From Persona to Person: Enhancing the Naturalness with Multiple Discourse Relations Graph Learning in Personalized Dialogue Generation

Anonymous submission

Abstract

In dialogue generation, the naturalness of responses is key for effective human-machine interaction, significantly enhancing user experience. Personalized response generation poses even greater challenges, as the responses must be coherent and consistent with the user's personal traits or persona descriptions. In this study, we propose a novel method named **MUDI** (**Mu**ltiple **Di**scourse Relations Graph Learning) aimed at effectively modeling and integrating discourse relations and persona information within the context of personalized dialogue generation. We initially utilize Large Language Models to assist in annotating discourse relations and to transform dialogue data into structured dialogue graphs. We employ our newly proposed DialogueGAT as the graph encoder, which captures implicit discourse relations within this structure. Persona descriptions are also encoded into completed persona graphs, facilitating the capture of semantic relationships between persona elements. An Attention-Based Feature Fusion method integrates data from both graphs, creating a personalized, coherence-aware dialogue representation. In the personalized response generation phase, a Prompt-based mechanism and a Coherence-Aware Attention strategy are implemented to enhance the decoder's consideration of discourse relations. Our experiments and case studies demonstrate significant improvements in the quality of personalized responses, making them more coherent, aligned with persona, and natural, thus resembling human-like dialogue exchanges.

Introduction

Dialogue Generation is a foundational technology for dialogue systems, primarily focusing on the task of Next Utterance Generation, also known as Next Response Generation. In multi-turn dialogue scenarios, the conversational agent's objective is to analyze the context of the multi-turn dialogue and the current query to produce a following appropriate response. A significant drawback of traditional dialogue systems is their limited ability to personalize responses based on specific user characteristics or preferences. This limitation often results in generic and less engaging interactions that fail to meet individual user needs effectively (Jiang and de Rijke 2018). Previous works (Song et al. 2020; Warren 2006) defined this problem as the naturalness issue of the Dialogue System. One effective solution to enhance the naturalness of dialogue systems is to integrate personality into the agents, referred to as "persona".

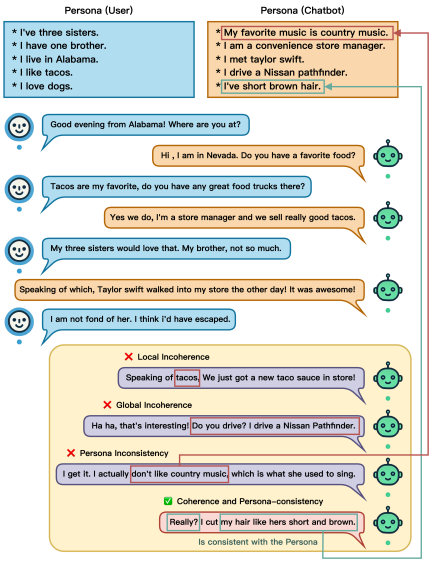

Figure 1: Example of Incoherence and Persona Inconsistency issue in Personalized Dialogue Generation. This figure highlights the challenges in maintaining coherence and persona consistency in personalized dialogue system.

Typically, a persona comprises several sentences describing the interlocutor's facts or background. This information is crucial for building a trustworthy and confident dialogue system. By endowing chatbot agents with human-like traits, the interactions become more realistic. Given these benefits, Personalized Dialogue Generation has emerged as a prominent research topic in recent years, focusing on improving user engagement and satisfaction within dialogue systems. This surge in interest is largely driven by the availability of large-scale personalized dialogue datasets, such as those from Zhang et al. (Zhang et al. 2018a) and Dinan et al. (Dinan et al. 2020). These datasets have significantly advanced efforts to enhance persona consistency and context understanding in generated responses. Innovative methods, including Liu et al. (Liu et al. 2020) have concentrated on improving dialogue system consistency through more complex modeling of interlocutor understanding. Song et al. (Song et al. 2021) tackled persona-based dialogue generation by

dividing it into two separate tasks: response generation and consistency understanding. They utilized unlikelihood training techniques to reduce the production of contradictory dialogue responses. Furthermore, Huang et al. (Huang et al. 2023) proposed a persona-adaptive attention that balances and regularizes the input information from persona and context, improving the consistent understanding of generated responses by persona and context.

Despite advances, challenges remain in enhancing engagement, coherence, and persona consistency. The focus has been predominantly on the trade-offs between persona consistency and discourse coherence. These challenges are primarily twofold. Firstly, many existing methods rely on sophisticated structures or external natural language inference (NLI) datasets to learn persona consistency. This approach, while effective, can sometimes lead the model to overly prioritize persona information at the expense of neglecting the broader dialogue context. Secondly, many dialogue-generating models fail to adequately consider the importance of discourse relations. They often neglect coherence, assuming that fluency alone can measure a dialogue's coherence. Discourse coherence, which focuses on how utterances are interconnected and the overall organization of dialogue to effectively convey information, is essential for effective conversation. Discourse coherence can be divided into local and global coherence. Local coherence refers to the logical connections between adjacent sentences, ensuring that they relate to each other and form a coherent sequence. Global coherence, on the other hand, extends beyond immediate sentence pairs to encompass higher-level relationships across the entire dialogue. This macro-linguistic capability allows conversational agents to maintain topic consistency and effectively convey meaning throughout an interaction. Poor global coherence can significantly impair the user's understanding of the discourse as a cohesive whole. As illustrated in Figure 1, the dialogue demonstrates various common issues encountered in personalized dialogue systems, including local and global incoherence as well as persona inconsistency.

This study introduces a novel approach utilizing graph learning to maintain both local and global coherence, along with coherence-guided prompt learning and coherence-aware attention mechanisms that use discourse relations as signals for conditional response generation. This approach aims to generate responses that are coherent with the context and consistent with the persona, thus enhancing the naturalness of the personalized dialogue generation.

Our contributions are summarized as follows:

- We introduce the novel framework **MUDI** (Multiple Discourse Relations Graph Learning), which enhances the naturalness of Personalized Dialogue Generation. To the best of our knowledge, **MUDI** is the first framework to jointly integrate Discourse Relations and Persona in Personalized Dialogue Generation.
- We utilize Prompt Learning and propose a Coherence-aware Attention mechanism to integrate discourse information, thereby guiding the conditional response generation process.
- We leverage Dynamic Weighting Aggregation to balance the discourse and persona information.
- Through extensive experiments and a case study on the ConvAI2 dataset, our model demonstrates performance that is comparable to, or even surpasses, established baselines. Our approach enables the generation of responses that are more natural, enhancing the overall personalized conversational quality.

Related Work

Persona-based Dialogue Generation

As open-domain dialogue generation has matured, the focus has shifted towards personalization to enhance engagement. The introduction of the PersonaChat dataset by Zhang et al. marked a significant milestone, initiating studies into embedding explicit persona traits into dialogue responses (Zhang et al. 2018a). This work was expanded by the creation of the ConvAI2 dataset by Dinan et al. (Dinan et al. 2020), which serves as a key benchmark in persona-based dialogue tasks. Jang et al. further enriched this domain by proposing a FoCus dataset that incorporates background knowledge alongside persona traits (Jang et al. 2022).

Before the advent of large personalized dialogue datasets (Zhang et al. 2018a), researchers explored diversifying generated responses by incorporating speaker information into models. For example, (Li et al. 2016b; Al-Rfou et al. 2016) defined a persona as a combination of background facts about a user, coupled with their language behavior and style of interaction. They integrated speaker information into dialogue generation by learning speaker embeddings.

With the advent of large datasets and advanced pre-trained language models (PLMs), new methods have emerged. For instance, Zhang et al. employed LSTMs to fuse persona with contextual information (Zhang et al. 2018a). TransferTransfo fine-tuned GPT-2 with concatenated persona and dialogue inputs (Wolf et al. 2019), while BoB utilized three BERT models and the MNLI dataset to improve response relevance and consistency (Song et al. 2021; Williams, Nangia, and Bowman 2018), this method of employing NLI datasets for consistency-learning has also become a commonly used technique in subsequent research. P^2BOT introduced a unique architecture that enhances mutual persona perception (Liu et al. 2020). Despite these advancements, maintaining persona consistency alongside coherence in responses remains a challenge. A few studies, like LMEDR, have attempted to address this by leveraging entailment and latent memory for discourse understanding (Chen et al. 2023), yet the effectiveness of relying on implication relations alone for coherence has been limited. Therefore, while current methods align responses with personas, there is substantial scope for improving discourse coherence evaluation.

Methodology

Task Formulation

The task involves generating a personalized response, denoted as $r_{|K|}$, given the persona descriptions $P = \{p_1, p_2, ..., p_{|P|}\}$ and a multi-turn dialogue context $C =$

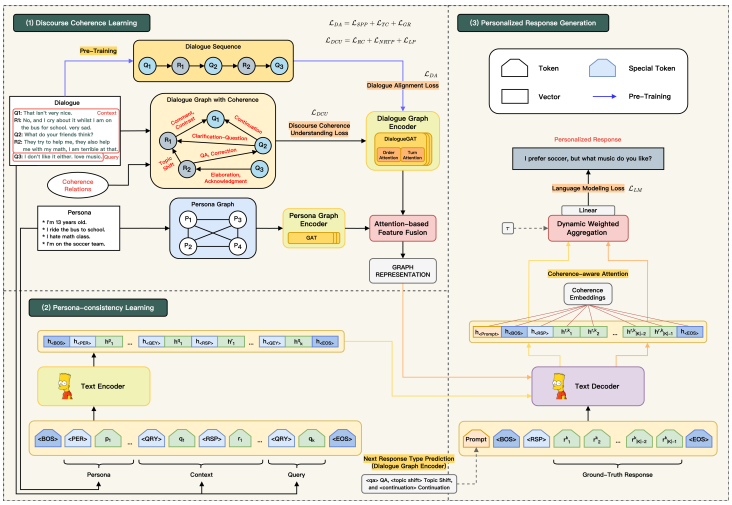

Figure 2: The overall architecture of our model - MUDI.

