CapOnImage: Context-driver Dense-Captioning On Image

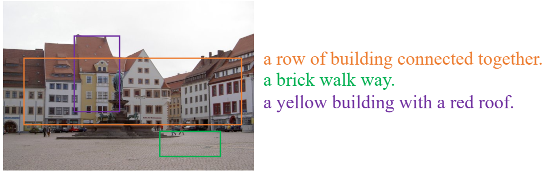
Yiqi Gao[1]*, **Xinglin Hou**[2], **Yuanmeng Zhang**[2], **Tiezheng Ge**[2]
Yuning Jiang[2], **Peng Wang**[1]†
[1]School of Computer Science, Northwestern Polytechnical University
[2]Alibaba Group
[1]`gyqjz@mail.nwpu.edu.cn`, [1]`peng.wang@nwpu.edu.cn`
[2]`{xinglin.hxl, zhangyuanmeng.zym, tiezheng.gtz, mengzhu.jyn}@alibaba-inc.com`

Abstract

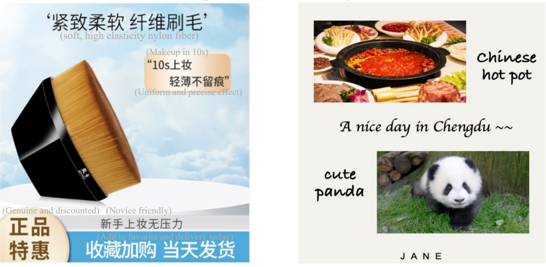
Existing image captioning systems are dedicated to generating narrative captions for images, which are spatially detached from the image in presentation. However, texts can also be used as decorations on the image to highlight the key points and increase the attractiveness of images. In this work, we introduce a new task called captioning on image (*CapOnImage*)[1], which aims to generate dense captions at different locations of the image based on contextual information. For this new task, we introduce a large-scale benchmark called CapOnImage2M, which contains 2.1 million product images, each with an average of 4.8 spatially localized captions. To fully exploit the surrounding visual context to generate the most suitable caption for each location, we propose a multimodal pre-training model with multi-level pre-training tasks that progressively learn the correspondence between texts and image locations from easy to hard. To avoid generating redundant captions for nearby locations, we further enhance the location embedding with neighbor locations . Compared with other image captioning model variants, our model achieves the best results in both captioning accuracy and diversity aspects.

1 Introduction

Building upon the advances in computer vision and natural language processing areas, the new research direction called vision-and-language has attracted more and more attentions, which pushes to tackle new problems that need to bridge the two areas to advance the concept comprehension and reasoning capabilities.


a row of building connected together.
a brick walk way.
a yellow building with a red roof.

(a) dense captioning on visual genome dataset



(b) e-commercial product image (c) social platform picture

Figure 1: Illustration of descriptive texts on image scenarios. The captions and images in existing tasks (*e.g* visual genome (a) (Krishna et al., 2017b)) are spatially detached, without any mutual association. In contrast, our CapOnImage task aims to generate descriptive text *on* the image, which has strong necessity and broad application prospects in some scenarios ((b) and (c)).

The image captioning task, as one of the most classic vision-and-language tasks, aims to generate natural language descriptions for images (Vinyals et al., 2015; Anderson et al., 2018; Zhang et al., 2021; Johnson et al., 2016). However, the captions and the images in this task are spatially detached in presentation, without any association between each other (Figure 1(a)). In fact, there are many scenarios where the image and text are tightly associated. For example, the product images on e-commercial website (Figure 1(b)) usually contain descriptive texts, explaining multiple perspectives of the product (e.g., product characteristics, selling points, etc) , which makes the image more informative and attractive. On the social media platform, users usually upload daily pictures with descriptive texts as decorations (Figure 1(c)). Therefore, it is significant to explore captioning ***on*** the image, which requires to consider not only the visual description, but also the text description placement on the image. Besides, to generate the informative captions in these scenarios, additional textual knowledge is usually needed, such as the product informa-

---

* Work done during an internship at Alibaba Group.

† Corresponding author.

[1] Code and dataset will be released at `https://github.com/YqGao716/CapOnImage`

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 3449–3465
December 7-11, 2022 ©2022 Association for Computational Linguistics

tion and the background story about the image etc. Therefore, in this work, we introduce a new task called *CapOnImage*, which aims to generate dense captions at different locations of the image based on contextual information. The CapOnImage task involves two steps, where the model needs to first predict a reasonable and aesthetic text layout (Arroyo et al., 2021; Gupta et al., 2020; Jyothi et al., 2019), and then generates a phrase or sentence for each text box. In this work, we simplify this task to generate captions for a provided list of text box locations, thereby removing the requirement for layout prediction. Therefore, the main focus of this work is to generate captions that are most suitable for the corresponding image locations.

The CapOnImage task involves two new challenges: *(i)* Better understanding of context: the captions at different locations can be greatly diverse. As shown in Figure 1(b), the texts around the product are descriptive captions describing the product features, while those at the bottom introduce the selling points. Therefore, the model needs to fully exploit the visual context around the text box to determine what caption is suitable to generate here. Since our task aims to generate captions *on* image, location context is vital for our task which is also validated on Table 2 . Overall, compared with traditional caption task, the CapOnImage task needs better understanding of context. *(ii)* Caption redundancy: some texts can be suitable for adjacent locations. Therefore, if the model can only "see" the isolated text box without surrounding ones, it tends to generate the same caption for nearby text boxes because it suits all of them, thus causing the problem of caption redundancy.

To solve the aforementioned challenges, we propose a multi-modal pre-training and fine-tuning framework which contains multi-level pre-training tasks to effectively exploit multi-modal contextual information. **First**, to better exploit context, we design multi-level pre-training tasks to help the model "feel" the context. It explicitly equip the model with the ability to distinguish which captions are appropriate for the current location and image while which are not. Besides, inspired by the learning progression of easy to complex biological vision systems, we further propose a progressive training strategy which learns multi-level pre-training tasks from easy to hard. **Second**, to solve the problem of caption redundancy, we introduce a neighbor-enhanced location encoding module, which utilizes the surrounding text box locations as context, so that our model can "see" the adjacent context. We show the captioning diversity results with different ranges of adjacent text boxes involved in the location encoding module, and demonstrate the importance of such neighbor context.

In order to evaluate our model and benchmark progress in the CapOnImage task, we introduce the *CapOnImage2M* dataset. It contains 2.1 million product images crawled from an e-commercial website, and each image contains multiple spatially localized captions describing the associated product. We automatically acquire the text contents and their spatial locations from the image via OCR (Li et al., 2017; Liu et al., 2018), and finally collect 4.8 captions for each image on average. We also crawl the product title and attributes as additional context information for caption generation. With the empirical analysis on CapOnImage2M dataset, we show that the visual context, location information and the additional product information are beneficial for the caption generation, and our model can generate corresponding types of captions at different spatial locations (Figure 3). Furthermore, we demonstrate that our proposed neighbor-enhanced location encoding module and multi-level pre-training tasks significantly improve the captioning accuracy and diversity.

The main contributions of this work are as follows:

- We introduce a new vision-and-language task called CapOnImage, which requires the model to tightly associate image and texts as a whole.
- We analyze the challenges of CapOnImage task and propose a context enhanced model with progressive training strategy, which achieves the best result compared with other image captioning model variants.
- We propose a large-scale multi-modal dataset called CapOnImage2M, with 50 categories images and localized captions to support the CapOnImage research.

2 Related Work

In recent years, significant progress has been made in the image captioning task (Vinyals et al., 2015; Anderson et al., 2018; Huang et al., 2019; Pan et al., 2020; Zhang et al., 2021), which aims to describe the image content in one natural sentence. With the advances in visual understanding abilities, researchers are not satisfied with generating dull and less informative captions and extend the traditional image captioning task along two directions.

The first direction is called dense captioning (Johnson et al., 2016; Melas-Kyriazi et al., 2018; Krishna et al., 2017a; Wang et al., 2021; Song et al., 2021), which targets to describe detailed visual content with a set of sentences. Johnson *et al.* (Johnson et al., 2016) propose a fully convolutional localization network to unify the object detection (Sermanet et al., 2014; Girshick et al., 2014; Ren et al., 2015) and image captioning in one framework to predict a set of descriptions across object regions. Krishna *et al.* (Krishna et al., 2017a) migrate it to the video, which aims to predict sequential event proposals and generate description for each clip. In these works, the dense captions deliver more details of visual content than traditional single sentence. The second direction is called text-aware image captioning (Biten et al., 2019; Sidorov et al., 2020; Yang et al., 2021), where the model generates captions not only according to the image, but also utilizes additional textual information as context. Besides , Sidorov *et al.* (Sidorov et al., 2020) and Gurari *et al.* (Gurari et al., 2020) propose to generate image captions with scene texts, which exploit OCR tokens as the textual context.

Although impressive progresses have been made along the two directions, they still remain separate. The proposed CapOnImage task can be considered as a combination of the two directions, where the model needs to first predict the text layout (spatial locations on the image) and then generate caption for each location conditioned on both the image and textual information.

