# Real20M: A Large-scale E-commerce Dataset for Cross-domain Retrieval

Yanzhe Chen
Wangxuan Institute of Computer Technology & National Key Laboratory for Multimedia Information Processing, Peking University
Beijing, China
chenyanzhe@stu.pku.edu.cn

Huasong Zhong
Kuaishou Technology
Beijing, China
zhonghuasong@kuaishou.com

Xiangteng He
Wangxuan Institute of Computer Technology & National Key Laboratory for Multimedia Information Processing, Peking University
Beijing, China
hexiangteng@pku.edu.cn

Yuxin Peng*
Wangxuan Institute of Computer Technology & National Key Laboratory for Multimedia Information Processing, Peking University
Beijing, China
pengyuxin@pku.edu.cn

Lele Cheng
Kuaishou Technology
Beijing, China
chenglele@kuaishou.com

## ABSTRACT

In e-commerce, products and micro-videos serve as two primary carriers. Introducing cross-domain retrieval between these carriers can establish associations, thereby leading to the advancement of specific scenarios, such as retrieving products based on micro-videos or recommending relevant videos based on products. However, existing datasets only focus on retrieval within the product domain while neglecting the micro-video domain and often ignore the multimodal characteristics of the product domain. Additionally, these datasets strictly limit their data scale through content alignment and use a content-based data organization format that hinders the inclusion of user retrieval intentions. To address these limitations, we propose the **PKU Real20M** dataset, a large-scale e-commerce dataset designed for cross-domain retrieval. We adopt a query-driven approach to efficiently gather over 20 million e-commerce products and micro-videos, including multimodal information. Additionally, we design a three-level entity prompt learning framework to align inter-modality information from coarse to fine. Moreover, we introduce the Query-driven Cross-Domain retrieval framework (QCD), which leverages user queries to facilitate efficient alignment between the product and micro-video domains. Extensive experiments on two downstream tasks validate the effectiveness of our proposed approaches. The dataset and source code are available at https://github.com/PKU-ICST-MIPL/Real20M_ACMMM2023.

## CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**; **Computer vision**; **Computer vision tasks**; **Visual content-based indexing and retrieval**;

## KEYWORDS

Large-scale data collection; E-commerce datasets; Cross-domain retrieval

**ACM Reference Format:**
Yanzhe Chen, Huasong Zhong, Xiangteng He, Yuxin Peng, and Lele Cheng. 2023. Real20M: A Large-scale E-commerce Dataset for Cross-domain Retrieval. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29-November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3581783.3612408

## 1 INTRODUCTION

E-commerce platforms heavily rely on retrieval and recommendation algorithms to match users with products they are interested in [11, 27]. In e-commerce scenarios, products and micro-videos serve as the main carriers. However, there is a lack of association between these two domains, which hinders the development of specific scenarios, such as retrieving products based on micro-videos or recommending relevant videos based on products. Cross-domain retrieval between e-commerce products and micro-videos has the potential to enhance retrieval and recommendation efficiency by establishing associations and overcoming data isolation. The synergistic interaction between them can not only provide more comprehensive information for users, leading to a more personalized and satisfying online shopping experience but also bring new challenges and opportunities to the academic community.

[*] Corresponding author.

*MM '23, October 29-November 3, 2023, Ottawa, ON, Canada*

ACM ISBN 979-8-4007-0108-5/23/10…$15.00
https://doi.org/10.1145/3581783.3612408

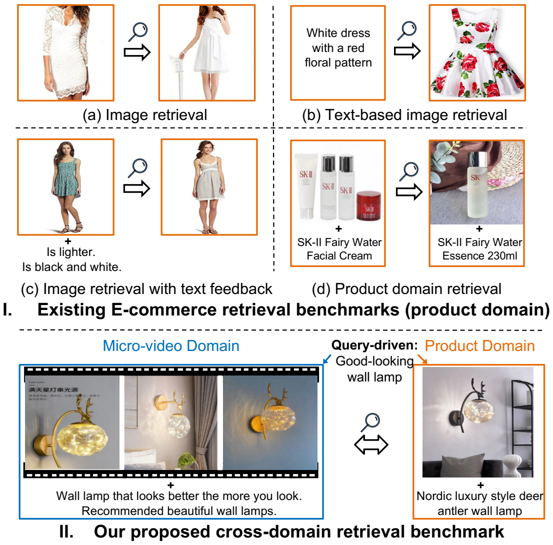
White dress
with a red
floral pattern
(a) Image retrieval
(b) Text-based image retrieval
+
Is lighter.
Is black and white.
+
SK-II Fairy Water
Facial Cream
+
SK-II Fairy Water
Essence 230ml
(c) Image retrieval with text feedback
(d) Product domain retrieval
I. Existing E-commerce retrieval benchmarks (product domain)
Micro-video Domain
Query-driven:
Good-looking
wall lamp
Product Domain
满天星灯串光源
+
Wall lamp that looks better the more you look.
Recommended beautiful wall lamps.
+
Nordic luxury style deer
antler wall lamp
II. Our proposed cross-domain retrieval benchmark

