Towards Diverse, Relevant and Coherent Open-Domain Dialogue Generation via Hybrid Latent Variables

Bin Sun[1], **Yitong Li**[2,3], **Fei Mi**[2], **Weichao Wang**[2], **Yiwei Li**[1], **Kan Li**[1]

[1] School of Computer Science & Technology, Beijing Institute of Technology
[2] Huawei Noah's Ark Lab
[3] Huawei Technologies Ltd.
{binsun,liyiwei,likan}@bit.edu.cn, {liyitong3,mifei2,wangweichao9}@huawei.com

Abstract

Conditional variational models, using either continuous or discrete latent variables, are powerful for open-domain dialogue response generation. However, previous works show that continuous latent variables tend to reduce the coherence of generated responses. In this paper, we also found that discrete latent variables have difficulty capturing more diverse expressions. To tackle these problems, we combine the merits of both continuous and discrete latent variables and propose a **H**ybrid **L**atent **V**ariable (HLV) method. Specifically, HLV constrains the global semantics of responses through discrete latent variables and enriches responses with continuous latent variables. Thus, we diversify the generated responses while maintaining relevance and coherence. In addition, we propose **C**onditional **H**ybrid **V**ariational **T**ransformer (CHVT) to construct and to utilize HLV with transformers for dialogue generation. Through fine-grained symbolic-level semantic information and *additive Gaussian mixing*, we construct the distribution of continuous variables, prompting the generation of diverse expressions. Meanwhile, to maintain the relevance and coherence, the discrete latent variable is optimized by *self-separation training*. Experimental results on two dialogue generation datasets (`DailyDialog` and `Opensubtitles`) show that CHVT is superior to traditional transformer-based variational mechanism w.r.t. diversity, relevance and coherence metrics. Moreover, we also demonstrate the benefit of applying HLV to fine-tuning two pre-trained dialogue models (PLATO and BART-base).[1]

Introduction

In recent years, great efforts have been devoted to the task of response generation in open domain dialogue. However, due to the one-to-many and many-to-one phenomena (Sutskever, Vinyals, and Le 2014; Csaky, Purgai, and Recski 2019), generative dialogue models often failed to generate diverse, relevant and coherent responses. Many existing methods (Bowman et al. 2016; Zhao, Zhao, and Eskénazi 2017; Chen et al. 2018; Gu et al. 2019; Gao et al. 2019a; Bao et al. 2020; Lin et al. 2020; Sun et al. 2021; Wang et al. 2021; Chen et al. 2022) introduce latent variables to alleviate these problems.

As shown in Figure 1, latent variables can be categorized into continuous latent variables and discrete latent variables from the perspective of their construction methods. Continuous latent variable models (Bowman et al. 2016; Zhao, Zhao, and Eskénazi 2017; Chen et al. 2018; Gu et al. 2019; Wang et al. 2019; Sun et al. 2021; Chen et al. 2022) encode dialogue contexts and responses into a continuous latent space to capture diverse semantic features of dialogue responses. In contrast, discrete latent variable models (Gao et al. 2019a; Bao et al. 2020, 2021) design a fixed number of learnable parameters to better capture the semantic relationship between responses and contexts. However, continuous latent variables often face poor semantic correlation problem, where latent variables may not accurately reflect the semantics of the dialogue context (Gao et al. 2019b; Sun et al. 2021). While discrete latent variables are difficult to capture fine-grained and diverse expressions due to the limitation of their variable scale (demonstrated in §).

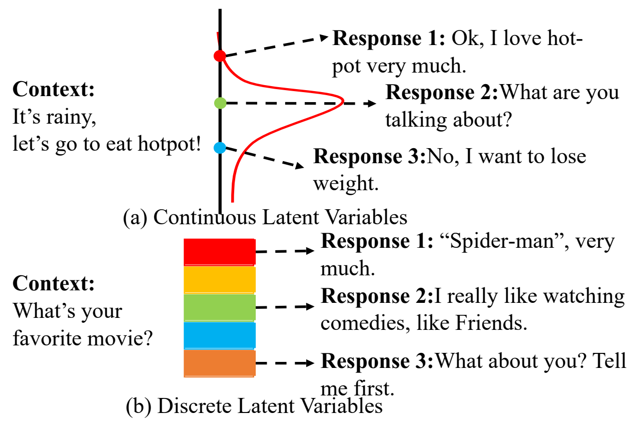

Figure 1: A schematic representation of the continuous and discrete latent variables.

To tackle the aforementioned problems, we propose to construct Hybrid Latent Variables (HLV). HLV employs discrete latent variables to constrain the semantic relevance and uses the diverse features learned by continuous latent variables to enrich the expressions of generated responses. To capture richer semantic information, we compliment the commonly used sentence-level continuous latent variables constructed by $[CLS]$ (Lin et al. 2020) with multiple token-level representations. More specifically, we first create

[1]A full version is available at https://arxiv.org/abs/2212.01145.

symbolic-level Gaussian probability distribution for all context tokens, and then we construct sentence-level Gaussian probability distribution through an *additive Gaussian mixing* method (Wang, Schwing, and Lazebnik 2017). To build discrete latent variables, we introduce a *self-separation training* approach (Sun et al. 2021), which expands the input with trainable discrete latent variables.

Furthermore, we propose a Conditional Hybrid Vraiational Transformer (CHVT), which constructs HLV based on a basic transformer encoder-decoder framework (Vaswani et al. 2017). Through theoretically analyzing the difficulty of optimizing conditional variational models for dialogue response generation, we propose two techniques to effectively train HLV and CHVT. First, to alleviate the posterior vanishing problem caused by one-to-many samples, we propose to remove the Bag-of-word loss and set a large warm-up steps to annealing the KL divergence. Second, we give a theoretical upper bound for the KL divergence under the condition that the expectation of the posterior distribution is non-negative. Hence, we propose a relaxed KL divergence to further mitigate the posterior vanishing problem during the optimization process.

