Counterfactual Reasoning Using Predicted Latent Personality Dimensions for Optimizing Persuasion Outcome

Donghuo Zeng[1 [0000−0002−6425−6270]], Roberto S. Legaspi[1 [0000−0001−8909−635X]], Yuewen Sun[2 [0009−0002−4842−1608]], Xinshuai Dong[2 [0000−0002−8593−0586]], Kazushi Ikeda[1 [0009−0000−9563−760X]], Peter Spirtes[2 [0000−0002−1385−190X]], and kun Zhang[2 [0000−0002−1343−9472]]

[1] KDDI Research, Inc., Saitama, Japan
{do-zeng, ro-legaspi, kz-ikeda}@kddi-research.jp
[2] Carnegie Mellon University, Forbes Avenue, Pittsburgh, 15213, Pennsylvania, USA
{xinshuad, ps7z, kunz1}@andrew.cmu.edu

Abstract. Customizing persuasive conversations related to the outcome of interest for specific users achieves better persuasion results. However, existing persuasive conversation systems rely on persuasive strategies and encounter challenges in dynamically adjusting dialogues to suit the evolving states of individual users during interactions. This limitation restricts the system's ability to deliver flexible or dynamic conversations and achieve suboptimal persuasion outcomes. In this paper, we present a novel approach that tracks a user's latent personality dimensions (LPDs) during ongoing persuasion conversation and generates tailored counterfactual utterances based on these LPDs to optimize the overall persuasion outcome. In particular, our proposed method leverages a Bi-directional Generative Adversarial Network (BiCoGAN) in tandem with a Dialogue-based Personality Prediction Regression (DPPR) model to generate counterfactual data $\tilde{D}$. This enables the system to formulate alternative persuasive utterances that are more suited to the user. Subsequently, we utilize the D3QN model to learn policies for optimized selection of system utterances on $\tilde{D}$. Experimental results we obtained from using the PersuasionForGood dataset demonstrate the superiority of our approach over the existing method, BiCoGAN. The cumulative rewards and Q-values produced by our method surpass ground truth benchmarks, showcasing the efficacy of employing counterfactual reasoning and LPDs to optimize reinforcement learning policy in online interactions.

Keywords: Counterfactual reasoning · Latent personality dimensions · Policy learning.

1 Introduction

Persuasive conversations [12, 18, 22, 25] aim to change users' opinions, attitudes, or behaviors by employing effective communication strategies, utilizing

techniques such as reasoning and emotional appeals. They involve deliberate communication with the specific goal of convincing or persuading the user to adopt a particular viewpoint or take action. In individual conversations, effective persuasion hinges on identifying unique, often hidden factors, or latent variables, such as personality (BigFive), beliefs, motivations, and experiences. Tailoring persuasive approaches to these individual nuances increases the likelihood of influencing opinions, attitudes, or behaviors.

Existing works on persuasive conversation systems have explored the impact of persuasive strategies, such as inquiry [16, 19], on the outcome of persuasion tasks, emphasizing strategy learning over personalized approaches. For instance, the work in [3] argues that when targeting demographics, the sequence of persuasive strategies might be immaterial and suggests using a fixed order of persuasive appeals. However, these approaches lack flexibility and dynamics, as it relies solely on pre-defined strategy sequences while overlooking the nuanced differences in natural language. Natural language presents a challenge for persuasive conversation systems in dynamically adapting their conversations to align with individual user personalities. The complexities involved in tracking user states hinder these systems' ability to deliver adaptive and flexible dialogues.

In this work, we introduce an approach designed for a system to dynamically track and leverage the latent personality dimensions (LPDs) of the users during ongoing persuasive conversations. Our proposed method employs a Bidirectional Generative Adversarial Networks (BiCoGAN) [4, 7] partnered with a Dialogue-based Personality Prediction Regression (DPPR) model to generate counterfactual data $\tilde{D}$. This enables our system to generate alternative responses fit to the user's current state, as predicted by the DPPR model. Subsequently, we optimize the system response selection using D3QN model, adapting the conversation flow to the inferred user traits.

The contribution of this work is three-fold: (1) we trained a DPPR model to uncover the hidden aspects of their personalities over time, facilitating the tracking of user states during persuasive conversations. (2) Leveraging the BiCoGAN model with estimated individual LPDs derived from the trained DPPR model, we constructed counterfactual data $\tilde{D}$. This dataset provides alternative system utterances based on LPDs, extending the original dialogues of PersuasionForGood [21] dataset. (3) Employing the D3QN [13] model to learn policies on the counterfactual data $\tilde{D}$, which improves the quality of persuasive conversations, particularly in terms of enhancing persuasion outcomes.

