PicPersona-TOD : A Dataset for Personalizing Utterance Style in Task-Oriented Dialogue with Image Persona

Jihyun Lee[1], **Yejin Jeon**[1], **Seungyeon Seo**[1], **Gary Geunbae Lee**[1,2]
[1]Graduate School of Artificial Intelligence, POSTECH, Republic of Korea
[2]Department of Computer Science and Engineering, POSTECH, Republic of Korea
{jihyunlee, jeonyj0612, ssy319, gblee}@postech.ac.kr

Abstract

Task-Oriented Dialogue (TOD) systems are designed to fulfill user requests through natural language interactions, yet existing systems often produce generic, monotonic responses that lack individuality and fail to adapt to users' personal attributes. To address this, we introduce PicPersona-TOD, a novel dataset that incorporates user images as part of the persona, enabling personalized responses tailored to user-specific factors such as age or emotional context. This is facilitated by first impressions, dialogue policy-guided prompting, and the use of external knowledge to reduce hallucinations. Human evaluations confirm that our dataset enhances user experience, with personalized responses contributing to a more engaging interaction. Additionally, we introduce a new NLG model, Pictor, which not only personalizes responses, but also demonstrates robust performance across unseen domains. [1]

1 Introduction

Task-oriented dialogue (TOD) is one of the core tasks of dialogue systems, which is designed to fulfill user requests, such as assisting users at customer service desks (Rastogi et al., 2020) and tourist centers (Zang et al., 2020). A TOD system is typically divided into the following sub-modules: (1) dialogue state tracking (DST) for tracking the user's requests, (2) policy module for determining system actions such as database (DB) searches or dialogue terminations, and (3) natural language generation module (NLG) for converting dialogue policies and DB results into natural language responses (Young et al., 2013). Among these components, the responses generated by the NLG module are used to directly interact with the user; therefore, NLG responses significantly influence the overall user experience.

Although extensive research has improved system responses (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), the primary focus has remained on enhancing the accuracy of information conveyance. As a result, the style of the generated response is monotonic and lacks individuality, which hinders the system from forming age-appropriate and emotionally resonant connections with users (McLean et al., 2021).

Figure 1: Example of PicPersona-TOD: Unlike existing TOD datasets (in grey), which lack user personas and personalization, PicPersona-TOD uses user images to generate tailored responses.

In an effort to improve the naturalness and engagement of system responses, recent approaches have curated new TOD datasets that support personalized response styles by incorporating user personas into the dialogue. For instance, Joshi et al. (2017) included age and gender information in the dialogues, Lin et al. (2023) integrated emotion, and Liu et al. (2024b) personalized system responses by mirroring users' noun and verb phrases. Although these approaches provide personalized responses to some degree, the persona modality has been limited to textual information, which lacks details and concurrency about the users they interact with.

Meanwhile, in the field of open dialogue systems, the integration of user personas has been a long-standing point of interest (Zhang et al., 2018a; Agrawal et al., 2023; Kim et al., 2024; Qian et al.,

[1]https://github.com/JihyunLee1/PicPersona

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 7937–7958
April 29 - May 4, 2025

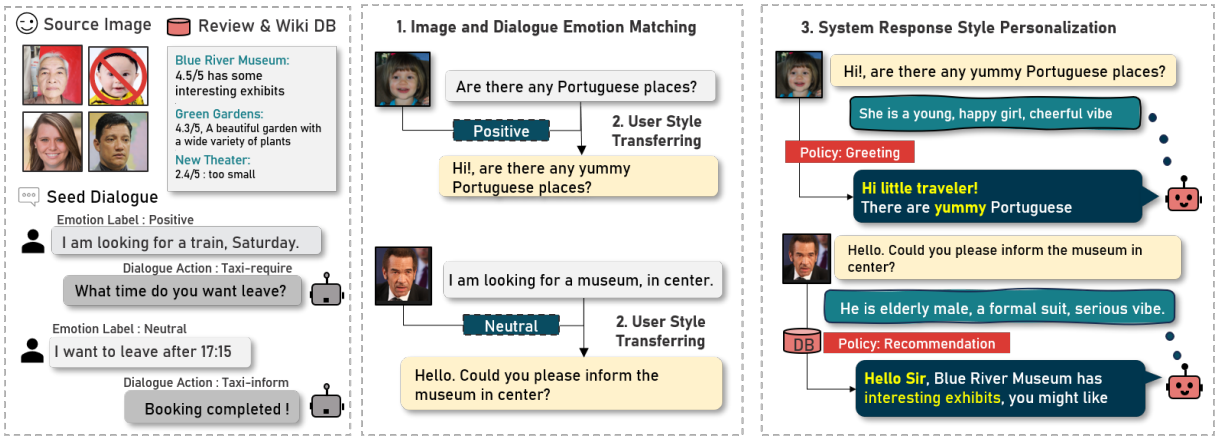


Figure 2: An overview of the automatic pipeline for generating PicPersona-TOD dataset.

2017; Zheng et al., 2019; Shen et al., 2024), with recent advancements highlighting the potential of the *visionary persona* approach (Ahn et al., 2023; Lee et al., 2024b). A visionary persona can capture subtle facial expressions and provide a rich understanding of the user's context, which is similar to how people interpret visual and non-verbal cues in real-life interactions. This approach is particularly valuable in situations where prior textual user profiles are unavailable, such as first-time interactions where no historical information is present. Despite these advantages, visionary personas have primarily been utilized in chit-chat (Poria et al., 2018; Firdaus et al., 2020) or counseling scenarios (Valstar et al., 2016; Li et al., 2023b) and have not yet been explored in current TOD systems.

Taking these considerations into account, we introduce a new TOD dataset that incorporates realistic user images as part of the user persona, enabling personalized system responses in terms of greetings, formality, age sensitivity, and emotional awareness (Figure 1). In constructing PicPersona-TOD, we use the user's first impression and dialogue policy-guided prompts, which effectively distill the personalization capabilities of Large Language Model (LLM). Additionally, we incorporate external knowledge from Google Maps and Wikipedia to reduce hallucinations in personalized responses. Furthermore, we implement a meticulous filtering process to ensure stylistic appropriateness, semantic accuracy, and overall naturalness, resulting in a well-refined personalized TOD dataset. Despite the highly automated process, our dataset demonstrates higher human preference in both user experience (§ 4.1) and personalization (§ 4.2) compared to other datasets and methods. From a label alignment perspective, analysis with DST and policy modules show that PicPersona-TOD maintains information accuracy. Specifically, we present an NLG model called **Pictor**, which demonstrates the ability to generate robustness in personalization, even in unseen domains.

In summary, the contributions of this work are threefold: first, we introduce PicPersona-TOD, a novel TOD dataset that integrates user images into personas and provides personalized system responses. Second, we present a highly automated dataset generation framework that efficiently creates realistic and personalized datasets. Third, through human evaluation, we demonstrate that our dataset enhances user experience through personalization, with benchmark results confirming that personalization does not compromise performance in other critical tasks.