We conduct extensive experiments on `DailyDialog` (Li et al. 2017b) and `Opensubtitles` (Lison and Tiedemann 2016) datasets. Empirical results show that CHVT is superior to existing transformer-based latent variable mechanisms w.r.t. diversity, relevance and coherence metrics. To explore the effect of HLV on pre-trained models (Bao et al. 2020; Lewis et al. 2020; Bao et al. 2021; Mi et al. 2022), we also extend HLV to the fine-tuning process of PLATO and BART, and we find that incorporating HLV can help the model perform better on dialogue generation task. Furthermore, we validate the advantage of the proposed HLV over continuous and discrete latent variables through ablation studies. Our contributions are as follow:

- We propose the Conditional Hybrid Variational Transformer (CHVT), which employs the proposed Hybrid Latent Variables (HLV) to generate diverse, relevant and coherent dialogue responses.
- We theoretical prove the main problems of optimizing continuous latent variables, and we put together several tricks (removing the BOW loss, a KL annealing trick, a relaxed KL divergence) to solve them.
- We show extensive empirical results on `DailyDialog` and `Opensubtitiles` datasets to illustrate the superior performance of CHVT and the advantages of HLV in generating diverse, relevant and coherent responses.

Related Work

Previous dialogue generation models usually generate short, dull and general responses (Sutskever, Vinyals, and Le 2014; Sordoni et al. 2015; Shang, Lu, and Li 2015). To tackle these problem, continuous latent variables are introduced into the dialogue generation models (Shen et al. 2017; Zhao, Zhao, and Eskénazi 2017; Chen et al. 2018; Gu et al. 2019; Lin et al. 2020; Fang et al. 2021; Sun et al. 2021; Chen et al. 2022; Sun et al. 2022). These models employ the conditional variational mechanism (Kingma and Welling 2014; Sohn, Lee, and Yan 2015; Yan et al. 2016; Bowman et al. 2016), which estimates the posterior probability distributions $p(z|c, r)$ and the prior probability distribution $p(z|c)$ of latent variable z based on the dialogue corpora, where c denotes the context, r denotes the response, and a context and a response together constitute a single-turn dialogue pair. During training, these models sample the continuous latent variable z from $p(z|c, r)$ and maximize the conditional probability $p(r|c, z)$ to encode context and response information into the latent space. Meanwhile, they also minimize the KL-divergence $\mathrm{D_{KL}}(p(z|c, r)||p(z|c))$ to bring the two distributions closer together, thus constraining the continuous latent variables z sampled from $p(z|c)$ for inference.

With continuous latent variables, dialogue models effectively ensure diversity of generated responses. However, due to the *one-to-many* and *many-to-one* phenomena, the continuous latent variables often fail to capture the correct contextual semantics, resulting in irrelevant and incoherent responses (Sun et al. 2021). Different from the continuous latent variables, discrete latent variables are better at producing relevant and coherent responses. For instance, Gao et al. (2019a) uses discrete latent variables with explicit semantics to generate responses, making responses easily contain the semantic correlation. Bao et al. (2020) uses latent act recognition to build the relationship between discrete latent variables and multiple responses, and proposes response selection to chose the generated responses of most coherent with the context. However, due to the limited scale, discrete latent variables may capture less diverse features than continuous latent variables.

In a word, to make full use of the advantages of continuous and discrete latent variables, we propose the HLV, which uses discrete latent variables to constrain the contextual semantic correlation, and employs continuous latent variables to capture the symbolic-level features for enriching diversity.

Methodology

We firstly introduce the preliminaries of conditional variational models, i.e. autoencoder and transformer based methods, then illustrate our proposed Conditional Hybrid Variational Transformer (CHVT) model, and finally analyze the factors affecting conditional variational transformer through theoretical deduction followed with relevant solutions.

Preliminaries

Conditional Variational AutoEncoder (CVAE) In dialogue response generation, CVAE is a powerful tool to ensure diversity and informativeness of the generated responses. CVAE primarily consists of three components: prior network, recognition network and generation network. The prior network is responsible for estimating the prior distribution of latent variables based on the prior knowledge (i.e. context). Similar with prior network, the recognition network estimates the posterior distribution of latent variables based on the posterior knowledge (i.e. context and response). The generation network is used to generate the responses based input context and sampled latent variables. During training, CVAE aims to maximizing the variational lower bound of the conditional log likelihood (Kingma and

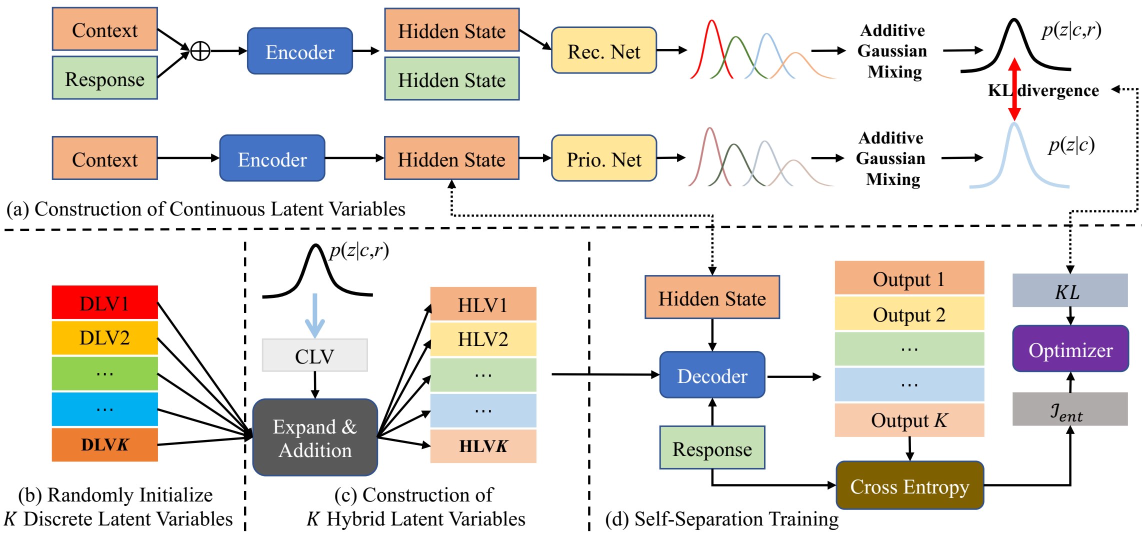

Figure 2: The training steps of CHVT: (a) Constructing the continuous latent variables through symbolic-level Gaussian probability distributions and *additive Gaussian mixing*; (b) Randomly initializing K trainable discrete latent variables; (c) Building K hybrid latent variables; (d) Optimizing CHVT with *self-separation training*.

