A Disentangled-Attention Based Framework with Persona-Aware Prompt Learning for Dialogue Generation

Pingsheng Liu[1]*, Zhengjie Huang[1]*, Xiechi Zhang[1]*, Linlin Wang[1]*†
Gerard de Melo[2], Xin Lin[1], Liang Pang[3], Liang He[1]

[1] East China Normal University
[2] Hasso Plattner Institute, University of Potsdam
[3] Institute of Computing Technology, CAS
{51205901014, 51215901069, 51255901060}@stu.ecnu.edu.cn
{llwang, xlin, lhe}@cs.ecnu.edu.cn, gdm@demelo.org, pangliang@ict.ac.cn

Abstract

Endowing dialogue agents with personas is the key to delivering more human-like conversations. However, existing persona-grounded dialogue systems still lack informative details of human conversations and tend to reply with inconsistent and generic responses. One of the main underlying causes is that pre-defined persona sentences are generally short and merely superficial descriptions of personal attributes, making appropriate persona selection and understanding non-trivial. Another challenge is that it is crucial to consider the context and the conversation flow to dynamically determine when to invoke different types of persona signals. To address these problems, we propose a disentangled-attention based pre-training architecture, which incorporates persona-aware prompt learning to bridge the connection between the selected persona and response generation. Our model first exploits the conversation flow to select context-relevant personas, and subsequently enriches the superficial persona descriptions with extra personality traits through persona-aware prompting. Finally, the decoder leverages a disentangled-attention mechanism to flexibly control the reliance on personas and dialogue contexts, and incorporates A*-like keyword-based heuristic estimates for controllable generation. Extensive experiments show that our approach can outperform strong baselines and deliver more consistent and engaging responses on the PERSONA-CHAT dataset.

Introduction

The massive proliferation of personal assistants, such as Apple's Siri and Microsoft's Cortana, has led to increased interest in enabling personalized dialogue systems with more engaging and realistic conversations (Mazaré et al. 2018; Gu et al. 2019). In the task of persona-grounded dialogue generation, chatbots are endowed with latent persona variables or predefined personal facts to enable more human-like responses with high personality consistency (Zhang et al. 2018; Song et al. 2021). One notable line of work on incorporating personas draws on PERSONA-CHAT, a benchmark that provides human-annotated persona descriptions (Zhang et al. 2018). An excerpt of personalized dialogue from PERSONA-CHAT is in Fig. 1, where predefined personas of an interlocutor are given by several sentences.

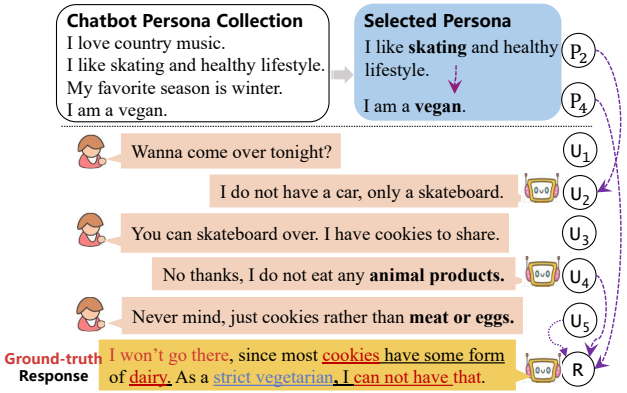

Figure 1: Example of Persona-grounded Dialogue. Different generated words result from multiple information sources, which are marked in red, blue, and underlined, respectively.

However, existing models tend to reply with generic responses that still lack consistency and informative details (Song et al. 2019; Zhao et al. 2020; Majumder et al. 2021; Xu et al. 2020). One of the main underlying causes is inappropriate persona selection, since conventional approaches often neglect the inherent persona transitions that exist in the conversation flow. As illustrated in Fig. 1, a number of complementary cues suggest the relevance of the persona trait P_4 concerning "vegan", which include the most semantically related dialogue contexts (e.g., "do not eat any animal products"), the semantic transition from another persona trait P_2, and the topic flow of the multi-turn conversation (e.g., "come over tonight" → "have cookies to share"). However, existing models merely exploit the semantic relevance between predefined persona candidates and dialogue contexts (Lian et al. 2019; Majumder et al. 2020), and rarely leverage the above conversation flow to track such transitions. Hence, most previous work struggles to select the most context-relevant per-

*These authors contributed equally.
†Corresponding author.

sona details, causing severe persona inconsistency in dialogue generation (Lian et al. 2019).

Second, predefined personas are mostly short and merely superficial descriptions of personal attributes (Majumder et al. 2021; Xu et al. 2020), which brings great difficulties for machines to understand the real personality traits (e.g., "introverted") of interlocutors with artificially generated persona sentences (e.g., "I do not like to talk"). For lack of an in-depth understanding of persona traits, existing models tend to copy the original description from the given persona when generating responses (Kim et al. 2022), making conversations less engaging, as part of the generated response trivially overlaps with the given persona sentence.

Another shortcoming of existing models is that they often leverage semantic ties between personas and dialogue contexts, resulting in an over-reliance on either personas or dialogue contexts during response generation (Jiang et al. 2020a). From a generative perspective, every generated word of the response in Fig. 1 may originate from one of three factors: (1) appropriate personas, (2) semantic relationships in the dialogue history (e.g., "animal products" $\rightarrow$ "meat or eggs"), (3) semantic connections between persona descriptions and specific dialogue contexts (e.g., "vegan" $\rightarrow$ U_4, U_5). However, prior work disregards such differences while generating responses, thereby failing to distinguish the individual effects of a specific persona, context, or persona-context connections towards generation.

To tackle the above issues and simulate the aforementioned generative process, we propose a disentangled-attention based pre-training architecture with persona-aware prompt learning, which first keeps track of the persona flows in multi-turn dialogues to select reasonable context-relevant personas, and subsequently leverages persona-aware prompting to enrich superficial persona descriptions with implicit personality-correlated knowledge extracted from pre-trained language models (Liu et al. 2021; Jiang et al. 2022). When generating responses, it draws on a disentangled-attention based decoder using a decoding manager to dynamically determine the individual effects of three adapters, which further incorporates A*-like keyword-based heuristic estimates to guarantee controllable generation.