**Figure 1: Existing e-commerce benchmarks and our proposed cross-domain benchmark.**

In this paper, we investigate existing e-commerce retrieval benchmarks with corresponding datasets and categorize them based on the included modalities of the product domain, as illustrated in Figure 1. As shown in Figure 1 (a), the first category is image retrieval, and the typical dataset for this task is DeepFashion [19], which only contains a single visual modality. The second category in Figure 1 (b) is text-based image retrieval, and FashionGen [26] is a representative dataset for this task. However, in real-world e-commerce scenarios, products are often multimodal [4, 15], with not only visual modalities such as product images, but also naturally paired textual modalities such as product names, titles, and descriptions. Combining information from both modalities can lead to more accurate and comprehensive retrieval results compared to using only one modality [21, 30, 35]. In Figure 1 (c), FashionIQ [32] proposed a task of retrieving new products based on a single product image and user feedback. However, the products in this gallery only contain visual modalities and lack associated textual information. Additionally, the dataset is limited to only three categories. In Figure 1 (d), Product1M [34] is a representative dataset that includes cosmetic product images with paired textual descriptions and proposes an instance-level multimodal product retrieval task. However, this dataset only contains cosmetic products and cannot represent a wider range of e-commerce product categories. And all of them are single-domain product retrieval, which neglects another important information carrier in real-world scenarios, namely the micro-video domain. Furthermore, the conventional data organization method based solely on content lacks the capability to comprehend user intentions, and necessitates manual association of each product, thereby being time-consuming and not readily scalable.

To bridge this gap and advance related research, we have collected a large dataset called PKU Real20M, which is more representative of cross-domain retrieval requirements in real-world e-commerce. To address the aforementioned challenges, we collect and filter over 20 million diverse products and micro-videos with corresponding multimodal information. Each sample includes either an image or a micro-video with its corresponding textual description. To ensure the correlation between the two domains and enable cross-domain retrieval, we construct a large amount of associated data in an efficient manner. Specifically, we collect a list of user queries and associated them with the products or micro-videos that the users click on based on their queries. When the click-through rate exceeds a predetermined threshold, we consider the query, products, and micro-video to be related. By constructing the dataset in this way, we efficiently couple the product and micro-video domains through the query and ultimately obtain a large-scale multi-modal e-commerce retrieval dataset with over 20 million samples, which is more effective than traditional content-based data organization methods that lack an understanding of user intentions and require manual association between each product.

In addition to constructing the large-scale dataset described above, we design a pre-training framework for the e-commerce scenes. Besides the three common pre-training tasks: Masked Language Model, Masked Frame Model, and Video-Text Matching, we also introduce a prompt learning task based on three-level entity words, achieving alignment from coarse to fine granularity between the visual modality and the textual modality according to the semantic granularity of the three-level entity words. Furthermore, we propose a Query-driven Cross-Domain retrieval framework(QCD). Our framework uses a model with parameter sharing to extract visual and textual features in the product and micro-video domains and combines them through a fusion module to obtain fused features. By aligning the query features with the fused features, visual features, and textual features separately, we achieve domain alignment between the product and micro-video domains. To prevent the fused features from being overly influenced by textual semantics, we propose a reconstruction loss based on text and image reconstruction. Our framework's effectiveness is validated by performing retrieval between the product and micro-video domains.

The main contributions are summarized as follows:

- We propose a cross-domain retrieval benchmark and a corresponding large-scale e-commerce dataset, which possesses several distinctive features: (1) cross-domain and multimodal, (2) query-driven, and (3) massive and diverse.
- We introduce a pre-training framework tailored to the e-commerce scene, which aligns visual and textual semantics from coarse to fine-grained levels using a three-level entity prompt learning task.
- We present a query-driven framework for aligning the product and micro-video domains, and demonstrate notable improvements by applying this framework to existing methods.

## 2 RELATED WORK

### 2.1 E-commerce Datasets

In recent years, a large number of datasets have been proposed for retrieval. Among which, the DeepFashion [3], DeepFashion2 [8], Fashion-Gen [26], Fashion200k [12], M5Product [5], FashionIQ [32] and Product1M [34] are commonly used multi-modal datasets. The DeepFashion [3] and DeepFashion2 [8] datasets have proposed retrieval tasks between user-taken images and product images. The

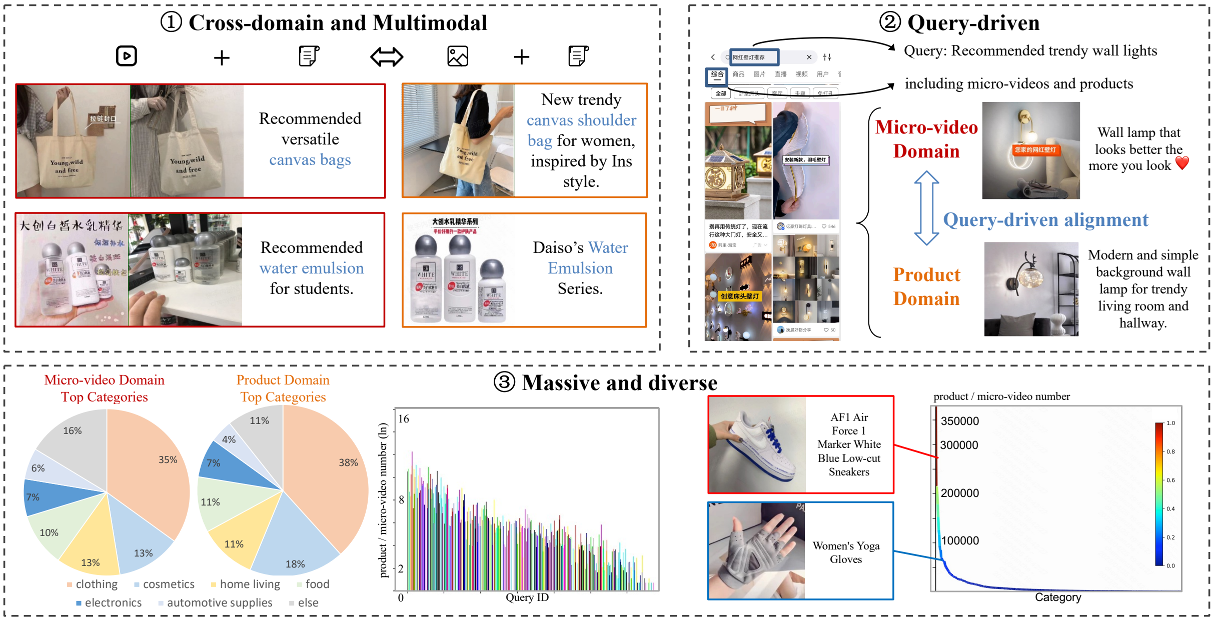
① Cross-domain and Multimodal
Recommended versatile canvas bags
New trendy canvas shoulder bag for women, inspired by Ins style.
Recommended water emulsion for students.
Daiso's Water Emulsion Series.
② Query-driven
Query: Recommended trendy wall lights
including micro-videos and products
Micro-video Domain
Wall lamp that looks better the more you look ❤
Query-driven alignment
Product Domain
Modern and simple background wall lamp for trendy living room and hallway.
③ Massive and diverse
Micro-video Domain Top Categories
Product Domain Top Categories
35%
13%
13%
10%
7%
6%
16%
38%
18%
11%
11%
7%
4%
11%
clothing
cosmetics
home living
food
electronics
automotive supplies
else
product / micro-video number (ln)
Query ID
16
8
2
0
AF1 Air Force 1 Marker White Blue Low-cut Sneakers
Women's Yoga Gloves
product / micro-video number
350000
300000
200000
100000
Category
1.0
0.8
0.6
0.4
0.2
0.0