2 PicPersona-TOD Dataset

In this section, we introduce **PicPersona-TOD**, the first personalized TOD dataset based on user image persona. To construct a high-quality personalized TOD dataset, we hypothesize that it should meet three criteria: (1) the user's utterances should be consistent with their image, (2) the system's responses should be appropriately personalized to the user image, and (3) the synthesized dataset should align with the sub-task labels of TOD tasks, such as DST and dialogue policy prediction, while maintaining the information. To address these criteria, our dataset construction pipeline comprises five key stages: (1) user image collection and dialogue dataset extension, (2) user image and utterance alignment, (3) user utterance style transfer, (4) system response personalization, and (5) data filtering.

For data construction, we mainly employed GPT-4o (Achiam et al., 2023) as the primary language model. The overall process is illustrated in Figure 2 and the used prompts are in Appendix I.

2.1 Collecting Images and Extending Dialogue

Initially, we select suitable user images that convey sufficient persona information. In order to effectively represent a user persona, each image should be a single person who is positioned in the center of the image, and close enough so that facial detail and clothing information are clearly visible. Based on these criteria, we selected the Flickr-Faces-HQ (Karras et al., 2019) as an image source, and made sure to exclude toddlers, as they are too young to engage in TOD interactions. After collecting the data, we used LLM to extract additional metadata for each image, including estimated age, gender, and formality.

For the dialogue dataset, we combined the MultiWOZ-2.2 (Zang et al., 2020) and SGD (Rastogi et al., 2020) datasets, which include 8,438 and 11,398 dialogues containing 18 service domains in total[2]. Additionally, since the movies, restaurants, hotels, and attractions in the dataset exist in the real world, we extended the dataset by collecting Google Maps reviews and Wikipedia entries for each location. Specifically, we added 2,474 sentences from 342 Wikipedia entries and 3,483 reviews from 406 locations on Google Maps. These results were added to the database to reduce hallucination in personalization.

2.2 Image and Dialogue Data Alignment

After selecting the image and dialogue datasets, we conducted user image and utterance alignment. Since the dialogue datasets lack details like age or gender, we chose emotion as a common attribute, as it is consistently present in both images and dialogues. We used a fine-tuned emotion classification model[3] to classify emotions in dialogues and prompted an LLM to classify images (positive, neutral, negative), and created image-dialogue pairs if they had the same predicted emotion label. The distribution of emotions across the dataset was 50.92% for positive, 52.44% for neutral, and 0.55% for negative.

[2]For the generalization test, we excluded the bus, home, and movie domains.

[3]Fine-tuned model from Hugging Face, `cardiffnlp/twitter-roberta-base-sentiment-latest`

2.3 Alignment User Utterance Style to Image

Next, we adapted the user utterances style to more closely align with the corresponding user images. We performed style transfer by prompting with user images, considering factors such as age, gender, emotion, and contextual cues. Formally, for each $i-th$ dialogue $D_i = (u_0, s_0, u_1, s_1 ... u_T, s_T)$, and its associated image Img_i, we generate a revised utterance $\tilde{u}_t$ at turn t; $\tilde{u}_t = \text{LLM}(s_{t-1}, u_t, \text{Img}_i)$ where T is the total number of turns, u is user utterance and s is system utterance in the dialogue.

2.4 System Response Style Personalization

In personalizing the system's responses, we categorized the process into three types and guided the prompts using first impressions and dialogue policy. 'Basic Personalization' was applied in most cases, while 'Greeting Personalization' was used for dialogue actions involving greetings. For recommendation-related actions, we implemented 'Recommendation Personalization', to provide less hallucinated suggestions.

Basic Personalization In order to generate personalized system utterances that align with the user's image, we use the *first impression* as a guide for the prompt. This process draws inspiration from human cognitive mechanisms in communication, which consist of two key steps. First, humans unconsciously infer a first impression of others within milliseconds (Borkenau et al., 2009; Willis and Todorov, 2006), and then adjust their communication tone and style to align with this impression, using it as an inferred persona (Rule and Ambady, 2008). Similar to this process, we first generate an impression from the user's image (Img_i) and then generate personalized system's utterances in terms of formality, age sensitivity, and emotional context, based on the inferred persona. Specifically, for a given dialogue D_i, Img_i and inferred Imp_i, the personalized system utterance $\tilde{s}_t$ is generated as follows; $\tilde{s}_t = \text{LLM}(s_t, \tilde{u}_t, \text{Img}_i, \text{Imp}_i)$.

Greeting Personalization Since greetings and closing remarks play a crucial role in creating personal interactions between speakers (McLean et al., 2021; Glas et al., 2017), we specifically tailor the system's greetings and farewells to provide a more engaging and personalized experience. This is achieved by prompting LLM to incorporate specific comments about the user's appearance, such as mentioning a distinctive feature of their outfit (e.g., "Nice red hat!" or "Congratulations on your

achievement") (Figure 2, Top-Right).

Recommendation Personalization TOD systems often make recommendations, such as "`How about [location]?`". In our preliminary experiments, we observed that when the model attempted to personalize these recommendations, it sometimes introduced hallucinated information (e.g., "`The [location] currently has a festival you might like`"). To mitigate these hallucinations, we performed retrieval-augmented generation (Lewis et al., 2020) by enhancing the prompt with authentic information gathered from online sources. Specifically, from the reviews about `[location]` in the database (DB) (§ 2.1), we retrieve the three reviews with the highest cosine similarity to $\tilde{u_t}$, by embedding them using Sentence-BERT(Reimers, 2019). These were incorporated into the prompt to guide the model in generating factually grounded responses(Figure 2, Bottom-Right). After these process we get the personalized TOD dialogue $\tilde{D}_i = \{\tilde{s}_{0:T}, \tilde{u}_{0:T}, \text{Img}_i, \text{Imp}_i\}$.

2.5 Dataset Quality Control by Filtering

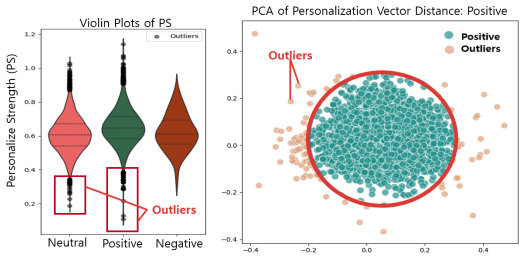


Figure 3: Visualizations of personalization strength and personalization direction filtering processes are shown on the left and right, respectively.

Despite the meticulous construction process in the previous subsections, some dialogues may still generate inappropriate utterance style, contain semantic inconsistencies, or lack overall naturalness. To address these potential issues and enhance the overall quality of the dataset, we implement several filtering processes.

Style Strength Filtering Dialogues are eliminated if the degree of personalization is too low (Figure 3, left). To do this, we calculate the strength of personalization for each dialogue i by defining the Personalization Strength (PS_i) as the average distance between the original system response $E(s_t)$ and the personalized response $E(\tilde{s}_t)$: $\text{PS}_i = \frac{1}{T}\sum_{t=0}^{T} \text{Dist}\left(E(\tilde{s}_t), E(s_t)\right)$, where Dist represents the Euclidean distance, E represents embedding with Sentence-BERT. Next, we collect the PS values for each metadata class (e.g., young, senior) and remove dialogues with PS values below the threshold, defined as less than $2.5 \times$ IQR (interquartile range). As a result, 1.49% of the dataset was filtered out.

