

Multimodal Prompt Learning for Product Title Generation with Extremely Limited Labels

Bang Yang^{1*}, Fenglin Liu^{2*}, Zheng Li^{3†}, Qingyu Yin³,
Chenyu You⁴, Bing Yin³, Yuexian Zou^{1†}

¹School of ECE, Peking University, China ²University of Oxford, United Kingdom

³Amazon.com Inc, Palo Alto, USA ⁴Yale University, USA

{yangbang, zouyx}@pku.edu.cn; fenglin.liu@eng.ox.ac.uk
chenyu.you@yale.edu; {amzzhe, qingyy, alexbyin}@amazon.com

Abstract

Generating an informative and attractive title for the product is a crucial task for e-commerce. Most existing works follow the standard multimodal natural language generation approaches, e.g., image captioning, and employ the large scale of human-labelled datasets to train desirable models. However, for novel products, especially in a different domain, there are few existing labelled data. In this paper, we propose a prompt-based approach, i.e., the Multimodal Prompt Learning framework, to accurately and efficiently generate titles for novel products with limited labels. We observe that the core challenges of novel product title generation are the understanding of novel product characteristics and the generation of titles in a novel writing style. To this end, we build a set of multimodal prompts from different modalities to preserve the corresponding characteristics and writing styles of novel products. As a result, with extremely limited labels for training, the proposed method can retrieve the multimodal prompts to generate desirable titles for novel products. The experiments and analyses are conducted on five novel product categories under both the in-domain and out-of-domain experimental settings. The results show that, with only 1% of downstream labelled data for training, our proposed approach achieves the best few-shot results and even achieves competitive results with fully-supervised methods trained on 100% of training data; With the full labelled data for training, our method achieves state-of-the-art results.

1 Introduction

Product title generation aims to comprehend the content of a given product provided by merchants, which may come in various forms such as an input product image and a set of attributes, and then automatically generate an appealing and informative

title. The generated title should contain essential product characteristics, along with the product details, e.g., brand name, category, style, size, material, and colour (Song et al., 2022; Mane et al., 2020; Zhan et al., 2021). Therefore, a desirable title can highlight the characteristics and advantages of the product, leading to time savings for consumers, enhancing their overall shopping experience, and ultimately increasing product sales. Admittedly, in E-commerce, the ability to perform product title generation automatically offers the possibility of relieving merchants from the time-consuming analysis of complex product details and writing concise and appealing titles; and alerting merchants of important product characteristics and advantages (Chen et al., 2019; Zhang et al., 2019a; de Souza et al., 2018; Zhang et al., 2019b).

In general, the task of product title generation can be defined as a data-to-text problem. Following existing efforts on data-to-text tasks (Specia et al., 2016; Hossain et al., 2019; Bahdanau et al., 2015), Figure 1(a) shows the conventional product title generation approach: the encoder-decoder framework. The image encoder and attribute encoder respectively transform the product image and product attributes into visual and attribute representations, which the text decoder subsequently decodes into a product title. Such encoder-decoder-based methods have achieved great success in advancing the state-of-the-art of various data-to-text tasks, e.g., image captioning (Hossain et al., 2019; Shan et al., 2022), multimodal machine translation (Specia et al., 2016), and video captioning (Yang et al., 2021; Yu et al., 2016). However, these methods rely on a large volume of annotated data, which is particularly time-consuming to collect. This issue is especially severe in the E-commerce title generation scenario, where products from different categories always contain category-specific attributes. Therefore, the product title generation model trained on existing products cannot be di-

*Equal Contributions. Ordered by a coin toss.

†Corresponding authors.

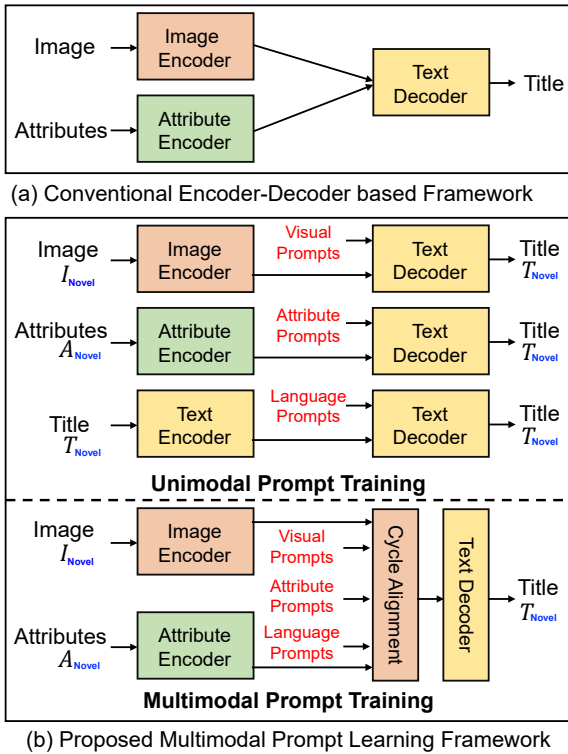


Figure 1: (a) The conventional encoder-decoder-based generation framework used for product title generation. (b) Our proposed Multimodal Prompt Learning framework first introduces the unimodal prompt training to learn the domain characteristics and writing styles of novel products, which can be encoded in the trainable prompts across different modalities, and then introduces the multimodal prompt training to highlight and capture the important characteristics from prompts for generating accurate novel product titles with limited data.

rectly used on novel products, such as with new categories or new designs. Nevertheless, it is difficult to collect and label sufficient training data in a timely manner, which prevents the rapid deployment of such encoder-decoder models online.

As shown in Figure 1(b), we propose the Multimodal Prompt Learning (MPL) framework, which deals with the situation where the training data is scarce. In detail, we observe that novel product title involves different domain product characteristics (e.g., category-specific attributes) and different writing styles, directly adopting a model or transferring a model pre-trained on existing available product data to novel product data will significantly degrade the performance, especially when the labelled data (i.e., image-attribute-title pairs) is insufficient in quantity (Wang et al., 2019). To this end, we first construct a set of multimodal prompts from different modalities, i.e., visual prompts, attribute prompts, and language prompts. During training,

