Supplementary Material for "Learning Instance-Level Representation for Large-Scale Multi-Modal Pretraining in E-commerce"

Yang Jin[1,2], Yongzhi Li[2], Zehuan Yuan[2], Yadong Mu[1*]
[1]Peking University [2]ByteDance Inc.
jiny@stu.pku.edu.cn, liyongzhi.ailab@bytedance.com,
yuanzehuan@bytedance.com, myd@pku.edu.cn

In this supplementary material, we first present more implementation details about the pretraining dataset and model architecture in Section 1. Then, more experimental details and results analysis on the downstream tasks are given in Section 2. To better demonstrate the superior generalization and grounding capability of ECLIP, we illustrate more visualization examples in Section 3. Finally, additional analysis and discussion are provided in Section 4.

1. More Pre-training Details

1.1. Pre-training Dataset Details

A large-scale dataset is indispensable for training a powerful foundation model. To this end, we construct a massive E-commerce pretraining dataset that consists of 100M image-text pairs and includes various product categories. All data samples are collected directly from a popular E-commerce website without further manual annotation. The details of the dataset collection are elaborated as follows:

Sample various products. We first collect a large number of products covering various categories: clothes, daily use, cosmetics, etc. To avoid long-tail distribution, these products are further uniformly sampled according to their categories. After processing, we harvest around 12M product items covering a total of 9K categories.

Collect images from different sources. For each collected product item, we sample several image samples from different sources: product details pages, customer comments, and attached advertisement videos. As shown in Figure 1, the detail pages contain 3-4 images provided by merchants to display the product being sold from multiple views. We also select the 3 images taken by customers from the hottest comments related to a product item. In addition, the attached advertisement videos showcase the product's appearance and usage to customers, and we randomly sample 5-6 video frames as image samples. For product items without comments and advertisement videos, we only collect images from product details pages. In summary, there are about 100M diverse E-commerce images.

Living room oil painting style tissue box

Figure 1. Images of one product from different sources.

Make positive training pairs. During pretraining, two image samples from different sources but belonging to the identical product are treated as positive pairs. Both positive image samples have the same text description. For the few items with less than two images, we treat their randomly augmented image as the positive pair. In each training batch, we ensure that only two images are from the identical product, and the others are from different ones.

1.2. Implementation Details

In our instance decoder, the head number of the self-attention module is 8 and the hidden dimension of the feed-

*Corresponding Author.

forward layers is 2048. The common embedding dim D is 512 for all variants. An AdamW optimizer [7] is used for training with weight decay 0.01. The learning rate for image and text encoders is set to $5e^{-5}$, and $1e^{-4}$ for the rest modules, which is first linearly warmed up and then exponentially decayed with a factor of 0.85. Besides, we set the exponential-moving-average update parameter to 0.998. Inspired by [5], we sample hard negative prompts based on global contrastive image-to-text and text-to-image similarities. This strategy contributes to learning a stronger and more robust representation. Following [3], the queue size for inter-product contrastive learning is set to 65,536. The overall pretraining procedures of ECLIP are two-stage: first freeze the weight of the instance decoder and only train two encoders with $\mathcal{L}_{itc}$ for 10 epochs. Then, train the whole components with all loss terms for another 5 epochs. In the second stage, the gradients of decoders are not propagated to two encoders.

2. More Experimental Details and Results

Additional details for transferring ECLIP to various downstream E-commerce tasks are described below.

2.1. Downstream Task Details

2.1.1 Zero-Shot Product Retrieval

For coarse-level product retrieval, the query and gallery set contains 24,410 and 1,197,905 product samples following [2]. During the evaluation, a product pair is considered a match if both belong to the same category. For instance-level retrieval, we follow the setting introduced in [11]: each query image encompasses multiple different kinds of product instances and the model needs to retrieve all the related products. There are 9,220 query samples and 40,033 gallery samples in the instance-level retrieval. Similar to the classification task, we conduct retrieval using the image-text pair of a product. It is worth noting that the text descriptions between the matched query and gallery samples in M5Product and Product1M benchmarks are very similar. In this case, text modality will dominate the retrieval performance, which is harmful to reflecting the role of visual features. Therefore, for the challenging fine-grained product retrieval, we build a dataset that contains 26,000 products as the query set and 130,000 products as the gallery set, where the text descriptions of matched pairs are adequately different. Besides, a product pair is considered a match if and only if they are the same products. The significance of instance features can be fully verified in this task.