To sum up, our key contributions are as follows:

- Our model can exploit conversation flows to select the most context-relevant personas by keeping track of the persona transitions as well as leveraging the response information as guiding signals during training.
- To enrich superficial descriptions of the selected persona, we leverage persona-aware prompt learning to generate instance-dependent prompts and elicit rich personality information from pre-trained language models, tying persona understanding to response generation.
- The disentangled-attention based decoder generates responses conditioned flexibly on the given persona and dialogue contexts, further incorporating A*-like keyword-based heuristic estimates for controllable generation.

Related Work

Persona-grounded dialogue generation has been extensively explored with both retrieval-based and generation-based models (Gu et al. 2019; Bao et al. 2020; Liu et al. 2020; Gu et al. 2021). Early generation-based methods attempt to capture chatbot personalities implicitly with user embeddings (Li et al. 2016b) or incorporate explicit structured profiles to maintain high coherence (Qian et al. 2018). The construction of the PERSONA-CHAT dataset gave rise to the development of sophisticated recurrent neural network based persona-grounded generative models, such as memory-augmented Persona-CVAE (Song et al. 2019) and DeepCopy (Yavuz et al. 2019).

Extending this line of work, large pre-trained language models (PLMs) have gradually become the dominant backbones of persona-grounded dialogue systems due to their powerful abilities (Wolf et al. 2019; Golovanov et al. 2019; Zheng et al. 2020; Song et al. 2021). By leveraging different pre-trained chatbots with sophisticated techniques (Lin et al. 2021), the quality of generated responses has improved substantially. However, these approaches generate responses by conditioning indiscriminately on both dialogue contexts and predefined personas, neglecting the importance of specific contents in the conversation history towards persona incorporation (Jiang et al. 2020a). Considerable attention has also been paid to the selection of appropriate personas, which is essential for personalized dialogue generation. Lian et al. (2019) leverage both prior and posterior distributions over personas to facilitate the selection, while Majumder et al. (2020) model the persona choice with a discrete latent random variable. However, these approaches ignore the inherent persona transitions in the conversation flow, making it hard to attend to the correct persona details.

Another challenging aspect is that the unstructured predefined persona descriptions are mostly short and limited in scope. Although the given personas enable dialogue models to increase the specificity of generated response, they do not provide sufficiently detailed experiences related to an interlocutor, often leaving conversations shallow and dull. Prior work leverages neural topic models to extract plausible topical keywords related to the given persona facts (Xu et al. 2020). Others enhance dialogue models with background stories related to a persona by leveraging fictional narratives from existing story datasets to make the conversation more engaging (Majumder et al. 2021). We instead propose a simple yet effective prompting method to automatically elicit additional knowledge from pre-trained language models to enrich the original persona. Finally, we also compare our model with other lines of work sharing seemingly similar goals. The first is a pre-training based model with an attention-routing mechanism for persona-sparse dialogue generation (Zheng et al. 2020), which, however, fails to distinguish the individual effects of a specific persona, context, or persona-context towards response generation (Zhang et al. 2019). Our model, in contrast, introduces a multi-head disentangled-attention enhanced decoder with persona-aware prompt learning, which dynamically determines the most pertinent factor for every generated word, using three adapters and A*-like keyword-based heuristic estimates. Recent studies consider sequential latent modeling for general knowledge selection (Kim, Ahn, and Kim 2020; Zhan et al. 2021). Unfortunately, sequential latent variables are insuf-

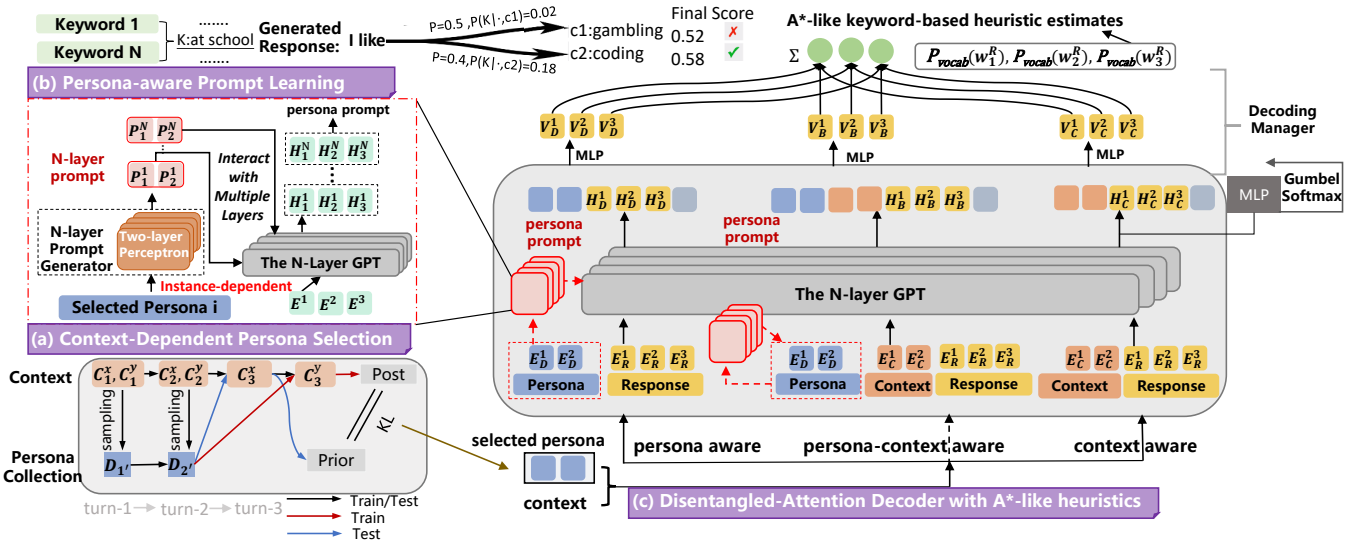

Figure 2: The overall framework of our model.

ficient to model context-relevant personas, since the predefined persona traits in PERSONA-CHAT are given as superficial descriptions. Thus, we leverage both the conversation flow for persona tracking and persona-aware prompting to extract additional persona signals from pre-trained language models, promoting higher quality response generation.

