Dual Diffusion Learning for Knowledge-Grounded Dialogue Generation

Anonymous ACL submission

Abstract

Knowledge-grounded dialogue generation plays a crucial role in the intelligent conversational agents. However, previous work suffers from inadequate control information in both knowledge selection and dialogue generation. Firstly, priori-based knowledge selection lacks a posteriori distribution, while posterior-based methods suffer from biases at the inference and training stages. Secondly, the conventional autoregressive generation lacks precise control over the injection of knowledge, leading to unintended shifts in focus of response. To address these limitations, we propose a Controllable Dual Diffusion Learning model, which serves as an enhanced framework for knowledge-grounded dialogue generation through the controllable modules. Our approach formulates response generation and knowledge generation as dual tasks to fully leverage prior and posterior knowledge, and to avoid training and inference biases. We optimize knowledge selection by employing knowledge labels generated by the dual module and iteratively update the generated dialogue with global-related knowledge information. Experimental results on two public datasets demonstrate that our approach achieves significant improvements in both automatic and manual evaluations[1].

1 Introduction

Open-domain dialogue systems have garnered significant attention in the literature, driven by advances in deep neural networks (Serban et al., 2015; Shum et al., 2018; Freitas et al., 2020). These systems are capable of generating fluent and grammatically correct responses based on dialogue history. However, despite their prowess, they still lag behind human-to-human dialogue in various aspects. In recent years, the researchers have witnessed a surge in interest surrounding knowledge-grounded

[1] Our code is available at https://released-when-published

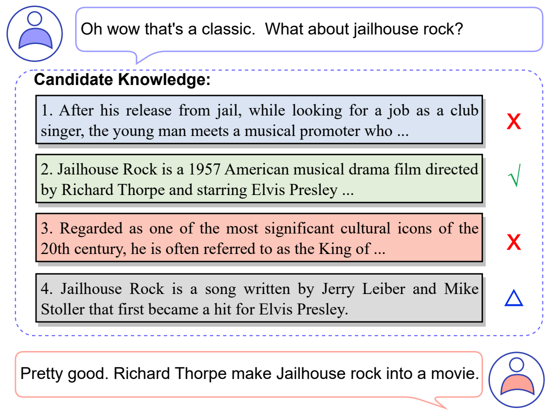

Figure 1: Knowledge selection depends on both the dialogue history and the following response. × marks those contextually irrelevant candidates. △ marks the knowledge which is similar to context in semantic but is undesired. ✓ denotes the selected knowledge when considering both dialogue history and potential responses. It perfectly meets the knowledge need of this dialogue.

dialogue generation (KGDG) (Dinan et al., 2018; Zhou et al., 2018b; Zhao et al., 2020; Yang et al., 2022; Zhan et al., 2021; Xu et al., 2022). These works aim to bridge the gap by enhancing the understanding and application of pertinent knowledge when generating responses, especially in conversations involving specific topics. Furthermore, the KGDG is currently effective in mitigating the hallucination problem under a large language model with a new paradigm (Zhang et al., 2023).

The KGDG is designed to integrate external knowledge into the dialogue generation process, which involves two key modules: Knowledge Selection and Response Generation. However, two unique characteristics of KGDG pose challenges to the existing works: **(I)** Knowledge selection can be influenced by both the dialogue history and the future response. Relying only on the context can easily lead to arbitrary knowledge due to the ambiguous reference and open topics, so combining the historical context (priori) and the future response (posteriori) can greatly improve the selection accuracy (Kim et al., 2019). As shown in Figure 1, if the model relies only on context selection, it may face a topic dilemma when choosing between knowledge (2) and (4). However, when both context and response are considered, knowledge (2) emerges as the correct choice. This important feature is largely ignored, which are based solely on prior information (i.e., dialogue history) and thus struggle to select the correct knowledge due to the lack of necessary posterior information (i.e., future response). For existing works (Kim et al., 2020a; Zheng et al., 2020; Chen et al., 2020) basically use a posterior distribution to approximate a prior distribution. These methods inevitably suffer from the fact that they cannot fully exploit the global information based on a posterior selection for training, but only have a prior selection for inference. During the inference, the responses are not available, which introduces a bias into the training. **(II)** KGDG needs to precisely inject knowledge into responses (Lian et al., 2019a). Existing research directly combines knowledge and context representations, which ignores the granularity of knowledge and leads to limited control over the use of desired knowledge (Meng et al., 2020). Furthermore, to the best of our knowledge, existing KGDG systems are implemented with autoregressive models that generate tokens sequentially. Even a single token that deviates from the desired knowledge can cause error accumulation, and these unintentional shifts cannot be revised, resulting in even incorrect knowledge injection. A high quality dialogue model should allow for flexible knowledge injection, from coarse-grained to fine-grained, through content planning for controllable generation.

To alleviate the above challenges, we propose a Controllable Dual Diffusion Learning (ConDDL) model. Specifically, our approach aims to mitigate the lack of control information in knowledge selection and dialogue generation. To address challenge (I), we take advantage of dual learning (He et al., 2016) from context and response for knowledge selection by using response generation and knowledge generation as dual tasks. In response generation, we utilize the context and selected knowledge to generate a response, while in knowledge generation, we generate globally relevant knowledge based on the context and response, which in turn updates the knowledge selection process and mitigates bias during training and inference. When ConDDL performs two tasks in a dual cycle, the model would learn to select knowledge that related to both history and response. For challenge (II), we replace the commonly used autoregressive model with the diffusion model, which performs generation in a first-plan-then-refine fashion, as verified in many works on image and text generation (Rombach et al., 2021; Feng et al., 2022). We argue that such a coarse-to-fine generation paradigm may reflect human cognitive behaviour, where people tend to continuously refine their statements from knowledge-based sketches, thereby improving precise knowledge incorporation. Through the response diffusion decoder and the knowledge diffusion decoder, we alleviate the precise knowledge injection challenge. The response diffusion decoder incorporates the selected knowledge into the dialogue generation process, resulting in more coherent and contextually appropriate responses. Simultaneously, the knowledge diffusion decoder predicts globally relevant knowledge based on the context and response, dynamically updating the previous knowledge selection. This iterative process ensures continuous refinement of the generated response with knowledge. In addition, we optimise the selection module using contrastive learning, enabling ConDDL to focus more accurately on specific knowledge while modeling the global content.