2 Related Work

Persuasive strategies identified from data gathered in persuasive online discussions and social media are frequently utilized to refine argument mining methodologies [1, 20, 23, 24] in the construction of dialogue systems. While these works have introduced several valuable persuasion strategies, none of them have explored an efficient and automated method for applying these strategies effectively. Contrasting these approaches, [21] introduced the PersuasionForGood

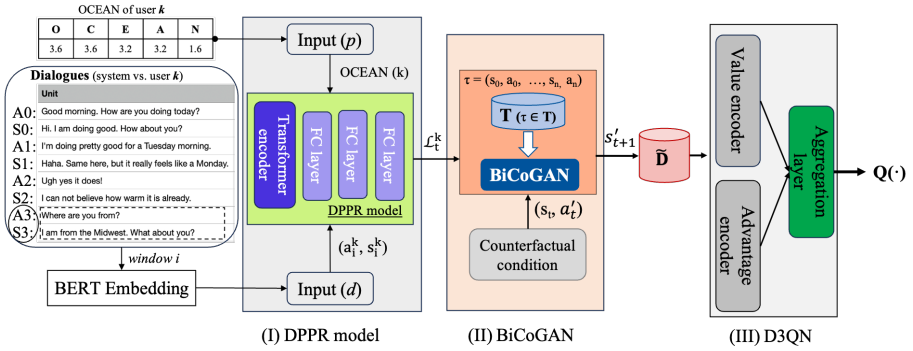


Fig. 1. The overview of our architecture.

dataset through crowd-sourcing, simulating donation-related persuasive scenarios in conversational formats. [16] extended on this dataset to develop an agenda-based persuasion conversation system. However, their strategy maintained a rigid sequence of persuasive appeal strategies and lacked user modeling. To address this issue, [19] adopts reinforcement learning with dynamic user modeling to optimize the sequence of persuasive appeals based on dialogue history and the user's inclination towards donation. Persuasive systems lack fluent communication and flexibility due to strategy reliance and absence of counterfactual cases. Leveraging GANs' success in persuasive dialogue [17], we generate counterfactual data to enhance adaptability.

Personalization in persuasion [3, 5, 8, 10, 14] is important for the persuasive system to adapt to individual scenarios. Tailoring persuasive messages to align with the interests and concerns of the users is a way to enhance system effectiveness. For instance, [5] explores the customization of persuasive technologies to individual users through persuasion profiling, utilizing both explicit measures from standardized questionnaires and implicit, behavioral measures of user traits. However, employing truthful personalization to enhance the persuasion dialogue may not be the optimal solution as it could impact responses over multiple steps and overlook the changing, hidden personality-related factors. Our research focuses on dynamically leveraging LPDs to capture the evolving, hidden personality-related factors during persuasive conversation. Subsequently, we employ counterfactual reasoning with LPDs to construct counterfactual data, thereby improving the persuasion outcomes.

3 Our Architecture

In this section, we introduce problem setting and three parts of our architecture: 1) estimation of individual latent personality dimensions, 2) counterfactual data $\tilde{D}$ establishment, and 3) policy learning.

3.1 Problem Setting

Let's assume that all the dialogues are padded to the same length, and $D = \{(s_t, a_t, s_{t+1})\}_{t=0}^{T-1}$ represents an observed episode within a decision-making process governed by the dynamics of a structural causal model (SCM) [11], which represents causal relationships between variables via structural equations. This aims to delineate the counterfactual outcome achievable by any alternative action sequence under the specific circumstances of the episode. We explore Individualized Markov Decision Pocesses (IMDPs), in which each IMDP is defined by $M = (S, A, \mathcal{L}, f, R, T)$, where $S = \{s_0, s_1, s_2, ..., s_{\lceil T/2 \rceil}\}$ and $A = \{a_0, a_1, a_2, ..., a_{\lfloor T/2 \rfloor}\}$ are finite state and action spaces, respectively, where $\lfloor \cdot \rfloor$ represents floor function and $\lceil \cdot \rceil$ denotes ceiling function. $\mathcal{L}$ is the individual LPDs space. R is the immediate reward computed by the reward model as introduced in the equation 5. Accordingly, the causal mechanism f is defined as

$$s_{t+1} = f(s_t, a_t, \mathcal{L}_t, \varepsilon_{t+1}) \tag{1}$$

where s_t, a_t, and $\mathcal{L}_t$ are the state, action, and estimated LPDs at time t, respectively, and ε_{t+1} is the noise term independent of (s_t, a_t). A graphical representation of the individual state transition process is depicted in Fig. 2.

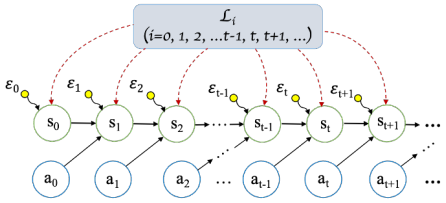


Fig. 2. The individualized transition dynamics model.

Suppose D is the real-world data, BiCoGAN by using estimated LPDs creates a set of N counterfactual datasets $\tilde{D} = \{\tilde{D}_0, \tilde{D}_1, ..., \tilde{D}_N\} = \{(s_t^{'}, a_t^{'})\}_{t=0}^{T-1}$. Our target is to learn policies to optimize the persuasion outcome on counterfactual data $\tilde{D}$.