**Figure 2: Visualization of the characteristics of our proposed benchmark dataset PKU Real20M.**

**Table 1: Comparison with other E-commerce datasets. V, T, and I respectively represent the video, text, and image modalities, and the arrow represents the retrieval direction.**

| Dataset | Samples | Product category | Retrieval type | Include video? | Cross domain retrieval ? | Query-driven? |
|---|---|---|---|---|---|---|
| M5Product [5] | 6,313,064 | Clothing, toys, etc | - | Yes | No | No |
| DeepFashion [19] | 54,642 | Clothing | I⟶I | No | No | No |
| DeepFashion2 [8] | 873234 | Clothing | I⟶I | No | No | No |
| Fashion-Gen [26] | 293,008 | Clothing | T⟶I | No | No | No |
| FashionIQ [32] | 77,684 | Clothing | (I+T)⟶I | No | No | No |
| Fashion200k [12] | 209,544 | Clothing | (I+T)⟶I | No | No | No |
| Product1M [34] | 1,182,083 | Cosmetics | (I+T) ↔ (I+T) | No | No | No |
| **PKU Real20M(Ours)** | 27,090,133 | Clothing, cosmetics, furniture, electronics, delicatessen, toys, etc. | (I+T)⇄(V+T) | Yes | Yes | Yes |

Weakly [3] and Fashion200k [12] contain 20,200 and 209,544 samples respectively and provide a basic retrieval task between text and images. The M5Product dataset [5] is a large-scale dataset in the e-commerce field, containing retrieval between 5 modalities. The FashionIQ [32] contains 77,684 samples and proposes a benchmark setting for retrieving target images based on image and text feedback. The Product1M [34] contains 1,182,083 cosmetic samples and performs cosmetic retrieval based on cosmetic images and related text information.

## 2.2 Multi-modal Pretraining for E-commerce

Several visual-linguistic pre-training models have been proposed for multi-modal learning of vision and language [13, 17, 28, 29]. In e-commerce, specific pre-training methods have also emerged. Kaleido-BERT [36] uses multi-granular and diverse image patch features for the visual modality and designs various pre-training tasks for these features. MEEK [23] employs five pre-training tasks to align visual and textual modalities, which include Masked Language Model (MLM), Masked Frame Model (MFM), Video-Text Matching (VTM), Category Classification, and Language Generation.

# 3 PKU REAL20M DATASET

## 3.1 Data Collections and Statistics

The data is collected from a popular e-commerce platform, based on user click feedback. When a user enters a query term, they are presented with two types of data: products and micro-videos. Users then interact with the relevant products or micro-videos by clicking and viewing them. By accumulating a large amount of user click data, we select products and micro-videos that exceed a certain threshold as content related to the query term. Therefore, we can easily collect a massive amount of <query, products> or <query, micro-video> data pairs. We remove duplicate queries and