There are several key distinctions between our proposed CapOnImage task and dense captioning: **1)** Dense image captioning aims to generate captions for subregions within an image, and there is no length limitation of captions. However, our task is to generate text and affix it to specific regions within an image, and the text length should be controlled according to the region size. **2)** The visual content is the only input of dense captioning. While for our task, there are three inputs: visual content, additional textual information, and the specified location to affix. All of them will impact the generated text content. **3)** The generated text is a plain description of the visual content for dense captioning. But for our task, the generated text is *on* the image, making it more informative, together with the visual content.

Table 1: Comparisons of different datasets. CI, IC and FC refer to *C*aption on *I*mage, *I*mage *C*aption and *F*ashion *C*aption.

dataset	#image	#text	avg_len	dense	task
CapOnImage	2.1M	10.07M	4.9	✓	CI
Flickr30K	30K	150K	12.3	-	IC
MSCOCO	123K	616K	10.4	-	IC
VG	108K	5M	5.7	✓	IC
TextCaps	28K	142K	12.4	-	IC
FACAD	993K	130K	21.0	-	FC

3 CapOnImage2M Dataset

In this section, we introduce our proposed CapOnImage2M dataset, which is the benchmark for the CapOnImage task. We first present an overview of the dataset collection and statistics, and then compare it with other related image captioning datasets. A more detailed datasheet describing the motivation, composition, and recommended uses of our CapOnImage2M dataset following (Gebru et al., 2018) can be found in the Appendix A.

3.1 Dataset Collection and Statistics

The CapOnImage2M dataset contains 2.1 million product images crawled from a Chinese e-commercial website[2], where each image contains both the product and descriptive captions describing the product features, efficacy, brand and so on.(detail information, *e.g* word cloud, can be found in Appendix A.) For each image, we employ an OCR toolkit to recognize the texts and their spatial locations on the image, and remove the noise with high perplexities by a pre-trained GPT.

3.2 Comparison with Other Datasets

In Table 1, we compare our CapOnImage2M dataset with other image captioning datasets. The CapOnImage2M dataset is substantially larger in both the number of images and texts. Unlike VG(Krishna et al., 2017b), where the dense captions independently describe different regions of the image, the CapOnImage2M dataset contains dense captions that describe the same product from different aspects. In addition to the dense captions on the image, each image also comes with a product title and attributes with an average length of

---

[2] https://taobao.com

34.8 characters as the textual context in the CapOnImage2M dataset. Therefore, it can also support the fashion captioning research as the FACAD(Yang et al., 2020) dataset does.

4 Model

In this section, we introduce our CapOnImage method based on the pre-training and fine-tuning framework as illustrated in Figure 2. First, we introduce the multi-modal representation of visual, location coordinates, and product information of the given input images. Then, a progressive training strategy with multi-level pre-training tasks is proposed to enhance the correspondence learning between textbox locations and captions for caption generation with a multi-layer transformer.

4.1 Input Representation

The inputs of our model include three parts from different modalities: the visual image, the textbox location coordinate and the textual product information. We independently encode each modality input as a sequence of d-dimensional feature vectors as follows.

Image representation. Given the image, we extract the grid features with standard ResNet-50 (He et al., 2016) backbone, which is further end-to-end fine-tuned with our model. We flatten the $k \times k$ feature map into a sequence and add spatial position embedding similarly as DETR (Carion et al., 2020). Specifically, for the i-th grid whose horizontal and vertical indexes are x_i and y_i, we add learnable spatial embedding and segment embedding which indicates the image modality to the appearance feature v_i as follows:

$$\hat{v}_i = v_i + [Emb_h(x_i); Emb_v(y_i)] + \mathrm{SE}_v, \quad (1)$$

where $Emb_h(\cdot)$ and $Emb_v(\cdot)$ are horizontal and vertical embedding layers with the output dimension of $\frac{d}{2}$, $[;]$ denotes concatenation and SE denotes segment embedding. Finally, we represent the image with a sequence of patch features $\hat{V} = \{\hat{v}_1, \cdots, \hat{v}_{k\times k}\}$.

Neighbor-enhanced location representation. We represent a text box location with 2D coordinates $\{(x_{min}, y_{min}), (x_{max}, y_{max})\}$, where (x_{min}, y_{min}) is the top left corner coordinate and (x_{max}, y_{max}) is the bottom right corner coordinate. We map the real value coordinates into the $k \times k$ grid and represent them with the same spatial position embeddings as image:

$$e_{cur} = [Emb_h(x_i); Emb_v(y_i); Emb_h(x_j); Emb_v(y_j)], \quad (2)$$

where $\{(x_i, y_i), (x_j, y_j)\}$ is the corresponding grid index.

Furthermore, to avoid the problem of caption redundancy for adjacent locations, we enhance the location representation with neighbor locations as context. We define the distance of two text boxes as the distance of their centers, the text boxes whose upper left corner are with smaller value of x-coordinate plus y-coordinate than the current one as the previous text boxes, and those larger than the current one as the next text boxes. Then, we employ the nearest previous textbox location and the nearest next textbox location as the neighbor context, and encode them similarly as e_{cur}. After encoding, we concatenate them with the current location embedding and add a segment embedding indicating the location modality as follows:

$$l = [W_1^T \cdot [e_{prev}; e_{next}]; W_2^T \cdot e_{cur}] + \mathrm{SE}_l, \quad (3)$$

where $W_1 \in \mathbb{R}^{4d\times\frac{d}{2}}$ and $W_2 \in \mathbb{R}^{2d\times\frac{d}{2}}$ are learned matrices, e_{prev} and e_{next} are the neighbor location embeddings, SE is the segment embedding.

Product information representation. To generate informative product descriptions, we also exploit product information as the textual context, which is the product title and attribute in this work. We concatenate them with a special $< SEP >$ token. Given the product information $X = \{x_1, \cdots, x_K\}$ with K words, we embed these words via the same word embedding matrix as the target caption words, and add positional and segment embeddings as follows:

$$w_i^{\text{info}} = W_e \cdot x_i + \mathrm{PE}_i + \mathrm{SE}_x, \quad (4)$$

where W_e is the word embedding matrix, PE denotes sequence positional embedding as in BERT (Devlin et al., 2019) and SE denotes segment embedding. Finally, we represent the product title with a sequence of d-dimensional feature vectors as $W^{\text{info}} = \{w_k^{\text{info}}\}_{k=1}^{K}$.

4.2 Pre-training Tasks and Strategy

After encoding each input modality into the common embedding space, we employ transformer layers on the multi-modal input to fuse the multi-modal information. To generate appropriate and diverse descriptions at different textbox locations,

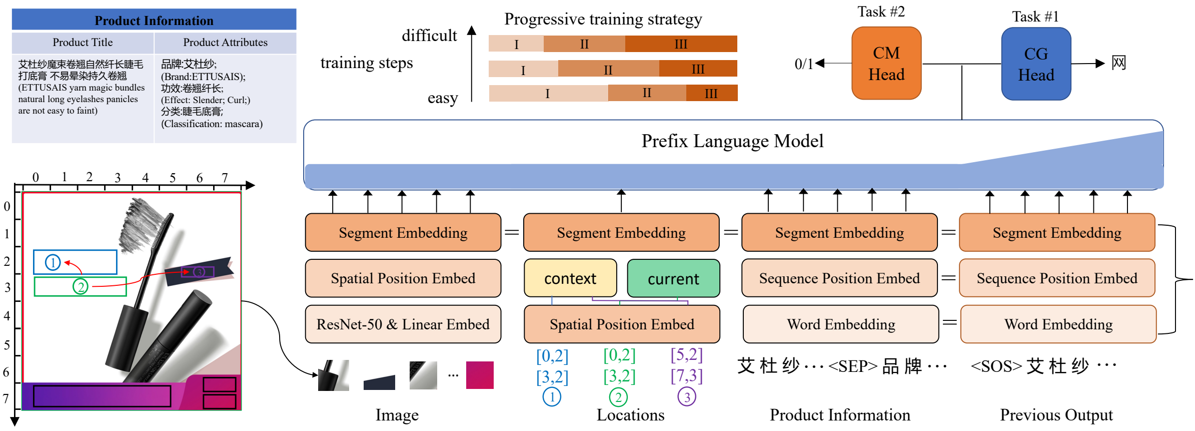

Product Information
Product Title
Product Attributes
艾杜纱魔束卷翘自然纤长睫毛打底膏 不易晕染持久卷翘
(ETTUSAIS yarn magic bundles natural long eyelashes panicles are not easy to faint)
品牌:艾杜纱;
(Brand:ETTUSAIS);
功效:卷翘纤长;
(Effect: Slender; Curl;)
分类:睫毛底膏;
(Classification: mascara)
difficult
training steps
easy
Progressive training strategy
I
II
III
Task #2
CM Head
Task #1
CG Head
0/1
网
Prefix Language Model
Segment Embedding
Spatial Position Embed
ResNet-50 & Linear Embed
context
current
Sequence Position Embed
Word Embedding
+
[0,2]
[3,2]
①
[0,2]
[3,2]
②
[5,2]
[7,3]
③
艾 杜 纱 ··· <SEP> 品 牌 ···
<SOS> 艾 杜 纱 ···
Image
Locations
Product Information
Previous Output

Figure 2: Illustration of our model with four input modalities: image patches, location coordinates, product information, and the predicted text tokens. Two pre-training tasks are employed to optimize the model with a progressive learning strategy from easy to difficult."=" denotes parameter sharing and "+" denotes addition. We add English translation for product information for better understanding.

we pre-train the model with two pre-training tasks, including Caption Generation (CG) and Caption Matching (CM). We first pre-train the model with both CG and CM tasks, and then fine-tune it only with the CG task for the final caption generation.

