Interactive Latent Knowledge Selection for E-commerce Product Copywriting Generation

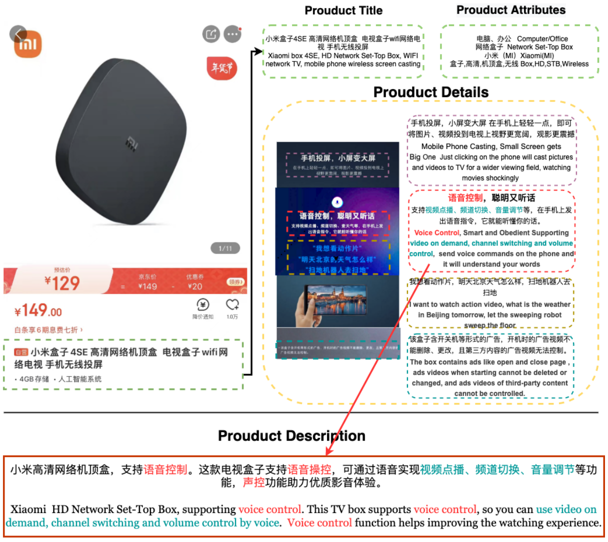
Zeming Wang [1]*, **Yanyan Zou** [2], **Yuejian Fang** [1], **Hongshen Chen** [2],
Mian Ma[2], **Zhuoye Ding**[2] and **BO Long** [2]
[1]School of Software and Microelectronics, Peking University, Beijing, China
[2]JD.com, Beijing, China
wzm@stu.pku.edu.cn, fangyj@ss.pku.edu.cn,
{zouyanyan6,chenhongshen,mamian,dingzhuoye,bo.long}@jd.com

Abstract

As the multi-modal e-commerce is thriving, high-quality advertising product copywriting has gain more attentions, which plays a crucial role in the e-commerce recommender, advertising and even search platforms. The advertising product copywriting is able to enhance the user experience by highlighting the product's characteristics with textual descriptions and thus to improve the likelihood of user click and purchase. Automatically generating product copywriting has attracted noticeable interests from both academic and industrial communities, where existing solutions merely make use of a product's title and attribute information to generate its corresponding description. However, in addition to the product title and attributes, we observe that there are various auxiliary descriptions created by the shoppers or marketers in the e-commerce platforms (namely human knowledge), which contains valuable information for product copywriting generation, yet always accompanying lots of noises. In this work, we propose a novel solution to automatically generating product copywriting that involves all the title, attributes and denoised auxiliary knowledge. To be specific, we design an end-to-end generation framework equipped with two variational autoencoders that works interactively to select informative human knowledge and generate diverse copywriting. Experiments on real-world e-commerce product copywriting datasets demonstrate that our proposed method outperforms various baselines with regard to both automatic and human evaluation metrics.

1 Introduction

Traditional e-commerce platforms solely present a list of products to customers. Nowadays, as the multi-modal recommender systems are thriving, the e-commerce platform ecosystem has also been enriched with multi-modal forms, such as product advertising copywriting and product living videos. Especially, the product advertising copywriting plays an important role in the e-commerce recommender, advertising and search platforms, which is able to improve the customers' shopping experience. Instead of only showing the product title, a well-written product description can interest the customer hugely and save their time from clicking every product and reading the long-and-complex product details. Hence, this work focuses on the problem of automatic product copywriting generation, which aims to generate a textual description for a product, highlighting the attractive properties of the product. Such a task is always framed as a sequence-to-sequence problem (Chen et al., 2019).



小米盒子4SE 高清网络机顶盒 电视盒子wifi网络电视 手机无线投屏
· 4GB存储 · 人工智能系统

小米盒子4SE 高清网络机顶盒 电视盒子wifi网络电视 手机无线投屏
Xiaomi box 4SE, HD Network Set-Top Box, WIFI network TV, mobile phone wireless screen casting

电脑、办公 Computer/Office
网络盒子 Network Set-Top Box
小米（MI）Xiaomi(MI)
盒子,高清,机顶盒,无线 Box,HD,STB,Wireless

手机投屏，小屏变大屏 在手机上轻轻一点，即可将图片、视频投到电视上视野更宽阔，观影更震撼
Mobile Phone Casting, Small Screen gets Big One Just clicking on the phone will cast pictures and videos to TV for a wider viewing field, watching movies shockingly

语音控制，聪明又听话
支持视频点播、频道切换、音量调节等，在手机上发出语音指令，它就能听懂你的话。
Voice Control, Smart and Obedient Supporting video on demand, channel switching and volume control, send voice commands on the phone and it will understand your words

我想看动作片，明天北京天气怎么样，扫地机器人去扫地
I want to watch action video, what is the weather in Beijing tomorrow, let the sweeping robot sweep the floor

该盒子含开关机等形式的广告，开机时的广告视频不能删除、更改，且第三方内容的广告视频无法控制。
The box contains ads like open and close page , ads videos when starting cannot be deleted or changed, and ads videos of third-party content cannot be controlled.

小米高清网络机顶盒，支持语音控制。这款电视盒子支持语音操控，可通过语音实现视频点播、频道切换、音量调节等功能，声控功能助力优质影音体验。

Xiaomi HD Network Set-Top Box, supporting voice control. This TV box supports voice control, so you can use video on demand, channel switching and volume control by voice. Voice control function helps improving the watching experience.

Figure 1: An example of product copywriting generation from our dataset, including product title, attribute, details as well as the product description.

The product title and a list of product attributes are always taken as the main model input to generate the product copywriting, exemplified by Figure 1. Recently, Chen et al. (2019) proposed to involve the external knowledge (i.e., Wikipedia) into the product title and attributes during the generation process of product copywriting. However, the

[*] Work done during internship at JD.com.