We conduct extensive experiments on two benchmark datasets to verify the effectiveness. Both automatic and manual evaluations demonstrate that our method significantly outperforms baselines. Our proposed model exhibits superior flexibility in knowledge selection, resulting in more accurate and informative responses. In summary, our contributions can be summarized as follows:

- We propose the Controllable Dual Diffusion Learning framework, which takes into account potential responses, thereby improving the selection of globally relevant knowledge.
- We take the advantage of diffusion models through iterative refinement manner to enhance the desired knowledge injection.
- Experiments conducted on two benchmark datasets show that our proposed method outperforms all baselines with limited training data.

2 Related Work

Knowledge-Grounded Dialogue Generation. Researchers have made significant progress in incorporating external knowledge sources to improve dialogue quality. Various knowledge sources have been utilised in this area of research, including knowledge graphs (Zhou et al., 2018a; Wu et al., 2019), document information (Dinan et al., 2018; Zhou et al., 2018b), and visual background (Das et al., 2016). These external knowledge sources play a crucial role in enhancing the dialogue generation process. Typically, knowledge-grounded dialogue generation tasks involve three main components: knowledge retrieval, knowledge selection, and dialogue generation based on knowledge (Kim et al., 2020a; Zheng et al., 2021). The first two tasks aim at ranking and selecting relevant knowledge based on context while avoiding noisy interferences. The last task focuses on improving the integration of knowledge during dialogue generation. Zhou et al. (2021) generates implicit knowledge sentences for further response generation. Liu et al. (2022) uses prompts for knowledge based on a large pre-trained language model. In this work, we focus on optimising the use of external knowledge in dialogue generation. We adopt a controlled diffusion strategy to gradually incorporate knowledge updates into the dialogue generation process.

Dual Learning. Dual learning is initially proposed in the context of neural machine translation (He et al., 2016) and has proven to be effective in various tasks, including neural machine translation and stylized dialogue generation (Li et al., 2021). Dual learning involves two models, namely the forward and backward models, which interact with each other and receive immediate rewards. (He et al., 2016) use one agent to represent the model for the primal task and the other agent to represent the model for the dual task, then ask them to teach each other through a reinforcement learning process. (Li et al., 2021) employ dual learning to work on a three-domain text related problem, then the contents of non-conversational text can be effectively utilized to enrich the dialogue generation. In this work, we apply the dual learning module to model the interdependence between external knowledge and the global conversational information from context and response.

Diffusion Learning. Diffusion learning has garnered significant attention in the field of machine learning and computer vision due to its effectiveness in modeling complex data distributions. Several noteworthy works have explored various aspects of diffusion models. Denoising diffusion probabilistic models (DDPM) have demonstrated promising capabilities in text-to-image generation using diffusion learning (Ho et al., 2020), opening up new possibilities for text generation. Li et al. (2022) propose the Diffusion-LM, which adopts the plug-and-play approaches to compose fine-grained constraints on the generated sentences, but it fails to condition on the whole source sentence in sequence-to-sequence tasks. Therefore, Gong et al. (2022) explore a diffusion-based approach for sequence-to-sequence tasks, showing strong potential for achieving a better trade-off between generation quality and diversity. Its gradual noise reduction characteristics are very consistent with human knowledge-based reply behavior. These studies have collectively propelled diffusion learning to the forefront of machine learning research, offering promising avenues for future developments in generative modeling and data analysis.

3 Methods

3.1 Preliminary

We introduce the structure of Controllable Dual Diffusion Learning (ConDDL) model, as shown in Figure 2, consisting of two main components: the Diffusion Learning module and the Dual Learning module. Suppose that we have a dialogue dataset $\mathcal{D} = \{(U_i, R_i, \mathbf{K}_i)\}_{i=1}^{n_D}$, where $U_i = (u_{i,1}, ..., u_{i,n_i})$ is the dialogue context and $u_{i,j}$ denotes the j-th utterance. R_i is the response regarding to U_i with the golden knowledge $K_i \in \mathbf{K}_i$, which consists of a set of candidate knowledge pieces (e.g., sentences from Wikipedia). Our goal is to train a response generation model f_{FW} to generate a knowledgeable response $R'_i = f_{FW}(U_i, \hat{K}_i)$ utilizing an input U_i and the selected knowledge $\hat{K}_i$ from $\mathbf{K}_i$, meanwhile addressing above mentioned two challenges.

3.2 Diffusion Learning Module

To control the injection of precise knowledge into conversation responses, we extend the diffusion learning module through sequence-to-sequence diffusion model (Yuan et al., 2022). The major benefit of this generation paradigm is that its generation involves denosining noise iteratively and this inherently involves the content planing, which promotes the precise knowledge incorporation. Training a diffusion model consists of a forward process and a reverse process. The forward process gradually adds quantitative noise to the original data z_0 towards data-irrelevant noise z_T in T time steps. By contrast, in the reverse process, the model learns