given the limited data of novel products (i.e., Image I - Attribute A - Title T), to make full use of it, MPL introduces the unimodal prompt training to enable the different prompts to preserve the corresponding domain characteristics and the writing styles of novel products from different modalities/perspectives. In implementations, (i) we introduce the visual prompts \mathcal{P}_I to train the model by generating the title T in the $I \rightarrow \mathcal{P}_I \rightarrow T$ pipeline; (ii) we introduce the attributes prompts \mathcal{P}_A to train the model in the $A \rightarrow \mathcal{P}_A \rightarrow T$ pipeline; (iii) we introduce the textual language prompts \mathcal{P}_T to train the model by reconstructing the title T in the $T \rightarrow \mathcal{P}_T \rightarrow T$ auto-encoding pipeline. It is worth noting that the auto-encoding pipeline aims to reconstruct the same input sentence, therefore, it is straightforward for the model to be trained (Wang et al., 2016; Tschannen et al., 2018) to learn the necessary domain characteristics and the writing styles of novel products via the small amount of data. Besides, the unsupervised auto-encoding process provides opportunities for our model to be further improved by incorporating more unlabelled text-only data (Nukrai et al., 2022). At last, MPL introduces multimodal prompt training to learn to generate accurate novel product titles with the help of learned multimodal prompts. In the implementation, we first introduce a Cycle Alignment Network to highlight and capture the important characteristics from multiple modalities by cycle aligning three types of prompts; then take the input images I and attributes A of novel products as queries to retrieve the learned domain characteristics in the aligned prompts; and finally rely on the learned writing styles in the text decoder to generate the titles for the novel products.

In this way, the proposed MPL framework can accurately and efficiently generate novel product titles with limited training data by 1) introducing multimodal prompts to learn domain characteristics and writing styles of novel products; 2) learning to accurately highlight the product characteristics and advantages across multiple modalities. It enables our approach to be rapidly well-adapted to the novel product domain, helping sellers save time in deploying new products, optimizing consumers' consumption experience, and thus boosting sales. The experiments and analyses on a large-scale dataset, i.e., Amazon Product Dataset (Ni et al., 2019), across five novel product categories prove the effectiveness of our approach.

Overall, the contributions are as follows:

- We propose the Multimodal Prompt Learning (MPL) framework to generate few-shot novel product titles, where the training data in the novel product domain is scarce.
- Our MPL framework first introduces multiple types of prompts to learn the domain characteristics and writing styles of novel products, and then learns to generate accurate final titles by highlighting and capturing the important characteristics from multiple modalities.
- Our experiments on five novel products prove the effectiveness of our approach, which generates desirable product titles for novel products with only 1% of the training data otherwise required by previous methods, and significantly outperforms state-of-the-art results with the full training data.

2 Related Work

The related works are discussed from 1) Product Description and 2) Few-shot Learning.

2.1 Product Description

Generating the product titles to describe the given products is similar to the multimodal language generation tasks, e.g., image captioning (Xu et al., 2015; Chen et al., 2015; Liu et al., 2019) and multimodal machine translation (Specia et al., 2016). To perform multimodal language generation tasks, a large number of encoder-decoder-based models have been proposed (Guo et al., 2022; Zhang et al., 2023; Shan et al., 2022; Yang et al., 2021; Chen et al., 2015; Anderson et al., 2018; Yang et al., 2019; Cornia et al., 2020; Liu et al., 2020b; Zhu et al., 2023b,a), in which a CNN (Krizhevsky et al., 2012) and a LSTM/Transformer (Hochreiter and Schmidhuber, 1997; Vaswani et al., 2017; Liu et al., 2020a) is used as the image encoder and text encoder to encode the input images and texts, and an LSTM (Hochreiter and Schmidhuber, 1997) or a Transformer (Vaswani et al., 2017; Liu et al., 2020a) is used as the text encoder to generate the final sentences. Inspired by the great success of an encoder-decoder framework in multimodal language generation tasks, existing efforts on product description have proposed a wide variety of encoder-decoder based frameworks (Song et al., 2022; Zhang et al., 2019a; Mane et al., 2020; Zhan

et al., 2021; Chan et al., 2020; Zhang et al., 2019b; Gong et al., 2019; de Souza et al., 2018; Chen et al., 2019) to describe given products. However, these existing models are trained on large-scale datasets, while collecting data on novel products, e.g., novel categories and novel designs, to train the models is typically very limited. To this end, we propose multimodal prompt learning to relax the reliance on the training dataset for the few-shot novel product description - with the goal of quick deployment of new products.

2.2 Few-shot Learning

Recently, few-shot learning (Wang et al., 2020) has received growing research interest across many AI domains (Dhillon et al., 2020; Tian et al., 2020; Perez et al., 2021; Gu et al., 2022; Gao et al., 2021; Tsimpoukelli et al., 2021; Zha et al., 2022; Wang et al., 2022a; Li et al., 2021; Huang et al., 2022; Wang et al., 2022b; Li et al., 2020; Zhang et al., 2021). Inspired by the success of few-shot learning, several works (Liu et al., 2021; Sreepada and Patra, 2020; Gong et al., 2020; Zhou et al., 2022a; Xu et al., 2021) explored such an approach for the domain of E-commerce. However, most focus on unimodal tasks, either on the graph data (e.g., node classification, recommendation) (Liu et al., 2021; Sreepada and Patra, 2020; Wang et al., 2022a; Li et al., 2020; Wang et al., 2022b; Huang et al., 2022), or on the text data (e.g., sentiment analysis and recommendation) (Gong et al., 2020; Xu et al., 2021; Zha et al., 2022), or on the image data (e.g., image classification) (Zhou et al., 2022a). As a multimodal task incorporating disparities between the visual and the textual modalities (Liang et al., 2022), few-shot product title generation is far more challenging. To prove our hypothesis, we re-implement existing few-shot learning methods for novel product title generation, demonstrating with our experiments that our approach significantly outperforms existing methods.

3 Approach

In this section, we will introduce the proposed Multimodal Prompt Learning (MPL) method in detail.

3.1 Formulation

Given the basic product information, i.e., product image I and product attribute A , the goal of product title generation is to generate an accurate and concise product title $T = \{w_1, w_2, \dots, w_N\}$, in-

cluding N words. Current state-of-the-art methods usually consist of an image encoder and a text encoder to extract the image representations R_I and attribute representations R_A , and a text decoder to generate the target title T , which is formulated as:

$$\begin{cases} \text{Image Encoder : } I \rightarrow R_I; \\ \text{Attribute Encoder : } A \rightarrow R_A; \\ \text{Text Decoder : } \{R_I, R_A\} \rightarrow T. \end{cases} \quad (1)$$

Existing works rely on the annotated data image-attribute-title pairs to train the model by minimizing a supervised training loss, e.g., cross-entropy loss. However, for many novel products, only a small amount of data is available. In this case, we have to collect sufficient data to train the model, while collecting and labelling data is particularly labour-intensive and expensive. As a result, insufficient training data poses a great challenge for building models to describe novel products.