Task #1: Caption Generation (CG). We generate captions using the same multi-modal transformer layers as decoder following the prefix LM (Raffel et al., 2020; Dong et al., 2019). Each word prediction can attend to all the image features, neighbor-enhanced location and product information embeddings, as well as previous generated words. We adopt the auto-regressive training objective for the CG task, which can be expressed as follows:

$$\mathcal{L}_{CG} = -\frac{1}{T}\sum_{t=1}^{T}\log p(y_t^*|y_{<t}^*, \hat{V}, l, W^{\text{info}}; \Theta), \tag{5}$$

where y_t^* denotes the t-th word of ground-truth caption for the current textbox location, and Θ denotes all learnable parameters of the pre-training model. During the inference phase, we first encode the image, location and product information embeddings, and then feed a special start token `[SOS]` to predict the caption word by word.

Task #2: Caption Matching (CM). To help the model learn which captions are appropriate for the current image and location while which are not, we further introduce another pre-training task called Caption Matching. It is similar to the ITM task commonly used in vision-and-language pre-training models (Chen et al., 2020; Li et al., 2020; Lu et al., 2019; Zhuge et al., 2021), which requires the model to predict if the image and caption are semantically aligned. A score s between 0 and 1 is predicted by the hidden output of the `[SOS]` token. The positive examples of this task are corresponding pairs in the dataset, while the negative examples can be diverse. In this work, we design three levels of negative example construction and progressively learn the task from easy to difficult.

Level-I: Image caption matching. The first negative level is to randomly replace the correct caption with descriptions of other images. Therefore, it is not consistent with the current image content. We expect the model can recognize such negative examples according to the visual image and product information, which are the easiest negative cases.

Level-II: Location caption matching. The second negative level is to replace the caption with those in other locations of the same image. It is more difficult than the Level-I because the negative caption exactly describes the current image but is not suitable for the current location. For example, the product efficacy descriptions may be inappropriate to appear on the left corner of the product image, while the product brand is more suitable. We expect the model can learn the relationship of texts and textbox locations according to the surrounding visual context.

Level-III: Neighbor-location caption matching. Since the captions in neighbor locations are the most confusing samples, we further introduce the third negative level, where we randomly replace the caption with those in neighbor locations, including the nearest previous location and the nearest next location defined in Section 4.1. It can be seen as a

Table 2: We report BLEU (B), METEOR (M), CIDEr and Diversity (D) scores for the captioning on image task on the CapOnImage2M dataset. Since the CapOnImage task is a newly proposed task in this work, we adapt conventional state-of-the-art image captioning models to this task by introducing text location and textual knowledge for comparison. We run our experiments 5 times under different random seed and report the average value.

	Validation						Test					
Methods	B@1	B@4	M	CIDEr	D@1	D@2	B@1	B@4	M	CIDEr	D@1	D@2
Up-down w/o TAtt	15.51	9.79	7.93	108.11	48.73	41.22	13.48	8.71	7.89	99.52	46.53	39.51
M2 w/o TAtt	24.32	20.96	14.58	200.32	54.83	49.29	23.19	18.74	12.16	184.34	54.11	47.23
RSTNet w/o TAtt	25.18	20.46	14.73	196.91	56.17	50.85	22.87	19.35	11.81	180.50	55.36	46.86
Up-down w/ TAtt	20.49	13.54	11.52	181.04	65.24	56.27	18.94	12.30	10.81	166.26	65.72	56.48
M2 w/ TAtt	34.18	24.31	18.14	273.35	63.29	53.43	31.63	22.11	17.71	265.22	62.81	54.19
RSTNet w/ TAtt	33.42	23.91	17.28	267.60	64.12	54.53	30.54	20.94	17.20	259.29	63.10	53.84
M4C w/o copying	36.46	27.08	20.19	296.73	65.69	55.69	35.73	26.24	19.8	287.61	64.31	55.01
M4C w/ copying	35.98	28.35	20.58	299.31	65.98	55.03	36.23	27.15	20.01	288.35	64.03	55.23
baseline	36.46	27.08	20.19	296.73	65.69	55.69	35.73	26.24	19.78	287.61	64.31	55.01
w/o locations	17.78	9.36	10.03	99.08	22.94	17.24	17.29	11.04	9.73	95.52	23.05	17.28
no-info (w/o info)	20.81	13.88	11.53	133.26	55.87	47.32	19.94	13.19	10.10	125.51	53.84	46.31
no-image	22.25	16.32	13.51	155.59	58.71	49.85	21.83	15.33	11.82	147.32	58.42	48.61
context	37.69	27.98	21.91	313.64	70.25	60.54	37.02	27.23	21.14	305.50	70.98	60.90
full	**41.77**	**32.20**	**24.52**	**357.03**	74.05	63.20	**40.95**	**31.45**	**23.49**	**345.41**	74.87	63.70
human	-	-	-	-	**90.13**	**75.53**	-	-	-	-	**89.91**	**74.18**

special case of Level-II, which limits the negative location to the neighboring locations and makes it more difficult to distinguish.

Progressive training strategy. Since the three levels of CM task are from easy to difficult, inspired by the human learning procedure, we propose a progressive training strategy to dynamically adjust the proportion of each level. Specifically, we randomly replace captions with 60% probability to form negative samples and leave 40% unchanged as positive ones. The negative captions come from the three levels with p_1, p_2 and p_3 probabilities respectively, where $p_1 + p_2 + p_3 = 1$. We vary the probabilities over the course of training, according to the following formula:

$$p_1 = \min(1, 2 \cdot \mathit{step\_num}^{-0.2}), \tag{6}$$

$$p_3 = \min(1, \mathit{step\_num} \cdot 5000^{-1.5}), \tag{7}$$

$$p_2 = \max(0, 1 - p_1 - p_3). \tag{8}$$

It corresponds to rapidly decreasing the probability of Level-I from 1 at the beginning and then slowly decreasing to 0, while linearly increasing the probability of Level-III from a very small value to 1. As a result, the probability of Level-II will increase first, and then decrease. Overall, the training objective of the CM task can be expressed as follows:

$$\mathcal{L}_{CM} = -\mathbb{E}_{(\hat{V}, l, W^{\text{info}}, Y)\sim\mathcal{D}}[r \log s + (1-r)\log(1-s)], \tag{9}$$

where s refers to the predicted matching score of a training sample and $r \in [0, 1]$ is the ground-truth label indicating whether it is a negative or positive sample.

5 Experiments

We carry out experiments to evaluate the ability of models for captioning on image given a provided text layout on the CapOnImage2M dataset. We evaluate the caption generation qualities from multiple aspects, including the *accuracy* measurement against the references, and the *diversity* measurement within an image. Since the correct caption for each textbox location is not unique, we further evaluate the *fitness* of generated captions to the corresponding textbox locations with respect to the caption length and type. Besides, we also conduct human evaluations. More results can be found in supplementary materials.

5.1 Experimental Setup

Evaluation metrics. For the accuracy measurement, we evaluate the generated captions against the ground-truth with standard metrics used in the image captioning task, including BLEU (Papineni et al., 2002), METEOR (Denkowski and Lavie, 2014) and CIDEr (Vedantam et al., 2015). For the diversity measurement, we concatenate the dense captions within an image as a paragraph and compute the ratio of unique n-grams, called Div@n (Shetty et al., 2017). For the fitness measurement, we show the relationship of generated caption length to the aspect ratio of text box, and

Table 3: Captioning results with different pre-training tasks and strategies. *fixed* denotes training the multi-level CM task with a fixed proportion, while *progressive* denotes varying the proportion according to the degree of difficulty and training steps.

Row	Pretrain tasks			Pretrain strategy		Validation				Test			
	Level-I	Level-II	Level-III	fixed	progressive	B@1	B@4	M	CIDEr	B@1	B@4	M	CIDEr
1	-	-	-	-	-	37.69	27.98	21.91	313.64	37.02	27.23	21.14	305.50
2	✓	-	-	✓	-	38.87	29.23	22.32	325.74	38.19	28.48	21.69	316.46
3	✓	✓	-	✓	-	40.22	30.36	22.88	339.72	39.36	29.54	22.03	329.77
4	✓	✓	✓	✓	-	40.55	30.71	23.34	347.30	39.40	29.57	22.51	335.49
5	✓	✓	✓	-	✓	**41.77**	**32.20**	**24.52**	**357.03**	**40.95**	**31.45**	**23.49**	**345.41**

the type distribution of generated captions.

