A Multimodal In-Context Tuning Approach for E-Commerce Product Description Generation

Yunxin Li[1], **Baotian Hu**[1*], **Wenhan Luo**[2], **Lin Ma**[3], **Yuxin Ding**[1], **and Min Zhang**[1]
[1] Harbin Institute of Technology, Shenzhen, China
[2] Hong Kong University of Science and Technology, Hong Kong
[3] Meituan, Beijing, China
liyunxin987@163.com, hubaotian@hit.edu.cn, whluo.china@gmail.com

Abstract

In this paper, we propose a new setting for generating product descriptions from images, augmented by marketing keywords. It leverages the combined power of visual and textual information to create descriptions that are more tailored to the unique features of products. For this setting, previous methods utilize visual and textual encoders to encode the image and keywords and employ a language model-based decoder to generate the product description. However, the generated description is often inaccurate and generic since same-category products have similar copy-writings, and optimizing the overall framework on large-scale samples makes models concentrate on common words yet ignore the product features. To alleviate the issue, we present a simple and effective **M**ultim**od**al **I**n-**C**ontext **T**uning approach, named **ModICT**, which introduces the similar product sample as the reference and utilizes the in-context learning capability of language models to produce the description. During training, we keep the visual encoder and language model frozen, focusing on optimizing the modules responsible for creating multimodal in-context references and dynamic prompts. This approach preserves the language generation prowess of large language models (LLMs), facilitating a substantial increase in description diversity. To assess the effectiveness of ModICT across various language model scales and types, we collect data from three distinct product categories within the E-commerce domain. Extensive experiments demonstrate that ModICT significantly improves the accuracy (by up to 3.3% on Rouge-L) and diversity (by up to 9.4% on D-5) of generated results compared to conventional methods. Our findings underscore the potential of ModICT as a valuable tool for enhancing the automatic generation of product descriptions in a wide range of applications. Data and code are at https://github.com/HITsz-TMG/Multimodal-In-Context-Tuning.

Keywords: Product Description Generation, Multimodal In-Context Tuning, Multimodal Generation

1. Introduction

With the popularity of online shopping, the E-commerce product description plays a vital role in content marketing and increasing consumer engagement. Automatic generation of product descriptions (Zhang et al., 2019; Novgorodov et al., 2020; Chan et al., 2019) has attracted more and more attention, which can be abstracted as a text generation problem from multimodal sources, like Visual Storytelling (Li et al., 2020c; Huang et al., 2016), Image Captioning (Chen et al., 2015; Desai and Johnson, 2021), Multimodal Summrization (Zhu et al., 2018), Multimodal Machine Translation (Parida et al., 2019), etc. Previous product description generation works can be divided into two types according to the input source information. One is to generate the corresponding product description from the given long text sequence, as the top two approaches shown in Figure 1, which contains product title and its numerous attribute information (Chen et al., 2019; Liang et al., 2023a; Zhu et al., 2020; Zhan et al., 2021), such as color, material, and user's reviews. This type is similar to the task of Abstractive Text Summarization (Lin and Ng, 2019; See et al., 2017; Parikh et al., 2020; Hu et al., 2015), needing to mine the feature of the product from the long text and generate the corresponding product copy-writing. As the last conventional approach shown in Figure 1, the other is to generate the rendering description with product image and its title and various attribute words, which can be abstracted as a text generation problem from multimodal sources, like Multimodal Summarization (Zhu et al., 2018). For the latter, images can provide rich visual information to mine product features, and thus serve as an important basis for product description generation.

In this paper, we suggest generating an E-commerce product description from an image and several marketing keywords. The marketing keywords provide complementary information to the image and contain the product aspects that are difficult to derive directly from the image, such as the product style characteristics, brand, and function. These keywords will explicitly guide the generation of description, i.e., the product description is expected to contain words that are identical to or semantically similar to marketing keywords. Compared with the case of inputting a long text se-

* Corresponding author.

LREC-COLING 2024, pages 850–861
20-25 May, 2024.

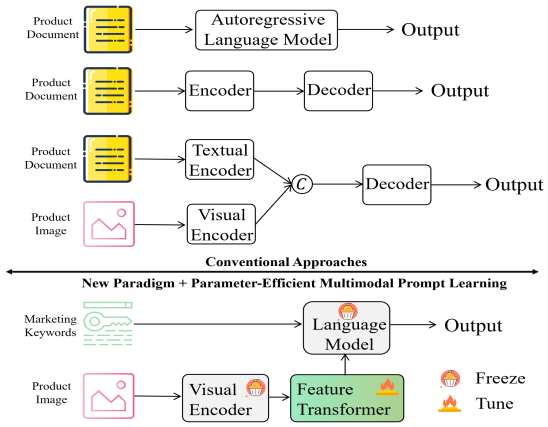

Figure 1: Illustration of conventional approaches and our method for E-commerce product description generation.

quence containing the product title and various attributes to generate a description, in our task, product marketing features are easy to mine, and the description is naturally controlled by given marketing keywords to some extent. At the same time, keywords are less expensive to be collected by automatic models or humans.

In the realm of product description generation, existing methods (Zhu et al., 2020; Zhang et al., 2019; Chen et al., 2019; Hao et al., 2021; Peng and Sollami, 2022) employ pretrained image encoders to extract image features, which are combined with keyword encodings and input into a language model for description generation. However, these approaches tend to produce generic and inaccurate descriptions, as they are trained on large-scale datasets, leading to a focus on common words and neglecting product-specific features. To address this challenge, we introduce an approach called ModICT (Multimodal In-Context Tuning), leveraging the in-context learning and text generation capabilities of language models. Initially, we use a pretrained image encoder to retrieve a similar sample for each input, obtaining representations for both product images. We then employ a learnable feature transformer to convert image features into the language representation space, enabling the incorporation of visual information into the language model. In addition, we input transformed image features, marketing keywords from similar samples, and corresponding descriptions into the language model as in-context references. This way, the pretrained language model learns to generate product descriptions based on similar samples through self-attention mechanisms. During training, we freeze the visual encoder and the generation portion of the language model (e.g., the decoder in a sequence-to-sequence model or all autoregressive model parameters), allowing us to harness the originally powerful generative capabilities of language models for multimodal generation.

