

PicPersona-TOD: 一种用于在任务导向对话中通过图像个性化话语风格的数据集

Jihyun Lee¹, Yejin Jeon¹, Seungyeon Seo¹, Gary Geunbae Lee^{1,2}

¹Graduate School of Artificial Intelligence, POSTECH, Republic of Korea

²Department of Computer Science and Engineering, POSTECH, Republic of Korea
{ jihyunlee, jeonyj0612, ssy319, gblee } @postech.ac.kr

Abstract

任务导向对话 (TOD) 系统旨在通过自然语言交互来满足用户请求, 但现有系统往往生成通用、单调的响应, 缺乏个性, 无法适应用户的个人属性。为了解决这一问题, 我们引入了 PicPersona-TOD, 这是一个新颖的数据集, 将用户图像作为个性的一部分, 能够生成基于用户特定因素 (如年龄或情感背景) 的个性化响应。这通过第一印象、对话策略指导提示以及使用外部知识减少幻觉来实现。人类评估证实, 我们的数据集增强了用户体验, 个性化的响应使互动更加引人入胜。此外, 我们引入了一个新的自然语言生成模型 Pictor, 它不仅能够个性化响应, 还展示出在未见领域中的强大性能。¹

1 介绍

任务导向对话 (TOD) 是对话系统的核心任务之一, 其设计目的是为了满足用户请求, 例如在客户服务台 (Rastogi et al., 2020) 和游客中心 (Zang et al., 2020) 协助用户。一个 TOD 系统通常分为以下子模块: (1) 对话状态追踪 (DST), 用于追踪用户的请求, (2) 策略模块, 用于确定系统的操作, 如数据库 (DB) 搜索或对话终止, 以及 (3) 自然语言生成模块 (NLG), 将对话策略和 DB 结果转换为自然语言响应 (Young et al., 2013)。在这些组件中, NLG 模块生成的响应用于直接与用户交互; 因此, NLG 响应对整体用户体验有显著影响。

尽管大量研究已经改善了系统的响应 (Peng et al., 2020; Lin et al., 2020; Hosseini-Asl et al., 2020; Su et al., 2021; Yang et al., 2021; Yu et al., 2022; Ohashi and Higashinaka, 2023), 但主要关注点仍然是提高信息传递的准确性。因此, 生成的响应风格是单调的, 缺乏个性, 这妨碍了系统与用户形成符合年龄和情感共鸣的连接 (McLean et al., 2021)。

为改善系统回复的自然性和参与度, 最近的方法通过将用户角色融入对话, 策划了新的面向任务的对话数据集, 以支持个性化的回复风

格。例如, Joshi et al. (2017) 在对话中加入了年龄和性别信息, Lin et al. (2023) 融入了情感元素, Liu et al. (2024) 通过模仿用户的名词和动词短语来个性化系统回复。尽管这些方法在某种程度上提供了个性化的回复, 但角色模式仅限于文本信息, 这缺乏与用户互动过程中的细节和并发性。

同时, 在开放对话系统领域, 用户角色的整合一直是一个长期关注的重点 (Zhang et al., 2018a; Agrawal et al., 2023; Kim et al., 2024; Qian et al., 2017; Zheng et al., 2019; Shen et al., 2024), 最近的进展突出了具前瞻性角色方法的潜力 (Ahn et al., 2023; Lee et al., 2024b)。具前瞻性角色可以捕捉细微的面部表情并提供对用户背景的丰富理解, 这类似于人们在现实互动中解读视觉和非语言线索的方式。这种方法在没有预先文本用户简介的情况下尤其有价值, 比如在初次互动时没有历史信息存在的情况下。尽管具有这些优势, 具前瞻性角色主要被应用于闲聊 (Poria et al., 2018; Firdaus et al., 2020) 或咨询场景 (Valstar et al., 2016; Li et al., 2023b), 并且尚未在当前的 TOD 系统中进行探索。

考虑到这些因素, 我们引入了一个新的 TOD 数据集, 该数据集将真实的用户图像作为用户角色的一部分, 能够在问候、正式程度、年龄敏感性和情感意识方面实现个性化系统响应 (图 ??)。在构建 PicPersona-TOD 时, 我们使用用户的第一印象和对话政策引导提示, 成功地提炼了大型语言模型 (LLM) 的个性化能力。此外, 我们结合来自 Google Maps 和 Wikipedia 的外部知识, 以减少个性化回应中的幻觉。此外, 我们实施了一项精细的过滤流程, 以确保风格的适当性、语义的准确性和整体的自然性, 最终形成一个精致的个性化 TOD 数据集。尽管过程高度自动化, 我们的数据集在用户体验 (§ ??) 和个性化 (§ 3.4) 方面相比其他数据集和方法表现出更高的人类偏好。从标签对齐的角度来看, 与 DST 和政策模块的分析表明 PicPersona-TOD 保持信息的准确性。具体来说, 我们提出了一种名为 Pictor 的 NLG 模型,

¹<https://github.com/JihyunLee1/PicPersona>

它展示了在个性化中生成鲁棒性的能力，即使是在未见过的领域。

总之，这项工作的贡献有三个方面：首先，我们引入了 PicPersona-TOD，这是一个将用户图像整合到人物设定中并提供个性化系统响应的新型 TOD 数据集。其次，我们提出了一个高度自动化的数据集生成框架，该框架能够高效地创建逼真且个性化的数据集。第三，通过人工评估，我们证明了我们的数据集通过个性化增强了用户体验，基准测试结果确认个性化并未在其他关键任务中影响性能。

2 PicPersona-TOD 数据集

在本节中，我们介绍了 PicPersona-TOD，这是第一个基于用户图像个性化的 TOD 数据集。为了构建一个高质量的个性化 TOD 数据集，我们假设它应满足三个标准：（1）用户的话语应与其图像保持一致，（2）系统的响应应适当地个性化到用户图像，以及（3）合成的数据集应与 TOD 任务的子任务标签（如 DST 和对话策略预测）对齐，同时保留信息。为了解决这些标准，我们的数据集构建流程包括五个关键阶段：（1）用户图像收集和对话数据集扩展，（2）用户图像与话语的对齐，（3）用户话语风格迁移，（4）系统响应个性化，以及（5）数据过滤。对于数据构建，我们主要使用 GPT-4o (Achiam et al., 2023) 作为主要的语言模型。整个过程如图 ?? 所示，所使用的提示在附录 ?? 中。

2.1 收集图像和扩展对话

首先，我们选择能传达足够个性信息的合适用户图像。为了有效地代表用户个性，每张图像应为单个人物，其位于图像中心，并且足够靠近，以使面部细节和衣着信息清晰可见。根据这些标准，我们选择了 Flickr-Faces-HQ (Karras et al., 2019) 作为图像来源，并确保排除幼儿，因为他们太小，无法参与 TOD 互动。在收集数据后，我们使用 LLM 为每张图像提取额外的元数据，包括估计年龄、性别和正式程度。

对于对话数据集，我们结合了 MultiWOZ-2.2 (Zang et al., 2020) 和 SGD (Rastogi et al., 2020) 数据集，这些数据集包括 8,438 和 11,398 个对话，总共涉及 18 个服务领域²。此外，由于数据集中的电影、餐厅、酒店和景点在现实世界中存在，我们通过收集每个地点的 Google Maps 评论和维基百科条目扩展了该数据集。具体来说，我们从 342 个维基百科条目中添加了 2,474 句子，并从 406 个 Google Maps 上的

位置中添加了 3,483 条评论。将这些结果添加到数据库中以减少个性化中的幻觉现象。

2.2 图像与对话数据对齐

在选择图像和对话数据集后，我们进行了用户图像和话语对齐。由于对话数据集缺少年龄或性别等细节，我们选择情感作为共同属性，因为情感在图像和对话中都一致存在。我们使用一个经过微调的情感分类模型³来对对话中的情感进行分类，并提示一个 LLM 对图像进行分类（积极、中性、消极），如果它们具有相同的预测情感标签，就创建图像-对话对。整个数据集的情感分布为积极占 50.92%，中性占 52.44%，消极占 0.55%。

2.3 对齐用户话语风格与图像

接下来，我们调整了用户话语的风格，以更紧密地与相应的用户图像一致。我们通过提示用户图像进行风格迁移，考虑到年龄、性别、情感和上下文线索等因素。形式化地说，对于每个 i -th 对话 $D_i = (u_0, s_0, u_1, s_1 \dots u_T, s_T)$ 及其相关图像 Img_i ，我们在轮次 t 生成修订后的话语 $\tilde{u}_t$ ； $\tilde{u}_t = \text{LLM}(s_{t-1}, u_t, \text{Img}_i)$ ，其中 T 是轮次总数， u 是对话中的用户话语， s 是系统话语。

2.4 系统响应风格个性化

在使系统的响应个性化时，我们将过程分为三种类型，并使用第一印象和对话策略来指导提示。“基本个性化”应用于大多数情况，而“问候个性化”用于涉及问候的对话动作。对于与推荐相关的动作，我们实施了“推荐个性化”，以提供更少幻觉的建议。

基础个性化为了生成与用户形象相一致的个性化系统话语，我们使用第一印象作为提示的指南。此过程借鉴了人类在交流中的认知机制，该机制包括两个关键步骤。首先，人类会在毫秒内不自觉地推测出他人的第一印象 (Borkenau et al., 2009; Willis and Todorov, 2006)，然后调整他们的交流语气和风格以符合这种印象，将其视作推测的个性 (Rule and Ambady, 2008)。类似于这一过程，我们首先从用户的形象 (Img_i) 产生一个印象，然后根据推测的个性在形式、年龄敏感性和情感背景方面生成个性化的系统话语。具体来说，对于给定的对话 D_i ， Img_i 和推测的 Imp_i ，个性化的系统话语 $\tilde{s}_t$ 产生如下； $\tilde{s}_t = \text{LLM}(s_t, \tilde{u}_t, \text{Img}_i, \text{Imp}_i)$ 。问候个性化由于问候和结束语在创造讲话者之间的个人互动中起着至关重要的作用 (McLean et al., 2021; Glas et al., 2017)，我们专门定制了