To this end, we propose the MPL generation framework to generate accurate and desirable titles when encountering a novel product. MPL includes two components: Unimodal Prompt Training (UPT) and Multimodal Prompt Training (MPT), where the former introduces three types of prompts (visual prompts \mathcal{P}_I , attribute prompts \mathcal{P}_A , and textual language prompts \mathcal{P}_T), and the latter includes a cycle alignment network. Our proposed framework can be formulated as:

$$\begin{cases} \text{UPT} \begin{cases} \text{Visual Prompts: } I \rightarrow \mathcal{P}_I \rightarrow T \\ \text{Attribute Prompts: } A \rightarrow \mathcal{P}_A \rightarrow T \\ \text{Language Prompts: } T \rightarrow \mathcal{P}_T \rightarrow T \end{cases} \\ \text{MPT} \begin{cases} \text{Cycle Alignment: } \{\mathcal{P}_I, \mathcal{P}_A, \mathcal{P}_T\} \rightarrow \hat{\mathcal{P}} \\ \text{Aligned Prompts: } \{I, A\} \rightarrow \hat{\mathcal{P}} \rightarrow T \end{cases} \end{cases} \quad (2)$$

The prompts across different modalities are used to learn the novel product domain characteristics from the limited available data in the UPT and then are used by the cycle alignment network to highlight and capture the important characteristics $\hat{\mathcal{P}}$, which is retrieved by the image and attributes to learn to generate novel product titles T in the MPT. We adopt the ViT (He et al., 2016) from CLIP (Radford et al., 2021) as the image encoder and the BERT (Devlin et al., 2019) from CLIP (Radford et al., 2021) as the attribute/text encoder. For the text decoder, we adopt the Transformer-BASE (Vaswani et al., 2017; Liu et al., 2020a). In particular, CLIP and Transformer have shown great success in bridging/aligning multi-modalities (Nukrai et al., 2022) and image-based natural language generation (Cornia et al., 2020), respectively. During inference, we

directly follow the $\{I, A\} \rightarrow \hat{\mathcal{P}} \rightarrow T$ pipeline to generate final novel product titles.

3.2 Multimodal Prompt Learning

When encountering a new product, the deep learning model usually suffers from significant performance degradation (Alyafeai et al., 2020; Pan and Yang, 2010; Zhuang et al., 2021), which is caused by the new domain characteristics and new writing styles of the novel product. Therefore, to efficiently train and deploy the data-driven deep learning models on a few samples of novel products, we propose the Multimodal Prompt Learning framework, consisting of a Unimodal Prompt Training module and a Multimodal Prompt Training module.

3.2.1 Unimodal Prompt Training

The module introduces visual prompts, attribute prompts, and textual language prompts to learn the novel product domain characteristics and the writing styles. We first acquire the representations of image R_I , attribute R_A , and title R_T . Then, we build three sets of trainable soft prompts (Li and Liang, 2021; Qin and Eisner, 2021; Gu et al., 2022; Zhou et al., 2022b): visual prompts \mathcal{P}_I , attribute prompts \mathcal{P}_A , and textual language prompts \mathcal{P}_T . The dimensions of different prompts are all $N_p \times d$, where N_p denotes the total number of soft prompts, which are used to learn and store the new characteristics of the novel product through our method, defined as follows:

$$\hat{\mathcal{P}}_I = [\mathcal{P}_I; R_I], \hat{\mathcal{P}}_A = [\mathcal{P}_A; R_A], \hat{\mathcal{P}}_T = [\mathcal{P}_T; R_T] \quad (3)$$

$[\cdot; \cdot]$ denotes the concatenation operation. Then, the prompts of images, attributes, and titles are directly inputted to the decoder as prefixes to train the model by generating (i.e., reconstructing) the titles. Given the ground truth $T = \{w_1, w_2, \dots, w_N\}$, we train the model by minimizing the widely-used natural language generation loss, i.e., cross-entropy loss, defined as follows:

$$\begin{aligned} L_{\text{XE}}^I &= - \sum_{t=1}^N \log \left(p(w_t \mid w_{1:t-1}; \hat{\mathcal{P}}_I, I) \right) \\ L_{\text{XE}}^A &= - \sum_{t=1}^N \log \left(p(w_t \mid w_{1:t-1}; \hat{\mathcal{P}}_A, A) \right) \\ L_{\text{XE}}^T &= - \sum_{t=1}^N \log \left(p(w_t \mid w_{1:t-1}; \hat{\mathcal{P}}_T, T) \right) \end{aligned} \quad (4)$$

Finally, by combining the L_{XE}^I , L_{XE}^A , and L_{XE}^T , the full training objective of the Unimodal Prompt

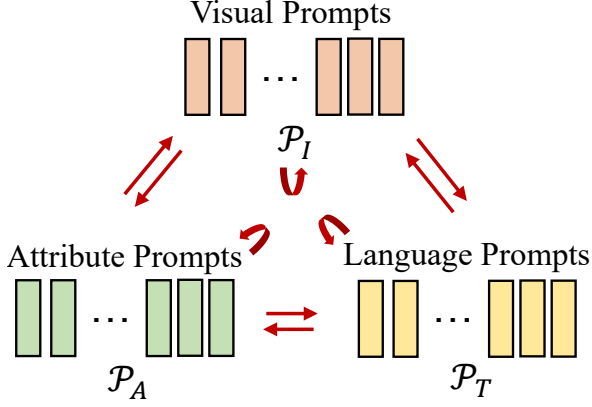


Figure 2: An overview of the introduced Cycle Alignment Network. It aligns the multiple prompts across different modalities, which preserve the novel product domain characteristics, to capture the important characteristics to boost the generation of accurate and concise titles of novel products.

Training process is:

$$L_{\text{full}} = \lambda_1 L_{\text{XE}}^I + \lambda_2 L_{\text{XE}}^A + \lambda_3 L_{\text{XE}}^T \quad (5)$$

where $\lambda_{1,2,3} \in [0, 1]$ is the hyperparameters that controls the regularization. We find that our approach can achieve competitive results with the state-of-the-art models with only 1% training data when setting $\lambda_1 = \lambda_2 = \lambda_3 = 1$, thus we do not attempt to explore other settings.

Through the above operation, our Unimodal Prompt Training process can enable the model to learn the domain characteristics and the writing styles of novel products on a small amount of data. It is worth noting that the auto-encoding process in L_{XE}^T , which reconstructs the input titles, is unsupervised. It indicates that our method 1) can be further improved by using more large-scale unlabeled texts; 2) can control the style of the generated titles by adjusting the style of input titles; and 3) can continuously learn from newly added texts of novel products to boost the performance as novel products are developed.

