DiffusionDialog: A Diffusion Model for Diverse Dialog Generation with Latent Space

Jianxiang Xiang[1], **Zhenhua Liu**[1], **Haodong Liu**[2],
Yin Bai[2], **Jia Cheng**[2], **Wenliang Chen**[1]*
[1]School of Computer Science and Technology, Soochow University, China
[2]Meituan
{jxxiang0720, zhliu0106}@stu.suda.edu.cn, wlchen@suda.edu.cn
{liuhaodong05, baiyin, jia.cheng.sh}@meituan.com

Speaker:Jianxiang Xiang

Code: https://github.com/Jxxiang99/DiffusionDialog

Outline

- Introduction
- Method
- Experiment
- Conclusion

Introduction

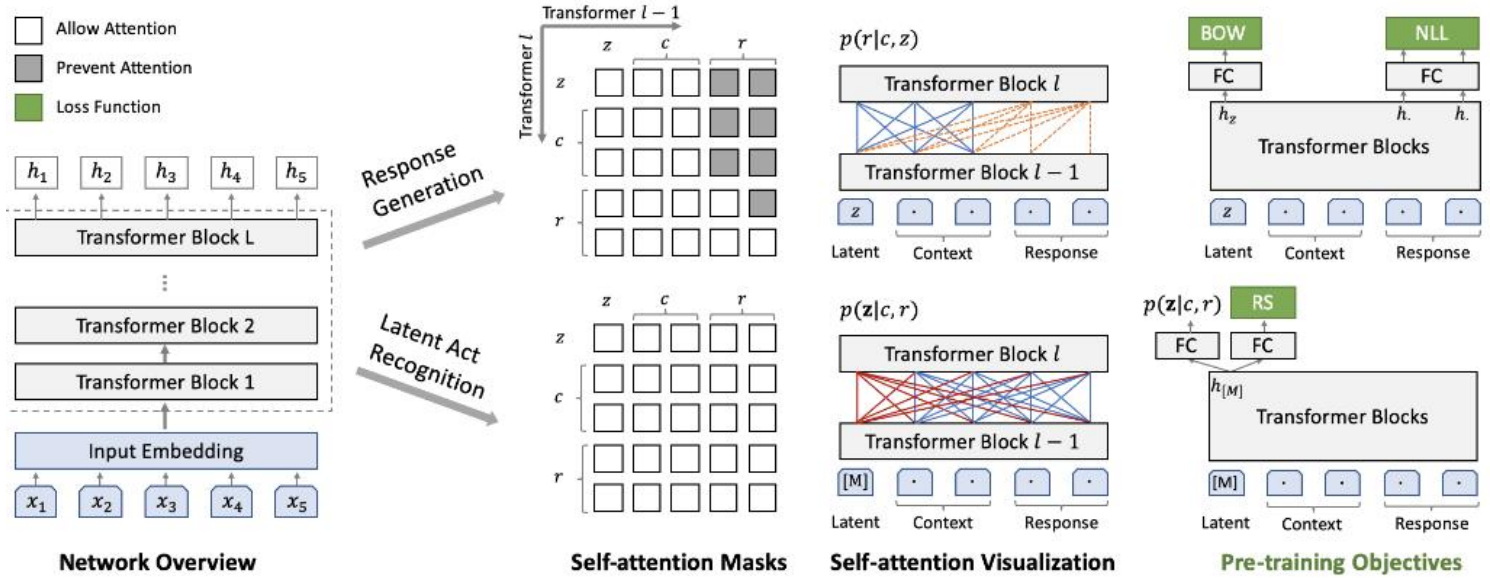
Dialogue tasks have one-to-many problems

- In real-life conversations, the content is diverse, and there exists the one-to-many problem that requires diverse generation.

1.1 Background

Dialogue tasks have one-to-many problems

- Previous studies attempted to introduce discrete or Gaussian-based continuous latent variables to address the one-to-many problem.