Model

Task Formulation

Suppose User x and Agent y are involved in a t-turn conversation. Our task aims to generate a response C_t^y for y based on all previous dialogue contexts $C = ((C_1^x, C_1^y), (C_2^x, C_2^y), \ldots, (C_t^x))$ and a predefined persona collection $D = \{D_1, D_2, \ldots, D_m\}$ with m sentences.

Model Overview

In this paper, we propose a disentangled-attention based architecture with persona-aware prompt learning for dialogue generation. As illustrated in Fig. 2, our model first conducts a selection of context-relevant personas, followed by persona-aware prompting so as to enrich the superficial description of the selected persona. A custom-designed decoder is finally invoked to generate an engaging response.

Context-Dependent Persona Selection

To select context-dependent personas, our model keeps track of inherent persona transitions in multi-turn dialogues to maintain consistency. We regard persona selection as a sequential decision process that is formulated using a latent variable z with the conditional probability $p(z|C, D)$, where C is the dialogue context and D is the pre-defined persona.

Basic Input Encoding We leverage BERT (Devlin et al. 2019) to encode the t-th turn dialogue contexts C_t^x, C_t^y, and the k-th persona sentence D_k by averaging subword embeddings, which respectively results in $e(C_t^x)$, $e(C_t^y)$, and $e(D_k)$. Each utterance pair $C_t^{xy} = [C_t^x; C_t^y]$ at the t-th dialogue turn is jointly represented by using a GRU (Cho et al. 2014) layer that follows:

$$d_t^{xy} = \text{GRU}_{\text{context}}(d_{t-1}^{xy}, e(C_t^{xy})) \in \mathbb{R}^{768}. \quad (1)$$

Sequential Persona Selection Fig. 2 (a) shows an example of the sequential persona selection process with the maximum number of dialogue turns $t = 3$. For the dialogue history, given the current utterance representation d_u^{xy} at turn u ($u < t$), we sample the corresponding persona sentence $D_{u'}$ from the persona collection $D = \{D_1, ..., D_m\}$ with:

$$u' = \text{Gumbel\_softmax}(W_1 d_u^{xy}, \tau), \quad (2)$$

where $W_1 \in \mathbb{R}^{n \times 768}$ is a trainable parameter, Gumbel-Softmax is the re-parametrization trick for a discrete sampling process (Jang, Gu, and Poole 2017), and τ is the temperature such that u' approaches a one-hot vector as $\tau \to 0$.

At the last turn t, we invoke an attention mechanism to calculate the persona distribution given the dialogue context and previously selected persona. The prior persona distribution $\eta_\alpha(D_t) \in \mathbb{R}^n$ and posterior persona distribution $q_\phi(D_t) \in \mathbb{R}^n$ are modeled as:

$$\eta_\alpha(D_t|C_{\le t}^x, C_{<t}^y, D_{<t}) = \text{softmax}(w_{\text{prior}}\,\mathbf{D}), \quad (3)$$

$$q_\phi(D_t|C_{\le t}^x, C_{\le t}^y, D_{<t}) = \text{softmax}(w_{\text{post}}\,\mathbf{D}), \quad (4)$$

where $\mathbf{D} = [e(D_1), ..., e(D_m)]^T$, and

$$w_{\text{prior}} = W_2([d_{t-1}^{xy}; e(C_t^x); d_{t-1}^D]), \quad (5)$$

$$w_{\text{post}} = W_3([d_t^{xy}; d_{t-1}^D]), \quad (6)$$

$$d_{t-1}^D = \text{GRU}_{\text{persona}}(d_{t-2}^D, e(D_{t-1'})). \quad (7)$$

Here, d_t^D is the hidden state of $\text{GRU}_{\text{persona}}$, $d_0^D \in \mathbb{R}^{768}$, and $W_2 \in \mathbb{R}^{768 \times (768 \times 3)}$, $W_3 \in \mathbb{R}^{768 \times (768 \times 2)}$ are parameters.

Finally, we obtain the persona sentence D_z with the highest probability over the posterior distribution in Eq. (4) in training. During testing, we obtain the persona with the highest probability over the prior distribution in Eq. (3).

Persona-Aware Prompting

Since the predefined personas in PERSONA-CHAT are given as superficial descriptions (i.e., using artificially generated sentences to represent underlying personality traits), it is essential to enrich the selected persona descriptions with additional pertinent personality trait features to infer more informative responses (Majumder et al. 2021). Language models pre-trained on large-scale corpora can serve as potential resources as they possess common-sense knowledge and also exhibit behavior that may match human personality traits (Jiang et al. 2022). Therefore, we elicit additional implicit information (e.g., "*a part of healthy lifestyles*") from pre-trained language models (i.e., GPT) (Petroni et al. 2019; Jiang et al. 2020b), serving as supplementary evidence for pre-defined persona sentences (e.g., "*I am a vegan*").

Specifically, we draw on prompt learning to elicit such information. However, conventional prompt learning merely adopts a single unified prompt, which is not sufficient to achieve this goal. We instead propose persona-aware prompting that can generate instance-dependent prompts to guide the learning, steering the language models towards the current persona. As depicted in Fig. 2 (b), we use a N-layer prompt generator to iteratively generate instance-dependent prompts. Specifically, for each layer, we employ a two-layer perceptron as the basic component. We thereby obtain:

$$O_n = \mathrm{G}(O_{n-1}), \tag{8}$$

where $n \in [1, N]$ represents the n-th layer, and $G(\cdot)$ refers to the prompt generator. We use $O = [O_1, \ldots, O_n, \ldots, O_N]$ to stand for the instance-dependent prompt. Subsequently, we prepend individual prompts to the corresponding GPT layers, aiming to discover personality-correlated traits at various levels from GPT. We will explain this step as follows.