Implementation details. We initialize the ResNet-50 backbone pre-trained on ImageNet, and fine-tune it with our model in an end-to-end manner. Our model has $L = 6$ transformer layers with the hidden dimension of $d = 1024$ and attention head $A = 8$. In the pre-training stage, we sample the batch of CG and CM tasks with a proportion of 3:1 for 200K steps. We adopt a warming-up strategy for the first 4K steps. For text processing, we tokenize Chinese captions into characters and build a vocabulary with 6263 tokens. We implement our method using pytorch (Paszke et al., 2019). We manually search hyper-parameter.

5.2 Comparison with Baseline Models

Since the CapOnImage task is a newly proposed task in this work, we adapt conventional state-of-the-art image captioning models (Up-down (Anderson et al., 2018), RSTNet (Zhang et al., 2021), M2 (Cornia et al., 2020), M4C-Captioner (Sidorov et al., 2020)) to this task as the baselines for comparison. The details of compared baseline methods are expanded on the supplementary materials.

Variants of our model. We also compare with different variants of our model. Since all the words to be generated are already in the vocabulary, the copy mechanism bring no significant improvement (Table 2), so we remove it and use M4C-Captioner w/o copying as our *baseline* model. The *no-info* model adopts the same architecture as the *baseline* model except that the product information input is removed. Similarly, the *no-image* and the *no-locations* model share the same baseline model architecture but with the image and locations input removed respectively. Our *context* model is the *baseline* model enhanced with neighbor location contexts, which is still trained only with the CG task. The *full* model is our complete model with progressive pre-training by both CG and CM tasks on the same dataset.

Table 2 reports the captioning on image results of different models on the CapOnImage2M validation and test sets. It is shown that the conventional image captioning model without any adaptation perform poorly on the CapOnImage task. This is because these models lack the textual context that can provide rich information for caption generation. Enhancing the Up-down, M^2, and RSTNet with textual attentions on the additional product information, the captioning results are significantly improved. However, they are still inferior to the adapted text-aware image captioning model, which has a good ability of multi-modal fusion with the cross transformer encoder. Therefore, it stands for a strong baseline for our model. Compared with the *baseline* model, our *context* model enhances the text location embedding with neighbor location contexts, which brings significant improvements on both accuracy and diversity metrics. It demonstrates the importance of location relationship modeling especially for reducing caption redundancy. Although good results have been achieved, the model is only trained with caption generation objective against the ground-truth, which is not sufficient to help the model learn complex correspondence between texts and image locations. Therefore, when pre-training the *context* model with both CG and CM tasks in a progressive manner, our *full* model achieves the state-of-the-art results. Nevertheless, there is still a gap with the human annotations on the captioning diversity metrics.

To further explore the contribution of each input modalities, we also report the captioning results with some input removed(*no-locations*, *no-info*, and *no-image*). It shows that the location information is more important than the visual image and the textual information for the CapOnImage task. However, these three models are severely inferior to the *baseline* model with multi-modal input, which shows the necessity of multi-modal fusion for this new task.

5.3 Ablative Analysis

Parameters determination.

In Figure 4, we conduct ablation studies to investigate the suitable parameters for our model. We use our *context* model and operate our experiment on test set. The parameter that need to be determined are number of layers of transformer, hidden dimension of transformer and grid feature size of resnet backbone. In Figure 4(a), we study the impact of these three parameters for caption performance(BLEU@4 and CIDEr). We choose 6 layer transformer with hidden dimension 1024 and 8×8 grid size resnet for the intuition of Accuracy-Efficiency Trade-Offs.

Pre-training tasks and strategy. In Table 3, we ablate the proposed multi-level pre-training tasks and progressive training strategy. It shows that pre-training with only Level-I of the CM task (row 2) can significantly improve the non-pretrained model with only CG objective (row 1). It demonstrates the importance of multi-modal alignment to the CapOnImage task. Upgrading the CM task with more difficult negatives in Level-II helps the model better learn the relationship of captions and text locations and thus yields better results (row 3). Further incorporating negatives in Level-III bring additional gains (row4), which confirms the importance of context information in CapOnImage task. However, since three levels of negative samples are built from easy to difficult, we seek to boost the learning process of CM task in an adaptive fashion: the ratio of pre-training tasks need to be adapted to the training status and vary in a progressive manner(as opposed to a fixed proportion of 30%:40%:30%). Therefore, we propose a progressive training strategy with the proportion of easy task decreased and hard task increased in the training process. It boosts the results stably (row 5).

5.4 Caption length to textbox aspect ratio.

Given a textbox location, the generated caption should exactly fit in with it for visual aesthetic. The text length and font size are the influencing factors. Since the short side of the textbox determines the font size, the aspect ratio (long side length / short side length) can reflect the most suitable text length.

Table 4: Human evaluation of our *full* model vs. *baseline* model on the test set w.r.t. relevance, diversity and informativeness.

	Base wins (%)	Full wins (%)	Delta
relevance	31.2	68.8	**+37.6**
diversity	34.5	65.5	**+ 31.0**
informativeness	32.8	67.2	**+34.4**

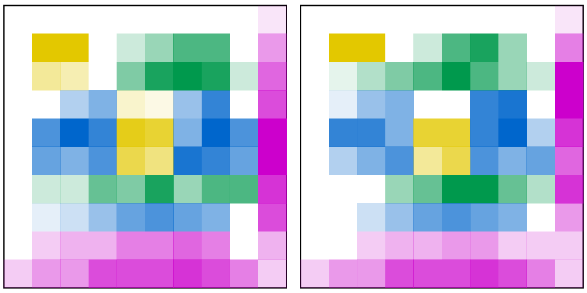
Therefore, we show the relationship of our generated caption length with aspect ratio of the corresponding textbox in the Figure 4(b). It shows that with the textbox aspect ratio increased, our model generates longer captions almost linearly, which demonstrates the controllability of the text box size to the length of the caption generated by our model.

5.5 Caption type to textbox location.

Besides the caption length, the types of captions in CapOnImage2M dataset are diverse at different image locations. To explore whether the type of our generated captions is suitable to the given locations, we visualize the caption type distribution on the image. Since the caption type annotations are not available, we automatically group the ground-truth captions into 4 categories by k-means based on their sentence-level BERT (Devlin et al., 2019) embeddings. We then display the same type of captions using the same color on an image. As shown in Figure 3(a), the captions with the same type are located together, which shows that the caption type is very related to its location. We assign our generated captions to the 4 clusters and visualize them in the same way in Figure 3(b). It looks very similar to the ground-truth type distribution map, which shows that our model effectively learns the relationship of text location and text type. The meaning of each color are illustrated in Figure 3(c).

5.6 Human Evaluation

In addition to the objective evaluation, we also conduct human evaluation on 400 randomly sampled images from the test set. We render the generated dense captions from *baseline* model and our *full* model on the image via opencv. We instruct 5 workers to choose which one is better or they are not distinguishable based on relevance, diversity and informativeness respectively and do the majority voting. To avoid the prior bias, we anonymize the model names and shuffle the predictions randomly. Table 4 shows the human evaluation results. Our *full* model significantly outperforms


(a)

(b)

Color	Illustration	Examples(EN)	Examples(CN)
🟨	product brand	Valida,Shiseido, Avene, Clinique, ...	万利达，资生堂，雅漾，倩碧, ...
🟩	product name	Anti-robbery sawtooth, milk facial cleanser, ...	防抢夺锯齿，牛奶洁面乳, ...
🟦	short description of product features	Baby grade PP material, Strong load-bearing without dropping,...	婴儿级PP材质，强承重不掉落, ...
🟪	long description of product features	Long-lasting and waterproof design,compact and soft fiber bristle, ...	持久续航防水设计,紧凑柔软纤维刷子, ...

(c)

Figure 3: *(a)*:Distribution of the GT caption types. *(b)*: Distribution of generated caption types. *(c)*: Illustration of the four caption types in CapOnImage2M dataset via automatic clustering.

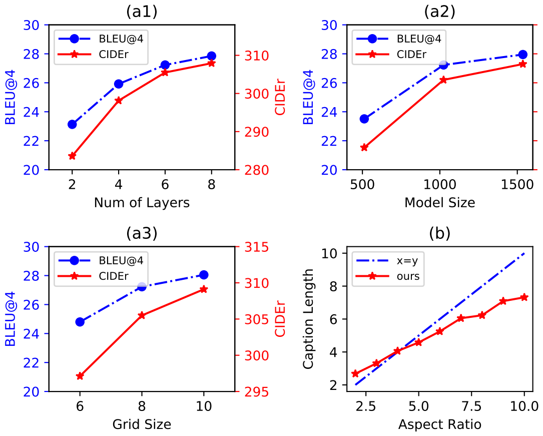


Figure 4: *(a1)*: Ablation of number of transformer layer. *(a2)*: Ablation of the hidden dimension of transformer model. *(a3)*: Ablation of grid size of resnet backbone. *(b)*: The average captioning length of our model for the text box with different aspect ratios.