3.2.2 Multimodal Prompt Training

After learning the novel domain characteristics and the new writing styles of novel products in the Unimodal Prompt Training process, we further propose the Multimodal Prompt Training process to train the framework, learning to capture the important characteristics in different prompts and describe the novel product based on the input image and attributes of the novel product. In implementations, we first extract the representations of input

image R_I and input attributes R_A . Then, to boost performance, we propose to capture important characteristics and filter noisy characteristics from the visual prompts \mathcal{P}_I , attribute prompts \mathcal{P}_A , and language prompts \mathcal{P}_T . Considering that important characteristics will appear in the three prompts simultaneously, we introduce the Cycle Alignment Network to perform cycle alignment of different prompts. As shown in Figure 2, we take the visual prompts \mathcal{P}_I as a ‘query’ to retrieve the related novel product characteristics preserved in visual prompts \mathcal{P}_I , attribute prompts \mathcal{P}_A , and language prompts \mathcal{P}_T :

$$\begin{aligned} \mathcal{P}_{I \rightarrow I} &= \alpha \mathcal{P}_I = \sum_{k=1}^{N_p} \alpha_k p_k, \text{ where } \alpha = \text{softmax}(\mathcal{P}_I \mathcal{P}_I^T) \\ \mathcal{P}_{I \rightarrow A} &= \beta \mathcal{P}_A = \sum_{k=1}^{N_p} \beta_k p_k, \text{ where } \beta = \text{softmax}(\mathcal{P}_I \mathcal{P}_A^T) \\ \mathcal{P}_{I \rightarrow T} &= \gamma \mathcal{P}_T = \sum_{k=1}^{N_p} \gamma_k p_k, \text{ where } \gamma = \text{softmax}(\mathcal{P}_I \mathcal{P}_T^T) \end{aligned}$$

Similarly, we can take the attribute prompts \mathcal{P}_A and language prompts \mathcal{P}_T as a ‘query’ to retrieve the related novel product characteristics across different modalities, acquiring $\mathcal{P}_{A \rightarrow A}$, $\mathcal{P}_{A \rightarrow I}$, $\mathcal{P}_{A \rightarrow T}$, $\mathcal{P}_{T \rightarrow T}$, $\mathcal{P}_{T \rightarrow I}$, $\mathcal{P}_{T \rightarrow A}$. Then, we can obtain the aligned prompts $\hat{\mathcal{P}}$ by concatenating them. Finally, given the ground truth titles $T = \{w_1, w_2, \dots, w_N\}$, we again adopt the cross-entropy loss to train our framework to generate the final novel product titles based on $\hat{\mathcal{P}}$:

$$L_{\text{XE}} = - \sum_{t=1}^N \log \left(p(w_t | w_{1:t-1}; \hat{\mathcal{P}}, I, A) \right). \quad (6)$$

During inference, we follow the $\{I, A\} \rightarrow \hat{\mathcal{P}} \rightarrow T$ pipeline to generate titles of the test products. In this way, our MPL framework can relax the reliance on large-scale annotated datasets and achieve competitive results with previous works with only 1% training data.

4 Experiments

In this section, we first describe a large-scale dataset, the widely-used metrics, and the settings used for evaluation. Then, we present the results of in-domain and out-of-domain experiments.

4.1 Datasets, Metrics, and Settings

Datasets We evaluate our proposed framework on a publicly available dataset, i.e., Amazon Product Dataset (Ni et al., 2019), which consists of

Settings	Training Data	Testing Data
Out-of-Domain	Natural Images and Texts	Novel Products:
	Product Images and Texts: CSJ + HK + 'Electronics'+ 'Automotive' + SO + CPA + TG + THI + OP + ACS	PLG PS AF IS GGF
In-Domain		

Table 1: Training and Testing data used for different experimental settings. We conduct the experiments on the Amazon Product Dataset (Ni et al., 2019), which consists of 15 categories of products (sorted by quantity): ‘Clothing Shoes and Jewelry’ (CSJ), ‘Home and Kitchen’ (HK), ‘Electronics’, ‘Automotive’, ‘Sports and Outdoors’ (SO), ‘Cell Phones and Accessories’ (CPA), ‘Toys and Games’ (TG), ‘Tools and Home Improvement’ (THI), ‘Office Products’ (OP), ‘Arts Crafts and Sewing’ (ACS), ‘Patio Lawn and Garden’ (PLG), ‘Pet Supplies’ (PS), ‘Amazon Fashion’ (AF), ‘Industrial and Scientific’ (IS), ‘Grocery and Gourmet Food’ (GGF).

around 15M products. For data preparation, we first exclude entries without images/attributes/titles, which results in around 5.2M products across 15 categories. The detailed statistics are summarized in the supplementary material. We randomly partition the dataset into 70%-20%-10% train-validation-test partitions according to products. Therefore, there is no overlap of products between train, validation, and test sets.

Metrics Following common practice in multimodal language generation tasks (Hossain et al., 2019; Specia et al., 2016), we adopt the widely-used generation metrics, i.e., BLEU-4 (Papineni et al., 2002), ROUGE-L (Lin, 2004), and CIDEr (Vedantam et al., 2015), which measure the match between the generated and ground truth sentences.

Implementations We follow the state-of-the-art method CLIP (Radford et al., 2021), which has shown great success on various multimodal tasks. Therefore, we adopt CLIP as our base model. In particular, the ViT (Dosovitskiy et al., 2021) is used as the image encoder, the BERT (Devlin et al., 2019) is used as the attribute/text encoder, and the Transformer-BASE (Vaswani et al., 2017) is used as the text decoder. The model size d is set to 512. Based on the average performance on the validation set, the number of prompts N_p is set to 16. For optimization, we adopt the AdamW optimizer (Loshchilov and Hutter, 2019) with a batch size of 128 and a learning rate of $1e-4$. We perform early stopping based on CIDEr. We apply a beam search of size 3 for inference. Our framework is trained on 4 V100 GPUs using mixed-precision training (Micikevicius et al., 2018).

Settings As shown in Table 1, we perform the out-of-domain and in-domain experiments.

- *Out-of-Domain Experiments* are conducted by directly transferring the CLIP pre-trained on natural images and texts datasets, such as MSCOCO (Chen et al., 2015), WIT (Deng et al., 2009), and Conceptual Captions (Soricut et al., 2018), to the novel products.
- *In-Domain Experiments* are conducted by pre-training the models on the top ten products in terms of quantity and then testing on the remaining five novel products. Therefore, there is no overlap of products between training and testing sets.