Style Direction Filtering We remove outliers that have a different direction of personalization within the same metadata class (Figure 3, right). We calculate the Personalization Vector (PV) for each dialogue i as $\text{PV}_i = \frac{1}{T}\sum_{t=0}^{T}\left(E(\tilde{s}_t) - E(s_t)\right)$. Then, we compute the mean personalization vector for each metadata class, PV_{class}, by averaging the PV vectors within that class. To detect outliers, we calculate the distance (PD_i) between the class mean and the personalization vector of each dialogue : $\text{PD}_i = \text{Dist}(\text{PV}_{class}, \text{PV}_i)$. We define outliers as those with an exceptionally large distance from the mean of the class style vector. We set the threshold as $4.5 \times$ IQR, resulting in the removal of 1.98% of the dataset.

Semantic Filtering We filter out semantically misaligned user and system utterances by comparing them with the corresponding DST and dialogue policy labels. For user utterances, we check their alignment with the DST labels. For example, if the label is `hotel-east, restaurant-expensive`, the user's utterance should reflect this, such as by saying, "I need a hotel in the east and an expensive restaurant." Similarly, we verify that system responses align with the dialogue policy label. Semantically misaligned data were filtered by prompting the LLM to check for inconsistencies, resulting in the removal of 2.37% of the dataset.

Overall Naturalness Filtering Lastly, we filter out dialogues that do not exhibit naturalness. Since the system and user utterances are generated turn by turn, some parts of the dialogue may not flow naturally. To remove such unnatural instances, we provide the entire dialogue to the LLM to assess its flow. As a result, we remove 4.39% of the dialogues. After these individual filtering stages, 92.59% of the initial dataset is retained.

2.6 Case Study for Filtering

Figure 4 shows examples of filtered-out results for the style strength filter (left) and the style direction filter (right). The style strength filter removes instances with minimal or unchanged personalization, while the style direction filter excludes cases where personalization led to inappropriate changes,

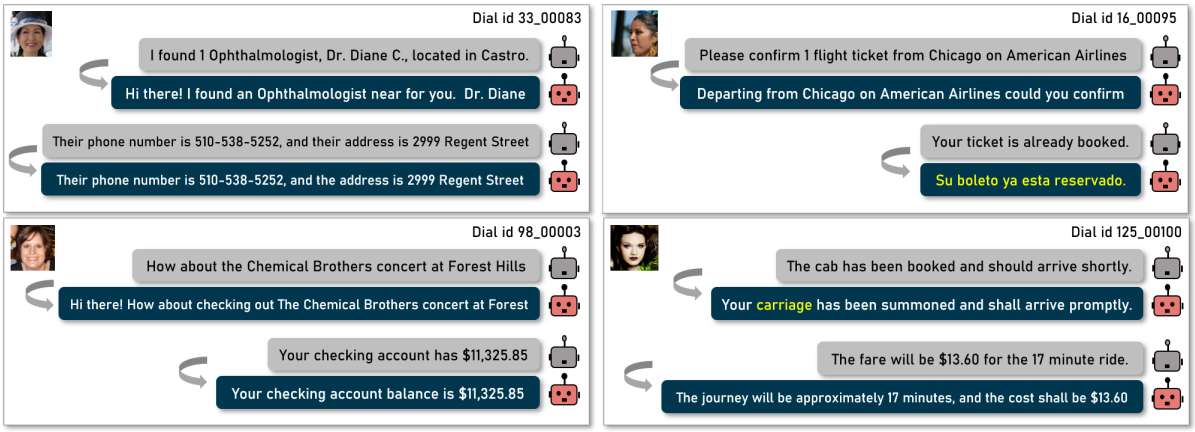


Figure 4: Examples of filtered-out results: style strength filtering (left) and style direction filtering (right).

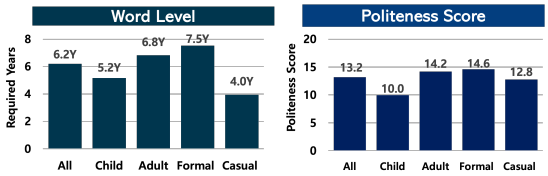


Figure 5: Lexical analysis of PicPersona-TOD. "Word level" refers to the required years of education needed to understand, while the politeness score represents the average use of politeness strategies.

such as medieval-style language or switching to a different language.

3 PicPersona-TOD Analysis

In this section, we analyze our dataset in comparison to other datasets(§ 3.1), followed by an analysis across key dimensions such as word difficulty, politeness (§ 3.2), and emotions (§ 3.3).

3.1 Comparison with Existing Datasets

In Table 1, we compare PicPersona-TOD with other datasets in terms of personalization and dialogue data modality. As shown in the results, our dataset holds a unique position as the only TOD dataset that incorporates a visionary persona. Additionally, by integrating multiple TOD datasets, it covers a wide range of services with a substantial amount of dialogue data. We also enhance personalization through the incorporation of external sources, such as Google Reviews and Wiki information, which constitutes a notable aspect of our dataset.

3.2 Word Complexity & Politeness

Word Complexity We analyzed the complexity of system responses across different user scenarios, including interactions with children, adults, and within formal and informal contexts (Figure 5). To assess lexical difficulty, we evaluate the number of years of education required to comprehend the system utterance[4]. On average, the system's responses require 6.2 years of education to be understood. Responses for children require 5.2 years of education, while those for adults require 6.17 years. This indicates that the system effectively personalized its word choice based on the user's age.

Politeness Politeness of responses is assessed by measuring the average number of politeness strategies used per sentence, where higher scores indicate greater politeness[5]. Overall, PicPersona-TOD achieved a politeness score of 13.2. In formal contexts, the score rises to 14.6, while in casual contexts it drops to 12.8. For children, the politeness score decreases to 10.0, while for adults, it increases to 14.2. This demonstrates the system's ability to adjust its politeness based on the user's context.

3.3 Emotion Awareness in Responses

Table 2 compares the distribution of emotions identified in the system responses across the MultiWoZ and PicPersona-TOD datasets, categorized into 27 emotions using GoEmotions' taxonomy (Demszky et al., 2020) with GPT-4 as the classifier. While the neutral emotion is most common in both datasets, PicPersona-TOD exhibits a significantly lower proportion of neutral responses (61.50%) compared to MultiWoZ (74.97%), which indicates a more emotive reaction. Furthermore, we analyze how

[4]We used the Gunning Fog Scale from the Textstat Python library https://pypi.org/project/textstat/.

[5]We used the PolitenessStrategies library from ConvoKit https://convokit.cornell.edu/, and measure the average number of strategy.

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., 2024b)	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: Comparison of statistics with other datasets in terms of personalization and modality. 'Mod.', 'Serv.', and 'Pol.' stand for modality, service, and policy prediction, respectively.

the system personalizes its responses based on the user's emotional state (third column of Table 2). We found that when the user's image has positive emotions, the system responds with a broader range of emotions, such as joy and gratitude. In contrast, when the user's emotions are neutral or negative, the system tends to generate more neutral responses, with a greater emphasis on empathy and care.