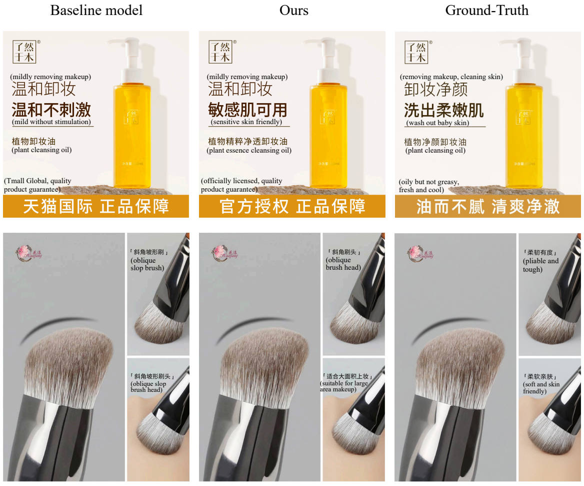

Baseline model
Ours
Ground-Truth
(mildly removing makeup)
温和卸妆
温和不刺激
(mild without stimulation)
植物卸妆油
(plant cleansing oil)
(Tmall Global, quality product guarantee)
天猫国际 正品保障
(mildly removing makeup)
温和卸妆
敏感肌可用
(sensitive skin friendly)
植物精粹净透卸妆油
(plant essence cleansing oil)
(officially licensed, quality product guarantee)
官方授权 正品保障
(removing makeup, cleaning skin)
卸妆净颜
洗出柔嫩肌
(wash out baby skin)
植物净颜卸妆油
(plant cleansing oil)
(oily but not greasy, fresh and cool)
油而不腻 清爽净澈
「斜角坡形刷」(oblique slop brush)
「斜角坡形刷头」(oblique slop brush head)
「斜角刷头」(oblique brush head)
「适合大面积上妆」(suitable for large area makeup)
「柔韧有度」(pliable and tough)
「柔软亲肤」(soft and skin friendly)

Figure 5: Qualitative dense-captioning on image results of our *full* model and *baseline* model. We add the English translation for each Chinese caption for better comprehension.

the *baseline* model especially on all three aspects, which demonstrates the effectiveness of the proposed neighbor-enhanced location embedding and multi-level progressive pre-training.

5.7 Qualitative Results

Figure 5 visualizes some results of our *full* model and *baseline* model. The *baseline* model is shown to generate repetitive captions due to the lack of global layout awareness. For example, for adjacent locations, the *baseline* model repeats the concept of "mild", while our model generates more informative caption of "sensitive skin friendly". Furthermore, our model is also shown to better exploit the visual context to generate more suitable captions. In the second example, our model generates the text "suitable for large area makeup" for the down-right region where a hand appears, while the *baseline* model fails to distinguish it with the up-right region and generates similar descriptions about the "oblique slop brush". More visualization results can be found in the supplementary material.

6 Conclusion

In this work, we propose a new vision-and-language task called CapOnImage, which aims to generate dense captions at different locations on an image with visual and textual context. We propose a multi-modal pre-training and fine-tuning model with multi-level pre-training tasks from easy to difficult for the correspondence learning between image location and text, and enhance the current location embedding with neighboring locations to reduce captioning redundancy. Experimental results shown that our model can generate controllable length and type of captions at different image locations. In the future work, we will explore to generate dense captions with self-predicted text layout and combine the layout generation with caption generation in one joint framework to benefit from each other.

Limitations

The definition of our proposed task, *i.e.*, generating text on image locations based on visual and textual context, can be found in many scenarios, such as billboard photos, posters, social platform images, etc. In this paper, we only report performance on our collected e-commercial dataset for the convenience of validating our key idea and our proposed task does not rely on any priors of specific inputs, so it can be expanded to a wide range of scenarios. In the future, we plan to collect more types of datasets, which can help us to apply our approach to more scenarios. Also, our dataset only contains caption annotations in Chinese.

Acknowledgement

This work was supported by National Key R&D Program of China (No. 2020AAA0106900), the National Natural Science Foundation of China (No. U19B2037, No. 61876152), Shaanxi Provincial Key R&D Program (No. 2021KWZ-03), Natural Science Basic Research Program of Shaanxi (No. 2021JCW-03) and Alibaba Group through Alibaba Innovative Research Program. We thank Yuqing Song, Wei Suo, Mengyang Sun, and Peng Wu for their helpful discussion.

References

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 visual question answering. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 6077–6086.

Diego Martín Arroyo, Janis Postels, and Federico Tombari. 2021. Variational transformer networks for layout generation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 13642–13652.

Ali Furkan Biten, Lluís Gómez, Marçal Rusiñol, and Dimosthenis Karatzas. 2019. Good news, everyone! context driven entity-aware captioning for news images. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 12466–12475.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. UNITER: universal image-text representation learning. In *European Conference on Computer Vision*, pages 104–120.

Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. 2020. Meshed-memory transformer for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10578–10587.

Michael J. Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 376–380.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Conference of the North American Chapter of the Association for Computational Linguistics*, pages 4171–4186.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. In *Annual Conference on Neural Information Processing Systems*, pages 13042–13054.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna M. Wallach, Hal Daumé III, and Kate Crawford. 2018. Datasheets for datasets. *CoRR*, abs/1803.09010.

Ross B. Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 580–587.

Kamal Gupta, Alessandro Achille, Justin Lazarow, Larry Davis, Vijay Mahadevan, and Abhinav Shrivastava. 2020. Layout generation and completion with self-attention. *CoRR*, abs/2006.14615.

Danna Gurari, Yinan Zhao, Meng Zhang, and Nilavra Bhattacharya. 2020. Captioning images taken by people who are blind. In *European Conference on Computer Vision*, pages 417–434.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778.

Lun Huang, Wenmin Wang, Jie Chen, and Xiaoyong Wei. 2019. Attention on attention for image captioning. In *IEEE International Conference on Computer Vision*, pages 4633–4642.

Justin Johnson, Andrej Karpathy, and Li Fei-Fei. 2016. Densecap: Fully convolutional localization networks for dense captioning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4565–4574.

Akash Abdu Jyothi, Thibaut Durand, Jiawei He, Leonid Sigal, and Greg Mori. 2019. Layoutvae: Stochastic scene layout generation from a label set. In *IEEE International Conference on Computer Vision*, pages 9894–9903.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017a. Dense-captioning events in videos. In *IEEE International Conference on Computer Vision*, pages 706–715.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2017b. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vis.*, 123:32–73.

Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. 2020. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence*, pages 11336–11344.

Hui Li, Peng Wang, and Chunhua Shen. 2017. Towards end-to-end text spotting with convolutional recurrent neural networks. In *IEEE International Conference on Computer Vision*, pages 5248–5256.

Xuebo Liu, Ding Liang, Shi Yan, Dagui Chen, Yu Qiao, and Junjie Yan. 2018. FOTS: fast oriented text spotting with a unified network. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5676–5685.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Annual Conference on Neural Information Processing Systems*, pages 13–23.

Luke Melas-Kyriazi, Alexander M. Rush, and George Han. 2018. Training for diversity in image paragraph captioning. In *Conference on Empirical Methods in Natural Language Processing*, pages 757–761.

Yingwei Pan, Ting Yao, Yehao Li, and Tao Mei. 2020. X-linear attention networks for image captioning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 10968–10977.

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.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2015. Faster R-CNN: towards real-time object detection with region proposal networks. In *Annual Conference on Neural Information Processing Systems*, pages 91–99.

Pierre Sermanet, David Eigen, Xiang Zhang, Michaël Mathieu, Rob Fergus, and Yann LeCun. 2014. Overfeat: Integrated recognition, localization and detection using convolutional networks. In *International Conference on Learning Representations*.

Rakshith Shetty, Marcus Rohrbach, Lisa Anne Hendricks, Mario Fritz, and Bernt Schiele. 2017. Speaking the same language: Matching machine to human captions by adversarial training. In *IEEE International Conference on Computer Vision*, pages 4155–4164.

Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. 2020. Textcaps: A dataset for image captioning with reading comprehension. In *European Conference on Computer Vision*, pages 742–758.

Yuqing Song, Shizhe Chen, and Qin Jin. 2021. Towards diverse paragraph captioning for untrimmed videos. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 11245–11254.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4566–4575.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3156–3164.

Teng Wang, Ruimao Zhang, Zhichao Lu, Feng Zheng, Ran Cheng, and Ping Luo. 2021. End-to-end dense video captioning with parallel decoding. In *IEEE International Conference on Computer Vision*, pages 6847–6857.

Xuewen Yang, Heming Zhang, Di Jin, Yingru Liu, Chi-Hao Wu, Jianchao Tan, Dongliang Xie, Jue Wang, and Xin Wang. 2020. Fashion captioning: Towards generating accurate descriptions with semantic rewards. In *European Conference on Computer Vision*, pages 1–17.

Zhengyuan Yang, Yijuan Lu, Jianfeng Wang, Xi Yin, Dinei Florêncio, Lijuan Wang, Cha Zhang, Lei Zhang, and Jiebo Luo. 2021. TAP: text-aware pre-training for text-vqa and text-caption. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8751–8761.