²在泛化测试中，我们排除了公交、家庭和电影领域。

³来自 Hugging Face 的微调模型，[cardiffnlp/twitter-roberta-base-sentiment-latest](#)

系统的问候语和告别语，以提供更具吸引力和个性化的体验。这是通过提示 LLM 加入关于用户外观的具体评论来实现的，例如提到他们服装的独特特征（例如，“漂亮的红帽子！”或“恭喜你取得的成就”）（图 ??，右上角）。推荐系统个性化 TOD 系统经常做出推荐，例如“*How about [location]?*”。在我们的初步实验中，我们观察到当模型尝试个性化这些推荐时，有时会引入臆想的信息（例如，“*The [location] currently has a festival you might like*”）。为了减轻这些臆想，我们通过从在线来源收集真实信息来增强提示，从而进行了检索增强生成 (Lewis et al., 2020)。具体而言，我们从数据库 (DB) 中的关于 [location] 的评论 (§ 2.1) 中检索出与 $\tilde{u}_i$ 余弦相似度最高的三个评论，通过使用 Sentence-BERT (Reimers, 2019) 对其进行嵌入。这些评论被整合到提示中，以指导模型生成以事实为基础的反应（图 ??，底部右侧）。在这些过程之后，我们得到了个性化的 TOD 对话 $\tilde{D}_i = \{\tilde{s}_{0:T}, \tilde{u}_{0:T}, \text{Img}_i, \text{Imp}_i\}$ 。

2.5 通过过滤进行数据集质量控制

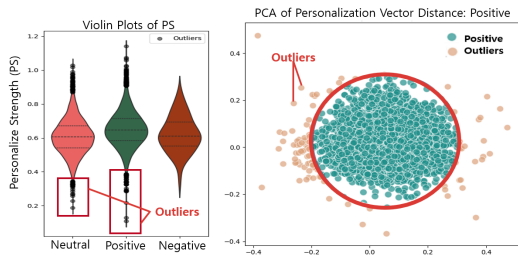


Figure 1: 个性化强度和个性化方向过滤过程的可视化分别显示在左侧和右侧。

尽管在前面的子节中进行了细致的构建过程，一些对话仍可能产生不恰当的语句风格，包含语义不一致，或者缺乏整体的自然性。为了解决这些潜在问题并提高数据集的整体质量，我们实施了几种过滤过程。

如果个性化程度过低（图 1，左侧），则过滤掉这些对话。为此，我们通过定义个性化强度 (PS_i) 为原始系统响应 $E(s_t)$ 和个性化响应 $E(\tilde{s}_t)$ 之间的平均距离 $PS_i = \frac{1}{T} \sum_{t=0}^T \text{Dist}(E(\tilde{s}_t), E(s_t))$ 来计算每个对话 i 的个性化强度，这里 Dist 表示欧几里得距离， E 表示用 Sentence-BERT 进行嵌入。接下来，我们收集每个元数据类别（例如，年轻、年老）的 PS 值，并删除低于阈值的对话，该阈值定义为低于 $2.5 \times \text{IQR}$ （四分位距）。结果，数据集中有 1.49 % 被过滤掉。

风格方向过滤我们移除在相同元数据类别中个

性化方向不同的异常值（图 1，右）。我们为每个对话计算个性化向量 (PV)，如 i 所示。然后，通过对该类别内的 PV 向量求平均值，计算每个元数据类别的平均个性化向量， PV_{class} 。为了检测异常值，我们计算类别均值与每个对话的个性化向量之间的距离 (PD_i): $PD_i = \text{Dist}(PV_{class}, PV_i)$ 。我们将离类别风格向量的均值距离异常大的定义为异常值。我们将阈值设定为 $4.5 \times \text{IQR}$ ，从而移除 1.98 % 的数据集。语义过滤我们通过将用户和系统发言与相应的 DST 和对话策略标签进行比较，过滤掉语义上不一致的用户和系统发言。对于用户发言，我们检查其与 DST 标签的对齐情况。例如，如果标签是 *hotel-east*, *restaurant-expensive*，那么用户的发言应反映这一点，比如说：“我需要有一个在东边的酒店和一个高档餐馆。”同样，我们验证系统响应与对话策略标签的一致性。语义不一致的数据被通过提示 LLM 检查不一致性进行过滤，从而移除数据集的 2.37%。

整体自然性过滤最后，我们过滤掉不展现自然性的对话。由于系统和用户的发言是逐句生成的，因此对话的某些部分可能不会自然衔接。为了移除这些不自然的例子，我们将整个对话提供给 LLM 以评估其流畅性。结果，我们删除了 4.39 % 的对话。经过这些单独的过滤阶段后，初始数据集的 92.59 % 被保留。

2.6 过滤的案例研究

图 2 展示了样式强度过滤器（左）和样式方向过滤器（右）的过滤结果示例。样式强度过滤器移除那些最小或未变化个性化的实例，而样式方向过滤器排除那些个性化导致不恰当变化的案例，例如中世纪风格的语言或转换成不同语言。

3 PicPersona-TOD 分析

在本节中，我们将对我们的数据集与其他数据集 (§ 3.1) 进行比较分析，随后分析关键维度，如词语难度、礼貌性 (§ 3.2) 和情感 (§ 3.3)。

3.1 与现有数据集的比较

在表格 1 中，我们在个性化和对话数据模态方面将 PicPersona-TOD 与其他数据集进行比较。正如结果所示，我们的数据集以唯一一个结合了远见性人格的 TOD 数据集而具有独特地位。此外，通过整合多个 TOD 数据集，它涵盖了广泛的服务，并包含大量的对话数据。我们还通过引入外部资源（如谷歌评论和维基信息）增强了个性化，这是我们数据集的一个显著方面。

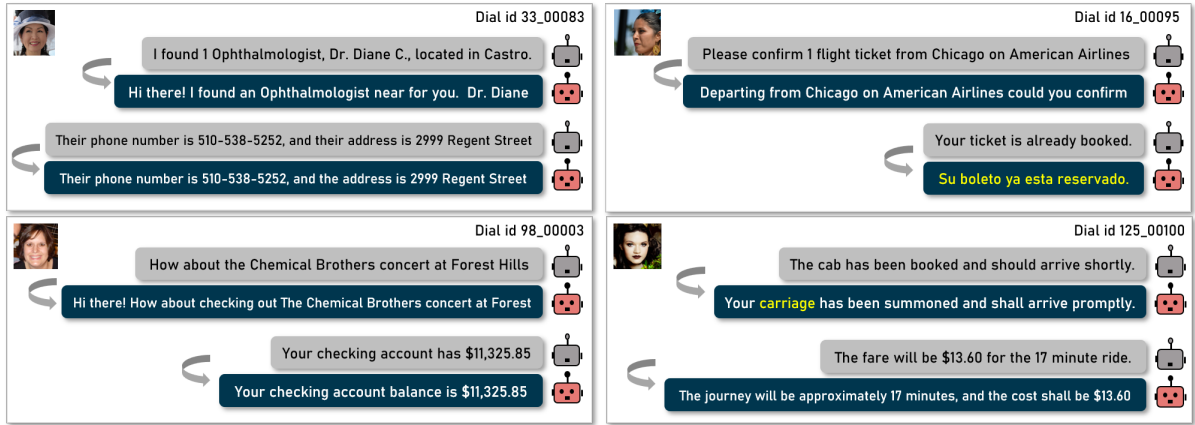


Figure 2: 被过滤结果的示例：样式强度过滤（左）和样式方向过滤（右）。

Dataset	Persona Mod.	Dialogue Mod.	Dialogue Type	Subtask	Collection	# of Dial	# of Serv.	Avg Turn	Avg Tok
MultiWoZ (Budzianowski et al., 2018)	-	Text	TOD	DST, Pol	Human	8,438	7	13.46	13.13
ABCD (Chen et al., 2021)	-	Text	TOD	Pol	Human	8,034	30	22.08	9.17
SGD (Rastogi et al., 2020)	-	Text	TOD	DST, Pol	Bot+Human	16,142	16	20.44	9.75
STAR (Mosig et al., 2020)	-	Text	TOD	Pol	Human	5,820	13	21.71	11.2
TOAD (Liu et al., 2024)	Text	Text	TOD	DST, Pol	GPT3.5	8,087	11	9.23	10.6
SIMMC-2.0 (Kottur et al., 2021)	-	Text, Vision	TOD	Disamb, Coref., DST	Bot+Human	11,244	2	10.4	13.7
DialogCC (Lee et al., 2024a)	-	Text, Vision	Open	Image Ret, Response Pred.	GPT4, CLIP	83k	-	8.20	-
MPChat (Ahn et al., 2023)	Text, Vision	Text, Vision	Open	Image Ret.	Reddit	15k	-	2.85	18.5
STARK (Lee et al., 2024b)	Text, Vision	Text, Vision	Open	Image Ret.	GPT4, Diffusion	0.5M	-	5.30	-
PicPersona-TOD (ours)	Text, Vision	Text, Vision	TOD	DST, Pol	GPT4, Google Map, Wiki	18,148	18	17.23	12.67