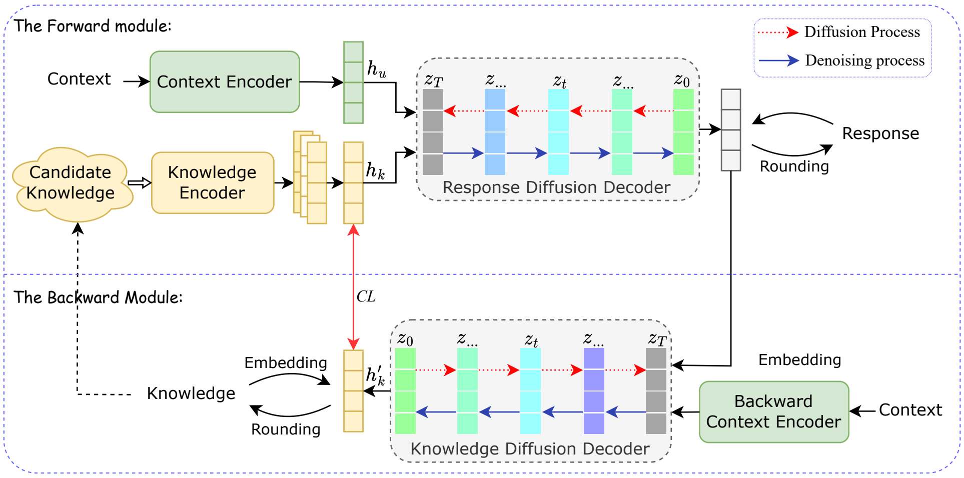

Figure 2: Overview of Controllable Dual Diffusion Learning Model. The method includes response generation (Forward) and knowledge generation (Backward) models, which is based on sequence-to-sequence diffusion module.

to conditionally denoise a corrupted data towards desired content by T steps.

Forward process. During the forward process, we employ an embedding function g_ϕ (Li et al., 2022) to map the discrete word tokens to continuous word embeddings. We define z_0 as a sequence of token representations corresponding to the response R, parameterized by Markov variants $q_\phi(z_0|R) = \mathcal{N}(z_0; g_\phi(R), \beta_0 I)$. We add Gaussian noise to the initial distribution sample z_0 step by step as follows:

$$q(z_t|z_{t-1}) = \mathcal{N}(z_t; \sqrt{1-\beta_t} z_{t-1}, \beta_t I), \tag{1}$$

where $t \in [1, 2, ..., T]$ and $\beta_t \in (0, 1)$ is a pre-defined noise schedule that controls the noise scale added in each step.

Reverse process. The reverse process aims to gradually reconstruct the original data z_0 from the noised data z_T obtained in the forward process through a learned denoising distribution p_θ. In our response diffusion decoder, we use the context and the selected knowledge to gradually inject hidden states to generate the response. For each time step t, the denoising distribution is conditioned on the input context U and knowledge K, denoted as $p_\theta(z_{t-1}|z_t, U, K)$. The distribution for the variants at each time step is defined as follows:

$$p_\theta(z_{t-1}|z_t, UK) = \mathcal{N}(z_{t-1}; \mu_\theta(z_t, UKt), \sigma_\theta(Z_t, UKt)), \tag{2}$$

where $\mu_\theta(\cdot)$ and $\sigma_\theta(\cdot)$ are predicted mean and standard deviation, respectively. The context-aware and knowledge-aware representations are gradually injected into the reverse process as conditions to generate desired responses. The context representation h_u is fed into the response diffusion decoder $f_\varepsilon(\cdot)$ through an attention layer to achieve cross-modal interaction (Nichol et al., 2021):

$$\begin{aligned} Q &= f_\varepsilon(z_t) W_Q^{(i)}, \\ K &= [f_\varepsilon(z_t) W_{K_z}^{(i)}; h_u W_{K_H}^{(i)}], \\ V &= [f_\varepsilon(z_t) W_{V_z}^{(i)}; h_u W_{V_H}^{(i)}], \end{aligned} \tag{3}$$

where i is the index for Transformer layers, $[;]$ denotes the concatenation operator, $W_Q^{(i)}$, $W_{K_z}^{(i)}$, $W_{V_z}^{(i)} \in \mathbb{R}^{d*d}$ and $W_{K_H}^{(i)}$, $W_{V_H}^{(i)}$ are learnable projection layers. Similarly, the knowledge representation h_k is incorporated into the reverse process. When $t = 0$, we use the rounding function $p_\phi(R|z_0')$ to convert the generated z_0' into the embedding space for decoding the response. We optimize the parameters of denoising decoder by minimizing the variational bound of the data log-likelihood (Yuan et al., 2022):

$$\mathcal{L}_{VB} = \mathbb{E}_{q_\phi}[\log \frac{q(z_T|z_0)}{p(z_T)} + \sum_{t=2}^{T} \log \frac{q(z_{t-1}|z_0, z_t)}{p_\theta(z_{t-1}|z_t, U, K)} - \log p_\theta(z_0|z_1, U, K) + \log q_\phi(z_0|R) - \log p_\phi(R|z_0)]. \tag{4}$$

The proposed ConDDL model includes both response diffusion and knowledge diffusion, which share similar computation processes. The control conditions for knowledge diffusion are the response and the context. The diffusion modules improve performance by effectively handling different denoising stages using denoising networks.