To verify the effectiveness of the proposed method, we build a new large-scale E-commerce product description generation dataset, with images and marketing keywords, based on an existing multimodal product summarization corpus (Zhu et al., 2020). Experimental results show that ModICT outperforms other strong baselines regarding almost all evaluation metrics. Both quantitative and qualitative analyses indicate that ModICT improves the semantic accuracy and diversity of generated descriptions and small language models equipped with ModICT also achieve competitive performances compared to 10x bigger LLMs.

Our contributions are summarized as follows:

- We present a product description generation paradigm that is based only on the image and several marketing keywords. For this new setting, we propose a straightforward and effective multimodal in-context tuning approach, named ModICT, integrating the power from the frozen language model and visual encoder.
- To the best of our knowledge, our work is the first one to investigate utilizing the in-context learning and text generation capabilities of various frozen language models for multimodal E-commerce product description generation. ModICT can be plugged into various types of language models and the training process is parameter-efficient.
- We conduct extensive experiments on our newly built three-category product datasets. The experimental results indicate that the proposed method achieves state-of-the-art performance on a wide range of evaluation metrics. Using the proposed multimodal in-context tuning technical, small models also achieve competitive performance compared to LLMs.

2. Related Work

Text Generation from Multimodal Sources. Text generation tasks from multimodal sources involve the interaction and transformation of information across different modalities. Image captioning (Chen et al., 2015; Chowdhury et al., 2021; Shi et al., 2021; Wang et al., 2021) is a typical image-to-text generation task (Li et al., 2023b), and it requires the model to generate the description of an image. Visual Storytelling (Huang et al., 2016; Li et al., 2020c) requires models to generate a long story, given multiple images or a video. Multimodal Machine Translation (Parida et al., 2019; Elliott et al., 2016) aims to introduce images to improve the accuracy and diversity of translation, where images

could bridge the representation gap across multiple languages. Multimodal Summarization (Li et al., 2020b; Zhu et al., 2020; Jangra et al., 2020) usually aims at generating a short text to summarize the given long text with relevant images. While prior work has mainly focused on generating text from visual information or augmenting text with additional textual content, the exploration of text generation from visual input and keywords remains limited. This unexplored area holds practical potential since short keywords are readily available and can be automatically generated, making them valuable for applications.

Product Description Generation. E-commerce product description aims to describe the characteristics of a product in detail and has obtained significant gains in the E-commerce platform. To generate detailed, diverse, and accurate descriptions, Zhang et al. (2019) propose a pattern-controlled description generation method to control the generation content based on various product properties. They design multiple generation patterns to satisfy different conditions. Chen et al. (2019) enhanced product attribute comprehension by incorporating additional knowledge, such as product materials and brand background information. Novgorodov et al. (2020) utilized customer reviews and clicked product information to diversify generated descriptions. Xu et al. (2021) proposed the K-plug model, a pretrained natural understanding and generation model, demonstrating effective product summarization generation in E-commerce. Zhan et al. (2021) and Hao et al. (2021) improved description quality using posterior distillation and user preference information. In contrast, our approach leverages marketing keywords in conjunction with images, providing complementary guidance for description generation.

Vision-assisted Language Models. Different from Vision Language Models (VLMs) that are trained with enormous multimodal data such as image-text pairs, vision-assisted language models usually incorporate external visual information into the pretrained language model, which could be used to perform multimodal reasoning or generation. The visual information is obtained by a pretrained visual encoder such as ViT (Dosovitskiy et al., 2020), Faster-RCNN (Ren et al., 2015), and other variants. Some works (Shi et al., 2019; Lu et al., 2022; Li et al., 2023b,d) are proposed to retrieve images from the image corpus and employ visual knowledge to improve the performance of the language model on the downstream tasks. Recently, researchers (Long et al., 2021; Yang et al., 2021; Zhu et al., 2022) utilize the powerful text-to-image technical to obtain the corresponding image of text and inject them into the language model via the prefix-tuning (Li and Liang, 2021) way. Li et al. (2023a) and Koh et al. (2023); Li et al. (2023c); Liang et al. (2023b); Chen et al. (2023a,b) also enable frozen large language models to perform question-answering and image-text retrieval tasks. This work will explore utilizing frozen large language models to handle the multimodal generation problem in the E-commerce field.

3. Methodology

3.1. Overview

For the new problem of E-commerce product description generation, the input contains a product image I and corresponding text sequence $W = (w_1, ..., w_i, ..., w_N)$, where w_i represents the i-th token of input keyword sequence and N indicates the total number of tokens. The output is the generated description of the product, and we define the ground-truth description as $Y = (y_1, ..., y_i, ..., y_m)$, where y_i and m refer to the i-th token and the length of the description, respectively. The proposed ModICT is a parameter-efficient multimodal in-context tuning approach for employing the frozen language models to perform multimodal generation. First, we utilize a frozen visual encoder to retrieve a similar product, which will be used to construct the in-context reference (Sec. 3.2). Then, for different types of language models, we adopt two ways to improve the efficiency of multimodal in-context tuning according to the structures of LLMs, which will be shown in Sec. 3.3. Finally, we briefly present the training strategy of ModICT in Sec. 3.4.

3.2. In-Context Reference Construction

We choose samples with visual features similar to in-context references to enhance description diversity. Human-written product descriptions exhibit significant variations, particularly for similar products in the same category. Thus, the language model should imitate human-written text styles and generate diverse, accurate descriptions based on similar products.

Selection. We employ the frozen visual encoder of CLIP (Yang et al., 2022) to obtain its image representation, denoted as $h_I = (h_{Ig}, h_{p1}, ..., h_{pn})$, where h_{Ig} and h_{pi} refer to the global and i-patch representations of the image. We consider the same-category training set as the retrieval candidate pool and utilize the same visual encoder to obtain image representations for all products as shown in Figure 2. By calculating cosine similarity scores across global representations, we retrieve the most similar product from the same category candidate set. This provides us with the image, marketing keywords, and human-written description of similar products for constructing the one-shot multimodal in-context reference.