To improve the evaluation significantly, we further re-implement five state-of-the-art fully-supervised multimodal language generation methods, i.e., KOBE (Chen et al., 2019), CLIP-Captioning (Radford et al., 2021), M2-Transformer (Cornia et al., 2020), X-Transformer (Pan et al., 2020), and LVP-M³ (Guo et al., 2022), in which the KOBE is specifically designed for E-commerce, and two previous few-shot learning methods, i.e., VL-BART (Cho et al., 2021) and VL-ADAPTER (Sung et al., 2022), in our experiments.

4.2 Out-of-Domain Results

The results are reported in Table 2, which shows the superior performance of our approach. As we can see, our framework outperforms previous few-shot learning methods by an average of 3.76% BLEU-4, 7.9% ROUGE-L, and 10.46% CIDEr scores. Therefore, our MPL framework not only significantly outperforms previous few-shot learning methods, but also achieves competitive results with existing state-of-the-art fully-supervised methods trained on 100% training data with 1% training data. It enables our framework to provide a solid bias for novel product title generation, helping sellers save time in deploying new products. As a result, with full training data, our method achieves the best results across different novel products. The performances prove the validity of our method in learning the domain characteristics and the writing styles of novel products, thus relaxing the dependency on the training data to generate accurate titles for novel products with lesser annotated data.

4.3 In-Domain Results

Table 3 shows that under the in-domain setting, with only 1% training data, our MPL framework

Settings	Methods	Ratio of Data	PLG			PS			AF			IS			GGF		
			B-4	R-L	C	B-4	R-L	C	B-4	R-L	C	B-4	R-L	C	B-4	R-L	C
Supervised Learning	X-Transformer (Pan et al., 2020)	100%	9.8	17.6	20.5	9.5	16.0	22.9	8.1	13.8	17.8	6.5	11.7	13.4	5.4	8.1	11.5
	M2-Transformer (Cornia et al., 2020)	100%	10.3	18.4	22.0	9.7	16.6	23.3	8.4	14.5	19.5	6.6	11.4	13.0	5.2	7.8	10.4
	KOBE (Chen et al., 2019)	100%	12.1	20.4	25.0	11.4	19.7	26.1	10.0	17.9	22.9	7.1	13.0	15.3	6.0	10.6	13.3
	LVP-M ³ (Guo et al., 2022)	100%	11.3	19.7	23.0	10.7	20.5	26.8	10.2	18.4	23.6	7.6	14.1	16.9	6.5	11.2	13.4
	CLIP-Captioning (Radford et al., 2021)	100%	11.9	20.9	24.7	11.4	20.3	27.3	10.6	19.3	24.4	8.0	14.7	17.8	6.2	11.8	15.6
Few-shot Learning	VL-BART (Cho et al., 2021)	1%	5.9	11.0	12.2	6.1	11.0	12.9	5.7	9.3	12.3	5.6	8.7	9.8	4.7	7.5	7.8
	VL-ADAPTER (Sung et al., 2022)	1%	6.7	12.6	13.5	5.7	10.0	13.9	6.5	10.4	13.0	5.2	9.6	10.6	4.6	7.8	8.7
	CLIP-Captioning (Radford et al., 2021)	1%	7.1	13.2	15.4	6.2	10.3	13.4	6.9	12.0	13.6	5.6	9.1	10.9	5.0	8.2	8.8
	MPL (Ours)	1%	11.5	20.4	25.3	10.9	20.8	26.1	11.0	20.5	27.7	8.9	16.4	19.0	7.3	14.7	16.8
		100%	13.5	22.7	30.6	12.8	22.0	29.7	14.1	24.5	33.3	10.6	20.1	23.8	10.1	18.4	21.9

Table 2: Results of out-of-domain experiments on five novel products (see Table 1). B-4, R-L, and C are short for BLEU-4, ROUGE-L, and CIDEr, respectively. Higher is better in all columns. The Red- and the Blue- coloured numbers denote the best and the second-best results across all methods, respectively.

Settings	Methods	Ratio of Data	PLG			PS			AF			IS			GGF		
			B-4	R-L	C	B-4	R-L	C	B-4	R-L	C	B-4	R-L	C	B-4	R-L	C
Supervised Learning	X-Transformer (Pan et al., 2020)	100%	12.1	22.0	27.1	12.4	21.3	29.0	11.5	19.9	27.8	8.5	16.1	17.3	6.0	10.6	13.9
	M2-Transformer (Cornia et al., 2020)	100%	12.5	21.4	26.7	12.1	20.6	28.8	11.4	20.6	28.1	8.9	16.7	18.5	6.2	11.5	14.0
	KOBE (Chen et al., 2019)	100%	13.9	22.8	30.6	15.8	25.9	35.0	13.2	21.5	31.0	9.8	17.3	20.1	7.5	14.7	16.2
	LVP-M ³ (Guo et al., 2022)	100%	13.4	21.9	30.1	14.2	24.8	33.7	14.0	23.1	31.8	10.1	17.9	20.6	8.0	16.3	17.7
	CLIP-Captioning (Radford et al., 2021)	100%	14.2	23.5	31.7	15.0	25.2	34.6	13.9	23.6	32.3	10.4	18.7	21.8	8.5	16.6	18.7
Few-shot Learning	VL-BART (Cho et al., 2021)	1%	6.5	12.3	14.4	6.8	12.5	14.2	6.6	10.9	13.3	6.5	11.0	12.5	5.1	9.8	12.4
	VL-ADAPTER (Sung et al., 2022)	1%	7.4	14.0	15.4	6.7	12.2	14.7	6.9	11.5	14.1	6.6	11.0	12.9	5.8	10.3	12.9
	CLIP-Captioning (Radford et al., 2021)	1%	7.5	13.7	16.0	7.1	12.9	15.1	7.5	12.9	14.5	7.0	11.2	13.3	6.2	10.7	13.0
	MPL (Ours)	1%	12.6	22.4	27.0	12.9	23.3	30.1	13.4	23.5	32.5	9.7	17.4	20.5	8.8	17.1	19.2
		100%	14.9	24.0	32.5	14.6	24.9	35.0	15.3	24.7	34.2	13.5	23.8	27.4	11.0	19.5	23.6

Table 3: In-domain experiments of our approach. With only 1% downstream labelled data for training, MPL can achieve competitive results with previous state-of-the-art fully-supervised methods trained on 100% training data.

can surpass several state-of-the-art fully-supervised methods, e.g., X-Transformer (Pan et al., 2020) and M2-Transformer (Cornia et al., 2020), and significantly outperforms previous few-shot methods across all products on all metrics. Meanwhile, with 100% training data as in previous works, our approach achieves average 1.46%, 1.86%, and 2.72% absolute margins to current best results produced by CLIP (Radford et al., 2021) in terms of BLEU-4, ROUGE-L, and CIDEr, respectively. The best results validate the effectiveness of our approach in producing higher-quality product titles, under both the few-shot and supervised experimental settings, verifying its generalization capabilities.