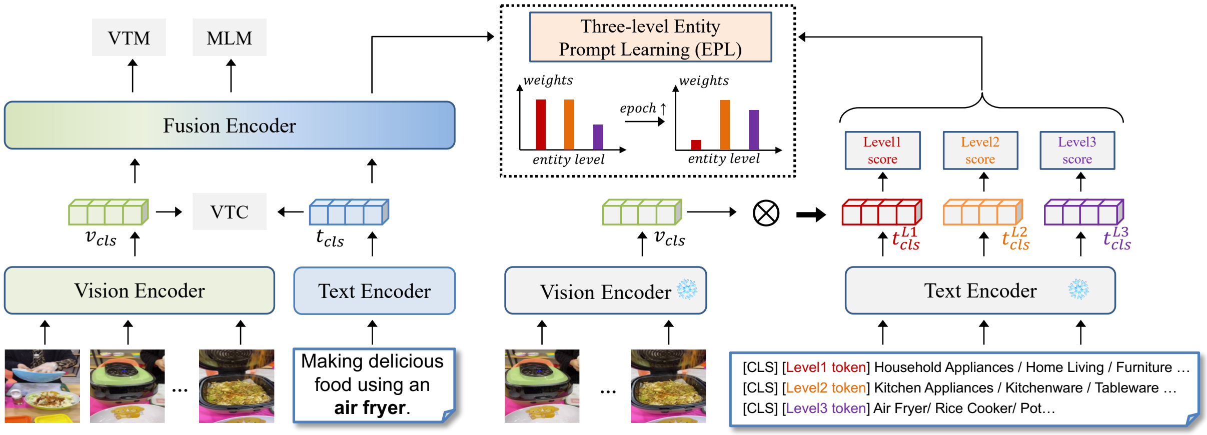
VTM
MLM
Fusion Encoder
Three-level Entity Prompt Learning (EPL)
weights
entity level
epoch
weights
entity level
Level1 score
Level2 score
Level3 score
$v_{cls}$
VTC
$t_{cls}$
$v_{cls}$
$t_{cls}^{L1}$
$t_{cls}^{L2}$
$t_{cls}^{L3}$
Vision Encoder
Text Encoder
Vision Encoder
Text Encoder
Making delicious food using an air fryer.
[CLS] [Level1 token] Household Appliances / Home Living / Furniture …
[CLS] [Level2 token] Kitchen Appliances / Kitchenware / Tableware …
[CLS] [Level3 token] Air Fryer/ Rice Cooker/ Pot…

**Figure 3: Overall architecture of our proposed pre-training framework.**

set a threshold for the number of containing samples to ensure query effectiveness. We collect a total of 1,000,000 user queries and built a high-quality dataset containing multimodal information, consisting of 27,090,133 samples and 1714 categories. The product domain contains 10,693,179 samples, while the micro-video domain contains 16,396,954 samples.

## 3.2 Benchmark Dataset Characteristics

Our proposed dataset PKU Real20M exhibits the following characteristics that closely resemble real e-commerce scenarios, which are visualized in Figure 2.

**1. Cross-domain and multimodal**. Unlike existing e-commerce datasets that often only contain image modalities, our dataset includes not only image or video modalities but also text modality information that is paired with them. And there is cross-domain alignment between products and micro-videos.

**2. Query-driven**. Our dataset is aligned in a query-driven manner, which is beneficial for building massive data. Additionally, incorporating user queries helps capture the intent behind micro-videos with complex scenes accurately.

**3. Massive and diverse.** Our dataset contains over 20 million samples, including various types of products and micro-videos. Our dataset contains 16,369,954 micro-videos with corresponding video titles and descriptions. The dataset is organized around 1714 major categories, covering a diverse range of product types and rich e-commerce scenarios. We visualize the number of samples included in the top categories of the two domains, the number of samples included in the query, and the number of samples under each category, as shown in Figure 2.

## 3.3 Comparison with Current Datasets

We compare our PKU Real20M dataset with commonly used existing datasets in e-commerce, as shown in Table 1. PKU Real20M differs from them mainly in the following four aspects: **1) Multimodal characteristics**: DeepFashion [19], FashionIQ [32] etc. ignore the multimodal attributes of the product domain in the retrieval task, while PKU Real20M is multimodal in both the product domain and the micro-video domain. **2) Retrieval task setting**: existing datasets ignore the important retrieval scenario between the product domain and the micro-video domain. Although the M5Product [5] dataset contains multiple modalities, they all belong to different forms of the product domain. At the same time, the way M5Product retrieves according to categories has significant differences from the actual retrieval scenario. The Product1M [34] dataset is limited to the retrieval within the cosmetics product domain. In contrast, PKU Real20M fills the gap in this field. **3) Data organization**: existing datasets are content-based, while our query-driven data organization approach can efficiently construct massive data and incorporate user intent information in the alignment relationship. **4) Data scale and diversity**: PKU Real20M contains a total of 27,090,133 samples, which is four times that of the M5Product dataset used for pre-training tasks and 20 times that of the Product1M dataset. Unlike existing retrieval datasets that are often limited to clothing and cosmetics categories, PKU Real20M also includes various real-world categories such as furniture, electronics, and more.

# 4 METHODOLOGY

## 4.1 Benchmark Description

The cross-domain benchmark involves two types of retrieval tasks: product retrieval based on micro-videos and micro-video retrieval based on products. The retrieval system under consideration involves the use of product and micro-video samples. The samples consist of a visual modality, represented by a product image or video, and a corresponding text modality, which can include a title or description, denoted as $(V, T)$. The retrieval gallery is a collection of such samples, denoted as $G = G_i \mid G_i = (V_G^i, T_G^i)$. The query sets, denoted as $Q = Q_i \mid Q_i = (V_Q^i, T_Q^i)$, are used for both types of retrieval. The aim of the retrieval system is to rank the samples in the gallery $G$ based on their relevance to the query set $Q$.