MultiWoZ		PicPersona-TOD		PicPersona-TOD (pos)	
neutral	74.97	neutral	61.50	neutral	44.66
curiosity	9.18	curiosity	11.79	approval	16.51
gratitude	6.31	approval	8.65	curiosity	11.65
approval	3.56	gratitude	6.52	gratitude	7.77
optimism	1.26	annoyance	2.26	excitement	4.86
apology	0.92	optimism	1.88	admiration	2.43
annoyance	0.69	excitement	1.88	PicPersona-TOD (neu/neg)	
confusion	0.57	admiration	1.00	neutral	66.69
caring	0.57	amusement	0.63	curiosity	11.84
disappointed	0.57	disappointed	0.63	gratitude	6.10
joy	0.35	caring	0.51	approval	5.90
excitement	0.35	joy	0.38	caring	2.54
admiration	0.23	confusion	0.38	optimism	1.36

Table 2: The proportion (%) of the most frequent emotions in system responses within the test dialogue.

4 Human Evaluation

4.1 Evaluation for Quality

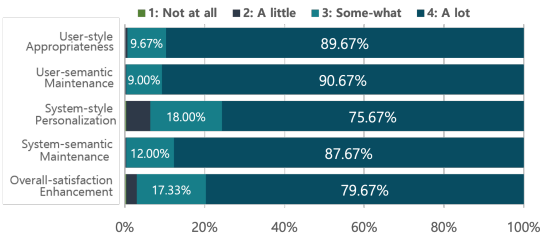


Figure 6: Human evaluation results for dataset quality using a 4-point Likert scale.

To assess user satisfaction with personalized results, we conducted human evaluation that focused on verifying the degree of personalized style and information retention. Three evaluators rated 100 randomly selected dialogues on a 4-point Likert scale and measured the scores for both user and system utterances using five questions (Appendix G). The results were highly positive, with average scores of 3.89 for user style appropriateness, 3.90 for user semantic consistency, 3.69 for system style personalization, 3.87 for system semantic consistency, and 3.76 for overall user satisfaction (Figure 6). Additionally, we observed a strong inter-rater agreement of 0.85, measured using Krippendorff's Alpha. These results confirm that the PicPersona-TOD meets our predefined criteria: (1) user and image consistency, (2) personalized system responses, and (3) maintaining the original information. We also performed the same evaluation using GPT-4, which showed a high inter-rater correlation with human evaluators, achieving a score of 0.84 (Appendix G.1).

4.2 Other Personalization Methods

We conducted a comparative evaluation of PicPersona-TOD against two other personalized methods. Since no method currently exists for incorporating visual persona into TOD, we compared PicPersona-TOD with methods that rely on textual modalities. The first baseline, Liu et al. (2024b), personalizes dialogue by mirroring the user's noun and verb phrases, while the second method, Joshi et al. (2017), personalizes interactions based on age and gender information. We sampled 120 dialogues from various scenarios and had three human judges evaluate them to determine the superior method based on personalization quality (Appendix G.2). As shown in Table 3, PicPersona-TOD consistently outperformed the text-based methods across diverse user scenarios. This result emphasizes the importance of rich, concurrent image personas for personalization, compared to relying on textual personas.

Figure 7 includes the distribution of key factors that influence the evaluators' preferences related to Table 3. In all user scenarios, appropriate formality is the most noticeable aspect of personalization

	All	Age			Emotion	
		Senior	Adult	Child	Pos	Neu&Neg
Liu et al. (2024b)	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: The winning ratio (%) when comparing PicPersona-TOD with personalization methods.

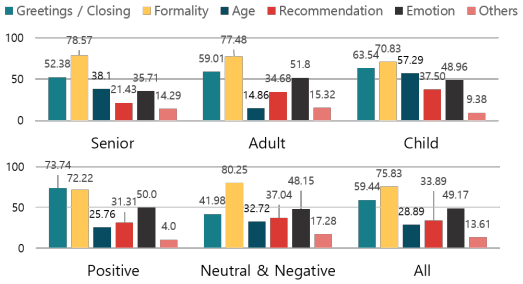


Figure 7: The winning characteristics analysis across ages and emotions.

for human evaluators. For children and in positive contexts, tailored greetings enhance the sense of personalization, while in neutral and negative environments, emotional awareness serves as a key factor in creating a personalized experience.

5 Baselines

In this section, we introduce **Pictor**, an NLG baseline trained on the PicPersona-TOD dataset to generate personalized responses. Additionally, we provide models for DST and policy prediction, allowing for comparisons with other datasets.

5.1 Baseline for NLG

Using the PicPersona-TOD dataset, we developed a multimodal TOD response generation model named **Pictor** (Figure 8). Pictor generates personalized responses ($\tilde{s}_i$) by leveraging user images and dialogue context. Specifically, the input to the Pictor model for turn t in D_i includes the turn progress

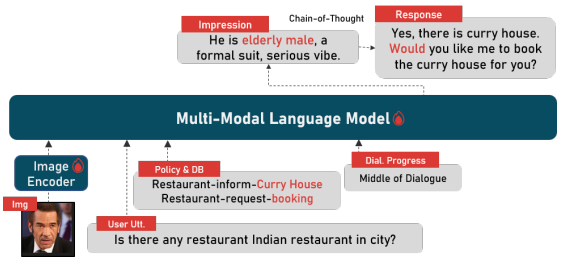


Figure 8: Overview of the proposed model Pictor.

(e.g., start, middle, end), user's utterance ($\tilde{u}_t$), dialogue policy (pol_t), DB results (DB_t), and the user's image (Img_i). Similar to the process used in constructing the PicPersona-TOD dataset, we first generate the user's impression and then, based on this impression, generate a personalized response $\tilde{s}_i$. Pictor is based on the LLaVA 7B (Li et al., 2023a) and 1.5B (Zhou et al., 2024) models, which are known for strong performance across various vision-language tasks. We train the Pictor model by utilizing a LoRA (Hu et al., 2021) adapter with rank of 16. Detailed information is in Appendix D.

5.2 Baseline for DST and Policy

The PicPersona-TOD dataset also supports a range of TOD tasks, including DST and policy prediction. To establish baselines for these tasks, we utilized the PPTOD model (Su et al., 2021) and trained the DST and policy prediction models using both T5-base and T5-small variants (Raffel et al., 2020). For the DST task, the input at turn t is defined as the concatenation of all user and system utterances up to and including turn t, which can be expressed as $\text{Input}_{\text{DST},t} = [\tilde{u}_1, \tilde{s}_1, \tilde{u}_2, \tilde{s}_2, \ldots, \tilde{u}_t]$. The output of the DST is represented as slot-value pairs (e.g., `hotel-name: Green Hotel`).

For the policy prediction task, the input at turn t is similarly constructed, but with the addition of the predicted dialogue state; $\text{Input}_{\text{POL},t} = [\tilde{u}_1, \tilde{s}_1, \tilde{u}_2, \tilde{s}_2, \ldots, \tilde{u}_t, \text{DST}_t]$. The policy prediction model generates the appropriate system action, such as a request for further information (e.g., `Request-restaurant-foodtype`).