Disentangled-Attention Decoder

Given the selected persona D_z and dialogue context C, we generate the response by using a disentangled-attention decoder. The generated response is expected to comprise various features during decoding for delivering more personalized conversations. To make response generation more flexible, we propose a disentangled-attention based decoder, enabling every word of the response to be decoded based on different information sources. Specifically, by extending the vanilla multi-head attention module of the Transformer, three attention adapters are devised to fuse different input signals. In addition, our decoder further incorporates A*-like keyword-based heuristic estimates for controllable response generation. During training, the inputs of our Transformer-based decoder include the selected persona D_z, the dialogue context C, and the target response R, where C denotes the concatenation of historical dialogue utterances. We use segment indicators to distinguish these tokens, and obtain the corresponding encodings of the persona, context, and target response tokens as E_{D_z}, E_C, and E_R, respectively. E_R^t refers to the embedding of the t-th token in the response, and E_{pre} represents the embeddings of previously decoded tokens in the target response. As illustrated in Fig. 2 (c), our decoder consists of a N-layer GPT. In the following, we will explain the working principles in further detail.

Persona-Aware Adapter With this adapter, only the persona-aware prompt O, the persona D_z and the target response R are concatenated as the input of the Transformer layers. Using E_R^{t-1} as the query, and attending to O_n, E_{D_z}, E_{pre}, the t-th token in R is encoded as:

$$H_{D_z,n}^t = \begin{cases} \mathrm{MultiHead}(E_R^{t-1}, [O_n; E_{D_z}; E_{\mathrm{pre}}], \\ \qquad [O_n; E_{D_z}; E_{\mathrm{pre}}]) & n = 1 \\ \mathrm{MultiHead}(E_R^{t-1}, [O_n; H_{D_z,n-1}^t], \\ \qquad [O_n; H_{D_z,n-1}^t]) & n > 1 \end{cases} \tag{9}$$

Note that $O_n \in \mathbb{R}^{l_O \times d}$, $E_{D_z} \in \mathbb{R}^{l_{D_z} \times d}$, $E_{\mathrm{pre}} \in \mathbb{R}^{l_{\mathrm{pre}} \times d}$, and $[O_n; E_{D_z}; E_{\mathrm{pre}}] \in \mathbb{R}^{(l_O + l_{D_z} + l_{\mathrm{pre}}) \times d}$, where l_O is the length of prompt, l_{D_z} is the length of persona tokens, l_{pre} is the length of previously decoded tokens, and d is the size of hidden state. Eq. (9) ensures that the output $H_{D_z,N}^t$ can incorporate abundant persona information. The final generation probability for the t-th response token is:

$$V_{D_z}^t = P_{\mathrm{vocab}}(w_t^R | D_z, w_{1:t-1}^R) = \mathrm{MLP}_{\theta'}(H_{D_z,N}^t). \tag{10}$$

Context-Aware Adapter This adapter highlights the importance of contextual information, omitting persona features. Without adding persona tokens, the dialogue context C is prepended to the target response R directly as the input. Using E_R^{t-1} as the query, and attending to E_C, E_{pre}, the t-th token in R can be encoded as:

$$H_C^t = \mathrm{MultiHead}(E_R^{t-1}, [E_C; E_{\mathrm{pre}}], [E_C; E_{\mathrm{pre}}]). \tag{11}$$

The generation probability of the t-th response token is:

$$V_C^t = P_{\mathrm{vocab}}(w_t^R | C, w_{1:t-1}^R) = \mathrm{MLP}_{\theta'}(H_C^t). \tag{12}$$

Persona-Context Aware Adapter The persona-aware prompt O, persona D_z, dialogue context C, and target response R are concatenated and fed into the Transformer. Using E_R^{t-1} as the query, and attending to O_n, E_{D_z}, E_C, E_{pre}, the t-th token in R is encoded as:

$$H_{B,n}^t = \begin{cases} \mathrm{MultiHead}(E_R^{t-1}, [O_n; E_{D_z}; E_C; \\ \qquad E_{\mathrm{pre}}], [O_n; E_{D_z}; E_C; E_{\mathrm{pre}}]) & n = 1 \\ \mathrm{MultiHead}(E_R^{t-1}, [O_n; H_{B,n-1}^t], \\ \qquad [O_n; H_{B,n-1}^t]) & n > 1 \end{cases} \tag{13}$$

This enables the output to incorporate the joint information from both the persona and context. Finally, the generation probability of the t-th response token is formulated as:

$$V_B^t = P_{\mathrm{vocab}}(w_t^R | D_z, C, w_{1:t-1}^R) = \mathrm{MLP}_{\theta'}(H_{B,N}^t). \tag{14}$$

To avoid feeding in the gold-standard tokens, we use a masked self-attention mechanism that only attends to known tokens for the current response prediction step.

Decoding Manager We subsequently propose a decoding manager that determines which of the above three adapter components is suitable for predicting each word in the response. The probability to predict word w_t^R is calculated as:

$$\begin{aligned} P_{\mathrm{vocab}}(w_t^R) = [&P_{\mathrm{vocab}}(w_t^R | D_z, w_{1:t-1}^R); \\ &P_{\mathrm{vocab}}(w_t^R | C, w_{1:t-1}^R); \\ &P_{\mathrm{vocab}}(w_t^R | D_z, C, w_{1:t-1}^R)] \cdot \psi_t. \end{aligned} \tag{15}$$

We adopt the Gumbel-Softmax trick (Jang, Gu, and Poole 2017) to handle the discrete and non-differentiable process, which defines ψ_t as:

$$\psi_t = \text{Gumbel\_softmax}(\text{MLP}_{\theta_\psi}(H_C^t), \tau) \in \mathbb{R}^3, \quad (16)$$

where τ is the temperature such that ψ_t approaches a one-hot vector when $\tau \to 0$. In training, we start from a high temperature and reduce it gradually. During testing, we discretize ψ_t as a one-hot vector with the distribution in Eq. (16).