Table 1: 在个性化和模态方面与其他数据集的统计比较。‘Mod.’、‘Serv.’ 和 ‘Pol.’ 分别代表模态、服务和政策预测。

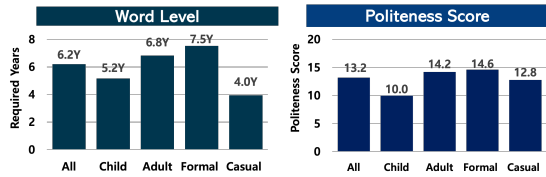


Figure 3: PicPersona-TOD 的词汇分析。“字词级别”指的是理解所需的教育年限，而礼貌评分则代表礼貌策略的平均使用频率。

3.2 词汇复杂度 & 礼貌

单词复杂性我们分析了系统响应在不同用户场景中的复杂性，包括与儿童、成人的互动，以及在正式和非正式环境中的互动（图 3）。为了评估词汇的难度，我们评估了理解系统话语所需的教育年限⁴。平均而言，系统的响应需要 6.2 年的教育才能理解。对于儿童的响应需要 5.2 年的教育，而对于成人则需要 6.17 年。这表明系统根据用户的年龄有效地个性化了词汇选择。

礼貌通过测量每句使用的礼貌策略的平均数量来评估响应的礼貌度，其中较高的得分表示更大的礼貌⁵。总体而言，PicPersona-TOD 的

⁴我们使用了来自 Textstat Python 库的 Gunning Fog Scale <https://pypi.org/project/textstat/>。

⁵我们使用了来自 ConvoKit 的 PolitenessStrategies 库 <https://convokit.cornell.edu/>，并测量策略的平均数量。

礼貌得分为 13.2。在正式场合中，得分上升到 14.6，而在非正式场合中降至 12.8。对于儿童，礼貌得分下降到 10.0，而对于成人，则增加到 14.2。这表明系统能够根据用户的环境调整其礼貌度。

3.3 回应中的情绪意识

表格 2 比较了在 MultiWoZ 和 PicPersona-TOD 数据集上的系统响应的情感分布，这些情感使用 GoEmotions 的分类体系 (Demszky et al., 2020) 进行分类，共分为 27 种情感，GPT-4 被用作分类器。虽然在两个数据集中中性情感最为常见，但 PicPersona-TOD 的中性响应比例明显低于 MultiWoZ (61.50 % 对比 74.97 %)，这表明其反应更加情感化。此外，我们分析了系统如何根据用户的情感状态个性化其响应（表格 2 的第三列）。我们发现，当用户的图像表现出积极的情感时，系统会以更广泛的情感响应，如快乐和感激。相反，当用户的情感是中性或消极时，系统倾向于生成更多中性的响应，更加注重同情和关怀。

为了评估用户对个性化结果的满意度，我们进行了人工评估，重点验证个性化风格的程度和信息保留。三位评估者对 100 个随机选择的对话进行了 4 点 Likert 量表评分，并使用五个问题测量了用户和系统发言的得分（附录 E）。结果非常积极，用户风格适当性平均得分为

MultiWoZ	PicPersona-TOD	PicPersona-TOD (pos)
neutral	74.97	neutral 44.66
curiosity	9.18	approval 16.51
gratitude	6.31	curiosity 11.65
approval	3.56	gratitude 7.77
optimism	1.26	excitement 4.86
apology	0.92	admiration 2.43
annoyance	0.69	PicPersona-TOD (neu/neg)
confusion	0.57	neutral 66.69
caring	0.57	curiosity 11.84
disappointed	0.57	gratitude 6.10
joy	0.35	approval 5.90
excitement	0.35	caring 2.54
admiration	0.23	confusion 0.38
		optimism 1.36

Table 2: 测试对话中系统响应的最常见情绪比例 (%)。

	All	Age			Emotion	
		Senior	Adult	Child	Pos	Neu & Neg
Liu et al. (2024)	2.22	0.00	1.80	4.17	1.01	3.70
Tie	13.33	9.52	14.41	12.50	6.06	22.22
PicPersona-TOD	84.44	90.48	83.78	83.33	92.93	74.07
Joshi et al. (2017)	9.44	14.29	9.01	8.33	4.04	16.05
Tie	23.33	33.33	22.52	20.83	15.15	33.33
PicPersona-TOD	67.22	52.38	68.47	70.83	80.81	50.62

Table 3: 比较 PicPersona-TOD 与个性化方法时的胜率 (%)。

3.89，用户语义一致性为 3.90，系统风格个性化为 3.69，系统语义一致性为 3.87，总体用户满意度为 3.76 (图 ??)。此外，我们观察到评估者之间的强一致性为 0.85，使用 Krippendorff’s Alpha 测量。这些结果证实了 PicPersona-TOD 符合我们预定的标准：(1) 用户和图像的一致性，(2) 个性化系统回应，以及 (3) 保持原信息。我们还使用 GPT-4 进行了相同的评估，显示出与人工评估者的高度一致性，得分为 0.84 (附录 E.1)。

3.4 其他个性化方法

我们对 PicPersona-TOD 与另外两种个性化方法进行了比较评估。由于目前没有将视觉人格纳入 TOD 的方法，我们将 PicPersona-TOD 与依赖文本模式的方法进行了比较。第一个基线，Liu et al. (2024)，通过模仿用户的名词和动词短语来个性化对话，而第二种方法，Joshi et al. (2017)，则根据年龄和性别信息个性化互动。我们从各种场景中抽样了 120 个对话，并请三位人类评审来评估它们，以基于个性化质量确定更优方法 (附录 ??)。如表 3 所示，在各种用户场景中，PicPersona-TOD 始终优于基于文本的方法。此结果强调了相对于依赖文本人格，丰富的，同时存在的图像人格对于个性化的重要性。

图 4 包含影响评估者偏好的关键因素的分布，这与表格 3 相关。在所有用户场景中，适当的正式性是人类评估者个性化最明显的方面。对于儿童和积极情境中，量身定制的问候语增强了个性化感，而在中性和消极环境中，

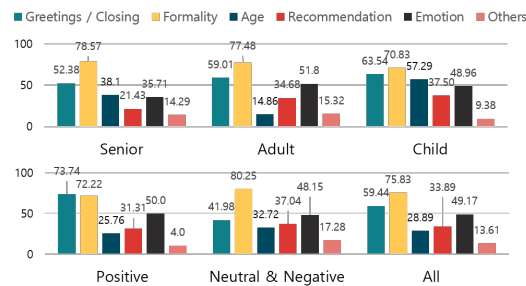


Figure 4: 各年龄段和情感的获胜特征分析。

情感意识是创造个性化体验的关键因素。

在本节中，我们介绍了 Pictor，这是一种在 PicPersona-TOD 数据集上训练的自然语言生成 (NLG) 基线，用于生成个性化的响应。此外，我们提供了用于对话状态跟踪 (DST) 和策略预测的模型，以便于与其他数据集进行比较。

使用 PicPersona-TOD 数据集，我们开发了一种名为 Pictor 的多模态任务导向对话 (TOD) 响应生成模型 (如图 ?? 所示)。Pictor 通过利用用户图像和对话上下文来生成个性化的响应 ($\tilde{s}_i$)。具体来说，Pictor 模型在对话回合 D_i 的第 t 回合的输入包括对话进展 (例如，开始、中间、结束)、用户的表达 ($\tilde{u}_t$)、对话策略 (pol_t)、数据库结果 (DB_t) 和用户的图像 (Img_i)。与构建 PicPersona-TOD 数据集的过程类似，我们首先生成用户印象，然后基于此印象生成个性化响应 $\tilde{s}_i$ 。Pictor 基于 LLaVA 7B (Li et al., 2023a) 和 1.5B (Zhou et al., 2024) 模型，这些模型在各种视觉语言任务中表现出色。我们通过使用秩为 16 的 LoRA (Hu et al., 2021) 适配器来训练 Pictor 模型。详细信息见附录 B。

3.5 DST 和策略的基线

PicPersona-TOD 数据集还支持一系列 TOD 任务，包括 DST 和策略预测。为了为这些任务建立基线，我们使用了 PPTOD 模型 (Su et al., 2021)，并使用 T5-base 和 T5-small 变体 (Raffel et al., 2020) 训练了 DST 和策略预测模型。对于 DST 任务，在回合 t 的输入被定义为所有用户和系统话语的连接，包括回合 t ，这可以表示为 $\text{Input}_{\text{DST},t} = [\tilde{u}_1, \tilde{s}_1, \tilde{u}_2, \tilde{s}_2, \dots, \tilde{u}_t]$ 。DST 的输出表示为槽值对 (例如，hotel-name: Green Hotel)。

对于策略预测任务，回合 t 的输入构建方法类似，但增加了预测的对话状态； $\text{Input}_{\text{POL},t} = [\tilde{u}_1, \tilde{s}_1, \tilde{u}_2, \tilde{s}_2, \dots, \tilde{u}_t, \text{DST}_t]$ 。策略预测模型生成适当的系统动作，例如请求进一步的信息 (例如，Request-restaurant-foodtype)。

4 基线评估

我们对我们的 Pictor 7B 模型与几个知名的视觉-LLM 进行了综合比较，包括 Llama3-8b、LLaVA 7B、InstructBLIP 7B 和 GPT-4o-mini。为了进行评估，我们抽取了 100 个对话，由 GPT-4 模型进行评估。如图所示，Pictor 在个性化质量方面，在各个类别中始终优于其他 sLLM 模型。相比于可能拥有更多参数的 GPT-4o-mini，Pictor 表现出更好的结果，除非是在中性/消极的情况下。这些发现突出了专为个性化设计的数据集（如 PicPersona-TOD）的重要性，以在个性化场景中实现最佳性能。

4.1 泛化性能评估

Domain	Natural.	Fluency	Personalize	Semantic	User Satisfaction
BUS	3.67	3.79	3.51	3.78	3.70
MOVIE	3.72	3.79	3.58	3.82	3.81
HOME	3.90	3.96	3.78	3.93	3.88

