

# Cross-Domain Product Representation Learning for Rich-Content E-Commerce

Xuehan Bai\* Yan Li\* Yanhua Cheng Wenjie Yang† Quan Chen† Han Li  
Kuaishou Technology

{baixuehan03, liyan26, chengyanhua, chenquan06, lihan08}@kuaishou.com, wenjie.yang@nlpr.ia.ac.cn

<https://github.com/adxcreative/COPE>

## Abstract

The proliferation of short video and live-streaming platforms has revolutionized how consumers engage in online shopping. Instead of browsing product pages, consumers are now turning to rich-content e-commerce, where they can purchase products through dynamic and interactive media like short videos and live streams. This emerging form of online shopping has introduced technical challenges, as products may be presented differently across various media domains. Therefore, a unified product representation is essential for achieving cross-domain product recognition to ensure an optimal user search experience and effective product recommendations. Despite the urgent industrial need for a unified cross-domain product representation, previous studies have predominantly focused only on product pages without taking into account short videos and live streams. To fill the gap in the rich-content e-commerce area, in this paper, we introduce a large-scale **c**Ross-**d**Omain **P**roduct **r**ecognition dataset, called **ROPE**. **ROPE** covers a wide range of product categories and contains over 180,000 products, corresponding to millions of short videos and live streams. It is the first dataset to cover product pages, short videos, and live streams simultaneously, providing the basis for establishing a unified product representation across different media domains. Furthermore, we propose a **C**ross-**d**Omain **P**roduct **r**epresentation framework, namely **COPE**, which unifies product representations in different domains through multimodal learning including text and vision. Extensive experiments on downstream tasks demonstrate the effectiveness of **COPE** in learning a joint feature space for all product domains.

## 1. Introduction

Recently, with the vicissitude in time spent by customers on entertainment media, the way consumers shop online

\*Equal contribution.

†Corresponding authors.

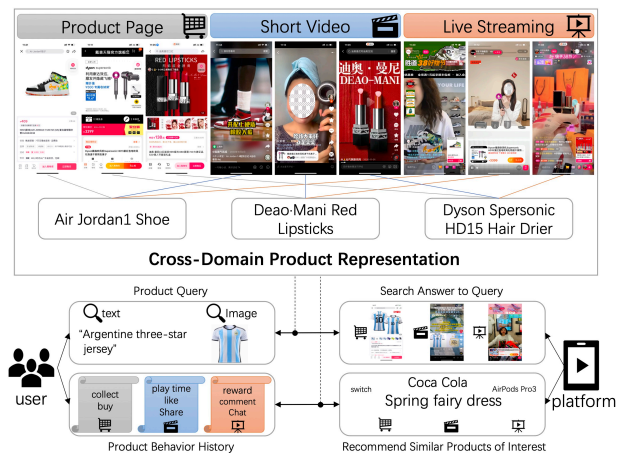


Figure 1. Illustrate the importance of cross-domain product representation for rich-content e-commerce. There are two solid demands for such new e-commerce: 1) The platform needs to provide accurate product results of product page, short video, and live streaming corresponding to user’s query; 2) The platform is able to recommend similar products of interest according to user’s behavior history. Both of the two tasks highly depend on a high-performance cross-domain product representation. Examples are from the popular rich-content e-commerce platforms, including TikTok, Kwai and Taobao.

has transformed significantly, and *rich-content e-commerce* is becoming increasingly popular. In the rich-content e-commerce area, the products are sold not only with traditional product pages but also with dynamic and interactive media formats, *i.e.*, short videos and live streams. As a result, consumers are increasingly relying on these formats to make informed purchase decisions. This shift has facilitated a more engaging shopping experience, bridging the gap between consumers and sellers while presenting new opportunities for platforms to capitalize on.

Despite the advantages of rich-content e-commerce, it presents several technical challenges. One of the most significant challenges is the inconsistency in product presentation across different media domains. For instance, a product may appear entirely different in a live stream than on

a traditional product page. Establishing a unified product representation across different domains is curial and desperately needed in industrial scenarios to address the inconsistency problem. As shown in Figure 1, when users search for a particular product, the unified product representation ensures an enjoyable search experience that the returned product pages, short videos, and live streams precisely describe the same product. When the platform recommends products for users, the unified product representations are beneficial to exploiting users’ consuming behaviors in different media for comprehensive product recommendations.

In spite of the urgent industrial need for a unified cross-domain product representation, prior efforts have concentrated solely on the product page domain. The most common way to learn the product representations is to train a product classification model with product images and titles [12, 14, 26, 27]. However, such representations are far from acceptable in rich-content e-commerce. Specifically, the pictures displayed on the product pages are generally well-shot by professionals, while in short videos and live streaming, the posture of the products and the positions they occupy in the scene often change a lot. Moreover, in live streams and short videos, it is not always guaranteed that products are visible at every moment. Short videos may be mixed with the story plot, while live streams may contain chats between the sellers and their audiences. These contents are generally irrelevant to the products. To bridge this gap and push forward the related research, we collect a large amount of real data from online shopping platforms and present the first large-scale **c**Ross-**d**Omain **P**roduct **r**Ecognition dataset, ROPE. Our dataset contains 3,056,624 product pages, 5,867,526 short videos, and 3,495,097 live streams of 189,958 different products. It covers all product categories of online shopping scenarios. To the best of our knowledge, ROPE is the rich-content e-commerce dataset, including product pages, short videos, and live streams. We hope that the publication of ROPE will attract more researchers to the field of content commerce and drive the development of related technologies.