$\{q_1, r_1, q_2, r_2, ..., q_{|K|1}, r_{|K|1}, q_{|K|}\}$. In this context, q and r represent the user query and the chatbot response, respectively. The core goal of personalized response generation is to accurately estimate the probability distribution $p(r|C, P)$, facilitating the generation of specifically tailored responses that reflect the persona information and dialogue history.

An ideal personalized response should be not only natural but also consistent with the persona. To generate a more coherent response, we incorporate the discourse relations in dialogue. With specific response types $T = \{t_1, t_2, ..., t_{|T|}\}$ identified, our goal extends to producing a response $r_{|K|}$ that seamlessly integrates these types across the dialogue. Consequently, we aim to optimize the probability $p(r|C, P, T)$, enhancing both the personalization and coherence of the responses generated.

Model Overview

We propose a framework based on Multiple Discourse Relations Graph Learning, namely **MUDI**, as illustrated in Fig. 2. Initially, we discuss Discourse Coherence Learning, which employs a dialogue-enhanced graph encoder to learn coherence information from the dialogue. Subsequently, we explore Persona Consistency Learning, examining the relationship between persona and context. Finally, we describe our Personalized Response Generation process, which integrates information from the aforementioned steps. This process utilizes a coherence-aware mechanism to guide the generation process and balance the persona and coherence information effectively.

Discourse Coherence Learning

Our method leverages discourse relations to enhance the coherence of dialogue generation. We employ a Graph Neural Network (GNN) model specifically designed to learn these relations. To further improve the GNN's ability to understand dialogue structure, we have enhanced the existing model by incorporating a mechanism to capture dialogue structure. Additionally, we adopt a pretrain-finetune strategy to optimize performance. The detailed description of these enhancements is as follows.

Coherence Relations Annotation To facilitate the model's understanding of how two sentences in a conversation are effectively and coherently connected, we employ Large Language Models (LLMs) such as GPT-4, Mixtral-8x7b, and LLaMA-3 to assist in annotating coherence relations. According to STAC (Asher et al. 2016), there are 16 discourse relations proposed. To these, we add a **topic-shift** to represent coherent topic transitions between conversations. Each pair of utterances can be annotated with up to three different relations.

Dialogue Graph Modeling To enable the model to capture discourse coherence information within conversations when generating responses, and inspired by the success of previous graph-based discourse modeling efforts (Dong, Mircea, and Cheung 2021; Feng, Feng, and Geng 2021; Li et al. 2021), we employ a Graph encoder to learn the interactive relationships between discourses. To account for sentence-level semantics, we utilize the Sentence-Transformer (Reimers and Gurevych 2019) as an encoder to extract contextualized global semantics from both utterances and personas, thereby initializing the node features.

In our method, we found that the powerful GNN models that have been proposed are not specifically designed for dialogue structure and may not fully capture the intricate structure and complex long-term interactions in conversations. To overcome this, we enhance the GATv2 (Brody, Alon, and Yahav 2022) model by incorporating structures that specifically capture dialogue information. Specifically, we introduce two key modifications to capture the dialogue structure: Order information and Turn information, both integrated via an attention mechanism. We call this dialogue-enhanced GNN is **DialogueGAT**. The specific introduction to these two mechanisms is as follows:

Order-Attention To model the sequential nature of dialogues, we introduce auxiliary edges connecting each ut-

terance to its k-hop neighboring utterances based on their order. This indicator could formalized as Eq. 1. Then $d = k + 1$, where d represents the difference.

$$I(i,j,d) = \begin{cases} 1 & \text{if } \text{order}(j) > \text{order}(i) \\ & \quad \text{and } |\,\text{order}(i) - \text{order}(j)| < d \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

The attention scores between nodes are calculated based on the exponential decay of the order difference, as described in Eq. 2 and 3. Here, λ represents the decay rate. Moreover, wZe utilize the standard methods for computing attention scores and updating hidden states.

$$s_{ij} = \exp(-\lambda \cdot |\,\text{order}(i) - \text{order}(j)|) \cdot I(i,j,d) \quad (2)$$

$$e(h_i, h_j) = (\alpha^T \cdot \text{LeakyReLU}(W \cdot [h_i \parallel h_j])) \cdot s_{ij} \quad (3)$$

Turn-Attention We also incorporate turn information by adding bidirectional auxiliary edges between utterance nodes within the same turn, as described in Eq. 4

$$t_{ij} = \begin{cases} 1 & \text{if turn}(i) = \text{turn}(j) \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

Then, we calculate the attention scores between nodes of the same conversational turn in the same manner. These calculations are detailed in Eq. 5.

$$e(h_i, h_j) = (\alpha^T \cdot \text{LeakyReLU}(W \cdot [h_i \parallel h_j])) \cdot t_{ij} \quad (5)$$

Pre-training Phase During the pre-training phase, our objective is to enhance the Graph encoder's ability to effectively comprehend and capture dialogue structure. To achieve this, we perform pre-training on the large-scale dialogue dataset Reddit Conversation Corpus (5-turns) (Dziri et al. 2019).

$$H = \text{GNN}_\theta(X^{\text{pre}}, A^{\text{pre}}) \quad (6)$$

Drawing inspiration from the strategies outlined in Wu et al. (Wu et al. 2023), we have designed three specific self-supervised pretraining tasks to aid the model in understanding the intricate dialogue structure. These tasks include:

- **Shortest Path Prediction**: This task enables the model to infer the most direct connections within dialogue sequences, using Mean Squared Error loss to enhance its ability to identify and predict the shortest paths between dialogue nodes. The associated loss is denoted as $\mathcal{L}_{\text{SPP}}$.
- **Turn Classification**: This task helps recognize the speaker's changes and continuities within dialogues. It is optimized with Cross-Entropy loss to accurately classify whether successive utterances belong to the same speaker. The loss denoted as $\mathcal{L}_{\text{TC}}$.
- **Graph Reconstruction**: This task aimed at enabling the model to rebuild dialogue sequences from scattered data points, this task utilizes Cross-Entropy loss to improve the model's capacity for reconstructing accurate dialogue structures. The loss denoted as $\mathcal{L}_{\text{GR}}$.

The total loss for the pre-training tasks, denoted as $\mathcal{L}_{\text{DA}}$ (Dialogue Alignment) is then given by the sum of these individual losses:

$$\mathcal{L}_{\text{DA}} = \mathcal{L}_{\text{SPP}} + \mathcal{L}_{\text{TC}} + \mathcal{L}_{\text{GR}} \quad (7)$$

Fine-tuning Phase During the fine-tuning stage, we use a personalized dialogue dataset annotated with coherence relations R to refine our pretrained Graph Encoder, GNN_θ. Each node connects to its k-hop neighbors, reflecting the local conversational structure and aiding in capturing dialogue dynamics. To address class imbalance in coherence relations, such as the overrepresented "Topic Shift," we prune edges labeled exclusively with high-frequency categories to balance class distribution and enhance the model's understanding of diverse conversational patterns. The fine-tuning process updates the encoder to $\text{GNN}_{\theta'}$, better adapting it to personalized dialogues as shown:

$$H_{\text{C}} = \text{GNN}_{\theta'}(X_C^{\text{ft}}, A_C^{\text{ft}}, R) \quad (8)$$

In addition, we transform persona sentences from the dialogue into a complete graph, allowing us to leverage a GAT (Brody, Alon, and Yahav 2022), denoted as GNN_ψ. This graph encoder applies its attention mechanism across all connections to better capture the nuances and importance of each persona sentence, enhancing sensitivity to persona-specific information.

$$H_{\text{P}} = \text{GNN}_\psi(X_P, A_P) \quad (9)$$

Next, we employ an attention-based feature fusion mechanism to integrate the utterance node representations of a specific speaker in the dialogue graph with those in the persona graph. This allows the model to focus on persona information relevant to the utterance. The feature fusion is performed using a multi-head attention mechanism to obtain personalized node representations, denoted as H_{D}.

$$\begin{aligned} H_{\text{D}} &= \text{MultiHead}(Q, K, V) \\ \text{where} \quad Q &= H_{\text{C}} \cdot W_i^Q, \quad K = H_{\text{P}} \cdot W_i^K, \\ V &= H_{\text{P}} \cdot W_i^V \end{aligned} \quad (10)$$

Furthermore, we learn coherence through three tasks:

- **Coherence Relations Classification**: This task involves multi-label classification where the graph encoder predicts which of the 17 coherence relations exist between two nodes. The task utilizes Cross-Entropy for the loss, denoted as $\mathcal{L}_{\text{RC}}$, to enhance the encoder's ability to identify and distinguish relation types within the dialogue. We encountered severe label imbalance in coherence relations, notably with Topic Shift being overly dominant. To address this, we adopted the Class-balance loss from (Cui et al. 2019), which adjusts weights between classes to reduce bias toward frequently occurring labels.
- **Next Response Type Prediction**: This task aims to predict the possible types of the next response. We first derive node representations from H_D with direct sequential relationships in the dialogue to form a sequence S. The model then uses two approaches for prediction: (1) Direct Prediction, where it determines the next response type

based solely on the current utterance node, and (2) Sequential Prediction (auto-regressive style), where it considers all previous utterances. The losses for these methods, denoted as $\mathcal{L}_{\text{NRTP}}^{\text{direct}}$ and $\mathcal{L}_{\text{NRTP}}^{\text{seq}}$, are calculated using Cross-Entropy.

- **Link Prediction**: This task is similar to the Graph Reconstruction task in the pretraining phase, aiming to capture the discourse structure of the dialogue. The model predicts whether an edge exists between two adjacent utterance nodes in the dialogue graph, thereby capturing the underlying structure and coherence relations between the utterances. The prediction method follows the same approach as the Graph Reconstruction task, using an inner product decoder to estimate the existence of an edge based on the representations of two nodes. The link prediction loss, $\mathcal{L}_{\text{LP}}$, is then calculated using Cross-Entropy.