Xuying Zhang, Xiaoshuai Sun, Yunpeng Luo, Jiayi Ji, Yiyi Zhou, Yongjian Wu, Feiyue Huang, and Rongrong Ji. 2021. Rstnet: Captioning with adaptive attention on visual and non-visual words. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 15465–15474.

Mingchen Zhuge, Dehong Gao, Deng-Ping Fan, Linbo Jin, Ben Chen, Haoming Zhou, Minghui Qiu, and Ling Shao. 2021. Kaleido-bert: Vision-language pretraining on fashion domain. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 12647–12657.

A Appendix

In the appendix, we first conduct further analysis of the choices of contextual locations and the diversity of our generated captions. Then claim the motivation and the challenge for the novel CapOnImage task. At last, we take hierarchical presentation of the CaptionOnImage2M from different perspectives.

A.1 Choice of contextual locations.

Table A5: Captioning results on the test set with different contextual locations. The number in () means how many locations used as the context to enhance the current location embedding.

Methods	B@1	B@4	CIDEr	D@1	D@2
w/o context (0)	35.73	26.24	287.61	69.31	59.01
w/ two random (2)	33.98	24.79	283.73	67.43	57.82
w/ top-1 nearest (2)	37.02	**27.23**	305.50	**70.98**	**60.90**
w/ top-2 nearest (4)	**37.88**	27.10	**307.61**	70.12	60.34

In Table A5, we take a further study on the neighbor-enhanced location embedding module with different contextual locations. With the nearest neighbor (previous and next) locations used as the context as described in Section 4.1, our model significantly improves the accuracy and diversity metrics. To figure out where the benefit comes from, we compare with the model using the same amount of randomly selected locations as context. Experimental results show that the randomly selected locations cannot improve the results and may even bring noise, which demonstrates the effectiveness of our model in encoding neighboring layout information to generate more appropriate and diverse captions. When further expanding the contextual range from the top-1 nearest to the top-2 nearest (top-2 previous and top-2 next), the model achieves slightly better result on the accuracy metric. To balance the efficiency and quality, we finally use the top-1 nearest locations as the context in our model.

A.2 Diversity from the textual input.

In Table A6, we calculate the Bleu score between the input product information and our generated captions. Results show that there is only a small percentage of copy text in our generated captions, demonstrating that our model is not just simply "copying text from the input product information", but generating diverse captions conditioned on the multi-modality input.

Table A6: Bleu score between the input product information and our generated captions.

#	Bleu1	Bleu2	Bleu3	Bleu4
Test	0.032	0.021	0.013	0.007

A.3 Motivation

For what purpose was the dataset created? The dataset was created to support the research on the captioning on image (CapOnImage) task, which aims to generate informative captions at different appropriate locations in the given image. CapOnImage is a valuable task for both vision-and-language research and industrial applications. We show the pipeline of our task on Figure A6.

A.4 Composition

What do the instances that comprise the dataset represent? Each instance in the CapOnImage2M dataset contains 50 categories product image, a sentence of product title, several product attributes, and multiple spatially localized captions with bounding box coordinates, describing the product from multiple aspects. We show the word cloud of product categories on Figure A7. Figure A8 shows some examples of the CapOnImage2M dataset.

How many instances are there in total? The dataset consists of 2.1M images and 10.07M texts in total. Each image contains an average of 4.8 spatially localized captions.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? CapOnImage2M is a new independent dataset. The instances in CapOnImage2M dataset are crawled from an e-commercial website. New product images with texts will continue to emerge. Therefore, the cur-

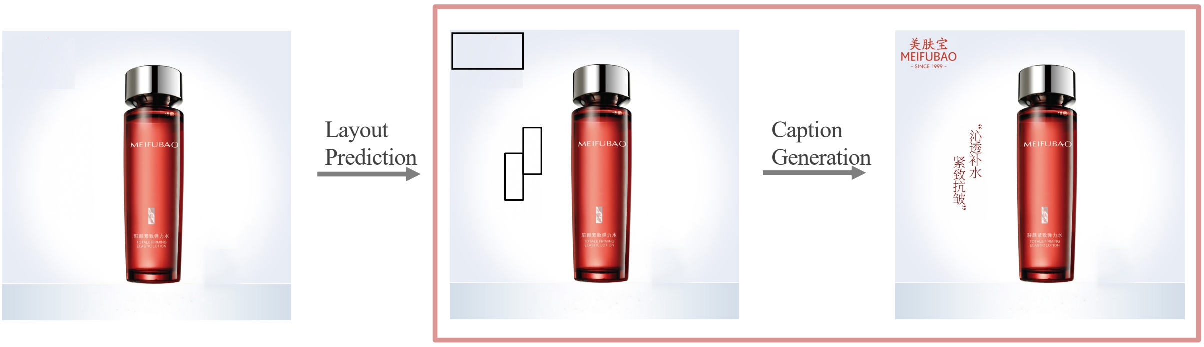


Figure A6: Pipeline of our CapOnImage task. We focus on generating captions for a provided list of text box locations in this work.

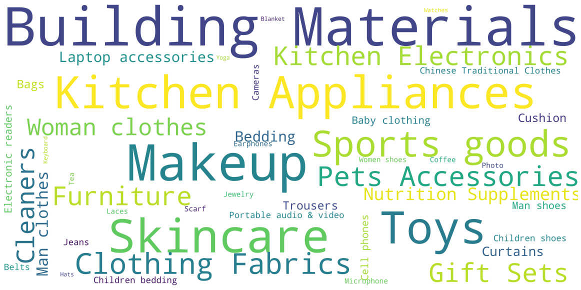


Figure A7: Word Cloud of 50 product categories. We show the Word Cloud according to the instances number of each category.

rent version of the dataset does not contain all possible instances.

What data does each instance consist of? Raw data (e.g., unprocessed text or images) or features? Each instance contains a product image with short side as 256 pixel in PNG format, the processed product title and captions. The original image with high resolution can be downloaded by the provided image URL.

Is any information missing from individual instances? No. Everything is included in the dataset.

Are there recommended data splits (e.g., training, development/validation, testing)? Yes. The dataset is split into 2.06M images for training, 20K for validation, and 20K for testing.

Are there any errors, sources of noise, or redundancies in the dataset? We have removed redundant or similar images whose captions are overlapped over a threshold. Therefore, the instance redundancy in the dataset is low. The captions are automatically recognized by an OCR model, therefore, there are noise in the captions of training set. To ensure the evaluation accuracy, we further manually clean the OCR errors for the validation and testing sets.

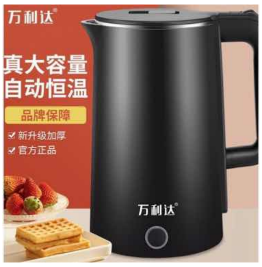
Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? The dataset is self-contained, except that the original high resolution images are linked to a public website. Nevertheless, the self-contained images are enough to reproduce the results, and the high resolution images are used for better visualization.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctorpatient confidentiality, data that includes the content of individuals non-public communications)? No. All data was collected from a publicly available e-commercial website.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? No. The dataset only consists of product images.

Does the dataset relate to people? Most of the images only contain products, and only a few images contain public product spokesmen.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? No. The dataset does not contain confidential information since all information was crawled from a public e-commercial website.

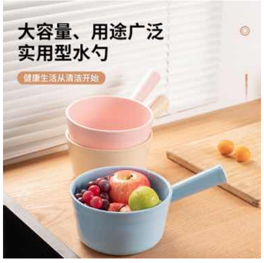


Product Title:
万利达烧水壶大容量快烧壶煲水宿舍学生小型防烫不锈钢家用煮水壶(Valida Boiling Water Large Capacity Quick Boiling Students Anti-Scalding Stainless Steel Kettle)

Product Attributes:
品牌:万利达(Brand: Valida);材质:不锈钢(material: stainless steel);锅盖类型:不锈钢盖(Lid Type: stainless steel lid);

Dense Spatially Localized Captions:

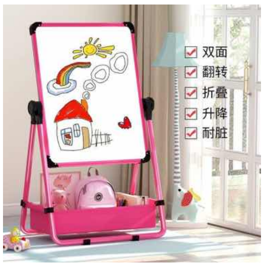
万利达(Valida)	(15, 16, 161, 61)
真大容量(large capacity)	(0, 152, 268, 218)
自动恒温 (automatic constant temperature)	(5, 225, 267, 292)
品牌保障(brand guarantee)	(58, 323, 213, 362)
新升级加厚 (upgrade thickening)	(47, 390, 199, 422)
官方正品(official product)	(51, 436, 168, 467)



Product Title:
家用塑料水勺加深长柄水瓢可挂透明塑料水勺浴室宝宝沐浴洗头水舀(Plastic Water Scoop, Plastic Water Ladle Bath Ladle Dipper Shampoo Ladle Cup Household Accessories for Kitchen Bathroom)

Dense Spatially Localized Captions:

大容量 (large Capacity)	(46, 71, 248, 119)
实用型水勺 (Practical Water Scoop)	(45, 138, 323, 184)
健康生活从清洁开始 (Healthy live starts with cleanliness)	(60, 226, 299, 254)
用途广泛 (Widely Usage)	(275, 71, 494, 118)