In addition to the ROPE dataset, we propose a **C**ross-**d**Omain **P**roduct **r**Epresentation baseline, COPE, that maps product pages, short videos, and live streams into the same feature space to build a unified product representation. Based on the ROPE dataset, we evaluate the COPE model on the cross-domain retrieval and few-shot classification tasks. The experimental results show significant improvement over the existing state-of-the-arts.

In summary, our contributions are as follows:

1) As far as we know, our work is the first exploration that tries to build a unified product representation across the product pages, short videos, and live streams to meet the urgent industrial need of the emerging rich-content e-commerce.

2) We collect realistic data from online e-commerce platforms and build a large-scale **c**Ross-**d**Omain **P**roduct **r**Ecognition dataset, ROPE. It contains 3,056,624 product pages, 5,867,526 short videos, and 3,495,097 live streams belonging to 189,958 different products. The included product categories cover the full spectrum of online shopping scenarios.

3) A **C**ross-**d**Omain **P**roduct **r**Epresentation model, COPE, is proposed to learn the cross-domain product representations. The experimental results prove the superiority of the COPE model to the existing methods.

## 2. Related Work

### 2.1. E-Commerce Datasets

A large number of e-commerce datasets have been proposed to advance the technical developments in the area [2, 6, 8, 11, 25, 32, 33]. The earlier datasets traditionally have limited size. Corbiere *et al.* introduce the Dress Retrieval [6] dataset in 2017, which has 20,000 pairs of the product image and text pairs. Rostamzadeh *et al.* propose the FashionGen [25] dataset, which includes 293,000 samples, covering only 48 product categories. In recent years, large-scale product recognition datasets have been introduced with the development of deep-learning-based methods. Product 1M [33] increases the scale of the training samples to a million level, but all samples come from 48 cosmetic brands. The coverage of the products is quite limited. MEP-3M [2] dataset includes more than three million samples, and each sample consists of the product image, product title, and hierarchical classification labels. However, all these datasets focus solely on the product page domain. In the experiment section, we will demonstrate that the representations learned on the product page domain are insufficient to handle the cross-domain product recognition task. The most related datasets to our ROPE dataset are M5Product [8] and MovingFashion [11]. M5Product comprises six million samples, and for each sample, it provides product images, product titles, category labels, attribute tables, assigned advertising videos, and audios extracted from videos. However, the provided videos in M5Product are quite different from the live streams introduced in our ROPE dataset. The videos in M5Product all come from the product page and are usually closely related to the advertised products, with products displayed in the center and described throughout. By contrast, there are many chat contents between the sellers and audiences in live streams of ROPE, which are unrelated to the products. Furthermore, the poses and locations of the products vary significantly in live streams, making the ROPE dataset more challenging for product recognition. MovingFashion [11] also focuses on aligning videos and product pages. It only comprises 15,000 videos, covering 13 product categories. The scale of

Dataset	Samples	Categories	Products	Domains
FashionGen [25]	293,008	48	78850	product page
Dress Retrieval [6]	20,200	50	20,200	product page
Product1M [33]	1,182,083	458	92,200	product page
MEP-3M [2]	3,012,959	599	-	product page
M5Product [8]	6,313,067	6,232	-	product page
MovingFashion [11]	15,000	-	-	product page/short video
ROPE(ours)	12,027,068	1,396	187,431	product page/short video/live streaming

Table 1. Comparisons with other product datasets. “-” means not mentioned.

the MovingFashion is much smaller than our ROPE dataset, which covers more than 1,000 product categories and provides million-level samples of the product page, short video, and live streaming domains.

## 2.2. Cross-Domain Retrieval Methods

Existing cross-domain retrieval methods typically learn unified representations between the visual and text domains. Some of the most popular models following the single-stream architecture, such as VL-bert [28], Imagebert [23], Videobert [29], Visualbert [13], and Uniter [4]. These models concatenate visual and text features and then use a binary classifier to predict whether the image-text pairs match. Although these methods usually perform better, they suffer from inferior inference efficiency. The ViLBERT [18], LXMERT [30], CLIP [24], and CoOp [34] utilize the two-stream architecture. In this approach, the visual and text features are extracted using independent encoders, and the visual-text similarity is efficiently calculated using the dot product operation. The proposed COPE model learns representations of different domains using contrastive loss to ensure efficient cross-domain retrieval.

## 3. ROPE Dataset

### 3.1. Data Collection and Cleaning

We collect data from the online e-commerce platform over 1,300 product categories. Three steps are taken to construct the ROPE dataset. Firstly, we collect a large amount of unsupervised multi-modal samples from the product page domain, short video domain, and live streaming domain. For the product page domain, we offer the product images and titles; for the short video and live streaming domains, the extracted frames and ASR (automatic speed recognition) texts are provided. The resulting dataset includes over 200 million samples and is defined as  $\mathcal{D}_{raw}$ .