A*-like Heuristics When decoding the next response token in Fig. 2, the key phrase "*at school*" (extracted from dialogue context) provides a crucial cue for choosing the most appropriate candidate "*coding*" instead of "*gambling*". Therefore, we additionally incorporate A*-like heuristic estimates of keywords to further ensure controllable response generation, akin to the A* search algorithm (Lu et al. 2022).

Specifically, we first extract k keywords from both dialogue contexts and the persona sentences, and subsequently incorporate keyword-based heuristics $h(w_t^R)$ to refine the probability distribution as Eq. (15). Finally, the refined probability distribution $P^F_{\text{vocab}}(w_t^R)$ is defined as:

$$P^F_{\text{vocab}}(w_t^R) = P_{\text{vocab}}(w_t^R) + h(w_t^R). \quad (17)$$

For the keyword-based heuristics $h(w_t^R)$, when decoding the next response token w_t^R, we temporarily concatenate the previously generated response (e.g.,"I like") with every candidate word (e.g., "gambling" or "coding") from the vocabulary, and calculate the probability of generating the extracted keywords (e.g., "at school") based on the concatenated sequence. For each candidate word v_i from the vocabulary, we multiply the probability with the corresponding importance coefficient of every keyword, obtaining $h(w_t^R)$ as follows:

$$H_{i,j} = P(K_j \mid D_z, C, R^P, v_i), \quad (18)$$

$$h(w_t^R) = HI. \quad (19)$$

where R^P is the previously generated response, K_j is the j-th keyword extracted from dialogue contexts and persona sentences, $I \in \mathbb{R}^k$ is the importance of each keyword, and $H \in \mathbb{R}^{\text{vocab}\times\text{k}}$ is the probability of using the candidate v_i to generate keyword K_j. As for keyword extraction, we convert multi-turn dialogues to successive context-response pairs, and leverage the DialoGPT annotator (Zhang et al. 2020) to predict the most important keywords from the given contexts. Since the given persona sentences often have semantically irrelevant meanings, we extract persona keywords with YAKE (Campos et al. 2020), an unsupervised method that relies on statistical linguistic features.

Training and Inference

In light of the absence of ground-truth persona labels, the persona selection and response generation are optimized jointly. The training objective is the *variational lower-bound* formulated as follows:

$$\begin{aligned}\mathcal{L}_{\text{ELBO}} = \mathbb{E}_{q_\phi(D_t)}[\log P_\theta(C_t^y | C^x_{\leq t}, C^y_{<t}, D_t)] \\ -\text{KL}(q_\phi(D_t)||\eta_\alpha(D_t)).\end{aligned} \quad (20)$$

Here, $P_\theta(C_t^y|\cdot)$ denotes the decoder network. $\eta_\alpha(D_t)$ and $q_\phi(D_t)$ correspond to Eq. (3) and (4), respectively. We approximate the expectation $\mathbb{E}_{q_\phi(D_t)}$ by drawing one sample D_z from the posterior distribution $q_\phi(D_t|C^x_{\leq t}, C^y_{\leq t}, D_{<t})$ with the Gumbel-Softmax function (Jang, Gu, and Poole 2017). Note that after the persona D_z is sampled, we conduct persona-aware prompting. The generated persona-aware prompt O and selected persona D_z are then fed into the decoder along with the context for training. For inference, the persona sentence with the highest probability is selected from the prior distribution $\eta_\alpha(D_t|C^x_{\leq t}, C^y_{<t}, D_{<t})$. Subsequently, the selected persona D_z is fed into the decoder along with the context for generation. After obtaining the preliminary probability distribution $P_{\text{vocab}}(w_t^R)$, we further use the A*-like algorithm to control response generation by incorporating heuristic estimates based on keywords.

Experiments

Dataset and Baselines

We conduct a series of experiments on the commonly used PERSONA-CHAT dataset[1] and its revised version (Zhang et al. 2018). Each dataset comprises 10,907 dialogues between pairs of interlocutors with pre-defined personas, out of which 968 dialogues are used for validation and 1,000 for testing. Each interlocutor has 4 to 5 persona sentences.

Several representative models are chosen as baselines for experiments, including **LIC** (Golovanov et al. 2019) and **TTransfo** (Wolf et al. 2019), which respectively topped the human and automatic evaluation results in the ConvAI2 competition (Dinan et al. 2020), the attention-routing model **(AR)** (Zheng et al. 2020), and $\boldsymbol{P^2}$ **Bot** (Liu et al. 2020), which obtain the previous state-of-the-art results.

Evaluation Metrics

Automatic Metrics. We adopt several standard metrics to conduct the evaluation. (1) Perplexity (**PPL**) is used to measure how the model fits the test data. (2) **BLEU** (Papineni et al. 2002) is used to evaluate how many n-grams (n=1) in the generated responses overlap with those in the reference responses. (3) **F1** (Dinan et al. 2020) measures the accuracy of the generated responses considering both the precision and recall at the character level. (4) Distinct-2 (**Dist-2**), the ratio of distinct bi-grams, is used to assess the diversity of generated responses (Li et al. 2016a).