6 Baseline Evaluation

6.1 Comparison with Other LLMs

We conducted a comprehensive comparison of our Pictor 7B model with several prominent vision-LLMs, including Llama3-8b(Dubey et al., 2024) LLaVA 7B, InstructBLIP 7B (Liu et al., 2024a), and GPT-4o-mini (OpenAI). For the evaluation, we sampled 100 dialogues, with assessments performed by the GPT-4 model. As illustrated in Figure 9, Pictor consistently outperforms the other sLLM models in terms of personalization quality across categories. In comparison to GPT-4o-mini, which likely has more parameters than Pictor, Pictor demonstrates better results, except in neutral/negative cases. These findings highlight the importance of datasets specifically designed for personalization, such as PicPersona-TOD, to achieve

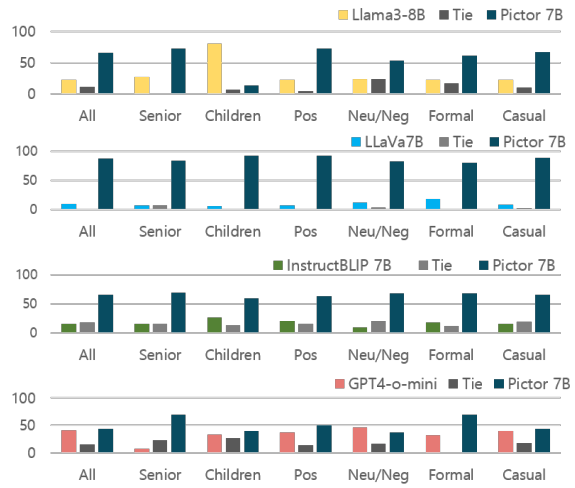


Figure 9: Performance comparison of LLama 3-8B, Pictor 7B with LLaVA 7B, InstrucBLIP 7B, and GPT-4o mini across various user scenarios.

optimal performance in personalized scenarios.

6.2 Generalization Performance Evaluation

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: Human evaluation for of the Pictor in unseen domains.

In Table 4, we perform the human evaluation on the generalization performance of the Pictor model across the Bus, Movie, and Home domains from the SGD dataset, which were not included in the model's training data. Note that these datasets were constructed using the same process as the dialogues in the PicPersona dataset, which also includes user images. We sampled 100 dialogues from each domain and performed zero-shot inference, followed by human evaluations conducted by three annotators using a 4-point Likert scale (Appendix G.1). The results demonstrate that, despite these domains not being part of Pictor's training set, the model is able to achieve strong personalization with strong user satisfaction close to 4. We attribute this to the inclusion of two large-scale TOD datasets as a dataset source, which cover a wide range of domains.

6.3 Ablation Study on Pictor

We conducted an ablation study to examine how different components affect the generation performance of Pictor. We evaluated the model using BLEU scores[6] and GPT-4 assessments[7] for semantic accuracy, style, and overall satisfaction (Table 5). Results reveal an interesting finding about impressions; while omitting to generate impressions resulted in higher BLEU scores, incorporating impressions significantly improved personalization. This trend was even more pronounced in larger models, with the style score increasing from 3.48 to 3.51 when impressions were included.

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

Table 5: Ablation study for training Pictor model. Each experiment was repeated three times, and the results were averaged.

6.4 DST and Policy Inference Results

Dataset	DST						Policy
	JGA	Rest.	Hotel	Att.	Train	Taxi	Entity-F1
T5-small							
Mwoz	47.17	83.0	79.6	86.4	88.7	94.5	46.18
PicPersona-TOD	49.18	83.8	80.4	88.3	88.9	96.3	41.26
Difference	△ 2.01	△ 0.8	△ 0.8	△ 1.9	△ 0.2	△ 1.8	△ 4.92
T5-base							
Mwoz	49.81	85.9	79.5	88.1	88.3	94.7	44.19
PicPersona-TOD	47.55	84.8	79.5	87.7	86.5	96.0	46.57
Difference	△ 2.26	△ 1.1	△ 0.0	△ 0.4	△ 1.8	△ 1.3	△ 2.38

Table 6: Result of DST and Policy Inference Tasks. Metric details are in Appendix E.

We conducted experiments to evaluate the information accuracy of the PicPersona-TOD dataset by testing DST and policy models using T5-small and T5-base. For comparison, we also present results using the MultiWOZ dataset. Table 6 shows that our dataset yields comparable performance to MultiWOZ across most metrics, with only minor differences observed. This consistency suggests that despite the added complexity of our dataset by personalized the user and system, PicPersona-TOD maintain information accuracy on par with human-curated datasets.

7 Related Works

Advancements in TOD Datasets Task-oriented dialogue (TOD) systems have long been a fo-

[6] nltk.translate.bleu_score

[7] Note that the GPT4 scores use the same question formats detailed in § 4.1.

cus of research, with early datasets like ATIS (Hemphill et al., 1990), WOZ2.0 (Wen et al., 2016), and DSTC2 (Henderson et al., 2014) limited to single domains. Later, multi-domain datasets such as M2M (Shah et al., 2018), MultiWOZ (Budzianowski et al., 2018), SGD (Rastogi et al., 2020), and ABCD (Chen et al., 2021) aimed to improve accuracy but often overlooked user satisfaction. Recent work has sought to enhance user experience by integrating chitchat (Sun et al., 2020; Young et al., 2022; Stricker and Paroubek, 2024), providing detailed explanations in system responses (Kim et al., 2023; Qian et al., 2021), and considering users' emotional states (Abolghasemi et al., 2024; Feng et al., 2024), though few studies address individual personalization. The most relevant work to our work involves personalization efforts, such as incorporating age and gender (Joshi et al., 2017) or linguistic patterns (Liu et al., 2024b). While some approaches include emotional states (Lin et al., 2023; Feng et al., 2024), our method introduces a visionary persona that provides richer and more concurrent information, leading to enhanced user satisfaction.

Integrating Persona into Dialogue While personalized dialogue systems have been widely researched to improve user experience, they have traditionally relied on textual information. Methods include constructing personas through narrative sentences (Zhang et al., 2018b; Zhong et al., 2020), key-value pair dictionaries (Qian et al., 2017; Zheng et al., 2019), or users' review histories (Kim et al., 2024). Recently, multimodal approaches have emerged, incorporating user images to create richer personas (Ahn et al., 2023; Lee et al., 2024b; Agrawal et al., 2023). Building on these advances, we introduce a novel method that uses user images as the primary basis for personas in TOD datasets, enabling more contextually appropriate and personalized responses.