Secondly, a small portion of  $\mathcal{D}_{raw}$  (0.1%, 200K data points) is sampled and defined as  $\mathcal{D}_{sample}$ . For each sample in  $\mathcal{D}_{sample}$ , we ask the human annotators to find other samples from  $\mathcal{D}_{raw}$  that shares the *same* product. To reduce the annotation costs, the extracted features with the public

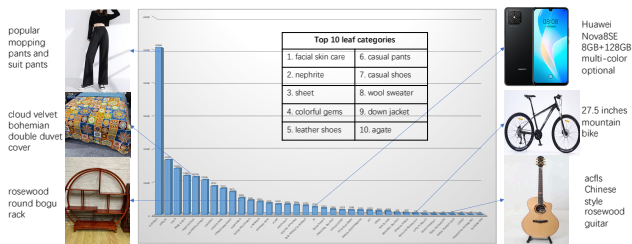


Figure 2. The distribution of training samples over product categories. It is biased and long-tailed.

Chinese CLIP model [31]<sup>1</sup> are utilized to find relevant samples for further human checkout. The annotated samples are used to train a baseline COPE model.

Thirdly, for remaining unannotated samples in  $\mathcal{D}_{raw}$ , the baseline COPE model is employed to filter out relevant samples, and only the samples whose matching scores are higher than 0.7 are kept. Afterward, the product pages, short videos, and live streams belonging to the same product are aggregated. We only retain the completed paired samples, including data from all three domains.

### 3.2. Datasets Statistics

The final ROPE dataset comprises 3,056,624 product pages, 5,867,526 short videos, and 3,495,097 live streams associated with 189,958 products. Table 1 compares the ROPE and previous product datasets. We divided the ROPE dataset into train and test sets. The train set has 187,431 products with 3,025,236 product pages, 5,837,716 short videos, and 3,464,116 live streams. On average, each product has 16 product pages, 31 short videos, and 18 live streams. The distribution of training samples across product categories is illustrated in Figure 2, showing a long-tailed pattern that reflects online shopping interests. The top five categories are Carpet, Calligraphy/Painting, Quilt Cover, Emerald, and Sheet. The test set contains 2,527 products,

<sup>1</sup>For short videos and live streamings, the average of frame-level image features are adopted as the visual representations. The visual and text features are concatenated as the final multi-model representations for retrieving relevant samples.

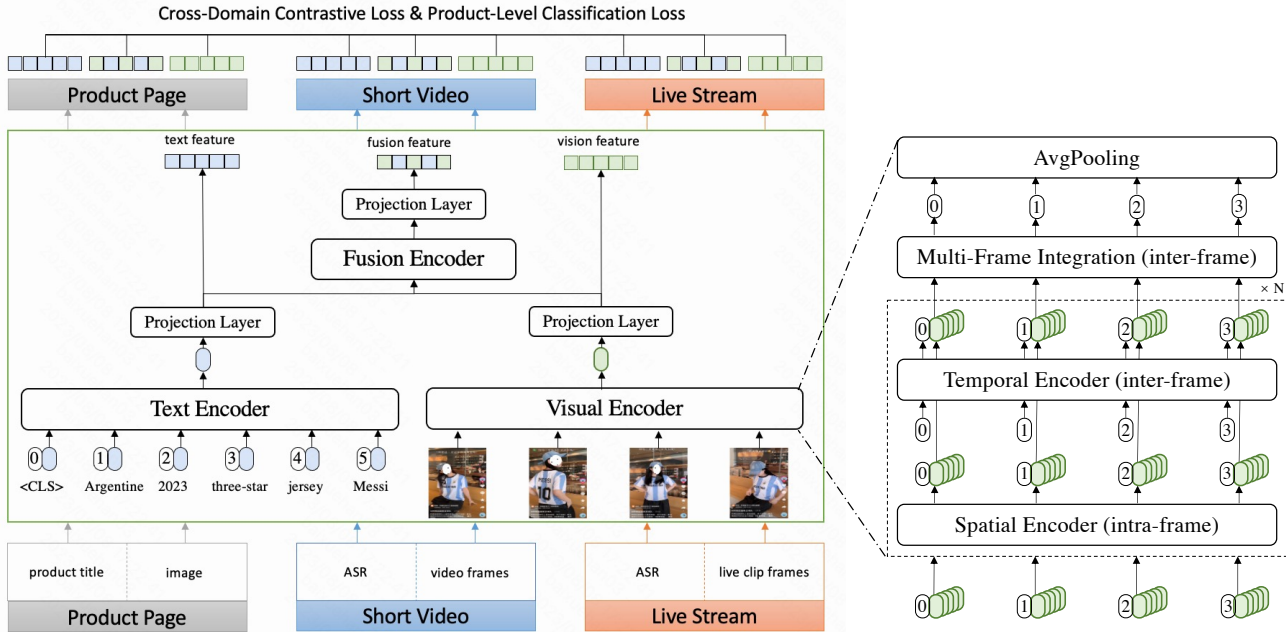


Figure 3. The overall framework of the proposed COPE model. The text encoder and visual encoder are utilized to extract features from the single modality, and the fusion encoder is adopted to aggregate the two features. To model the temporal information in videos and live streams, we insert the cross-frame communication transformer into each block of the visual encoder. The multi-frame integration transformer is placed at the top of the visual encoder to summarize the whole video’s representation.