3.3 Dual Learning Module

For addressing the challenge of knowledge selection, we consider response generation and knowledge generation as dual tasks from the dual learning viewpoint to fully leverage external knowledge and use global information to enhance knowledge selection and alignment. In the primal task (response generation), the context encoder and the knowledge encoder provide context representations $h_u \in \mathbb{R}^{ni \times d}$ and knowledge representations $\{h_{k_1}, ...h_{k_l}\}$, respectively. The selected knowledge h_k guides the response generation of the response diffusion decoder based on h_u. In the dual task (knowledge generation), the knowledge generation component generates globally relevant knowledge to enhance the selection of the golden knowledge. In our framework, both response diffusion and knowledge diffusion share similar computation processes as mentioned above. Specifically, the response generation model is defined as the forward model f_{FW}. Given the context U and the selected knowledge K, the forward model generates the response $R' = f_{FW}(U, K)$. The knowledge generation model serves as the backward model f_{BW}. It aims to generate the golden knowledge K based on the response R and the original context U, producing $K' = f_{BW}(U, R)$. The forward model is trained by $\mathcal{L}_{VB}$ and maximizing the log-likelihood with a cross-entropy loss between the hypothetical response R' and the golden response R:

$$\mathcal{L}_{\text{FW}} = -\sum_{i=1}^{T_r} \log P(R_i | R'_{<i}, U, K), \tag{5}$$

Similarly, the loss $\mathcal{L}_{\text{BW}}$ of backward model is defined as follows:

$$\mathcal{L}_{\text{BW}} = -\sum_{i=1}^{T_k} \log P(K_i | K'_{<i}, R, U), \tag{6}$$

where T_r, T_k is the length of R and K, respectively. Theoretically, without additional constraints, parameter optimization problems can arise due to gradient chain breaks, resulting in poor alignment of the knowledge vector representation. To improve alignment, we introduce contrastive learning. The knowledge representation h_k is obtained by encoding the selected knowledge using the knowledge encoder. Similarly, the backward model generates the representation h'_k through the knowledge diffusion decoder. Intuitively, these two hidden representations should indicate the same input data in the knowledge domain, making them positive instances of each other (where h_k acts as the anchor), but dissimilar to all other instances in a training batch. Formally, the contrastive loss with a mini-batch of N pairs is defined as follows:

$$\mathcal{L}_{\text{CL}} = -\log \frac{\exp(\text{sim}({h_k}^{(i)}, {h_k}'^{(i)})/\tau)}{\sum_{j=1}^{N} \exp(\text{sim}({h_k}^{(i)}, {h_k}'^{(j)})/\tau)}, \tag{7}$$

where $\text{sim}(\cdot)$ and τ are the cosine similarity function and temperature parameter, respectively. By optimizing the forward and backward models with the cross-entropy losses and incorporating the contrastive loss for knowledge alignment, the dual learning module helps to improve the quality and alignment of the generated responses and knowledge representations.

3.4 Iterative Jointly Generation Strategy

In order to enhance the performance of model by effectively utilizing global information from the context and responses, we propose an iterative jointly generation strategy to improve knowledge selection. During each iteration of this strategy, we begin by feeding a sampled example from the training set to the ConDDL model. Initially, the knowledge is obtained through similarity retrieval. Subsequently, the backward model predicts the knowledge, and this prediction is updated to improve the accuracy of knowledge retrieval. Formally, given the current context representation h_u and the candidate knowledge representations $[h_{k_1}, ..., h_{k_l}]$, we compute the cosine similarity between the context and each knowledge, denoted as $\text{sim}(h_u, h_{k_i})$. When the knowledge diffusion decoder predicts a more accurate knowledge distribution z_0^k, the knowledge selector updates the cosine similarity as follows:

$$h_K = \tanh(W_u h_u + W_k z_0^k), \tag{8}$$

$$\text{Score}_i = \text{sim}(h_K, h_{k_i}), \tag{9}$$

where W_u and W_k are trainable parameters. We rank the similarity scores and retain only the best knowledge for this step. In each training step, we first update the forward model and backward model by optimizing their respective losses using a batch of training samples (U, K, R). Furthermore, we sample a batch of generated responses R' from the forward model. Together with the context U, they are fed to the backward model to generate the knowledge K'. These pseudo pairs $([U, R'], K)$ are utilized to train the forward model with the loss $\mathcal{L}_{\text{FW}}$. Additionally, to balance the bias between training and inference, we employ teacher forcing (Bengio et al., 2015). During the early stages

Methods	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-1	PPL↓	DIST-1	DIST-2
ITDD	15.8	7.1	4.0	2.5	16.2	17.8	-	-
BART$_{cat}$	23.1	11.4	6.7	4.3	19.3	19.7	7.1	29.9
BART$_{skt}$	23.2	11.9	7.6	4.4	19.4	20.3	6.8	30.3
DRD	21.8	11.5	7.5	5.5	18.0	23.0	-	-
TAKE	20.8	-	-	3.6	**27.1**	-	-	-
ZRKGC	22.2	7.3	2.8	1.8	18.6	40.4	5.4	22.5
KAT	25.5	13.9	9.0	6.6	21.6	14.5	9.3	37.0
KAT-TSLF	25.5	13.9	9.1	6.7	21.7	**14.4**	9.5	38.3
ConDDL	**25.9**	**14.1**	**9.5**	**6.9**	26.5	14.5	**10.9**	**43.7**

Table 1: The automatic evaluation results on Wizard of Wikipedia (WoW) test seen dataset.

of training, the backward model receives golden responses as inputs, while in the later stages of training, the inputs are pseudo responses generated by the forward model. During inference, when a knowledge is given to the forward model, ConDDL can predict a knowledge-grounded response with controllable generation, thereby providing a meaningful and contextually relevant output. The overall optimization objective of the model is the loss-weighted sum described above, with weights calculated based on the performance of ConDDL on the validation set.

4 Experiments

4.1 Experimental Setup

Datasets. We conducted experiments on two knowledge-grounded dialogue generation datasets: Wizard of Wikipedia (WoW) (Dinan et al., 2018) and CMU Document Grounded Conversations (CMU_DoG) (Zhou et al., 2018b). These datasets are collected via Amazon Mechanical Turk. WoW covers a wide range of topics and involves a wizard who possesses knowledge about a specific topic, and an apprentice who seeks to learn from the wizard. CMU_DoG focuses specifically on the movie domain and involves two workers who are knowledgeable about a document and engage in in-depth discussions about its content.