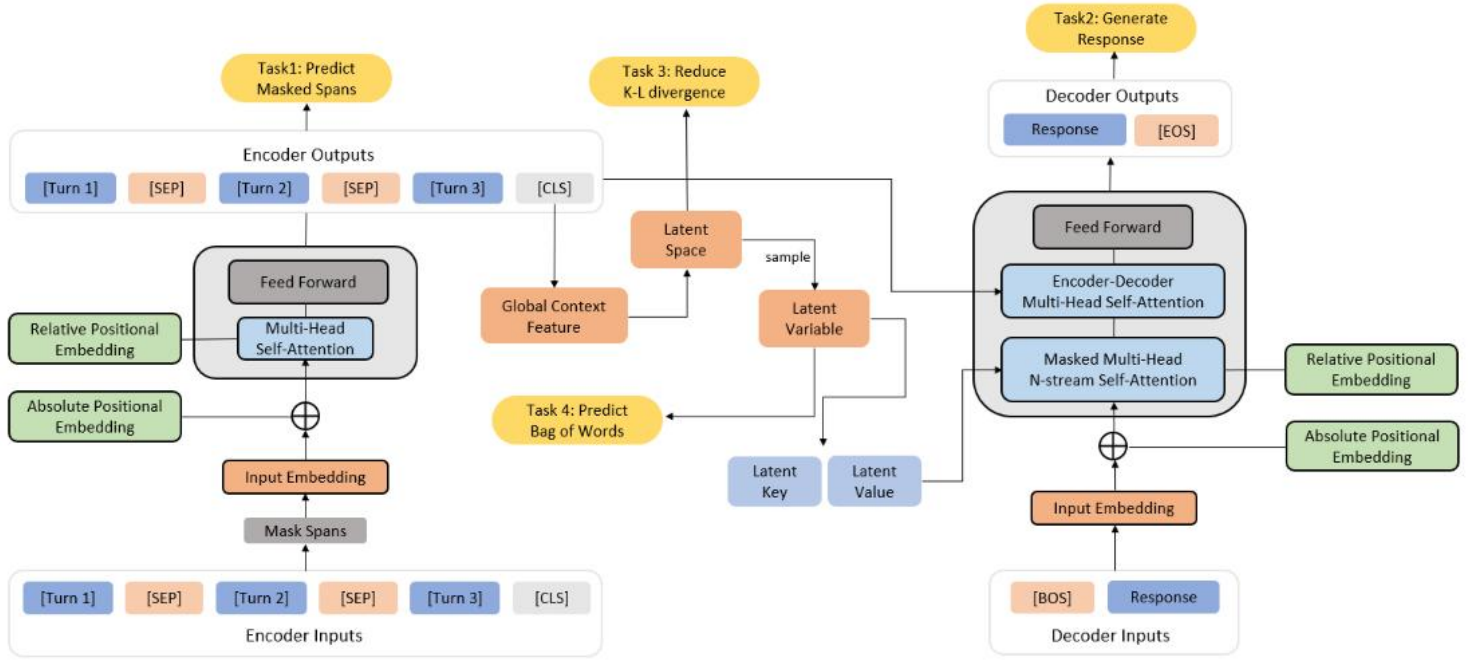


[1] Siqi Bao, Huang He, Fan Wang, Hua Wu, and Haifeng Wang. 2019. Plato: Pre-trained dialogue generation model with discrete latent variable. arXiv preprint arXiv:1910.07931.

1.1 Background

Dialogue tasks have one-to-many problems

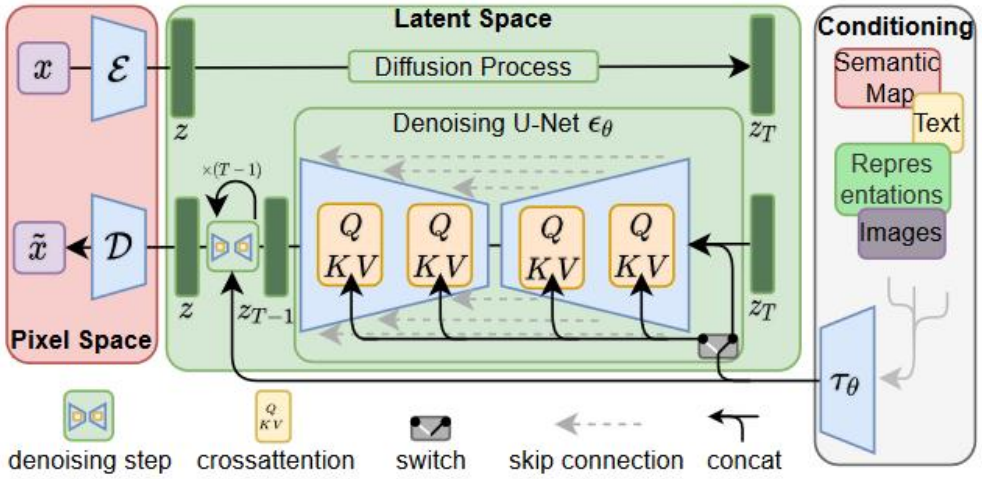
- Previous studies attempted to introduce discrete or Gaussian-based continuous latent variables to address the one-to-many problem.