with 31,388 product pages, 29,810 short videos, and 30,981 live streams. The average duration of short videos and live streams is 31.78 seconds and 129.09 seconds, respectively, and each product has an average of 12 product pages, 11 short videos, and 12 live streamings. The product categories in the test set are different from those in the train set to ensure an accurate evaluation, and human annotators have thoroughly reviewed the test set.

### 3.3. Evaluation Tasks

We propose two evaluation tasks based on the ROPE dataset to verify the unified cross-domain product representations. The first is the cross-domain product retrieval task, which aims to find the matched samples from the identical product from the two domains. There are six variations of the task:  $P \rightarrow V$ ,  $V \rightarrow P$ ,  $P \rightarrow L$ ,  $L \rightarrow V$ ,  $V \rightarrow L$ , and  $L \rightarrow V$ , where  $P$ ,  $V$ , and  $L$  indicate product page domain, short video domain, and live streaming domain, respectively. The second one is the cross-domain few-shot ( $k=1$ ) classification task. Similar to the retrieval task, it also has six variations.

Taking the  $P \rightarrow V$  variation as an example, we elaborate on the detailed evaluation processes for the two tasks. For the retrieval task, we collect all the short videos in the test set as the gallery set  $G_V$  and regard all the product pages in the test set as the query set  $Q_P$ . For each query in  $Q_P$ ,

the goal is to find a matched short video from  $G_V$ , whose product label is the same as the query product page. For the few-shot ( $k=1$ ) classification task, we randomly sample one short video from each product in the test set. The sampled shot videos are considered anchors. Then we try to classify all the product pages in test set by finding the nearest short video anchor.

## 4. Method

The overall framework of the proposed COPE model is illustrated in Figure 3. It comprises the visual encoder, the text encoder, the fusion encoder, and the domain projection layers. The visual and text encoders are shared between the three domains, and the parameters of the domain projection layers in each domain are not shared.

### 4.1. Architectural Design

As stated in Section 3.2, we provide training samples with multiple modalities for each domain. Specifically, we offer product titles and images for the product domain, while for the short video and live streaming domain, we provide extracted frames and ASR (automatic speech recognition) texts. The COPE model is designed with a two-stream pipeline to handle both visual and textual modalities. At the bottom of the model, we utilize a shared text encoder



and visual encoder to extract representations for raw texts and images/frames for each domain. These extracted features are fed into three domain-specific projection layers to obtain domain-specific representations. Additionally, we employ a fusion encoder module, followed by a projection layer, to aggregate visual and text features. The parameters of the fusion encoder are shared across domains, while the projection layers are domain-specific. It is important to note that we do not utilize the ASR texts, and we remove text-modal related modules for the short video and live streaming domains in our initial version of COPE. The excessive noise information in raw ASR texts can negatively impact the final presentations for videos and live streams. In our future work, we will explore possible approaches to utilize the ASR texts by extracting product-related keywords from the raw texts.

The visual encoder follows the same architecture as in [21], which consists of  $N$  cross-frame communication transformer (CCT) modules and a multi-frame integration transformer (MIT) module. The CCT module is a revised ViT [9] block by inserting the temporal encoder to enable temporal information exchangeability. The MIT module is placed at the top of  $N$  stack CCT modules to integrate the sequence of frame-level features into a unified video representation.

Given an input video  $\mathbf{V} \in \mathbb{R}^{T*H*W*3}$  (the product image can be regarded as a video with only one frame), where  $T$  denotes the number of frames.  $H$  and  $W$  indicate the spatial resolution of the video; we split the  $t$ -th frame into  $M$  non-overlapping patches,  $\mathbf{X}_{vis}^t \in \mathbb{R}^{M*3}$ . The learnable class token is inserted at the beginning of the patch sequence, and the spatial position encoding is added to the patch sequence. Formally,

$$\mathbf{z}_t^{(0)} = [e_{vis}^{cls}; \mathbf{X}_{vis}] + e^{spa} \quad (1)$$

Then we feed  $\mathbf{z}_t^{(0)}$  into  $N$  CCT modules to obtain the frame-level representations:

$$\begin{aligned} \mathbf{z}_t^{(n)} &= \text{CCT}^{(n)}(\mathbf{z}_t^{(n-1)}), \quad n = 1, \dots, N \\ &= [h_{t,cls}^{(n),vis}, h_{t,1}^{(n),vis}, h_{t,2}^{(n),vis}, \dots, h_{t,M}^{(n),vis}] \end{aligned} \quad (2)$$

where  $n$  denotes the CCT module index.