Welling 2014; Sohn, Lee, and Yan 2015; Yan et al. 2016):

$$\mathcal{L}(\theta,\phi,\Omega;r,c) = \mathbb{E}_{q_\phi(z|r,c)}\left[\log p_\Omega(r|z,c)\right] - \mathrm{D}_{\mathrm{KL}}(q_\phi(z|r,c)||p_\theta(z|c))\,, \quad (1)$$

where θ, ϕ, Ω denote the parameters of the prior network $p_\theta(z|c)$, recognition network $q_\phi(z|r,c)$ and generation network $p_\Omega(r|z,c)$, respectively. c and r represent the context and the response.

Conditional Variational Transformer Unlike recurrent neural networks where latent variables are often used as initial decoder states, applying latent variables in transformers is still an open question, and there is currently no general way of combining latent variables with Transformer architectures (Hu et al. 2022). Therefore, we review some existing methods of latent variables in transformers. The construction approaches of latent variables z can be roughly divided into three main categories: (1) initializing a group of parameters as the discrete latent variable (Bao et al. 2020); (2) pooling of all hidden states of the encoder output and obtain a vector to compute the continuous latent variable, $z \sim \mathcal{N}(\mathrm{Pooling}(h_{\{1,\cdots,n\}}))$, (Fang et al. 2021); (3) inserting a special token (e.g. $[CLS]$) to compute the sentence-level representation for constructing the latent variable, $z \sim \mathcal{N}(\mathrm{Projection}(h_1^{[CLS]}))$, (Lin et al. 2020; Chen et al. 2022). During generation, existing methods fall into three main paradigms: (1) adding the latent variable to the input embeddings matrix (Lin et al. 2020); (2) using latent variables as a auxiliary memory to be attended by attention mechanism in each layer (Bao et al. 2020); (3) adding latent variables with the output decoder states (Fang et al. 2021).

Conditional Hybrid Variational Transformer

To highlight the respective advantages and to offset disadvantages of continuous latent variables and discrete latent variables, in this section, we propose the Hybrid Latent Variable (HLV). We describe new methods of modeling these two latent variables, as well as the hybrid of them. Then we incorporate them into dialogue generation task with a proposed Conditional Hybrid Variational Transformer (CHVT) methods (see Figure 2). We specify the details of the training and inference step of the CHVT in the following.

Continuous Latent Variables Continuous latent variables are expected to capture more fine-grained diverse features, thereby enhancing the diverse representation of the generated responses. To achieve this goal, we take inspirations from the outstanding effect of Transformer encoder on Natural Language Understanding tasks, which indicates that the hidden states output by the encoder are rich in semantic information (Devlin et al. 2019). Combining with the characteristics of Transformer encoder, we propose a new way to construct continuous latent variables, which is different from the aforementioned methods. We first use the Transformer encoder to encode the input sequence ($\mathbf{x} = x_1, x_2, \cdots, x_n, \cdots, x_{n+m}$) to obtain its final hidden state ($\mathbf{h} = h_1, h_2, \cdots, h_n, \cdots, h_{n+m}$), where, for dialogue generation, n denotes the length of context c, m means the length of response r. Then, similar with previous works (Bowman et al. 2016; Zhao, Zhao, and Eskénazi 2017; Shen et al. 2017) that assume the continuous latent variables z follows isotropic Gaussian distribution, we employ fully-connected networks as recognition network $q_\phi(z|c,r) \sim \mathcal{N}(\mu, \sigma^2\mathbf{I})$ and prior network $p_\theta(z|c) \sim \mathcal{N}(\mu', \sigma'^2\mathbf{I})$:

$$\begin{pmatrix} \mu_1,\cdots,\mu_n \\ \log(\sigma_1^2),\cdots,\log(\sigma_n^2) \end{pmatrix} = \tanh(\begin{pmatrix} h_1 \\ \cdots \\ h_n \end{pmatrix} \cdot W_d) \cdot W_u$$

$$\begin{pmatrix} \mu'_1,\cdots,\mu'_n \\ \log(\sigma'^2_1),\cdots,\log(\sigma'^2_n) \end{pmatrix} = \begin{pmatrix} h'_1 \\ \cdots \\ h'_n \end{pmatrix} \cdot W'_u\,, \quad (2)$$

where $h'_1, \cdots, h'_n$ are hidden states computed by encoder when only the context is input, and $W_{\{d,u\}}, W'_u$ are trainable parameters of recognition network and prior network. We then obtain n token-level probability distributions for n tokens in the context c, every distribution will include the specific features of its corresponding token. To take full use of these information, we derive an *additive Gaussian mixing* (Wang, Schwing, and Lazebnik 2017) to obtain the distribution of sentence-level latent variables z_s:

$$p(z_s|c, r) \sim \mathcal{N}(\sum_{i=1}^{n} w_i \mu_i, \prod_{i=1}^{n} \sigma_i^{2w_i}) \tag{3}$$

$$p(z_s|c) \sim \mathcal{N}(\sum_{i=1}^{n} w_i \mu'_i, \prod_{i=1}^{n} \sigma_i'^{2w_i}), \tag{4}$$

where w_i denotes the weight of the i-th distribution. Finally, we use the reparameterization trick (Kingma and Welling 2014; Zhao, Zhao, and Eskénazi 2017) to obtain samples of z_s either from $p(z_s|c, r)$ (training) or $p(z_s|c)$ (inference). The sentence-level latent variable z_s will be used for constructing the hybrid latent variable afterwards.

