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

Abstract

This paper aims to establish a generic multi-modal foundation model that has the scalable capability to massive downstream applications in E-commerce. Recently, large-scale vision-language pretraining approaches have achieved remarkable advances in the general domain. However, due to the significant differences between natural and product images, directly applying these frameworks for modeling image-level representations to E-commerce will be inevitably sub-optimal. To this end, we propose an instance-centric multi-modal pretraining paradigm called ECLIP in this work. In detail, we craft a decoder architecture that introduces a set of learnable instance queries to explicitly aggregate instance-level semantics. Moreover, to enable the model to focus on the desired product instance without reliance on expensive manual annotations, two specially configured pretext tasks are further proposed. Pretrained on the 100 million E-commerce-related data, ECLIP successfully extracts more generic, semantic-rich, and robust representations. Extensive experimental results show that, without further fine-tuning, ECLIP surpasses existing methods by a large margin on a broad range of downstream tasks, demonstrating the strong transferability to real-world E-commerce applications.

1. Introduction

Nowadays, the flourishing growth of E-commerce has brought great convenience to people's daily life. And a wide range of product-based application tasks has subsequently emerged, such as item classification [16, 27], product retrieval [7, 32], commodity recommendation [19, 26], and so on. Compared to developing individual task-specific models, building a general-purpose foundation model that works for massive E-commercial applications simultaneously can enhance applicability and reduce training costs.

*Corresponding Author.

Figure 1. Domain difference between natural and product images. For natural images, it is the frequent case that most pixels are semantically correlated to the textual sentence. However, in E-commerce, such correlation is much more sparse (*e.g.*, "frying pan" or "coffee machine" only occupy small portions of the entire images). Moreover, images for a product are often provided in a group from multiple sources such as (a) advertisement videos, (b) product pages, (c) customer comments (see the bottom examples).

Recent developments in vision-language pretraining (VLP) [9, 11, 15, 18, 29] have demonstrated remarkable advances in diverse VL downstream tasks. Profiting from large-scale image-text pairs, these methods are able to learn generic multimodal representations that are reused across various tasks. In E-commerce scenario, the related data naturally contains cross-modal information to describe a corresponding product. Motivated by the tremendous success achieved by VL modeling, several approaches [4, 31, 32, 34] have made attempts at designing a commerce-specific multimodal representation learning paradigm. They imitate the existing VLP methods (e.g., CLIP [18], VilBERT [15]) to learn the *image-level* representations of the product via pretraining on abundant commerce image-text pairs.

Though promising results have been achieved, directly applying these VLP methods in the general domain to E-commerce still suffers from inherent deficiencies. The properties of natural and product images appear to be dramatically different. Given a natural image-text pair, almost every pixel in the natural image is mentioned by the corresponding textual description. In contrast, as shown in Figure 1, in a real E-commerce scenario, the images are mostly product-oriented. Only very few instances are related to the product description. Simply treating the whole image as a monolithic entity to perform cross-modal alignment with text will inevitably confound the foreground and noisy background. Hence, to establish a foundation model that generalizes well to diverse E-commerce applications, it is of great significance to learn the product-related instance-level representation. With this goal in mind, a crucial challenge needs to be addressed: How can we enable the model to focus on the product instance in the presence of background interference?

A straightforward way to tackle this problem would be to resort to object-level human annotations, but it is laborious and infeasible to scale on larger data from the Internet. In this work, we strive to derive the capability of grounding product instances from uncurated data. Our motivation is built on the natural characteristics of E-commerce data itself. As illustrated in Figure 1, a product usually has multiple image samples from different sources (*e.g.*, merchant, customer comments, attached advertisement videos, etc.). Although the appearance of these samples may be diverse due to the changes of camera view or scenes, they all include the identical product entity. This fact strongly spurs us to pursue an instance-centric multi-modal learning paradigm by leveraging such explicit correlation.

The proposed pretraining framework, dubbed as **ECLIP** (E for "E-commerce"), employs two separate encoders to embed the images and texts of products. Our key idea is to develop a decoder architecture built upon the above-mentioned encoders, which aims to aggregate the instance-centric product representations without additional hand-crafted annotation. Inspired by [1,13,28], the decoder introduces a set of learnable tokens that we refer to as *instance query*. At each decoder block, these instance queries are updated via interacting with the encoded visual features. Through the stack of multiple blocks, they will gradually probe the potential product instance from the entire image. Moreover, each instance query is conditioned on a concrete text or image called *multi-modal prompt*. Such a design renders it dedicated to a particular instance type indicated by the content of its associated prompt. Therefore, by specifying the content of multi-modal prompt, the decoder can adaptively discover the corresponding instance. During pretraining, there is only one positive prompt for a given sample. The rest are negative ones sampled from other products.

To effectively optimize the generated instance representations, we newly craft two pretext tasks: inter-product and intra-product multi-modal learning. The first one is in charge of pulling the representations of the identical product closer to each other and pushing away the unmatched ones. It is noteworthy that the appearance of the positive image samples varies a lot except for the presented product. Bringing their representations closer than negative pairs in the feature space will implicitly encourage the instance query to focus on the visual region that corresponds to the desired product. The second one aims to ensure that only positive queries can aggregate the semantics of the foreground instance, rather than negative ones. Coupling these two novel pretext tasks together, we find that the whole framework is capable of learning a generic product representation. Our core contributions can be summarized as follows:

(1) We propose ECLIP, an effective and simple multi-modal representation learning paradigm in the E-commerce scenario. Going beyond regular global representations, it can successfully obtain instance-centric product representations via a decoder architecture.