Baseline. To evaluate the effectiveness of our proposed method, we compare it against several baselines followed (Liu et al., 2021): **ITDD** is an Transformer-based architecture which incrementally represents multi-turn dialogues and knowledge (Li et al., 2019). **BART$_{cat}$** is a BART-based model that take the concatenation of dialogue context and all knowledge as the input of BART for response generation (Lewis et al., 2020a). **BART$_{skt}$** is variational model that introduced BART on the basis of (Lian et al., 2019b) and considered the knowledge selection history in multi-turn dialogue (Kim et al., 2020b). **DRD** (Zhao et al., 2020) intends to combat low-resource settings with pre-trained techniques. **ZRKGC** (Li et al., 2020) explores the response generation problem without leveraging the matching annotations between the context and knowledge during training. **KAT** (Liu et al., 2021) has a knowledge-aware decoder which could obtains information from the dialogue context and background documents through cross-attention and integrates them through a controller. **KAT-TSLF** propose a three-stage learning framework based on weakly supervised learning which benefits from large scale ungrounded dialogues and unstructured knowledge base.

Implementation Details. We implemente the experiments using PyTorch on an NVIDIA A100 GPU. Our code is based on the Huggingface[2]. The batch size is set to 8, and input utterances are padded or truncated to contain 512 tokens. The maximum decoding length is set to 40. BART (Lewis et al., 2020b) is initialized by BART$_{Large}$ and choose the Adam optimizer with the warm-up steps. The learning rate for the generators is set to 2×10^{-4}. We use the grid search to tune the hyper-parameters. The search ranges for learning rate and batch size are $\{1\times10^{-4}, 2\times10^{-4}, 4 \times 10^{-4}, 6 \times 10^{-4}\}$ and $\{4, 8, 16, 32\}$, respectively. We choose the parameter combination with the lowest perplexity in the validation set.

4.2 Results and Analysis

Automatic Evaluation. For automatic evaluation, we employed commonly used metrics, which include BLEU (Papineni et al., 2002) (BLEU-1, BLEU-2, BLEU-3 and BLEU-4), ROUGE (Lin, 2004) (ROUGE-1), perplexity (PPL), and DISTINCT (Li et al., 2015) in the knowledge-grounded dialogue generation following (Liu et al., 2021). The computation of ROUGE scores is based on

[2]https://huggingface.co

Methods	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-1	PPL↓	DIST-1	DIST-2
ITDD	13.4	4.7	2.1	1.1	11.4	44.8	-	-
BART$_{cat}$	23.2	11.0	6.3	4.1	18.9	24.5	5.3	22.2
BART$_{skt}$	23.4	10.9	6.8	4.6	19.0	22.3	5.2	24.5
DRD	20.7	10.1	6.2	4.3	16.5	25.6	-	-
TAKE	20.1	-	-	3.3	**26.2**	-	-	-
ZRKGC	21.8	7.1	2.7	1.1	18.5	41.5	3.4	15.6
KAT	24.4	12.5	7.8	6.6	20.5	**15.8**	10.1	39.1
KAT-TSLF	24.1	12.9	8.3	6.0	20.7	**15.8**	6.7	26.0
ConDDL	**24.6**	**13.1**	**8.5**	**6.3**	25.9	16.1	**10.5**	**42.3**

Table 2: The automatic evaluation results on Wizard of Wikipedia (WoW) test unseen dataset.

Methods	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-1	PPL↓	DIST-1	DIST-2
ITDD	9.5	3.6	1.7	0.9	10.4	26.0	-	-
BART$_{cat}$	17.0	8.6	5.3	3.4	13.6	36.4	1.5	7.3
BART$_{skt}$	16.2	8.3	5.1	3.1	12.7	40.1	1.2	7.3
DRD	15.0	5.7	2.5	1.2	10.7	54.4	-	-
ZRKGC	15.1	4.2	1.2	0.4	12.5	53.5	1.2	8.1
KAT	19.4	10.5	6.9	4.7	14.4	22.2	1.8	8.9
KAT-TSLF	20.4	10.6	6.7	4.4	15.1	21.7	2.0	11.1
ConDDL	**21.0**	**10.8**	**6.8**	**4.5**	**17.8**	**16.9**	**2.3**	**12.6**

Table 3: The automatic evaluation results on CMU_DoG dataset.

n-grams and is performed using pyrouge package[3].

The overall results under full-dataset scenarios of both WoW and CMU_DoG datasets are shown in Table 1, Table 2 and Table 3, respectively. Our proposed ConDDL model consistently outperforms other dialogue models, including the state-of-the-art models TAKE and KAT. On the WoW dataset, ConDDL achieves improvements of 1.57%, 3.70%, and 14.74% than KAT-TSLF in terms of BLEU-1, ROUGE-1 and DIST-1, respectively. Notably, ConDDL also demonstrates consistently better performance in terms of DIST compared to the baselines. These results highlight the effectiveness of our approach in enhancing semantic and diversity modeling. Our model achieves a lower ROUGE-1 score than TAKE and comparable PPL to KAT on the WOW dataset, which is attributed to the difference in the non-autoregressive generation method and the pre-trained language model. Furthermore, on the CMU_DoG dataset, ConDDL achieves significant performance improvement in terms of the all automatic metrics. The improvement can be attributed to the better utilization of historical information and the diffusion learning mechanism employed in our method. ConDDL shows substantial improvements across all generation metrics, indicating its ability to generate more informative and engaging responses. Statistical significance tests using t-tests confirm that ConDDL outperforms the baselines with a $p-$value of less than 0.05.

[3]https://pypi.org/project/pyrouge

Manual Evaluation. To complement the automatic metrics, we conducted a manual evaluation focusing on fluency, coherence, and informativeness of the generated responses. The results presented in Table 4 demonstrate that ConDDL outperforms the baseline models in both manual metrics. The kappa statistics[4] measuring the agreement between annotators are 0.71, 0.63, and 0.66 for fluency, coherence, and informativeness, respectively, indicating substantial agreement. Importantly, ConDDL exhibited a significant improvement in informativeness, indicating that the integration of external data enhanced the ability of model to comprehend. ConDDL is capable of generating responses that incorporate flexible knowledge and leverage global information to produce more relevant and informative responses.