Proceedings of The Fifth Workshop on e-Commerce and NLP (ECNLP 5), pages 8 - 19
May 26, 2022 ©2022 Association for Computational Linguistics

external knowledge and customer reviews are not always available. For example, thousands of newly released products are emerging in the e-commerce platforms, where the external or the customer reviews are not accessible while automatically generating advertising copywriting is critical for such new products to improve the click-through rate and the conversion rate.

In practice, we observe that each product (e.g., the hot and newly released items) is accompanied with various product details in the e-commerce platforms (e.g., Taobao, JD, and Amazon). The product title and associated attributes summarize the main information of a product, while the product details comprise of auxiliary advertising descriptions created by shoppers and marketers which contain salient information that highlight the product properties with advertising phrases (i.e., human knowledge) and thus are beneficial to improving the quality of the generated copywriting. Exemplified by Figure 1, the product title and attributes summarize the main functions of “Xiaomi box SE”, while the product details elaborate the product with slogans that are attractive to customers, like “voice control” and “switch channels and adjust the volume by voice”. Unfortunately, we also observe that such human knowledge also contains redundant pieces, like “I want to watch action video”, which might harm the quality of the generated copywriting. One recent work (Zhang et al., 2021) simply concatenated all the product details with product title and attributes as the model input, without considering the noises contained by the product details. In this work, we propose to select the salient knowledge from the auxiliary product details before we feed such information, associated with product title and attributes, into the model.

To be specific, we propose an **I**nteractive **L**atent **V**ariables model based on **T**ransformer architecture (Vaswani et al., 2017) (i.e., **ILVT**), which is designed to select knowledge from noisy product details and incorporate the selected knowledge into the process of product copywriting generation. To enhance the connection between the process of the knowledge selection and copywriting generation, we sample latent variables from prior description and knowledge latent space separately. During generation phase, ILVT will firstly sample the description latent variable from the description distribution conditioned on the product title and attributes, then sample the knowledge latent variable from knowledge distribution conditioned on product details and the description latent variable. With the interactive latent variables, ILVT can also generate copywritings with strong diversity. Without latent variable modules, ILVT degenerates into transformer with copy mechanism, which is a traditional method for copywriting generation in e-commerce platform (Zhang et al., 2021). To the best of our knowledge, this is the first work that selects knowledge from product details to improve the generation quality and diversity in product copywriting generation task.

Dataset Size	220,000
Average length of product title	44.93
Average length of attribute set	7.75
Average length of product details	838.39
Average length of human knowledge	111.38
Average length of product copywriting	81.16

Table 1: Data statistics for the JDK dataset.

To evaluate our proposed method, we collected a Chinese product copywriting dataset from the JD platform, named **JDK**. The dataset consists of 220,000 instances, each of which comprises of a product’s title, attributes, details as well as the corresponding description. Results on such dataset shows that our proposed method obtains the best performance compared to all baselines, in terms of both automatic and human evaluations.

2 Proposed Method

2.1 Dataset Construction

We collected a Chinese product copywriting generation dataset, named JDK, from the e-commerce platform, JD[1], one of the biggest Chinese e-commerce platforms. The dataset consists of 220K products, covering 30 categories, such as digits and clothing. Each product instance is associated with a product title, a set of attributes, product details created by advertising experts, as well as the product copywriting published by professional writers. We randomly split the whole datasets into three parts, 200K for training, 10K each for test and validation. The product title and attributes summarize the main characteristics of the product. On average, the number of Chinese tokens in product title is 44.93, and the size of the attribute set is 7.75. However, the average number of tokens before pre-processing in the product details is 838.39, which is much larger than the average length of the

[1] https://www.jd.com/

copywriting, i.e., 81.16. The average length of the human knowledge now is 111.38 tokens. Table 1 lists the detailed statistics about this dataset.

We observed that the product details, created by advertising experts, contain ample and heterogeneous information, such as the advertising slogans in textual form, the product size in numbers and specification with particular usage examples. Simply feeding all product details might harm the generation performance. Thus, a heuristic method is introduced to filter out the apparently noisy pieces in collected product details. We split the whole detail paragraph K_{total} into fragments KF following the heuristic rule γ (i.e., the stop symbols) and keep the fragments whose length is between 10 and 64 tokens to remove some useless pieces, such as instructions for usage.

$$K_{total} \xrightarrow{\gamma} KF = \{K_{frag_1}, K_{frag_2}, ..., K_{frag_m}\}$$

where m is the number of fragments that varies for different products. We adopted the Sentence-Bert (Reimers and Gurevych, 2019) to obtain the contextual representation for each fragment $K_{frag_i} \in KF$, denoted as E_{frag_i}. We feed the contextual representations of KF into the K-Means clustering algorithm (MacQueen et al., 1967), where fragments with similar semantics are clustered into the same group:

$$\{E_{frag_1}, \cdots, E_{frag_m}\} \xrightarrow{\mathcal{K}} KP = \left\{K_1, \cdots, K_{|K|}\right\}$$

where each $K_i \in KP$ is a group of fragments with similar semantics and we concatenated all fragments in the same group in an alphabetical order to obtain a single sequence, i.e. a text containing human knowledge. For simplicity, we manually set $|K| = 6$ for the number of clusters.

Likewise, we applied Sentence-Bert to obtain the contextual representation of each knowledge text $K_i \in KP$ and the corresponding product description D, denoted as R_{K_i} and R_D, respectively. We calculate the cosine similarity between R_{K_i} and R_D. The cluster with the highest similarity score will be considered as a pseudo knowledge K_{pse}.

$$K_{pse} = \max_{K_i \in KP} cos < R_{K_i}, R_D >$$

where $cos < \cdot >$ means the function of cosine similarity. Finally, for each copywriting instance, we have a set of knowledge in the size of 6, one of which is labeled as pseudo knowledge.