3.2 Estimation of Individual Latent Personality Dimensions

The OCEAN model [3] is widely recognized in psychology and social sciences for its reliability in assessing and understanding personality traits, which is validated across different cultures, age groups, and demographic backgrounds. However, critics argue that it falls short in representing the nuanced dynamics of personalities such as in ongoing dialogues. In addressing this, we endeavor to estimate individual LPDs in real-time to foster adaptability during conversations. It helps in dynamically adjusting persuasive approaches, such as system utterances, based on the inferred user traits for individualized conversation. For this purpose, we developed a dialogue-based personality prediction regression (DPPR) model that consists of a transformer encoder followed by three fully connected layers (refer to (I) in Fig. 1). During model training, the inputs comprise utterances between system and user, alongside the annotated 5-dimensional

[3] The "Big Five" includes Openness, Conscientiousness, Extroversion, Agreeableness, and Neuroticism.

personalities OCEAN values [15] attributed to the users. Each input is a one-turn utterance constituting a one-time exchange within a dialogue between the system and user. When a new user engages, the model progressively infers the user's LPDs over time to better understand them during the conversation. This enables the system to dynamically adapt the new utterances in the persuasion dialogue, optimizing the outcome based on the inferred individual LPDs.

3.3 Counterfactual Data $\tilde{D}$

Persuasive systems heavily rely on past online interactions, often leading to sub-optimal outcomes due to the absence of sufficient actual observed results. Counterfactual reasoning enables the assessment of unobserved scenarios, encompassing various conditions and user-diverse reactions. Constructing counterfactual data facilitates enhanced policy learning, optimizing the decision-making process to achieve better outcomes.

We assume that the state s_{t+1} satisfies the SCM: $s_{t+1} = f(s_t, a_t, \mathcal{L}_t, \varepsilon_{t+1})$, then, we employ BiCoGAN (refer to (II) in Fig. 1) to learn the function f by minimizing the disparity between the input real data and the generated data, maintaining realistic counterfactual states aligned with observed scenarios for practical relevance. Simultaneously, it estimates the value of noise term ε_{t+1}, representing disturbances arising from unobserved factors, seen in Fig. 3. Specifically, the BiCoGAN proceeds in two directions: 1) mapping from (s_t, a_t, $\mathcal{L}_t$, ε_{t+1}) to s_{t+1} in the generator G, and 2) estimating (s_t, a_t, $\mathcal{L}_t$, ε_{t+1}) from s_{t+1} via the encoder E. The discriminator D is trained to distinguish between real and inferred data. The decoder and encoder distributions are formulated as follows, respectively.

$$\begin{aligned} P(\hat{s}_{t+1}, s_t, a_t, \mathcal{L}_t, \varepsilon_{t+1}) &= P(s_t, a_t, \mathcal{L}_t, \varepsilon_{t+1}) P(\hat{s}_{t+1} | s_t, a_t, \mathcal{L}_t, \varepsilon_{t+1}), \\ P(s_{t+1}, \hat{s}_t, \hat{a}_t, \hat{\mathcal{L}}_t, \hat{\varepsilon}_{t+1}) &= P(s_{t+1}) P(\hat{s}_t, \hat{a}_t, \hat{\mathcal{L}}_t, \hat{\varepsilon}_{t+1} | s_{t+1}) \end{aligned} \tag{2}$$

where $\hat{s}_t$, $\hat{s}_{t+1}$, $\hat{a}_t$, $\hat{\mathcal{L}}_t$, and $\hat{\varepsilon}_t$ are estimations of s_t, s_{t+1}, a_t, $\mathcal{L}_t$, and ε_{t+1}, respectively. The $P(\hat{s}_{t+1} | s_t, a_t, \mathcal{L}_t, \varepsilon_{t+1})$ and $P(\hat{s}_t, \hat{a}_t, \hat{\mathcal{L}}_t, \hat{\varepsilon}_t | s_{t+1})$ are respectively the conditional distributions of the decoder and encoder. To deceive the discriminator model, the objective function is optimized as a minimax game defined as

$$\begin{aligned} \min_G \max_D V(D, G, E) = \min_G \max_D \{ & \mathbb{E}_{s_{t+1} \sim p_{\text{data}}(s_{t+1})} [\log D(E(s_{t+1}), s_{t+1})] \\ & + \mathbb{E}_{z_t \sim p(z_t)} [\log(1 - D(G(z_t), z_t))] \\ & + \lambda E_{(s_t, a_t, s_{t+1}) \sim p_{\text{data}}(s_t, a_t, s_{t+1})} [R((s_t, a_t), E(s_{t+1}))] \} \end{aligned} \tag{3}$$

where $z_t = (s_t, a_t, \mathcal{L}_t, \varepsilon_{t+1})$, R is a regularizer with its hyperparameter λ to avoid overfitting issues.

After learning the SCM, counterfactual reasoning can be carried out to build the counterfactual data $\tilde{D}$. Suppose at time t, we have the tuple (s_t, a_t, $\mathcal{L}_t$, s_{t+1}), we want to know what would the state s'_{t+1} be if we take an alternative action a'_t.