In summary, through the above training process, we enhance the Dialogue Graph Encoder's ability to understand the structure of dialogues and improve its capability to grasp the implicit discourse relations between utterances. Considering the fine-tuning tasks and their respective losses, the total loss $\mathcal{L}_{\text{DCU}}$ (Discourse Coherence Understanding) of this Dialogue Graph Encoder is then given by the weighted sum of these individual losses:

$$\mathcal{L}_{\text{DCU}} = \alpha\mathcal{L}_{\text{RC}} + \beta\mathcal{L}_{\text{NRTP}}^{\text{direct}} + \gamma\mathcal{L}_{\text{NRTP}}^{\text{seq}} + \delta\mathcal{L}_{\text{LP}} \tag{11}$$

where α, β, γ, and δ are the weights for the respective loss components.

Persona-consistency Learning

In this stage, the objective is to learn the implicit relationships between persona and dialogue. We use BART (Lewis et al. 2020) as the backbone model for this stage and for subsequent personalized response generation. Following the approaches used in previous research (Chen et al. 2023), the input to the Text encoder is the concatenation of the persona descriptions P and dialogue context C, which is structured as follows: $E_{\text{TextEncoder}} = [e_{\text{[BOS]}}, e_{\text{[PER]}}, e_{\text{p1}}, e_{\text{p2}}, ..., e_{\text{[QRY]}}, e_{\text{q1}}, e_{\text{[RSP]}}, e_{\text{r1}}, ..., e_{\text{[QRY]}}, e_{\text{q}|\text{K}|}, e_{\text{[EOS]}}]$. where [PER], [QRY], and [RSP] are three special tokens that indicate the beginning of persona, query, and response, respectively.

Personalized Response Generation

After completing the dialogue representation learning processes, we obtain coherence-aware dialogue representations through the Dialogue Graph Encoder and persona-aware representations through the Text Encoder. Our objective is to generate personalized responses informed by these implicit representations. Initially, we employ a prompt-based conditional dialogue generation approach, designing a prompt that provides guiding signals for the response generation process. Building on the predictions from the Dialogue Graph Encoder about the next response type, our Prompt Tuning module generates a comprehensive description. This description instructs the response generator on task approach and guides the generation process by specifying response types, leveraging both dialogue context and persona information. Therefore, the input sequence for the personalized response generator (Text Decoder) is structured as follows: $E_{\text{Generator}} = [e_{\text{[PROMPT]}}, e_{\text{[BOS]}}, e_{\text{[RSP]}}, e_1^{\text{k}}, e_2^{\text{k}}, ..., e_{|\text{K}|-1}^{\text{k}}, e_{|\text{K}|}^{\text{k}}, e_{\text{[EOS]}}]$.

To enhance coherence in the next token prediction, we apply cross-attention to dialogue representations from the Dialogue Graph Encoder at each transformer block. Consequently, each layer of the decoder performs cross-attention on both the standard encoder outputs and the coherence-aware dialogue representation. Furthermore, we introduce a Coherence-aware Attention mechanism that employs learnable embeddings to capture the semantic nuances of coherence relations. Special tokens representing these relations are integrated into the prompt, combining their embeddings with those of coherence relations. This approach enables the generator to consider the type of response being predicted, such as words aligning with an Acknowledgment response type. This dual cross-attention mechanism allows the decoder to leverage both context and persona information more effectively, enhancing the coherence and personalization of generated responses. Finally, inspired by (Huang et al. 2023), we enhance the generator's ability to incorporate both persona-aware and coherence-aware context information during generation through a Dynamic Weighted Aggregation mechanism. Specifically, we aim to address the trade-off between coherence and persona consistency by balancing the information from the Graph Encoder and Text Encoder. We use a threshold hyperparameter τ and a learnable mask to control the proportion of persona-aware and coherence-aware information considered.

For personalized response generation, we use a language modeling task to compute the probability distribution over the vocabulary for generating the next word given the current context. This approach encourages the model to generate responses that are contextually relevant. By minimizing this loss function, the model learns to produce responses that are fluent, personalized, coherent with the dialogue history, and consistent with the persona. The loss function can be formally defined as:

$$\mathcal{L}_{\text{LM}} = -\sum_{t=1}^{T} \log P(y_t | y_{<t}, O_{\text{Generator}}) \tag{12}$$

Experiments

Experimental Setup

Dataset We perform experiments on the ConvAI2 dataset. ConvAI2 (Dinan et al. 2020) is a chit-chat dataset based on PersonaChat (Zhang et al. 2018a). The dataset comprises 17,878 and 1,000 multi-turn dialogues for the training and development sets, respectively, with a total of 131,438 and 7,801 utterances. It features 1,155 and 100 unique persona descriptions for the training and development sets, respectively. Each conversation participant who are assigned at least five persona descriptions, which are selected from these unique personas.

Baselines For comparison, we have selected the following baselines: (1) **Personalized Dialogue Generation**

Model		Text Similarity			Personalization	Coherence
		BLEU-1 ↑	ROUGE-1 ↑	BERTScore ↑	C.Score ↑	QuantiDCE ↑
Large Language Model (Prompting)						
GPT-4		7.47	13.52	84.05	2.86	3.41 / 2.92
General Dialogue Generation						
DialoGPT		7.34	9.46	83.31	4.53	**3.23** / 2.79
PLATO	w/ persona	4.35	4.88	82.77	0.56	1.68 / 1.57
	w/o persona	6.82	4.99	81.44	0.18	1.87 / 1.77
Persona-based Dialogue Generation						
BoB		15.30	13.21	83.77	0.51	2.99 /2.76
LMEDR		15.47	13.28	85.00	7.83	2.89 / 2.90
PAA		<u>16.55</u>	13.53	84.42	**15.19**	2.70 / <u>2.93</u>
MUDI (ours)	$SP_{\tau=0.2}$	15.14	14.87	85.07	<u>11.87</u>	3.05 / 2.84
	$Emb_{\tau=0.2}$	<u>16.55</u>	**17.10**	<u>85.42</u>	9.70	**3.23** / **2.94**
	$SP+Emb_{\tau=0.2}$	**18.19**	<u>16.59</u>	**85.53**	9.75	<u>3.21</u> / 2.92

Table 1: Automatic evaluation results on ConvAI2 dataset over our implemented approach. The best results in each column are in bold, while the second is underlined.

methods: We compare our approach with established text-description-based personalized models such as BOB (Song et al. 2021), LMEDR (Chen et al. 2023), and PAA (Huang et al. 2023), which are recognized for their strong performance. (2) **General Dialogue Generation methods**: To ensure a fair evaluation, we include PLATO (Bao et al. 2020) and DialoGPT (Zhang et al. 2020). For PLATO, following the original publication's methodology, we prepend the persona descriptions as part of the knowledge to the entire context during inference. For DialoGPT, we adopt a post-processing approach as suggested by (Zhou et al. 2023). We first generate multiple responses as candidates from DialoGPT for the same context and then select the response that is most consistent with the persona and most relevant to the context as the final output. (3) **Large Language Models (LLMs)**: We also test the ability of LLMs, specifically GPT-4, in personalized response generation, which utilizes prompting technologies.

Implementation Details The **MUDI**'s backbone Generator is BART-large[1]. For Generator training, we train it on 1 NVIDIA A100 80GB GPU. For the Dialogue Graph Encoder training, it is conducted on a single NVIDIA RTX 4090 GPU. In the Dialogue Graph Encoder, we initially employ the SBERT model[2] to encode both the utterances and persona sentences, thereby initializing the node embeddings. We construct the Dialogue Graph by keeping the 3-hop neighbors. The model employs a 2-layer GNN. For training the generator, we retain the most recent 5 turns of dialogue as historical context and choose the top-3 predicted response types for the prompt. We set $\tau = 0.2$ for the Dynamic Weighted Aggregation. Other detailed configurations of each method are given in Appendix[3].

Automatic Evaluation We assess the quality of dialogue responses from four perspectives: (1) **Text-similarity**: We using BLEU (Papineni et al. 2002), ROUGE (Lin 2004), and BERTScore (Zhang* et al. 2020) to evaluate lexical overlap and semantic similarity between generated and ground-truth responses. (2) **Diversity**: Measured using Distinct-n (Dist-n) (Li et al. 2016a) metrics for unique n-grams, Entropy-n (Ent-n) (Zhang et al. 2018b) to assess n-gram distribution, and Unique Sentence Ratio (USR) (Li, Zhang, and Chen 2020) to evaluate sentence uniqueness. (3) **Coherence**: We use QuantiDCE metrics (Ye et al. 2021) to assess dialogue coherence, focusing on how responses align with preceding queries and the overall context, and the unity of utterances within the conversation. This approach addresses the limitations of traditional fluency-based evaluations. (4) **Personalization**: We use Consistency Score (C.Score) (Madotto et al. 2019) to evaluate how well responses align with the persona based on predictions from an NLI model.

Main Results

In Table we report experimental results on Text Similarity, Personalization, and Coherence evaluation. Our method offers better BLEU, ROUGE, and BERTScore compared with baseline methods. Specifically, MUDI's BLEU-1 and ROUGE-1 scores reach 18.19 and 17.10, respectively, outperforming existing methods by 1.64 and 3.57. As shown in Table 6, we report the Diversity evaluation. MUDI's USR

[1]https://huggingface.co/facebook/bart-large

[2]https://huggingface.co/sentence-transformers/all-mpnet-base-v2

[3]The detailed appendix will be available online.