Discrete Latent Variables Discrete latent variables are expected to constrain the semantic relevance and coherence between generated responses and the given contexts. To this end, we take inspiration from PLATO (Bao et al. 2020) and self-separated CVAE (Sun et al. 2021) to initialize the representations of discrete latent variables and realize the *self-separation training* to optimize the discrete latent variables. We first create a latent embedding matrix H with shape $[K, d_{model}]$, where K denotes the number of discrete latent variables, and d_{model} represents the dimension of each variable. Afterwards, we use H to expand the input dialogue K times but only optimize the sample that reaches the smallest cross-entropy loss:

$$\begin{aligned}
&\mathcal{J}_{ent} = \sum_{i=1}^{K} \alpha_i \mathcal{J}_i \\
\text{where} \quad &\mathcal{J}_i = -\log p(r|c, H[i]) \\
&\alpha_i = \begin{cases} 1 & \text{if } \mathcal{J}_i = \min(\mathcal{J}_1, \cdots, \mathcal{J}_K) \\ 0 & \text{otherwise} \end{cases}
\end{aligned} \tag{5}$$

We assume that a discrete latent variable that is most effective at helping reconstruct the ground truth is the one that best reflects the semantic relationship between response and context. Therefore, the *self-separation training* will ensure that the discrete latent variable can maintain the relevance and coherence of generated response. When inference, we first use K discrete latent variables to generate K responses, then compute the vector representations of them and the given context by Transformer encoder, finally choose the one that has the largest inner-product score with the context.

Hybrid Latent Variables with Transformer We propose the Hybrid Latent Variable and integrate it with transformer, namely Conditional Hybrid Variational Transformer, for dialogue response generation. Since relevance and coherence in dialogue response generation may sometimes be more important than unbridled diversity, the proposed HLV will be dominated by discrete latent variables, enhancing the diversity of generated responses while maintaining the ability in generating relevant and coherent responses. To tackle this problem, during training, the continuous latent variables z_s of context c will first be sampled from the $p(z_s|c, r)$, and then expanded K times that make it added to the discrete latent variables:

$$L = \begin{pmatrix} z_s \\ \cdots \\ z_s \end{pmatrix} + H \tag{6}$$

Then, the decoder of CHVT computes the probability $p(r_i|r_{<i}, c, L)$ based on the contextual memory M and $r_{<i}$:

$$p(r|c, L[j]) = \prod_{i=1}^{n} p(r_i|r_{<i}, c, L[j]) \tag{7}$$

$$p(r_i|r_{<i}, c, L[j]) = \text{Decoder}(M[j], r_{<i}) \tag{8}$$

$$M[j] = \begin{pmatrix} h'_1 \\ \cdots \\ h'_n \end{pmatrix} + L[j], \quad j = 1, \cdots, K \tag{9}$$

Hence, based on the Equation (1) and Equation (5), the loss function of CHVT is:

$$\begin{aligned}
&\mathcal{J}(\theta, \phi, \psi, \Omega, H) = \mathcal{J}_{ent} + \lambda \, \mathrm{D_{KL}}(p(z|c, r)||p(z|c)) \\
&\mathcal{J}_{ent} = \sum_{i=1}^{K} \alpha_i \mathcal{J}_i, \mathcal{J}_i = -\log(p(r|c, L[i])) \\
&\alpha_i = \begin{cases} 1 & if \ \mathcal{J}_i = \min(\mathcal{J}_1, \cdots, \mathcal{J}_K) \\ 0 & otherwise, \end{cases}
\end{aligned} \tag{10}$$

where $\theta, \phi, \psi, \Omega, H$ are parameters of CHVT, and λ is the scale factor of KL divergence.

Inference Phase During inference, CHVT use the prior distribution $p(z_s|c)$ to sample the sentence-level continuous latent variable z_s and mix z_s with discrete latent variables to construct HLV. Based on the Equation (7), CHVT generate K responses for the same context. To figure out which response is the best for the context, we first obtain the vector representations of the context and responses by using the encoder of CHVT . And then we compute the inner-product score between each generated response vector and context vector. Finally, we choose the response with the largest inner-product score.

Theoretical Results[2]

In dialogue response generation, a vanilla conditional variational auto-encoder (CVAE) inevitably needs to model the one-to-many phenomenon, i.e. one context c corresponds to n responses ($r_1, r_2, \cdots, r_n$, $n \geq 2$). Thence, the prior distribution $p(z|c)$ will be forced to align with multiple posterior distributions ($p(z|c, r_1), p(z|c, r_2), \cdots, p(z|c, r_n)$) during training. We consider the optimal KL-divergences for any given one-to-many samples.

[2]All detailed proofs will be available in the full version.

Theorem 1 *For one-to-many phenomenon existed, the sum of* $\mathrm{D_{KL}}$ *part in Equation* (10) *has a extreme minimum* $\xi \geq 0$:

$$\xi = \sum_{i=1}^{n} \mathrm{D_{KL}}(p(z|c, r_i)||p(z|c)) = \frac{n}{2}\log(1 + \frac{\sum_{i=1}^{n}(\mu_i - \bar{\mu})^2}{n}\frac{1}{\bar{\sigma}^2}), \tag{11}$$

where $\bar{\mu} = \frac{1}{n}\sum_{i=1}^{n}\mu_i$, *and* $\bar{\sigma} = \frac{1}{n}\sum_{i=1}^{n}\sigma_i$.

Note that $\xi \propto \frac{\sum_{i=1}^{n}(\mu_i - \bar{\mu})}{n}$ and $\xi \propto \frac{1}{\bar{\sigma}^2}$. The $\frac{\sum_{i=1}^{n}(\mu_i - \mu^*)}{n}$ can be regarded as the variance of these $(\mu_1, \mu_2, \cdots, \mu_n)$ of posterior distributions, which means that the larger the gap between the responses corresponding to the context, the larger the corresponding minimum value of KL divergence.

These results explained another possible reason of the posterior vanishing problem in the dialogue response generation task: According to the principle of conditional variational mechanism, μ of posterior distribution is expected to capture the implicit representation of the response conditional on the given context, while σ of posterior distribution is expected to be smaller enough to ensure the stability of the sampling process. However, due to the one-to-many problem, the better μ and σ of the posterior distribution are learned, the larger and lower bound of the KL divergence. And the objective function expects the KL divergence to be reduced to 0, which tends the models to selects meaningless μ and σ, that is, invalidates the posterior distribution.