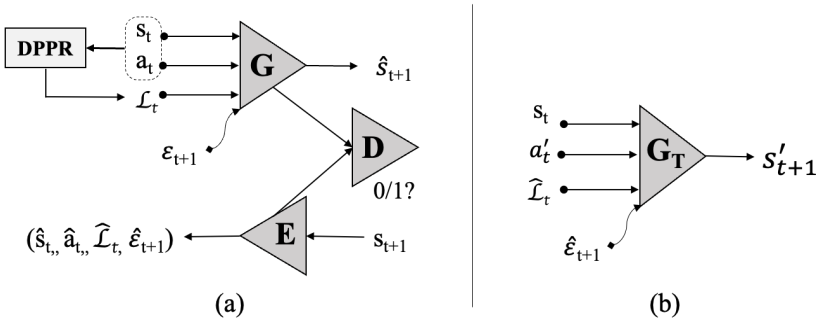


Fig. 3. (a) Trained DPPR model, Generator G, Encoder E, and Discriminator D in the training. (b) Trained Generator G_T in counterfactual states generating.

To determine this, we take (s_t, $a_t^{'}$, $\mathcal{L}_t$), as input for the trained generator model G_T, which consequently outputs the counterfactual state $s_{t+1}^{'}$. Additionally, to achieve this process, we first need to build the set of counterfactual actions $\{a_t^{'}\}$. We derived the counterfactual actions $\{a_t^{'}\}$ from the real-world actions $\{a_t\}$ through random selection.

3.4 Policy learning

Following the generation of counterfactual data $\tilde{D}$, our approach involves learning policies on $\tilde{D}$ to maximize future rewards. We employ the Dueling Double-Deep Q-Network (D3QN) [13] as an enhanced variant of the standard Deep Q-Networks (DQNs) [9] to address Q-value overestimation (refer to (III) in Fig. 1). D3QN mitigates this issue by segregating the value function into state-dependent advantage function $A(s, a)$ and value function $V(s)$, where $A(s, a)$ capitalizes on how much better one action is than the other actions, and $V(s)$ indicates how much reward will be achieved from state s. The Q-values can be calculated on $\tilde{D}$ as follows:

$$Q(s', a'; \theta) = \mathbb{E}\left[r(s', a') + \gamma \max_{a'} Q(s', a'; \theta^{-}) \mid s^{'}, a^{'}\right], \tag{4}$$

where $r(s^{'}, a^{'})$ is the reward of taking action $a^{'}$ at the state $s^{'}$, the γ is the discount factor of the max q value among all possible actions from the next state. The $r(\cdot)$ is a reward function that is used for the calculation of reward value for a given action input. The reward function is an LSTM-based architecture trained using dialogues and their corresponding donation values, the reward value of (s'_t, a'_t) is calculated as

$$r(s', a') = \begin{cases} 0, & \text{if } t < T-1, \\ LSTM(BERT_{embedding}(\{s'_t, a'_t\}_{t=0}^{T-1})), & \text{otherwise,} \end{cases} \tag{5}$$

where T is the length of one dialogue in time unit. The target Q-values are derived from actions obtained through a feed-forward pass on the main network, diverging from direct estimation from the target network. During policy

learning, the state transition starts with state $s_0^{'}$, and computing $argmax_{a_{0i}^{'}} Q(s_0^{'}, a_{0i}^{'}; \theta, \alpha, \beta)$ leads to selecting the optimal action a_0^* for state $s_0^{'}$ in $\tilde{D}$, where $i = 0, 1, 2, .., N-1$. In the end, we apply the Mean Squared Error (MSE) as a loss function to update the weights of D3QN neural networks.

4 Experiment

4.1 Dataset

ER:	Good Evening
EE:	Hello there. how are you?
ER:	I am doing well! How are doing today?
EE:	I am doing pretty well. thanks for asking!
ER:	I'd like to tell you about a great program I am working on! Have you ever heard of Save the Children?
EE:	I may have in passing, but could you tell me more information about it?
ER:	Save the Children is a non-governmental organization that operates worldwide wide raising funds through partners and donations to fight for children's rights and provide relief and support for children in developing countries.
EE:	Sounds interesting and something worth donating towards a good cause. Would you donate to this fund?
ER:	I would and I have! 9 out of every 10 dollars goes directly to a child in need.
EE:	That is great to know. Any additional information you could tell me about the Save the Children program?
ER:	Absolutely! There are a variety of ways to be a part of the program outside of donations. They have a program to sponsor a child in need, you can throw a fundraising event or participate in events such as triatholons to raise funds, or help advocate for children.
EE:	Thank you for the information. What is the best way I could get involved with this charity?
ER:	You are more than welcome to donate today and your donation will be sent to Save the Children to go towards programs that help the children directly.
EE:	Great. How much are the donations?
ER:	You can donate up to two dollars of your payment, but more is always appreciated!
EE:	Alright. $2 sounds good enough. Could I get a web address to the fund if I have any more questions?
ER:	Thank you so much for you generous donation! To answer any more questions the link is URL
EE:	Thanks you. I will be more than glad to donate.

Fig. 4. An example (ID: 20180904154250_98_*live*, donation: $2.0, OCEAN values: 3, 3.2, 3, 3.6, 3) persuasive dialogue between persuader (ER) and persuadee (EE) from PersuasionForGood dataset. Dynamic modeling of the dialogue, utterances of ER as actions (grey), utterance of EE as states (white).