## 4.2 Multimodal Pretraining

We first adopt three commonly used pre-training tasks, Video-Text Contrastive Learning (VTC), Video-Text Matching (VTM), and

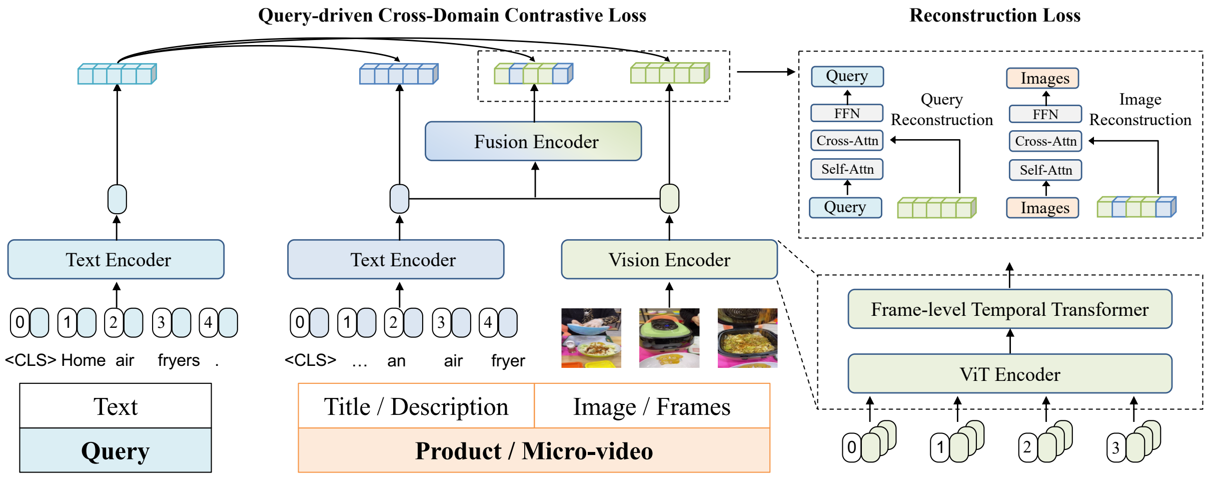

**Figure 4: Overall architecture of our proposed query-guided cross-domain retrieval framework.**

Masked Language Model (MLM), internally in the product/micro-video domain to further align the visual and textual backbones of e-commerce data. These three loss functions are denoted as $L_{vtc}$, $L_{vtm}$ and $L_{mlm}$, refer to [13, 13, 28, 29].

*Three-level Entity Prompt Learning (EPL).* During this process, we freeze the vision encoder and the text encoder trained by the previous three pre-training methods and train a fusion encoder that contains more fine-grained semantic information. Specifically, we extracted entities of three different semantic granularities from the e-commerce scene, and prompter maintained lists of these entities with lengths of $M_1$, $M_2$, and $M_3$ respectively. The first-level entities have the largest semantic granularity, such as home living, furniture, etc., while the third-level entities have the smallest semantic granularity, such as air fryer, rice cooker, etc. For each level, the text prompt is instantiated as follows: "[level] ENTITY", where "level" is the token identifying the entity level, and "ENTITY" is the specific word in the entity list. For each text prompt, we can compute the [CLS] embedding representation for each entity as $\{e^1_{\text{cls}}, e^2_{\text{cls}}, \dots, e^{M_l}_{\text{cls}}$ $\}$, $l \in \{L_1, L_2, L_3\}$. We extract visual features $v$ from random frames of the videos and use the prompter to generate pseudo-labels $p^l$ for the three-level entity words. Specifically, we calculate the softmax-normalized similarity between $v$ and the prompt embeddings of each level $e^l_m$, $l \in \{L_1, L_2, L_3\}$, as follows:

$$p^l_m = \frac{\exp\left(s(v, e^l_m)/\tau\right)}{\sum_{m=1}^{M_l} \exp\left(s(v, e^l_m)/\tau\right)} \tag{1}$$

Finally, we apply mean pooling to the fused feature $u$ computed by the Fusion encoder and obtain $c$ through a linear mapping layer. We then compute the cross-entropy between $c$ and $p$ as the EPL loss, which can be expressed as follows:

$$\begin{aligned} L_{epl} = &- (1 - w_t) \times \sum_{m_1=1}^{M_1} p^{L_1}_{m_1} \cdot \log c^{L_1}_{m_1} - \sum_{m_2=1}^{M_2} p^{L_2}_{m_2} \cdot \log c^{L_2}_{m_2} \\ &- w_t \sum_{m_3=1}^{M_3} p^{L_3}_{m_3} \cdot \log c^{L_3}_{m_3} \end{aligned} \tag{2}$$

In order to ensure that the fused features conform to coarse-grained semantic information in the early stage of training and focus more on fine-grained semantic information in the later stage, we adopt a progressive weight adjustment strategy: $w_t = \frac{1}{1+e^{-\alpha t}}$, where $t$ represents the number of epochs in training, and $\alpha$ is a hyper-parameter used to control the speed of decline for $w_t$. Through Three-level Entity Prompt Learning, we naturally associate the fusion features with three-level entity words and pay more attention to fine-grained semantic information as training progresses. Compared with the product categories, the three-level entity words have more generalization characteristics for retrieval from users, and can better fit the intent-based retrieval framework. During the pre-training stage, the overall loss function can be represented as follows:

$$L_{pretrain} = L_{vtc} + L_{vtm} + L_{mlm} + L_{epl} \tag{3}$$

## 4.3 Query-driven Cross-domain Retrieval Framework

Our proposed framework consists of two branches: a query branch for guiding alignment and a multi-modal feature extraction branch for extracting product/micro-video features, as shown in Figure 4. In the query branch, we first encode the query text to obtain the query embedding denoted as $q = \{q_{cls}, q_1, q_2, ..., q_n\}$. In the multi-modal feature extraction branch, we extract both textual and visual features. Specifically, we use the same textual backbone as in the query branch to extract textual information contained in the product, represented as $t = \{t_{cls}, t_1, t_2, ..., t_n\}$. For different visual inputs, such as images or videos, we use the same visual backbone to learn visual features and establish the association between the product domain and micro-video domain during the alignment with the query. We treat images as a special case of videos that only contain one frame. Unlike commonly used video datasets, e-commerce videos are more concerned with the items appearing in the video rather than coherent actions. Therefore, we introduce a parameter-efficient approach to model video features, leveraging the parameters learned in the image backbone as much as possible.