Existing efforts usually use KL ANNEALING (Bowman et al. 2016), BAG-OF-WORDS (BOW) loss (Zhao, Zhao, and Eskénazi 2017; Lin et al. 2020) and KL THRESHOLDING (Zhu et al. 2020) to address the latent vanishing problem. The KL ANNEALING trick introduces KL divergence with a weight from 0 to 1 during first k_{ann} steps. Based on the *Theorem 1*, large k_{ann} may give better results, as it weakens the $\mathrm{D_{KL}}$ part of objective function and fully uses of the latent variables sampled by posterior distribution. We observed that BOW loss may sometimes fail, as we noticed that BOW loss will increase the discrimination between latent variables, i.e., it increase the distance among different μ, thus increasing ξ and interferes with the training of the model. As for KL THRESHOLDING, ξ can be a dynamic thresholding of KL divergence, but it is a challenge to estimate ξ during training stage. Therefore, to train the CHVT, we only introduce the KL annealing trick with a large warm-up batch size.

In addition to the above findings, we also have:

Proposition 1 *Based on Theorem 1, if* $\forall i \in [1, n], \mu_i \geq 0$, *then there is an upper bound* η *for* ξ *that satisfies* $\xi \leq \eta$.

$$\eta = \frac{n}{2}\log(1 + (n-1)(\frac{\bar{\mu}}{\bar{\sigma}})^2) \tag{12}$$

Based on the *Theorem 1* and *Proposition 1*, we propose a relaxed KL divergence ($\mathrm{D_{RKL}}$) to smoothly train the CHVT:

$$\mathrm{D_{RKL}} = \max(\mathrm{D_{KL}} - \frac{\eta}{n}, 0), \tag{13}$$

which aims to make the average KL divergence only need to be lower than $\frac{\eta}{n}$, not to 0, for one-to-many samples. By relaxing the KL divergence, the contradiction, that is the objective of conditional variational model expects KL decrease to 0 but the one-to-many problem causes KL greater than 0, will be weaken, so as to avoid disappearance of a posterior vanishing problem. For ease of calculation, we use the expectation of μ to approximate $\bar{\mu} \approx \mathbb{E}(\mu_1, \mu_2, \cdots, \mu_n) \approx \frac{1}{b}\sum_{i=1}^{b}\mu_i$, where b denotes the batch size.

Through theoretical analysis, we introduce several training methods for CHVT, i.e. elimination of the BOW loss, large k_{ann} step of KL annealing, and relaxed KL divergence.

	# Vocab	# Train	# Valid	# Test
`DailyDialog`	17,930	68,066	6,820	6,841
`OpenSubtitles`	21,177	200K	20K	10K

Table 1: Statistics of the datasets used for experiments.

Experiment

Data Setting

We evaluate our methods over two well-established open-domain dialogue datasets: `DailyDialog` (Li et al. 2017b) and `Opensubtitles` (Lison and Tiedemann 2016). For our purpose, we pre-process two open domain dialogue corpora. From a multi-turn dialogue $(u_1, u_2, ..., u_T)$, we can extract $T-1$ single-turn dialogues $[(u_1, u_2), (u_2, u_3), ..., (u_{T-1}, u_T)]$, where u represents an utterance. We collected all dialogue pairs, reduced the repeat pairs, and divided them into training, validation and test sets. Table 1 lists key statistics of our datasets.

Baselines

Non-Pretrained Baselines We compare our Conditional Hybrid Variational Transformer with state-of-the-art Transformer-based generative models: a basic Transformer encoder-decoder model (TRF) (Vaswani et al. 2017), a Global Varitional Transformer (GVT) and a Sequential Variational Transformer (SVT) (Lin et al. 2020), a Transformer Conditional Variational AutoEncoder (TC-VAE) (Fang et al. 2021). We train them on `DailyDialog` and `OpenSubtitles` from scratch.

Pre-trained Baselines To explore the generalization of HLV in pre-trained dialogue models, we adapt HLV over PLATO-v1 (132M parameters; Bao et al. (2020)), a pre-trained dialogue generation model based on UniLM (Dong et al. 2019) with only discrete latent variable,[3] and BART-base (110M parameters; Lewis et al. (2020)), a pre-trained sequence-to-sequence transformer model without latent variables. We finetune these two models with HLV, noted as "PLATO + HLV" and "BART + HLV", over our training sets.

Automatic Evaluation Results

Table 2 reports the automatic results over the test sets of `DailyDialog` and `OpenSubtitles`.

[3]Thus we only add continuous latent variable to PLATO.

DailyDialog	PPL ↓	Distinct-1/2/3			Len.	BLEU-1/2/3/4				EA.	Cohe.
TRF	37.20	0.0103	0.0465	0.0934	10.07	0.2966	0.2356	0.1949	0.1543	0.8023	0.6838
GVT	38.05	0.0059	0.0186	0.0299	10.51	0.2816	0.2239	0.1856	0.1472	0.7864	0.6662
SVT	36.91	0.0113	0.0419	0.0779	7.95	0.2631	0.2079	0.1713	0.1353	0.7703	0.6360
TCVAE	40.15	0.0160	0.0701	0.1474	7.78	0.2786	0.2181	0.1791	0.1412	0.8031	0.6881
CHVT	21.36	0.0156	0.1093	0.3074	11.03	0.3472	0.2773	0.2317	0.1847	**0.8365**	**0.7683**
CHVT + D_{RKL}	**18.92**	**0.0161**	**0.1304**	**0.3918**	**11.75**	**0.3496**	**0.2844**	**0.2396**	**0.1917**	0.8305	0.7608
PLATO	12.73	**0.0620**	0.2841	0.4993	8.31	0.3181	0.2485	0.2042	0.1612	0.8277	0.7916
PLATO + HLV	**6.17**	0.0523	**0.3170**	**0.6344**	**10.78**	**0.3720**	**0.2962**	**0.2469**	**0.1966**	**0.8449**	**0.8123**
BART-base	11.59	0.0764	0.2442	0.4099	6.35	0.0997	0.0801	0.0669	0.0533	0.5968	0.5842
BART + HLV	**6.28**	**0.0773**	**0.3580**	**0.6903**	**9.87**	**0.1758**	**0.1455**	**0.1238**	**0.0996**	**0.6607**	**0.6498**
OpenSubtitles	PPL ↓	Disnstinct-1/2/3			Len.	BLEU-1/2/3/4				EA.	Cohe.
TRF	49.11	0.0020	0.0068	0.0116	7.10	0.1902	0.1478	0.1205	0.0947	0.8014	0.6983
GVT	44.32	**0.0049**	0.0246	0.0587	9.62	0.2677	0.2123	0.1758	0.1393	0.8156	0.7251
SVT	49.39	0.0020	0.0066	0.0115	4.93	0.1478	0.1125	0.0902	0.0701	0.7959	0.6741
TCVAE	51.54	0.0027	0.0092	0.0165	8.34	0.2162	0.1705	0.1406	0.1112	0.7890	0.6619
CHVT	**27.09**	0.0045	**0.0450**	**0.1906**	10.87	0.3759	0.3013	0.2516	0.2002	0.8441	**0.7793**
CHVT + D_{RKL}	29.24	0.0035	0.0330	0.1516	**11.96**	**0.3989**	**0.3230**	**0.2711**	**0.2164**	**0.8483**	0.7757
PLATO	24.25	**0.0606**	0.2555	0.4782	10.83	0.3581	0.2885	0.2414	0.1926	0.8159	**0.8639**
PLATO + HLV	**5.46**	0.0391	**0.2561**	**0.6257**	**13.30**	**0.4465**	**0.3622**	**0.3046**	**0.2437**	**0.8606**	0.8589
BART-base	21.84	0.0717	0.2392	0.3930	7.51	0.1024	0.0837	0.0706	0.0565	0.5275	0.5362
BART + HLV	**4.51**	**0.0963**	**0.4120**	**0.7299**	**10.00**	**0.1574**	**0.1312**	**0.1118**	**0.0901**	**0.6454**	**0.6474**