We take the final output of the class token at the  $N$ -th CCT module,  $h_{t,cls}^{(N),vis}$ , to represent the  $t$ -th frame. Then the global representation of the video is obtained by aggregating frame-level features with the MIT module. Formally,

$$Z_{vis} = \text{AvgPool}(\text{MIT}([h_{1,cls}^{(N),vis}, \dots, h_{T,cls}^{(N),vis}] + e^{temp})) \quad (3)$$

where AvgPool and  $e^{temp}$  denote the average pooling operator and temporal position encoding, respectively.  $Z_{vis} \in$

$\mathbb{R}^d$  is utilized as the visual representation for the input product image or videos.

The text encoder is a three-layer RoBERTa [7, 15] model. The input raw texts are firstly tokenized and defined as  $\mathbf{X}_{txt} \in \mathbb{R}^L$  where  $L$  indicates the length of the token sequence. Then the class token is inserted at the beginning of the sequence, and the position embeddings are added to retrain positional information. The final obtained text sequence is fed into the text encoder to extract text representations. Formally,

$$\begin{aligned} \mathbf{H}_{txt} &= \text{RoBERTa}([e^{cls}; \mathbf{X}_{txt}] + e^{pos}) \\ &= [h_{cls}^{txt}, h_1^{txt}, h_2^{txt}, \dots, h_L^{txt}] \end{aligned} \quad (4)$$

where  $e^{cls}$  and  $e^{pos}$  denote the input class token embedding and position embeddings, respectively.  $h_{cls}^{txt} \in \mathbb{R}^d$  indicates the extracted feature of the class token. We utilize  $h_{cls}^{txt}$  as the text representation for input raw texts.

The visual representation  $Z_{vis}$  and text representation  $h_{cls}^{txt}$  of the three domains are extracted with the shared visual and text encoders, even though the samples in different domains vary significantly. Such a scheme is expected to enhance the generalization capability of the basic feature extractors. The characteristics of each domain are retained and magnified by utilizing different projection layers that are not shared across domains to transform the general representations into domain-specific representations. For each domain, the projection layer is a linear layer with weight  $\mathbf{W}$  and bias  $b$ , and the domain-specific representations are obtained as:

$$\begin{aligned} E_{vis}^P &= \mathbf{W}_{vis}^P Z_{vis}^P + b_{vis}^P \\ E_{txt}^P &= \mathbf{W}_{txt}^P h_{txt}^P + b_{txt}^P \\ E_{vis}^V &= \mathbf{W}_{vis}^V Z_{vis}^V + b_{vis}^V \\ E_{vis}^L &= \mathbf{W}_{vis}^L Z_{vis}^L + b_{vis}^L \end{aligned} \quad (5)$$

where P, V, and L denote the product page domain, short video domain, and the live streaming domain. It should be noted that in the short video domain and live streaming domain, we do not include the text modality, and only visual representations, *i.e.*,  $E_{vis}^V$  and  $E_{vis}^L$ , are utilized for the two domains.

Finally, the fusion encoder, followed by a projection layer, is proposed to aggregate the visual and text representations. The fusion encoder is implemented with a self-attention layer, and the projection layer is a linear layer. Also, in our initial version of COPE, the fusion operation is only applied to the product page domain. Formally,

$$\begin{aligned} E_{fus}^P &= \text{SelfAttn}([E_{vis}^P; E_{txt}^P]) \\ E_{fus}^P &= \mathbf{W}_{fus}^P E_{fus}^P + b_{fus}^P \end{aligned} \quad (6)$$

where SelfAttn denotes the self-attention layer.  $E_{fus}^P$  is the obtained multi-modal representation for the product page

domain P. For the other two domains V and L, the visual representations  $E_{vis}^V$  and  $E_{vis}^L$  are the final obtained representations for them.

## 4.2. Training Objective

To learn a unified product representation across the different domains, we first leverage contrastive learning to train the proposed COPE model following previous self-supervised learning methods [3, 10, 22]. The basic formulation of the contrastive loss function [22] is defined as:

$$\mathcal{L}_{con} = -\log \frac{\exp(s_{qk_+}/\tau)}{\sum_{i=0}^{i=K} \exp(s_{qk_i}/\tau)} \quad (7)$$

where  $s_{qk_i}$  denotes the cosine similarity between the sample  $q$  and the sample  $k_i$ . The positive sample  $k_+$  indicates the sample that has the same product label with  $q$ .

The similarity  $s_{qk}$  can be calculated with a different form of representations (vision, text, or fusion), and the samples  $q$  and  $k$  can come from different domains (product page, short video, or live streaming). In this paper, we choose seven different implementations of the similarity  $s_{qk}$ , resulting in seven different contrastive loss functions. The details of the implementations are summarized in Table 2. Based on the seven contrastive loss functions, we define cross-domain loss as the sum of them. Formally,

$$\mathcal{L}_{cd} = \sum_{n=1}^{n=7} \alpha_n \mathcal{L}_{con}^n \quad (8)$$

where  $\alpha_n$  is the weight of  $n$ -th contrastive learning loss function.