For a video clip $V \in \mathbb{R}^{T \times H \times W \times 3}$, of $T$ sampled frames with $H$ and $W$ denote the spatial resolution, following ViT [6], we utilize the [cls] token obtained by passing each image through the ViT encoder as the frame-level representation $h$, denoted as $h = [h_1, h_2, ..., h_T]$. We propose a Frame-level Temporal Transformer(FTT) for generating video-level features $v = \{v_{cls}, v_1, v_2, ..., v_m\}$, which are shown below.

$$v = \text{Avg}(\text{FTT}(h + e^{temp})) \tag{4}$$

where Avg and $e^{temp}$ denote the average pooling and temporal position encoding, respectively. For $e^{temp}$, we adopt standard learnable absolute position embeddings. The FTT is constructed using the standard multi-head self-attention and feed-forward networks [31].

Furthermore, we design a Fusion encoder to merge the textual and visual features into a unified feature representation, denoted as $u = \{u_{cls}, u_1, u_2, ..., u_i\}$. Consistent with existing multimodal fusion models [16, 18], we use the last 6 layers of the $\text{BERT}_{\text{base}}$ model as the fusion model. The image features $v$ are then fused with the text features $t$ through cross-attention at each layer of the fusion model.

Since the fusion feature $u_{cls}$, visual feature $v_{cls}$, and text feature $t_{cls}$ of a product should express the same semantic information as much as possible, we design a three-way alignment contrastive loss guided by the query feature $q_{cls}$, which is expressed as:

$$L_{q,x} = -\log \frac{\exp s(q_{cls}, x_{cls})/\tau}{\sum_{i=1}^{N} \exp s(q_{cls}, x_{cls,i})/\tau} \tag{5}$$

where $q_{cls}$ and $x_{cls}$ are the input feature vectors for the query and positive/negative samples, respectively, and $x \in \{u, v, t\}$. $N$ is the batch size, and $s(q_{cls}, x_{cls})$ is the similarity between the two vectors, $\tau$ is a learnable temperature parameter. The total cross-domain contrastive loss $L_{cross}$ based on the query and product features can be defined as:

$$L_{cross} = L_{q,u} + L_{q,v} + L_{q,t} \tag{6}$$

As the query that guides the retrieval process and the title or description contained in the product are both in the text modality, to prevent the search results from being dominated by text and ignoring the visual information, we adopt a generation-based strategy to enhance the visual information. We enhance the visual information using a generation-based strategy, as the guiding query and product text features belong to the same textual modality. This is done to prevent the search results from being dominated by the text and ignoring the visual information. We implement the enhancement strategy through two aspects: text-based reconstruction and vision-based reconstruction. For the former, we take the query text as input to the self-attention layer (SA), and use the output of SA as queries, with $v_{cls}$ as keys and values input to the cross-attention layer (CA). The reconstructed text feature can be represented as:

$$\begin{cases} z'_l = \text{SA}(\text{LN}(z_{l-1})) + z_{l-1}, l = 1...L, \\ z_{re,query} = \text{MLP}(\text{LN}((\text{CA}(z'_l, v_{cls})))) \end{cases} \tag{7}$$

We employ the text-based reconstruction loss defined as $L_{re,q}$ to minimize the distance between the generated text $z_{re,query}$ and the target text $q$. Specifically, we use the cross-entropy loss function as a metric to measure the distance. Similarly, to enhance the weight of visual information in the fusion features, based on visual reconstruction, we use images or video frames as the input of the SA layer and use the output of the SA layer as the queries of the CA layer. We use the fusion feature $u_{cls}$ as the keys and values of the CA layer to obtain the reconstructed visual features $z_{re,vis}$. The loss function for visual reconstruction can be expressed as $L_{re,v}$. $L_{recon}$ is defined as the sum of $L_{re,q}$ and $L_{re,v}$. The loss function of the entire cross-domain retrieval framework $L_{retrieval}$ can be represented as:

$$L_{retrieval} = L_{cross} + L_{recon} \tag{8}$$

# 5 EXPERIMENTS

## 5.1 Experimental Setup

**Implementation Details.** We adopt $\text{Chinese}-\text{CLIP}_{\text{ViT}-\text{B}/16}$[33] to initialize the vision and text backbone of our pretraining framework, while the remaining modules are randomly initialized. The visual backbone used in our approach is the ViT model, which contains 86M parameters, and the textual backbone is the RoBERTa-wwm-Base model, which contains 102M parameters. We set the hidden state size and other baselines to 768. We set the maximum sequence length of the query to 20 and that of the product text to 90. We train the pretraining framework and the retrieval framework with a batch size of 80 for 10 epochs on 32 Tesla V100 GPUs. Temperature parameter $\tau$ is set to 0.07. We extract 5 frames for each video. We use Adam [14] optimizer with an initial learning rate of 1e-2 for the pretraining framework. In the retrieval framework, the learning rate for the visual backbone and the generative model is set to 1e-5, while the learning rate for the remaining components is set to 1e-4. A linear learning rate decay schedule is adopted.

**Baselines.** We use 5 commonly used pre-trained models as baselines in our experiments. Among them, XCLIP [22] is the pre-trained model for videos, and we treat images as 1-frame videos for input. CLIP [25], ChineseCLIP [33], CyCLIP [9] are pre-trained vision-language models, and FashionCLIP [2] is a model tailored for e-commerce scenarios. We use the output of video frames with Average Pooling as the video domain representation.