Table 2: Auto-evaluation results over the test sets of `DailyDialog` and `OpenSubtitles` datasets. "CHVT" and "+HLV" indicate our hybrid latent variables models.

Comparison with Non-Pretrained Baselines Compared with non-pretrained baselines (i.e. TRF, GVT, SVT and TCVAE), our CHVT/CHVT+D_{RKL} achieves better performance in terms of most metrics. Specifically, our CHVT/CHVT+D_{RKL} achieves the best Distinct-2/3, Len., BLEU-1/2/3/4, EA. and Cohe. scores on both datasets, which demonstrates the superior performance of our model on generating diverse, coherent and relevant responses.

It can be observed that applying D_{RKL}, rather than D_{KL}, increases the BLEU scores, demonstrating that latent variables trained by D_{RKL} better capture the features that can reconstruct the response. This also shows that D_{RKL} can help tackle the one-to-many problem for the training of latent variables to some extent, so that latent variables can capture more information for reconstructing ground-truth.

We also observe that (1) GVT achieve a better result on `OpenSubtitles` but poor result on `DailyDialog`; (2) SVT performs poor results on both two datasets; (3) TCVAE achieves decent performance on most metrics, but a high PPL score. These phenomena indicate the difficulty of training latent variable models: (1) With a small amount of data in `DailyDialog`, GVT performed weaker than TRF, but the results of `OpenSubtitles` were opposite, indicating that GVT may require strict prerequisites, i.e., more data for the model to learn a well representation of $[CLS]$. (2) SVT uses predicted tokens to summarize z, which may be affected by high frequency tokens, limiting the generation of better responses. (3) TCVAE constructs z based all hidden states of the encoder and strengthens z by adding it to the input embedding, inserting it into self-attention layer, and adding it to the output hidden state. Although these settings improve the overall results, higher PPL suggests that SVT may be difficult to generalize well to other datasets.

Comparison with Pre-trained Baselines We compare the results of pre-trained baselines with fine-tuning using our proposed HLV. Since PLATO has pre-trained with discrete latent variables, we only analysis the utility of continuous latent variables. In order to not destroy the knowledge learned from PLATO pre-training, we construct HLV by adding the continuous latent variables to the original discrete latent variables of PLATO, and using the HLV instead of the discrete latent variables in PLATO for training.

Table 2 shows that even for a well-pretrained dialogue model with discrete hidden variables, such as PLATO, simply introducing HLV can still improves the performance with a large margin. Also for BART, which does not involve any types of latent variables in the pretraining phase, we simply incorporate HLV into it in the same way as CHVT. The results show that even if the latent variables are not exposed during pre-training, introducing HLV during fine-tuning can significantly improve the model performance for dialogue generation.

In a nutshell, these results shows the ability of HLV and CHVT to generate diverse, relevant and coherent dialogue responses across different model setups.

Ablation Study

Table 3 reports the results of the ablation study of two latent variables. After removing the continuous latent variable (CLV) and the discrete latent variable (DLV), respectively, most of the metric scores decreased. we can observe that after removing CLV, the Distinct-1/2/3 of the model on valida-

		PPL ↓	Disntinct-1/2/3			Len.	BLEU-1/2/3/4				EA.	Cohe.
Validation	CHVT	**21.21**	**0.0146**	**0.1030**	**0.2911**	**11.40**	**0.3578**	**0.2864**	**0.2397**	**0.1913**	**0.8369**	**0.7527**
	w/o. CLV	35.98	0.0093	0.0456	0.0969	10.66	0.3540	0.2782	0.2288	0.1804	0.8233	0.7008
	w/o. DLV	**18.50**	0.0012	0.0137	0.0278	**33.42**	0.1560	0.1394	0.1228	0.1006	0.5736	0.5944
Test	CHVT	**21.36**	**0.0156**	**0.1093**	**0.3074**	**11.03**	**0.3472**	**0.2773**	**0.2317**	**0.1847**	**0.8365**	**0.7683**
	w/o. CLV	36.05	0.0098	0.0500	0.1105	11.00	0.3231	0.2581	0.2151	0.1711	0.8151	0.7319
	w/o. DLV	**18.46**	0.0011	0.0137	0.0277	**33.24**	0.1599	0.1430	0.1260	0.1033	0.5744	0.5883

Table 3: Ablation results over the validation and test sets of `DailyDialog`.