Table 4: 对 Pictor 在未见领域的人类评价。

在表 4 中，我们对 SGD 数据集中不包含在模型训练数据中的公交、电影和家庭领域的 Pictor 模型进行泛化性能的人为评估。请注意，这些数据集是使用与 PicPersona 数据集中的对话相同的流程构建的，后者也包含用户图像。我们从每个领域抽取了 100 个对话，进行零样本推理，接着由三位注释者使用 4 分 Likert 量表（附录 E.1）进行人为评估。结果显示，尽管这些领域不是 Pictor 的训练集的一部分，但该模型能够实现强的个性化，与用户满意度接近 4 的高分。这归功于两个大规模 TOD 数据集作为数据集来源的引入，它们涵盖了广泛的领域。

4.2 对 Pictor 的消融研究

我们进行了一个消融研究，以检查不同组件如何影响 Pictor 的生成性能。我们使用 BLEU 分数⁶和 GPT-4 评估⁷对语义准确性、风格和总体满意度进行评估（表 5）。结果揭示了一个关于印象的有趣发现；虽然省略生成印象导致了更高的 BLEU 分数，但加入印象显著提高了个性化。在更大的模型中，这一趋势更加明显，当包含印象时，风格评分从 3.48 增加到 3.51。

4.3 DST 和政策推断结果

我们进行了实验，通过使用 T5-small 和 T5-base 来测试 DST 和策略模型，以评估 PicPersona-TOD 数据集的信息准确性。为了比较，我们还

⁶nlk.translate.bleu_score
⁷请注意，GPT4 分数使用与 § 到 ?? 中详细描述的同问题格式。

Input	BLEU	Style	Semantic	Overall
LLaVA 1.5B				
Pol +DB	8.75	2.71	2.95	2.60
Pol +DB + $\bar{u}$	14.28	3.15	3.52	3.1
Pol +DB + $\bar{u}$ + Img	16.18	3.47	3.74	3.41
Pol +DB + $\bar{u}$ + Img + Imp (Pictor)	14.96	3.47	3.76	3.41
LLaVA 7B				
Pol +DB	15.46	3.00	3.49	2.99
Pol +DB + $\bar{u}$	20.21	3.18	3.63	3.22
Pol +DB + $\bar{u}$ + Img	22.01	3.48	3.82	3.50
Pol +DB + $\bar{u}$ + Img + Imp (Pictor)	20.77	3.51	3.89	3.53

Table 5: 对 Pictor 模型进行训练的消融研究。每个实验重复三次，结果取平均值。

展示了使用 MultiWOZ 数据集的结果。表 ?? 显示，我们的数据集在大多数指标上表现与 MultiWOZ 相当，仅观察到微小差异。这种一致性表明，尽管我们的数据集因个性化用户和系统而更复杂，但 PicPersona-TOD 在信息准确性方面仍与人工策划的数据集保持一致。

在 TOD 数据集方面的进展任务导向对话 (TOD) 系统长期以来一直是研究的重点，早期的数据集如 ATIS (Hemphill et al., 1990)、WOZ2.0 (Wen et al., 2016) 和 DSTC2 (Henderson et al., 2014) 仅限于单一领域。后来，多领域的数据集如 M2M (Shah et al., 2018)、MultiWOZ (Budzianowski et al., 2018)、SGD (Rastogi et al., 2020) 和 ABCD (Chen et al., 2021) 旨在提高准确性，但往往忽视了用户满意度。最近的研究通过整合闲聊 (Sun et al., 2020; Young et al., 2022; Stricker and Paroubek, 2024)、在系统响应中提供详细解释 (Kim et al., 2023; Qian et al., 2021)、并考虑用户的情感状态 (Abolghasemi et al., 2024; Feng et al., 2024) 等方式，试图增强用户体验，尽管很少有研究关注个性化定制。与我们工作最相关的是个性化努力，比如结合年龄和性别 (Joshi et al., 2017) 或语言模式 (Liu et al., 2024)。虽然一些方法包括情感状态 (Lin et al., 2023; Feng et al., 2024)，但我们的方法引入了一个具有愿景的人物，为用户提供更丰富和更同步的信息，从而提高用户满意度。

将角色融入对话虽然个性化对话系统已经被广泛研究以改善用户体验，但它们传统上依赖于文本信息。方法包括通过叙述句子 (Zhang et al., 2018b; Zhong et al., 2020)、键值对字典 (Qian et al., 2017; Zheng et al., 2019) 或用户评论历史 (Kim et al., 2024) 构建角色。最近，多模态方法已经出现，结合用户图片以创建更丰富的角色 (Ahn et al., 2023; Lee et al., 2024b; Agrawal et al., 2023)。基于这些进展，我们介绍了一种新方法，在 TOD 数据集中使用用户图片作为角色的主要基础，从而实现更具上下文适应性和个性化的响应。

通过蒸馏大型语言模型 (LLM) 进行数据生

成收集对话数据面临隐私问题、高成本以及需要多个参与者的挑战。为了解决这些问题,许多研究利用 LLM 进行数据集创建。例子包括编制种子对话 (Ahn et al., 2023; Kim et al., 2022b)、构建社会事件图 (Kim et al., 2022a) 以及生成长期对话 (Jang et al., 2023)。其他人则使用 LLM 生成具备常识的对话 (Chae et al., 2023)、亲社会对话 (Kim et al., 2022b) 和任务导向的对话 (TOD) 话语 (Kulkarni et al., 2024)。LLM 生成的数据集具有成本效益、多样性,且通常比人工编制的数据集更受欢迎 (Kim et al., 2022a; Lee et al., 2024b, 2021)。在此基础上,我们使用 LLM 生成个性化的、注重隐私且多样化的用户场景。

在本文中,我们介绍了 PicPersona-TOD,这是一个在任务导向对话 (TOD) 领域中基于用户视觉角色定制系统响应的新颖数据集。具体来说, PicPersona-TOD 在问候语、年龄、礼貌和情感方面结合了个性化响应。通过用户满意度实验,我们证明了 PicPersona-TOD 在保持原始信息的同时增强了个性化。此外,我们提出并分析了一个基线模型,该模型包括自然语言生成 (NLG)、对话状态跟踪 (DST) 和策略预测。我们的结果表明,这种方法在不影响其他关键任务性能的情况下改善了个性化体验。我们相信这项工作推动了具备多模态用户角色的个性化任务导向对话研究,使得交流更加自然和类似人类。

5

限制

缺乏直接的基准比较通常,由 LLM 提取的数据库,比如用于开放对话 (Kim et al., 2022a) 或聊天中的图像发送 (Lee et al., 2024a) 的数据库,通常会与传统的人工创建的测试集直接进行比较,以展示新数据库的实际优势。然而,在我们的案例中,没有传统的、人工创建的 TOD (任务导向对话) 数据库存在,这些数据库能结合用户角色,因为创建这样的 TOD 数据库相较于开放对话数据库需要显著更多的精力和更高的标注成本。

由于这种环境,我们无法与标准数据集进行直接比较。相反,我们在第 ?? 节中评估了我们模型相对于其他著名视觉-大语言模型的个性化能力,这些模型已经在大规模视觉文本数据集上进行了训练。我们进行了 GPT4 评估,以评估在这些比较中的性能,结果显示对在我们的数据集上训练的模型表现出了强烈的偏好。尽管这不能作为与传统 TOD 基准的直接比较,我们相信它提供了一种有效的替代方法,因为结果突显了专门为个性化设计的数据集的重要

性,例如 PicPersona-TOD,显示了我们数据集的优势。

其他 LLM 使用的局限性在我们的研究中,我们仅使用 GPT-4o 进行数据集策划。这个决定是基于初步实验做出的,其中其他视觉-LLM 未能生成与 GPT-4o 质量相同的个性化响应。然而,随着开源视觉-LLM 模型的不断改进,它们可能成为开发高质量、成本高效的数据集生成管道的可行选项。

面部识别要求虽然我们的数据集在个性化方面展现了强大的潜力,但其全部能力只能在配备面部识别技术的系统中实现,例如自助服务终端或机器人。没有这些设备, PicPersona-TOD 的基于图像的个性化功能无法得到有效利用。简单应用检索结果在构建我们的数据集时,我们将评论和维基百科结果整合为一种基于检索的生成方法,以增强系统响应。尽管这种方法有助于提高响应质量,但它缺乏复杂性。未来的研究可以专注于开发更先进的基于图像个性化的检索方法,从而实现更深入的个性化和对用户个性的更复杂理解,这最终将提升响应质量。

6

伦理考虑 在构建数据集时,我们使用了两个涉及真实用户的来源,这可能会引发关于隐私和同意的担忧。对于图像数据,我们使用了 FFHQ 数据集。根据原始论文 (Karras et al., 2019),这个数据集是通过抓取 Flickr 网站上的图片创建的,其中仅收集了在宽松许可下的图片。具体来说,这些图片的许可为 CC BY-NC-SA 2.0,允许只要使用是非商业性质,就可以自由分发。

对于谷歌地图评论数据,我们使用谷歌地图 API 检索评论。谷歌地图用户同意通过谷歌 API 使用条款收集他们的数据,该条款指出用户了解他们的帖子将会公开可用并可以通过 API 访问。此外,为了保护用户隐私,我们没有收集任何个人可识别信息 (PII),以确保匿名性。从同意的角度来看,我们遵循两个来源的许可条款,并采取措施保护用户隐私。此外,我们的数据使用限制于商业和营利目的。