Model		Diversity		
		Ent-1 ↑	Dist-1 ↑	USR ↑
Persona-based Dialogue Generation				
BoB		7.89	41.75	0.99
LMEDR		7.14	43.08	0.94
PAA		6.66	40.27	0.87
MUDI (ours)	$SP_{\tau=0.2}$	**8.13**	46.76	**1.00**
	$Emb_{\tau=0.2}$	7.65	**51.03**	**1.00**
	$SP+Emb_{\tau=0.2}$	7.66	47.68	**1.00**

Table 2: Automatic evaluation results for diversity on the ConvAI2 dataset. Best results are bolded and second-best are underlined.

is 1.0, indicating that it can generate completely unique responses under different queries and personas. Furthermore, our approach achieves the highest scores among all Persona-based dialogue generation methods and significantly surpasses other baselines in Dist-1, outperforming them by 7.95. This performance establishes our method could generate varied and engaging responses. In addition, the results of the Coherence evaluation. Compared to other persona-based methods, MUDI has made significant progress in QuantiDCE, particularly in assessing the coherence between the query and response (left-side scores). This indicates that our approach indeed enables the model to generate responses with enhanced local coherence. Furthermore, MUDI also achieves excellent results in global coherence, which evaluates the coherence between the entire dialogue context and the response (right-side scores). Finally, the results of evaluating Personalization and Feature Coverage. PAA significantly outperforms other methods in scores for Personalization. Upon further examination, we discovered that this is because PAA frequently generates sentences that are exact restatements of the persona description, often ignoring the relevance to the query. As a result, its high scores in Personalization can be attributed to this tendency. Excluding the special case of PAA, MUDI achieves excellent results in Personalization compared to other methods. Combined with the previously discussed results from the Coherence evaluation, this demonstrates that our approach successfully balances discourse relations and persona. It generates responses that effectively consider both aspects simultaneously. Furthermore, our model achieves comparable scores in the Coherence evaluation compared to DialoGPT, which focuses on general dialogue generation.

In summary, compared to existing methods, our approach MUDI not only significantly improves performance in Text Similarity scores but also excels at integrating discourse relations and persona information. This enables us to generate personalized responses that are not only rich in content and diverse but also encompass these aspects. Moreover, in Coherence evaluations, our method achieves scores comparable to those of state-of-the-art models specialized in open-domain dialogue generation.

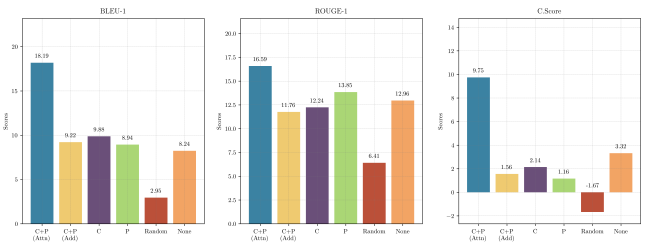

Figure 3: Performance analysis of Dialogue Graph Encoder across different settings. Here, "C" represents Context, and "P" represents Persona.

Analysis

The Effect of Dialogue Graph Encoder We evaluated the Dialogue Graph Encoder's effectiveness under various configurations: (1) **Context+Persona (Attention)** using an Attention-based Feature Fusion, (2) **Context+Persona (Add)** with a simple addition of representations, (3) **Context** only, (4) **Persona** only, (5) **Random** vector substitution, and (6) **None** where the encoder is removed. Analysis results in Figure 3 show that the attention-based approach excels in BLEU-1, ROUGE-1, and C.Socre metrics, emphasizing the importance of integrating contextual and persona information effectively. Methods that focus solely on one aspect or do not employ integration decrease performance, highlighting the necessity of meaningful input for coherent responses.

For more detailed experimental results, analysis, and case studies, please refer to the Appendix.

Conclusion

In this work, we propose a new method, **MUDI**, to effectively model discourse relations in personalized dialogue generation. To the best of our knowledge, **MUDI** is the first framework to jointly integrate Discourse Relations and Persona in Personalized Dialogue. Firstly, we propose DialogueGAT, a dialogue-enhanced GNN, as a Dialogue Graph Encoder, designed to capture dialogue structure and contextual discourse relations. Additionally, we utilize an Attention-Based Feature Fusion method to effectively integrate context relations and persona information. We further enhance our model by employing a Text Encoder to capture persona-aware dialogue representations. We increase the decoder's ability to consider coherent information while predicting the next token by leveraging both a prompt-based conditional dialogue generation mechanism, which uses prompts to guide the response generation process, and our coherence-aware attention mechanism, which incorporates learnable embeddings and token representations. Finally, we leverage Dynamic Weighting Aggregation to balance the information between coherence-aware and persona-aware dialogue representations, ensuring a robust integration of both elements. Extensive experiments and analyses demonstrate that our method, **MUDI**, significantly improves the quality of personalized responses by making them more coherent, informative, and aligned with the user's persona traits, as well as more human-like.

References

AI@Meta. 2024. Llama 3 Model Card.

Al-Rfou, R.; Pickett, M.; Snaider, J.; Sung, Y.-H.; Strope, B.; and Kurzweil, R. 2016. Conversational contextual cues: The case of personalization and history for response ranking. *arXiv preprint arXiv:1606.00372*.

Asher, N.; Hunter, J.; Morey, M.; Farah, B.; and Afantenos, S. 2016. Discourse Structure and Dialogue Acts in Multiparty Dialogue: the STAC Corpus. In Calzolari, N.; Choukri, K.; Declerck, T.; Goggi, S.; Grobelnik, M.; Maegaard, B.; Mariani, J.; Mazo, H.; Moreno, A.; Odijk, J.; and Piperidis, S., eds., *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 2721–2727. Portorož, Slovenia: European Language Resources Association (ELRA).

Bao, S.; He, H.; Wang, F.; Wu, H.; and Wang, H. 2020. PLATO: Pre-trained Dialogue Generation Model with Discrete Latent Variable. In Jurafsky, D.; Chai, J.; Schluter, N.; and Tetreault, J., eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 85–96. Online: Association for Computational Linguistics.

Brody, S.; Alon, U.; and Yahav, E. 2022. How Attentive are Graph Attention Networks? In *International Conference on Learning Representations*.

Chen, R.; Wang, J.; Yu, L.-C.; and Zhang, X. 2023. Learning to memorize entailment and discourse relations for persona-consistent dialogues. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press. ISBN 978-1-57735-880-0.

Cui, Y.; Jia, M.; Lin, T.; Song, Y.; and Belongie, S. 2019. Class-Balanced Loss Based on Effective Number of Samples. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9260–9269. Los Alamitos, CA, USA: IEEE Computer Society.

Dinan, E.; Logacheva, V.; Malykh, V.; Miller, A.; Shuster, K.; Urbanek, J.; Kiela, D.; Szlam, A.; Serban, I.; Lowe, R.; Prabhumoye, S.; Black, A. W.; Rudnicky, A.; Williams, J.; Pineau, J.; Burtsev, M.; and Weston, J. 2020. The Second Conversational Intelligence Challenge (ConvAI2). In Escalera, S.; and Herbrich, R., eds., *The NeurIPS '18 Competition*, 187–208. Cham: Springer International Publishing. ISBN 978-3-030-29135-8.

Dong, Y.; Mircea, A.; and Cheung, J. C. K. 2021. Discourse-Aware Unsupervised Summarization for Long Scientific Documents. In Merlo, P.; Tiedemann, J.; and Tsarfaty, R., eds., *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 1089–1102. Online: Association for Computational Linguistics.

Dziri, N.; Kamalloo, E.; Mathewson, K.; and Zaiane, O. 2019. Augmenting Neural Response Generation with Context-Aware Topical Attention. In Chen, Y.-N.; Bedrax-Weiss, T.; Hakkani-Tur, D.; Kumar, A.; Lewis, M.; Luong, T.-M.; Su, P.-H.; and Wen, T.-H., eds., *Proceedings of the First Workshop on NLP for Conversational AI*, 18–31. Florence, Italy: Association for Computational Linguistics.

Feng, X.; Feng, X.; and Geng, X. 2021. Dialogue Discourse-Aware Graph Model and Data Augmentation for Meeting Summarization. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI-21)*, 3808–3814.

Ghazarian, S.; Wen, N.; Galstyan, A.; and Peng, N. 2022. DEAM: Dialogue Coherence Evaluation using AMR-based Semantic Manipulations. In Muresan, S.; Nakov, P.; and Villavicencio, A., eds., *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 771–785. Dublin, Ireland: Association for Computational Linguistics.

Huang, Q.; Zhang, Y.; Ko, T.; Liu, X.; Wu, B.; Wang, W.; and Tang, H. 2023. Personalized dialogue generation with persona-adaptive attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 12916–12923.

Jang, Y.; Lim, J.; Hur, Y.; Oh, D.; Son, S.; Lee, Y.; Shin, D.; Kim, S.; and Lim, H. 2022. Call for Customized Conversation: Customized Conversation Grounding Persona and Knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 10803–10812.

Jiang, S.; and de Rijke, M. 2018. Why are Sequence-to-Sequence Models So Dull? Understanding the Low-Diversity Problem of Chatbots. In Chuklin, A.; Dalton, J.; Kiseleva, J.; Borisov, A.; and Burtsev, M., eds., *Proceedings of the 2018 EMNLP Workshop SCAI: The 2nd International Workshop on Search-Oriented Conversational AI*, 81–86. Brussels, Belgium: Association for Computational Linguistics.

Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; and Zettlemoyer, L. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In Jurafsky, D.; Chai, J.; Schluter, N.; and Tetreault, J., eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7871–7880. Online: Association for Computational Linguistics.

Li, J.; Galley, M.; Brockett, C.; Gao, J.; and Dolan, B. 2016a. A Diversity-Promoting Objective Function for Neural Conversation Models. In Knight, K.; Nenkova, A.; and Rambow, O., eds., *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 110–119. San Diego, California: Association for Computational Linguistics.

Li, J.; Galley, M.; Brockett, C.; Spithourakis, G.; Gao, J.; and Dolan, B. 2016b. A Persona-Based Neural Conversation Model. In Erk, K.; and Smith, N. A., eds., *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 994–1003. Berlin, Germany: Association for Computational Linguistics.

Li, J.; Liu, M.; Zheng, Z.; Zhang, H.; Qin, B.; Kan, M.-Y.; and Liu, T. 2021. DADgraph: A Discourse-aware Dialogue

Graph Neural Network for Multiparty Dialogue Machine Reading Comprehension. *2021 International Joint Conference on Neural Networks (IJCNN)*, 1–8.

Li, L.; Zhang, Y.; and Chen, L. 2020. Generate Neural Template Explanations for Recommendation. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, 755–764. New York, NY, USA: Association for Computing Machinery. ISBN 9781450368599.