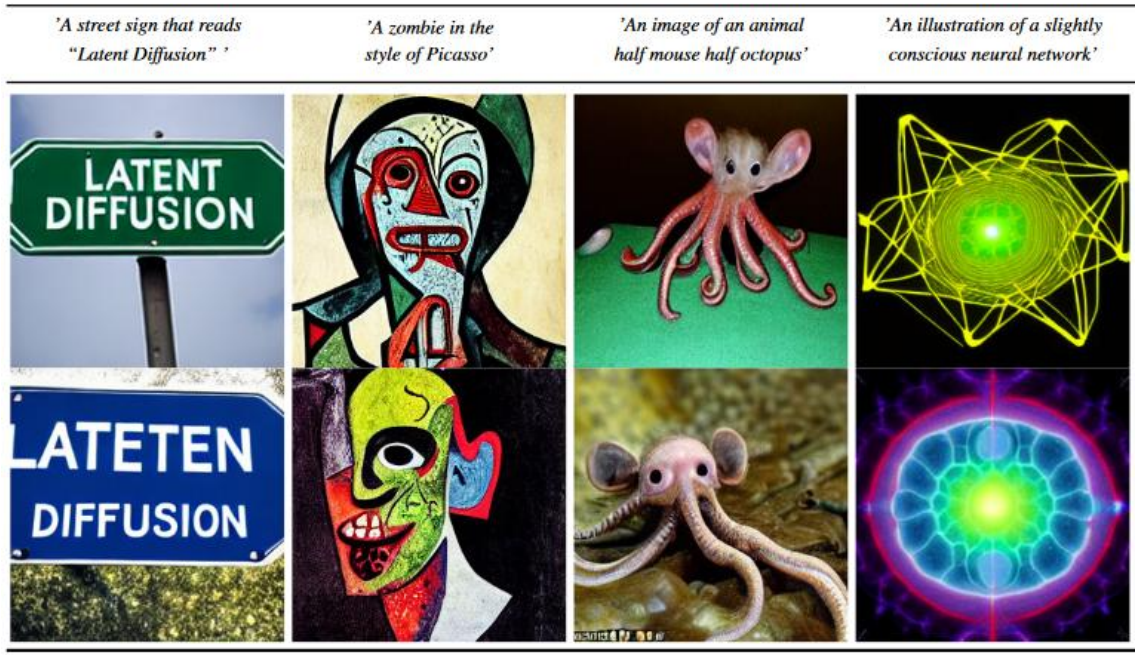


[2] Wei Chen, Yeyun Gong, Song Wang, Bolun Yao, Weizhen Qi, Zhongyu Wei, Xiaowu Hu, Bartuer Zhou, Yi Mao, Weizhu Chen, et al. 2022b. Dialogved: A pre-trained latent variable encoder decoder model for dialog response generation. arXiv preprint arXiv:2204.13031.

Introduce the Diffusion model into Dialogue tasks

- Diffusion model have shown its' superiority of generating high-quality and diverse results in the fields of image and audio.

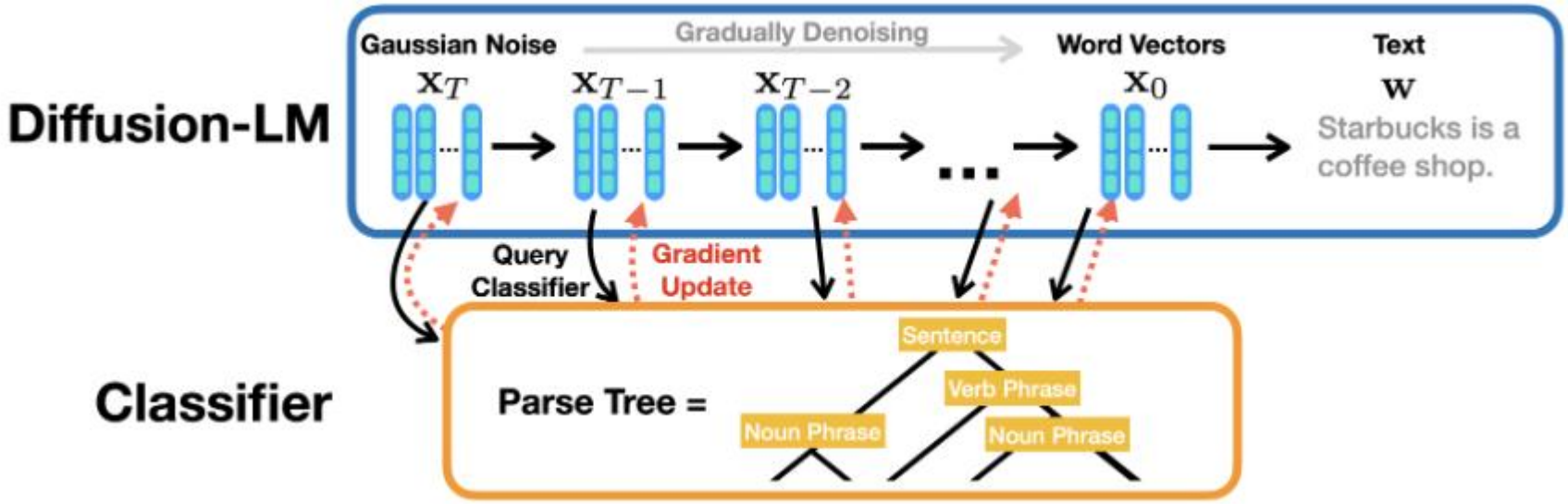




[3] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 10684–10695.