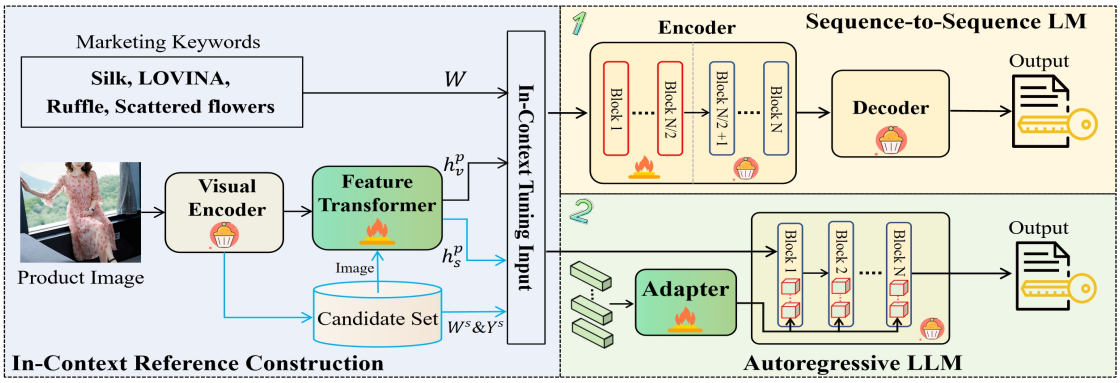

Figure 2: The overall workflow of ModICT. The left part depicts the process of in-context reference construction. The right parts show the efficient multimodal in-context tuning ways for the sequence-to-sequence language model (1) and autoregressive language model (2). Blocks with red lines are learnable.

Construction. The obtained image encodings lie in a representation space different from the language model due to the discrepancy between the frozen language model and the image encoder. To address this issue, we employ a learnable Feature Transformer to convert the image feature into the corresponding language space as the visual prefix vector. Specifically, we feed the global image representation h_{Ig} into a two-layer perceptron with Tanh activation function to obtain the fixed-length visual prefix, denoted as $h_v^p = (h_{v_1}^p, ..., h_{v_L}^p)$. L is the total length of visual prefix and $h_{v_L}^p$ represents the L-th prefix embedding. After obtaining the visual prefixes of two images, we construct a simple in-context template: *"Input Image: <img> and Marketing Keywords: W^s, output description is Y^s \n Input Image: <img> and Marketing Keywords: W, output description is "*, where <img> is a new token to represent the position of visual prefix input and will be frozen during training. To represent the in-context reference, here, we introduce W^s, Y^s, and h_s^p to represent the marketing keywords, description, and transformed image feature of the similar sample, respectively. All text inputs (including marketing keywords and the human-written description) are projected into corresponding word vectors via looking up the embedding table of the language model and the representations of <img> are added by visual prefix vectors. The whole sequence representation is input to the following blocks of the language model for multimodal generation.

3.3. Efficient Multimodal In-Context Tuning

Instead of optimizing the overall parameter of LLMs, we adopt two parameter-efficient in-context tuning methods according to the structure of LLMs. For sequence-to-sequence language models such as BART and T5, we freeze the decoder and only optimize some parameters of the encoder. In this way, we do not corrupt the generative structure of the language model and do not introduce more parameters. As the top right shown in Figure 2, we optimize the first $N/2$ blocks in the encoder to allow the model to adapt to the multimodal input, where N refers to the total number of blocks in the encoder. For the decoder-only LLMs such as BLOOM (Scao et al., 2022), GPT (Brown et al., 2020), and GLM (Du et al., 2022), inspired by the deep prompt tuning approach (Liu et al., 2021; Tang et al., 2022; Wu et al., 2022), we introduce a learnable adapter to allow LLMs to quickly adapt to this multimodal generation task without finetuning any pretrained parameters. Specifically, we randomly initialize M learnable vectors $h^v = (h_1^v, ..., h_M^v)$ and utilize a two-layer perceptron with the ReLU activation function as the adapter to project them into continuous prompts. The specific calculation process is as follows:

$$\mathbf{h}_{cp} = \mathbf{W}^a(ReLU(\mathbf{W}^1 h^v + \mathbf{b}^1)) + \mathbf{b}^a, \quad (1)$$

where $\mathbf{W}^1$, $\mathbf{W}^a$, $\mathbf{b}^1$ and $\mathbf{b}^a$ are learnable parameters. The obtained dynamic prompt sequence is $h_{cp} = (h_{cp}^1, ..., h_{cp}^{M*N})$, where N is the number of layers of the large language model. As the bottom part shown in Figure 2, these dynamic prompts are inserted into each layer of the large model and participate in the self-attention calculation process. The length of inserted vectors for each layer is equal to $M = (M*N)/N$. These continuous prompts are concatenated directly in front of the sequence of input hidden states for each layer of LLMs. Hence, we do not modify any structure of large language models and the training process is parameter-efficient.

3.4. Training and Inference

Training. For all models, we adopt the cross-entropy generation loss to train them and the spe-

cific process is given in Eq. 2,

$$\mathcal{L} = -\sum_{i=1}^{m} \log P_i(\hat{y}_i = y_i | W^s, h_s^p, Y^s; W, h_v^p; y_1, ..., y_{i-1}). \quad (2)$$

For autoregressive language models with fewer parameters (<1B), we also update their overall parameters due to that continuous prompt tuning is used for efficiently training LLMs.

Inference. For each testing sample, the corresponding similar sample is retrieved from the training set. The inferring process is similar to Eq. 2 and is equipped with some common generation methods, e.g., beam sample strategy.

4. Experiment

4.1. Dataset: MD2T

Construction. In previous product description generation tasks, input data often included product titles, attributes, images, and many others. Some information was redundantly reflected in both text and images. To address this, we build a new product description generation dataset with images and keywords, named MD2T [1], based on the large-scale millions of multimodal Chinese E-commerce product summarization corpus (Li et al., 2020a). To be specific, we collect the product style, brand, color, material, and popular element dictionaries from the released text-based E-commerce product summarization dataset (Yuan et al., 2020). The collected product attribute aspects of Cases & Bags. We then use the Chinese word segmentation tool Jieba[2] with the above dictionary to segment the long text sequence of product samples. For the obtained word segmentation set, we filter out words that can be easily derived from images (e.g., color, size, and shape) via word matching. We select the style, popular element, brand, material, and a few other randomly sampling words (20% of the number of other remaining words) from the remaining word set as the marketing keywords. Finally, we remove instances whose product descriptions do not contain any marketing keywords (exact matching) to ensure that marketing keywords can guide the description generation.