Lin, C.-Y. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, 74–81. Barcelona, Spain: Association for Computational Linguistics.

Liu, Q.; Chen, Y.; Chen, B.; Lou, J.-G.; Chen, Z.; Zhou, B.; and Zhang, D. 2020. You Impress Me: Dialogue Generation via Mutual Persona Perception. In Jurafsky, D.; Chai, J.; Schluter, N.; and Tetreault, J., eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 1417–1427. Online: Association for Computational Linguistics.

Madotto, A.; Lin, Z.; Wu, C.-S.; and Fung, P. 2019. Personalizing Dialogue Agents via Meta-Learning. In Korhonen, A.; Traum, D.; and Màrquez, L., eds., *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 5454–5459. Florence, Italy: Association for Computational Linguistics.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In Isabelle, P.; Charniak, E.; and Lin, D., eds., *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 311–318. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics.

Reimers, N.; and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Song, H.; Wang, Y.; Zhang, K.; Zhang, W.-N.; and Liu, T. 2021. BoB: BERT Over BERT for Training Persona-based Dialogue Models from Limited Personalized Data. In Zong, C.; Xia, F.; Li, W.; and Navigli, R., eds., *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 167–177. Online: Association for Computational Linguistics.

Song, H.; Zhang, W.-N.; Hu, J.; and Liu, T. 2020. Generating Persona Consistent Dialogues by Exploiting Natural Language Inference. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05): 8878–8885.

Warren, M. J. 2006. Features of Naturalness in Conversation.

Williams, A.; Nangia, N.; and Bowman, S. 2018. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In Walker, M.; Ji, H.; and Stent, A., eds., *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 1112–1122. New Orleans, Louisiana: Association for Computational Linguistics.

Wolf, T.; Sanh, V.; Chaumond, J.; and Delangue, C. 2019. Transfertransfo: A transfer learning approach for neural network based conversational agents. *arXiv preprint arXiv:1901.08149*.

Wu, L.; Lin, H.; Tan, C.; Gao, Z.; and Li, S. Z. 2023. Self-Supervised Learning on Graphs: Contrastive, Generative, or Predictive. *IEEE Trans. on Knowl. and Data Eng.*, 35(4): 4216–4235.

Ye, Z.; Lu, L.; Huang, L.; Lin, L.; and Liang, X. 2021. Towards Quantifiable Dialogue Coherence Evaluation. In Zong, C.; Xia, F.; Li, W.; and Navigli, R., eds., *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2718–2729. Online: Association for Computational Linguistics.

Yuan, W.; Neubig, G.; and Liu, P. 2021. BARTScore: Evaluating Generated Text as Text Generation. In Ranzato, M.; Beygelzimer, A.; Dauphin, Y.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems*, volume 34, 27263–27277. Curran Associates, Inc.

Zhang, S.; Dinan, E.; Urbanek, J.; Szlam, A.; Kiela, D.; and Weston, J. 2018a. Personalizing Dialogue Agents: I have a dog, do you have pets too? In Gurevych, I.; and Miyao, Y., eds., *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2204–2213. Melbourne, Australia: Association for Computational Linguistics.

Zhang*, T.; Kishore*, V.; Wu*, F.; Weinberger, K. Q.; and Artzi, Y. 2020. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*.

Zhang, Y.; Galley, M.; Gao, J.; Gan, Z.; Li, X.; Brockett, C.; and Dolan, B. 2018b. Generating Informative and Diverse Conversational Responses via Adversarial Information Maximization. In Bengio, S.; Wallach, H.; Larochelle, H.; Grauman, K.; Cesa-Bianchi, N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Zhang, Y.; Sun, S.; Galley, M.; Chen, Y.-C.; Brockett, C.; Gao, X.; Gao, J.; Liu, J.; and Dolan, B. 2020. DialoGPT: Large-Scale Generative Pre-training for Conversational Response Generation. In *ACL, system demonstration*.

Zhou, J.; Pang, L.; Shen, H.; and Cheng, X. 2023. SimOAP: Improve Coherence and Consistency in Persona-based Dialogue Generation via Over-sampling and Post-evaluation. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 9945–9959. Toronto, Canada: Association for Computational Linguistics.

Reproducibility Checklist

This paper:

- Includes a conceptual outline and/or pseudocode description of AI methods introduced (yes)
- Clearly delineates statements that are opinions, hypothesis, and speculation from objective facts and results (yes)
- Provides well marked pedagogical references for less-familiare readers to gain background necessary to replicate the paper (yes)

Does this paper make theoretical contributions? (no)
If yes, please complete the list below.

- All assumptions and restrictions are stated clearly and formally. (yes/partial/no)
- All novel claims are stated formally (e.g., in theorem statements). (yes/partial/no)
- Proofs of all novel claims are included. (yes/partial/no)
- Proof sketches or intuitions are given for complex and/or novel results. (yes/partial/no)
- Appropriate citations to theoretical tools used are given. (yes/partial/no)
- All theoretical claims are demonstrated empirically to hold. (yes/partial/no/NA)
- All experimental code used to eliminate or disprove claims is included. (yes/no/NA)

Does this paper rely on one or more datasets? (yes)
If yes, please complete the list below.

- A motivation is given for why the experiments are conducted on the selected datasets (yes)
- All novel datasets introduced in this paper are included in a data appendix. (yes)
- All novel datasets introduced in this paper will be made publicly available upon publication of the paper with a license that allows free usage for research purposes. (yes)
- All datasets drawn from the existing literature (potentially including authors' own previously published work) are accompanied by appropriate citations. (yes)
- All datasets drawn from the existing literature (potentially including authors' own previously published work) are publicly available. (yes)
- All datasets that are not publicly available are described in detail, with explanation why publicly available alternatives are not scientifically satisficing. (NA)

Does this paper include computational experiments? (yes)
If yes, please complete the list below.

- Any code required for pre-processing data is included in the appendix. (yes).
- All source code required for conducting and analyzing the experiments is included in a code appendix. (yes)
- All source code required for conducting and analyzing the experiments will be made publicly available upon publication of the paper with a license that allows free usage for research purposes. (yes)
- All source code implementing new methods have comments detailing the implementation, with references to the paper where each step comes from (yes)
- If an algorithm depends on randomness, then the method used for setting seeds is described in a way sufficient to allow replication of results. (yes)
- This paper specifies the computing infrastructure used for running experiments (hardware and software), including GPU/CPU models; amount of memory; operating system; names and versions of relevant software libraries and frameworks. (yes)
- This paper formally describes evaluation metrics used and explains the motivation for choosing these metrics. (yes)
- This paper states the number of algorithm runs used to compute each reported result. (yes)
- Analysis of experiments goes beyond single-dimensional summaries of performance (e.g., average; median) to include measures of variation, confidence, or other distributional information. (no)
- The significance of any improvement or decrease in performance is judged using appropriate statistical tests (e.g., Wilcoxon signed-rank). (no)
- This paper lists all final (hyper-)parameters used for each model/algorithm in the paper's experiments. (yes)
- This paper states the number and range of values tried per (hyper-) parameter during development of the paper, along with the criterion used for selecting the final parameter setting. (yes)

Appendix

Coherence Relations Annotation

To facilitate the model's understanding of how two sentences in a conversation are effectively connected, we employ Large Language Models (LLMs) such as GPT-4, Mixtral-8x7b, and LLaMA-3 to assist in annotating coherence relations. There are, in total, 16 discourse relations according to STAC (Asher et al. 2016), namely, **comment**, **clarification-question**, **elaboration**, **acknowledgment**, **continuation**, **explanation**, **conditional**, **question-answer**, **alternation**, **question-elaboration**, **result**, **background**, **narration**, **correction**, **parallel** and **contrast**. On top of these relationships, we add **topic-shift** to represent coherent topic transitions between conversations.

Each pair of utterances could be annotated with zero to three different relations. In total, we have annotated 1,942,177 pairs of utterances for their coherence relations. An example of annotated results can be seen in Table 3. The prompt for coherence relations annotations is shown in Figure 4

{SYSTEM PROMPT}

You are provided with two sentences from a dialogue. Your task is to analyze the coherence relationship between these two sentences, specifically focusing on how Sentence 2 responds to or relates to Sentence 1. Assign appropriate coherence labels to these relationships. The coherence labels you may use are:
– **Acknowledgment**: Sentence 2 acknowledges or confirms what was said in Sentence 1.
– **Elaboration**: Sentence 2 provides more details or expands on Sentence 1.
– **Explanation**: Sentence 2 explains or provides a reason for Sentence 1.
– **QA**: Sentence 2 answers a question posed in Sentence 1.
– **Comment**: Sentence 2 adds a comment to Sentence 1.
– **Clarification–Question**: Sentence 2 asks for clarification about Sentence 1.
– **Continuation**: Sentence 2 continues the idea or topic introduced in Sentence 1.
– **Conditional**: Sentence 2 provides a condition related to Sentence 1.
– **Alternation**: Sentence 2 presents an alternative to Sentence 1.
– **Question–Elaboration**: Sentence 2 elaborates on a question posed in Sentence 1.
– **Result**: Sentence 2 states a result or outcome of Sentence 1.
– **Background**: Sentence 2 provides background information related to Sentence 1.
– **Narration**: Sentence 2 narrates a sequence of events related to Sentence 1.
– **Correction**: Sentence 2 corrects information in Sentence 1.
– **Parallel**: Sentence 2 presents parallel or similar information to Sentence 1.
– **Contrast**: Sentence 2 contrasts with Sentence 1.
– **Topic Shift**: Sentence 2 shifts to a different topic from Sentence 1.

A relationship can have multiple labels if it serves multiple functions or has multiple characteristics. However, you should assign no more than four labels to any relationship. If the two sentences are incoherent or do not have any relationship (difficult to assign a coherent relationship), return "None".

Ensure that your labeling accurately reflects how Sentence 2 responds to or relates to Sentence 1. For each coherence labels separated by commas. Please refer to the following examples for the output format of the coherence labels.