To verify our method, we use the *PersuasionForGood* [4] dataset of human-human dialogues. This dataset aims to facilitate the development of intelligent

[4] https://convokit.cornell.edu/documentation/persuasionforgood.html

persuasive conversational agents and focuses on altering users' opinions and actions regarding donations to a specific charity. Through the online persuasion task, one participant as *persuader* (ER) was asked to persuade the other as *persuadee* (EE) to donate to a charity, Save the Children [5]. This large dataset comprises 1,017 dialogues, including annotated emerging persuasion strategies in a subset. Notably, the recorded participants' demographic backgrounds, such as the Big-Five personality traits (OCEAN), offer an opportunity to estimate the LPDs of users (*persuadees*). Additionally, each dialogue includes donation records for both *persuader* and *persuadee*. Our focus in this work is primarily on the donation behavior of the *persuadee*, with 545 (54%) recorded as donors and 472 (46%) as non-donors. The dialogue data utilized in this study is represented by BERT embeddings extracted by a pre-trained BERT model [2] from Hugging Face[6]. Each natural language utterance of EE or ER in the dialogue is represented by a 768-dimensional BERT feature vector. The OCEAN are the real values sourced from the original dataset. For instance, in Fig. 4, the OCEAN of a specific *persuadee* (ID=180904154250_98_*live*) is presented as a 5-dimensional vector.

4.2 Experimental setup

All implementations were executed using PyTorch and models were trained on a GeForce RTX 3080 GPU (10G). Our method is a teamwork of three models, which are trained separately by using different formats of the input data. Hyperparameters shared by the models are listed in Table 1. A five-fold cross-validation is employed to ensure the reliability of the DPPR model, which allows us to assess its generalizability to unseen data. For BiCoGAN, we set the dimension of the noise item to be the same as BERT's embedding, which is 768. In the end, we generated a total of 100 counterfactual data by using different sets of counterfactual actions. To guarantee the robustness of the reward model training, dialogues featuring donation amounts surpassing $20.0 were intentionally removed. The aim behind removing these dialogues is to reduce potential bias and ensure a more balanced training for the model. The final number of dialogues is 997, featuring 25 exchanges of utterances between EE and ER, alternating between the two roles in each dialogue. We split the data into training and testing sets using a 80/20 split. All the model is optimized using Adam [6] with a learning rate of 0.0001. In addition, we set the discount factor γ in the D3QN to 0.9.

Table 1. The hyperparameters across various models.

Model	hidden units	batch size	lr	epochs
DPPR	1024	64	0.0001	100
BiCoGAN	100	100	0.0001	10
RM	256	64	0.0001	1,000
D3QN	256	60	0.001	20

[5] https://www.savethechildren.org/
[6] https://huggingface.co/bert-base-uncased

4.3 Results

To facilitate the creation of the counterfactual world data $\tilde{D}$, we introduce the counterfactual actions $\{a_t^{'}\}$ as inputs into the trained BiCoGAN for $\tilde{D}$ generation. An initial step in this process involves establishing the counterfactual actions set $\{a_t^{'}\}$, which is derived through random selection from real-world action set $\{a_t\}$. However, when randomly selected as counterfactual actions, some greeting utterances will appear in the middle or end of a conversation, which is inappropriate.

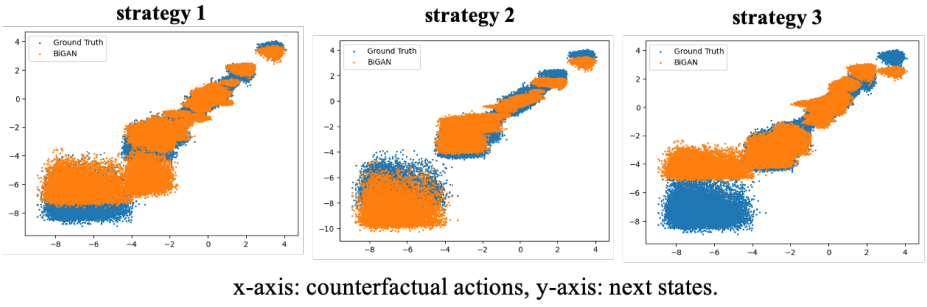


Fig. 5. The relationship between the counterfactual action $a_t^{'}$ and the next state: counterfactual case $s_{t+1}^{'}$ generated by BiCoGAN or ground truth s_{t+1}

To address this issue, we define three strategies of counterfactual action selection based on whether or not the greetings are selected: In strategy 1, all the greetings are selected, we shuffle all the real-world actions as $\{a_t^{'}\}$; in strategy 2, the first greeting is not selected, we remove the first utterance of persuader $\{a_0\}$ per dialogue, then sample from the rest actions $\{a_t\}$ - $\{a_0\}$ as $\{a_t^{'}\}$; in strategy 3, all the greetings are not selected, we remove the first three utterances $\{a_0, a_1, a_2\}$ per dialogue, then sample from the rest actions $\{a_t\}$ - $\{a_0, a_1, a_2\}$ as $\{a_t^{'}\}$. To leverage the quality of the counterfactual data across the three strategies of counterfactual action sets, we conducted a comparison between the generated counterfactual states by the trained BiCoGAN generator and the ground truth states, seen in Fig. 5. We can observe that strategy 2 yields superior results as it better aligns the real world's next state, s_{t+1}, and counterfactual next state, $s_{t+1}^{'}$, compared to the other two strategies. Finally, counterfactual data $\tilde{D}$ can be represented as

$$\begin{aligned}\tilde{D} &= \{\tilde{D}_0, \tilde{D}_1, ..., \tilde{D}_i, ..., \tilde{D}_{N-1}\} \\ &= \{(s_0^{'}, a_0^{'}, s_1^{'}, a_1^{'}, ..., s_t^{'}, a_t^{'}, ..., s_{T-1}^{'})_{j=0}^{\Sigma-1}\}_{i=0}^{N-1},\end{aligned} \tag{6}$$

where T typically represents the total number of steps or time points in the sequence, Σ is the number of dialogues, and N is the number of counterfactual databases.