Statistic. The detailed statistics are shown in Table 1. The total number of samples across the three categories is approximately 300,000. Product descriptions are relatively long, containing about five marketing keywords and spanning around 80 words. The keyword length varies across categories, with Clothing and Home Appliances having longer keywords compared to Bags & Cases. This difference arises from challenges in collecting category-specific dictionaries for keyword segmentation and filtering, such as brand, style, and material. While some samples may contain a few noisy words (up to 2), they do not significantly impact our overall analysis.

MD2T	Cases&Bags	Clothing	Home Appliances
#Train	$18,711$	$200,000$	$86,858$
#Dev	983	$6,120$	$1,794$
#Test	$1,000$	$8,700$	$2,200$
Avg$_N$#MP	5.41	6.57	5.48
Avg$_L$#MP	13.50	20.34	18.30
Av$_L$#Desp	80.05	79.03	80.13

Table 1: The detailed statistics of MD2T. Avg$_N$ and Avg$_L$ represent the average number and length respectively. MP and Desp indicate the marketing keywords and description.

4.2. Experimental Settings

Evaluation Metrics. In our experiments, we adopt the widely-used automatic overlap-based metrics BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) in the text generation area to evaluate the generated description. However, these word-overlap based metrics ignore the contextual semantic discrepancy. Therefore, we use a popular word embedding-based similarity evaluation metric BERTScore (Zhang et al., 2020), dubbed as BS, to capture the fine-grained semantic difference, which employs the contextualized representation of words. We adopt the pretrained BERT-base-Chinese parameters to initialize the model for BERTScore. We also design a simple yet effective evaluation metric to measure the whole **diversity** of models in the testing set. Specifically, we concatenate all generated descriptions into a long sequence and remove punctuation. Then this sequence is converted into a list L based on Chinese character segmentation. The whole list sequence is divided by the Distinct N-gram segmentation method (Li et al., 2015) and thus we can obtain the no repeated n-gram word set S_n, where the maximum number of S_n does not exceed the length of L. The n-gram word diversity of generated descriptions can be calculated as follows: $D\text{-}n = \frac{Number\_of(S_n)}{Length\_of(L)} * 100$. In this way, $D\text{-}n$ can intuitively show the diverse generation capability of a model.

Comparative Models. We compare our method with several SOTA multimodal product description generation methods. MMPG (Zhu et al., 2020) is a multimodal E-commerce product summarization model based on LSTM (Hochreiter and Schmidhuber, 1997). We use it with the DecInit (D) and Copying (C) mechanism on this task, and

[1]We will release the preprocessing codes and data sources.

[2]https://github.com/fxsjy/jieba

Model	#TunedPara	#TotalPara	B@1↑	B@2↑	R@1↑	R@L↑	BS↑	D-2 ↑	D-3 ↑	D-4 ↑	D-5 ↑
MMPG+D (Zhu et al., 2020)	96M	96M	30.91	10.0	30.92	10.98	32.0	-	-	-	-
MMPG+D+C (Zhu et al., 2020)	100M	100M	33.31	10.72	31.47	21.25	31.8	2.66	4.73	7.55	10.54
M-kplug (Xu et al., 2021)	225M	225M	30.96	8.76	29.43	19.45	30.0	9.34	16.54	22.57	27.48
Oscar (Zhang et al., 2021)	110M	110M	29.52	7.58	28.32	17.93	28.2	-	-	-	-
Oscar-GPT (Cho et al., 2021)	279M	279M	33.70	10.81	32.12	21.19	32.1	11.35	22.68	33.21	44.23
ModICT (BART-L)	232M	521M	**36.54**	**13.28**	**35.43**	**24.50**	34.9	12.76	24.11	36.23	**47.31**
w/o MICT	232M	521M	36.30	13.15	35.39	24.45	34.7	12.35	23.38	34.51	45.21
w/o MICT (full)	435M	521M	34.95	11.94	33.63	22.93	33.4	10.59	20.34	31.58	42.34
ModICT (BART-RD)	181M	874M	35.40	12.56	34.10	23.92	33.9	13.07	23.86	34.40	43.67
ModICT (GLM-L)	364M	450M	28.20	10.28	32.43	22.55	32.1	10.86	20.85	30.17	38.67
ModICT (GLM-10B)	511M	10.6B	28.83	10.35	32.15	22.78	32.7	12.20	23.06	32.56	41.11
ModICT (BLOOM-1.1B)	118M	1.4B	32.61	11.88	34.08	24.23	34.6	13.23	24.07	35.38	45.74
ModICT (BLOOM-1.7B)	156M	2B	33.62	12.07	34.50	24.31	34.5	11.12	19.89	29.14	37.83
ModICT (BLOOM-3B)	226M	3.4B	33.15	12.08	33.88	24.06	34.4	12.91	23.68	34.60	44.62
ModICT (BLOOM-7.1B)	360M	7.6B	32.86	12.06	34.48	24.28	**35.3**	12.95	23.42	34.33	44.46
w/o Adapter	108M	7.6B	32.36	12.18	34.28	24.10	34.4	**13.8**	**25.08**	**36.45**	46.85
w/o Adapter+MICT	108M	7.6B	32.00	11.62	33.84	24.19	33.1	10.0	18.45	28.0	37.26

Table 2: Automatic evaluation on the testing set of Cases&Bags. The bold indicates the best performance. "#TunedPara" and "#TotalPara" represent the trainable and overall parameters of models respectively. "BS" refers to the BertScore evaluation metric. "MICT" represents the proposed multimodal in-context tuning way. "full" indicates that the overall parameters of language models are tuned during training. "Adapter" shows the parameter-efficient prefix tuning method for autoregressive LLMs.