Human Metrics. We perform pairwise comparisons between the model-generated responses and the gold responses. Five human annotators are employed to rate 200 test examples per model (in total 2.0k responses, 1.0k per dataset) from three aspects that are critical for practical use: (1) **Fluency** measures whether the response is fluent and human-like. (2) **Coherency** measures whether the response is coherent with the dialogue context. (3) **Consistency** measures whether the response is consistent with the target persona. The rating scale ranges from 0 to 2 (higher scores indicating better results). The inter-annotator agreement for flu-

[1] http://parl.ai/downloads/personachat/personachat.tgz

Models	Original Persona Data							Revised Persona Data						
	Automatic				Human			Automatic				Human		
	PPL	BLEU	F1	Dist-2	Flu.	Coh.	Con.	PPL	BLEU	F1	Dist-2	Flu.	Coh.	Con.
LIC	-	24.06	17.79	0.248	**1.843**	1.804†	1.442†	-	23.94	16.83	0.246	1.813	1.786	1.318†
TTransfo	17.87	25.49	19.09	0.273	1.812	1.756†	1.368†	19.96	25.46	18.15	0.258	1.785†	1.725†	1.247†
AR	15.84	26.28	18.86	0.270	1.768†	1.783†	1.383†	18.72	26.13	18.03	0.254	1.762†	1.692†	1.342†
P^2 Bot	**15.13**	26.90	19.67	0.276	1.804	1.810	1.413†	18.98	26.25	19.14	0.266	1.796	1.779†	1.325†
Ours	18.47	**28.54**	**19.84**	**0.349**	1.831	**1.862**	**1.593**	**18.46**	**29.68**	**19.84**	**0.344**	**1.827**	**1.842**	**1.527**

Table 1: Automatic and Human Evaluation on PERSONA-CHAT test set, where *Flu.*, *Coh.*, *Con.* are Fluency, Coherency and Consistency. We run our model five times and report the average results. † indicates a significant difference with the best result (t-test, p-value$<$0.05).

ency, coherency, and consistency is measured with Cohen's κ (Cohen 1960), obtaining 0.62, 0.71, and 0.73, respectively.

Experimental Results

Table 1 shows the experimental results on automatic metrics and human evaluation. Our approach achieves good performance on the original data, obtaining a relative improvement of 6.1% on BLEU and 0.9% on F1 compared to the previous strongest baseline P^2 Bot. On the revised data, our approach achieves the best performance on all metrics. These results suggest that our model produces higher-quality dialogue responses. In the human evaluation, our approach again achieves strong results along all metrics. Most notably, it substantially outperforms all baselines by a large margin on persona consistency. This demonstrates that our model can generate more persona-consistent responses by making the context correlate with appropriate persona texts. The coherency of generated responses also attains a high level since the disentangled-attention decoder can generate responses flexibly based on specific dialogue contexts and persona traits as well as incorporated keyword constraints.

Quantitative Analysis

Ablation Study We compare the following variants with the full model: (1) **w/o SPS** removes Sequential Persona Selection. (2) **w/o Prompt** drops persona-aware prompting. (3) **w/o DA** omits the Disentangled-Attention decoder. (4) **w/o P (DA)** removes the Persona-aware adapter. (5) **w/o C (DA)** removes the Context-aware adapter. (6) **w/o PC (DA)** omits the Persona-Context aware adapter.

Models	PPL	BLEU	F1
Ours	**18.46**	**29.68**	**19.84**
w/o SPS	18.47	28.20 (↓ 5.0%)	19.74
w/o Prompt	19.08	28.76 (↓ 3.1%)	19.17
w/o DA	18.53	28.45 (↓ 4.2%)	19.66
w/o P (DA)	18.58	27.38 (↓ 7.8%)	19.47
w/o C (DA)	18.60	28.65 (↓ 3.5%)	19.81
w/o PC (DA)	18.59	27.87 (↓ 6.1%)	19.74
w/o A*	18.47	28.93 (↓ 2.5%)	18.89

Table 2: Ablation study on the PERSONA-CHAT revised data, where the relative degradations are reported.

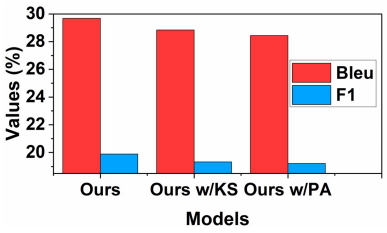

Figure 3: Comparison with two previous single-step persona selection methods on the PERSONA-CHAT revised data.

Impact of Sequential Persona Selection The removal of SPS in Table 2 causes a relative 5.0% performance degradation on BLEU, showing the validity of our persona selection. We further compare our SPS with **KS** (Lian et al. 2019) and **PA** (Majumder et al. 2020), two representative methods that adopt single-step mechanisms for persona selection. Fig. 3 demonstrates that our SPS outperforms the previous single-step method, and the quality of generated responses is greatly improved by tracking conversation flows.

Effect of Persona-Aware Prompting The removal of prompting causes a relative 3.4% performance drop in terms of F1 in Table 2, suggesting the validity of our prompt learning. We further conduct a comparison with Prefix-Tuning (Li and Liang 2021). As observed in Fig. 4 (left), the persona-aware prompting approach outperforms Prefix-Tuning with a unified prompt for response generation, showing the superiority of instance-dependent prompts generated by our model. We report the impact of using prompts with different sizes in Fig. 4 (right). To further investigate whether our

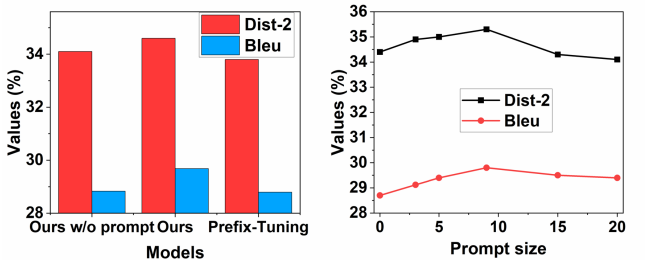

Figure 4: Quantitative study of our persona-aware prompting on the PERSONA-CHAT revised data.

Persona	I like stuffed animals.	I love eating raw fish with rice.
Context	I love dogs! Not a big fan of cats though.	I eat granola daily younger folks can not stand it.
Ours w/o prompt:	I had a dog in high school.	I really enjoy eating raw fish.
Ours:	They are **adorable**, and I have a huge **german shepard** and two tigers.	I eat the **products of nature**, I am a **health nut**.
Reference:	I love dogs too, do you happen to own any pets?	That is what is good for the body and spirit.