In addition to the cross-domain loss, we also adopt the product classification loss to train our COPE model. Specifically, we use an MLP (multi-layer perceptron) with shared parameters to predict product classification scores for each domain with the domain-specific representations. For the product page domain, the multi-modal representation  $E_{fus}^P$  is utilized. For the short video and live streaming domains, the visual representations  $E_{vis}^V$  and  $E_{vis}^L$  are adopted. Formally,

$$\begin{aligned} s^P &= \text{MLP}(E_{fus}^P) \\ s^V &= \text{MLP}(E_{vis}^V) \\ s^L &= \text{MLP}(E_{vis}^L) \end{aligned} \quad (9)$$

where  $s$  denotes the classification score for each domain. Then the standard softmax loss is used to train the model. Formally,

$$\mathcal{L}_{cls} = -\left(\log \frac{e^{s_i^P}}{\sum_j^C e^{s_j^P}} + \log \frac{e^{s_i^V}}{\sum_j^C e^{s_j^V}} + \log \frac{e^{s_i^L}}{\sum_j^C e^{s_j^L}}\right) \quad (10)$$

similarity $s_{qk}$	domain	modality
$\langle E_{fus}^P(q), E_{vis}^V(k) \rangle$	product-video	fusion-vision
$\langle E_{fus}^P(q), E_{vis}^L(k) \rangle$	product-live	fusion-vision
$\langle E_{vis}^V(q), E_{vis}^L(k) \rangle$	video-live	vision-vision
$\langle E_{fus}^P(q), E_{vis}^V(k) \rangle$	product-video	vision-vision
$\langle E_{txt}^P(q), E_{vis}^V(k) \rangle$	product-video	text-vision
$\langle E_{vis}^P(q), E_{vis}^L(k) \rangle$	product-live	vision-vision
$\langle E_{txt}^P(q), E_{vis}^L(k) \rangle$	product-live	text-vision

Table 2. The implementations of different similarity functions  $s_{qk}$ .

The total loss to train the COPE model is the combination of the cross-domain loss and the classification loss. Formally,

$$\mathcal{L}_f = \mathcal{L}_{cd} + \beta \mathcal{L}_{cls} \quad (11)$$

where  $\beta$  indicates the weight of classification loss.

## 4.3. Implementation Details.

We initialize the text encoder with the public Chinese RoBERTa model [7, 15] and the visual encoder with the pre-trained model in [21]. Eight frames are extracted to obtain features for short videos and live streams. The training batch size is set to 84, and the training process continues for 80 epochs. We optimize the model using AdamW [17], and the cosine schedule with a linear warmup is used for adjusting the learning rate. The warmup approach continues for two epochs, and the max learning rates are set to 5e-5, 5e-7, and 5e-3 for the text encoder, visual encoder, and other layers.

## 4.4. Experimental Results

In this section, we evaluate our proposed COPE model and compare it with the state-of-the-arts on the ROPE dataset. The cross-domain product retrieval task and one-shot cross-domain classification task are considered. Since no existing methods are precisely suitable to our cross-domain setting, we compare the COPE model with the multi-modal vision-language models [5, 16, 19, 20, 31], which are not fine-tuned on our dataset. The product page representation is obtained for these models by averaging the image and text features. The short video and live streaming representations are extracted by averaging the representations of all frames.

The vision-language models trained with general image-text pairs are compared in the first compartment of Table 3. We can see that all of them obtain inferior performance to our COPE model on each setting of the two evaluation tasks. In both the retrieval and classification tasks, the performance in the live related settings, *i.e.*, P→L, L→P, V→L, and L→V, is obviously lower than others. For example, the