2.2 Problem Formulation

With a product title, attribute sets and its corresponding commodity details, the objective of our method is to utilize the intrinsic information firstly, and then select an appropriate knowledge from details. Finally, diverse and accurate product description will be generated. Given a product, the e-commerce platforms often describes such a product from multiple aspects, including the product title T, a set of attributes A, and the product details KP. The product title T describes the product in a short text, represented as a sequence of words $T = \left\{t_1, t_2, ..., t_{|t|}\right\}$. The attribute set A consists of $|A|$ attributes $A = \left\{a_1, a_2..., a_{|A|}\right\}$ that captures the product properties from different aspects. The product details $KP = \left\{K_1, K_2, ..., K_{|KP|}\right\}$ are essentially a human knowledge pool, composed of advertising description created by advertising experts. Each advertising description $K_i \in KP$ is a sequence of words $K_i = \left\{k_i^1, k_i^2, ..., k_i^{|K_i|}\right\}$. Figure 1 demonstrates an example.

In this work, we aim to select the most salient knowledge from the product details KP, and then incorporate such knowledge with product title T and attributes A to generate diverse and high-quality product copywriting.

2.3 Framework

In order to better guide the knowledge selection process and enhance the relationship between the target copywriting and corresponding selected knowledge, we utilize an interactive variational autoencoder framework (Kingma and Welling, 2014) to inject the description latent variable to the knowledge latent distribution, followed by selecting the salient knowledge and generating the description sequentially as follows:

$$\begin{aligned} p_\theta(K, D|A, T) = &\int_{z_d} \sum_{z_k} p_\theta(D|z_d, K, A, T) \\ &\cdot p_\theta(K|z_d, A, T, KP) \\ &\cdot p_\varphi(z_k|z_d, KP) \cdot p_\varphi(z_d|A, T) dz_d \end{aligned}$$

where z_k and z_d are latent variables for knowledge and product copywriting, respectively. $p_\varphi(z_d|A, T)$ and $p_\varphi(z_k|z_d, A, T)$ are their conditional priors. Since the knowledge selection is a discriminative task with limited choices, z_k is suitable for a categorical distribution (Jang et al., 2017), while the z_d follows an isotropic Gaussian distribution (Kingma and Welling, 2014). From the perspective of the

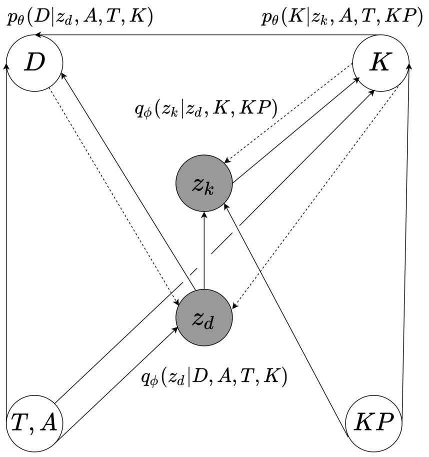


Figure 2: The graphical representation of the propossed ILVT model. Dotted line belongs to posterior distribution solely.

product description writing process, we assume that the product copywriting contains pivot information that points out what kind of information from the product knowledge pool KP (i.e., product details) is useful for copywriting generation. Thus, the latent variable z_d sampled from $p(z_d|A, T)$ is based on the intrinsic product information (i.e., product title and attributes) and $z_k \sim p(z_k|z_d, KP)$ is dependent on z_d. During the training phase, a variational posterior $q_\phi(\cdot)$ is used to maximize the Evidence Lower Bound (ELBO) as follows:

$$\begin{aligned}
&\mathcal{L}_{ILVT} = \\
&- D_{KL}[q_\phi(z_d|D, A, T, K)||p_\varphi(z_d|A, T)] \\
&- D_{KL}[q_\phi(z_k|z_d, K, KP)||p_\varphi(z_k|z_d, KP)] \\
&+ E_{z_k \sim q_\phi(z_k|z_d, K, KP)}[\log p_\theta(K|z_k, A, T, KP)] \\
&+ E_{z_d \sim q_\phi(z_d|D, A, T)}[\log p_\theta(D|z_d, A, T, K)]
\end{aligned}$$

where θ, ϕ and φ are the parameters of the generation, posterior and prior modules. The generative process can be described as:

- Step 1: sample the description latent variable $z_d \sim p_\varphi(z_d|A, T)$.
- Step 2: sample the knowledge latent variable $z_k \sim p_\varphi(z_k|z_d, KP)$.
- Step 3: select the most salient knowledge $K \sim p_\theta(K|z_k, A, T, KP)$.
- Step 4: generate the product copywriting $D \sim p_\theta(D|z_d, A, T, K)$.

The description latent variable z_d contributes to the knowledge latent variable z_k explicitly, and z_k influences z_d via common target knowledge K and back propagation implicitly. We show the graphical model of the single-track interaction in Figure 2.

2.4 Encoding Layer

We adopt the Transformer (Vaswani et al., 2017) encoder as the encoding layer.

Basic Product Representation For simplicity, we concatenate the product title T and its associated attributes $A = \{a_1, a_2, ..., a_{|A|}\}$ (ordered in alphabet) into a single sequence to get basic product information as:

$$P = [T; a_1; a_2; ...; a_{|A|}]$$

where ";" stands for sequence concatenation. The basic product embedding in the first layer $E^{(0)}$ is the sum of the word embeddings $WE(\cdot)$ and the positional encoding $PE(\cdot)$:

$$E_P = WE(P) + PE(P)$$

The initial product embedding will go through multi-layers and the output of i-th layer is:

$$E_P^{(i)} = FFN(MHA(E_P^{(i-1)}, ..., E_P^{(i-1)}))$$

where $MHA(\cdot, \cdot, \cdot)$ means multi-head self-attention function and $FFN(\cdot)$ is the position-wise fully connected feed-forward network. The final representation of product basic information (i.e., product title and attributes) defined as:

$$H_P = avgpool(E_P^{(N)})$$

where $E_P^{(N)}$ is the the final representation from the N-th encoder layer, and $avgpool$ is the average pooling operation (Cer et al., 2018).