Model	#TunedPara	#TotalPara	B@1	B@2	R@1	R@L	BS	D-2	D-3	D-4	D-5
MMPG+D (Zhu et al., 2020)	96M	96M	24.94	6.22	26.51	17.49	22.1	-	-	-	-
MMPG+D+C (Zhu et al., 2020)	100M	100M	25.04	7.05	26.12	18.16	24.8	-	-	-	-
M-kplug (Xu et al., 2021)	225M	225M	31.64	10.35	30.48	20.42	29.7	10.03	22.00	34.60	45.81
Oscar (Zhang et al., 2021)	110M	110M	27.91	6.89	26.61	16.51	25.2	-	-	-	-
Oscar-GPT (Cho et al., 2021)	279M	279M	31.22	10.15	30.22	20.27	29.6	12.22	24.15	36.43	47.41
ModICT (BART-L)	232M	521M	**34.27**	**12.56**	**33.24**	**23.42**	**32.1**	14.73	29.36	44.24	56.82
w/o MICT	232M	521M	32.57	11.43	32.04	22.90	30.7	9.34	18.82	30.43	41.28
w/o MICT (full)	435M	521M	33.56	11.89	32.73	23.08	31.5	9.01	17.11	29.17	39.98
ModICT (BART-RD)	181M	874M	30.62	10.47	31.75	22.05	29.7	16.20	32.27	46.73	58.25
ModICT (GLM-L)	364M	450M	26.07	9.65	29.40	20.43	29.1	12.62	25.74	37.46	47.41
ModICT (GLM-10B)	511M	10.6B	25.75	9.11	29.72	20.88	29.5	13.23	27.31	39.73	50.11
ModICT (BLOOM-1.1B)	118M	1.4B	30.10	10.45	30.99	22.25	29.9	13.97	27.80	41.81	53.71
ModICT (BLOOM-1.7B)	156M	2.0B	30.25	10.65	31.21	22.24	30.3	15.16	29.57	43.30	54.64
ModICT (BLOOM-3B)	226M	3.4B	29.77	10.54	31.14	22.47	30.2	14.46	28.31	42.11	53.81
ModICT (BLOOM-7.1B)	360M	7.6B	30.91	10.89	31.46	22.20	30.5	14.12	28.09	42.12	53.88
-Adapter	108M	7.6B	30.36	10.80	31.42	22.50	30.3	**17.63**	**34.31**	**48.89**	**60.50**
-Adapter-MICT	108M	7.6B	28.68	9.64	29.63	21.99	27.1	10.72	21.10	31.99	41.69

Table 3: Automatic evaluation on the testing set of Home Appliances.

Zhu et al. (2020) have verified their effectiveness to improve the quality of generated results and MMPG + D + C performs the best on the Case & Bag category in multimodal E-commerce product summarization (Li et al., 2020a). M-kplug is an extension of the text-based pretrained model k-plug (Xu et al., 2021) in E-commerce, which injects the visual signals into the decoder layer. Oscar (Li et al., 2020d; Zhang et al., 2021) is a transformer-based pretrained conditional cross-modal model, having achieved great success in many vision-language tasks. Oscar-GPT (Kayser et al., 2021) is a sequence-to-sequence vision-language generation model.

Implementation Details. We train all models on two A100-40G GPUs with the python environment. We train models with an initial learning rate $1e^{-4}$ and the learning rate declines via the linear way. The total training steps are 10 epochs with $1,000$ warm-up steps. For baselines and ModICT variants equipped with very small autoregressive language models (like GLM-L(Du et al., 2022)), we update the overall parameters of language models. The batch sizes of training and inference are set to 32 and 10 respectively. We adopt the Chinese CLIP-ViT-16 (Yang et al., 2022) as the image feature extractor, which contains 84M parameters. We use the validation set to select the best parameter when training all models. For inference, we adopt the beam sample generation method and

Model	#TunedPara	#TotalPara	B@1	B@2	R@1	R@L	BS	D-2	D-3	D-4	D-5
MMPG+D (Zhu et al., 2020)	96M	96M	29.01	8.30	27.03	17.92	30.2	-	-	-	-
MMPG+D+C (Zhu et al., 2020)	100M	100M	29.63	8.65	27.61	18.26	30.5	1.63	2.24	4.01	5.28
M-kplug (Xu et al., 2021)	225M	225M	33.38	10.91	31.42	20.34	33.0	4.5	10.05	18.54	27.22
Oscar (Zhang et al., 2021)	110M	110M	29.69	7.74	28.38	16.71	29.4	-	-	-	-
Oscar-GPT (Cho et al., 2021)	279M	279M	33.54	10.98	31.92	20.51	33.1	5.1	12.18	20.31	28.15
ModICT (BART-L)	232M	521M	**35.07**	**12.57**	**33.71**	**22.52**	**34.9**	4.17	9.28	18.05	24.45
w/o MICT	232M	521M	34.93	12.30	33.34	21.91	34.6	2.95	6.84	12.92	20.0
ModICT (BART-RD)	181M	874M	33.48	11.17	32.41	21.86	33.3	5.82	13.45	22.48	31.54
ModICT (GLM-L)	364M	450M	26.03	9.10	29.51	19.44	31.4	3.67	8.71	14.57	20.62
ModICT (GLM-10B)	511M	10.6B	26.45	9.21	29.66	19.74	31.8	**6.71**	**17.18**	**28.16**	**37.61**
ModICT (BLOOM-1.1B)	118M	1.4B	29.14	10.07	31.59	21.40	34.2	4.07	9.27	16.84	25.04
ModICT (BLOOM-1.7B)	156M	2.0B	30.25	10.17	31.26	20.83	33.6	3.75	8.16	14.25	20.76
ModICT (BLOOM-3B)	226M	3.4B	29.34	10.19	31.56	21.15	33.9	4.39	9.58	17.18	25.17
ModICT (BLOOM-7.1B)	360M	7.6B	30.58	10.52	32.12	21.65	**34.9**	4.64	9.13	16.44	25.21
w/o Adapter	108M	7.6B	30.45	10.49	31.91	21.36	34.2	5.97	13.21	22.71	32.08
w/o Adapter+MICT	108M	7.6B	30.16	10.50	31.61	21.05	33.1	3.39	7.67	14.43	21.89

Table 4: Model performances (accuracy and diversity) on the testing set of Clothing.

set the beam and sample sizes to 4 and 20, respectively. The visual prefix length L is set to 5 and the length of continuous prompts (M) is set to 10. GLM-10B and BLOOM-7B are trained in mixed precision (half-precision for the forward/backward computations, full-precision for the gradient update) with the AdamW optimizer. We test various models with different parameter sizes on our newly built dataset of three categories of products.