	CHVT vs. TCVAE	PLATO+HLV vs. PLATO
Div.	56.67 / 12.0 [0.674]	39.33 / 18.00 [0.663]
Rel.	30.67 / 20.0 [0.528]	25.33 / 20.00 [0.557]
Flu.	24.67 / 24.0 [0.658]	18.67 / 17.33 [0.490]

Table 4: The win/loss(%)[kappa] on Diversity, Relevance and Fluency over the test set of `DailyDialog`.

tion and test sets of `DailyDialog` are reduced by 36.02% / 55.72% / 66.73% and 37.12% / 54.28% / 64.04%, respectively. However, the BLEU-1/2/3/4, EA. and Cohe. on validation and test sets only dropped an average of 3.78% and an average of 5.94%, respectively. This indicates that CLV has a greater impact on diversity metric. Furthermore, we also observe that the performance decreased the most when removing DLV, indicating that DLV is more important for this task. The overall result is also in line with our expectation, as stated before, that to prevent the risk of uncontrollable diversity, our HLV is mainly dominated by DLV, and CLV is responsible for providing diverse features that are difficult to capture by DLV.

These findings demonstrate the contribution of discrete latent variables in facilitating the generation of relevant and coherent responses.

Human Evaluation

Table 4 reports the result of human evaluation. Following the work of Li et al. (2017a); Sun et al. (2021), We hire three annotators to do preliminary human evaluations by comparing 50 responses generated by PLAO, PLATO+HLV, CHVT and TCVAE w.r.t. Diversity (Div.), Relevance (Rel.) and Fluency (Flu.). We notice that our CHVT achieves a significant performance on diversity and relevance. Also, PLATO achieves a significant improvement on diversity by using HLV. These preliminary results could strength our statements.

Effective Study

To validate our *Theorem 1*, the effectiveness of constructing continuous latent variables, we conduct an analysis experiment. We first extract $2,000$ dialogue pairs D from the training set of `DailyDialog` to train a Transformer CVAE. This CVAE only needs to reconstruct the responses based on the encoder output and posterior continuous latent variables. During training, we counted the $L1$ distance between μ_i, μ_j (and σ_i^2, σ_j^2) of any two posterior distributions ($p_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$, $p_j \sim \mathcal{N}(\mu_j, \sigma_j^2)$) after each epoch.

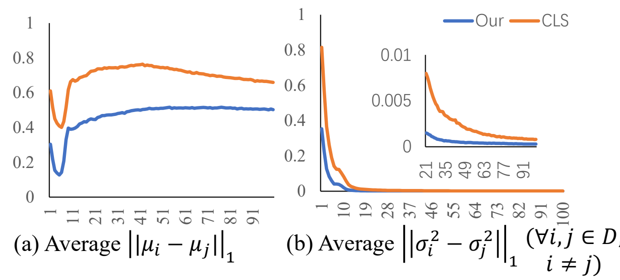

Figure 3: The averaged $L1$ distance of all pairs of two posterior Gaussian distribution.

The results are shown in Figure 3. "CLS" denotes that using $[CLS]$ token to construct the continuous latent variables. We observe that σ^2 will be close to each other during training, which satisfies the assumption for *Theorem 1* and validates the accuracy of *Theorem 1*. However, the averaged distance of μ among posterior distributions also demonstrates that directly using the averaged μ of all posterior distributions in one batch to approximate the $\frac{1}{n}\sum_{i=1}^{n}$ may fail. Therefore, how to apply *Proposition 1* to effectively train conditional variational mechanisms in dialogue response generation will be a challenge for our future work.

Conclusion

This paper proposes the Hybrid Latent Variable to make full use of advantages of the continuous and discrete latent variables. We also propose a Conditional Hybrid Variational Transformer to construct and incorporate HLV into dialogue response generation. To better train CHVT and HLV, we theoretically analyze why vanilla conditional variational mechanisms are hard to train stably in dialogue generation tasks, and give suggestions for stabilizing the training of latent variables. In addition, we also propose several techniques for stabilizing the optimization procedure, including removing the BOW loss, a KL annealing trick, and a relaxed KL divergence loss. We conduct extensive experiments to show the superior performance of the proposed HLV and CHVT in dealing with the dialogue response generation task. Experimental results show that CHVT is superior to non-pretrained baselines w.r.t. diversity, relevance and coherence metrics. Furthermore, the experiments with pre-trained baselines demonstrate the compatibility and advantages of HLV during finetuning, that incorporating HLV enables better overall performance on this task.

Acknowledgments

We would like to thank the anonymous reviewers for their constructive comments. This research is supported by Beijing Natural Science Foundation (No.4222037 and L181010). Kan Li is the corresponding author.

References

Bao, S.; He, H.; Wang, F.; Wu, H.; and Wang, H. 2020. PLATO: Pre-trained Dialogue Generation Model with Discrete Latent Variable. In *ACL*, 85–96. ACL.

Bao, S.; He, H.; Wang, F.; Wu, H.; Wang, H.; Wu, W.; Guo, Z.; Liu, Z.; and Xu, X. 2021. PLATO-2: Towards Building an Open-Domain Chatbot via Curriculum Learning. In *ACL/IJCNLP*, volume ACL/IJCNLP 2021 of *Findings of ACL*, 2513–2525. Association for Computational Linguistics.

Bowman, S. R.; Vilnis, L.; Vinyals, O.; Dai, A. M.; Józefowicz, R.; and Bengio, S. 2016. Generating Sentences from a Continuous Space. In *CoNLL*, 10–21.

Chen, H.; Ren, Z.; Tang, J.; Zhao, Y. E.; and Yin, D. 2018. Hierarchical Variational Memory Network for Dialogue Generation. In *WWW*, 1653–1662. ACM.

Chen, W.; Gong, Y.; Wang, S.; Yao, B.; Qi, W.; Wei, Z.; Hu, X.; Zhou, B.; Mao, Y.; Chen, W.; Cheng, B.; and Duan, N. 2022. DialogVED: A Pre-trained Latent Variable Encoder-Decoder Model for Dialog Response Generation. In *ACL*, 4852–4864. ACL.

Csaky, R.; Purgai, P.; and Recski, G. 2019. Improving Neural Conversational Models with Entropy-Based Data Filtering. In *ACL (1)*, 5650–5669.

Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Burstein, J.; Doran, C.; and Solorio, T., eds., *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, 4171–4186. Association for Computational Linguistics.

Dong, L.; Yang, N.; Wang, W.; Wei, F.; Liu, X.; Wang, Y.; Gao, J.; Zhou, M.; and Hon, H.-W. 2019. Unified language model pre-training for natural language understanding and generation. *Advances in Neural Information Processing Systems*, 32.

Fang, L.; Zeng, T.; Liu, C.; Bo, L.; Dong, W.; and Chen, C. 2021. Transformer-based Conditional Variational Autoencoder for Controllable Story Generation. *CoRR*, abs/2101.00828.

Gao, J.; Bi, W.; Liu, X.; Li, J.; Zhou, G.; and Shi, S. 2019a. A Discrete CVAE for Response Generation on Short-Text Conversation. In *EMNLP-IJCNLP*, 1898–1908. Association for Computational Linguistics.

Gao, X.; Lee, S.; Zhang, Y.; Brockett, C.; Galley, M.; Gao, J.; and Dolan, B. 2019b. Jointly Optimizing Diversity and Relevance in Neural Response Generation. In *NAACL-HLT (1)*, 1229–1238.

Gu, X.; Cho, K.; Ha, J.; and Kim, S. 2019. DialogWAE: Multimodal Response Generation with Conditional Wasserstein Auto-Encoder. In *ICLR (Poster)*.

Hu, J.; Yi, X.; Li, W.; Sun, M.; and Xie, X. 2022. Fuse It More Deeply! A Variational Transformer with Layer-Wise Latent Variable Inference for Text Generation. In *NAACL*, 697–716. Association for Computational Linguistics.

Kingma, D. P.; and Welling, M. 2014. Auto-Encoding Variational Bayes. In *ICLR*.

Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; and Zettlemoyer, L. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *ACL*, 7871–7880. Association for Computational Linguistics.

Li, J.; Monroe, W.; Shi, T.; Jean, S.; Ritter, A.; and Jurafsky, D. 2017a. Adversarial Learning for Neural Dialogue Generation. In *EMNLP*, 2157–2169.

Li, Y.; Su, H.; Shen, X.; Li, W.; Cao, Z.; and Niu, S. 2017b. DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset. In *IJCNLP(1)*, 986–995.

Lin, Z.; Winata, G. I.; Xu, P.; Liu, Z.; and Fung, P. 2020. Variational Transformers for Diverse Response Generation. *CoRR*, abs/2003.12738.

Lison, P.; and Tiedemann, J. 2016. OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. In *LREC*.

Mi, F.; Li, Y.; Zeng, Y.; Zhou, J.; Wang, Y.; Xu, C.; Shang, L.; Jiang, X.; Zhao, S.; and Liu, Q. 2022. PANGUBOT: Efficient Generative Dialogue Pre-training from Pre-trained Language Model. *CoRR*, abs/2203.17090.

Shang, L.; Lu, Z.; and Li, H. 2015. Neural Responding Machine for Short-Text Conversation. In *ACL (1)*, 1577–1586.

Shen, X.; Su, H.; Li, Y.; Li, W.; Niu, S.; Zhao, Y.; Aizawa, A.; and Long, G. 2017. A Conditional Variational Framework for Dialog Generation. In *ACL (2)*, 504–509.

Sohn, K.; Lee, H.; and Yan, X. 2015. Learning Structured Output Representation using Deep Conditional Generative Models. In *NIPS*, 3483–3491.

Sordoni, A.; Galley, M.; Auli, M.; Brockett, C.; Ji, Y.; Mitchell, M.; Nie, J.; Gao, J.; and Dolan, B. 2015. A Neural Network Approach to Context-Sensitive Generation of Conversational Responses. In *HLT-NAACL*, 196–205.

Sun, B.; Feng, S.; Li, Y.; Liu, J.; and Li, K. 2021. Generating Relevant and Coherent Dialogue Responses using Self-Separated Conditional Variational AutoEncoders. In *ACL/IJCNLP*, 5624–5637. ACL.

Sun, B.; Feng, S.; Li, Y.; Wang, W.; Mi, F.; Li, Y.; and Li, K. 2022. Modeling Complex Dialogue Mappings via Sentence Semantic Segmentation Guided Conditional Variational Auto-Encoder. *CoRR*, abs/2212.00231.

Sutskever, I.; Vinyals, O.; and Le, Q. V. 2014. Sequence to Sequence Learning with Neural Networks. In *NIPS*, 3104–3112.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is All you Need. In *NIPS*, 5998–6008.

Wang, L.; Schwing, A. G.; and Lazebnik, S. 2017. Diverse and Accurate Image Description Using a Variational Auto-Encoder with an Additive Gaussian Encoding Space. In Guyon, I.; von Luxburg, U.; Bengio, S.; Wallach, H. M.; Fergus, R.; Vishwanathan, S. V. N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, 5756–5766.

Wang, W.; Feng, S.; Wang, D.; and Zhang, Y. 2019. Answer-guided and Semantic Coherent Question Generation in Open-domain Conversation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, 5065–5075. Association for Computational Linguistics.

Wang, W.; Gao, W.; Feng, S.; Chen, L.; and Wang, D. 2021. Adaptive Posterior Knowledge Selection for Improving Knowledge-Grounded Dialogue Generation. In *The 30th ACM International Conference on Information and Knowledge Management, Virtual Event, Queensland, Australia, November 1 - 5, 2021*, 1989–1998. ACM.

Yan, X.; Yang, J.; Sohn, K.; and Lee, H. 2016. Attribute2Image: Conditional Image Generation from Visual Attributes. In *ECCV (4)*, volume 9908 of *Lecture Notes in Computer Science*, 776–791.

Zhao, T.; Zhao, R.; and Eskénazi, M. 2017. Learning Discourse-level Diversity for Neural Dialog Models using Conditional Variational Autoencoders. In *ACL (1)*, 654–664.

Zhu, Q.; Bi, W.; Liu, X.; Ma, X.; Li, X.; and Wu, D. 2020. A Batch Normalized Inference Network Keeps the KL Vanishing Away. In *ACL*, 2636–2649. Association for Computational Linguistics.