**Evaluation Metrics.** Consistent with existing cross-modal retrieval tasks [1, 7, 10], we use the standard top-K recall metric to evaluate the performance of our model, denoted as R@K. Specifically, we adopt R@10, R@20, R@50, R@100 and their mean as our evaluation metrics.

## 5.2 Experimental Results

**Quantitative results.** To facilitate experimental comparisons and future related work, we first construct a smaller dataset called PKU Real400K, which has the same characteristics as Real20M but on a smaller scale. We select 100,000 queries with their corresponding products and micro-videos from the 1,000,000 queries in Real20M, which are also multimodal. We conduct comparisons with baseline methods and ablation experiments on the Real400K dataset and demonstrate our approach's performance on Real400K and Real20M separately in the ablation experiments. Real400K is used as the training set. In the test set, the query part consists of 400/3000 aligned and manually labeled products/micro-videos. The gallery set contains a total of 200,000 products/micro-videos separately.

We conduct four sets of experiments to showcase the effectiveness of our proposed approach. *In the first set*, we directly test five

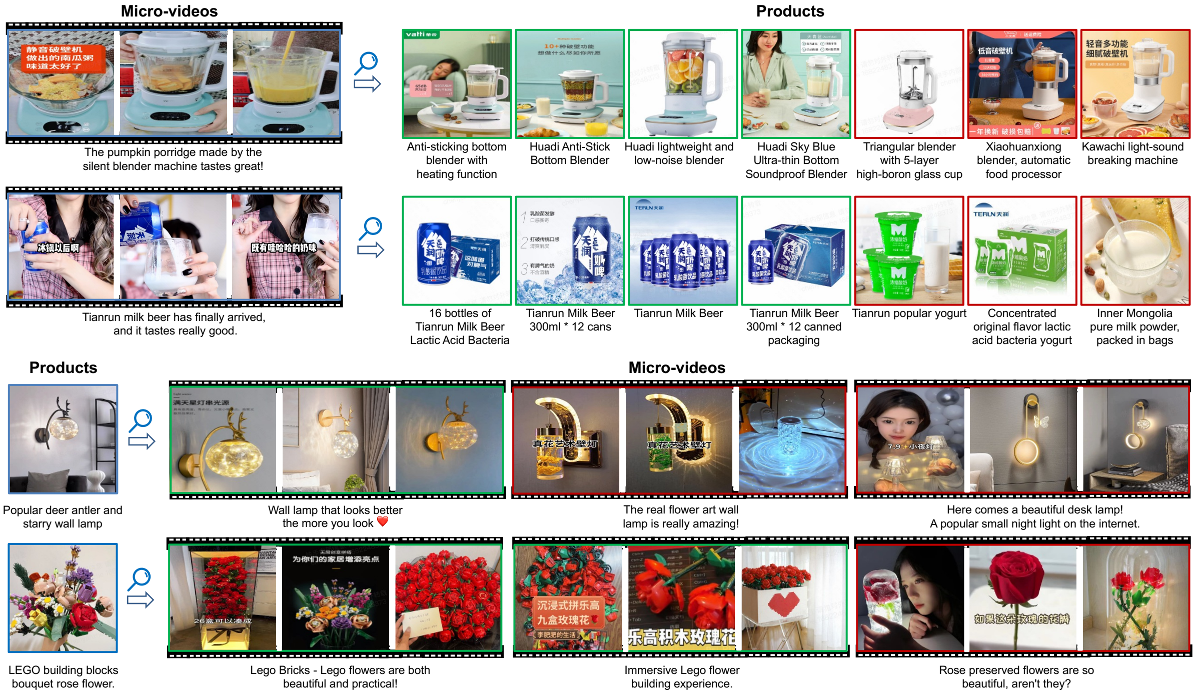
Micro-videos
Products
The pumpkin porridge made by the silent blender machine tastes great!
Anti-sticking bottom blender with heating function
Huadi Anti-Stick Bottom Blender
Huadi lightweight and low-noise blender
Huadi Sky Blue Ultra-thin Bottom Soundproof Blender
Triangular blender with 5-layer high-boron glass cup
Xiaohuanxiong blender, automatic food processor
Kawachi light-sound breaking machine
Tianrun milk beer has finally arrived, and it tastes really good.
16 bottles of Tianrun Milk Beer Lactic Acid Bacteria
Tianrun Milk Beer 300ml * 12 cans
Tianrun Milk Beer
Tianrun Milk Beer 300ml * 12 canned packaging
Tianrun popular yogurt
Concentrated original flavor lactic acid bacteria yogurt
Inner Mongolia pure milk powder, packed in bags
Products
Micro-videos
Popular deer antler and starry wall lamp
Wall lamp that looks better the more you look ❤
The real flower art wall lamp is really amazing!
Here comes a beautiful desk lamp! A popular small night light on the internet.
LEGO building blocks bouquet rose flower.
Lego Bricks - Lego flowers are both beautiful and practical!
Immersive Lego flower building experience.
Rose preserved flowers are so beautiful, aren't they?