TrainS	R@1	R@L	BS	D-3	D-4	D-5
200k	**33.71**	**22.52**	34.9	9.28	18.05	24.45
50k	33.30	22.10	34.6	**10.64**	**18.65**	27.13
40k	33.54	22.16	**34.9**	10.37	18.57	**27.57**
30k	33.49	22.11	34.8	9.86	17.55	25.89

Table 5: ModICT(BART-L) performances on Clothing with various scales of training samples . "TrainS" refers to the size scale of training samples.

4.3. Quantitative Analysis

Content Accuracy. Evaluation results in Tables 2, 3, and 4 reveal ModICT(BART-L) excels in content accuracy across most metrics, significantly outperforming Oscar-GPT and M-kplug (e.g., R@L ↑ 3.31, BERTScore ↑ 2.8). Frozen autoregressive BLOOM also outperforms baselines in essential metrics (R@L, BERTScore). This demonstrates the effectiveness of the learnable multimodal in-context tuning approach across various language models. Additionally, BLOOM outperforms GLM-10B in all product categories, despite its larger parameter count, possibly due to its multilingual training data. Increasing language model parameters slightly enhances performance in these fixed E-commerce product descriptions, achieved through the proposed in-context tuning approach. As the parameters of the language model increase, the performance of models on the three product categories increases slightly, which may be due to the fixed description paradigm of E-commerce products, and the small model could acquire corresponding generation capability through the proposed multimodal in-context tuning approach.

Diversity. To assess diversity, we analyze Tables 2, 3, and 4. ModICT variants show varying performance across product categories but consistently outperform strong baselines (M-kplug and Oscar-GPT) in diversity evaluation. In Cases & Bags, ModICT (BART-L) improves D-5 score by 3.08, while ModICT(BLOOM-7B) and ModICT (BART-L) achieve impressive improvements of **3.93** and **9.41** for Clothing and Home Appliances. Notably, ModICT variants perform less well in Clothing, likely due to overfitting on common words in large-scale training sets, especially for variants with more parameters. Table 5 reveals that ModICT (BART-L) achieves superior content accuracy and diversity in Clothing with fewer training samples, showcasing the feasibility of training small models using multimodal in-context tuning for practical, cost-effective applications.

4.4. Ablation Study

Effectiveness of ModICT. From all experimental Tables, it is observed that, in the case of both the sequence-sequence language model and the autoregressive large language model, the MICT mainly improves the diversity of generated content. For various ModICT (BLOOM) variants, it also advances the content accuracy and substantially promotes the diversity of content (model performance comparison: -Adapter vs. -Adapter-MICT), especially for the category of Home Appliances.

Impact of Adapter. By ablation experiments on ModICT(BLOOM), we observe that the adapter

Model	NRE	R@L	BS	D-3	D-4	D-5
BLOOM-1.7B	1-shot	22.24	30.3	29.57	43.40	54.64
BLOOM-1.7B	2-shot	22.07	29.3	30.75	44.40	55.60
BLOOM-1.7B	3-shot	21.70	28.0	30.40	44.10	55.40
BLOOM-1.7B*	0-shot	21.50	26.0	20.68	31.24	40.64
BLOOM-1.7B*	1-shot	19.52	26.3	41.16	55.29	64.53
BLOOM-1.7B*	2-shot	19.73	27.3	40.23	54.39	63.96
BLOOM-1.7B*	3-shot	19.63	26.6	38.78	52.74	62.54
BLOOM-7.1B	1-shot	22.20	30.5	28.09	42.12	53.88
BLOOM-7.1B	2-shot	21.94	30.3	32.21	45.88	56.80
BLOOM-7.1B	3-shot	21.83	30.0	33.65	47.48	58.56
BLOOM-7.1B*	0-shot	21.99	27.1	21.10	31.99	41.69
BLOOM-7.1B*	1-shot	20.15	26.8	42.20	56.38	65.85
BLOOM-7.1B*	2-shot	20.12	24.8	36.77	50.42	60.40
BLOOM-7.1B*	3-shot	19.80	22.1	36.76	50.41	60.52

Table 6: Model performances with various in-context reference examples on the testing set of Home Appliances. * refers to that the corresponding model without adapter are not trained with MICT. "NRE" refers to the Number of Reference Examples.

mostly improves the overall content accuracy yet sometimes leads to a slight decrease in diversity. It may be attributed to that more parameters are introduced and we add the continuous prompts in each layer of LLMs. The performance comparisons between ModICT (BLOOM) and its "-Adapter-MICT" variant indicate that it is useful for improving the content diversity and accuracy by introducing the adapter and MICT together.

Tuned Parameter vs. Performance. Small language models (<1B) also excel in high-quality product description generation through multimodal in-context tuning. However, when fine-tuning generation-related parameters, diversity decreases (e.g., -MICT vs. -MICT(full) in Tables 2, 3, and 4), and content accuracy declines when training the overall BART-L parameters for Cases & Bags. To enhance overall accuracy and diversity, we recommend freezing the LLMs and using multimodal in-context tuning approach with one-shot reference.

Training Data Size. In Table 5, ModICT (BART-L) trained on 40k samples achieves content accuracy similar to the 200k-sample model but with better diversity. Comparing all models on the 18k sets of Clothing and Cases & Bags, they perform better in Cases & Bags with a small-scale training set. This suggests that our multimodal in-context tuning approach is effective with limited labeled data.

Analysis of In-Context References. Table 6 displays LLM performance with varying in-context examples. ModICT improves content accuracy and maintains diversity compared to ModICT(BLOOM)* with one-shot input. As in-context samples increase, ModICT diversity rises while content accuracy slightly decreases, in line with our motivation to enhance description diversity. However, it highlights the instability of large multi-modal models based on LLMs. Increasing LLM parameters results in a wider range of outcomes, yielding high diversity but lower accuracy. It may be attributed to the fact that the generation of larger language models is more diverse and less controllable.