Ablation Study. To evaluate the effectiveness of each module in ConDDL, we perform an ablation study where we remove key modules from our framework one by one. The results are presented in Table 5 and the ablation models are evaluated using all metrics. The removed modules include the dual learning module (DualLearning), diffusion learning (DiffLearning), and contrastive learning (ContLearning). The results indicate that all ablation models performe worse than ConDDL across all metrics, highlighting the superiority of

[4]Landis and Koch (1977) characterize kappa values < 0 as no agreement, 0-0.20 as slight, 0.21-0.40 as fair, 0.41-0.60 as moderate, 0.61-0.80 as substantial, and 0.81-1 as almost perfect agreement.

Methods	WoW Test Seen			WoW Test Unseen		
	Information	Coherence	Fluency	Information	Coherence	Fluency
BART$_{skt}$	0.52	0.61	0.52	0.49	0.61	0.49
ZRKGC	0.53	0.59	0.56	0.52	0.57	0.51
KAT-TSLF	0.55	0.66	0.63	0.54	0.61	**0.69**
ConDDL	**0.61**	**0.70**	**0.66**	**0.60**	**0.71**	0.65

Table 4: Manual Evaluation results on the WoW dataset.

Methods	WoW Test Seen				WoW Test Unseen			
	BL-1	BL-4	RG-1	DIST-1	BL-1	BL-4	RG-1	DIST-1
ConDDL	**25.9**	**6.9**	**26.5**	**10.9**	**24.6**	**6.3**	**25.9**	**10.5**
w/o DualLearning	24.3	6.1	26.1	10.1	24.1	5.9	25.6	10.0
w/o DiffLearning	23.9	5.9	25.6	9.7	23.5	5.4	25.1	9.9
w/o ContLearning	24.7	6.4	25.9	10.3	24.5	6.1	25.3	10.2

Table 5: The Ablation study on the WoW dataset.

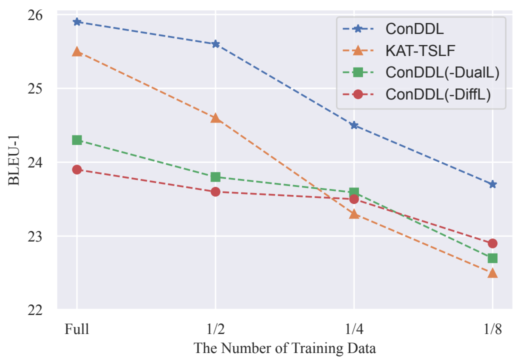

Figure 3: Performance of the proposed model with different number of training data on the WoW.

ConDDL. In particular, the diffusion learning is found to contribute the most to the overall performance, demonstrating the need to incorporate multi-stage diffusion learning with external knowledge. This phenomenon is mainly due to the fact that knowledge-based dialogue generation requires more precise information injection, as opposed to non-knowledge dialogue. This finding is consistent with our hypothesis that introducing a large amount of external knowledge without effective constraints can reduce the impact of the primary goal.

Low-Resource Settings. To investigate the factors influencing the performance of ConDDL in low-resource scenarios, we randomly select different numbers of training samples for all datasets and evaluate the performance of our model using the ablation study. Figure 3 illustrates the results of the baseline KAT-TSLF and the ablation study in terms of ConDDL. The experimental results show that the data requirements of ConDDL are significantly lower than the baseline model. These findings emphasize the significance of the proposed model for low-resource knowledge-grounded dialogue generation. Removing any component from ConDDL resulted in a performance drop when the training data was limited. Furthermore, the dual learning module was found to be the most sensitive component, and its removal has a greater impact on the overall performance. This also shows that the iterative dual structure helps the model to make full use of the available data in low-resource scenarios. Due to the presence of the dual module, our method can automatically expand the training data and improve its performance when data is limited. While the diffusion learning module played a larger role when the training data exceeded a quarter of the available dataset, it can inject knowledge into the response more accurately.

5 Conclusions

In this paper, we propose a knowledge-grounded dialogue generation framework called Controllable Dual Diffusion Learning (ConDDL) to mitigate the problem of knowledge selection and dialogue generation. ConDDL leverages dual learning and diffusion learning to effectively exploit knowledge beyond the generation process. We incorporate knowledge information into the diffusion process, which guides the model to allocate pay more attention to the precise knowledge in the training process. We formulate response generation and knowledge generation as dual tasks to fully leverage the prior and posterior knowledge. Experimental results on two public datasets, Wizard of Wikipedia and CMU_DoG, demonstrate the significant performance of the proposed ConDDL model, validating its effectiveness. Furthermore, we aim to explore methods for utilizing limited data, which is crucial for the exploitation of unlabelled knowledge data.

Limitations

The limitations focus on text length, the number of dialogue characters, and GPU resources in this work. We know that text length limits the modeling ability of the model in the natural language process, and the same is valid for dialogue generation. If the dialogue involves more than 20 or 30 rounds, it will significantly reduce the ability of the model to capture important information. In addition, the number of interlocutors involved in the conversation is significant. We do not discuss the loss of modeling effectiveness with more than three interlocutors due to space, but we will explore this issue in more detail in the future. Finally, due to the limitation of GPU resources, we could not set a larger batch size, resulting in the model lacking more super-reference search space. These issues will be investigated in more depth in the future. Besides, if there is malicious information in the dataset, it might generate harmful responses like most generative models. This phenomenon is a potential risk in data-driven models and requires us to explore additional control techniques.