(2) By fully exploiting the natural characteristics of E-commerce data and the proposed pretext tasks, ECLIP obtains the fine-grained alignment capability to ground the desired product instance (see Figure 4a) without reliance on any manual annotation.

(3) Pre-trained on large-scale product data, the resulting foundation model can seamlessly generalize to downstream E-commerce applications. Comprehensive experimental results further demonstrate the superiority of ECLIP: without any fine-tuning, it achieves substantial improvements over the existing state-of-the-art methods on diverse real-world E-commerce tasks.

2. Related Work

Vision-Language Representation Learning. In recent years, vision-language pretraining (VLP) has attracted the attention of numerous researchers and has been widely explored [6], which aims to learn from tremendous image-text paired data to obtain knowledge that can be generalized to downstream tasks. Some pioneer works (e.g. LXMERT [21], UNITER [2], VinVL [33]) rely on pre-trained object detection modules such as Faster-RCNN [20] to extract visual representations. Later efforts such as ViLT [10] and VLMo [24] unify the vision and language transformers, and train a multimodal transformer from scratch. Then, CLIP [18] and ALIGN [9] demonstrate that dual-encoder models pretrained with contrastive objectives on noisy image-text pairs can learn strong image and text representations for crossmodal alignment tasks and zero-shot image classification. While ALBEF [11] additionally trains a fusion-encoder to jointly learn the multi-modal representations. GLIP [12] unifies object detection and phrase grounding for pretraining and surpasses many baselines in the detection field. Another line of researches [17,23,25,30] develop encoder-decoder models that are trained using generative losses and show strong generation performances in vision-language benchmarks, while the visual encoder still performs competitively on image classification. But most aforementioned VLP methods devote themselves to the coarse correlation between the text and the entire image and ignore the instance-level information, which is critical in an e-commerce scenario (as shown in Figure 1).

MultiModal Pre-training for E-commerce. Early works like FashionBERT [7], Kaleido-BERT [35] leverage a transformer-based model and tailored masking strategy to perform pretraining to generate more fine-grained features for cloth retrieval. Then CAPTURE [32] generates discriminative instance features via masked multi-modal learning as well as cross-modal contrastive pretraining achieve surprising performance in the instance-level product retrieval task. K3M [34] further introduces the knowledge modality in multi-modal pretraining to correct the noise and supplement the missing of image and text modalities. SCALE [4] proposes a self-harmonized contrastive learning framework that can integrate six different modalities into a unified model. Recent CommerceMM in [31] design a contrastive and MLM-based pretraining paradigm on 14 different tasks. However, all the existing methods only consider the global alignment between images and text, without exploring the special characteristic contained in the e-commercial data for learning instance-centric representation.

3. Approach

In this section, we begin with the overview of our proposed ECLIP in Section 3.1. Then the core decoder architecture that aims at aggregating instance-level representation of the desired product is introduced in Section 3.2. To optimize the entire framework, we carefully designed several pretraining objectives in Section 3.3. Finally, we delineate how to transfer the resulting foundation model to various downstream tasks in Section 3.4.

3.1. Model Overview

As illustrated in Figure 2, ECLIP is composed of an image encoder, a text encoder, and an instance decoder. Given an input sample $x = (x^I, x^T)$, where x^I and x^T are image and text describing the corresponding product information, respectively. These two encoders first encode the image-text pair as a sequence of feature embeddings. Then, a modality-dependent projection layer is employed to map them linearly into a joint multi-modal feature space. These projected embeddings are further decoded to produce an instance-centric representation. Details of the two unimodal encoders are elaborated as follows.

Image Encoder. Following the vision transformers [5], the product image $x^I \in \mathcal{R}^{H\times W\times C}$ is partitioned into N non-overlapping patches. These patches are flattened to 1D input tokens and then projected linearly, with position embeddings added. Through hierarchical feature encoding, we can obtain a sequence of visual embeddings $\{v_{cls}, v_1, ..., v_N\}$, where v_{cls} indicates the special token [CLS] for encoding the entire image information.

Text Encoder. This encoder adopts analogous transformer-style architecture. For the input product description x_T, it tokenizes the text to M subwords as in BERT [3]. Similar to the image encoder, a special [CLS] token is appended to the beginning of the textual input to summarize the text semantics. After encoding, the resulting linguistic embedding sequence is denoted as $\{w_{cls}, w_1, ..., w_M\}$.