Table 3: Comparison of responses with and without persona-aware prompting.

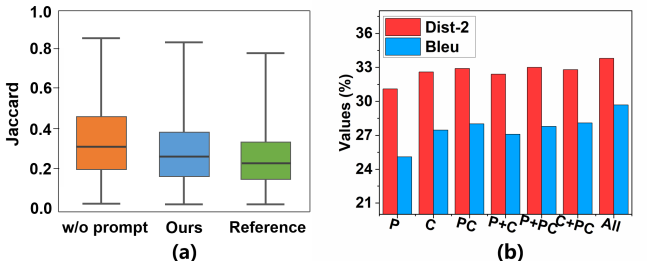

Figure 5: (a) Similarity of persona-response alignments on the PERSONA-CHAT for different methods. (b) Results of different combinations in our decoder on the PERSONA-CHAT revised data. Here, P, C, and PC indicate Persona, Context, and Persona-Context aware adapter, respectively.

prompt learning can extract personality-correlated knowledge from pre-trained language models (i.e., GPT), we randomly pick two examples for qualitative analysis in Table 3. In the leftmost example, our model leverages persona-aware prompting to discover insightful details ("*adorable*" and "*german shepard*") to deepen the understanding of a persona phrase ("*stuffed animals*"), delivering a more engaging and informative response, whereas the generated response of the non-prompting architecture trivially overlaps with the original persona description. We plot the linguistic similarity between persona-response pair alignments with Jaccard similarity scores (Park et al. 2019) in Fig. 5 (a). Hence, our persona-aware prompting enhancement is beneficial for high-quality response generation, as a lower similarity score between a alignment pairs indicates less linguistic bias.

Influence of Three Adapters in Decoder Finally, we further investigate the influence of the three adapters. When removing the Disentangled-Attention decoder (DA), we observe more obvious performance changes in Table 2, causing a relative decrease of 4.2% on BLEU. As for P(DA), C(DA), and PC(DA), removing any of them entails a drop in performance, suggesting the effectiveness of all three components. We observe that P(DA)>PC(DA)>C(DA) in terms of importance. In Fig. 5 (b), we conduct a further study on different combinations of three adapters, which shows that the combination of all three adapters yields the best results.

For a more intuitive comparison, we show randomly sampled examples in Table 4. For instance, we observe that the persona-context adapter generates the integrated semantic content *"I have a lot of friends"* by considering the persona *"helping others"* and context *"alone"* jointly. Additionally, Table 5 shows that our disentangled attention mechanism can generate reasonable responses according to the specific context by evading distracting features and reducing the over-reliance on persona modeling in previous work.

Persona	I enjoy helping people.
Context	You play checkers? I do. I like to be alone.
AR: **P^2 Bot:**	Yes, I like it. It's good. I often play checkers in my spare time.
Ours:	Yes, I play checkers on my computer. I prefer helping others, **I have a lot of friends.**
Reference:	Yes, it's interesting. I often help people, which makes me happy.

Table 4: Case study on our Disentangled Attention decoder.

Persona	(×) I have two dogs and one cat. (×) My favorite color is blue.
Context	I am major meat eater to build muscles. I'm a bodybuilder.
Reference:	Cool. I think scotch goes good with meat.
AR: **P^2 Bot:**	I like meat ok. I have two dogs. I like steak as well.
Ours:	**Oh, I like meat but not as much as vegetables.**
P+PC:	Cool. What is your favorite color?
PC:	That's cool. Do you have any dogs?

Table 5: Reducing the over-reliance on persona sentences.

Conclusion

In this paper, we propose an end-to-end persona-grounded dialogue generation framework. Specifically, we first conduct a primary selection of context-relevant personas, followed by the novel persona-aware prompting to enrich the superficial description of the selected persona. Finally, a disentangled-attention decoder is devised for response generation, which dynamically decodes each token by extracting features from different sources and incorporates A*-like heuristic estimates of keywords. Experimental results on the large-scale standard benchmark dataset PERSONA-CHAT confirm the effectiveness of our approach, suggesting that our method can deliver more consistent and informative responses in persona-grounded conversations.

Acknowledgments

This work was supported by the National Innovation 2030 Major S&T Project of China (No. 2020AAA0104200 & 2020AAA0104205), National Natural Science Foundation of China (No. 62006077 & 62276248), and Shanghai Sailing Program (No. 20YF1411800).

References

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

Campos, R.; Mangaravite, V.; Pasquali, A.; Jorge, A.; Nunes, C.; and Jatowt, A. 2020. YAKE! Keyword extraction from single documents using multiple local features. *Information Sciences*, 509: 257–289.

Cho, K.; van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; and Bengio, Y. 2014. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1724–1734.

Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1): 37–46.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 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)*, 4171–4186.

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.

Golovanov, S.; Kurbanov, R.; Nikolenko, S.; Truskovskyi, K.; Tselousov, A.; and Wolf, T. 2019. Large-scale transfer learning for natural language generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 6053–6058.

Gu, J.-C.; Ling, Z.-H.; Zhu, X.; and Liu, Q. 2019. Dually interactive matching network for personalized response selection in retrieval-based chatbots. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 1845–1854.

Gu, J.-C.; Liu, H.; Ling, Z.-H.; Liu, Q.; Chen, Z.; and Zhu, X. 2021. *Partner Matters! An Empirical Study on Fusing Personas for Personalized Response Selection in Retrieval-Based Chatbots*, 565–574.

Jang, E.; Gu, S.; and Poole, B. 2017. Categorical Reparameterization with Gumbel-Softmax. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.

Jiang, B.; Zhou, W.; Yang, J.; Yang, C.; Wang, S.; and Pang, L. 2020a. PEDNet: A persona enhanced dual alternating learning network for conversational response generation. In *Proceedings of the 28th International Conference on Computational Linguistics*, 4089–4099.

Jiang, G.; Xu, M.; Zhu, S.-C.; Han, W.; Zhang, C.; and Zhu, Y. 2022. MPI: Evaluating and Inducing Personality in Pre-trained Language Models. *arXiv preprint arXiv:2206.07550*.