Ethics Statement

This paper proposes a knowledge-grounded dialogue generation framework called Controllable Dual Diffusion Learning. The research will not pose ethical issues. We have considered the ethical implications of our research across different frameworks. We have ensured fair compensation for human evaluators, used publicly available datasets, and minimized the introduction of biases. By addressing these ethical considerations, we aim to contribute to responsible and impactful advancements in dialogue generation, knowledge distillation, and open-domain question answering.

References

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam M. Shazeer. 2015. Scheduled sampling for sequence prediction with recurrent neural networks. *ArXiv*, abs/1506.03099.

Xiuyi Chen, Fandong Meng, Peng Li, Feilong Chen, Shuang Xu, Bo Xu, and Jie Zhou. 2020. Bridging the gap between prior and posterior knowledge selection for knowledge-grounded dialogue generation. In *Conference on Empirical Methods in Natural Language Processing*.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M. F. Moura, Devi Parikh, and Dhruv Batra. 2016. Visual dialog. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1080–1089.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2018. Wizard of wikipedia: Knowledge-powered conversational agents. *ArXiv*, abs/1811.01241.

Zhidan Feng, Zhenyu Zhang, Xintong Yu, Yewei Fang, Lanxin Li, Xuyi Chen, Yuxiang Lu, Jiaxiang Liu, Weichong Yin, Shi Feng, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2022. Ernie-vilg 2.0: Improving text-to-image diffusion model with knowledge-enhanced mixture-of-denoising-experts. *ArXiv*, abs/2210.15257.

Daniel De Freitas, Minh-Thang Luong, David R. So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, and Quoc V. Le. 2020. Towards a human-like open-domain chatbot. *ArXiv*, abs/2001.09977.

Shansan Gong, Mukai Li, Jiangtao Feng, Zhiyong Wu, and Lingpeng Kong. 2022. Diffuseq: Sequence to sequence text generation with diffusion models. *ArXiv*, abs/2210.08933.

Di He, Yingce Xia, Tao Qin, L. Wang, N. Yu, T. Liu, and W. Ma. 2016. Dual learning for machine translation. In *NIPS*.

Jonathan Ho, Ajay Jain, and P. Abbeel. 2020. Denoising diffusion probabilistic models. *ArXiv*, abs/2006.11239.

Byeongchang Kim, Jaewoo Ahn, and Gunhee Kim. 2019. Sequential latent knowledge selection for knowledge-grounded dialogue. In *International Conference on Learning Representations*.

Byeongchang Kim, Jaewoo Ahn, and Gunhee Kim. 2020a. Sequential latent knowledge selection for knowledge-grounded dialogue. *ArXiv*, abs/2002.07510.

Byeongchang Kim, Jaewoo Ahn, and Gunhee Kim. 2020b. Sequential latent knowledge selection for knowledge-grounded dialogue. *arXiv preprint arXiv:2002.07510*.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020a. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020b. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *ACL*.

Jinpeng Li, Yingce Xia, Rui Yan, Hongda Sun, Dongyan Zhao, and Tie-Yan Liu. 2021. Stylized dialogue generation with multi-pass dual learning. In *NeurIPS*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2015. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*.

Lin-Xiao Li, Can Xu, Wei Wu, Yufan Zhao, Xueliang Zhao, and Chongyang Tao. 2020. Zero-resource knowledge-grounded dialogue generation. *ArXiv*, abs/2008.12918.

Xiang Lisa Li, John Thickstun, Ishaan Gulrajani, Percy Liang, and Tatsunori Hashimoto. 2022. Diffusion-lm improves controllable text generation. *ArXiv*, abs/2205.14217.

Zekang Li, Cheng Niu, Fandong Meng, Yang Feng, Qian Li, and Jie Zhou. 2019. Incremental transformer with deliberation decoder for document grounded conversations. *arXiv preprint arXiv:1907.08854*.

Rongzhong Lian, Min Xie, Fan Wang, Jinhua Peng, and Hua Wu. 2019a. Learning to select knowledge for response generation in dialog systems. In *International Joint Conference on Artificial Intelligence*.

Rongzhong Lian, Min Xie, Fan Wang, Jinhua Peng, and Hua Wu. 2019b. Learning to select knowledge for response generation in dialog systems. *arXiv preprint arXiv:1902.04911*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*.

Shilei Liu, Xiaofeng Zhao, Bochao Li, Feiliang Ren, Longhui Zhang, and Shujuan Yin. 2021. A three-stage learning framework for low-resource knowledge-grounded dialogue generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2262–2272.

Zihan Liu, Mostofa Patwary, Ryan J. Prenger, Shrimai Prabhumoye, Wei Ping, Mohammad Shoeybi, and Bryan Catanzaro. 2022. Multi-stage prompting for knowledgeable dialogue generation. *ArXiv*, abs/2203.08745.

Chuan Meng, Pengjie Ren, Zhumin Chen, Weiwei Sun, Zhaochun Ren, Zhaopeng Tu, and M. de Rijke. 2020. Dukenet: A dual knowledge interaction network for knowledge-grounded conversation. *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. 2021. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *International Conference on Machine Learning*.

Kishore Papineni, S. Roukos, T. Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*.

Robin Rombach, A. Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2021. High-resolution image synthesis with latent diffusion models. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685.

Iulian Serban, Alessandro Sordoni, Yoshua Bengio, Aaron C. Courville, and Joelle Pineau. 2015. Building end-to-end dialogue systems using generative hierarchical neural network models. In *AAAI Conference on Artificial Intelligence*.

Harry Shum, Xiaodong He, and Di Li. 2018. From eliza to xiaoice: challenges and opportunities with social chatbots. *Frontiers of Information Technology & Electronic Engineering*, 19:10–26.