3.2. Extract Instance-Centric Representation

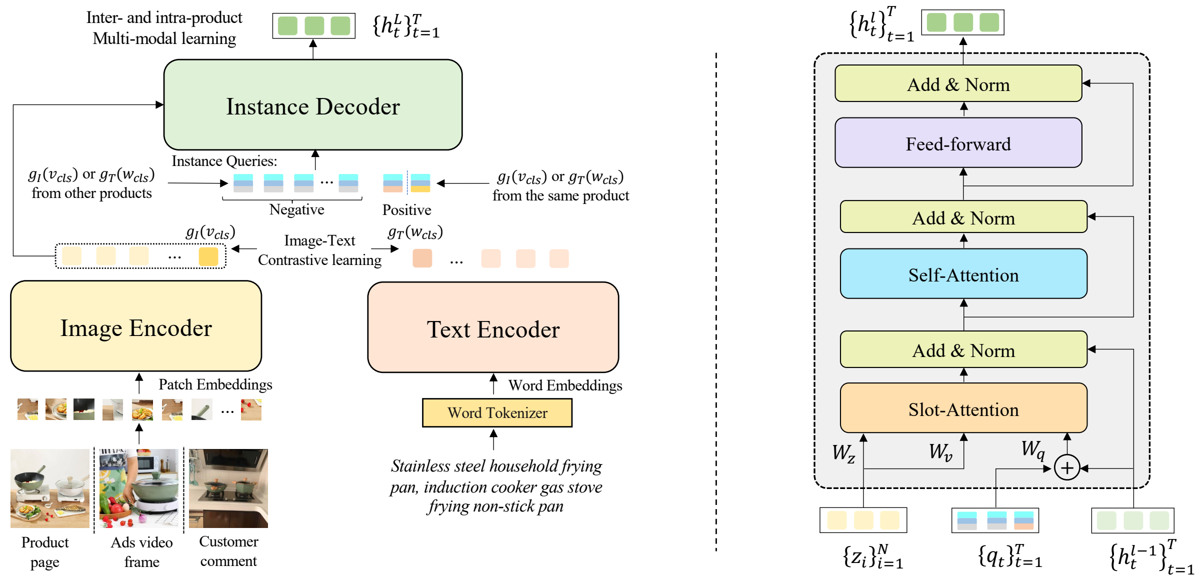
After obtaining the contextualized embeddings, existing VLP approaches leverage $g_I(v_{cls}) \in \mathcal{R}^D$ and $g_T(w_{cls}) \in \mathcal{R}^D$ to align positive image-text pairs via contrastive learning. Here $g_I(\cdot)$ and $g_T(\cdot)$ are the aforementioned projections. While effective in the general domain, this design only considers the alignment between the global image-text semantics. However, in the E-commerce image, only several regions containing the desired product instance are informative foregrounds corresponding to the text description. Modeling such image-level alignment will fail to learn strong and robust product semantics. Hence, we are committed to learning instance-centric representation.

Instance Query. Motivated by [1, 13], a set of learnable tokens called *instance query* are introduced to ground the potential instance in the product image. As Figure 2 shows, each query is correlated to a specific text or image that we refer to as *multi-modal prompt*. The insight behind this design is that we wish the instance that a query should probe to be specified by the prompt content. Formally, the proposed instance query is denoted as $\boldsymbol{Q} = \{q_t \in \mathcal{R}^D\}_{t=1}^{T}$, which can be obtained by:

$$q_t = q_t^{\text{prompt}} + q_t^{\text{pos}} + q_t^{\text{type}} \tag{1}$$

Here, q_t^{prompt} denotes $g_I(v_{cls})$ or $g_T(w_{cls})$, q_t^{pos} and q_t^{type} are learnable position and type embedding, indicating the probing area of a query and the modality type of the binding prompt. These queries are responsible for aggregating the instance-centric representations $\boldsymbol{H} = \{h_t\}_{t=1}^{T}$ from the encoded visual features via a decoder architecture. During pretraining, there is only one positive prompt (w.r.t the same product) for a given sample, and the rest $T-1$ are negative ones sampled from other products.

Instance Decoder. We first project all the encoded $\{v_i\}_{i=1}^{N}$ into the same feature space as the prompt, yielding an embedding sequence $\boldsymbol{Z} = \{z_i \in \mathcal{R}^D\}_{i=1}^{N}$. Moreover, the instance representations $\boldsymbol{H}^0$ are zero-initialized, and then before feeding to the decoder. The proposed decoder then



(a) The overall Pipeline

(b) The l-th Decoder Block

Figure 2. (a) The architecture of the proposed instance-level representation learning paradigm (ECLIP), which is composed of an image encoder, a text encoder, and an instance decoder. The whole framework is optimized by three pretext tasks when pretraining on a large-scale E-commerce dataset. (b) The core decoder architecture that aims at aggregating the instance representation of the desired product.

reads all the above-described embeddings: $\boldsymbol{Z}$, $\boldsymbol{Q}$ and $\boldsymbol{H}^0$ as its input. It has L stacked blocks, and each one consists of a slot-attention and a self-attention layer.