Model	Coh	Acc	Rich	Rel
MMPG+D	4.42	2.75	3.15	2.07
M-kplug	4.48	3.21	3.18	3.19
Oscar-GPT	4.50	3.25	3.26	3.17
ModICT(BART-L)	**4.63**	3.53	**3.73**	3.49
ModICT(BLOOM-7B)	4.54	**3.61**	3.65	**3.53**
Human	4.81	3.72	4.33	3.62

Table 7: Human evaluation results on the randomly selected sample set.

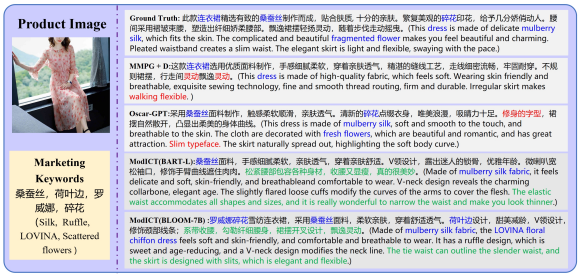

Figure 3: An illustration of descriptions generated by several models. The blue words represent keyphrases related to marketing keywords. Words in red show the inaccurate expression. The green-colored sentences are the eye-catching statements.

4.5. Qualitative Analysis

Case Study. We compare descriptions generated by various models in Figure 3. MMPG+D generates descriptions with universal characteristics but less relevance to marketing keywords. Oscar-GPT emphasizes keywords but can produce general and inaccurate statements similar to MMPG+D. The generated results of ModICT variants cover more aspects of the product and are more relevant to the marketing keywords, such as the words marked blue and green in the last generated example. In conclusion, ModICT variants could achieve superior performance in the accuracy, diversity, and vividness of the generated content.

Human Evaluation. We conduct human evaluation on content accuracy (Acc), contextual coherence (Coh), richness (Rich), and relevance (Rel) to marketing keywords. Four master students rate ground truth and model-generated descriptions on 150 randomly selected testing samples. During scoring, model-generated and human-written descriptions are randomly shuffled and reviewed blindly. ModICT outperforms strong baselines in all aspects, particularly in content coherence and richness. However, there is still room for improvement

in content richness compared to human-written descriptions. (See Table 7).

5. Conclusion

In this work, we suggest a setting of E-commerce product description generation from images and marketing keywords, where the marketing keywords provide complementary information to the image. It could guide the generation of product descriptions to some extent by providing marketing keywords. To improve the accuracy and diversity of generated descriptions, we propose a simple and effective parameter-efficient multimodal in-context tuning approach, ModICT.

Ethics Statement

The dataset included in our work is human-annotated E-commerce data that can be used for academic research, and we will release the preprocessing codes and data download source.

Acknowledge

Thanks for the efforts from reviewers and action editors. This work is supported by grants: Natural Science Foundation of China (No. 62376067).

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Zhangming Chan, Xiuying Chen, Yongliang Wang, Juntao Li, Zhiqiang Zhang, Kun Gai, Dongyan Zhao, and Rui Yan. 2019. Stick to the facts: Learning towards a fidelity-oriented e-commerce product description generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4959–4968.

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*, pages 3040–3050.

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*.

Ziyang Chen, Dongfang Li, Xiang Zhao, Baotian Hu, and Min Zhang. 2023a. Temporal knowledge question answering via abstract reasoning induction. *arXiv preprint arXiv:2311.09149*.

Ziyang Chen, Jinzhi Liao, and Xiang Zhao. 2023b. Multi-granularity temporal question answering over knowledge graphs. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11378–11392, Toronto, Canada. Association for Computational Linguistics.

Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. 2021. Unifying vision-and-language tasks via text generation. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 1931–1942. PMLR.

Sreyasi Nag Chowdhury, Rajarshi Bhowmik, Hareesh Ravi, Gerard de Melo, Simon Razniewski, and Gerhard Weikum. 2021. Exploiting image–text synergy for contextual image captioning. In *Proceedings of the Third Workshop on Beyond Vision and LANguage: inTEgrating Real-world kNowledge (LANTERN)*, pages 30–37.

Karan Desai and Justin Johnson. 2021. Virtex: Learning visual representations from textual annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11162–11173.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. GLM: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335, Dublin, Ireland. Association for Computational Linguistics.

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. Multi30k: Multilingual english-german image descriptions. *arXiv preprint arXiv:1605.00459*.

Shaoyang Hao, Bin Guo, Hao Wang, Yunji Liang, Lina Yao, Qianru Wang, and Zhiwen Yu. 2021. Deepdepict: Enabling information rich, personalized product description generation with the deep multiple pointer generator network. *ACM Trans. Knowl. Discov. Data*, 15(5).

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Baotian Hu, Qingcai Chen, and Fangze Zhu. 2015. Lcsts: A large scale chinese short text summarization dataset. *arXiv preprint arXiv:1506.05865*.

Ting-Hao Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, et al. 2016. Visual storytelling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1233–1239.

Anubhav Jangra, Sriparna Saha, Adam Jatowt, and Mohammad Hasanuzzaman. 2020. Multi-modal summary generation using multi-objective optimization. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1745–1748.

Maxime Kayser, Oana-Maria Camburu, Leonard Salewski, Cornelius Emde, Virginie Do, Zeynep Akata, and Thomas Lukasiewicz. 2021. e-vil: A dataset and benchmark for natural language explanations in vision-language tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1244–1254.

Jing Yu Koh, Ruslan Salakhutdinov, and Daniel Fried. 2023. Grounding language models to images for multimodal generation. *arXiv preprint arXiv:2301.13823*.

Haoran Li, Peng Yuan, Song Xu, Youzheng Wu, Xiaodong He, and Bowen Zhou. 2020a. Aspect-aware multimodal summarization for chinese e-commerce products. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8188–8195.

Haoran Li, Junnan Zhu, Jiajun Zhang, Xiaodong He, and Chengqing Zong. 2020b. Multimodal sentence summarization via multimodal selective encoding. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5655–5667.

Jiacheng Li, Siliang Tang, Juncheng Li, Jun Xiao, Fei Wu, Shiliang Pu, and Yueting Zhuang. 2020c. Topic adaptation and prototype encoding for few-shot visual storytelling. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 4208–4216.

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

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, Online. Association for Computational Linguistics.