Wenquan Wu, Zhen Guo, Xiangyang Zhou, Hua Wu, Xiyuan Zhang, Rongzhong Lian, and Haifeng Wang. 2019. Proactive human-machine conversation with explicit conversation goal. In *Annual Meeting of the Association for Computational Linguistics*.

Lin Xu, Qixian Zhou, Jinlan Fu, Min-Yen Kan, and See-Kiong Ng. 2022. Corefdiffs: Co-referential and differential knowledge flow in document grounded conversations. *arXiv preprint arXiv:2210.02223*.

Chenxu Yang, Zheng Lin, Jiangnan Li, Fandong Meng, Weiping Wang, Lanrui Wang, and Jie Zhou. 2022. Take: topic-shift aware knowledge selection for dialogue generation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 253–265.

Hongyi Yuan, Zheng Yuan, Chuanqi Tan, Fei Huang, and Songfang Huang. 2022. Seqdiffuseq: Text diffusion with encoder-decoder transformers. *ArXiv*, abs/2212.10325.

Haolan Zhan, Lei Shen, Hongshen Chen, and Hainan Zhang. 2021. Colv: A collaborative latent variable model for knowledge-grounded dialogue generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2250–2261.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. 2023. Siren's song in the ai ocean: A survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.

Xueliang Zhao, Wei Wu, Chongyang Tao, Can Xu, Dongyan Zhao, and Rui Yan. 2020. Low-resource knowledge-grounded dialogue generation. *ArXiv*, abs/2002.10348.

Chujie Zheng, Yunbo Cao, Daxin Jiang, and Minlie Huang. 2020. Difference-aware knowledge selection for knowledge-grounded conversation generation. *ArXiv*, abs/2009.09378.

Wen Zheng, Natasa Milic-Frayling, and Ke Zhou. 2021. Knowledge-grounded dialogue generation with term-level de-noising. In *FINDINGS*.

Hao Zhou, Tom Young, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2018a. Commonsense knowledge aware conversation generation with graph attention. In *International Joint Conference on Artificial Intelligence*.

Kangyan Zhou, Shrimai Prabhumoye, and Alan W. Black. 2018b. A dataset for document grounded conversations. In *Conference on Empirical Methods in Natural Language Processing*.

Pei Zhou, Behnam Hedayatnia, Karthik Gopalakrishnan, Seokhwan Kim, Jay Pujara, Xiang Ren, Yang Liu, and Dilek Z. Hakkani-Tür. 2021. Think before you speak: Learning to generate implicit knowledge for response generation by self-talk. *Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI*.

A Appendix

A.1 Case Study

Table 6 presents a comparison between the reference response and the outputs of various models on the WoW dataset. It can be observed that the knowledge sentences contained differences in judgments of hair length based on gender (indicated in red and blue text). The baselines fail to capture this vital information during the dialogue generation, leading to less incorporation of knowledge in their responses. In contrast, ConDDL produced a response that is fluent and consistent with the knowledge. ConDDL does not rely on complex structure or fact analysis modules but rather relies on end-to-end training to understand the original text well. When enhanced knowledge is introduced externally, the major challenge lies in the inconsistency of data distribution. Manual evaluation and case analysis revealed that ConDDL demonstrated a better understanding of dialogue history and generated responses containing more factual information. This is because the ConDDL model can automatically capture the central representation and global-related knowledge through dual diffusion learning module, thereby strengthening the main modeling content.

Topic	Brown hair
Knowledge	For example, a woman with chin-length hair in some cultures may be said to have short hair, while a man with the same length of hair in some of the same cultures would be said to have long hair.
Dialogues	A: Hi. I have brown hair. B: That's nice. Is it long or short hair? A: That depends on where you're from. Long hair is a concept that varies culturally. B: That's true. I guess I would rephrase it as do you have shoulder-length hair or longer?
Reference	It's about to my shoulders. I think most Americans would say medium. On a guy it would be "long".
BART$_{skt}$	I think it is a long hair.
KAT-TSLF	I'm not sure but the hair is brown.
ConDDL	It's shoulder-length hair. I guess some cultures would say it's short hair. But for men it's long.

Table 6: Examples of the generated response by ConDDL and other models on the WoW dataset.

A.2 Manual Evaluation

We evaluate the models with the followed three metrics: (i) *Fluency*, which indicates the smoothness of the sentence. (ii) *Coherence*, which measures the consistency of the response with the context.

Datasets	WoW				CMU_DoG		
	Train	**Valid**	**Test Seen**	**Test Unseen**	**Train**	**Valid**	**Test**
# of Utterances	166,787	17,715	8,715	8,782	74,717	4,993	13,646
# of Dialogues	18,430	1,948	965	968	3,373	229	619
# of Topics	1,247	599	533	58	30	30	30
Avg. # Turn	9.0	9.1	9.0	9.1	22.2	21.8	22.0
Avg. # words/turn	16.4	16.4	16.4	16.1	18.6	20.1	18.1
Avg. # knowledge entries	61.2	61.5	60.8	61.0	31.3	30.4	31.8
Avg. # words/knowledge	37.2	37.6	36.9	37.0	27.2	28.2	27.0

Table 7: The statistics for WoW and CMU_DoG datasets. "#" means the number of pairs.

(iii) *Informativeness*, which evaluates how well the response aligns with the target informativeness. We randomly selected 100 responses from the WoW test seen set and 100 responses from the test unseen set. Three well-educated annotators independently judged the responses generated by ConDDL and the baseline models. The annotators rated the responses based on fluency and informativeness, using scores ranging from 0 to 1 (with 1 being the best). To give a fair salary, we first evaluate 50 samples by ourselves, calculate the time and effort, and set this amount (samples evaluated by ourselves are just for evaluating the salary, which is not given to evaluators and not reported in the final results). The dialogues are presented to the annotators in a random order. All generated responses were fairly capitalized and detokenized.

A.3 Dataset Statistics

Table 7 provides statistics for both datasets.