The goal of the slot-attention layer is to adaptively update query representations through the interaction with the encoded visual embeddings. In detail, for the l-th slot-attention layer, it first calculates a similarity matrix $M \in \mathcal{R}^{N\times T}$, which is implemented by the dot-product attention mechanism [22]. Formally, it is formulated by:

$$M = \frac{1}{\sqrt{D}}(ZW_z)\cdot((Q+H^{l-1})W_q)^\top \tag{2}$$

where W_z and W_q are the learnable projections parameter matrices, $\boldsymbol{H}^{l-1}$ is the instance representation produced by the $(l-1)^{\text{th}}$ decoder block. The similarity matrix M is further normalized by a softmax function over T queries:

$$M_{ij} = \frac{\exp(M_{ij})}{\sum_{t=1}^{T}\exp(M_{it})} \tag{3}$$

The generated matrix M actually performs soft assignment via computing semantic similarity between N visual tokens and T instance queries. In doing so, it is capable of distributing each visual token to a specific query according to their similarity score. To aggregate the information of the visual tokens into their assigned input query, we compute a weighted mean update based on M:

$$\Delta h_t^{l-1} = \frac{1}{\sum_{i=1}^{N} M_{it}}\sum_{i=1}^{N} M_{it}(W_v z_i) \tag{4}$$

Finally, the instance representation $\boldsymbol{H}^l$ at the l-th layer can be updated by a residual connection:

$$h_t^l = h_t^{l-1} + W_o \Delta h_t^{l-1} \tag{5}$$

where W_v and W_o are linear tranformation parameters.

On the top of slot-attention layer, there is a self-attention module that performs information propagation between each query. In detail, given the previously updated $\boldsymbol{H}^l$, it employs standard multi-head self-attention (MSA) followed by a fully connected feed-forward network as in [22]. After L successive decoder blocks, we can obtain the final instance representation $\boldsymbol{H}^L$. It is noteworthy that, since the multi-modal prompts just participate in the similarity calculation in Eq. 2, the resulting $\boldsymbol{H}^L$ thus contains only visual-modality information.

Discussion: The proposed decoder works like conducting a clustering on the image tokens, where each instance query serves as the centroid of a cluster. In each decoder block, it determines where each token belongs by measuring its distance from the centroids in the semantic space. The cluster centroid is then updated via a soft manner (Eq.4) based on the calculated distance. By stacking multiple decoder blocks, it can implicitly force each query to attend to a specific region and aggregate instance-level representations.

3.3. Multi-Modal Pretraining Objectives

Our ECLIP is optimized on large-scale uncurated product data with several pretraining proxy tasks. In the following, we describe each task in detail.

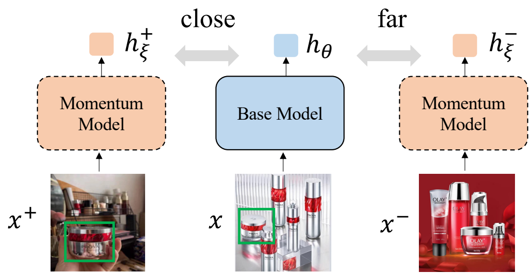


Figure 3. The illustration of inter-product contrastive learning. It aims to bring the positive samples from the same product (x and x^+) closer than the negative ones (x and x^-), which contributes to grounding the desired product (outlined with the green box).

Image-Text Contrastive Learning. As in [9, 11, 18], this task contributes to learning better unimodal representations. Given a batch of product samples $\{(x_i^I, x_i^T)\}_{i=1}^B$, the similarity between image x^I and text x^T is estimated as:

$$s(x^I, x^T) = g_I(v_{cls})^\top g_T(w_{cls}).$$

This pretraining objective bring the image-text pairs of the same product in the embedding space closer than the unmatched ones, which consists of image-to-text term $\mathcal{L}_{i2t}$:

$$\mathcal{L}_{i2t} = -\sum_{i=1}^{B} \log \frac{\exp(s(x_i^I, x_i^T)/\tau)}{\sum_{j=1}^{B} \exp(s(x_i^I, x_j^T)/\tau)}, \quad (6)$$

and a text-to-image term $\mathcal{L}_{t2i}$:

$$\mathcal{L}_{t2i} = -\sum_{i=1}^{B} \log \frac{\exp(s(x_i^T, x_i^I)/\tau)}{\sum_{j=1}^{B} \exp(s(x_i^T, x_j^I)/\tau)}, \quad (7)$$

where τ is a learnable temperature parameter. The whole objective is then defined as $\mathcal{L}_{itc} = \frac{1}{2}(\mathcal{L}_{i2t} + \mathcal{L}_{t2i})$.

Inter-Product Multi-modal Learning. As shown in Figure 3, we maintain a momentum model during pretraining that is an exponential-moving-average of the origin model like [8]. For a product sample x_i, we denote the representation of the positive prompt produced by base and momentum model as h_θ^i and h_ξ^i. An inter-product contrastive loss $\mathcal{L}_{inter}$ is computed by:

$$\mathcal{L}_{inter} = -\sum_{i=1}^{B} \log \frac{\exp({h_\theta^i}^\top h_\xi^j/\tau)}{\exp({h_\theta^i}^\top h_\xi^j/\tau) + \sum_{k\in\mathcal{N}^-} \exp({h_\theta^i}^\top h_\xi^k/\tau)}, \quad (8)$$

where sample i and j are a positive pair, $\mathcal{N}^-$ is a negative sample set that belongs to other products. This objective maximizes the similarity between different samples of the identical product, while minimizing those of the unmatched ones. Since the images of a product are collected from different sources, their background appearances are usually diverse. Hence, $\mathcal{L}_{inter}$ will encourage the yielded representation to be highly correlated with the desired product and thus contributes to aligning the positive prompt with the corresponding image tokens in a fine-grained manner.