Introduce the Diffusion model into Dialogue tasks

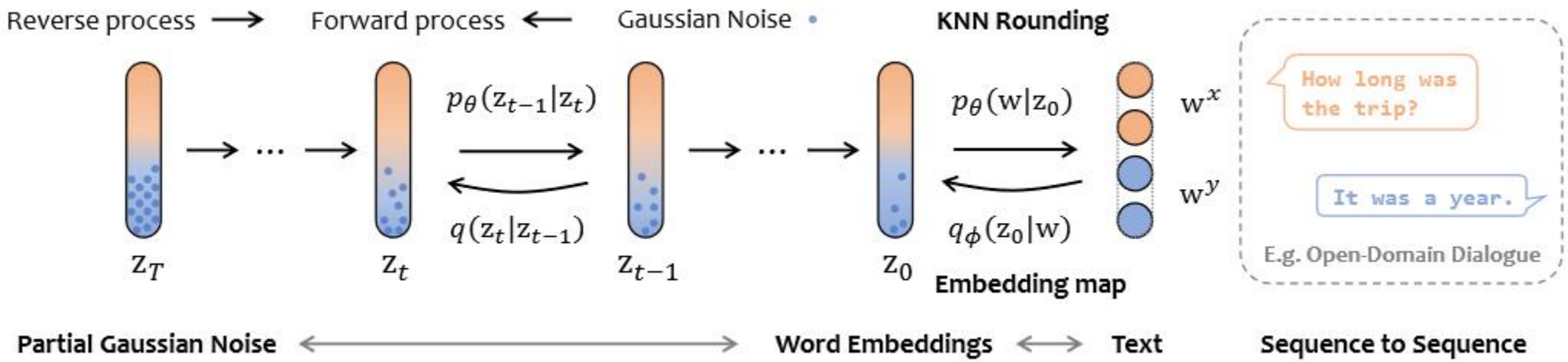
- Some attempts have been made in text generation tasks.



[4] Xiang Li, John Thickstun, Ishaan Gulrajani, Percy S Liang, and Tatsunori B Hashimoto. 2022. Diffusion-lm improves controllable text generation. Advances in Neural Information Processing Systems, 35:4328–4343.

Introduce the Diffusion model into Dialogue tasks

- Some attempts have been made in text generation tasks.



[5] Shansan Gong, Mukai Li, Jiangtao Feng, ZhiyongWu, and LingPeng Kong. 2022. Diffuseq: Sequence to sequence text generation with diffusion models. arXiv preprint arXiv:2210.08933.

1.2 Contributions

- We propose a novel approach to address the one-to-many problem in dialogue using a combination of a latent-based diffusion model and a pre-trained language model.
- To the best of our knowledge, our work is the first to apply a latent diffusion model to dialog generation. By reasoning in the latent space, the inference efficiency of our diffusion model is significantly improved.
- Through comparative experiments, we demonstrate the effectiveness of our model, which can generate responses that are rich in diversity while ensuring fluency and coherence.

Method

Dialog Generation with Latent Variable

- Variables:z-latent vatiable, c-dialog context, x responce.