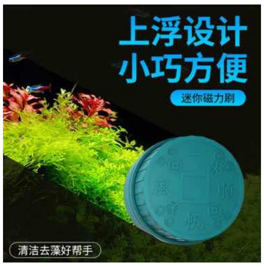


Product Title: 创意女孩礼品智力手绘板多功能美术绘画板素描便携式画板儿童(Creativity Girls Gifts Children's Drawing Creative Card Stickers Portable)

Product Attributes:
适用性别:中性(Gender:Neutral);
材质:合金(Material: Alloy)

Dense Spatially Localized Captions:

双面 (double side)	(617, 129, 696, 165)
翻转 (rotate)	(618, 189, 696, 227)
折叠 (fold)	(618, 252, 697, 291)
升降 (lift)	(618, 318, 695, 356)
耐脏 (dirt resistant)	(617, 379, 696, 416)

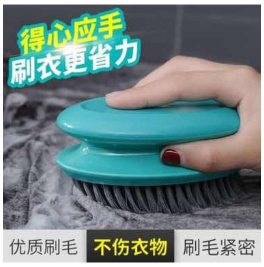


Product Title:
森森迷你磁力刷水族双面清洁刷除苔玻璃悬浮刮刀蓝色小巧磁力刷(SENSEN Magnetic Aquarium Fish Tank Brushes Floating Clean Glass Window Algae Scraper Cleaner Brush Easy To Use Fish Tank Clean Tools)

Product Attributes:
品牌:森森 (Brand: SENSEN)

Dense Spatially Localized Captions:

上浮设计 (float design)	(356, 31, 760, 131)
小巧方便 (tiny and easy to use)	(350, 145, 761, 247)
迷你磁力刷 (mini magnetic brush)	(552, 274, 718, 306)
清洁去藻好帮手 (algae clean tools)	(45, 746, 277, 779)

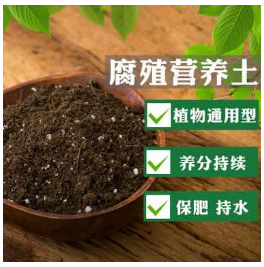


Product Title:
洗衣刷软毛家用刷子洗衣服的鞋子洗鞋专用硬毛清洁多功能板刷鞋刷(Multi-function Silicone Laundry Brush Soft Hair Cleaning Shoes Brush Underwear Brushes Cleaning Laundry Underwear Brushes)

Product Attributes:
尺寸(size): 60*60*75mm
颜色(color): 绿(green), 红(red), 白(white)

Dense Spatially Localized Captions:

得心应手(handy)	(38, 70, 386, 96)
刷衣更省力(easy to brush clothes)	(5, 168, 434, 185)
优质刷毛(high quality brush hair)	(19, 726, 233, 778)
不伤衣物(no harm for clothes)	(276_723_513_780)
刷毛紧密(dense brush hair)	(557, 725, 784, 779)

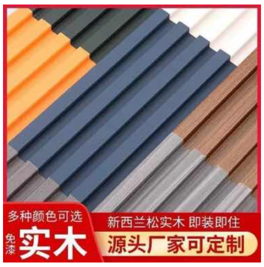


Product Title:
营养土通用型100斤花土大包50斤园土100斤阳台种菜多肉用种植土(50kg Nutrient Soil Rich Fertilizer NPK For Plant Flower Succulent Garden Bonsai)

Product Attributes:
植物营养类型(type): 营养土(Nutrient Soil)

Dense Spatially Localized Captions:

腐殖营养土 (humic nutrient soil)	(327, 165, 795, 252)
植物通用型 (plant universal)	(412, 317, 788, 368)
养分持续 (nutrients last)	(546, 466, 757, 517)
保肥持水 (keep fertilizer and water)	(535, 605, 764, 654)

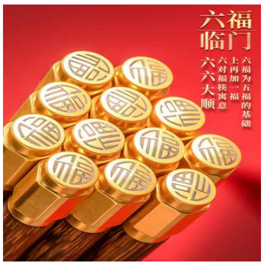


Product Title:
实木格栅板实木格栅电视背景墙板网格栅实木格栅板护墙板(TV Background Wall Border Decorative Strips Stickers Skirting Waist Line Self-adhesive Skirting Wall Stickers Soft Lines)

Product Attributes:
材质(material): XPe

Dense Spatially Localized Captions:

多种颜色可选 (multiple colors to choose)	(11, 638, 248, 673)
实木 (wood)	(57, 698, 246, 780)
免漆 (paint free)	(7, 706, 38, 780)
新西兰松实木即装即住 (Newzealand pine solid wood anytime ready to live)	(320, 655, 736, 692)
源头厂家可定制(customizable)	(327, 717, 747, 767)

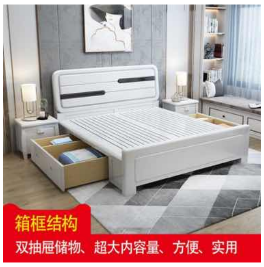


Product Title:
红木鸡翅木高档礼品家用筷子套装防霉防滑可刻字中国风(Red Wood High-quality Household Alloy Chopstick Home Non-slip Tableware Reusable Food Sticks Traditional Chinese Style)

Product Attributes:
品牌:本纳(Brand: BENNA)
材质: 合金(material: alloy)

Dense Spatially Localized Captions:

六福	(566, 18, 728, 95)
临门 (about to come)	(570, 101, 728, 181)
六对福筷寓意 (the meaning of this stick)	(621, 178, 656, 416)
上再加一福(plus additional luck)	(656, 197, 688, 426)
六福为五福的基础	(689, 183, 723, 506)
六六大顺(good luck)	(570, 189, 619, 420)

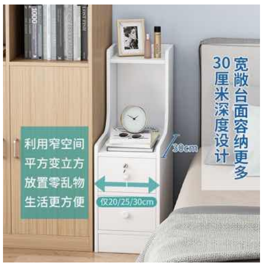


Product Title:
居家现代约简实床1.8m木米双人床白橡木黑白色床卧室家家用(Solid wood bed double bed large bed solid wood master bedroom 1.8 m wedding bed)

Product Attributes:
风格: 欧式(Style: European)
材质:橡胶木(Material: Wood)

Dense Spatially Localized Captions:

箱框结构 (box frame structure)	(36, 654, 229, 702)
双抽屉储物 (double drawer storage)	(37, 733, 256, 772)
超大内容量 (extra large amount of content)	(264, 733, 498, 772)
方便(convience)	(523, 733, 624, 773)
实用(practical)	(630, 733, 727, 771)



Product Title:
高款床头柜置物架简约现代北欧风卧室收纳小型床边储物小柜(Bedside Cabinet Modern Minimalist Small Storage Cabinet Bedside Cabinet Storage European Style Cabinet Bedroom)

Product Attributes:
深度(depth): 30cm

Dense Spatially Localized Captions:

利用窄空间 (take advantage of narrow spaces)	(63, 414, 254, 449)
平方变立方(square to cube)	(64, 474, 253, 509)
放置零乱物(place messy)	(64, 534, 255, 569)
生活更方便(convience life)	(64, 594, 255, 630)
30厘米深度设计 (30cm depth design)	(659, 200, 707, 491)
宽敞台面容纳更多 (width surface to accommodate more)	(711, 162, 769, 541)

Figure A8: Examples of the CapOnImage2M dataset. We add the English translation for better understanding.