This pretext task also incorporates an additional instance-text matching loss that predicts whether an instance and a text description is matched. Formally, given an instance-text pair, we obtain their match logits is defined by: $f(h_\theta^i \odot g_T(w_{\text{cls}}^i))$, where $\odot$ is Hadamard product and $f(\cdot)$ is a mapping layer: $\mathcal{R}^D \to \mathcal{R}^2$. This match logits is optimized by a typical binary cross entropy loss $\mathcal{L}_{itm}$.

Intra-Product Multi-modal Learning. For a product sample, there is only one positive prompt describing the presented product during pretraining, and the rest $T-1$ prompts are sampled from other products. The motivation behind this pretext task is to ensure that only positive queries can probe the foreground instance, rather than negative ones. To this end, we apply an intra-product contrastive loss using text supervision. Suppose index r indicates the positive query, then $\mathcal{L}_{intra}$ can be formulated as:

$$\mathcal{L}_{intra} = -\sum_{i=1}^{B} \log \frac{\exp({h_r^i}^\top g_T(w_{\text{cls}}^i)/\tau)}{\sum_{t=1}^{T} \exp({h_t^i}^\top g_T(w_{\text{cls}}^i)/\tau)}, \quad (9)$$

which serves to bring the positive query and the product description closer than all $T-1$ negative ones. Moreover, we also introduce an entropy regularization term for M:

$$\mathcal{L}_{\mathcal{R}} = -\sum_{i=1}^{N}\sum_{j=1}^{T} M_{i,j} \log(M_{i,j}), \quad (10)$$

which prevents too smooth similarity score over all queries. Finally, the overall pretraining objective of ECLIP is the sum of all aforementioned loss terms.

3.4. Transfer to Downstream Tasks

Once pretrained, the resulting foundation model can be leveraged to extract the product instance representation with minimal surgery. Specifically, given a product sample (x_i^I, x_i^T), we first encode the image-text pair into an embedding sequence via the unimodal encoders. Then, the global representation of text description $g_T(w_{\text{cls}})$ is treated as the positive query and fed into the decoder concatenated with $T-1$ negative ones. Here, the negative queries $\{q_t\}_{t=2}^T$ are sampled from a standard Gaussian distribution for convenience. We also explore different negative query setting manners in Section 4.3. The yielded representation h_0^L belonging to the positive query is then applied to a wide range of E-commerce downstream tasks.

4. Experiments

4.1. Pretraining Details

Pretraining Dataset We collected a large-scale pretraining dataset from a popular E-commerce website. It consists of

15M different products and over 100M various images, covering about 9K diverse categories such as clothes, daily necessities, instruments, and so on. For each product item, it has a corresponding textual description and several images from the product details pages, customer comments, and attached advertisement videos. During pretraining, the positive data pairs are constructed by sampling images belonging to the same product from different sources.

Implementation Details. The image encoder adopts the same network configuration as the standard ViT [5] and is initialized from the weight pre-trained on ImageNet. Our text encoder is implemented with the same architecture as $\text{BERT}_{\text{base}}$ [3]. The decoder has 6 identical blocks and 20 instance queries. We here explore two variants of ViTs: ViT-B/16 and ViT-L/16. There is a total of 250 / 480 million parameters for the base and large version. During pretraining, the input images are resized to 224×224 with random crop and horizontal flip augmentation, and the texts are tokenized by WordPiece with a maximum length of 55. We pretrain for 15 epochs with a batch size of 4096 (ViT-B) / 3072 (ViT-L) on 32 NVIDIA A100 GPUs. The whole framework is learned using AdamW [14] optimizer and the learning rate is warmed up to 1e-4 and then decayed linearly. More details are elaborated in the supplementary.

Compared Baselines. We mainly compare ECLIP with two state-of-the-art VLP methods: CLIP [18] and ALBEF [11], which is designed to learn global representation. For fair comparison, we also leverage ViT-B/16 as the image encoder, $\text{BERT}_{\text{base}}$ as the text encoder, and pretrain these baselines on the same 100M E-commerce data using their public official implementations.

4.2. Evaluation on Downstream Tasks

Next, we delineate the evaluation performances for five specific E-commerce downstream tasks in turn.

4.2.1 Zero-Shot / Finetuned Product Classification

We first transfer ECLIP to product item classification. It is a recognition task that aims to map a product sample to a specific category. We evaluate the performance on a large-scale publicly available E-commerce dataset called M5Product [4], which covers 1.1M images and 5,679 various product categories. Here, we consider two different settings: image and multimodal classification. The former uses only the image-modality representation and the latter uses the mulimodal representation from image and text.

To demonstrate the strong zero-shot ability, we apply ECLIP directly to the classification evaluation without further finetune. It is achieved by measuring the similarity between the category text like CLIP [18]. Table 1 summarizes the comparison results. As presented, our ECLIP greatly exceeds all the existing baselines by a large margin (e.g., +10.16% v.s. CLIP), demonstrating the superiority of instance-level representation. We also test the finetuned performance compared to the existing pretrained approaches in E-commerce [4]. It can be observed in Table 1 that ECLIP still achieves substantial performance gains.