Jiang, Z.; Xu, F. F.; Araki, J.; and Neubig, G. 2020b. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8: 423–438.

Kim, B.; Ahn, J.; and Kim, G. 2020. Sequential Latent Knowledge Selection for Knowledge-Grounded Dialogue. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*.

Kim, M.; Kwak, B.-w.; Kim, Y.; Lee, H.-i.; Hwang, S.-w.; and Yeo, J. 2022. Dual Task Framework for Improving Persona-grounded Dialogue Dataset. *Proceedings of the AAAI Conference on Artificial Intelligence*.

Li, J.; Galley, M.; Brockett, C.; Gao, D., Jianfeng; and Bill. 2016a. A diversity-promoting objective function for neural conversation models. In *Proceedings of NAACL-HLT 2016*, 110–119.

Li, J.; Galley, M.; Brockett, C.; Spithourakis, G. P.; Gao, J.; and Dolan, B. 2016b. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, 994–1003.

Li, X. L.; and Liang, P. 2021. Prefix-Tuning: Optimizing Continuous Prompts for Generation. In *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)*, 4582–4597.

Lian, R.; Xie, M.; Wang, F.; Peng, J.; and Wu, H. 2019. Learning to Select Knowledge for Response Generation in Dialog Systems. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, 5081–5087.

Lin, Z.; Madotto, A.; Bang, Y.; and Fung, P. 2021. The Adapter-Bot: All-In-One Controllable Conversational Model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 16081–16083.

Liu, P.; Yuan, W.; Fu, J.; Jiang, Z.; Hayashi, H.; and Neubig, G. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*.

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 *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 1417–1427.

Lu, X.; Welleck, S.; West, P.; Jiang, L.; Kasai, J.; Khashabi, D.; Bras, R. L.; Qin, L.; Yu, Y.; Zellers, R.; et al. 2022. Neurologic a* esque decoding: Constrained text generation with lookahead heuristics. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for*

Computational Linguistics: Human Language Technologies, 780–799.

Majumder, B. P.; Berg-Kirkpatrick, T.; McAuley, J.; and Jhamtani, H. 2021. Unsupervised Enrichment of Persona-grounded Dialog with Background Stories. In *Proceedings of the ACL-IJCNLP (Volume 2: Short Papers)*, 585–592.

Majumder, B. P.; Jhamtani, H.; Berg-Kirkpatrick, T.; and McAuley, J. 2020. Like Hiking? You Probably Enjoy Nature: Persona-grounded Dialog with Commonsense Expansions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 9194–9206.

Mazaré, P.-E.; Humeau, S.; Raison, M.; and Bordes, A. 2018. Training millions of personalized dialogue agents. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2775–2779.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL'02, 311–318.

Park, S.; Hwang, S.-w.; Chen, F.; Choo, J.; Ha, J.-W.; Kim, S.; and Yim, J. 2019. Paraphrase Diversification Using Counterfactual Debiasing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01): 6883–6891.

Petroni, F.; Rocktäschel, T.; Riedel, S.; Lewis, P.; Bakhtin, A.; Wu, Y.; and Miller, A. 2019. Language Models as Knowledge Bases? In *Proceedings of the 2019 Conference on EMNLP-IJCNLP*, 2463–2473.

Qian, Q.; Huang, M.; Zhao, H.; Xu, J.; and Zhu, X. 2018. Assigning Personality/Profile to a Chatting Machine for Coherent Conversation Generation. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, 4279–4285.

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 *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, 167–177.

Song, H.; Zhang, W.-N.; Cui, Y.; Wang, D.; and Liu, T. 2019. Exploiting persona information for diverse generation of conversational responses. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, 5190–5196.

Wolf, T.; Sanh, V.; Chaumond, J.; and Delangue, C. 2019. Transfertransfo: A transfer learning approach for neural network based conversational agents. In *Proceedings of the 32nd Conference on Neural Information Processing Systems CAI Workshop*.

Xu, M.; Li, P.; Yang, H.; Ren, P.; Ren, Z.; Chen, Z.; and Ma, J. 2020. A Neural Topical Expansion Framework for Unstructured Persona-Oriented Dialogue Generation. In *ECAI 2020*, 2244–2251.

Yavuz, S.; Rastogi, A.; Chao, G.; and Hakkani-Tür, D. 2019. DeepCopy: Grounded Response Generation with Hierarchical Pointer Networks. In Nakamura, S.; Gasic, M.; Zuckerman, I.; Skantze, G.; Nakano, M.; Papangelis, A.; Ultes, S.; and Yoshino, K., eds., *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue, SIGdial 2019, Stockholm, Sweden, September 11-13, 2019*, 122–132.

Zhan, H.; Zhang, H.; Chen, H.; Ding, Z.; Bao, Y.; and Lan, Y. 2021. Augmenting knowledge-grounded conversations with sequential knowledge transition. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 5621–5630.

Zhang, S.; Dinan, E.; Urbanek, J.; Szlam, A.; Kiela, D.; and Weston, J. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 2204–2213.

Zhang, Y.; Sun, S.; Galley, M.; Chen, Y.-C.; Brockett, C.; Gao, X.; Gao, J.; Liu, J.; and Dolan, B. 2019. Dialogpt: Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 270–278.

Zhang, Y.; Sun, S.; Galley, M.; Chen, Y.-C.; Brockett, C.; Gao, X.; Gao, J.; Liu, J.; and Dolan, W. B. 2020. DIALOGPT: Large-Scale Generative Pre-training for Conversational Response Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 270–278.

Zhao, X.; Wu, W.; Xu, C.; Tao, C.; Zhao, D.; and Yan, R. 2020. Knowledge-grounded dialogue generation with pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 3377–3390.

Zheng, Y.; Zhang, R.; Huang, M.; and Mao, X. 2020. A pre-training based personalized dialogue generation model with persona-sparse data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 9693–9700.