Following the estimation of the dynamics model and the creation of an augmented dataset $\tilde{D}$ through counterfactual reasoning, the subsequent step involves

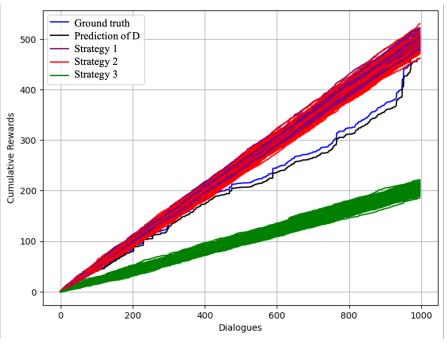


Fig. 6. The (BiCoGAN + DPPR) reward predictions of counterfactual data on three strategies and real data D (black), compared with ground-truth (blue).

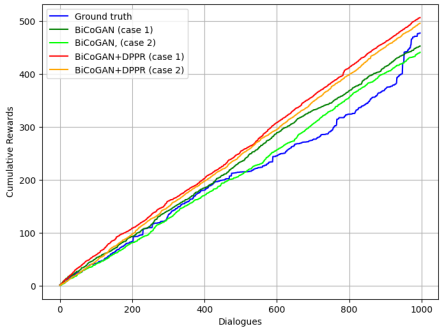


Fig. 7. Comparison of cumulative rewards in dialogues between ground truth and counterfactual cases utilizing BiCoGAN or BiCoGAN+DPPR.

training policies on the counterfactual data $\tilde{D}$ to optimize the Q values and improve the future predicted cumulative rewards. To ensure fairness in policy learning, we aim for balanced predicted cumulative rewards of $\tilde{D}$, with 50 counterfactual databases exceeding and 50 counterfactual databases falling short of the ground truth, thus forming our final counterfactual data $\tilde{D}$. (cf: Section 3.3). For cumulative reward computation, we first input the dialogue into the reward model (Equation (5)), then perform a cumulative sum operation on the predicted reward of the trained reward model. Fig. 6 shows the cumulative reward prediction for the three strategies, where strategies 1 and 2 are mostly higher than the ground truth's, while strategy 3 is lower than the ground truth.

During policy learning with the D3QN model, the process involves computing Q values for 100 action candidates from 100 counterfactual databases at each state. The model selects the action with maximum Q value for each time step. As a result, this process generates a counterfactual dialogue and then utilizes the trained reward model (cf: Equation (5)) to predict the reward. Determining whether the counterfactual dialogue in natural language signifies a donation or not will be our focus for future work. For the optimization of D3QN, we established two distinct cases based on the optimization time for the loss function. In **Case 1**, where the loss function is optimized once per dialogue. For example, the sequence of j-th dialogue starts from state $s^{'}_{j,0}$. It proceeds with a series of action-state pairs $a^{'}_{j,0}$, $s^{'}_{j,1}$, ..., $a^{'}_{j,i}$, $s^{'}_{j,i}$, ..., $a^{'}_{j,12}$, $s^{'}_{j,12}$, for each state $s^{'}_{j,i}$, based on $k = argmax(Q(s^{'}_{j,i}, a^{'}_{j})^{N}_{i=0})$ (N=99) to select the next action $a^{'}_{j,k}$ from the j-th dialogue of $\tilde{D}_k$. In contrast, **Case 2** differs as the loss function optimization takes place once per state. The action selection process is illustrated in Fig. 8

After the policy learning, we obtain a corresponding counterfactual dialogue for donation prediction, aligning with each real-world dialogue. The sequences of real-world dialogues begin from s_0, while the counterfactual dialogues starts from $s^{'}_0$, where maintaining the equivalence of $s_0 = s^{'}_0$. In Fig. 7, the cumulative rewards of both BiCoGAN and BiCoGAN+DPPR models in two cases are higher

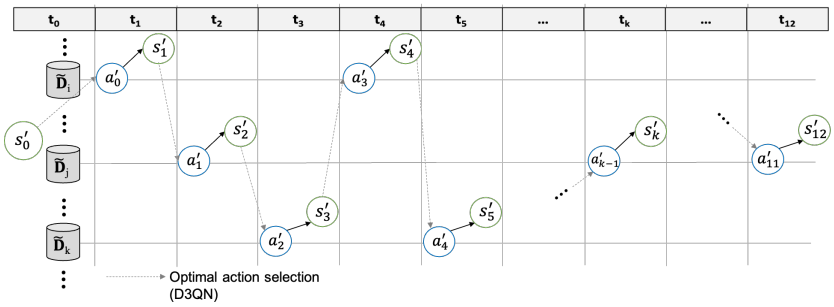


Fig. 8. The process of counterfactual dialogue produced for reward prediction during the policy learning.