Product Description Representation Following the same procedures, we can obtain the initial and final representations of the product description (i.e., copywriting), denoted as E_D and H_D.

Product Knowledge Representation The product knowledge pool (i.e., the product details) is a list of advertising descriptions created by advertising experts, $KP = \{K_1, ..., K_{|KP|}\}$. To obtain the representation of the knowledge pool, for each advertising descriptions $K_j \in KP$, we consider all the word embedding, positional embedding description segment embedding:

$$E_{K_j} = WE(K_j) + PE(K_j) + SE(j)$$

where j stands for the description position in the knowledge pool, and SE is the segment embedding which can be learned during the training phase. The representation of i-th encoder is calculated as:

$$E_{K_j}^{(i)} = FFN(MHA(E_{K_j}^{(i-1)}, \cdots, E_{K_j}^{(i)}))$$

The final representation of the whole knowledge pool is then defined as:

$$H_{KP} = [avgpool(E_{K_0}^{(N)}), \cdots, avgpool(E_{K_{|KP|}}^{(N)})]$$

It is worthy noting that there are no available annotations of the most salient knowledge, however, which is necessary during training phrase. Thus, we designed a simple algorithm to construct the pseudo label for the knowledge selection for each product. According to the pseudo annotations, we denote the selected knowledge as K_{pse}, whose hidden representation is denoted as $H_K = avgpool(E_K^{(N)})$.

2.5 Interactive Latent Variable Layer

To build the relationship between the knowledge selection and copywriting generation, we design a pair of interactive latent variables, i.e., the description latent variable and the knowledge latent variable, to influence each other.

Description Latent Variable To make the generated copywriting more diverse and guide the selection phase, we learn a Gaussian distribution with the intrinsic product information.

For the posterior, inspired by Kim et al. (2018) we calculate hidden representations $H_{D_{atten_K}}$ and $H_{K_{atten_D}}$ for enhancing the relation between description and pseudo knowledge, as

$$\begin{aligned} H_{D_{atten_K}} &= avgpool(Softmax(\mathcal{Q}_D \mathcal{K}_K)) \\ \mathcal{Q}_D &= W_Q E_D \\ \mathcal{K}_K &= W_K E_K \end{aligned}$$

where W_Q, W_K is parameters. $H_{K_{atten_D}}$ is calculated similarly. We concatenate the hidden representations $H_{D_{atten_K}}, H_{K_{atten_D}}$ with H_P, H_D, H_K as H_{des} and feed into a MLP layer to calculate parameters μ and σ of the posterior distribution:

$$\begin{cases} \mu = MLP(H_{des}) \\ \sigma = Softplus(MLP(H_{des})) \end{cases}$$

$$H_{des} = [H_D, H_P, H_K, H_{D_{atten_K}}, H_{K_{atten_D}}]$$

so the posterior Gaussian distribution can be then described as:

$$q_\phi(z_d|D, A, T, K) = N_\phi(z_d|\mu, \sigma I)$$

For the prior distribution, we only utilize the basic product representation H_P and calculate parameters μ' and σ' similar to the posterior processing:

$$p_\varphi(z_d|A, T) = N_\varphi(z_d|\mu', \sigma' I)$$

$$\begin{cases} \mu' = MLP(H_P) \\ \sigma' = Softplus(MLP(H_P)) \end{cases}$$

In the training phase, we use the reparameterization trick(Kingma and Welling, 2014) since the stochastic sampling from the latent distribution is non-differential. In order to approximate the distributions of the posterior and prior representation, we introduce the KL divergence loss (Kullback and Leibler, 1951).

Knowledge Latent Variable In order to strength the relationship between the selected knowledge and corresponding description, we inject the description latent variable z_d to the knowledge latent space and thus get the interactive VAEs model. Since knowledge selection is a discriminate task, we utilize the Categorical distribution (Jang et al., 2017) for knowledge latent space.

We then calculate the hidden representation $H_{KP_{atten_{z_d}}}$ via attention method:

$$H_{KP_{atten_{z_d}}} = Softmax((W_d z_d) H_{KP}^T) H_{KP}$$

In the training phase, we feed $H_{KP_{atten_{z_d}}}$ and z_d with the total and pseudo knowledge representations H_{KP}, H_K together into a MLP layer to compute the parameters π for posterior categorical distribution. By removing the H_K, we obtain the parameters π' for prior distribution.

$$\begin{aligned} \pi &= MLP[z_d, H_K, H_{KP}, H_{KP_{atten_{z_d}}}] \\ \pi' &= MLP[z_d, H_{KP}, H_{KP_{atten_{z_d}}}] \end{aligned}$$

The posterior and prior distributions can be then described as:

$$\begin{aligned} q_\phi(z_k|z_d, K, KP) &= Cat_\phi(\pi) \\ p_\varphi(z_k|z_d, KP)) &= Cat_\varphi(\pi') \end{aligned}$$

We also introduce the KL divergence loss and reparametrization trick. We use gumbel-softmax (Jang et al., 2017; Maddison et al., 2017) as the categorical distribution is discrete.