7

致谢

本研究受到韩国科学和信息通信技术部 (MSIT) 通过信息通信技术计划与评估研究所 (IITP) 的资助: RS-2019-II191906 (POSTECH 人工智能研究生院计划, 50%) 和 RS-2024-00437866 (ITRC 计划 50%)。

References

- Amin Abolghasemi, Zhaochun Ren, Arian Askari, Mohammad Aliannejadi, Maarten de Rijke, and Suzan Verberne. 2024. Cause: Counterfactual assessment of user satisfaction estimation in task-oriented dialogue systems. *arXiv preprint arXiv:2403.19056*.
- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Harsh Agrawal, Aditya Mishra, Manish Gupta, et al. 2023. Multimodal persona based generation of comic dialogs. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14150–14164.
- Jaewoo Ahn, Yeda Song, Sangdoo Yun, and Gunhee Kim. 2023. Mpmchat: Towards multimodal persona-grounded conversation. *arXiv preprint arXiv:2305.17388*.
- Peter Borkenau, Steffi Brecke, Christine Möttig, and Marko Paelecke. 2009. Extraversion is accurately perceived after a 50-ms exposure to a face. *Journal of research in personality*, 43(4):703–706.
- Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Ultes Stefan, Ramadan Osman, and Milica Gašić. 2018. Multiwoz - a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Hyunjoo Chae, Yongho Song, Kai Ong, Taeyoon Kwon, Minjin Kim, Youngjae Yu, Dongha Lee, Dongyeop Kang, and Jinyoung Yeo. 2023. Dialogue chain-of-thought distillation for commonsense-aware conversational agents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5606–5632.
- Derek Chen, Howard Chen, Yi Yang, Alex Lin, and Zhou Yu. 2021. Action-based conversations dataset: A corpus for building more in-depth task-oriented dialogue systems. *arXiv preprint arXiv:2104.00783*.
- Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. Goemotions: A dataset of fine-grained emotions. *arXiv preprint arXiv:2005.00547*.
- Mihail Eric and Christopher D Manning. 2017. Key-value retrieval networks for task-oriented dialogue. *arXiv preprint arXiv:1705.05414*.
- Shutong Feng, Hsien-chin Lin, Christian Geishauser, Nurul Lubis, Carel van Niekerk, Michael Heck, Benjamin Ruppik, Renato Vukovic, and Milica Gašić. 2024. Infusing emotions into task-oriented dialogue systems: Understanding, management, and generation. *arXiv preprint arXiv:2408.02417*.
- Mauajama Firdaus, Hardik Chauhan, Asif Ekbal, and Pushpak Bhattacharyya. 2020. Meisd: A multi-modal multi-label emotion, intensity and sentiment dialogue dataset for emotion recognition and sentiment analysis in conversations. In *Proceedings of the 28th international conference on computational linguistics*, pages 4441–4453.
- Dylan F Glas, Kanae Wada, Masahiro Shiomi, Takayuki Kanda, Hiroshi Ishiguro, and Norihiro Hagita. 2017. Personal greetings: Personalizing robot utterances based on novelty of observed behavior. *International Journal of Social Robotics*, 9:181–198.
- Charles T Hemphill, John J Godfrey, and George R Doddington. 1990. The atis spoken language systems pilot corpus. In *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27, 1990*.
- Matthew Henderson, Blaise Thomson, and Jason D Williams. 2014. The second dialog state tracking challenge. In *Proceedings of the 15th annual meeting of the special interest group on discourse and dialogue (SIGDIAL)*, pages 263–272.
- Ehsan Hosseini-Asl, Bryan McCann, Chien-Sheng Wu, Semih Yavuz, and Richard Socher. 2020. A simple language model for task-oriented dialogue. *Advances in Neural Information Processing Systems*, 33:20179–20191.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.
- Jihyoung Jang, Minseong Boo, and Hyoungun Kim. 2023. Conversation chronicles: Towards diverse temporal and relational dynamics in multi-session conversations. *arXiv preprint arXiv:2310.13420*.
- Chaitanya K Joshi, Fei Mi, and Boi Faltings. 2017. Personalization in goal-oriented dialog. *arXiv preprint arXiv:1706.07503*.
- Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410.
- Hyunwoo Kim, Jack Hessel, Liwei Jiang, Peter West, Ximing Lu, Youngjae Yu, Pei Zhou, Ronan Le Bras, Malihe Alikhani, Gunhee Kim, et al. 2022a. Soda: Million-scale dialogue distillation with social commonsense contextualization. *arXiv preprint arXiv:2212.10465*.
- Hyunwoo Kim, Youngjae Yu, Liwei Jiang, Ximing Lu, Daniel Khashabi, Gunhee Kim, Yejin Choi, and Maarten Sap. 2022b. Prosocialdialog: A prosocial backbone for conversational agents. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4005–4029.

- Minjin Kim, Minju Kim, Hana Kim, Beong-woo Kwak, Soyeon Chun, Hyunseo Kim, SeongKu Kang, Youngjae Yu, Jinyoung Yeo, and Dongha Lee. 2024. Pearl: A review-driven persona-knowledge grounded conversational recommendation dataset. *arXiv preprint arXiv:2403.04460*.
- Yongil Kim, Yerin Hwang, Joongbo Shin, Hyunkyung Bae, and Kyomin Jung. 2023. Injecting comparison skills in task-oriented dialogue systems for database search results disambiguation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4047–4065.
- Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Satwik Kottur, Seungwhan Moon, Alborz Geramifard, and Babak Damavandi. 2021. Simmc 2.0: A task-oriented dialog dataset for immersive multimodal conversations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4903–4912.
- Atharva Kulkarni, Bo-Hsiang Tseng, Joel Moniz, Dhivya Piraviperumal, Hong Yu, and Shruti Bhargava. 2024. Synthdst: Synthetic data is all you need for few-shot dialog state tracking. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1988–2001.
- Chia-Hsuan Lee, Hao Cheng, and Mari Ostendorf. 2021. Dialogue state tracking with a language model using schema-driven prompting. *arXiv preprint arXiv:2109.07506*.
- Young-Jun Lee, Byungsoo Ko, Han-Gyu Kim, Jonghwan Hyeon, and Ho-Jin Choi. 2024a. Dialogcc: An automated pipeline for creating high-quality multimodal dialogue dataset. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1938–1963.
- Young-Jun Lee, Dokyong Lee, Junyoung Youn, Kyeongjin Oh, Byungsoo Ko, Jonghwan Hyeon, and Ho-Jin Choi. 2024b. Stark: Social long-term multi-modal conversation with persona common-sense knowledge. *arXiv preprint arXiv:2407.03958*.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich K ttler, Mike Lewis, Wen-tau Yih, Tim Rockt schel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.
- Lei Li, Yuwei Yin, Shicheng Li, Liang Chen, Peiyi Wang, Shuhuai Ren, Mukai Li, Yazheng Yang, Jingjing Xu, Xu Sun, et al. 2023a. M³ it: A large-scale dataset towards multi-modal multilingual instruction tuning. *arXiv preprint arXiv:2306.04387*.
- Xin Li, Jicai Pan, Yufei Xiao, Yanan Chang, Feiyi Zheng, Shangfei Wang, et al. 2023b. Medic: A multimodal empathy dataset in counseling. *arXiv preprint arXiv:2305.02842*.
- Hsien-Chin Lin, Shutong Feng, Christian Geishauser, Nurul Lubis, Carel van Niekerk, Michael Heck, Benjamin Ruppik, Renato Vukovic, and Milica Gasi c. 2023. Emous: Simulating user emotions in task-oriented dialogues. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2526–2531.
- Zhaojiang Lin, Andrea Madotto, Genta Indra Winata, and Pascale Fung. 2020. Mintl: Minimalist transfer learning for task-oriented dialogue systems. *arXiv preprint arXiv:2009.12005*.
- Yinhong Liu, Yimai Fang, David Vandyke, and Nigel Collier. 2024. Toad: Task-oriented automatic dialogs with diverse response styles. *arXiv preprint arXiv:2402.10137*.
- Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.
- Graeme McLean, Kofi Osei-Frimpong, and Jennifer Barhorst. 2021. Alexa, do voice assistants influence consumer brand engagement?—examining the role of ai powered voice assistants in influencing consumer brand engagement. *Journal of Business Research*, 124:312–328.
- Johannes EM Mosig, Shikib Mehri, and Thomas Kober. 2020. Star: A schema-guided dialog dataset for transfer learning. *arXiv preprint arXiv:2010.11853*.
- Atsumoto Ohashi and Ryuichiro Higashinaka. 2023. Enhancing task-oriented dialogue systems with generative post-processing networks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3815–3828.
- Baolin Peng, Chunyuan Li, Jinchao Li, Shahin Shayan-deh, Lars Liden, and Jianfeng Gao. 2020. Soloist: Few-shot task-oriented dialog with a single pre-trained auto-regressive model. *arXiv preprint arXiv:2005.05298*, 3.
- Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2018. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508*.
- Kun Qian, Ahmad Beirami, Satwik Kottur, Shahin Shayan-deh, Paul Crook, Alborz Geramifard, Zhou Yu, and Chinnadhurai Sankar. 2021. Database search results disambiguation for task-oriented dialog systems. *arXiv preprint arXiv:2112.08351*.