Data Generation by Distillation LLM Collecting dialogue data is challenging due to privacy concerns, high costs, and the need for multiple participants. To address this, many works have leveraged LLMs for dataset creation. Examples include compiling seed dialogues (Ahn et al., 2023; Kim et al., 2022b), constructing social event graphs (Kim et al., 2022a), and generating long-term dialogues (Jang et al., 2023). Others have used LLMs for commonsense-aware dialogues (Chae et al., 2023), prosocial dialogues (Kim et al., 2022b), and task-oriented dialogue (TOD) utterances (Kulkarni et al., 2024). LLM-generated datasets are cost-effective, diverse, and often preferred over human-curated datasets (Kim et al., 2022a; Lee et al., 2024b, 2021). Building on this, we use LLMs to generate personalized, privacy-conscious, and diverse user scenarios.

8 Conclusion

In this paper, we have introduced PicPersona-TOD, a novel dataset that personalizes system responses based on a user's visual persona in the TOD domain. Specifically, PicPersona-TOD incorporates personalized responses in terms of greetings, age, politeness, and emotions. Through user satisfaction experiments, we have demonstrated that PicPersona-TOD enhances personalization while retaining the original information. Additionally, we have proposed and analyzed a baseline model, which includes NLG (Pictor), DST, and policy prediction. Our results show that this method improves personalization without compromising performance in other critical tasks. We believe this work advances research on personalized TOD with multimodal user personas, enabling more natural and human-like interactions.

Limitations

Lack of Direct Benchmark Comparison Typically, datasets distilled from LLMs, such as those used for open dialogue (Kim et al., 2022a) or image-sending in chat (Lee et al., 2024a), are directly compared with traditional, human-created test sets to demonstrate the practical advantages of the new dataset. However, in our case, no traditional, human-made dataset exists for TOD that incorporates user personas, as creating a TOD dataset requires significantly more effort and higher labeling costs compared to open dialogue datasets.

Due to this environment, we were unable to perform direct comparisons with standard datasets. Instead, we evaluated our model's personalization capabilities against other prominent vision-LLM models in Section 6.1, which have been trained on large-scale vision-text datasets. We conducted GPT4 evaluations to assess performance in these comparisons, and the results show a strong preference for the model trained on our dataset. While this does not serve as a direct comparison with a traditional TOD benchmark, we believe it offers a valid alternative, as the results highlight the impor-

tance of datasets specifically designed for personalization, such as PicPersona-TOD, showing the advantages of our dataset.

Limitations in the Use of Other LLMs In our study, we conducted dataset curation exclusively using GPT-4o. This decision was based on preliminary experiments, where other vision-LLMs failed to generate personalized responses with the same quality as GPT-4o. However, as open-source vision-LLM models continue to improve, they may become viable options for developing high-quality, cost-efficient dataset-generation pipelines.

Facial Recognition Requirement Although our dataset shows strong potential for personalization, its full capability can only be realized in systems equipped with facial recognition technology, such as kiosks or robots. Without such equipment, the image-based personalization features of PicPersona-TOD cannot be effectively utilized.

Naive Application of Retrieved Results In the construction of our dataset, we integrated reviews and Wikipedia results as a retrieval-based generation method to enhance system responses. While this approach contributed to improving response quality, it lacked sophistication. Future research could focus on developing more advanced image persona-based retrieval methods, enabling deeper personalization and a more sophisticated understanding of the user's persona, which would ultimately lead to improved response quality.

Ethical Considerations

In constructing the dataset, we use two sources that involve real users and may raise concerns about privacy and consent. For the image data, we utilize the FFHQ dataset. According to the original paper (Karras et al., 2019), this dataset was created by crawling images from the Flickr website, where only images under permissive licenses were collected. Specifically, the license for these images is CC BY-NC-SA 2.0, which allows for free distribution as long as the use is non-commercial.

For the Google Map review data, we retrieve reviews using the Google Maps API. Consent from Google Maps users for collecting their data is provided through the Google API Terms of Use, which state that users understand their posts will be publicly available and can be accessed via the API. Additionally, to protect user privacy, we did not collect any personally identifiable information (PII) to ensure anonymity. From a consent perspective, we comply with the licensing terms for both sources and take steps to safeguard user privacy. Furthermore, the use of our data for commercial and for-profit purposes is restricted.

Acknowledgements

This work was supported by the MSIT (Ministry of Science and ICT), Korea, through the IITP (Institute for Information & Communications Technology Planning & Evaluation) grants: RS-2019-II191906 (Artificial Intelligence Graduate School Program at POSTECH, 50%) and RS-2024-00437866 (ITRC Program 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 Altenschmidt, 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. Mpchat: 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)*.

Hyungjoo 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*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

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 multimodal 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 Hyounghun 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 commonsense knowledge. *arXiv preprint arXiv:2407.03958*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, 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. M^3 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 Gasić. 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*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024a. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Yinhong Liu, Yimai Fang, David Vandyke, and Nigel Collier. 2024b. 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.

OpenAI. Gpt-4o mini: advancing cost-efficient intelligence. https://openai.com. Accessed: 2024-10-15.

Baolin Peng, Chunyuan Li, Jinchao Li, Shahin Shayandeh, Lars Liden, and Jianfeng Gao. 2020. Soloist: Few-shot task-oriented dialog with a single pretrained 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 Shayandeh, 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 Pandelea, 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 multimodal models. *arXiv preprint arXiv:2402.14289*.

Figure 10: Examples of greeting personalization (left) and recommendation personalization (right).

A Case Study

Figure 10 illustrates different methods of system response personalization (§ 2.4). For Greetings Personalization (left), the system personalizes greetings and farewells by incorporating specific characteristics from the user's first impressions. For Recommendation Personalization (right), the system retrieves information from reviews and tailors the response based on that information.

B Demographics

We analyze the distribution of PicPersona-TOD by using V-LLM to obtain information on gender, age, and formality from the user image. The gender distribution shows that 35.82% of the dataset consists of females, while males represent a slightly larger percentage at 47.71%. Regarding age distribution, the adult group constitutes the largest proportion at 74.24%. The proportion decreases with age, with users in the child group and the senior group making up 19.66%, and 6.68% of the dataset, respectively. Formality is divided into formal and casual groups, with the casual group accounting for the majority at 86.54% and the formal group making up 13.46%. The distribution of emotions across the dataset is 50.92% for positive, 52.44% for neutral, and 0.55% for negative. These results highlight the diverse demographic representation of the PicPersona-TOD dataset across gender, age, formality, and emotion.

C Experiments with GPT-4

We conducted two parallel evaluation tasks using GPT-4. In the Personalization Quality Evaluation task (§ 4.1), GPT-4 assigned the following scores: Q1: 3.79, Q2: 3.99, Q3: 3.66, Q4: 3.97, and Q5: 3.75, which achieves a high inter-rater reliability with a Krippendorff's alpha of 0.84.

D Baseline Training Details

We trained the Pictor model using both the LLaVA 1.5B and 7B models. For the 1.5B model, we employed LoRA with a rank of 16, an alpha value of 64, a batch size of 16, and a learning rate of 2e-5 over 5 epochs. In the case of the 7B model, LoRA was configured with a rank of 16, an alpha value of 32, a batch size of 16, and a learning rate of 5e-5, also for 3 epochs. Both models employed the Adam optimizer (Kingma and Ba, 2014) with no weight decay and utilized a cosine learning rate schedule with a 3% warmup ratio. All training for the Pictor model was conducted on an NVIDIA A100 GPU.