2.1.2 Zero-Shot Visual Grounding

Our ECLIP learns the fine-grained cross-modal alignment capability to localize the desired instance indicated by the text. To validate this, we transfer ECLIP to zero-shot visual grounding task. The collected grounding dataset has 484,385 image-text pairs, where each image contains only one ground truth box corresponding to the textual description. To evaluate the grounding performance, following [9], we first leverage an off-the-shelf region proposal network to extract a set of bounding-box proposals for each image. This network is pretrained on a human-annotated object detection dataset. Then, we estimate the similarity score between the text and each 16×16 image patch. The obtained 2D score map is further interpolated to the origin input resolution: $\mathcal{S} \in \mathcal{R}^{H \times W}$. Each box proposal $b = (x1, y1, x2, y2)$ is ranked based on $r(\cdot)$:

$$r(b) = \frac{1}{\sqrt{\text{area}(b)}} \sum_{x=x_1}^{x2} \sum_{y=y_1}^{y2} S_{x,y} \tag{1}$$

We select box with the maximum $r(\cdot)$ as the grounding result. The performance is finally evaluated by accuracy at IoU thresholds $\{0.5, 0.7\}$ with the ground truth box.

Setting	Methods	IoU Thresh	
		Acc@0.5	Acc@0.7
Zero-Shot	CLIP [8]	79.8	72.1
	ALBEF [5]	80.2	74.8
	$\text{Ours}_{\text{ViT-B/16}}$	**88.7**	**85.6**

Table 1. Performance comparisons of different approaches to zero-shot image-conditioned visual grounding.

2.1.3 Object Detection

On object detection, an image usually contains multiple foreground instances. Hence, we increase the query number to 40. The position and type embeddings of newly added 20 queries are copied directly from the pre-trained ones. Besides, the box prediction head is implemented as a 3-layer MLP as in DETR [1]. For the compared baselines, we initialize the image encoder with the pre-trained weight and train other parameters from scratch. The collected object detection dataset contains 146,813 samples from 18 classes for training and 25,909 samples for evaluation. During fine-tuning, the image resolution is increased to 640×640 and the batch size is set to 128 on 8 A100 GPUs. We finetune the entire model for 80 epochs, with the learning rate of $1e{-}5$ for the image encoder and $8e^{-5}$ for the remaining ones.

2.2. Additional Experimental Results

2.2.1 Zero-Shot Image-Conditioned Grounding

The image-conditioned Grounding requires the model to localize the instance depicted by a query image. Different from traditional visual grounding, it is suitable when the query is difficult to describe through text. Since ECLIP also

Figure 2. Qualitative examples of object detection.

incorporates image prompts during pretraining, we can thus transfer it to image-conditioned grounding without further finetuning. To this end, we construct a grounding dataset consisting of 100K samples, where each one has a test and query image. The test image contains multiple instances and the query image has only one instance that needs to be grounded. During evaluation, each query is first embedded into $g_I(v_{\text{cls}})$ by the image encoder. Similar to Section 2.1.2, we can estimate the 2D similarity map and then rank each proposal to select the candidate with maximum $r(\cdot)$. Table 1 shows detailed results for ECLIP and compared baselines. Notably, ECLIP is still highly competitive in image-conditioned visual grounding. By contrast, approaches that learn global representation ignore fine-grained feature alignment and thus achieve worse performance.

2.2.2 Ablation of Instance Query Number

We also explore the effect of instance query number T of the ECLIP decoder. Due to the huge training costs, all experiments are conducted on a smaller pretraining subset that includes only 5M images. As shown in Table 2, increasing the query number from 20 to 40 will slightly boost performance on instance-level product retrieval. It is intuitive because more queries bring the potential to focus on more instances. Accordingly, the computation and GPU memory load will increase as well. Therefore, we set $T = 20$ during pretraining for all other experiments.

2.2.3 The effect of Slot-Attention

The slot-attention distributes each visual token to one of T queries according to their similarity, which explicitly divides an image into T different parts. Such a mechanism will encourage the positive query to focus on the region that contains correlated instances and the negative ones to be distracted by the background. In contrast, the cross-attention weights tend to be smoothed over all image tokens (See Figure 3). We verify the effectiveness of slot-attention via replacing it with cross-attention in the decoder layer. The results in Table 3 present clear performance gain brought by slot-attention. We also explore the effect of loss terms $\mathcal{L}_{itm}$ and $\mathcal{L}_{\mathcal{R}}$. As in Table 3, all the components contribute to the final performance.