First, give you an example to help you understand the task:
{DEMONSTRATIONS}

Then give you the dialogues of the task:
{INPUT}

Figure 4: The prompt of Coherence Relations Annotation.

Implementation Details

The **MUDI** is mainly implemented in PyTorch. Our backbone Generator is BART-large[2]. For Generator training, we train it using a batch size of 4 on 1 NVIDIA A100 80GB GPU via the AdamW optimizer with a learning rate of 5×10^{-6} and a weight decay of 0.01. For the Dialogue Graph Encoder training, it is conducted on a single NVIDIA RTX 4090 GPU using a batch size of 512. The training also employs the AdamW optimizer, but with a learning rate of 2×10^{-5} and the same weight decay of 0.01. In the Dialogue Graph Encoder, we initially employ the SBERT model "all-mpnet-base-v2"[3] to encode both the utterances and persona sentences, thereby initializing the node embeddings. We construct the Dialogue Graph by keeping the 3-hop neighbors. The model employs a 2-layer GNN with 4 multi-heads and a hidden dimension of 512. The weights for the different tasks are as follows: the weight of the Coherence Relation Classification task is 1.5, the weight of both types of Next Response Type Prediction tasks is 1.5, and the weight of the Link Prediction task is 1.2. For training the generator, we retain the most recent 5 turns of dialogue as historical context and choose the top-3 predicted response types for the prompt. We set $\tau = 0.2$ for the Dynamic Weighted Aggregation.

Prompt Structure for LLM Prompting in Evaluation

The prompt used for GPT-4 can be seen in Figure 5.

{SYSTEM PROMPT}

You are provided with a dialogue transcript between User1 and User2, along with User2's (your) persona information. Your task is to analyze the last question from the user in the dialogue and determine which personas are relevant for crafting a personalized response to this question.

Following the relevance determination, craft a personalized response to the user's last question based on the relevant personas. Your response should reflect the characteristics and preferences indicated by the relevant personas, and engage with the content of the user's last question in a manner that aligns with these personas. During the conversation, please ensure the natural flow and coherence of the topics, even when faced with sudden changes or situations.

First, give you an example to help you understand the task:

[Persona]

I read twenty books a year.

I'm a stunt double as my second job.

I only eat kosher.

I was raised in a single parent household.

[Dialogue]

User1: Hello what are doing today?

User2: I am good , i just got off work and tired, i have two jobs.

User1: I just got done watching a horror movie.

User2: I rather read, I've read about 20 books this year.

User1: Wow! I do love a good horror movie. Loving this cooler weather.

User2: But a good movie is always good.

User1: Yes ! my son is in junior high and i just started letting him watch them too.

User2: I work in the movies as well .

User1: Neat ! ! i used to work in the human services field

[Response]

Yes it is neat, i stunt double , it is so much fun and hard work.

Then give you the personas and dialogue history of the task:

{INPUT}

[Response]

Figure 5: The prompt of LLM inference on Personalized Dialogue Generation task.

[1]https://huggingface.co/meta-llama/Meta-Llama-3-70B-Instruct

[2]https://huggingface.co/facebook/bart-large

[3]https://huggingface.co/sentence-transformers/all-mpnet-base-v2

Index	Dialogue	
0	[PERSON 1:] Hello what are doing today?	
1	[PERSON 2:] I am good, I just got off work and tired, I have two jobs.	
2	[PERSON 1:] I just got done watching a horror movie.	
3	[PERSON 2:] Wow! I do love a good horror movie. Loving this cooler weather.	
4	[PERSON 1:] But a good movie is always good.	
5	...	
Index	**Utterance**	**Coherence Relations**
0	Hello what are doing today?	QA, Explanation
1	I am good, I just got off work and tired, I have two jobs.	
0	Hello what are doing today?	QA
2	I just got done watching a horror movie.	
...		
1	I am good, I just got off work and tired, I have two jobs.	Topic Shift
2	I just got done watching a horror movie.	
...		

Table 3: Examples of coherence relations annotated by the LLaMA-3-70B[1](AI@Meta 2024). We annotated all utterance pairs in the dialogue, and the examples shown here represent only a subset of the complete dataset.

Automatic Metrics

More specifically, aside from the evaluation metrics discussed in the Experiments chapter, we assess the quality of dialogue responses from four perspectives:

(1) **Text-similarity**: To evaluate the similarity between the generated responses and the ground-truth responses, we employ BLEU (Papineni et al. 2002) and ROUGE (Lin 2004) metrics, which focus on the word overlap level. Additionally, we utilize BERTScore (Zhang* et al. 2020), which measures the semantic similarity between generated responses and ground-truth using contextual embeddings from BERT. This helps capture nuances that traditional overlap-based metrics might miss.

(2) **Diversity**: We assess the diversity and informativeness of the generated responses at token, sentence, and corpus levels. We utilize Distinct-n (Dist-n) metrics (Li et al. 2016a) to measure the proportion of unique n-grams (n=1) relative to the total number of n-grams in the generated responses. Additionally, we employ Entropy-n (Ent-n) (Zhang et al. 2018b) to evaluate the uncertainty or randomness of the distribution of n-grams (n=1) in the generated text. We also calculate the Unique Sentence Ratio (USR) (Li, Zhang, and Chen 2020), which quantifies text diversity by measuring the proportion of unique sentences among all predicted responses.

(3) **Coherence**: Previous studies often omit explicit evaluations of Coherence, assuming that Fluency can also measure the coherence of dialogues. However, this approach does not directly account for the coherence of generated responses with the context and query, an aspect often presumed to be covered by fluency. To measure this unexplored dimension, our research incorporates specific coherence metrics to provide a more accurate and holistic assessment of response quality. Specifically, We employ several metrics. Firstly, we use QuantiDCE (Ye et al. 2021) and DEAM (Ghazarian et al. 2022), which are state-of-the-art metrics in dialogue coherence evaluation. These metrics assess both how logically the responses align with the preceding query or the entire context within the dialogue, and how well the utterances in a conversation are unified, leading to a consistent and coherent interaction. Additionally, we utilize BARTScore (Yuan, Neubig, and Liu 2021) to measure coherence under the faithfulness setting discussed in the paper.

(4) **Personalization**: To assess the personalization of the generated responses, we first measure the alignment between the persona and responses. First, we apply Consistency Score (C.Score) (Madotto et al. 2019), which leverages a NLI model to predict consistency between response and persona. Additionally, we utilize the BARTScore (Yuan, Neubig, and Liu 2021), which provides a method to calculate the semantic overlaps between texts. Specifically, we compute: (a) Precision (persona $\rightarrow$ response): This measures how closely the generated responses adhere to the input persona, reflecting the degree to which the model captures persona-specific attributes in the response. (b) Recall (response $\rightarrow$ persona): This assesses whether all aspects of the persona are sufficiently covered by the responses, indicating the comprehensiveness of the persona information in the generated text. we report the F1 Score is then calculated

Persona description	Features
I read twenty books a year.	read books, twenty books a year
I'm a stunt double as my second job.	stunt double, second job
I only eat kosher.	only eat kosher
I was raised in a single parent household.	single parent household
Utterance	**Features**
Oh wow. All i've is a dog. That's enough for me.	having a dog
I am good, I just got off work and tired, I have two jobs.	got off work, two jobs, tired
But a good movie is always good.	good movie

Table 4: Examples of feature annotations used for calculating the Feature Coverage Ratio (FCR). All personas and queries in the dialogue have been annotated.

from these Precision and Recall scores to provide a balanced measure of personalization, capturing both the accuracy and completeness of the generated responses in reflecting the specified persona.

Moreover, to better assess whether the generated responses accurately incorporate important features from the persona or the query, we utilize Large Language Models (LLMs)[4] to identify key terms (features) from these persona descriptions or dialogue utterances. We then employ a NLI model to calculate the Feature Coverage Ratio (FCR), ensuring that critical features are effectively represented in the responses. The example of the feature-annotated table mentioned earlier is shown in Table 4.

Main Results

Apart from DialoGPT, we use the context of the most recent 5 turns as dialogue history. After testing, we found that DialoGPT starts to talk nonsense when given more than 2 turns of context. Therefore, for DialoGPT, we only use the most recent 2 turns.

In Table 5, we report experimental results on Text Similarity evaluation. Our method offers better BLEU, ROUGE, and BERTScore compared with baseline methods. Specifically, MUDI's BLEU-1, ROUGE-1, and ROUGE-L scores reach 18.19, 17.10, and 16.13, respectively, outperforming existing methods by 1.64, 3.57, and 3.45. As shown in Table 6, we report the Diversity evaluation. MUDI's USR is 1.0, indicating that it can generate completely unique responses under different queries and personas. Additionally, we achieved the second-highest scores in Ent-1 and Dist-1. Upon examining the outputs from PLATO, we found that they often generate shorter sentences, such as 'me to!!'. Shorter sentences tend to inflate certain metrics, like Dist-1, because they typically involve less repetition of words within a single response, leading to high distinct scores. This suggests that while our method ranks second, it could provide more substantial and contextually rich responses compared to PLATO. Furthermore, our approach achieves the highest scores among all Persona-based dialogue generation methods and significantly surpasses other baselines in Dist-1, outperforming them by 7.95. This performance establishes our method could generate varied and engaging responses.

In addition, Table 7 presents the results of the Coherence evaluation. Compared to other persona-based methods, MUDI has made significant progress in QuantiDCE, DEAM, and BARTScore, particularly in assessing the coherence between the query and response (left-side scores). This indicates that our approach indeed enables the model to generate responses with enhanced local coherence. Furthermore, MUDI also achieves excellent results in global coherence, which evaluates the coherence between the entire dialogue context and the response (right-side scores).

Finally, Table 8 presents the results of evaluating Personalization and Feature Coverage. PAA significantly outperforms other methods in scores for Personalization and FCR_p. Upon further examination, we discovered that this is because PAA frequently generates sentences that are exact restatements of the persona description, often ignoring the relevance to the query. As a result, its high scores in Personalization can be attributed to this tendency. Excluding the special case of PAA, MUDI achieves excellent results in Personalization compared to other methods. Combined with the previously discussed results from the Coherence evaluation (Table 8), this demonstrates that our approach successfully balances discourse relations and persona. It generates responses that effectively consider both aspects simultaneously. Furthermore, our model achieves comparable scores in the Coherence evaluation compared to DialoGPT, which focuses on general dialogue generation.