For the DST and policy models (T5-small and base variants), we used a batch size of 16, learning rate of 1e-3, and trained them for 10 epochs using the AdamW (Loshchilov and Hutter, 2017) optimizer with no weight decay. These models were trained on an NVIDIA A6000 GPU.

E Metric for DST and Policy Prediction

When evaluating DST performance, we used two metrics: Joint Goal Accuracy (JGA) and domain-specific JGA. JGA is considered correct if all dialogue states in a turn are accurate. Domain-specific JGA is marked as correct if the dialogue state for the targeted domain is accurate, regardless of other domains (Wu et al., 2019). For Dialogue Policy evaluation, we used Entity F1, which calculates the F1 score for each turn and then averages these scores across all turns (Eric and Manning, 2017; Shi et al., 2023).

F License

PicPersona-TOD is synthesized using the MultiWoZ 2.2, SGD, and FFHQ datasets. MultiWoZ 2.2 is released under an MIT license, while SGD

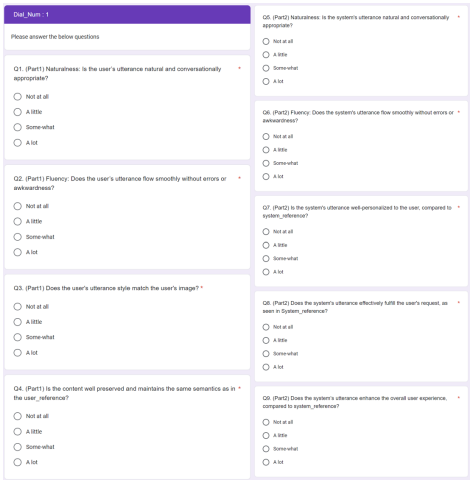


Figure 11: A screenshot of the Personalization Quality Evaluation.

is under a CC BY-SA 4.0 license and the images in the FFHQ dataset are licensed under Creative Commons BY 2.0, Creative Commons BY-NC 2.0, Public Domain Mark 1.0, Public Domain CC0 1.0, or U.S. Government Works licenses. These licenses permit free use, copying, modification, and publication for non-commercial purposes.

G Human Evaluation Details

For human evaluation, we hired three native English-speaking evaluators through the Upwork[8] platform. They were informed that all personal information would remain anonymous and that their submitted responses would be used solely for research purposes.

G.1 Section 4.1 and 6.2

The evaluators were asked to rate the quality of personalization in PicPersona-TOD or Pictor's utterances by selecting one of the provided *Options* in response to the following questions, which pertained to user utterances (U1–U4) and system utterances (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?
- **S3.** Is the system's utterance well-personalized to the user, compared to the original system utterance?

[8]www.upwork.com

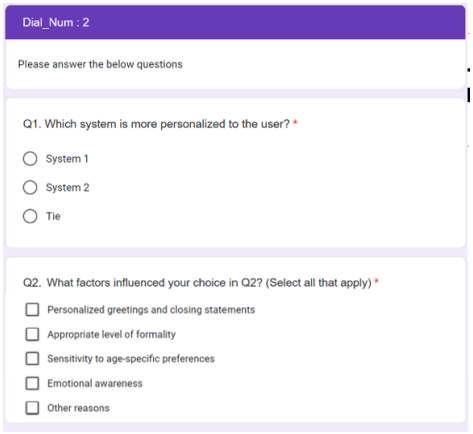


Figure 12: A screenshot of the Comparison with Other Personalization Methods.

- **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).

We used Google Forms for the evaluation, and Figure 11 shows the sample screenshot that the evaluators performed. The scores corresponding to each option were used to calculate the results.

G.2 For Section 4.2

The evaluators are asked to compare two anonymous systems, and indicate which part has been improved:

- **Question:** Which system is more personalized to the user?

  Options: System 1 / System 2 / Tie

- **Question:** Is the content well preserved and maintains the same semantics as in the original user utterance?

  Options:

  - **Personalized greetings and ending statements:** A personalized response could include customized greetings or closing remarks.
  - **Formality:** A personalized response should be appropriately formal or informal, depending on the situation.
  - **Age sensitivity:** A personalized response should be age-sensitive.
  - **Emotional awareness:** A personalized response should be emotionally aware.
  - **Other reasons**

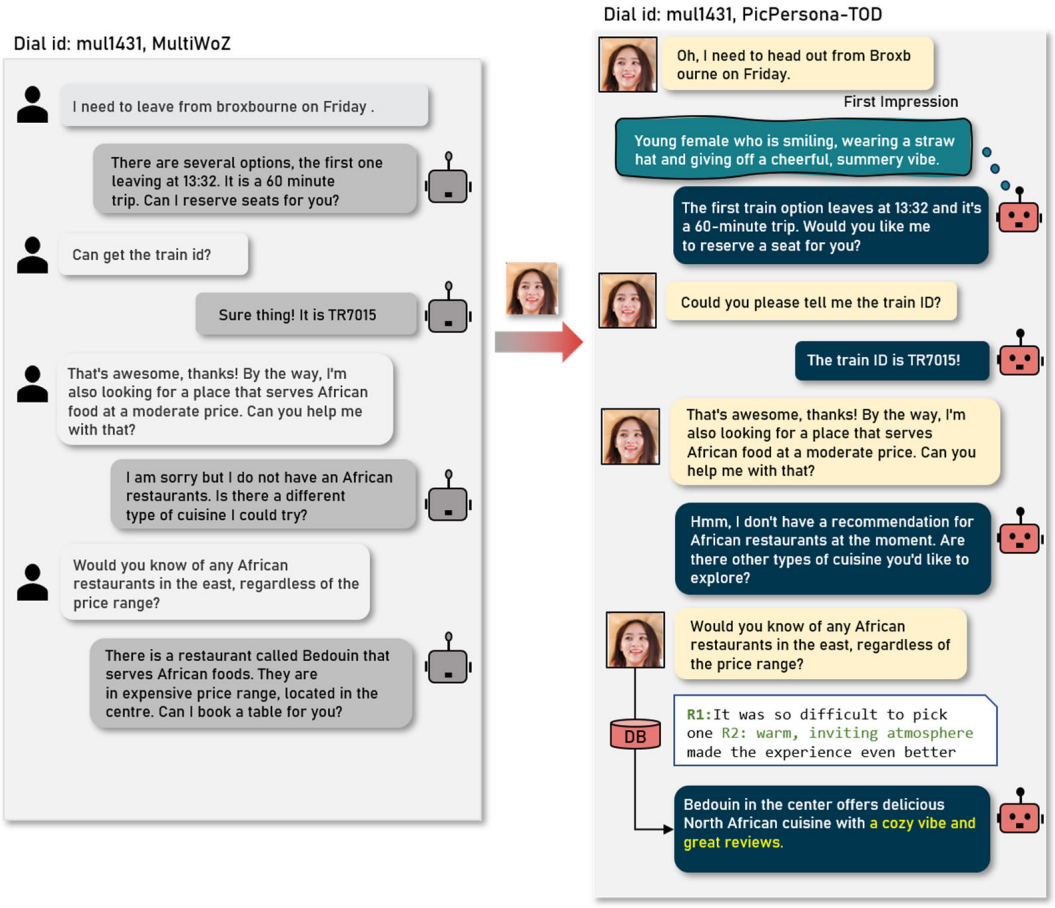
Figure 12 illustrates a screenshot of the questions used for evaluation.