T	Peak GPU Memory	mAP@10	mAP@50	mAP@100
10	47.1 GB	87.5	82.7	81.3
20	48.6 GB	89.6	84.6	**82.2**
40	52.3 GB	**89.9**	**84.7**	81.9

Table 2. Ablation results of different number of instance queries on instance-level product retrieval (Visual Modality).

Setting	Visual Grounding		Coarse-Level Retrieval	
	Acc@0.5	Acc@0.7	mAP@1	mAP@5
Cross-Attention	73.9	66.3	53.2	55.4
Slot-Attention	**78.7**	**70.5**	**54.7**	**57.8**
Base	77.4	69.6	53.2	56.4
Base+$\mathcal{L}_{itm}$	78.1	70.0	54.2	56.9
Base+$\mathcal{L}_{itm}$+$\mathcal{L}_{\mathcal{R}}$	**78.7**	**70.5**	**54.7**	**57.8**

Table 3. Ablation results of different loss terms and slot attention.

3. More Visualization Results

In order to qualitatively demonstrate the strong generalization of ECLIP on downstream E-commerce tasks, we provide more visualizations in this section.

Object Detection. The finetuned object detection results

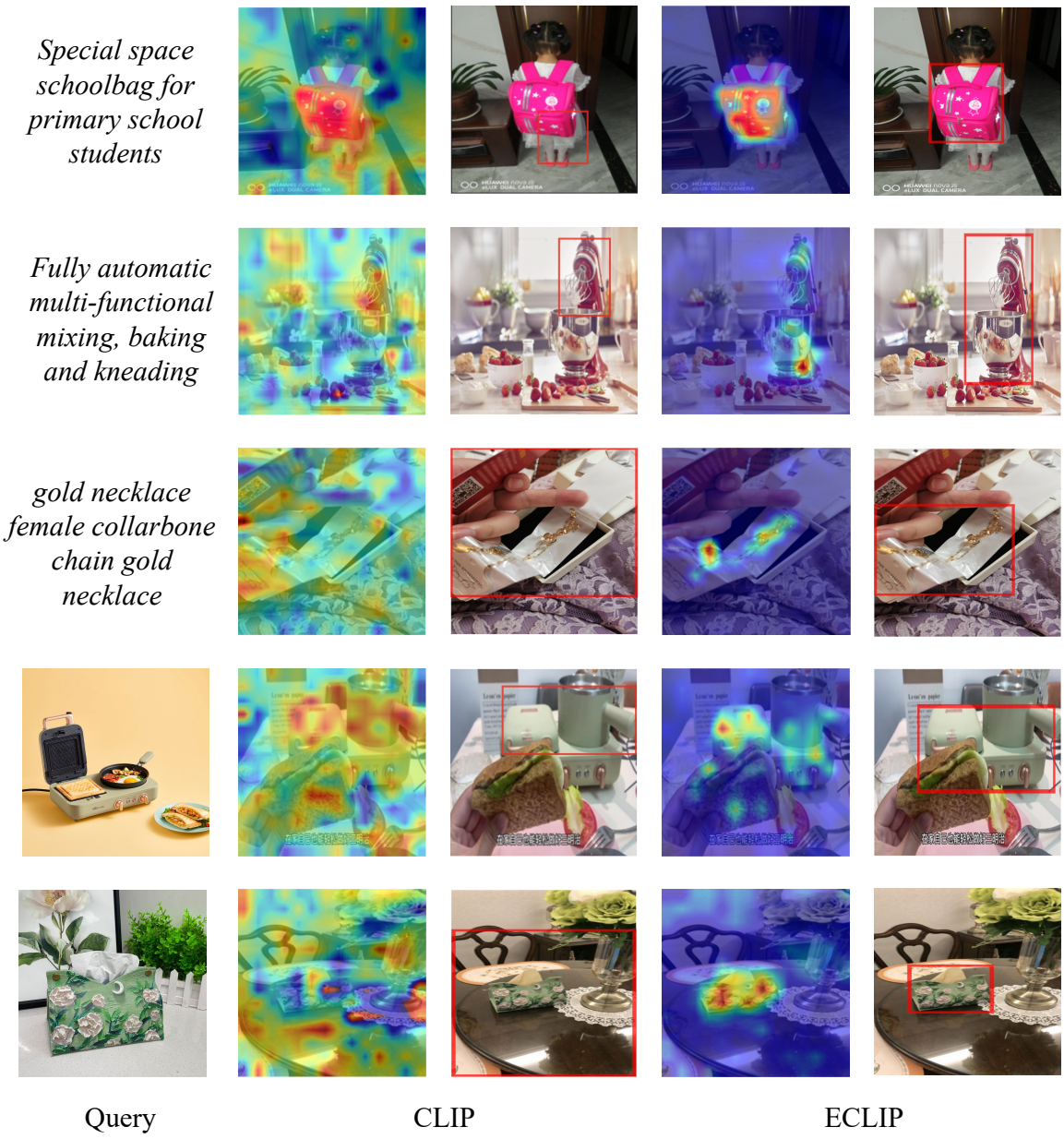

Figure 3. Visualizations of zero-shot text- and image-conditioned visual grounding achieved by CLIP and ECLIP.

achieved by ECLIP are shown in Figure 2. It further illustrates the promising fine-grained understanding capability in real-world E-commerce applications. Even for some complex scenes, ECLIP is able to successfully detect the various product instances appearing in the given image.

Zero-Shot Visual Grounding. We here provide additional qualitative examples of zero-shot visual grounding. In Figure 3, we illustrate the comparison of text- and image-conditioned grounding with CLIP. Obviously, ECLIP can properly attend to the desired instance depicted by the text or image query. Moreover, compared to CLIP, the similarity map is highly concentrated on the target regions of interest. However, CLIP's similarity score map is distracted by many unrelated background instances, due to ignorance of the instance-level modeling.

Zero-Shot Product retrieval. We also illustrate the coarse- and instance-level product retrieval examples in Figure 4 and Figure 5. The images marked in green are the samples correctly retrieved, while images marked in red are mismatched ones. It can be observed that ECLIP successfully returns satisfactory retrieval results. Especially for challenging instance-level retrieval, it can still recall the existing product instances in a query image from a large gallery set.