- Qiao Qian, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2017. Assigning personality/identity to a chatting machine for coherent conversation generation. *arXiv preprint arXiv:1706.02861*.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.
- Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8689–8696.
- N Reimers. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.
- Nicholas O Rule and Nalini Ambady. 2008. First impressions: Peeking at the neural correlates. In *First impressions: Peeking at the neural correlates*. Guilford Press.
- Pararth Shah, Dilek Hakkani-Tür, Gokhan Tür, Abhinav Rastogi, Ankur Bapna, Neha Nayak, and Larry Heck. 2018. Building a conversational agent overnight with dialogue self-play. *arXiv preprint arXiv:1801.04871*.
- Xiaoteng Shen, Rui Zhang, Xiaoyan Zhao, Jieming Zhu, and Xi Xiao. 2024. Pmg: Personalized multimodal generation with large language models. In *Proceedings of the ACM on Web Conference 2024*, pages 3833–3843.
- Tianyuan Shi, Liangzhi Li, Zijian Lin, Tao Yang, Xiaojun Quan, and Qifan Wang. 2023. Dual-feedback knowledge retrieval for task-oriented dialogue systems. *arXiv preprint arXiv:2310.14528*.
- Armand Stricker and Patrick Paroubek. 2024. Chitchat as interference: Adding user backstories to task-oriented dialogues. *arXiv preprint arXiv:2402.15248*.
- Yixuan Su, Lei Shu, Elman Mansimov, Arshit Gupta, Deng Cai, Yi-An Lai, and Yi Zhang. 2021. Multi-task pre-training for plug-and-play task-oriented dialogue system. *arXiv preprint arXiv:2109.14739*.
- Kai Sun, Seungwhan Moon, Paul Crook, Stephen Roller, Becka Silvert, Bing Liu, Zhiguang Wang, Honglei Liu, Eunjoon Cho, and Claire Cardie. 2020. Adding chit-chat to enhance task-oriented dialogues. *arXiv preprint arXiv:2010.12757*.
- Michel Valstar, Jonathan Gratch, Björn Schuller, Fabien Ringeval, Denis Lalanne, Mercedes Torres Torres, Stefan Scherer, Giota Stratou, Roddy Cowie, and Maja Pantic. 2016. Avec 2016: Depression, mood, and emotion recognition workshop and challenge. In *Proceedings of the 6th international workshop on audio/visual emotion challenge*, pages 3–10.
- Tsung-Hsien Wen, David Vandyke, Nikola Mrksic, Milica Gasic, Lina M Rojas-Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. 2016. A network-based end-to-end trainable task-oriented dialogue system. *arXiv preprint arXiv:1604.04562*.
- Janine Willis and Alexander Todorov. 2006. First impressions: Making up your mind after a 100-ms exposure to a face. *Psychological science*, 17(7):592–598.
- Chien-Sheng Wu, Andrea Madotto, Ehsan Hosseini-Asl, Caiming Xiong, Richard Socher, and Pascale Fung. 2019. Transferable multi-domain state generator for task-oriented dialogue systems. *arXiv preprint arXiv:1905.08743*.
- Yunyi Yang, Yunhao Li, and Xiaojun Quan. 2021. Ubar: Towards fully end-to-end task-oriented dialog system with gpt-2. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 14230–14238.
- Steve Young, Milica Gašić, Blaise Thomson, and Jason D Williams. 2013. Pomdp-based statistical spoken dialog systems: A review. *Proceedings of the IEEE*, 101(5):1160–1179.
- Tom Young, Frank Xing, Vlad Pandealea, Jinjie Ni, and Erik Cambria. 2022. Fusing task-oriented and open-domain dialogues in conversational agents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11622–11629.
- Xiao Yu, Qingyang Wu, Kun Qian, and Zhou Yu. 2022. Krls: Improving end-to-end response generation in task oriented dialog with reinforced keywords learning. *arXiv preprint arXiv:2211.16773*.
- Xiaoxue Zang, Abhinav Rastogi, Srinivas Sunkara, Raghav Gupta, Jianguo Zhang, and Jindong Chen. 2020. Multiwoz 2.2: A dialogue dataset with additional annotation corrections and state tracking baselines. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI, ACL 2020*, pages 109–117.
- Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018a. Personalizing dialogue agents: I have a dog, do you have pets too? *arXiv preprint arXiv:1801.07243*.
- Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018b. Personalizing dialogue agents: I have a dog, do you have pets too? *arXiv preprint arXiv:1801.07243*.
- Yinhe Zheng, Guanyi Chen, Minlie Huang, Song Liu, and Xuan Zhu. 2019. Personalized dialogue generation with diversified traits. *arXiv preprint arXiv:1901.09672*.

Peixiang Zhong, Chen Zhang, Hao Wang, Yong Liu, and Chunyan Miao. 2020. Towards persona-based empathetic conversational models. *arXiv preprint arXiv:2004.12316*.

Baichuan Zhou, Ying Hu, Xi Weng, Junlong Jia, Jie Luo, Xien Liu, Ji Wu, and Lei Huang. 2024. Tinyllava: A framework of small-scale large multi-modal models. *arXiv preprint arXiv:2402.14289*.

A 案例研究

图 5 展示了系统响应个性化的不同方法 (§ 2.4)。对于问候个性化 (左侧), 系统通过结合用户的第一印象中的特定特征来个性化问候和告别。对于推荐个性化 (右侧), 系统从评论中检索信息并根据这些信息定制响应。

我们使用 V-LLM 来获取用户图像中的性别、年龄和正式程度信息, 从而分析 PicPersona-TOD 的分布。性别分布显示, 数据集中 35.82 % 为女性, 而男性的比例稍大, 为 47.71 %。关于年龄分布, 成年人群体占据最大的比例, 为 74.24 %。比例随着年龄的增加而减少, 儿童组和老年组分别占数据集的 19.66 % 和 6.68 %。正式程度被划分为正式和非正式组, 其中非正式组占多数, 为 86.54 %, 正式组占 13.46 %。整个数据集的情感分布为, 积极情感占 50.92 %, 中性情感占 52.44 %, 消极情感占 0.55 %。这些结果突出了 PicPersona-TOD 数据集中性别、年龄、正式程度和情感的多样化人口表现。

我们使用 GPT-4 进行了两项平行的评估任务。在个性化质量评估任务 (§ ??) 中, GPT-4 给出的评分是: Q1: 3.79, Q2: 3.99, Q3: 3.66, Q4: 3.97, Q5: 3.75, 评分之间的可靠性指标 Krippendorff's alpha 达到了 0.84, 显示出高度的一致性。

B 基线训练详情

我们同时使用 LLaVA 1.5B 和 7B 模型训练了 Pictor 模型。对于 1.5B 模型, 我们采用了 LoRA, 设定了 16 的秩、64 的 alpha 值、16 的批次大小以及 $2e-5$ 的学习率, 总共 5 轮。在 7B 模型的情况下, LoRA 配置为 16 的秩、32 的 alpha 值、16 的批次大小以及 $5e-5$ 的学习率, 同样进行了 3 轮训练。两个模型都使用了 Adam 优化器 (Kingma and Ba, 2014), 且没有进行权重衰减, 并采用了余弦学习率调度, 启用 3 % 的预热比率。所有 Pictor 模型的训练都在 NVIDIA A100 GPU 上进行。

对于 DST 和策略模型 (T5-small 和 base 变体), 我们使用了 16 的批量大小, $1e-3$ 的学习率, 并使用不带权重衰减的 AdamW (Loshchilov and Hutter, 2017) 优化器训练了 10 个 epoch。这些模型是在 NVIDIA A6000 GPU 上训练的。

C DST 和策略预测的指标

在评估 DST 性能时, 我们使用了两个指标: 联合目标准确率 (JGA) 和领域特定 JGA。只有当一个对话轮次中的所有对话状态都准确时, JGA 才被认为是正确的。领域特定 JGA 则是

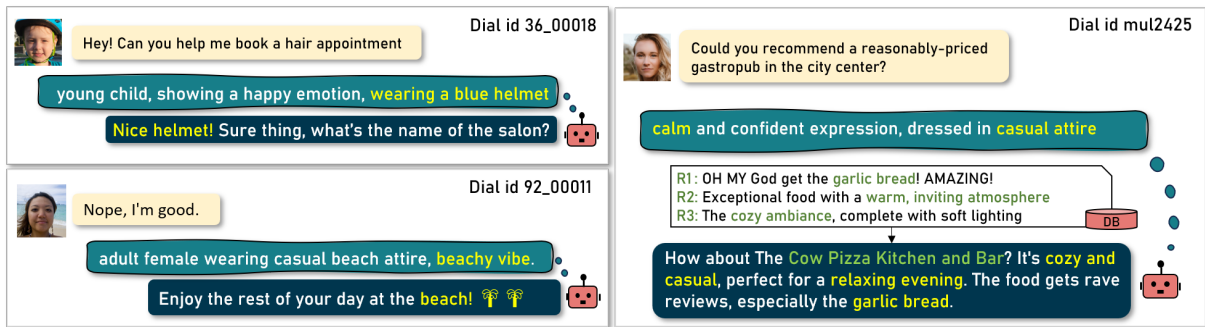


Figure 5: 问候个性化（左）和推荐个性化（右）的示例。

指针对目标领域的对话状态准确时标记为正确的，而不考虑其他领域 (Wu et al., 2019)。对于对话策略的评估，我们使用了实体 F1 分数，该分数计算每一轮次的 F1 分数，然后在所有轮次之间取平均 (Eric and Manning, 2017; Shi et al., 2023)。

D 许可

PicPersona-TOD 是使用 MultiWoZ 2.2、SGD 和 FFHQ 数据集集成的。MultiWoZ 2.2 是在 MIT 许可证下发布的，而 SGD 在 CC BY-SA 4.0 许可证下，FFHQ 数据集中的图像则根据 Creative Commons BY 2.0、Creative Commons BY-NC 2.0、Public Domain Mark 1.0、Public Domain CC0 1.0，或美国政府作品许可证进行许可。这些许可证允许免费使用、复制、修改和为非商业目的发布。

E 人工评估详情

对于人工评估，我们通过 Upwork⁸ 平台聘请了三位母语为英语的评估员。他们被告知所有个人信息将保持匿名，且他们提交的回复将仅用于研究目的。

E.1 第 ?? 节和第 4.1 节

评估者被要求通过选择提供的选项来评价 PicPersona-TOD 或 Pictor 的语句中个性化的质量，以回应以下问题，这些问题涉及用户语句 (U1-U4) 和系统语句 (S1-S5)。

- U1 & S1. Naturalness: Is the { user/system } 's utterance natural and conversationally appropriate?
- U2 & S2. Fluency: Does the { user/system } 's utterance flow smoothly without errors or awkwardness?
- U3. Does the user's utterance style match the user's image?
- U4. Is the content well preserved and maintains the same semantics as in the original user utterance?