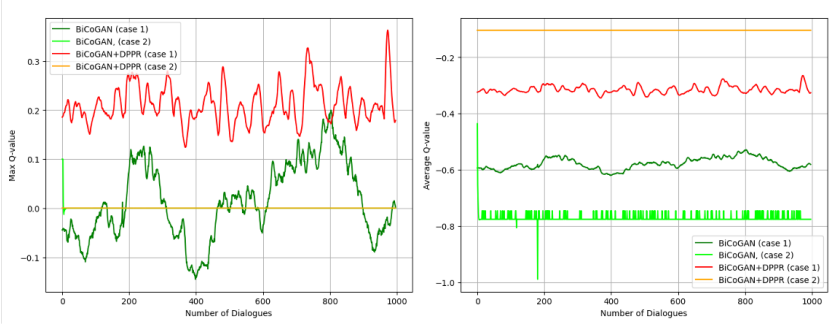


Fig. 9. Comparison of maximum and average Q values in dialogues between ground truth and counterfactual cases utilizing BiCoGAN or BiCoGAN+DPPR.

than the ground truth overall. It can be observed that the BiCoGAN+DPPR (case 1) gets the best cumulative rewards and BiCoGAN+DPPR (case 2) obtains the second results over dialogues. The final cumulative rewards/donation amounts are \$506.82 (+\$29.2) and \$496.15 (+\$18.53) respectively, higher than ground truth cumulative rewards \$477.62 and the BiCoGAN in case 1 and case 2, achieved \$453.02 (-\$24.6) and \$441.11 (-\$36.51), respectively.

Furthermore, when estimating the Q-values of the learned optimal policy (Fig. 9), the optimal policy derived from BiCoGAN with LPDs exhibits higher estimated Q-values, both in maximum and average, compared to the single BiCoGAN model. In Case 1 of BiCoGAN+DPPR, the maximum Q-values surpass Case 2, while the average Q-values are lower in Case 1 than in Case 2. This discrepancy arises from weight optimization in D3QN primarily at the dialogue's end, potentially leading to actions with exceptionally high Q-values.

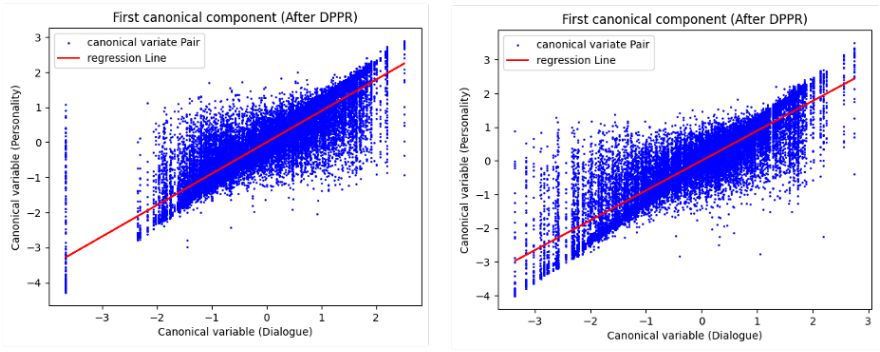
4.4 Ablation Study

Impact of window size plays an essential role in the DPPR model, it will influence the accuracy of prediction and determine how much information is

meaningful for the individual LPDs estimation. We separately set the window size as 1, 2, 3, and 4 turns, each turn includes exchange utterances between a *persuader* and a *persuadee*. The evaluation[7] results of DPPR on different window size is shown in Table 2, we can observe that when the window sizes decrease, the precision increase, so we choose one-turn to train DPPR model for counterfactual data building. To further investigate the one-turn trained model, we leverage the first top two canonical components, shown in Fig. 10, the CCA values for each are 0.894 and 0.888, respectively, which indicate there exist relatively high correlations between the one-turn utterance and personality.

Table 2. Performance of DPPR model with varied window sizes.

Win size	MSE	RMSE	MAPE	R2	MAE
1 turn	0.166	0.407	0.092	0.830	0.254
2 turns	0.258	0.508	0.119	0.641	0.323
4 turns	0.441	0.664	0.178	0.387	0.474
8 turns	0.488	0.698	0.195	0.319	0.518



CCA=0.894 CCA=0.888

Fig. 10. The canonical coefficient of the top two CCA components.

5 Conclusion

In this paper, we highlight the limitations of existing persuasive conversation systems, particularly in adapting to individual user traits. We augmented the BiCoGAN model, which creates counterfactual data that simulates an alternative world, with our DPPR models to enable awareness of latent user dimensions during persuasive interactions. This facilitated an optimized conversation flow using the D3QN model. In the experiments conducted on the PersuasionForGood dataset, our approach showcased superiority over existing method, BiCoGAN, and ground truth, demonstrating the efficacy of leveraging latent traits to enhance persuasion outcomes.