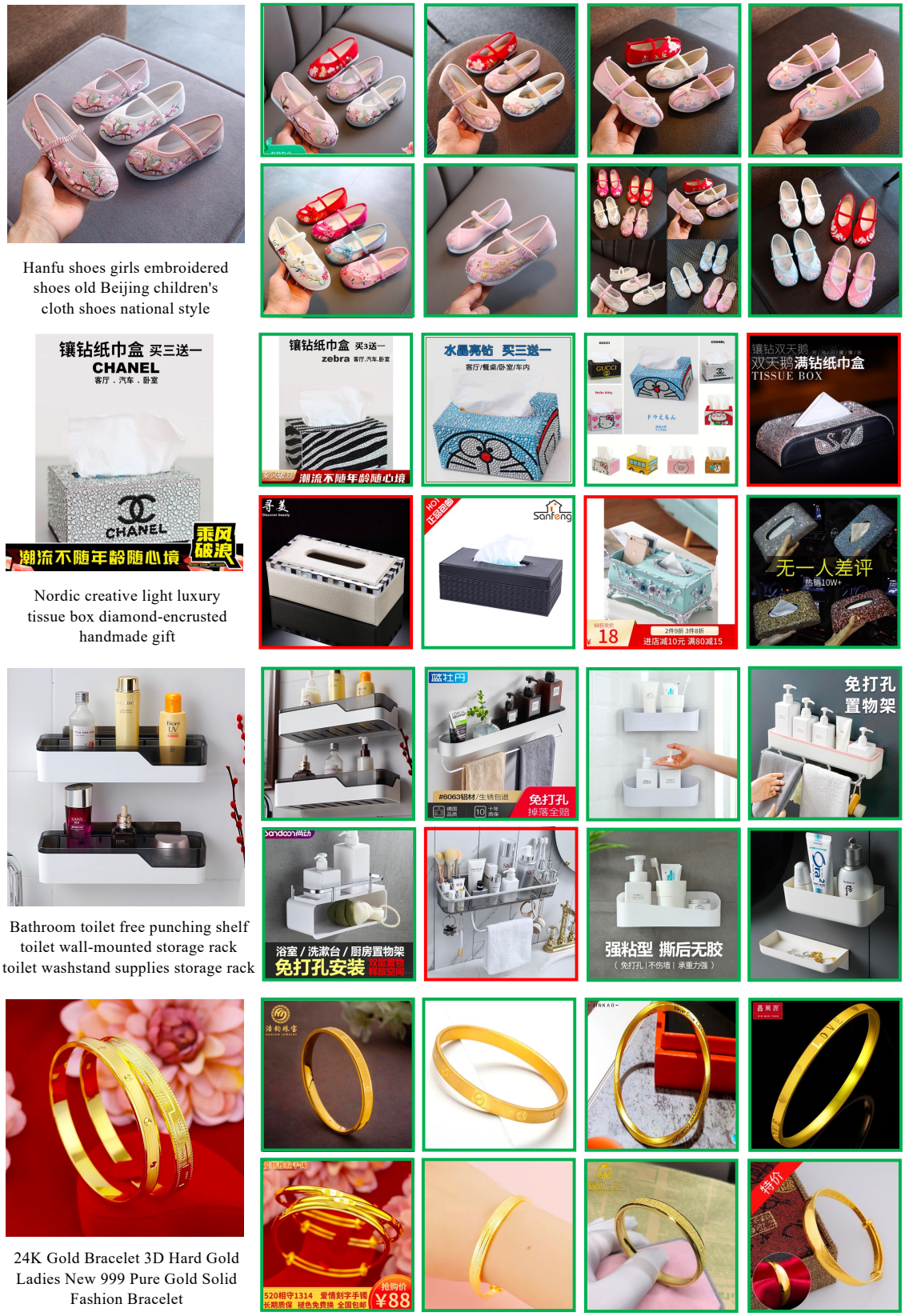

Figure 4. Visualizations of zero-shot coarse-level retrieval results achieved by ECLIP.

Estee Lauder Red Pomegranate Three-piece Skin Care Set Hydrating Moisturizing Nourishing Set

Pechoin Herbal Essence (Essence) Surprise Set Hydrating Nourishing Moisturizing Refreshing Translucent

Caudalie Grape Source Hydrating Beauty Set Spray 75ml+Essence 30ml+Aqua Cream 15ml

Hanokfang duty-free skin key new version of water milk three-piece box moisturizing water day milk night milk

Figure 5. Visualizations of zero-shot instance-level retrieval results achieved by ECLIP.

4. Boarder Impact

This work provides a novel perspective for learning prompt-based visual representation. The unique contributions of ECLIP as follows: (1) This work mainly focuses on E-commerce scenarios. E-commerce oriented model, despite its high practical importance, still remain inadequately studied. We identify the unique challenge by comparing natural / product images, *i.e.*, the gap between the demand for instance-level representation and the lack of box annotations in large-scale raw E-commerce data. (2) The proposed instance decoder innovatively correlates the multi-modal prompt with input queries and adopts the slot-attention to implicitly force each query to attend to a specific image region. The developed two proxy tasks can fully exploit the natural characteristics of E-commerce data itself as supervision. These novel designs collectively enable ECLIP to effectively ground a desired instance (see Figure 3). In contrast, existing VLP models (e.g., X-VLM [10], GLIP [6], MDETR [4], etc) obtain such ability by relying on object-level annotations.

References

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 2

[2] Xiao Dong, Xunlin Zhan, Yangxin Wu, Yunchao Wei, Michael C Kampffmeyer, Xiaoyong Wei, Minlong Lu, Yaowei Wang, and Xiaodan Liang. M5product: Self-harmonized contrastive learning for e-commercial multi-modal pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21252–21262, 2022. 2

[3] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 2

[4] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1780–1790, 2021. 7

[5] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021. 2

[6] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10965–10975, June 2022. 7

[7] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 2

[8] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2

[9] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. Mattnet: Modular attention network for referring expression comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1307–1315, 2018. 2

[10] Yan Zeng, Xinsong Zhang, and Hang Li. Multi-grained vision language pre-training: Aligning texts with visual concepts. *arXiv preprint arXiv:2111.08276*, 2021. 7

[11] Xunlin Zhan, Yangxin Wu, Xiao Dong, Yunchao Wei, Minlong Lu, Yichi Zhang, Hang Xu, and Xiaodan Liang. Product1m: Towards weakly supervised instance-level product retrieval via cross-modal pretraining. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11782–11791, 2021. 2