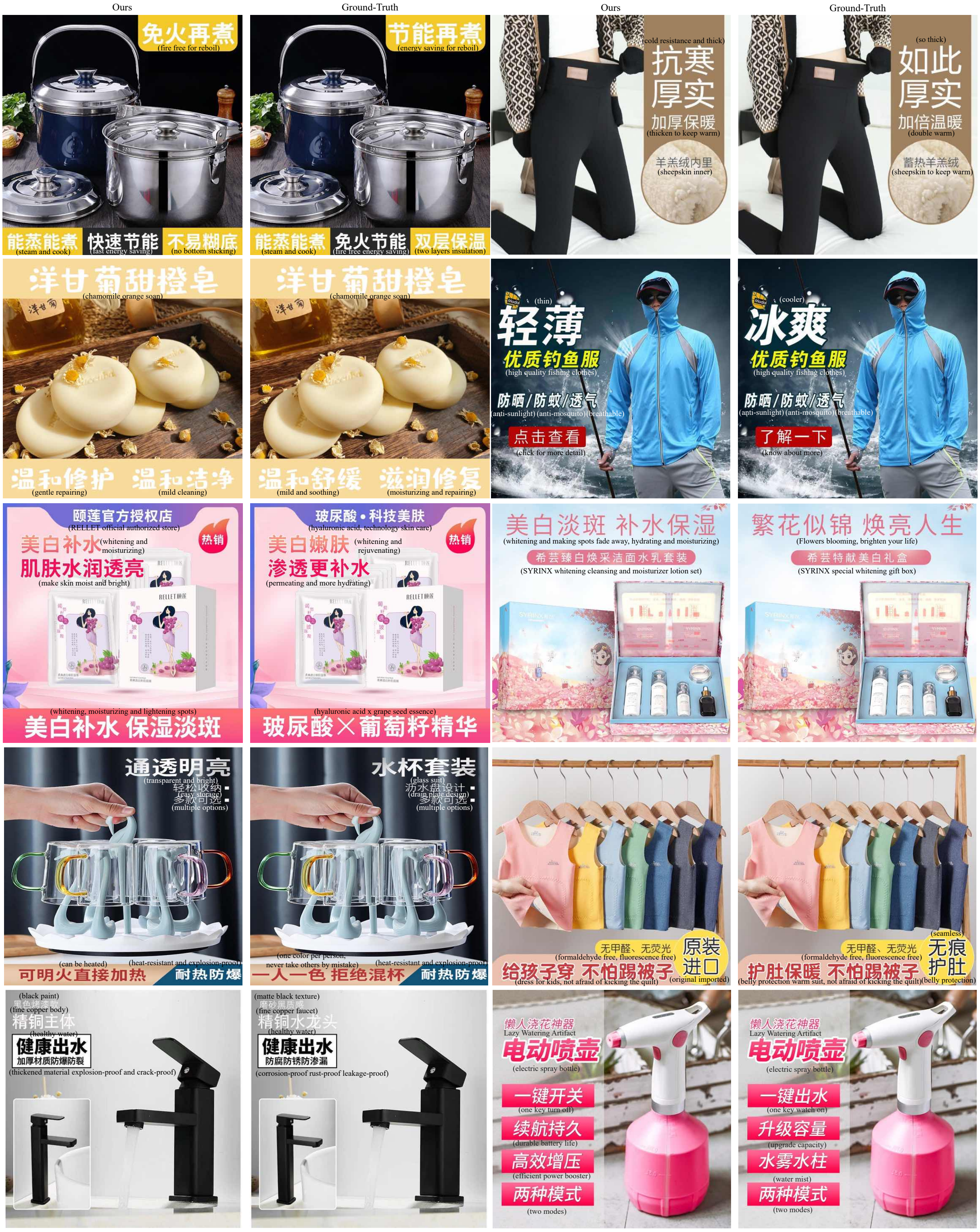

Ours
Ground-Truth
Ours
Ground-Truth
免火再煮
(fire free for reboil)
能蒸能煮
(steam and cook)
快速节能
(fast energy saving)
不易糊底
(no bottom sticking)
节能再煮
(energy saving for reboil)
能蒸能煮
(steam and cook)
免火节能
(fire free energy saving)
双层保温
(two layers insulation)
(cold resistance and thick)
抗寒厚实
加厚保暖
(thicken to keep warm)
羊羔绒内里
(sheepskin inner)
(so thick)
如此厚实
加倍温暖
(double warm)
蓄热羊羔绒
(sheepskin to keep warm)
洋甘菊甜橙皂
(chamomile orange soap)
温和修护
(gentle repairing)
温和洁净
(mild cleaning)
洋甘菊甜橙皂
(chamomile orange soap)
温和舒缓
(mild and soothing)
滋润修复
(moisturizing and repairing)
(thin)
轻薄
优质钓鱼服
(high quality fishing clothes)
防晒/防蚊/透气
(anti-sunlight) (anti-mosquito) (breathable)
点击查看
(click for more detail)
(cooler)
冰爽
优质钓鱼服
(high quality fishing clothes)
防晒/防蚊/透气
(anti-sunlight) (anti-mosquito) (breathable)
了解一下
(know about more)
颐莲官方授权店
(RELLET official authorized store)
美白补水
(whitening and moisturizing)
肌肤水润透亮
(make skin moist and bright)
热销
(whitening, moisturizing and lightening spots)
美白补水 保湿淡斑
玻尿酸 • 科技美肤
(hyaluronic acid, technology skin care)
美白嫩肤
(whitening and rejuvenating)
渗透更补水
(permeating and more hydrating)
热销
(hyaluronic acid x grape seed essence)
玻尿酸╳葡萄籽精华
美白淡斑 补水保湿
(whitening and making spots fade away, hydrating and moisturizing)
希芸臻白焕采洁面水乳套装
(SYRINX whitening cleansing and moisturizer lotion set)
繁花似锦 焕亮人生
(Flowers blooming, brighten your life)
希芸特献美白礼盒
(SYRINX special whitening gift box)
通透明亮
(transparent and bright)
轻松收纳
(easy storage)
多款可选
(multiple options)
(can be heated)
可明火直接加热
(heat-resistant and explosion-proof)
耐热防爆
水杯套装
(glass suit)
沥水盘设计
(drain plate design)
多款可选
(multiple options)
(one color per person, never take others by mistake)
一人一色 拒绝混杯
(heat-resistant and explosion-proof)
耐热防爆
无甲醛、无荧光
(formaldehyde free, fluorescence free)
给孩子穿 不怕踢被子
(dress for kids, not afraid of kicking the quilt)
原装进口
(original imported)
无甲醛、无荧光
(formaldehyde free, fluorescence free)
护肚保暖 不怕踢被子
(belly protection warm suit, not afraid of kicking the quilt)
(seamless)
无痕护肚
(belly protection)
(black paint)
黑色烤漆款
(fine copper body)
精铜主体
(healthy water)
健康出水
加厚材质防爆防裂
(thickened material explosion-proof and crack-proof)
(matte black texture)
磨砂黑质感
(fine copper faucet)
精铜水龙头
(healthy water)
健康出水
防腐防锈防渗漏
(corrosion-proof rust-proof leakage-proof)
懒人浇花神器
Lazy Watering Artifact
电动喷壶
(electric spray bottle)
一键开关
(one key turn off)
续航持久
(durable battery life)
高效增压
(efficient power booster)
两种模式
(two modes)
懒人浇花神器
Lazy Watering Artifact
电动喷壶
(electric spray bottle)
一键出水
(one key watch on)
升级容量
(upgrade capacity)
水雾水柱
(water mist)
两种模式
(two modes)

Figure A9: Qualitative results of our *full* model on the CapOnImage2M test set. We add the English translation for better understanding.

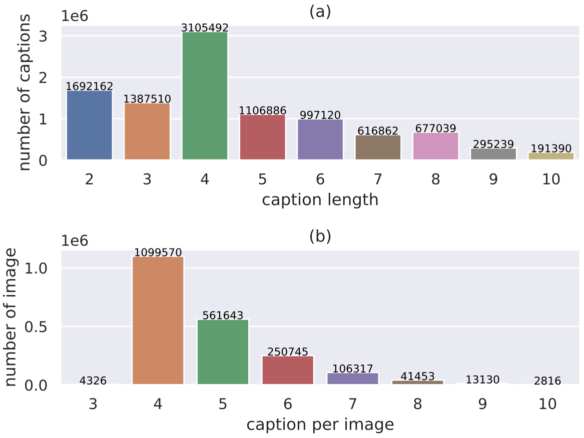


Figure A10: *(a)*: Length distribution of caption. *(b)*: Distribution of captions per image.

A.5 Collection Process

What mechanisms or procedures were used to collect the data? The raw images and product title sentences in this dataset were automatically crawled from Taobao website [3]. The spatially localized captions are extracted from the image by an OCR model, and further manually cleaned for the validation and testing sets.

A.6 Preprocessing

Was any preprocessing/cleaning/labeling of the data done? The following steps were taken to process the data: (1) *Crawling raw images and product titles*. We first crawl the raw product images and titles from the e-commercial website, and then resize the images with short side as 256. (2) *Detecting texts on the image*. For each image, we employ an OCR toolkit to automatically detect the texts on the image as well as their bounding box coordinates. (3) *Removing redundant instances*. We remove redundant instances whose images contain similar captions that exceed the overlap threshold. (4) *Removing discount information*. We remove redundant instances that have high correlation with discount information. (5) *Cleaning instances*. We remove the texts longer than 10 or shorter than 2 characters, and remove the instances with only one caption on the image. Then, we input the automatically recognized captions into a pre-trained GPT model, and remove the captions with high generation perplexities. (6) *Manual labeling*. We further manually clean the captions with OCR errors in the validation and testing sets for accurate evaluation.

Is the software used to preprocess/clean/label the instances available? Yes. All software used to process the data is open source and has been mentioned above.

A.7 Uses

Has the dataset been used for any tasks already? No, the dataset is newly collected from scratch in this work to support the proposed new task.

What (other) tasks could the dataset be used for? The dataset was created to support the CapOnImage task. In addition, since each product image is accompanied by a product title sentence, it can be directly used for the Fashion Captioning (Yang et al., 2020) task. It may support a wider range of vision-and-language tasks as well.

A.8 Distribution

We show the distribution of caption length and number of captions per image in Figure A10.

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? Yes. The dataset will be released publicly.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? The dataset can be downloaded from Google Drive and Baidu Disk as a gzipped tar file.

When will the dataset be distributed? The dataset will be released upon the publication of this work.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? There will be no license. Users only need to fill in an agreement form regarding the dataset not to be used for commercial purposes and citation suggestions etc.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? No. There are no fees or restrictions.

A.9 Maintenance

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? Yes. The dataset will be updated for fair comparison with future works if there is any kind of changes.

---

[3] https://taobao.com