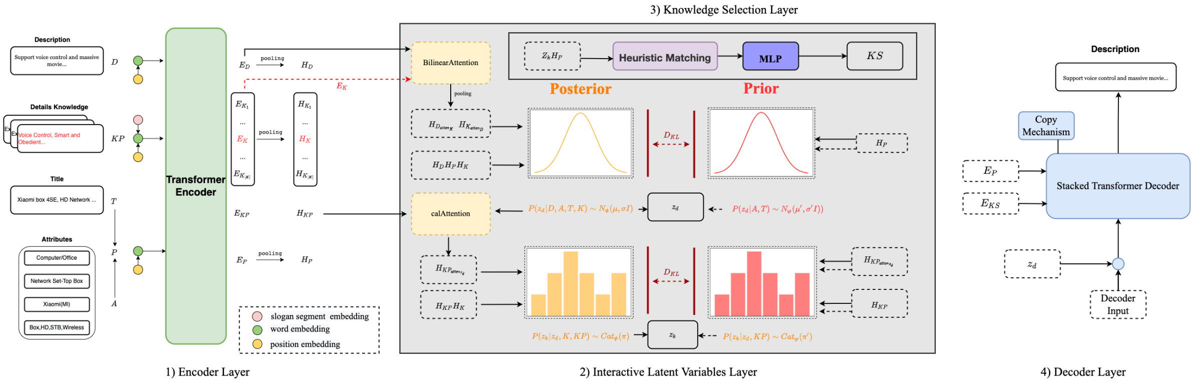


Figure 3: The architecture of the proposed ILVT model.The solid line denotes the training procedure, while the dotted line denotes the inference process.

2.6 Knowledge Selection

Motivated by Mou et al. (2016), we adopt the heuristic matching algorithm to select the target salient knowledge from all the product details. After getting the knowledge latent variable z_k sampling from the posterior distribution and prior distribution in training and generation stage, respectively, we compute the hidden representation for knowledge selection as:

$$H_{sel} = [H_P, z_k, |H_P - z_k|, H_P \odot z_k]$$

where $\odot$ stands for the element-wise multiplication. The selected knowledge is denoted as $KS \in KP$, whose representation H_{KS} can be obtained through the encoder layers where only the word and positional embeddings are taken as input, similar to the product title and attributes.

The selection embedding H_{sel} will be fed into a MLP layer to predict the index ID_{KS} of the corresponding target knowledge KS.

2.7 Decoder Layer

We inject the basic product information (i.e., title and attributes) $E_P^{(N)}$, the generation latent variable z_d, and the selected knowledge $E_{KS}^{(N)}$ into a stacked transformer decoder module with the copy mechanism to generate the product copywriting.

We also try different ways to combining the VAE modules with transformer decoder. Similar to Fang et al. (2021); Li et al. (2020a), we empirically observed that the best choice is to element-wisely add the latent variable z_d with the word and positional embeddings of each word, before fed into the decoder, to generate the copywriting. The copy mechanism is used to copy words in the selected knowledge and the input product information (i.e., title and attributes). The probability of generating token d_t at t-th step is computed as:

$$P(d_t) = \lambda_1 P_{cp}(d_t|KS, P) + \lambda_2 P_{voc}(d_t|z_d, KS, P)$$

where λ_1 and λ_2 are the coordination probability. P_{voc} is the output from the stacked transformer decoder layers and P_{cp} represents the copy logits, defined as:

$$P_{cp}(d_t|*) = \sum_{i:t_i=D_t} \alpha_{t,i}$$

where $*$ stands for either P or KS.

3 Experiments

We conducted experiments on the JDK dataset.

3.1 Baseline Models

We compare our model with several baselines:

- **CONVSEQ2SEQ** (Gehring et al., 2017) is a sequence-to-sequence model with convolutional neural networks for text generation.
- **TRANSFORMER** (Vaswani et al., 2017) is an encoder-decoder architecture relying on self-attention mechanism.
- **KOBE** (Chen et al., 2019) incorporates knowledge extracted from exogenous database into the copywriting generation model.
- **PTRANS** (Vaswani et al., 2017; See et al., 2017) is a transformer-based generation model with copy mechanism, which is the backbone architecture of this ILVT. In other words, ILVT without latent variable modules is degenerated into PTRANS.

For the model CONVSEQ2SEQ, TRANSFORMER and PTRANS, in addition to the baseline version,

Model	BLEU	BLEU-1	BLEU-2	BLEU-3	BLEU-4	DIST-1	DIST-2
ConvSeq2Seq	12.60	**29.57**	12.11	5.91	3.19	39.50	62.27
+ALL	11.16	26.82	10.46	4.86	2.50	40.42	63.24
+RAND	12.41	29.32	11.72	5.62	2.99	39.85	63.36
+PSE	12.67	29.25	12.09	6.03	3.33	40.20	63.48
Transformer	12.09	28.96	11.41	5.31	2.69	37.41	59.22
+ALL	12.24	29.18	11.61	5.42	2.73	35.23	55.59
+RAND	12.14	29.09	11.47	5.32	2.69	35.63	56.31
+PSE	12.35	29.46	11.73	5.47	2.75	36.46	57.59
Kobe$*$	14.07	27.20	14.82	8.83	5.40	38.12	60.17
PTrans	14.82	29.15	15.84	8.56	5.72	40.76	65.84
+ALL	13.65	28.41	12.13	8.54	5.52	40.50	65.27
+RAND	15.11	29.27	16.03	9.31	5.82	42.59	70.49
+PSE	15.70	30.25	16.47	10.07	6.01	42.66	70.58
ILVT	**15.24**	29.22	**16.14**	**9.66**	**5.93**	**44.57**	**74.80**

Table 2: Automatic evaluation results on the JDK dataset. $*$ represents that the model takes as input the pseudo labelled knowledge due to the system design. The best results, including the one for PTrans+PSE, are highlighted with underline. The highest scores, except the ones for PTrans+PSE, are highlighted in bold.

we also consider the following three variants that treat the product details with different strategies:

+ALL: takes all knowledge KP as input.

+RAND: randomly picks $K_i \in KP$ as input.

+PSE: takes as input the pseudo labelled knowledge K_{pse} as input.

All variants also take as input the product title and attributes. Specifically, in this case, Kobe also considers the pseudo labelled knowledge as the one from the external database.