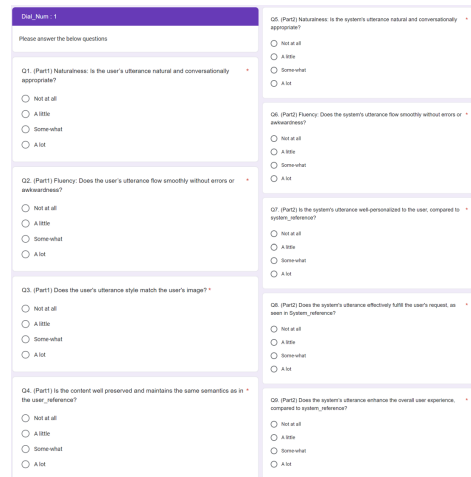


Figure 6: 个性化质量评估的屏幕截图。

- S3. Is the system's utterance well-personalized to the user, compared to the original system utterance?
- S4. Does the system's utterance effectively fulfill the user's request, as seen in the original system utterance?
- S5. Does the system's utterance enhance the overall user experience, compared to the original system utterance?
- Options: Not at all (1) / A little (2) / Some-what (3) / A lot (4).

我们使用 Google 表单进行评估，图 6 显示了评估者执行的样本截图。每个选项对应的分数用于计算结果。

评估者被要求比较两个匿名系统，并指出哪个部分得到了改进：

- 问题：哪个系统对用户来说更个性化？选项：系统 1 / 系统 2 / 相同
- 问题：内容是否被很好地保留并保持与用户原始话语中相同的语义？选项：
 - 个性化的问候和结束语句：个性化的回复可以包含自定义的问候语或结束语。
 - 形式：个性化的回复应根据情况恰当地正式或非正式。
 - 年龄敏感性：个性化的响应应对年龄敏感。

⁸www.upwork.com

- 情感意识：个性化的回应应具有情感意识。
- 其他原因

图 ?? 显示了用于评估的问题的屏幕截图。

F PicPersona-TOD 的样本

图 7 展示了来自 PicPersona-TOD 数据集的样本。左侧显示了来自原始 MultiWoZ 数据集的示例，而右侧则展示了来自 PicPersona-TOD 的样本。

示例提示模板已从第 18 页开始包括。

Dial id: mul1431, MultiWoZ

 I need to leave from broxbourne on Friday .

 There are several options, the first one leaving at 13:32. It is a 60 minute trip. Can I reserve seats for you?

 Can get the train id?

 Sure thing! It is TR7015

 That's awesome, thanks! By the way, I'm also looking for a place that serves African food at a moderate price. Can you help me with that?

 I am sorry but I do not have an African restaurants. Is there a different type of cuisine I could try?

 Would you know of any African restaurants in the east, regardless of the price range?

 There is a restaurant called Bedouin that serves African foods. They are in expensive price range, located in the centre. Can I book a table for you?

Dial id: mul1431, PicPersona-TOD

 Oh, I need to head out from Broxbourne on Friday.

First Impression

 Young female who is smiling, wearing a straw hat and giving off a cheerful, summery vibe.

 The first train option leaves at 13:32 and it's a 60-minute trip. Would you like me to reserve a seat for you?

 Could you please tell me the train ID?

 The train ID is TR7015!

 That's awesome, thanks! By the way, I'm also looking for a place that serves African food at a moderate price. Can you help me with that?

 Hmm, I don't have a recommendation for African restaurants at the moment. Are there other types of cuisine you'd like to explore?

 Would you know of any African restaurants in the east, regardless of the price range?

 **R1:** It was so difficult to pick one **R2:** warm, inviting atmosphere made the experience even better

 Bedouin in the center offers delicious North African cuisine with a cozy vibe and great reviews.

Dial id: pmul2026, MultiWoZ

 I am looking for a Chinese restaurant please.

 What area of town would you like the restaurant to be in?

 I would like to be in the centre of town .

 Jinling noodle bar is in the centre area serving Chinese in the moderate price range. Would you like to reserve a table?

 I want something expensive.

 I think you would enjoy ugly duckling. Want more information?

Dial id: pmul2026, PicPersona-TOD

 Could you recommend a spot that serves Chinese food?

First Impression

 Adult male who looks content, with a greenish beard and hair

 Hey there! That's an awesome beard, by the way. What part of town are you thinking for your Chinese food adventure?

 Let's hit up the city center.

 Jinling Noodle Bar is right in the city center and offers delicious Chinese food at a moderate price. How does that sound for a meal?

 I'm looking for a more upscale option.

 How about trying Ugly Duckling? I think it'd be a great fit. Want more details?

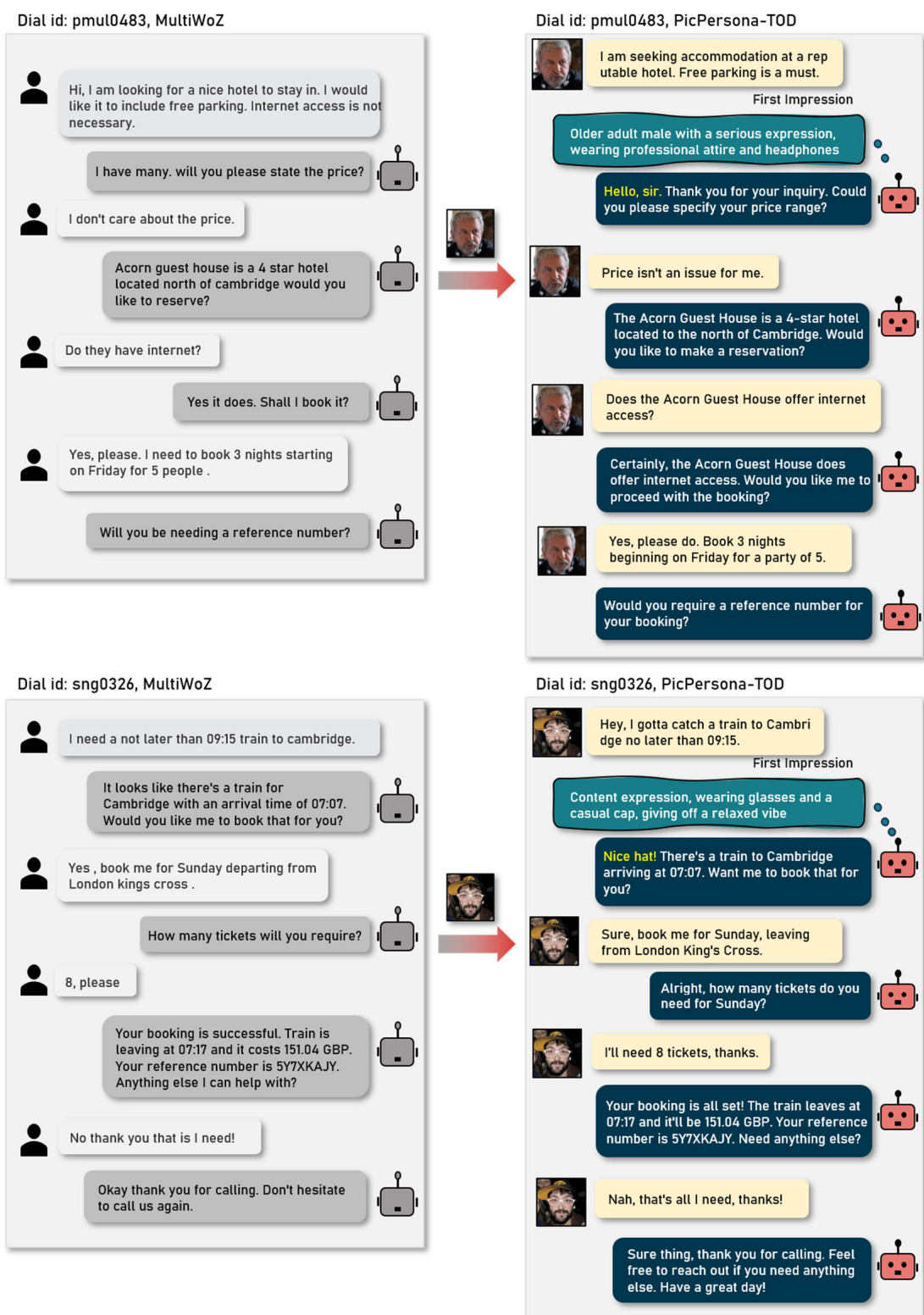


Figure 7: PicPersona-TOD 对话示例。

Prompt for Classifying Emotion in an Image

Classify the image's sentiment as negative, neutral, or positive. Answer should be one word.