- Goal: $p_\theta(x|c) = \int p_\theta(z|c)p_\theta(x|z,c)d_z$

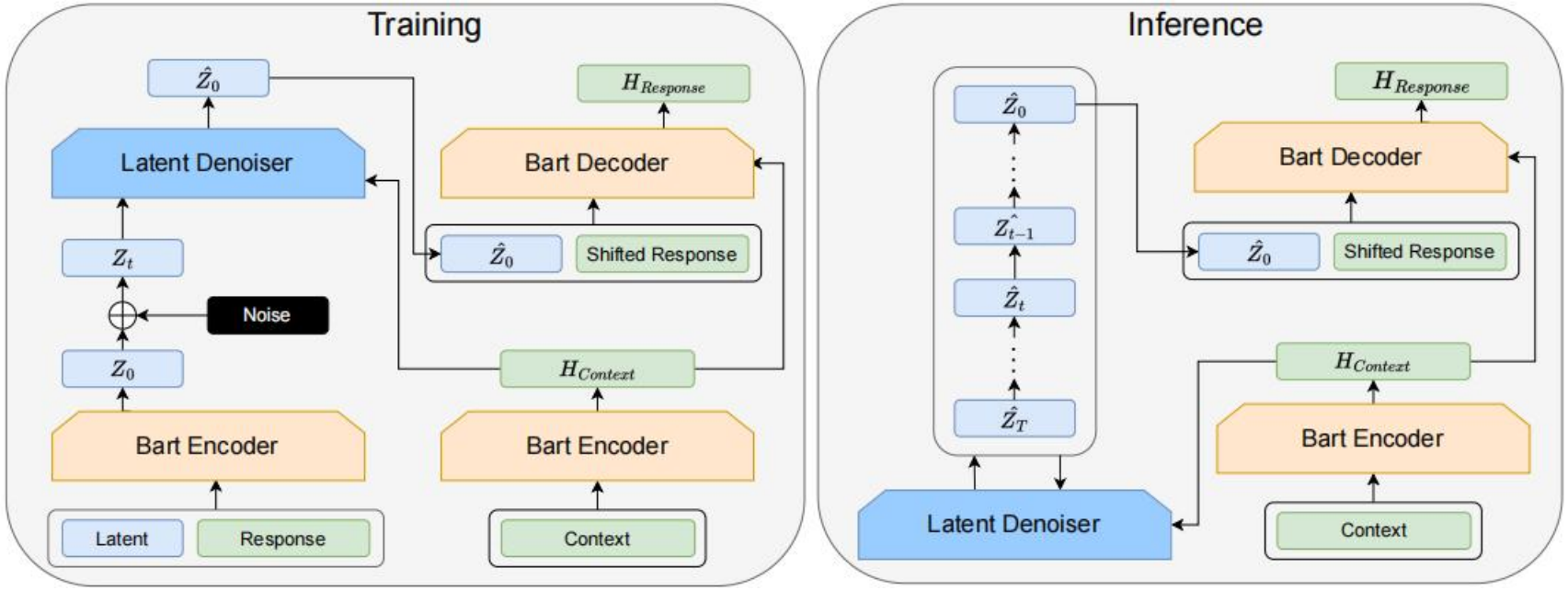
2.1 Background Knowledge

Diffusion Model in Latent Space

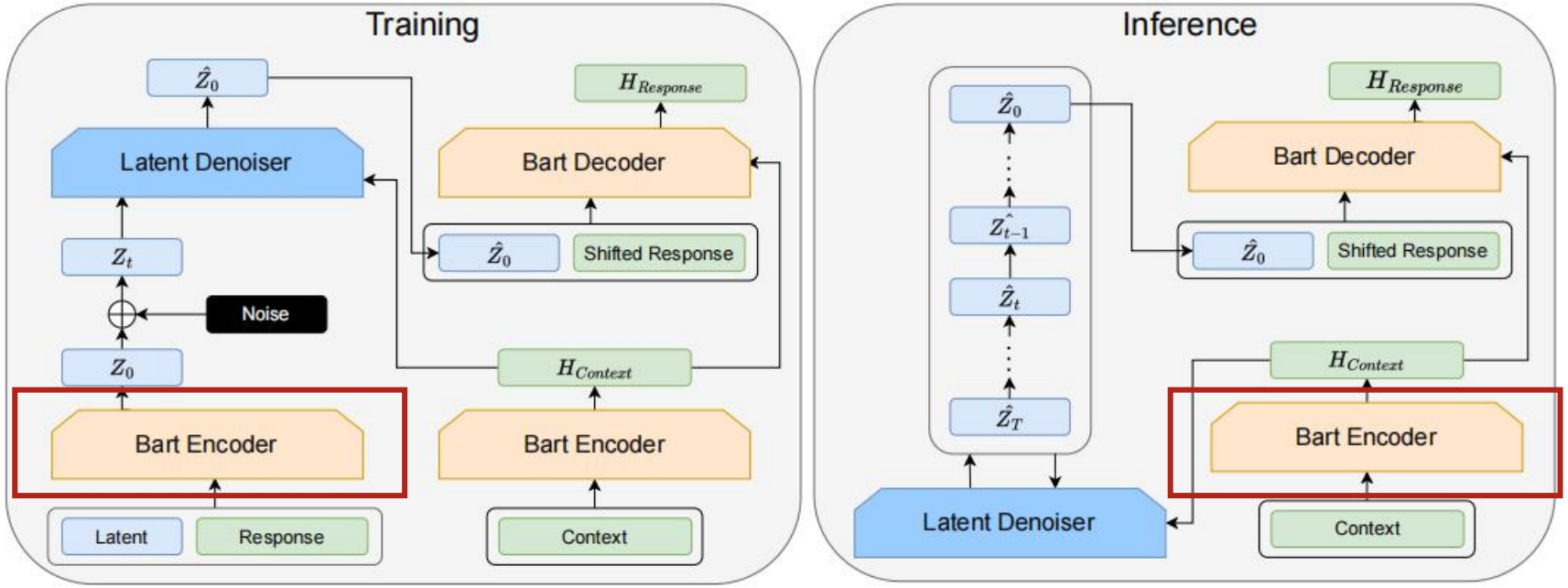
- Forward Process: $q(z_t|z_{t-1}) = N(z_t; \sqrt{1-\beta_t} z_{t-1}, \beta_t I)$

- Backward Process: $\mathcal{L}_{simple}(z_0) = \sum_{t=1}^{T} E_{q(z_t|z_0)} \|\mu_\theta(z_t, t) - \hat{\mu}(z_t, z_0)\|^2$

2.2 Model


Training
$\hat{Z}_0$
Latent Denoiser
Z_t
Noise
Z_0
Bart Encoder
Latent
Response
$H_{Response}$
Bart Decoder
$\hat{Z}_0$
Shifted Response
$H_{Context}$
Bart Encoder
Context
Inference
$\hat{Z}_0$
$\hat{Z}_{t-1}$
$\hat{Z}_t$
$\hat{Z}_T$
Latent Denoiser
$H_{Response}$
Bart Decoder
$\hat{Z}_0$
Shifted Response
$H_{Context}$
Bart Encoder
Context