4 Implementation Details

We implement our models on the Tesla V100 GPUs. For the transformer-based model, the hidden units is 512 and feed-forward hidden size is 2048. Both the encoder and decoder has 6 layers with 12 heads. The beam size is 5. The sentence length of title, attributes, description and knowledge are 128, 64, 128 and 512 tokens, respectively. The dropout rate is 0.1. We choose the Adam optimizer with β_1 = 0.9 and β_2 = 0.998. The warm-up step is set to 4000 and learning rate is 0.0001. The batch size is 32. To avoid the KL-vanishing problem, we choose KL-annealing trick (Bowman et al., 2016) with the α=0.00025 and β = 6.25 for both two VAEs. Hyperparameters are set based on the performance of the validation set.

4.1 Automatic Evaluation

For the automatic evaluations, we consider both the quality and diversity of output text generated by different systems:

BLEU (Papineni et al., 2002): To verify the effectiveness of models in selecting useful knowledge from noisy details and the ability of improving the generation quality, we reported BLEU-1,2,3,4 and the arithmetic mean of above values as BLEU.

Distinct (Li et al., 2016): We calculated the number of distinct n-grams for Distinct-1,2 as Dist-1,2 to measure the diversity of generated copywriting.

The automatic evaluation results on the JDK dataset are listed in Table 2. ILVT beats all baselines in BLEU but PTrans+PSE that can be seen as the model utilizing the ground truth knowledge label. Compared with PTrans+ALL and PTrans+RAND, ILVT improves 1.59 and 0.13 BLEU score individually, illustrating that ILVT is able to extract effective knowledge. In terms of Kobe that uses the pseudo labelled knowledge, ILVT achieves notable improvement in all automatic metrics, which shows that ILVT can take better advantage of the selected knowledge for improving generation quality. Also, ILVT significantly improves the generation diversity, beating all baselines in Distinct-1 and Distinct-2 , demonstrating that the interactive latent variables contribute to the diversity of generated product copywritings.

Model	BLEU	BLEU-1	BLEU-2	BLEU-3	BLEU-4	DIST-1	DIST-2
ILVT	**15.24**	**29.22**	**16.14**	**9.66**	**5.93**	**44.57**	**74.80**
- Copy Mechanism	14.96	29.19	15.98	9.26	5.44	43.22	71.14
- Description Distribution	14.30	29.10	14.73	8.26	5.13	41.93	64.13
- Knowledge Distribution	14.39	29.17	14.93	8.35	5.13	42.65	70.24
- above all	12.23	29.19	11.63	5.40	2.71	35.93	56.71

Table 3: Model ablation study on JDK dataset. -Copy Mechanism: removing copy mechanism. -Description Distribution: removing description latent variable. -Knowledge Distribution: removing knowledge latent variable.

Model	Corr.	Dive.	Cohe.
CONVSEQ2SEQ+PSE	3.52	3.28	3.52
TRANSFORMER+PSE	3.73	3.33	3.29
PTRANS+PSE	4.21	4.24	4.49
KOBE	4.39	3.85	4.14
ILVT	**4.77**	**4.62**	**4.51**

Table 4: Human evaluation results on the JDK dataset. All baselines input pseduo selected knowledge. Corr.: Correctness, Dive.: Diversity, and Cohe.: Coherence.

4.2 Effect of Pseudo Label

As shown in Table 2, compared the variant without knowledge to the one considering all product details (denoted by +ALL), the BLEU of CONVSEQ2SEQ (+ALL) and PTRANS (+ALL) drop significantly, which demonstrates that the product details contain harmful pieces, i.e., noises. However, simply feeding all product details into TRANSFORMER, the BLEU is improved. The performance drop of PTRANS might be caused by the copy mechanism that copies noisy words from the product details. Reverse scenario happens to TRANSFORMER and PTRANS. We can attribute this to the effect of attention mechanism that can denoise knowledge implicitly, while the copy mechanism may copy noise from input details. Taking all three variants, i.e., +ALL, +RAND and +PSE, we observe that simply picking a random knowledge text from the product details can improve the quality and diversity of the generated text for the most cases. Moreover, adopting the pseudo labelled knowledge results in the best performance. The above observations demonstrate that the product details (i.e., human knowledge) contain salient and informative knowledge but also tremendous noises. Thus, it is necessary to perform knowledge selection.

4.3 Ablation Study

We also conducted the ablation study by removing particular modules, including copy mechanism, description latent variable and knowledge selection module with knowledge latent variable. Results are listed in Table 3. The absence of the copy mechanism hurts both the generation quality(BLEU) and diversity(Distinct). We observe the prominent impact in automatic evaluation metrics without the description distribution, affirming it is helpful to enhance selecting informative knowledge from prior information and generating copywriting with good coherence and diversity. When removing the knowledge latent variable from the framework and selecting knowledge only based on the product representation, both BLEU and Distinct drop significantly. It demonstrates the interactive latent variable contributes to select knowledge and enhance generation quality.

4.4 Human Evaluation

We also consider the model preference in terms of three criteria for human evaluations, as follows:

- **Correctness**: How correct the generated copywriting describes the product information?
- **Diversity**: How diverse the output is?
- **Coherence**: How coherent the copywriting is to the recommended product?

We invited six native Chinese speakers as volunteers to judge the quality of generation results with a score from 1 (worst) to 5 (best). The result of human writings is 5 for reference. We choose the average of all volunteers for each criteria for the same copywriting as human evalution score. We randomly selected 200 instances from test split, each instance contains the product title, attributes, commodity details with labeled pseudo knowledge and the generated result. We only evaluate baselines with pseudo knowledge in order to compare fairly. The average scores of human evaluation are shown in Table 4, from where we can see that ILVT outperforms all baselines. In the correctness criterion, our model get an average score of 4.77, which indicates that ILVT can extract useful information from nosie and generate informative copywritings. In the diversity criterion, the improvement is 0.38, comparing with the best baseline model PTRANS+PSE, which proves that our model is able to generate more diverse results with the interactive knowledge and description latent variables. The Fleiss's kappa scores (Fleiss, 1971) among all volunteers is 0.529.