In evaluating powerful LLMs like GPT-4, we find that they excel at generating lengthy responses and maintaining coherence between questions and dialogue context. Consequently, GPT-4 stands out in Coherence evaluation, scores highly in Ent-1, and performs well in FCR_q. However, this also results in a lower overlap with the ground-truth responses, leading to lower Text Similarity scores. Moreover, the longer sentences generated lead to a lower Dist-1 score. Another observation is that GPT-4 performs poorly in evaluations related to Personalization and FCR_p. This indicates that although existing LLMs are powerful and adept at generating fluent sentences, tasks such as personalized dialogue generation require deeper understanding and inferential capabilities, areas in which they currently fall short.

In summary, compared to existing methods, our approach MUDI not only significantly improves performance in Text Similarity scores but also excels at integrating discourse relations and persona information. This enables us to generate personalized responses that are not only rich in content and diverse but also encompass these aspects. Moreover, in Coherence evaluations, our method achieves scores compa-

[4]https://huggingface.co/meta-llama/Meta-Llama-3-70B-Instruct

Model		Text Similarity				
		BLEU-1 ↑	BLEU-2 ↑	ROUGE-1 ↑	ROUGE-L ↑	BERTScore ↑
Large Language Model (Prompting)						
GPT-4		7.47	2.40	13.52	11.06	84.05
General Dialogue Generation						
DialoGPT		7.34	1.54	9.46	8.41	83.31
PLATO	w/ persona	4.35	1.01	4.88	4.80	82.77
	w/o persona	6.82	1.86	4.99	4.77	81.44
Persona-based Dialogue Generation						
BoB		15.30	5.39	13.21	12.48	83.77
LMEDR		15.47	5.83	13.28	12.26	85.00
PAA		<u>16.55</u>	6.28	13.53	12.68	84.42
MUDI (ours)	$SP_{\tau=0.2}$	15.14	6.43	14.87	13.87	85.07
	$Emb_{\tau=0.2}$	<u>16.55</u>	<u>7.34</u>	**17.10**	**16.13**	<u>85.42</u>
	$SP+Emb_{\tau=0.2}$	**18.19**	**7.77**	<u>16.59</u>	<u>15.46</u>	**85.53**

Table 5: Automatic evaluation results on ConvAI2 dataset over our implemented approach. The best results in each column are in bold, while the second is underlined.

rable to those of state-of-the-art models specialized in open-domain dialogue generation.

Analysis

The Effect of Dialogue Graph Encoder We analyze the effectiveness of the Dialogue Graph Encoder (discussed in Section 3.2.2 fine-tuning phase) under various settings. (1) **Context+Persona (Attention)**: This is our primary method where we utilize an Attention-based Feature Fusion approach to integrate representations from both the Dialogue Graph and the Persona Graph. (2) **Context+Persona (Add)**: We replace the Attention-based Feature Fusion approach with a simple addition of the persona and context representations. (3) **Context**: In this setting, we solely rely on the representation from the Dialogue Graph. (4) **Persona**: Here, we use only the representation from the Persona Graph. (5) **Random**: A random vector replaces the output of the Dialogue Graph Encoder. (6) **None**: The Dialogue Graph Encoder is completely removed from the process, allowing the generator to receive only the context and persona information from the Text Encoder. The comprehensive experimental results can be found in Figure 3. We report the metrics for BLEU-1, ROUGE-1, and Consistency Score (C.Score).

Our experimental results demonstrate that the attention-based feature fusion approach (Context+Persona (Attention)) significantly outperforms other methods in terms of BLEU-1 and ROUGE-1 scores. These findings confirm that effectively integrating contextual and persona information through an attention mechanism enhances the similarity of the generated responses to the ground-truth responses. In contrast, simpler methods such as addition (Context+Persona (Add)) or those relying solely on context or persona information exhibit lower performance. The scores

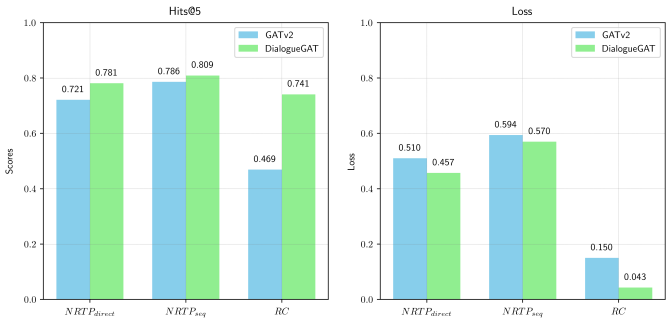

Figure 6: Comparison of GATv2 and DialogueGAT (our proposed) as Dialogue Graph Encoders in Different Tasks. *NRTP* denotes Next Response Type Prediction; *RC* denotes Relations Classification.

drastically decrease when random vectors are employed or when the dialogue graph encoder is omitted entirely, which emphasizes the crucial role of structured and meaningful input in producing coherent responses. The consistency scores further elucidate the models' ability to generate responses that align with persona traits. The superior performance of the attention-based method suggests its effectiveness in maintaining persona consistency within the responses. Notably, employing random vectors results in negative C.Score values, indicating that some generated responses are not only irrelevant but also contradictory to the defined personas.

These outcomes reinforce the importance of utilizing attention-based feature fusion in dialogue systems, especially in tasks that require a nuanced understanding of both

Model		Diversity		
		Ent-1 ↑	Dist-1 ↑	USR ↑
Large Language Model (Prompting)				
GPT-4		9.40	16.09	1.00
General Dialogue Generation				
DialoGPT		**9.05**	36.84	0.99
PLATO	w/ persona	4.67	**58.18**	0.61
	w/o persona	6.73	43.34	0.85
Persona-based Dialogue Generation				
BoB		7.89	41.75	0.99
LMEDR		7.14	43.08	0.94
PAA		6.66	40.27	0.87
MUDI (ours)	$SP_{\tau=0.2}$	8.13	46.76	**1.00**
	$Emb_{\tau=0.2}$	7.65	51.03	**1.00**
	$SP+Emb_{\tau=0.2}$	7.66	47.68	**1.00**

Table 6: Automatic evaluation results for diversity tested on ConvAI2 dataset over our implemented approach. The best results in each column are in bold, while the second is underlined.

context and persona. Additionally, the inferior results associated with random inputs and the complete removal of the dialogue graph encoder highlight potential risks of response incoherence and contradiction when inputs are not integrated thoughtfully. We present examples of generated results for these settings in Table 9.

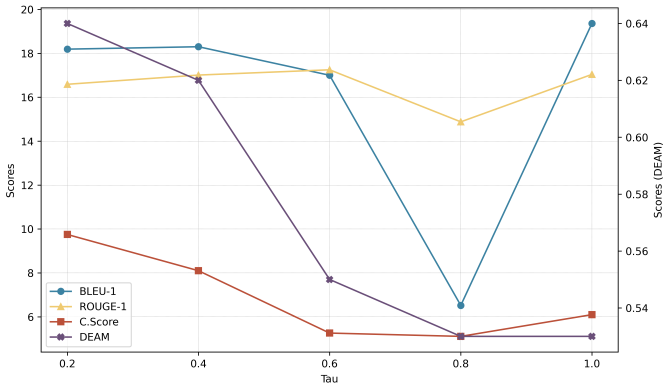

Figure 7: Performance analysis of τ in Dynamic Weighted Aggregation.

The Effect of Tau Values in Dynamic Weighted Aggregation We analyze the effectiveness of τ in Dynamic Weighted Aggregation (discussed in Section 3.4). According to our results, shown in Figure 7, the scores gradually decrease as τ increases, with the most significant drop occurring at 0.8. Additionally, we observed that the scores, particularly for BLEU-1, actually increase when τ reaches 1.0. We hypothesize that this is due to our approach, which involves a residual connection (Eq. **??**) between the outputs of the Text Encoder and the Dialogue Graph Encoder with the results of the dynamic weighting. When τ is set to 1.0, it effectively considers only the original outputs, thus retaining a certain level of generative capability. However, this does not enhance persona consistency as much as when τ is optimally adjusted. This observation confirms that through Dynamic Weighted Aggregation, we can appropriately balance discourse and persona information.

Ablation Study In our approach, to enable the dialogue graph encoder to understand the dialogue structure and adapt to subsequent coherence-related fine-tuning tasks, we designed several pre-training tasks aimed at capturing the dialogue structure. Therefore, we conducted an ablation study on these tasks to examine their impact on generating personalized responses later on. We present the results for three important metrics, coherence, personalization, and diversity, to demonstrate the impact of the pretraining task on the generated outcomes. The results of this ablation study are shown in Table 10, we observed that first, after removing the Shortest Path Prediction task, the scores for local coherence (left side of the QuantiDCE score), persona-consistency (C.Score), and diversity (Dist-1) began to drop sharply. The goal of the Turn Classification task is to help the model capture the local structure. When this task was removed, there was a dramatic decline in scores for local coherence and persona-consistency, proving that this task aids the model in effectively capturing semantic similarities in dialogue data. Furthermore, if the Graph Reconstruction task is removed, there is a continuing downward trend in personalization scores. This confirms that our three self-supervised pretraining tasks are beneficial for the model's understanding of dialogue structure and assist in coherence-related fine-tuning tasks, ultimately impacting the model's ability to balance coherence and persona information when generating personalized responses.

Models	QuantiDCE	C.Score	Dist-1
$MUDI_{SP}$	3.05 / 2.84	11.87	46.76
w/o SPP	2.72 / 2.92	7.05	34.07
w/o TC	2.69 / 2.89	1.18	33.07
w/o GR	2.72 / 2.92	6.49	33.94

Table 10: Ablation study of the Dialogue Graph Encoder for pre-training tasks.

