sentiment Analysis via Disentangled Representation and Prototypical Learning. *IEEE Transactions on Affective Computing* 16, 1 (2025), 264–276. doi:10.1109/TAFFC.2024.3431946

[45] Daksitha Withanage Don, Marius Funk, Michal Balazia, Huajian Qiu, Shogo Okada, François Brémond, Jan Alexandersson, Andreas Bulling, Elisabeth André, and Philipp Müller. 2025. MultiMediate '25: Cross-cultural Multi-domain Engagement Estimation. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25)* (Dublin, Ireland) *(MM '25)*. Association for Computing Machinery. doi:10.1145/3746027.3762076

[46] Sophie Wohltjen and Thalia Wheatley. 2024. Interpersonal eye-tracking reveals the dynamics of interacting minds. *Frontiers in human neuroscience* 18 (2024), 1356680.

[47] Jinbao Xie, Yulong Wang, Tianxin Meng, Jianqiao Tai, Yueqian Zheng, and Yury I Varatnitski. 2025. Multimodal Emotion Recognition Method Based on Domain Generalization and Graph Neural Networks. *Electronics* 14, 5 (2025), 885.

[48] Chen Yu, Paul M Aoki, and Allison Woodruff. 2004. Detecting user engagement in everyday conversations. *arXiv preprint cs/0410027* (2004).

[49] Jun Yu, Keda Lu, Mohan Jing, Ziqi Liang, Bingyuan Zhang, Jianqing Sun, and Jiaen Liang. 2023. Sliding window seq2seq modeling for engagement estimation. In *Proceedings of the 31st ACM International Conference on Multimedia*. 9496–9500.

[50] Yuhao Zhang, Ying Zhang, Wenya Guo, Xiangrui Cai, and Xiaojie Yuan. 2023. Learning Disentangled Representation for Multimodal Cross-Domain Sentiment Analysis. *IEEE Transactions on Neural Networks and Learning Systems* 34, 10 (2023), 7956–7966. doi:10.1109/TNNLS.2022.3147546