2.2 Model



Bart Encoder

Two tasks:

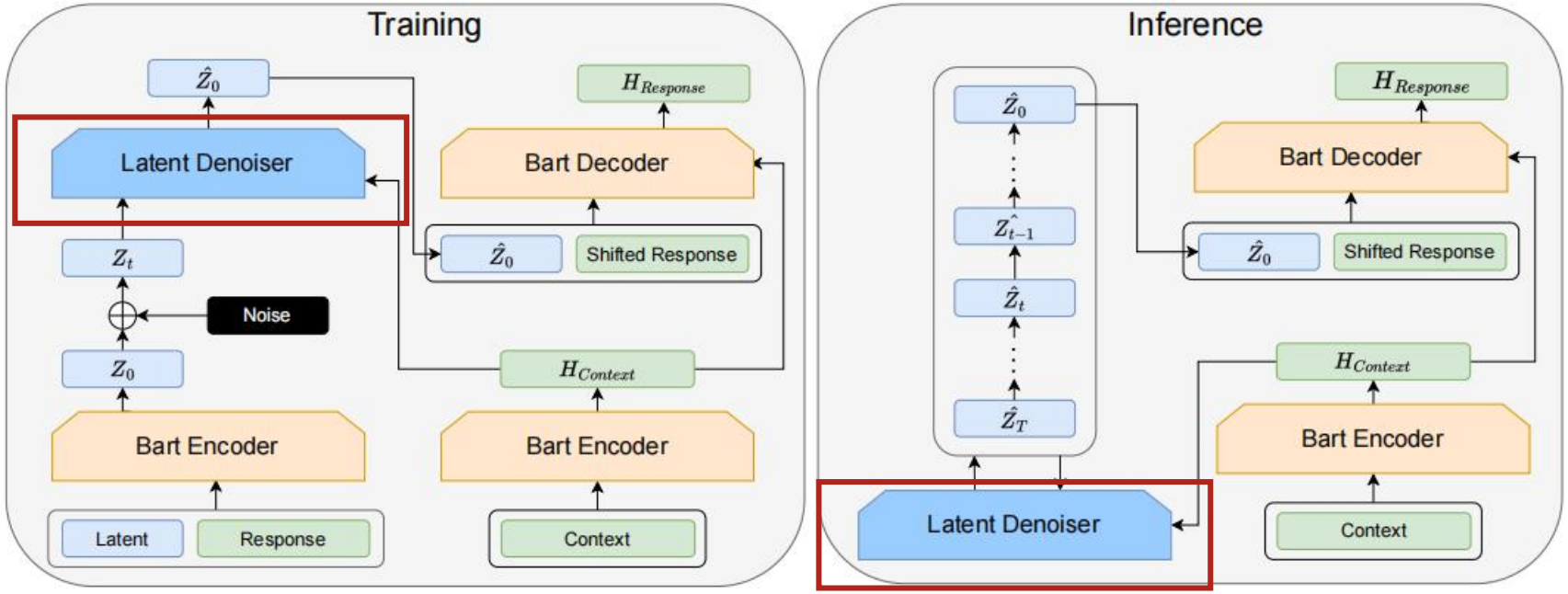
- Encode the context
- Encode the prior of the latent variable

Objective Function: **BOW Loss**

$$\mathcal{L}_{BOW} = -E_{z_0 \sim q(z|r)} \sum_{n=1}^{N} logp(r_t|z_0)$$

$$= -E_{z_0 \sim q(z|r)} \sum_{n=1}^{N} log \frac{e^{f_{r_n}}}{\sum_{v \in V} e^{f_v}}$$