models	cross domain setting	cross domain retrieval						few-shot classification
		R@1	R@5	R@10	R@20	R@50	R@mean	Top1 Acc
CLIP4CLIP [19]	P2V	59.06	79.31	86.02	91.01	95.03	82.08	27.94
	V2P	38.48	52.25	59.16	66.54	74.65	58.21	26.55
	P2L	23.68	38.14	45.32	54.27	66.79	45.64	9.97
	L2P	14.46	24.52	30.77	38.09	48.91	31.35	10.75
	V2L	18.10	29.83	35.65	42.22	52.01	35.56	9.47
	L2V	20.14	33.51	40.44	48.05	58.68	40.16	7.22
TS2-Net [16]	P2V	57.42	77.88	85.29	90.44	94.92	81.19	26.11
	V2P	36.56	50.93	58.02	65.12	73.89	56.90	24.09
	P2L	22.85	38.49	45.91	54.11	65.89	45.45	9.83
	L2P	14.16	24.52	30.50	37.52	48.37	31.01	10.57
	V2L	17.69	29.63	34.84	41.27	50.95	34.87	9.68
	L2V	20.55	33.80	40.91	48.46	59.16	40.57	7.40
X-CLIP [20]	P2V	56.61	77.46	84.84	90.11	94.51	80.70	26.97
	V2P	35.29	49.41	56.82	64.13	72.54	55.63	23.55
	P2L	22.66	37.47	44.33	52.11	63.38	43.98	9.72
	L2P	13.52	23.08	28.92	35.98	46.14	29.52	8.88
	V2L	17.64	28.71	34.03	40.17	49.67	34.04	9.05
	L2V	19.60	32.73	39.51	47.07	57.25	39.23	7.42
ChineseCLIP [31]	P2V	56.93	79.80	87.43	92.48	96.51	82.65	31.44
	V2P	40.48	57.85	66.74	75.25	84.03	64.87	29.10
	P2L	34.37	50.83	58.66	67.05	78.57	57.89	19.23
	L2P	22.49	37.11	46.78	56.30	68.14	46.16	15.73
	V2L	25.51	38.28	45.02	52.27	62.53	44.72	13.24
	L2V	28.28	45.87	53.67	62.18	72.27	52.45	14.16
FashionClip [5]	P2V	44.31	67.06	75.25	82.57	89.29	71.69	18.59
	V2P	25.51	40.75	48.71	56.63	65.94	47.50	15.88
	P2L	19.54	31.14	36.98	43.91	54.39	37.19	8.70
	L2P	11.22	24.23	31.90	40.05	50.96	31.67	7.57
	V2L	15.55	24.88	29.51	35.07	42.68	29.53	6.80
	L2V	21.20	35.72	42.55	49.60	58.77	41.56	10.40
COPE (Ours)	P2V	<b>82.58</b>	<b>94.88</b>	<b>97.54</b>	<b>98.89</b>	<b>99.65</b>	<b>94.70</b>	<b>59.84</b>
	V2P	<b>65.20</b>	<b>76.56</b>	<b>82.04</b>	<b>86.86</b>	<b>91.69</b>	<b>80.47</b>	<b>57.12</b>
	P2L	<b>54.06</b>	<b>71.07</b>	<b>77.14</b>	<b>82.86</b>	<b>89.70</b>	<b>74.96</b>	<b>34.95</b>
	L2P	<b>42.33</b>	<b>56.48</b>	<b>63.67</b>	<b>71.11</b>	<b>80.22</b>	<b>62.76</b>	<b>36.51</b>
	V2L	<b>45.95</b>	<b>63.63</b>	<b>70.64</b>	<b>77.50</b>	<b>85.47</b>	<b>68.63</b>	<b>30.43</b>
	L2V	<b>48.28</b>	<b>67.20</b>	<b>74.70</b>	<b>81.52</b>	<b>89.15</b>	<b>72.17</b>	<b>33.30</b>

Table 3. Retrieval and classification results on COPE. P, V, and L means product page, short video, and live stream domains.

COPE model obtains 82.58% R@1 in the P→V retrieval task and 59.84% Acc in the P→V classification task. By contrast, the performance of COPE in the P→L settings is 54.06% and 34.95%, respectively. The scales and views of products in live streams differ from the product pages. Moreover, the products are not always visible in the live stream durations. These situations significantly improve the difficulty of recognizing products in live streams. In the second compartment of Table 3, we compare the COPE model with the FashionClip model. Although the FashionClip model is trained with the product images and titles

rather than the general data, there are still large margins between the obtained results of FashionClip and COPE. As described in Section 2.1, the representations learned on the product page domain only are insufficient to deal with the cross-domain product recognition problem.

#### 4.5. Performance on other datasets

In order to verify the generalization of our ROPE dataset. We directly utilize the learned COPE model on ROPE to extract product representations and conduct evaluations on other datasets, such as Product1M [33] and M5Product [8].



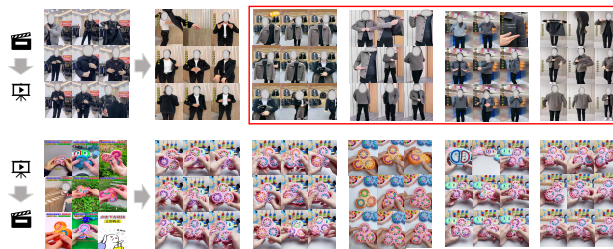




(a) P2V and V2P retrieval results



(b) P2L and L2P retrieval results



(c) V2L and L2V retrieval results

Figure 5. Visualization of the retrieval results, where the red box denotes false positive.

of our COPE approach in distinguishing between different products. Furthermore, Figure 5 displays some of our retrieval results. Notably, most of the false positive results belong to the same category as the query and possess similar visual characteristics.

## 5. Conclusion

To enable the creation of a unified cross-domain product representation, we introduce a large-scale E-commerce cross-domain dataset that includes three domains (product pages, short videos, and live streams) and two modalities (vision and language). It is the first dataset that encompasses various domains in the e-commerce scenario. We propose our COPE as the baseline and evaluate it on cross-domain retrieval and few-shot classification tasks. Finally, we provide an analysis and visualization of the results. This task applies to most e-commerce platforms, and both the dataset and the proposed framework will inspire research

on cross-domain product representation.