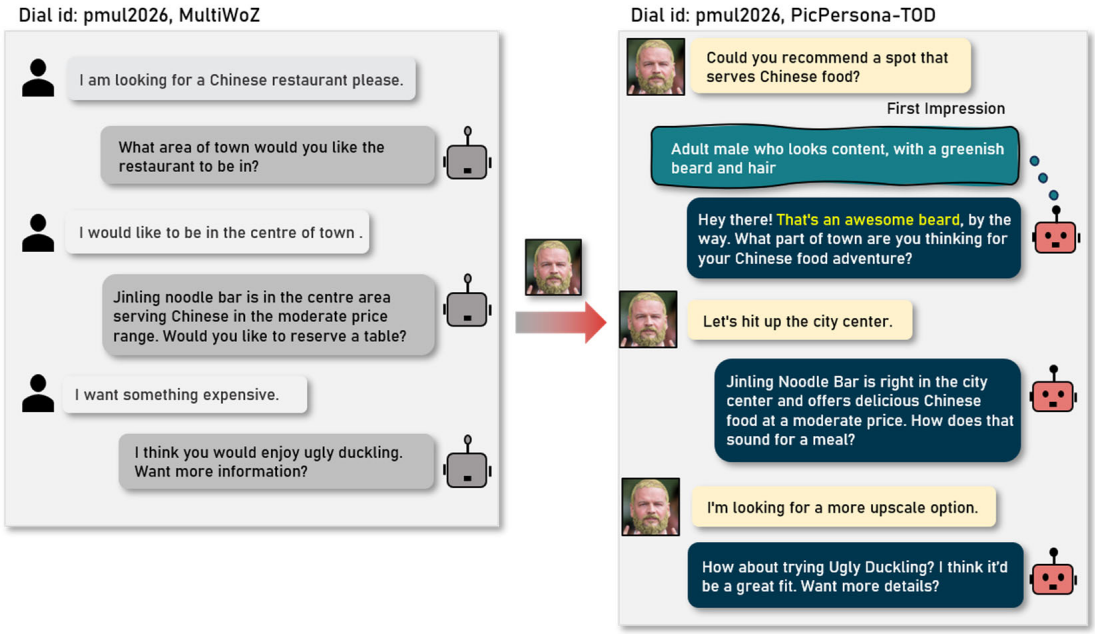
H Sample of PicPersona-TOD

Figure 13 displays samples from the PicPersona-TOD dataset. The left side shows examples from the original MultiWoZ dataset, while the right side presents samples from PicPersona-TOD.

I Prompt Templates

Example prompt templates have been included starting from page 18.


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 Broxb ourne 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?
DB
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?

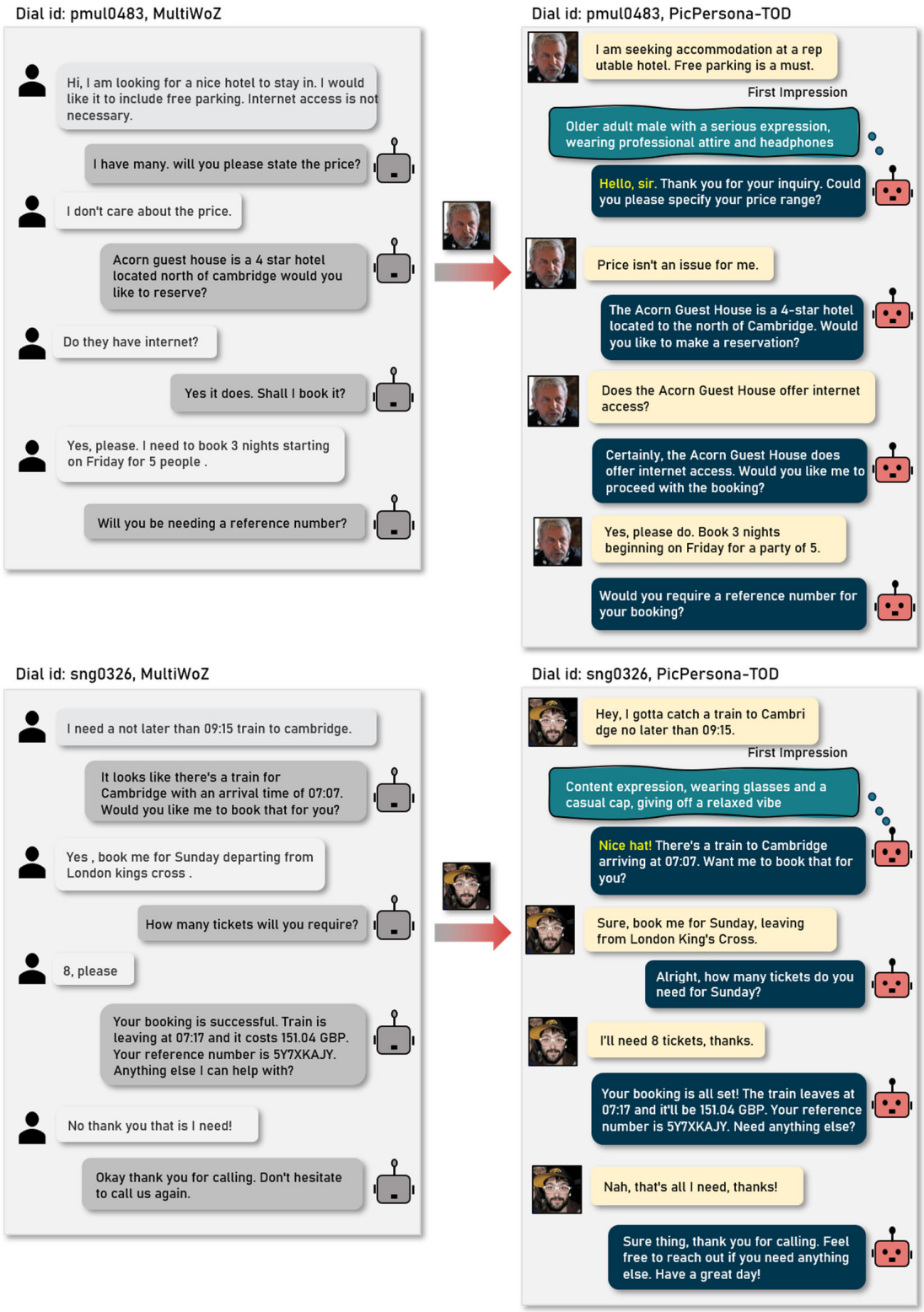


Figure 13: PicPersona-TOD dialogue sample.

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 looks like in a good mood, or it is formal setting, you
can just say "Hello" or "Hi" or Hello sir, madam, etc. Dont' 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) Dose 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 14: Promps templates.