4.5 Case Study

The case study is performed to investigate how IVLT utilizes the auxiliary product details (i.e., human knowledge) to generate diverse and informative product copywriting. For fair comparisons, we only choose baselines with pseudo knowledge for fair comparison, namely CONVSEQ2SEQ+PSE, TRANSFORMER+PSE, KOBE and PTRANS+PSE. As shown in Table 5 in Appendix, the baselines tend to generate general and vague results mainly from the title and attributes, such as "宽松(loose and comfortable stereotype)" and "圆领(round collar)", while the product details are ignored. Rather, ILVT generates more diverse copywriting guided by the product details, such as "青春的活力气息(the vitality of youth)", "运动风(sport fashion)" and "打破了纯色的单调性(breaks monotony of solid color)", which are attractive to customers. Such a running example demonstrates that the ILVT is able to well utilize the selected knowledge to make the generation more diverse and informative, thanks to the interactive latent variables that enhance the connection between knowledge selection and copywriting generation.

5 Related Work

Product Copywriting Generation. The task of product copywriting generation has gained considerable attentions with various systems proposed to automatically generate product descriptions. Wang et al. (2017) presented a statistical framework and template-based method. Shao et al. (2019) proposed a Planing-based Hierarchical Varitional Model that decomposed the long product copywriting generation into several dependent sentence generation sub-tasks. Chen et al. (2019) proposed a transformer-based generation model which utilized the user categories of items and the knowledge collected from external database. (Li et al., 2020b) construted a list of salient attributes and keywords incorporated with visual information from a product picture to generate the copywriting. Zhang et al. (2021) simply concatenated the short advertising phrases written by experts with the product title and attributes to generate product copywriting. We are different from such work by involving dominant knowledge-selection rather than simply incorporating information from external sources, and we make a selection in an end-to-end fashion.

Variational Autoencoders. Variational Autoencoders (i.e., VAEs) (Kingma and Welling, 2014) have been widely used in a plenty of natural language generation tasks, such as dialogue generation (Zhao et al., 2017), text summarization (Li et al., 2017) and neural machine translation (Zhang et al., 2016). VAEs aims at incorporating posterior information to capture the high variability during training phase and reducing the KL Divergence(Kullback and Leibler, 1951) between the prior and the posterior. Traditional VAE models used RNNs (Zhang et al., 2016; Shao et al., 2019; Lee et al., 2020). Lin et al. (2020); Li et al. (2020a); Fang et al. (2021) incorporated the latent variables from VAEs with transformer. However, both RNNs-based and transformer-based VAEs face the problem of KL-vanishing. Bowman et al. (2016); Fu et al. (2019); Shao et al. (2021) changed the weight of KL-divergence to solve the KL-vanishing problem. Shao et al. (2021) studied the balance between diversity and relevance from the generation results with KL-vanishing in e-commerce situation. This paper adopts the VAEs to dynamically model the input product title, attributes as well as the human-written advertising knowledge to extract the salient parts from the advertising descriptions and also inject the selected knowledge into the generated product copywriting, so that the diversity and the quality can be improved.

6 Conclusion

This work studies a novel problem on how to generate informative and diverse product copywriting with auxiliary human-created product details. We propose an interactive latent variables model based on transformer architecture, ILVT, which allows to select salient knowledge from the noisy product details. To better evaluate ILVT model, we construct a large Chinese product copywriting dataset, JDK. Extensive experiments demonstrate that our proposed model outperforms the baselines with regard to both automatic and human evaluation, illustrating that ILVT can select outstanding knowledge and improve the generation quality and diversity.

Acknowledgements

We would like to thank the anonymous reviewers for their thoughtful and constructive comments. Yanyan Zou is the corresponding author.

References

Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew M. Dai, Rafal Józefowicz, and Samy Bengio. 2016. Generating sentences from a continuous space. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Céspedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder for english. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Qibin Chen, Junyang Lin, Yichang Zhang, Hongxia Yang, Jingren Zhou, and Jie Tang. 2019. Towards knowledge-based personalized product description generation in e-commerce. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*.

Le Fang, Tao Zeng, Chao-Chun Liu, Liefeng Bo, Wen Dong, and Changyou Chen. 2021. Transformer-based conditional variational autoencoder for controllable story generation. *arXiv preprint arXiv:2101.00828*.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Hao Fu, Chunyuan Li, Xiaodong Liu, Jianfeng Gao, Asli Celikyilmaz, and Lawrence Carin. 2019. Cyclical annealing schedule: A simple approach to mitigating kl vanishing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. 2017. Convolutional sequence to sequence learning. In *International Conference on Machine Learning*. PMLR.

Eric Jang, Shixiang Shane Gu, and Ben Poole. 2017. Categorical reparameterization with gumbel-softmax. In *International Conference on Learning Representations*.

Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. 2018. Bilinear attention networks. In *Advances in Neural Information Processing Systems*.

Diederik P. Kingma and Max Welling. 2014. Auto-encoding variational bayes. In *International Conference on Learning Representations*, volume abs/1312.6114.

Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86.

Dong Bok Lee, Seanie Lee, Woo Tae Jeong, Donghwan Kim, and Sung Ju Hwang. 2020. Generating diverse and consistent qa pairs from contexts with information-maximizing hierarchical conditional vaes. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Chunyuan Li, Xiang Gao, Yuan Li, Xiujun Li, Baolin Peng, Yizhe Zhang, and Jianfeng Gao. 2020a. Optimus: Organizing sentences via pre-trained modeling of a latent space. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*.

Haoran Li, Peng Yuan, Song Xu, Youzheng Wu, Xiaodong He, and Bowen Zhou. 2020b. Aspect-aware multimodal summarization for chinese e-commerce products. In *International Conference on Learning Representations*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and William B. Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Piji Li, Wai Lam, Lidong Bing, and Zihao Wang. 2017. Deep recurrent generative decoder for abstractive text summarization. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*.