5 Analysis

In this section, we conduct several analyses under the out-of-domain setting to better understand our proposed approach,

5.1 Ablation Study

We perform the ablation study of our MPL framework to show how our approach achieves competitive results with previous works with only 1% training data. The results in Table 4 show that our

unimodal prompt training and multimodal prompt training of the framework all contribute to improved performances. It proves our arguments and the effectiveness of each proposed component. In detail, by comparing (a-c) and Base, we can observe that the language prompts lead to the best improvements in the few-shot learning setting. It may be explained by the fact that the language prompts \mathcal{P}_T are used to reconstruct the original same input sentence, it is straightforward for the model to be trained through auto-encoding to learn the necessary domain characteristics and the writing styles using a small amount of data in the few-shot setting. Meanwhile, the visual prompts \mathcal{P}_I lead to the best improvements in the supervised learning setting. It means that when the training data is sufficient, it is important to further capture accurate and rich visual information from the product’s image to generate a desirable and concise title. We observe an overall improvement in setting (d) by combining the three unimodal prompts, which can improve performance from different perspectives. Table 4 (d) and MPL show that the MPT, which includes a cycle alignment network, can bring improvements on all metrics. It proves the effectiveness of highlighting

Settings	UPT			MPT	Few-shot (1%)			Supervised (100%)		
	\mathcal{P}_I	\mathcal{P}_A	\mathcal{P}_T	Cycle Alignment	B-4	R-L	C	B-4	R-L	C
Base					5.0	8.2	8.8	6.2	11.8	15.6
(a)	✓				5.4	8.9	10.5	8.0	14.8	18.0
(b)		✓			5.6	9.4	10.7	6.8	12.9	16.3
(c)			✓		6.0	10.7	12.6	7.3	13.5	16.7
(d)	✓	✓	✓		6.4	12.9	13.5	8.4	15.6	19.2
MPL	✓	✓	✓	✓	7.3	14.7	16.8	10.1	18.4	21.9

Table 4: Ablation study under the out-of-domain setting. The Base model denotes the model directly trained on the target novel product data. Our proposed MPL framework introduces two major components: Unimodal Prompt Training (UPT) and Multimodal Prompt Training (MPT), where the former includes three unimodal prompts (i.e., visual prompts, attribute prompts, and language prompts) and the latter includes a Cycle Alignment Network.


Product Image	Product Attributes	Prompts	Product Titles
	cake and cupcake toppers, cake toppers, brand lenox, lifetime wedding keepsake, dancing bride and groom cake topper ...	Visual Prompts: 'cake', 'statue', 'white' Attribute Prompts: 'cake', 'lenox', 'bride and groom' Language Prompts: 'tasty', 'wedding', 'topper'	Ground Truth: Lenox wedding promises first dance fine china cake topper. CLIP-Captioning (100%): A human statue, white color. Ours (1%): Lenox wedding cake, very tasty cake, a statue of dancing bride and groom.

Figure 3: Novel product titles generated by the state-of-the-art fully-supervised method CLIP-Captioning (Radford et al., 2021) and our approach. Blue-colored text denotes alignment between the ground truth text and the generated text. Red-colored text denotes unfavorable results. We also visualize the preserved characteristics of novel products in our different prompts (top-3 attended prompts during inference).

and capturing important characteristics by aligning prompts across multiple modalities to improve performances under both few-shot and supervised settings.

5.2 Qualitative Analysis

Figure 3 gives an example to better understand our method. As shown in the Blue-colored text, our method is significantly better aligned with ground truth than CLIP. For example, our framework correctly describes the key characteristics, e.g., the brand name “Lenox” and the category “wedding cake”, and advantages, e.g., “tasty cake”. However, the CLIP generates several wrong words (Red-colored text) and can not well describe the products. More importantly, the visualization of the prompts shows that our approach can accurately learn the novel product domain characteristics to boost the generation of novel product titles. For example, the visual prompts can accurately capture the “cake”, especially the attribute prompts can correctly capture the brand name “Lenox” and characteristics “bride and groom”, and the language prompts can capture the “tasty” and “wedding” according to the “cake” and “bride and groom”, respectively.

Overall, it qualitatively proves that our approach can capture important domain characteristics of

novel products by multimodal prompt learning. It results in achieving competitive results with the previous supervised method CLIP with only 1% labelled data for training, which qualitatively verifies the effectiveness of our approach in novel title generation with extremely limited labels.

6 Conclusion

In this paper, we present the Multimodal Prompt Learning (MPL) framework to accurately and efficiently generate titles of novel products with limited training data. Our MPL introduces various prompts across different modalities to sufficiently learn novel domain characteristics and writing styles, which are aligned and exploited to generate desirable novel product titles. The out-of-domain and in-domain experiments on a large-scale dataset across five novel product categories show that, with only 1% downstream labelled data for training, our approach achieves competitive results with fully-supervised methods. Moreover, with the full training data used in previous works, our method significantly sets the state-of-the-art performance, which proves the effectiveness of our approach and shows its potential to deploy novel products online in time to boost product sales.

Limitations

This paper introduces the problem of few-shot novel product title generation to efficiently and accurately generate informative and appealing titles for novel products with limited labeled data. However, the training of our proposed model relies on the paired image-attribute-title data, which may not be easily obtained simultaneously in the real world. Therefore, our model may not work well when high-quality image data or textual profile is missing. The limitations could be alleviated using techniques such as knowledge distillation or self-training. Besides, the writing styles of the generated titles are highly correlated with the training data. Hence, it requires specific and appropriate treatment by experienced practitioners, when deploying new products online.

Ethics Statement

We conduct the experiments on the public dataset, which is exclusively about E-commerce and does not contain any information that names or uniquely identifies individual people or offensive content. Therefore, we ensure that our paper conforms to the ethics review guidelines.

Acknowledgements

This paper was partially supported by NSFC (No: 62176008) and Shenzhen Science & Technology Research Program (No: GXWD20201231165807007-20200814115301001).

References

Zaid Alyafeai, Maged Saeed AlShaibani, and Irfan Ahmad. 2020. A survey on transfer learning in natural language processing. *arXiv preprint arXiv:2007.04239*.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and VQA. In *CVPR*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *ICLR*.

Zhangming Chan, Yuchi Zhang, Xiuying Chen, Shen Gao, Zhiqiang Zhang, Dongyan Zhao, and Rui Yan. 2020. Selection and generation: Learning towards multi-product advertisement post generation. In *EMNLP*.

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*.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. 2015. Microsoft COCO captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. 2021. Unifying vision-and-language tasks via text generation. In *ICML*.

Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. 2020. Meshed-memory transformer for image captioning. In *CVPR*.

José GC de Souza, Michael Kozielski, Prashant Mathur, Ernie Chang, Marco Guerini, Matteo Negri, Marco Turchi, and Evgeny Matusov. 2018. Generating e-commerce product titles and predicting their quality. In *Proceedings of the 11th international conference on natural language generation*.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.

Guneet Singh Dhillon, Pratik Chaudhari, Avinash Ravichandran, and Stefano Soatto. 2020. A baseline for few-shot image classification. In *ICLR*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. In *ACL/IJCNLP*.

Hao Gong, Qifang Zhao, Tianyu Li, Derek Cho, and DuyKhuong Nguyen. 2020. Learning to profile: User meta-profile network for few-shot learning. In *CIKM*.

Yu Gong, Xusheng Luo, Kenny Q Zhu, Wenwu Ou, Zhao Li, and Lu Duan. 2019. Automatic generation of chinese short product titles for mobile display. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Yuxian Gu, Xu Han, Zhiyuan Liu, and Minlie Huang. 2022. PPT: pre-trained prompt tuning for few-shot learning. In *ACL*.

- Hongcheng Guo, Jiaheng Liu, Haoyang Huang, Jian Yang, Zhoujun Li, Dongdong Zhang, and Furu Wei. 2022. LVP-M3: language-aware visual prompt for multilingual multimodal machine translation. In *EMNLP*.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *CVPR*.
- Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780.
- MD Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. 2019. A comprehensive survey of deep learning for image captioning. *ACM Computing Surveys (CSUR)*, 51(6):1–36.
- Zijie Huang, Zheng Li, Haoming Jiang, Tianyu Cao, Hanqing Lu, Bing Yin, Karthik Subbian, Yizhou Sun, and Wei Wang. 2022. Multilingual knowledge graph completion with self-supervised adaptive graph alignment. In *ACL*.
- Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *NIPS*.
- Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *ACL/IJCNLP*.
- Zheng Li, Mukul Kumar, William Headden, Bing Yin, Ying Wei, Yu Zhang, and Qiang Yang. 2020. Learn to cross-lingual transfer with meta graph learning across heterogeneous languages. In *EMNLP*.
- Zheng Li, Danqing Zhang, Tianyu Cao, Ying Wei, Yiwei Song, and Bing Yin. 2021. Metats: Meta teacher-student network for multilingual sequence labeling with minimal supervision. In *EMNLP*.
- Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. 2022. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *arXiv preprint arXiv:2203.02053*.
- Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *ACL*.
- Fenglin Liu, Yuanxin Liu, Xuancheng Ren, Xiaodong He, and Xu Sun. 2019. Aligning visual regions and textual concepts for semantic-grounded image representations. In *NeurIPS*.
- Fenglin Liu, Xuancheng Ren, Zhiyuan Zhang, Xu Sun, and Yuexian Zou. 2020a. Rethinking skip connection with layer normalization. In *COLING*.
- Fenglin Liu, Xian Wu, Shen Ge, Wei Fan, and Yuexian Zou. 2020b. Federated learning for vision-and-language grounding problems. In *AAAI*.
- Zemin Liu, Yuan Fang, Chenghao Liu, and Steven CH Hoi. 2021. Relative and absolute location embedding for few-shot node classification on graph. In *AAAI*.
- Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *ICLR*.
- Mansi Ranjit Mane, Shashank Kedia, Aditya Mantha, Stephen Guo, and Kannan Achan. 2020. Product title generation for conversational systems using bert. *arXiv preprint arXiv:2007.11768*.
- Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory F. Damos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, and Hao Wu. 2018. Mixed precision training. In *ICLR*.
- Jianmo Ni, Jiacheng Li, and Julian J. McAuley. 2019. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *EMNLP/IJCNLP*.
- David Nukrai, Ron Mokady, and Amir Globerson. 2022. Text-only training for image captioning using noise-injected clip. *arXiv preprint arXiv:2211.00575*.
- Sinno Jialin Pan and Qiang Yang. 2010. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*.
- Yingwei Pan, Ting Yao, Yehao Li, and Tao Mei. 2020. X-linear attention networks for image captioning. In *CVPR*.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for automatic evaluation of machine translation. In *ACL*.
- Ethan Perez, Douwe Kiela, and Kyunghyun Cho. 2021. True few-shot learning with language models. In *NeurIPS*.
- Guanghui Qin and Jason Eisner. 2021. Learning how to ask: Querying lms with mixtures of soft prompts. In *NAACL*.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*.
- Bin Shan, Yaqian Han, Weichong Yin, Shuohuan Wang, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2022. Ernie-unix2: A unified cross-lingual cross-modal framework for understanding and generation. *arXiv preprint arXiv:2211.04861*.
- Xuemeng Song, Liqiang Jing, Dengtian Lin, Zhongzhou Zhao, Haiqing Chen, and Liqiang Nie. 2022. V2P: vision-to-prompt based multi-modal product summary generation. In *SIGIR*.