Setting	Methods	Modality	
		Image	Image + Text
Zero-Shot	CLIP [18]	25.27	38.26
	ALBEF [11]	27.36	43.13
	$\text{Ours}_{\text{ViT-B/16}}$	**35.43**	**47.42**
Finetuned	ViLBERT [15]	-	82.21
	UNITER [2]	-	82.37
	SCALE [4]	30.14	83.68
	CLIP [18]	69.56	82.85
	ALBEF [11]	71.38	83.79
	$\text{Ours}_{\text{ViT-B/16}}$	**83.38**	**84.67**

Table 1. Performance comparisons of zero-shot and finetuned product classification on M5Product benchmark [4].

4.2.2 Zero-Shot Image-Text Retrieval

ECLIP is also transferred to test zero-shot performance for image-to-text and text-to-image retrieval. To this end, we collect a large dataset that contains 205K image-text pairs of E-commerce products. Since only unimodal information is available in this task, we simply use our image and text encoders to embed image-text pairs and complete the retrieval based on their pairwise similarity. We utilize the widely-used Recall@K metric for evaluation. Detailed comparison results are shown in Table 2. We can see that, despite training on the same dataset, our method achieves superior performance owing to fine-grained alignment modeling between text and product instance.

Method	Image-to-Text			Text-to-Image		
	R@1	R@5	R@10	R@1	R@5	R@10
CLIP [18]	47.31	69.44	74.98	57.88	82.03	88.15
ALBEF [11]	48.54	70.82	75.34	58.10	82.16	88.63
$\text{Ours}_{\text{ViT-B/16}}$	49.81	71.34	76.91	58.46	**83.80**	**89.43**
$\text{Ours}_{\text{ViT-L/16}}$	**52.72**	**74.70**	**80.35**	**59.48**	83.32	86.62

Table 2. Performance comparisons of zero-shot image-to-text and text-to-image retrieval tasks.

4.2.3 Zero-Shot Product Retrieval

This task aims to find the most relevant target product given a query. It has a wide range of applications in real e-commerce scenarios such as recommending relevant products for users. We first evaluate the coarse-level retrieval. Following [4], a procut pair is considered a match if both belong to the same category during evaluation. The results on M5Product is reported on the left of Table 3. It can be seen that exploiting instance-centric representation significantly boost the performance. Especially for retrieval using

Method	Zero-Shot Product Retrieval								
	Coarse-Level Retrieval			Instance-Level Retrieval					
	mAP@1	mAP@5	mAP@10	mAP@10	mAP@50	mAP@100	mAR@10	mAR@50	mAR@100
Image+Text Modality:									
ViLBERT [15]	58.87	61.74	60.06	70.11	68.19	68.29	29.05	25.54	25.02
UNITER [2]	58.85	62.86	60.92	74.69	71.02	70.93	29.47	25.82	26.20
CAPTURE [32]	-	-	-	79.36	74.79	74.63	34.69	30.04	30.08
SCALE [4]	59.81	64.13	62.18	-	-	-	-	-	-
Ours$_{\text{ViT-B/16}}$	**68.82**	**73.95**	**71.41**	**90.30**	**87.72**	**86.31**	**54.43**	**61.54**	**66.28**
Image Modality:									
CAPTURE [32]	-	-	-	40.35	36.77	34.76	17.20	15.86	15.45
CLIP [18]	49.86	55.87	53.46	85.86	80.99	78.54	55.24	60.66	62.03
ALBEF [11]	51.18	57.40	54.88	88.41	83.25	80.52	55.94	60.04	62.29
Ours$_{\text{ViT-B/16}}$	**63.15**	**68.71**	**65.90**	89.62	**86.71**	**85.03**	**55.86**	**62.43**	66.23
Ours$_{\text{ViT-L/16}}$	62.64	68.18	65.58	**89.84**	86.33	84.85	55.74	62.42	**66.88**

Table 3. Performance comparisons of zero shot product retrieval task on M5Product [4] and Product1M [32] benchmarks.

Setting	Methods	IoU Thresh	
		Acc@0.5	Acc@0.7
Zero-Shot	CLIP [18]	80.14	74.23
	ALBEF [11]	80.57	74.08
	Ours$_{\text{ViT-B/16}}$	**85.29**	**80.83**
	Ours$_{\text{ViT-L/16}}$	84.19	80.13

Table 4. Performance comparisons of zero-shot visual grounding.

Setting	Methods	IoU Thresh	
		mAP@0.3	mAP@0.5
Finetuned	DETR [1]	16.63	14.11
	CLIP [18]	17.18	14.28
	ALBEF [11]	17.37	14.50
	Ours$_{\text{ViT-B/16}}$	**19.99**	**16.38**

Table 5. Performance comparisons of object detection.

only visual modality, ECLIP achieves a substantially improvement (63.15% v.s. 49.86% (CLIP) on mAP@1).

We also consider another setting introduced in [32], called instance-level retrieval, where the query image encompasses multiple different kinds of product instances. And the model needs to find all the related products from a large gallery. We evaluate the retrieval results using mean Average Precision (mAP@K) and Average Recall (mAR@N) on the Product1M [32]. As shown in Table 3, ECLIP still achieves superior performance than all the previous approaches. Although the CAPTURE leverages a specially-trained object detector to extract instances, ECLIP still surpasses it by a clear margin with no box annotation.