Case Study

Teable 11, 12, 13, and 14 present the personalized responses generated by various methods using the ConvAI2 dataset. The responses generated by the proposed method, **MUDI**, demonstrated greater consistency with their respective personas and showed higher coherence with both the context and the query, appearing more human-like. As illustrated in Table 11, the responses from BoB were incoherent with

Model		Coherence		
		QuantiDCE ↑	DEAM ↑	BARTScore$_{q\to r, c\to r}$ ↓
GOLD		3.19 / 2.83	0.64 / 0.85	5.74 / 5.70
Large Language Model (Prompting)				
GPT-4		3.41 / 2.92	0.87 / 0.96	4.24 / 3.73
General Dialogue Generation				
DialoGPT		**3.23** / 2.79	**0.70** / **0.88**	5.12 / 5.19
PLATO	w/ persona	1.68 / 1.57	0.22 / 0.75	6.42 / 6.63
	w/o persona	1.87 / 1.77	0.25 / 0.74	5.83 / 5.80
Persona-based Dialogue Generation				
BoB		2.99 / 2.76	0.39 / 0.86	5.82 / 5.92
LMEDR		2.89 / 2.90	0.43 / 0.78	5.32 / 4.99
PAA		2.70 / 2.93	0.56 / 0.83	5.17 / 4.62
MUDI (ours)	SP$_{\tau=0.2}$	3.05 / 2.84	0.63 / 0.85	4.67 / **4.61**
	Emb$_{\tau=0.2}$	**3.23** / **2.94**	0.56 / 0.82	4.80 / 4.83
	SP+Emb$_{\tau=0.2}$	3.21 / 2.92	0.64 / 0.86	**4.66** / 4.66

Table 7: Automatic evaluation results for coherence tested on the ConvAI2 dataset. The best results in each column are in bold, while the second-best results are underlined. The score on the left considers only the local coherence between the query and the response, while the score on the right takes into account the global coherence between the entire dialogue and the response.

the query, though they remained consistent with the persona. PAA often overfocused on the persona, leading to repetitive narratives of the persona description. Both LMEDR and our model produced responses that were more coherent, basing them on the user persona, but MUDI was able to generate more detailed replies, such as adding personal aspirations to the basic response about family support.

In Table 12, the responses generated by BoB were irrelevant to the personas and incoherent with the context. PAA exhibited a global incoherence from the dialogue context, failing to maintain a logical flow in the conversation. While LMEDR generated a somewhat relevant response indicating a dream car, it lacked depth and personal context. On the other hand, MUDI produced a more personalized and contextually rich response that not only mentions the dream car but also integrates family dynamics, demonstrating a deeper understanding of the persona's background and the complexities in their personal relationships.

In Table 13, BoB and PAA tended to overlook content from the ongoing dialogue, generating repetitive responses that rendered the entire conversation incoherent. Although LMEDR was able to generate adequate responses, our model excelled by producing more natural responses through the inclusion of responsive questions.

As shown in Table 14, BoB diverged from the persona's preference by mentioning classic rock instead of country music, showing a misalignment with the user's interests. LMEDR and PAA both correctly identified and responded with a generic appreciation for country music, aligning with the persona's interests, yet their responses lacked specific engagement with the user's mention of Taylor Swift or deeper personal nuances. MUDI's response, while initially not favoring Taylor Swift, cleverly circled back to acknowledge her music, demonstrating not only a nuanced understanding of the persona's tastes but also adding an interesting twist to the conversation. This illustrates MUDI's ability to generate engaging and lifelike responses, making the interaction more intriguing for users.

In summary, our model **MUDI** effectively combines dialogue and persona elements to generate appropriate responses. Furthermore, it excels at using responsive questions to enhance the conversation's naturalness.

Model		Personalization		Feature Coverage	
		C.Score ↑	BARTScore$_{p \leftrightarrow r}$ ↓	FCR$_q$ ↑	FCR$_p$ ↑
GOLD		4.10	5.68	7.69	4.52 / 3.60
Large Language Model (Prompting)					
GPT-4		2.86	4.06	6.95	3.60
General Dialogue Generation					
DialoGPT		4.53	5.39	**6.78**	4.45
PLATO	w/ persona	0.56	6.06	1.20	0.92
	w/o persona	0.18	5.83	3.23	0.28
Persona-based Dialogue Generation					
BoB		0.51	5.52	3.68	0.42
LMEDR		7.38	5.27	4.63	6.99
PAA		**15.19**	**4.26**	<u>4.66</u>	**13.84**
MUDI (ours)	SP$_{\tau=0.2}$	<u>11.87</u>	<u>4.53</u>	<u>4.66</u>	6.50
	Emb$_{\tau=0.2}$	9.70	4.75	4.36	<u>7.77</u>
	SP+Emb$_{\tau=0.2}$	9.75	4.76	4.05	5.01

Table 8: Automatic evaluation results for personalization and feature coverage tested on the ConvAI2 dataset. The best results in each column are in bold, while the second-best results are underlined.

Persona	
I have two dogs. I like to work on vintage cars. My favorite music is country. I own two vintage mustangs.	
Context	
Q: Hi! How are you doing tonight? R: I'm doing great. Just relaxing with my two dogs. Q: Great. In my spare time I do volunteer work. R: That's neat. What kind of volunteer work do you do? Q: I work in a homeless shelter in my town. R: Good for you. Do you like vintage cars? I've two older mustangs. Q: Cool. Not really into cars. My day job is wrestling. R: Cars are my thing. Vintage cars. I love working on them. Wrestling? Do you enjoy it? Q: Yes, I love the crowds, getting to know people. R: I didn't think about the crowd aspect of wrestling. I do not like crowds. Q: Understood. When I need to be alone, I work out a few times each week. R: I agree. Working out is a great way to burn off steam. Do you like country music?	
Query	
A little bit. I can get into taylor swift.	
Response	
Context+Persona (Attention)	I do not like her. She's a bit too uptight for my taste. But I like her music.
Context+Persona (Add)	I can get may may and m and may and j on for more of her
Context	What was she like?
Persona	I like country music too.
Random	you The A' A A played The old White lead A Barbara at The The 17…
None	I like country music as well.

Table 9: Case analysis of the effectiveness of Dialogue Graph Encoder.

Persona	
My dream car is a rolls royce ghost. I currently work for a watch dealer. My family does not support my career choices. I often shop for shoes and watches. I like shopping.	
Context	
Q: Good evening. What are you up to? R: Hello, how are you ? I am well. I like to shop for shoes and watches. Q: I don't shop much. Prefer to relax at home with my cow. R: What do you do ? I'm work for a watch dealer.	
Query	
No wonder you like watches ! I'm a commercial electrician.	
Response	
GOLD	wonderful! yeah, however, my family does not support my choice of career.
BoB	I work at a dealership, but I am not very good at it.
LMEDR	My family does not support my career choices.
PAA	I like to shop. I also like to shop.
MUDI	My family does not support my career choices. I want to drive a rolls royce ghost.

Table 11: Case analysis of personalized response generation on the ConvAI2 dataset.

Persona	
My dream car is a rolls royce ghost. I currently work for a watch dealer. My family does not support my career choices. I often shop for shoes and watches. I like shopping.	
Context	
Q: Good evening. What are you up to? R: Hello, how are you ? I am well. I like to shop for shoes and watches. Q: I don't shop much. Prefer to relax at home with my cow. R: What do you do ? I'm work for a watch dealer. Q: No wonder you like watches! I'm a commercial electrician. R: Wonderful! Yeah, however, my family does not support my choice of career.	
Query	
Why not? Family is important.	
Response	
GOLD	I'm not sure why they do not. Do you like cars?
BoB	I am not sure. I am a watch dealer.
LMEDR	I do not know. I do have a dream car though. It is a rolls royce ghost.
PAA	I am a watch dealer. I have a lot of money.
MUDI	yes agree! I want to drive a rolls royce ghost car, but my family does not like that.

Table 12: Case analysis of personalized response generation on the ConvAI2 dataset.

Persona	
I listen to rap music. I produce music for artists. I drive a 2015 Honda Civic. My favorite food is pizza.	
Context	
Q: Hi, how are you? Do you have any brothers or sisters? R: No I don't do you? Q: Yes, I'm 13 and I've an older brother. R: That's nice what kind of music do you like. Q: I do not have much time as I play soccer. You? R: I am a music producer for rap artists	
Query	
Cool I like rap. I hate maths though! Do you have other hobbies?	
Response	
GOLD	Work takes up a lot of my time.
BoB	I like music and I like to listen to music.
LMEDR	I like to drive my honda civic.
PAA	I like rap music, I produce it for artists.
MUDI	I like to go my old Honda Civic car and eat pizza. Do you have any hobbies?

Table 13: Case analysis of personalized response generation on the ConvAI2 dataset.

Persona	
I have two dogs. I like to work on vintage cars. My favorite music is country. I own two vintage mustangs.	
Context	
Q: Hi! How are you doing tonight? R: I'm doing great. Just relaxing with my two dogs. Q: Great. In my spare time I do volunteer work. R: That's neat. What kind of volunteer work do you do? Q: I work in a homeless shelter in my town. R: Good for you. Do you like vintage cars? I've two older mustangs. Q: Cool. Not really into cars. My day job is wrestling. R: Cars are my thing. Vintage cars. I love working on them. Wrestling? Do you enjoy it? Q: Yes, I love the crowds, getting to know people. R: I didn't think about the crowd aspect of wrestling. I do not like crowds. Q: Understood. When I need to be alone, I work out a few times each week. R: I agree. Working out is a great way to burn off steam. Do you like country music?	
Query	
A little bit. I can get into taylor swift.	
Response	
GOLD	lol. She's more pop now. Her old stuff was country. I like classic country.
BoB	I'm not much into music. I like classic rock.
LMEDR	Yes, I love country music.
PAA	I do. I love to listen to country music.
MUDI	I do not like her. She's a bit too uptight for my taste. But I like her music.

Table 14: Case analysis of personalized response generation on the ConvAI2 dataset.