[7] MSE - Mean Squared Error; RMSE - Root Mean Squared Error; MAPE - Mean Absolute Percentage Error; R2 - R-squared (Coefficient of Determination); MAE - Mean Absolute Error

Bibliography

[1] Chakrabarty, T., Hidey, C., Muresan, S., McKeown, K., Hwang, A.: AMPERSAND: Argument mining for PERSuAsive oNline discussions. In: EMNLP-IJCNLP. pp. 2933–2943. ACL, Hong Kong, China (Nov 2019)

[2] Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)

[3] Hirsh, J.B., Kang, S.K., Bodenhausen, G.V.: Personalized persuasion: Tailoring persuasive appeals to recipients' personality traits. Psychological science **23**(6), 578–581 (2012)

[4] Jaiswal, A., AbdAlmageed, W., Wu, Y., Natarajan, P.: Bidirectional conditional generative adversarial networks. In: Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part III 14. pp. 216–232. Springer (2019)

[5] Kaptein, M., Markopoulos, P., De Ruyter, B., Aarts, E.: Personalizing persuasive technologies: Explicit and implicit personalization using persuasion profiles. International Journal of Human-Computer Studies **77**, 38–51 (2015)

[6] Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)

[7] Lu, C., Huang, B., Wang, K., Hernández-Lobato, J.M., Zhang, K., Schölkopf, B.: Sample-efficient reinforcement learning via counterfactual-based data augmentation. arXiv preprint arXiv:2012.09092 (2020)

[8] Matz, S., Teeny, J., Vaid, S.S., Harari, G.M., Cerf, M.: The potential of generative ai for personalized persuasion at scale (2023)

[9] Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G., Graves, A., Riedmiller, M., Fidjeland, A.K., Ostrovski, G., et al.: Human-level control through deep reinforcement learning. nature **518**(7540), 529–533 (2015)

[10] Orji, R., Tondello, G.F., Nacke, L.E.: Personalizing persuasive strategies in gameful systems to gamification user types. In: Proceedings of the 2018 CHI conference on human factors in computing systems. pp. 1–14 (2018)

[11] Pearl, J., et al.: Models, reasoning and inference. Cambridge, UK: CambridgeUniversityPress **19**(2), 3 (2000)

[12] Prakken, H.: Formal systems for persuasion dialogue. The knowledge engineering review **21**(2), 163–188 (2006)

[13] Raghu, A., Komorowski, M., Ahmed, I., Celi, L., Szolovits, P., Ghassemi, M.: Deep reinforcement learning for sepsis treatment. arXiv preprint arXiv:1711.09602 (2017)

[14] Rieger, A., Shaheen, Q.U.A., Sierra, C., Theune, M., Tintarev, N.: Towards healthy engagement with online debates: An investigation of debate summaries and personalized persuasive suggestions. In: Adjunct Proceedings of

the 30th ACM Conference on User Modeling, Adaptation and Personalization. pp. 192–199 (2022)
[15] Roccas, S., Sagiv, L., Schwartz, S.H., Knafo, A.: The big five personality factors and personal values. Personality and social psychology bulletin **28**(6), 789–801 (2002)
[16] Shi, W., Wang, X., Oh, Y.J., Zhang, J., Sahay, S., Yu, Z.: Effects of persuasive dialogues: testing bot identities and inquiry strategies. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. pp. 1–13 (2020)
[17] Su, H., Shen, X., Hu, P., Li, W., Chen, Y.: Dialogue generation with gan. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 32 (2018)
[18] Torning, K., Oinas-Kukkonen, H.: Persuasive system design: state of the art and future directions. In: Proceedings of the 4th international conference on persuasive technology. pp. 1–8 (2009)
[19] Tran, N., Alikhani, M., Litman, D.: How to ask for donations? learning user-specific persuasive dialogue policies through online interactions. In: Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization. pp. 12–22 (2022)
[20] Wachsmuth, H., Naderi, N., Hou, Y., Bilu, Y., Prabhakaran, V., Thijm, T.A., Hirst, G., Stein, B.: Computational argumentation quality assessment in natural language. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers. pp. 176–187 (2017)
[21] Wang, X., Shi, W., Kim, R., Oh, Y., Yang, S., Zhang, J., Yu, Z.: Persuasion for good: Towards a personalized persuasive dialogue system for social good. In: Korhonen, A., Traum, D.R., Màrquez, L. (eds.) ACL 2019. pp. 5635–5649. ACL (2019)
[22] Wei, Z., Liu, Y., Li, Y.: Is this post persuasive? ranking argumentative comments in online forum. In: Erk, K., Smith, N.A. (eds.) Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 195–200. ACL, Berlin, Germany (Aug 2016)
[23] Wei, Z., Liu, Y., Li, Y.: Is this post persuasive? ranking argumentative comments in online forum. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 195–200 (2016)
[24] Yang, D., Chen, J., Yang, Z., Jurafsky, D., Hovy, E.: Let's make your request more persuasive: Modeling persuasive strategies via semi-supervised neural nets on crowdfunding platforms. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 3620–3630 (2019)
[25] Yoshino, K., Ishikawa, Y., Mizukami, M., Suzuki, Y., Sakti, S., Nakamura, S.: Dialogue scenario collection of persuasive dialogue with emotional expressions via crowdsourcing. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018) (2018)