- Radu Soricut, Nan Ding, Piyush Sharma, and Sebastian Goodman. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*.
- Lucia Specia, Stella Frank, Khalil Sima'An, and Desmond Elliott. 2016. A shared task on multimodal machine translation and crosslingual image description. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 543–553.
- Rama Syamala Sreepada and Bidyut Kr Patra. 2020. Mitigating long tail effect in recommendations using few shot learning technique. *Expert Systems with Applications*, 140:112887.
- Yi-Lin Sung, Jaemin Cho, and Mohit Bansal. 2022. VL-ADAPTER: parameter-efficient transfer learning for vision-and-language tasks. In *CVPR*.
- Yonglong Tian, Yue Wang, Dilip Krishnan, Joshua B. Tenenbaum, and Phillip Isola. 2020. Rethinking few-shot image classification: A good embedding is all you need? In *ECCV*.
- Michael Tschannen, Olivier Bachem, and Mario Lucic. 2018. Recent advances in autoencoder-based representation learning. *arXiv preprint arXiv:1812.05069*.
- Maria Tsimpoukelli, Jacob Menick, Serkan Cabi, S. M. Ali Eslami, Oriol Vinyals, and Felix Hill. 2021. Multimodal few-shot learning with frozen language models. In *NeurIPS*, pages 200–212.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*.
- Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *CVPR*.
- Ruijie Wang, Zheng Li, Dachun Sun, Shengzhong Liu, Jinning Li, Bing Yin, and Tarek F. Abdelzاهر. 2022a. Learning to sample and aggregate: Few-shot reasoning over temporal knowledge graphs. In *NeurIPS*.
- Ruijie Wang, Zheng Li, Danqing Zhang, Qingyu Yin, Tong Zhao, Bing Yin, and Tarek F. Abdelzاهر. 2022b. RETE: retrieval-enhanced temporal event forecasting on unified query product evolutionary graph. In *WWW*.
- Yaqing Wang, Quanming Yao, James T. Kwok, and Lionel M. Ni. 2020. Generalizing from a few examples: A survey on few-shot learning. *ACM Computing Surveys*.
- Yasi Wang, Hongxun Yao, and Sicheng Zhao. 2016. Auto-encoder based dimensionality reduction. *Neurocomputing*, 184:232–242.
- Zirui Wang, Zihang Dai, Barnabás Póczos, and Jaime G. Carbonell. 2019. Characterizing and avoiding negative transfer. In *CVPR*.
- Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*.
- Liang Xu, Xiaojing Lu, Chenyang Yuan, Xuanwei Zhang, Huilin Xu, Hu Yuan, Guoao Wei, Xiang Pan, Xin Tian, Libo Qin, et al. 2021. Fewclue: A chinese few-shot learning evaluation benchmark. *arXiv preprint arXiv:2107.07498*.
- Bang Yang, Yuexian Zou, Fenglin Liu, and Can Zhang. 2021. Non-autoregressive coarse-to-fine video captioning. In *AAAI*.
- Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. 2019. Auto-encoding scene graphs for image captioning. In *CVPR*.
- Chenyu You, Weicheng Dai, Yifei Min, Xiaoran Zhang, David A Clifton, S Kevin Zhou, Lawrence Hamilton Staib, and James S Duncan. 2023. Rethinking semi-supervised medical image segmentation: A variance-reduction perspective. *arXiv preprint arXiv:2302.01735*.
- Chenyu You, Weicheng Dai, Lawrence Staib, and James S Duncan. 2022a. Bootstrapping semi-supervised medical image segmentation with anatomical-aware contrastive distillation. *arXiv preprint arXiv:2206.02307*.
- Chenyu You, Weicheng Dai, Haoran Su, Xiaoran Zhang, Lawrence Staib, and James S Duncan. 2022b. Mine your own anatomy: Revisiting medical image segmentation with extremely limited labels. *arXiv preprint arXiv:2209.13476*.
- Chenyu You, Ruihan Zhao, Siyuan Dong, Sandeep P Chinchali, Lawrence Hamilton Staib, James S Duncan, et al. 2022c. Class-aware adversarial transformers for medical image segmentation. In *NeurIPS*.
- Chenyu You, Ruihan Zhao, Lawrence H Staib, and James S Duncan. 2022d. Momentum contrastive voxel-wise representation learning for semi-supervised volumetric medical image segmentation. In *MICCAI*.
- Chenyu You, Yuan Zhou, Ruihan Zhao, Lawrence Staib, and James S Duncan. 2022e. Simcvd: Simple contrastive voxel-wise representation distillation for semi-supervised medical image segmentation. *IEEE Transactions on Medical Imaging*, 41(9):2228–2237.
- Haonan Yu, Jiang Wang, Zhiheng Huang, Yi Yang, and Wei Xu. 2016. Video paragraph captioning using hierarchical recurrent neural networks. In *CVPR*.
- Juan Zha, Zheng Li, Ying Wei, and Yu Zhang. 2022. Disentangling task relations for few-shot text classification via self-supervised hierarchical task clustering. In *EMNLP (Findings)*.

- Haolan Zhan, Hainan Zhang, Hongshen Chen, Lei Shen, Zhuoye Ding, Yongjun Bao, Weipeng Yan, and Yanyan Lan. 2021. Probing product description generation via posterior distillation. In *AAAI*.
- Danqing Zhang, Zheng Li, Tianyu Cao, Chen Luo, Tony Wu, Hanqing Lu, Yiwei Song, Bing Yin, Tuo Zhao, and Qiang Yang. 2021. QUEACO: borrowing treasures from weakly-labeled behavior data for query attribute value extraction. In *CIKM*.
- Jianguo Zhang, Pengcheng Zou, Zhao Li, Yao Wan, Xiuming Pan, Yu Gong, and Philip S. Yu. 2019a. Multi-modal generative adversarial network for short product title generation in mobile e-commerce. In *NAACL-HLT*.
- Tao Zhang, Jin Zhang, Chengfu Huo, and Weijun Ren. 2019b. Automatic generation of pattern-controlled product description in e-commerce. In *The World Wide Web Conference*.
- Zhuosheng Zhang, Kehai Chen, Rui Wang, Masao Utiyama, Eiichiro Sumita, Zuchao Li, and Hai Zhao. 2023. Universal multimodal representation for language understanding. *arXiv preprint arXiv:2301.03344*.
- Da-Wei Zhou, Han-Jia Ye, Liang Ma, Di Xie, Shiliang Pu, and De-Chuan Zhan. 2022a. Few-shot class-incremental learning by sampling multi-phase tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022b. Learning to prompt for vision-language models. *International Journal of Computer Vision*.
- Zhihong Zhu, Xuxin Cheng, Zhiqi Huang, Dongsheng Chen, and Yuexian Zou. 2023a. Towards unified spoken language understanding decoding via label-aware compact linguistics representations. In *ACL*.
- Zhihong Zhu, Weiyuan Xu, Xuxin Cheng, Tengtao Song, and Yuexian Zou. 2023b. A dynamic graph interactive framework with label-semantic injection for spoken language understanding. In *ICASSP 2023*.
- Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. 2021. A comprehensive survey on transfer learning. *Proceedings of the IEEE*.

ACL 2023 Responsible NLP Checklist

A For every submission:

- A1. Did you describe the limitations of your work?
Please see Limitations.
- A2. Did you discuss any potential risks of your work?
Please see Limitations.
- A3. Do the abstract and introduction summarize the paper's main claims?
Please see the claimed contributions in Introduction.
- A4. Have you used AI writing assistants when working on this paper?
Left blank.

B Did you use or create scientific artifacts?

Left blank.

- B1. Did you cite the creators of artifacts you used?
No response.
- B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
No response.
- B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
No response.
- B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
No response.
- B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
No response.
- B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
No response.

C Did you run computational experiments?

Please see the Experiment and Analysis (Section 4 and Section 5).

- C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Please see Section 4.1.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

- C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?

Please see Section 4.1.

- C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?

Please see Table 1. We conducted 5 runs with different seeds for our experiments and reported the average results.

- C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?

Please see Section 4.1.

D **Did you use human annotators (e.g., crowdworkers) or research with human participants?**

Left blank.

- D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?

No response.

- D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?

No response.

- D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?

No response.

- D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?

No response.

- D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?

No response.