2.2 Model



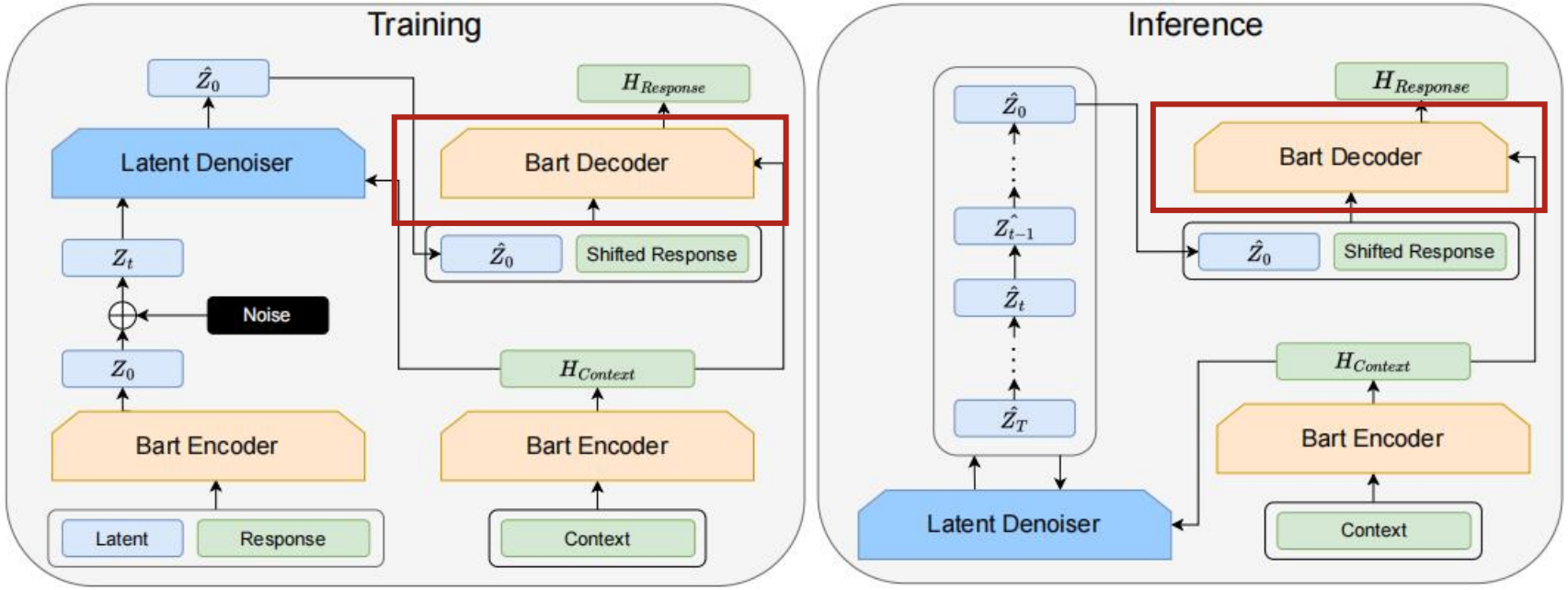
Latent Denoiser

Task: learn to denoise the noised latent.

Objective Function: **Latent Denoising Loss**

$$z_0 \sim q(z|r) = Encoder([l; x])[0]$$
$$q(x_t|x_0) = N(x_t; \sqrt{a_t}x_0, (1 - a_t)I)$$
$$\tilde{z}_0 = Denoiser(z_t, e_t, h_c)$$
$$\mathcal{L}_{LD} = - \|\tilde{z}_0 - z_0\|^2$$

2.2 Model



Bart Decoder

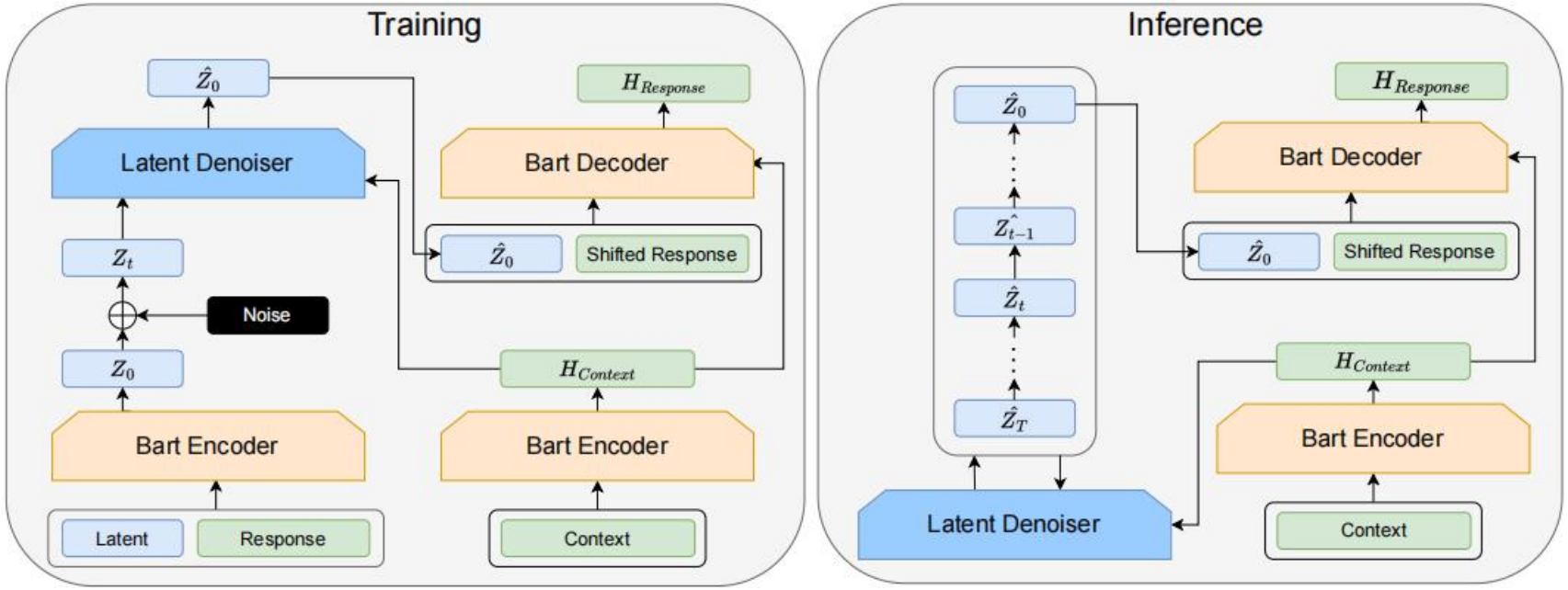
Task: learn to generate the response conditioning on both the latent variable and the context.