Prompt for User Utterance Style Alignment

```
prompt = f"""
**Objective:**
Adjust the tone, age, gender, emotion, and formality of the user's utterance to match the style of
the user depicted in the provided image.
Rephrase the utterance as if the user originally spoke in that style, while preserving the original
meaning.
Answer naturally, and Do not add any greetings, closing remarks, or expressions of thanks unless
they were part of the original utterance.
"""

if previous_system:
    prompt += f"""
    Ensure the revised utterance flows naturally as a response to the following system message:
    Previous system message: {previous_system}
    """

prompt += f"""
Original user utterance: {user}
Rephrased user utterance:
"""
```

Prompt for First Impression

What is the first impression of the person in the image in terms of age, gender, emotion, outfit, and overall vibe? Answer in one complete sentence and start with 'The person in the image appears to be'

Prompt for Personalized System Utterance Generation

```
prompt = f"""
You are given a dialogue between a user and a system, consisting of the latest user utterance, the current
system utterance, and the next user utterance.
Your task is to modify the tone, formality, and wording of the current system utterance to give a
personalized response to the user. The response should match the provided image and description.
This is user image description : {first_impression}
Below is the information you have:

**Latest User Utterance (user1):**
{user1}

**Current System Utterance (system1):**
{system1}
"""
if user2 is not None:
    prompt += f"""
    **And this is the next user answer (user2):**
    {user2}
    As this is future information, do not use it in your response, just keep it in mind.
    """
if strategy != None:
    if strategy['name'] == 'greeting':
        prompt += '''As this is the first turn in the conversation, make the greetings reflect the user's
image or highlight something extraordinary about their appearance, like "Nice hat!" or "Congratulations on your
graduation!"""
        prompt += """However, if the user doesn't look like in a good mood, or it is formal setting, you
can just say "Hello" or "Hi" or Hello sir, madam, etc. Don't say appearance related things."""
    if strategy['name'] == 'goodbye':
        prompt += '''As this is the final turn in the conversation, make the ending statement. If needed,
reflect the user's image or "Enjoy your vacation!". You can just say "Goodbye" things if no other information
is available.'''
```

```

if strategy['name'] == 'DB':
    prompt += '''The user is providing information from internet soruces. \n'''
    prompt += f'This is online {strategy["DB_type"]} information for {strategy["Key"]}. \n'
    prompt += '\n'.join(strategy['online'])
    prompt += '''\nIf this relates to the user's age, emotion, gender, formality and their style, make
a personalized response using that context. \
    Mention why you recommend this, connecting it to something specific about their age, emotion,
gender, formality and events. \
    For instance, you might say, 'This could perfectly match your cool mood,' or 'Given your
artistic taste, this seems ideal,' \
    or even 'It's a great fit for an academic setting with children-you might really enjoy it.' \
    You could also highlight occasions like celebrations, with phrases like 'This spot would be
perfect for celebration.'\
    If no connection exists, simply omit this step.\n'''

    prompt += f"""
    **Dialogue progress:**
    {dialogue_progress}
    """
    if dialogue_progress == "Middle of the dialogue":
        prompt += "Don't say celebration, thank you, or goodbye. In the middle of the conversation, it is not
natural."

    prompt += """

    **Objective:**

    Modify the current system utterance (system1) so that it matches the style described in the user image
description.
    Don't use 'craving' 'kindly' 'certainly', 'sure thing', 'hey there', 'hey and 'vibe' It is not natural.
    Keep in the information center staff role.
    Your Answer (no description needed):
    """

```

Prompt for Dialogue Accuracy Quality

```

prompt = "You are the proficient dialogue quality evaluator. Please evaluate the dialogue quality of the
following dialogue. "
prompt += "You will be given a two dialogue sets. The first one in original dialogue and the second one is
dialogue style transferred version. "
prompt += f"In the restyled version, the user's utterance is modified to reflect how they would say it
based on their first impression: {first_impression}."
prompt += "System utterance is changed to give personalized response to user, in terms of user's first
impression"
prompt += "You will also be given the dialogue actions for both user and system, which is the direction
user and system should follow."
prompt += "In the original dialogue, user and systems followed the action well."

prompt += "This is the original dialogue: \n"
for i in range(len(user)):
    prompt += f"Turn {i+1}\n"
    prompt += f"User: {user[i]}, UserAction {user_info[i]}\n"
    prompt += f"System: {sys[i]}, SystemAction {sys_info[i]}\n"

prompt += "This is the dialogue style transferred version."
for i in range(len(st_user)):
    prompt += f"Turn {i+1}\n"
    prompt += f"Transferred User: {st_user[i]}, UserAction {user_info[i]}\n"
    prompt += f"Transferred System: {st_sys[i]}, SystemAction {sys_info[i]}\n"

prompt += "Please evaluate the dialogue quality in two aspects: "
prompt += "1. User's dialogue quality : Does the transferred user dialogue follow the action well?"
prompt += "2. System's dialogue quality : Does the transferred system dialogue follow the action well?"

prompt += "Additionally, transferred systems sometimes provide personalized recommendation using the DB
results. Don not consider the DB results in the evaluation."
prompt += "Additionally, changes the booking time, such as 5:45 PM to 5:30 PM or 6PM should not be
considered as a failure."

```



```

prompt += "If there is any issue in the dialogue, please report it."
prompt += "Format of the report: \n"
prompt += "User's dialogue quality: <pass/fail>, System's dialogue quality: <pass/fail>,Reason:
<reason>"
prompt += "for example, User's dialogue quality: fail, System's dialogue quality: pass, Reason:
User's dialogue is not following the action"
prompt += "or User's dialogue quality: pass, System's dialogue quality: fail, Reason: System's
dialogue is not following the action"
prompt += "or User's dialogue quality: pass, System's dialogue quality: pass, Reason: transferred
dialogue contains all information as in original dialogue"
prompt += "Now, please evaluate the dialogue quality of the transferred dialogue."

```

Prompt for Dialogue Overall Quality

```

prompt = "You are the proficient dialogue quality evaluator. Please evaluate the dialogue quality of
the following dialogue. "
prompt += "You will be given a synthesized dialogue sets."
prompt += f"In this dialogue, the user's utterance is synthesized to reflect how they would say it
based on their first impression: {first_impression}."
prompt += "System utterance is synthesized to give personalized response to user, in terms of user's
first impression \n"

for i in range(len(st_user)):
    prompt += f"Turn {i+1}\n"
    prompt += f"User: {st_user[i]}\n"
    prompt += f"System: {st_sys[i]}\n"

prompt += "Please evaluate the quality of the dialogue's in two criteria\n"
prompt += "1. Flow: Does the dialogue flow as smoothly? Does it sound natural?"
prompt += "2. Logical: Does the dialogue and system response make sense in the context of the
conversation?"

prompt += 'Additionally, greetings and ending words can be some what overly sentimental over
personalized. However do not consider the greetings in the evaluation. It is intended to make the
dialogue more personalized.'
prompt += "Additionally, changes the booking time slightly, such as 5:45 PM to 5:30 PM or 6PM should
not be considered as a failure."

prompt += "If there is any issue in the dialogue, please report it."
prompt += "Format of the report: \n"

prompt += "Flow: <pass/fail>, Logical: <pass/fail> Reason: <reason>"
prompt += "for example, Flow: fail, Logical: pass, Reason: System's dialogue is too rude for the
user, in terms of user's first impression"
prompt += "for example, Flow: fail, Logical: pass, Reason: System's dialogue is too verbose and
gives too much information which makes the dialogue unnatural"
prompt += "for example, Flow: pass, Logical: fail, Reason: System's response is not logical or
coherent, as the answer is not related to the user's query"
prompt += "for example, Flow: pass, Logical: pass, Reason: Transferred dialogue contains all
information as in original dialogue, and flows naturally"
prompt += "Now, please evaluate the dialogue quality of the transferred dialogue."

```

Prompt for Dialogue Quality Test. (Section 4.1 and Section 6.2)

```

prompt = f"""
You are the proficient dialogue quality assessment. You are given a two dialogue a user and a system.
First one is the original dialogue and the second one is the paraphrased dialogue, to match the style
described in the user image description.

Please check the dialogue in five perspectives.
1) Dose the paraphrase user utterance is well matched to user description?
2) Dose the user paraphrased user utterance is semantically equivalent to the original user utterance ?
3) Dose the paraphrase system utterance is well personalized (style, tone, formality) to user description?
- 1: Not at all (The sentence paraphrased system utterance is not personalized with specific words, phrases, or
style to user description)
- 2: A little (changes formality or tone for according to user description (Please tell your plan ->
Could I know your plan?, Not specifically for user in description)
- 3: Somewhat (changes style, tone, formality, greeting words, etc. to user description, ex, Nice red
hat! or Your smile is beautiful!)

```

```

- 4: A lot (The paraphrase system utterance is well personalized with specific words, phrases, or style
to user description)
4) Does the system paraphrased system utterance is semantically equivalent to the original system
utterance?
5) Does the system's utterance enhance the overall user experience, compared to system_reference?

assess the dataset in four scales.
1) 1: Not at all
2) 2: A little
3) 3: Somewhat
4) 4: A lot

**Original Dialogue**
{dial1}
**User Image Description:**
{user_impression}

**Changed Dialogue:**
{dial2}

Your answer must be in the following format/. Below is just an example, not the actual answer.:
Score : (Q1:3, Q2:4, Q3:2, Q4:4, Q5:3)
Score : (Q1:3, Q2:3, Q3:4, Q4:2, Q5:4)
Score : (Q1:2, Q2:4, Q3:3, Q4:2, Q5:2)
Now your turn to make your own answer with brief reason.
"""

```

Prompt for Dialogue Personalization and Paraphrase Evaluation. (Section 6.1)

```

prompt = f"""
You are the proficient dialogue system quality assessment. You are given a two dialogue system.
Please evaluate the following two systems based on the personalization to the user image and image
description, in terms of personalized greetings, personalization to age, personalized recommendation,
emotion and formal context.

**User Image Description:**
{user_impression}
**dialogue**
{dial}

Your answer must be in the following format:
(Reason : [reason for selection], Winner :[System1, Tie, System2])
"""

```

Figure 8: 提示模板。