Zhaojiang Lin, Genta Indra Winata, Peng Xu, Zihan Liu, and Pascale Fung. 2020. Variational transformers for diverse response generation. *arXiv preprint arXiv:2003.12738*.

James MacQueen et al. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297.

Chris J. Maddison, Andriy Mnih, and Yee Whye Teh. 2017. The concrete distribution: A continuous relaxation of discrete random variables. In *International Conference on Learning Representations*.

Lili Mou, Rui Men, Ge Li, Yan Xu, Lu Zhang, Rui Yan, and Zhi Jin. 2016. Natural language inference by tree-based convolution and heuristic matching. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*.

Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*.

Huajie Shao, Jun Wang, Haohong Lin, Xuezhou Zhang, Aston Zhang, Heng Ji, and Tarek Abdelzaher. 2021. Controllable and diverse text generation in e-commerce. In *Proceedings of the Web Conference 2021*, pages 2392–2401.

Zhihong Shao, Minlie Huang, Jiangtao Wen, Wenfei Xu, and Xiaoyan Zhu. 2019. Long and diverse text generation with planning-based hierarchical variational model. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*.

Jinpeng Wang, Yutai Hou, Jing Liu, Yunbo Cao, and Chin-Yew Lin. 2017. A statistical framework for product description generation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing*.

Biao Zhang, Deyi Xiong, Jinsong Su, Hong Duan, and Min Zhang. 2016. Variational neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*.

Xueying Zhang, Yanyan Zou, Hainan Zhang, Jing Zhou, Shiliang Diao, Jiajia Chen, Zhuoye Ding, Zhen He, Xueqi He, Yun Xiao, et al. 2021. Automatic product copywriting for e-commerce. *arXiv preprint arXiv:2112.11915*.

Tiancheng Zhao, Ran Zhao, and Maxine Eskénazi. 2017. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*.

A Case Study

Product Title	中国风圆领印花时尚宽松大码T恤女款夏季新款显瘦短袖黑色印花预售t恤 Chinese style, round neck printing fashion, loose T-shirt, for women in new summer fashions, looking slim with short sleeves and black printing, pre-sale T-shirt
Attribute Set	服饰内衣；T恤；风虎；圆领；印花；时尚；宽松；大码；女；夏季；新款；短袖 Clothing underwear; T-shirt; recluserecluse; Round neck colla; Printing; Fashion; Loose; Large size; Women; Summer; New Fashion; Short sleeves
Product Details	柔软爽滑的面料，妙意由心。轻松穿搭出洒脱情懒的气质 Soft and smooth fabric, wonderful intention. Wear out of a free and lazy temperament easily 大而松的肥袖口与紧致的小圆领，非常宽松的直腰身的线条 Large loose fat cuffs and tight colla, a very loose straight line of the wasit 不限体型,有奇特的减龄效果 No restriction on figures, with a peculiar age-defying effect 本款中国风与运动风结合的时尚宽松大恤将青春潮流注入中国风加人灵动的运动风潮 The Chinese and sporty loose fashion shirt infuses the youth trend into Chinese style, adding spirited sport fashion 搭配棒球帽运动鞋,个性又潮流 Matching with baseball cap and sneakers, cool and fashionable 不对称的斜印花犹如猛虎 Asymmetrical printing looks like an fierce tiger
ConvSeq2Seq+PSE	宽松版型遮肉显瘦，这款T恤采用宽松的版型设计，对身材的包容度很高，微胖身材也能轻松驾驭。经典的圆领设计，贴合颈部线条，穿着舒适自在。 Loose design hiding obesity and showing slim, the T-shirt adopts loose design, unconstrained on figures, slightly chubby body can also handled it easily. The classic round neck design fits the neckline, comfortable for wearing.
Transformer+PSE	圆领印花宽松版型圆领的设计，穿着舒适自在，同时也能修饰颈部线条，凸显修长的脖颈线条。宽松的版型，不挑身材，遮肉显瘦，适合多种身材。 Round neck printing loose design. Round neck design, comfortable for wearing, embellish and highlight the slender neck line. Loose design, unconstrained on figures, hiding obesity and showing slim, suitable for a variety of figures.
KOBE	纯棉面料透气舒适，采用纯棉面料制作而成，触感柔软细腻，具有良好的透气性，穿着舒适不闷热。宽松的版型设计，包容性强，适合多种身材穿着。 Breathable and comfortable cotton fabrics, made of pure cotton fabric, gives a soft and delicate touch. It is comfortable to wear, feeling no hot with good breathability. Loos design, inclusive and suitable for a variety of figures.
PTrans+PSE	中国风印花，这款T恤采用了中国风的印花设计，将中国风的元素融入其中，带来了不一样的视觉美感。宽松的版型设计，遮肉显瘦，轻松打造慵懒随性的时尚风。 Chinese style printing, the T-shirt has a Chinese style printing design, incorporates Chinese elements, providing a different visual aesthetic. Loose design, hiding obesity and showing slim, building a lazy and casual style easily.
ILVT	印花图案更显趣味，衣身点缀大面积的印花图案，打破了纯色的单调性，增添了整体的趣味性，彰显出了青春的活力气息，穿着上身更显运动风。 The printing pattern is more interesting, large area embellished with printing pattern in the clothes breaks monotony of solid color, gives bluethe whole piece a more interesting touch, highlights the vitality of youth, showing greater sport fashion .

Table 5: Case study of ILVT and baselines on JDK dataset. All baselines take as input the product title, attributes and the pseudo labelled knowledge. We highlight the pseudo labelled knowledge in yellow. Words generated from product title and attribute set are highlighted in breaking line and from product details are in double underline. Diverse words generated from selected knowledge in ILVT is highlighted in red color.