## References

- [1] Xiang An, Xuhan Zhu, Yuan Gao, Yang Xiao, Yongle Zhao, Ziyong Feng, Lan Wu, Bin Qin, Ming Zhang, Debing Zhang, et al. Partial fc: Training 10 million identities on a single machine. In *ICCV*, pages 1445–1449, 2021. 8
- [2] Delong Chen, Fan Liu, Xiaoyu Du, Ruizhuo Gao, and Feng Xu. Mep-3m: A large-scale multi-modal e-commerce products dataset. In *IJCAI*, volume 21, 2021. 2, 3
- [3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 6
- [4] 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 *ECCV*, pages 104–120. Springer, 2020. 3
- [5] Patrick John Chia, Giuseppe Attanasio, Federico Bianchi, Silvia Terragni, Ana Rita Magalhães, Diogo Goncalves, Ciro Greco, and Jacopo Tagliabue. Fashionclip: Connecting language and images for product representations. *arXiv preprint arXiv:2204.03972*, 2022. 6, 7
- [6] Charles Corbiere, Hedi Ben-Younes, Alexandre Ramé, and Charles Ollion. Leveraging weakly annotated data for fashion image retrieval and label prediction. In *ICCV*, pages 2268–2274, 2017. 2, 3
- [7] Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. Pre-training with whole word masking for chinese bert. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3504–3514, 2021. 5, 6
- [8] 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 *CVPR*, pages 21252–21262, 2022. 2, 3, 7
- [9] 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. 5
- [10] Tianyu Gao, Xingcheng Yao, and Danqi Chen. Simcse: Simple contrastive learning of sentence embeddings. In *EMNLP*, pages 6894–6910, 2021. 6
- [11] Marco Godi, Christian Joppi, Geri Skenderi, and Marco Cristani. Movingfashion: a benchmark for the video-to-shop challenge. In *WACV*, pages 1678–1686, 2022. 2, 3
- [12] Brendan Kolisnik, Isaac Hogan, and Farhana Zulkernine. Condition-cnn: A hierarchical multi-label fashion image classification model. *Expert Systems with Applications*, 182:115195, 2021. 2
- [13] Liumian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019. 3

- [14] Huidong Liu, Shaoyuan Xu, Jinmiao Fu, Yang Liu, Ning Xie, Chien-Chih Wang, Bryan Wang, and Yi Sun. Cma-clip: Cross-modality attention clip for image-text classification. *arXiv preprint arXiv:2112.03562*, 2021. 2
- [15] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 5, 6
- [16] Yuqi Liu, Pengfei Xiong, Luhui Xu, Shengming Cao, and Qin Jin. Ts2-net: Token shift and selection transformer for text-video retrieval. In *ECCV*, pages 319–335. Springer, 2022. 6, 7
- [17] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*. 6
- [18] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*, pages 13–23, 2019. 3
- [19] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing*, 508:293–304, 2022. 6, 7
- [20] Yiwei Ma, Guohai Xu, Xiaoshuai Sun, Ming Yan, Ji Zhang, and Rongrong Ji. X-clip: End-to-end multi-grained contrastive learning for video-text retrieval. In *ACMMM*, pages 638–647, 2022. 6, 7
- [21] Bolin Ni, Houwen Peng, Minghao Chen, Songyang Zhang, Gaofeng Meng, Jianlong Fu, Shiming Xiang, and Haibin Ling. Expanding language-image pretrained models for general video recognition. 2022. 5, 6
- [22] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 6
- [23] Di Qi, Lin Su, Jia Song, Edward Cui, Taroon Bharti, and Arun Sacheti. Imagebert: Cross-modal pre-training with large-scale weak-supervised image-text data. *arXiv preprint arXiv:2001.07966*, 2020. 3
- [24] 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. 3
- [25] Negar Rostamzadeh, Seyedarian Hosseini, Thomas Boquet, Wojciech Stokowiec, Ying Zhang, Christian Jauvin, and Chris Pal. Fashion-gen: The generative fashion dataset and challenge. 2018. 2, 3
- [26] Majuran Shajini and Amirthalingam Ramanan. An improved landmark-driven and spatial-channel attentive convolutional neural network for fashion clothes classification. *The Visual Computer*, 37(6):1517–1526, 2021. 2
- [27] Majuran Shajini and Amirthalingam Ramanan. A knowledge-sharing semi-supervised approach for fashion clothes classification and attribute prediction. *The Visual Computer*, 38(11):3551–3561, 2022. 2
- [28] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vi-bert: Pre-training of generic visual-linguistic representations. In *ICLR*. 3
- [29] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *ICCV*, pages 7464–7473, 2019. 3
- [30] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *EMNLP-IJCNLP*, pages 5100–5111, 2019. 3
- [31] An Yang, Junshu Pan, Junyang Lin, Rui Men, Yichang Zhang, Jingren Zhou, and Chang Zhou. Chinese clip: Contrastive vision-language pretraining in chinese. *arXiv preprint arXiv:2211.01335*, 2022. 3, 6, 7
- [32] Wenjie Yang, Yiyi Chen, Yan Li, Yanhua Cheng, Xudong Liu, Quan Chen, and Han Li. Cross-view semantic alignment for livestreaming product recognition. *arXiv preprint arXiv:2308.04912*, 2023. 2
- [33] 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 *ICCV*, pages 11782–11791, 2021. 2, 3, 7
- [34] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *CVPR*, pages 16816–16825, 2022. 3