Xiujun Li, Xi Yin, Chunyuan Li, Xiaowei Hu, Pengchuan Zhang, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020d. Oscar: Object-semantics aligned pre-training for vision-language tasks. *ECCV 2020*.

Yunxin Li, Baotian Hu, Yuxin Ding, Lin Ma, and Min Zhang. 2023b. A neural divide-and-conquer reasoning framework for image retrieval from linguistically complex text. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16464–16476, Toronto, Canada. Association for Computational Linguistics.

Yunxin Li, Baotian Hu, Wei Wang, Xiaochun Cao, and Min Zhang. 2023c. Towards vision enhancing llms: Empowering multimodal knowledge storage and sharing in llms. *arXiv preprint arXiv:2311.15759*.

Yunxin Li, Baotian Hu, Chen Xinyu, Yuxin Ding, Lin Ma, and Min Zhang. 2023d. A multi-modal context reasoning approach for conditional inference on joint textual and visual clues. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10757–10770, Toronto, Canada. Association for Computational Linguistics.

Ke Liang, Yue Liu, Sihang Zhou, Wenxuan Tu, Yi Wen, Xihong Yang, Xiangjun Dong, and Xinwang Liu. 2023a. Knowledge graph contrastive learning based on relation-symmetrical structure. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–12.

Ke Liang, Sihang Zhou, Yue Liu, Lingyuan Meng, Meng Liu, and Xinwang Liu. 2023b. Structure guided multi-modal pre-trained transformer for knowledge graph reasoning. *arXiv preprint arXiv:2307.03591*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Hui Lin and Vincent Ng. 2019. Abstractive summarization: A survey of the state of the art. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):9815–9822.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2021. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. *arXiv preprint arXiv:2110.07602*.

Quanyu Long, Mingxuan Wang, and Lei Li. 2021. Generative imagination elevates machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5738–5748, Online. Association for Computational Linguistics.

Yujie Lu, Wanrong Zhu, Xin Eric Wang, Miguel Eckstein, and William Yang Wang. 2022. Imagination-augmented natural language understanding. *NACCL*.

Slava Novgorodov, Ido Guy, Guy Elad, and Kira Radinsky. 2020. Descriptions from the customers: Comparative analysis of review-based product description generation methods. 20(4).

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*, pages 311–318.

Shantipriya Parida, Ondřej Bojar, and Satya Ranjan Dash. 2019. Hindi visual genome: A dataset for multi-modal english to hindi machine translation. *Computación y Sistemas*, 23(4):1499–1505.

Ankur Parikh, Xuezhi Wang, Sebastian Gehrmann, Manaal Faruqui, Bhuwan Dhingra, Diyi Yang, and Dipanjan Das. 2020. ToTTo: A controlled table-to-text generation dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1173–1186, Online. Association for Computational Linguistics.

Xiangyu Peng and Michael Sollami. 2022. Xfboost: Improving text generation with controllable decoders. *arXiv preprint arXiv:2202.08124*.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

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 (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Haoyue Shi, Jiayuan Mao, Kevin Gimpel, and Karen Livescu. 2019. Visually grounded neural syntax acquisition. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1842–1861, Florence, Italy. Association for Computational Linguistics.

Zhan Shi, Hui Liu, and Xiaodan Zhu. 2021. Enhancing descriptive image captioning with natural language inference. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 269–277.

Zhengyang Tang, Benyou Wang, and Ting Yao. 2022. Dptdr: Deep prompt tuning for dense passage retrieval. *COLING*.

Jing Wang, Jinhui Tang, Mingkun Yang, Xiang Bai, and Jiebo Luo. 2021. Improving ocr-based image captioning by incorporating geometrical relationship. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1306–1315.

Yuxiang Wu, Yu Zhao, Baotian Hu, Pasquale Minervini, Pontus Stenetorp, and Sebastian Riedel. 2022. An efficient memory-augmented transformer for knowledge-intensive nlp tasks. *arXiv preprint arXiv:2210.16773*.

Song Xu, Haoran Li, Peng Yuan, Yujia Wang, Youzheng Wu, Xiaodong He, Ying Liu, and Bowen Zhou. 2021. K-PLUG: Knowledge-injected pre-trained language model for natural

language understanding and generation in E-commerce. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1–17, Punta Cana, Dominican Republic. Association for Computational Linguistics.

An Yang, Junshu Pan, Junyang Lin, Rui Men, Yichang Zhang, Jingren Zhou, and Chang Zhou. 2022. Chinese clip: Contrastive vision-language pretraining in chinese. *arXiv preprint arXiv:2211.01335*.

Ze Yang, Wei Wu, Huang Hu, Can Xu, Wei Wang, and Zhoujun Li. 2021. Open domain dialogue generation with latent images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14239–14247.

Peng Yuan, Haoran Li, Song Xu, Youzheng Wu, Xiaodong He, and Bowen Zhou. 2020. On the faithfulness for E-commerce product summarization. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5712–5717, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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 *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14301–14309.

Tao Zhang, Jin Zhang, Chengfu Huo, and Weijun Ren. 2019. Automatic generation of pattern-controlled product description in e-commerce. In *The World Wide Web Conference*, WWW '19, page 2355–2365, New York, NY, USA. Association for Computing Machinery.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *International Conference on Learning Representations (ICLR)*.

Zhuosheng Zhang, Hanqing Zhang, Keming Chen, Yuhang Guo, Jingyun Hua, Yulong Wang, and Ming Zhou. 2021. Mengzi: Towards lightweight yet ingenious pre-trained models for chinese. *arXiv preprint arXiv:2110.06696*.

Junnan Zhu, Haoran Li, Tianshang Liu, Yu Zhou, Jiajun Zhang, and Chengqing Zong. 2018. Msmo: Multimodal summarization with multimodal output. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 4154–4164.

Junnan Zhu, Yu Zhou, Jiajun Zhang, Haoran Li, Chengqing Zong, and Changliang Li. 2020. Multimodal summarization with guidance of multimodal reference. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):9749–9756.

Wanrong Zhu, An Yan, Yujie Lu, Wenda Xu, Xin Eric Wang, Miguel Eckstein, and William Yang Wang. 2022. Visualize before you write: Imagination-guided open-ended text generation. *arXiv preprint arXiv:2210.03765*.