Objective Function: **NLL loss**

$$\mathcal{L}_{NLL} = -E_{\tilde{z}_0 \sim p(z|c,z_t,t)} \log p(r|c,\tilde{z}_0)$$

$$= -E_{\tilde{z}_0 \sim p(z|c,z_t,t)} \sum_{n}^{N} \log p(r_t|c,\tilde{z}_0,r_{<t})$$

2.2 Model



Final Training Loss

$$\mathcal{L}=\mathcal{L}_{BOW}+\mathcal{L}_{NLL}+\mathcal{L}_{LD}$$

Experiment

3.1 Experimental Setup

Datasets and Evaluation Metrics

Datasets	DailyDialog	PersonaChat
train	76052 samples	122499 samples
dev	7069 samples	14602 samples
test	6740 samples	14056 samples

	Metrics
Accuracy	BLEU-1
	BLEU-2
Diversity	Distinct-1
	Distinct-2

3.2 Main Results

DailyDialog

模型	BLEU-1	BLEU-2	Distinct-1	Distinct-2
Seq2Seq	0.336	0.238	0.030	0.128
iVAE_MI	0.309	0.249	0.029	0.250
PLATO w/o latent	0.405	0.322	0.046	0.246
PLATO	0.397	0.311	0.054	0.291
DialogVED	0.481	0.421	0.042	0.232
DiffusionDialog w/o latent	**0.406**	**0.371**	0.046	0.217
DiffusionDialog	0.348	0.318	**0.072**	**0.372**
DiffusionDialog upper bound	0.471	0.424	0.063	0.348

PersonaChat

模型	BLEU-1	BLEU-2	Distinct-1	Distinct-2
Seq2Seq	0.448	0.353	0.004	0.016
LIC	0.405	0.320	0.019	0.113
PLATO w/o latent	**0.458**	0.357	0.012	0.064
PLATO	0.406	0.315	0.021	0.121
DialogVED	0.482	0.399	0.015	0.094
DiffusionDialog w/o latent	0.410	**0.359**	0.013	0.056
DiffusionDialog	0.383	0.329	**0.031**	**0.177**
DiffusionDialog upper bound	0.472	0.411	0.026	0.166

DiffusionDialog dramatically increases output diversity with a small loss of accuracy.

Case Analysis

Context	[P1] It's a lovely day out today, isn't it? [P2] It's beautiful. Enjoy it while it lasts. It's supposed to get cold tomorrow. [P1] What's the weather forecast for tomorrow?
DiffusionDialog w/o latent	It's supposed to snow.
DiffusionDialog	Cloudy in the morning and overcast in the afternoon
	The weatherman says it's going to snow tomorrow
	It's supposed to snow in the morning and in the afternoon. Is that possible?
	Dreadful. It's supposed to snow tomorrow.
	The weatherman says it's supposed to snow all day.
	The weatherman says a storm is coming.
	Yep. It's supposed to snow in the morning.
	Cloudy with a chance of showers.

The output of the DiffusionDialog is very diverse in style and wording.

3.3 Discussions

Inference Speed

Model	Speed：second/sample
DiffusionDialog w/o latent	0.068
PLATO	25.813
DialogVED	0.076
DiffusionDialog-10	0.072
DiffusionDialog-100	0.189
DiffusionDialog-1000	1.500
DiffuSeq-10	0.384
DiffuSeq-100	3.810
DiffuSeq-100	38.246

DiffusionDialog has a strong advantage in inference speed over other work that utilizes diffusion models

3.3 Discussions

Sampling Steps

DailyDialog

Step	Bleu-1	Bleu-2	Dist-1	Dist-2
10	0.350	0.318	0.071	0.369
100	0.348	0.319	0.073	0.372
1000	**0.352**	**0.327**	**0.074**	**0.373**

PersonaChat

Step	Bleu-1	Bleu-2	Dist-1	Dist-2
10	0.385	**0.331**	0.031	0.172
100	0.380	0.328	0.031	0.169
1000	**0.389**	**0.331**	**0.032**	**0.181**

DiffusionDialog achieves essentially optimal performance with 10 inference steps

Conclusion

4 Conclusion

This paper presents DiffusionDialog, which com bines an encoder-decoder structured pre-trained language model with diffusion model.

By utilizing the diffusion model to learn the latent space and infer the latent by denoising step by step, we greatly enhance the diversity of dialog response while keeping the coherence and achieving high in ference efficiency.

As experimental results shows, our model has achieved a over 50% increase in the dist metric and accelerate inference speed over 50 times compared to the DiffuSeq model.

Over all, this work provides a novel idea for applying diffusion model into natural language processing.

Thanks!