4.2.4 Zero-Shot Visual Grounding

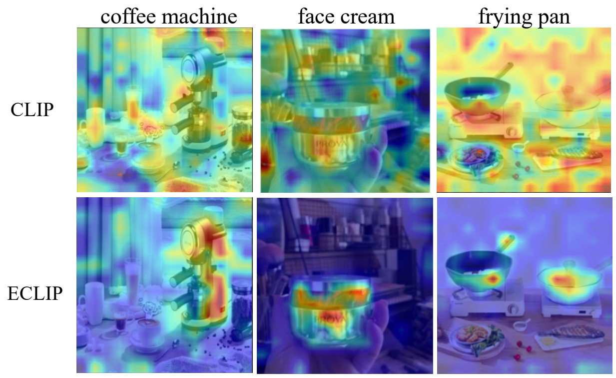
To demonstrate whether our model possesses the capability of localizing the desired product instance after pretraining, we further evaluate ECLIP on the zero-shot product grounding, which requires to localize the product instance in an image according to a textual description. Specifically, the input image-text pair is first fed to our ECLIP to obtain a score map $\mathcal{S} \in \mathcal{R}^{H\times W}$ that measures the similarity between the text and each image location. Then, we use $\mathcal{S}$ to rank the candidate regions produced by an off-the-shelf region proposal network. The performance is evaluated by the top-1 accuracy at IoU thresholds $\{0.5, 0.7\}$ on an annotated grounding dataset consisting of 450K product images. Detailed comparison results are listed in Table 4. As we can see, compared to methods aimed at global representation, our model has learned fine-grained cross-modal understanding ability and thus performs better. Since ECLIP supports image prompt during pretraining, we also conduct zero-shot image-conditioned grounding. The results and analysis can be found in the supplementary.

4.2.5 Transfer to Object Detection

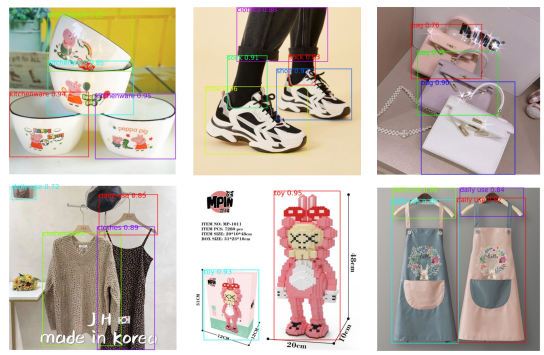
We also transfer ECLIP to object detection to further validate its fine-grained understanding ability. Following DETR [1], we utilize the image encoder to embed visual features, and utilize the decoder with a newly added prediction head to decode the potential objects. Moreover, we collect a manually annotated detection dataset covering 160K images. We split a 20K subset for evaluation and leave the rest for model finetuning. The supplementary provides experiment details of baselines and ECLIP. It can be observed from Table 5 and Figure 4b that it outperforms the existing VLP methods, demonstrating the superiority of ECLIP to learning fine-grained semantics in E-commerce.

4.3. Ablation Study

Effect of Pretext Task. To demonstrate the effectiveness of the inter- and intra- product learning pretext task, we designed the experiments on different task combinations for the product retrieval task. All the ablations were conducted



(a) Visualization of zero-shot visual grounding.



(b) Object detection examples after finetune.

Figure 4. Qualitative examples of visual grounding and object detection. See the main text for more explanation.

on a smaller pretraining dataset that includes only 5M images due to the costly training time. The complete results are listed in Table 6. One can observe that canceling either of these two tasks will result in worse performance. Notably, the inter-product task brings a more significant performance boost compared to the intra-product. We speculate that because the former contrasts with more negative samples from different product images.

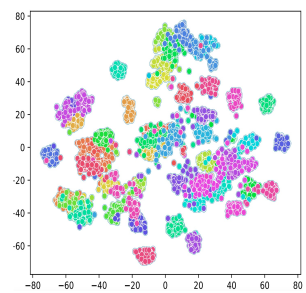
Inter-product	Intra-product	mAP@1 (%)	mAP@5 (%)
		61.63	65.24
✓		64.37	68.79
	✓	62.17	66.58
✓	✓	**65.23**	**69.26**

Table 6. Ablation of different pretraining pretext task combinations on coarse-level product retrieval task (ViT-B/16).

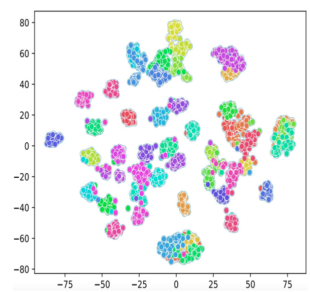
Metric	mAP@10	mAR@10
Negative Text	86.85	55.97
EMA Update	87.12	54.18
Random Sampling	**87.90**	**56.06**

Table 7. Ablation of different negative query setting manners on the instance-level product retrieval task (ViT-B/16).

Effect of Negative Query. Since the negative instance queries are used when transferring to classification and retrieval tasks, we also ran ablations on different ways of setting these negative ones. We tried the following cases: 1) Leverage the description text from other products and encode with the text encoder. 2) Randomly sample the negative queries from the standard Gaussian distribution. 3) Adopt the exponential moving average of queries on the whole dataset during pretraining. Table 7 summarizes the results of this ablation study on the product retrieval task. As observed, there is little to no difference between different manners of setting up negative queries. We thus adopt random sampling for implementation simplicity.



(a) CLIP



(b) ECLIP

Figure 5. The T-SNE visualization for the learned representations of CLIP and our ECLIP. Both of them are pretrained on the same dataset. We first randomly select 50 kinds of products and then draw 100 samples for each. Points belonging to the same category are of the same color. See the main text for more explanation.

4.4. Qualitative Analysis

In this section, we first qualitatively showcase that ECLIP can learn fine-grained cross-modal alignment to ground the desired product. Figure 4a presents the visualization of the similarity score map between a product image and its corresponding text description, where darker color indicates image locations with higher similarity to text. We can clearly observe that our model can rightly attend to the desired instance depicted by text. Furthermore, T-SNE is used for visualizing the visual embeddings of different kinds of product samples. As shown in Figure 5, compared to CLIP, our ECLIP can extract semantic-rich yet compact representations that better distinguish products belonging to different categories. More visualization examples and analysis are provided in our supplementary.

5. Conclusion

In this paper, we develop an effective large-scale multimodal pretraining paradigm called ECLIP in E-commerce. Beyond regular global representation, it instead aims to learn the instance-level representation via a novel decoder and the carefully-designed pretraining proxy tasks. Extensive experimental results further demonstrate the superior generalization capability of the proposed framework.

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, 3, 7

[2] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020. 2, 6, 7

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 3, 6

[4] 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. 1, 3, 6, 7

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3, 6

[6] Zhe Gan, Linjie Li, Chunyuan Li, Lijuan Wang, Zicheng Liu, and Jianfeng Gao. Vision-language pre-training: Basics, recent advances, and future trends. *arXiv preprint arXiv:2210.09263*, 2022. 2

[7] Dehong Gao, Linbo Jin, Ben Chen, Minghui Qiu, Peng Li, Yi Wei, Yi Hu, and Hao Wang. Fashionbert: Text and image matching with adaptive loss for cross-modal retrieval. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2251–2260, 2020. 1, 3

[8] 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. 5

[9] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021. 1, 2, 5

[10] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR, 2021. 2

[11] 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. 1, 2, 5, 6, 7

[12] 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. 2

[13] Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. *Advances in Neural Information Processing Systems*, 33:11525–11538, 2020. 2, 3

[14] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[15] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019. 1, 6, 7

[16] Mieczysław Pawłowski. Machine learning based product classification for ecommerce. *Journal of Computer Information Systems*, 62(4):730–739, 2022. 1

[17] AJ Piergiovanni, Wei Li, Weicheng Kuo, Mohammad Saffar, Fred Bertsch, and Anelia Angelova. Answer-me: Multi-task open-vocabulary visual question answering. *arXiv preprint arXiv:2205.00949*, 2022. 2

[18] 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. 1, 2, 5, 6, 7

[19] Puji Rahayu, Dana Indra Sensuse, Betty Purwandari, Indra Budi, F Khalid, and N Zulkarnaim. A systematic review of recommender system for e-portfolio domain. In *Proceedings of the 5th International Conference on Information and Education Technology*, pages 21–26, 2017. 1

[20] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 2

[21] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019. 2

[22] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 4

[23] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Unifying architectures, tasks, and modalities

through a simple sequence-to-sequence learning framework. *arXiv preprint arXiv:2202.03052*, 2022. 2

[24] Wenhui Wang, Hangbo Bao, Li Dong, and Furu Wei. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *arXiv preprint arXiv:2111.02358*, 2021. 2

[25] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*, 2021. 2

[26] Chi-Man Wong, Fan Feng, Wen Zhang, Chi-Man Vong, and Chen. Improving conversational recommender system by pretraining billion-scale knowledge graph. In *International Conference on Data Engineering (ICDE)*, pages 2607–2612, 2021. 1

[27] Hu Xu, Bing Liu, Lei Shu, and P Yu. Open-world learning and application to product classification. In *The World Wide Web Conference*, pages 3413–3419, 2019. 1

[28] Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. Groupvit: Semantic segmentation emerges from text supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18134–18144, 2022. 2

[29] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: Fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783*, 2021. 1

[30] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022. 2

[31] Licheng Yu, Jun Chen, Animesh Sinha, Mengjiao Wang, Yu Chen, Tamara L Berg, and Ning Zhang. Commercemm: Large-scale commerce multimodal representation learning with omni retrieval. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4433–4442, 2022. 1, 3

[32] 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. 1, 3, 7

[33] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5579–5588, 2021. 2

[34] Yushan Zhu, Huaixiao Zhao, Wen Zhang, Ganqiang Ye, Hui Chen, Ningyu Zhang, and Huajun Chen. Knowledge perceived multi-modal pretraining in e-commerce. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 2744–2752, 2021. 1, 3

[35] Mingchen Zhuge, Dehong Gao, Deng-Ping Fan, Linbo Jin, Ben Chen, Haoming Zhou, Minghui Qiu, and Ling Shao. Kaleido-bert: Vision-language pre-training on fashion domain. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12647–12657, 2021. 3