**Figure 5: Qualitative results on Real20M. We show reference products or micro-videos with blue boxes on the left and top-k retrievals with descending scores on the right. Ground truths are shown with green boxes, others are shown in red boxes.**

**Table 2: Best scores are highlighted in bold, and second-best scores are underlined. "infer" represents directly applying the model to testing, "ft" represents fine-tuning the model on the dataset, "Q" represents integrating the model with our Query-driven cross-domain retrieval framework.**

| Method | Video2goods | | | | Goods2video | | | | Overall | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | R@10 | R@20 | R@50 | R@100 | R@10 | R@20 | R@50 | R@100 | R@10 | R@20 | R@50 | R@100 | **Mean** |
| CLIP-infer [25] | 14.62 | 19.87 | 27.59 | 33.08 | 19.11 | 24.05 | 32.31 | 38.78 | 16.87 | 21.96 | 29.95 | 35.93 | 26.18 |
| CyCLIP-infer [9] | 15.22 | 19.89 | 27.69 | 34.56 | 20.11 | 24.59 | 32.55 | 38.92 | 17.67 | 22.24 | 30.12 | 36.74 | 26.69 |
| FashionCLIP-infer [2] | 10.47 | 14.03 | 20.09 | 24.43 | 15.32 | 19.08 | 25.06 | 30.40 | 12.90 | 16.92 | 22.58 | 27.42 | 19.86 |
| XCLIP-infer [22] | 9.21 | 11.15 | 13.73 | 15.85 | 17.20 | 18.82 | 21.85 | 24.95 | 13.21 | 14.99 | 17.79 | 20.27 | 16.60 |
| ChineseCLIP-infer [33] | 23.34 | 30.92 | 39.52 | 47.47 | 27.30 | 34.46 | 43.83 | 51.78 | 25.32 | 32.69 | 41.68 | 49.63 | 37.33 |
| CLIP-ft | 23.88 | 32.14 | 41.55 | 49.86 | 27.45 | 32.71 | 41.08 | 47.96 | 25.67 | 32.43 | 41.32 | 48.91 | 37.08 |
| CyCLIP-ft | 24.95 | 31.39 | 41.60 | 49.62 | 28.57 | 33.69 | 42.13 | 48.78 | 26.76 | 32.54 | 41.87 | 49.20 | 37.59 |
| FashionCLIP-ft | 24.09 | 32.12 | 41.33 | 49.05 | 27.85 | 34.64 | 44.78 | 52.94 | 25.97 | 33.38 | 43.06 | 51.00 | 38.35 |
| XCLIP-ft | 26.49 | 34.37 | 45.36 | 54.91 | 31.51 | 38.20 | 47.07 | 54.14 | 29.00 | 36.29 | 46.22 | 54.53 | 41.51 |
| ChineseCLIP-ft | 30.09 | 38.09 | 48.66 | 57.33 | 34.00 | 38.87 | 47.34 | 53.73 | 32.05 | 38.48 | 48.00 | 55.53 | 43.51 |
| CLIP-Q | 28.16 | 36.27 | 46.58 | 55.18 | 36.90 | 42.89 | 53.48 | 61.31 | 32.53 | 39.57 | 50.03 | 58.25 | 45.10 |
| CyCLIP-Q | 28.78 | 36.05 | 46.09 | 54.63 | 36.49 | 42.40 | 53.83 | 61.69 | 32.64 | 39.23 | 49.96 | 58.16 | 45.00 |
| FashionCLIP-Q | 34.84 | 44.20 | 55.63 | 64.84 | 38.54 | 45.93 | 55.08 | 63.11 | 36.69 | 45.07 | 55.36 | 63.98 | 50.27 |
| XCLIP-Q | 34.20 | 43.60 | 54.51 | 64.22 | 43.94 | 53.36 | 62.02 | 71.24 | 39.07 | 48.48 | 58.27 | 67.73 | 53.38 |
| ChineseCLIP-Q | 33.04 | 41.46 | 52.07 | 59.39 | 45.39 | 54.09 | 63.07 | 70.35 | 39.22 | 47.78 | 57.57 | 64.87 | 52.36 |
| **Ours-Q** | <u>36.66</u> | <u>46.97</u> | <u>57.91</u> | <u>66.85</u> | <u>47.91</u> | <u>55.64</u> | <u>64.61</u> | <u>72.45</u> | <u>42.29</u> | <u>51.31</u> | <u>61.26</u> | <u>69.65</u> | <u>56.13</u> |
| **Ours-Q + Pretrain** | **38.90** | **49.06** | **59.37** | **67.82** | **49.46** | **57.31** | **67.98** | **73.60** | **44.18** | **53.19** | **63.68** | **70.71** | **57.94** |

baseline methods on the testing dataset. However, the results indicated that these methods exhibited poor performance despite being pre-trained on luxury product data and a large-scale Chinese corpus as reported in [24]. This benchmark has brought new challenges to existing methods, mainly because our benchmark is different from previous single-domain retrieval tasks. We require cross-domain retrieval and the construction of connections between different domains. *In the second set of experiments*, we utilize the commonly employed content-based alignment technique, which uses a Siamese network to model the product and micro-video domains separately,

**Table 3: Ablation study on our proposed dataset.**

| Method | Video2goods | | | | Goods2video | | | | Overall | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | R@10 | R@20 | R@50 | R@100 | R@10 | R@20 | R@50 | R@100 | R@10 | R@20 | R@50 | R@100 | Mean |
| base (w/o Recon) | 31.11 | 40.03 | 51.10 | 58.20 | 42.17 | 51.78 | 61.43 | 68.02 | 36.64 | 45.91 | 56.27 | 63.11 | 50.48 |
| base (w Recon) | 33.04 | 41.46 | 52.07 | 59.39 | 45.39 | 54.09 | 63.07 | 70.35 | 39.22 | 47.78 | 57.57 | 64.87 | 52.36 |
| base + FTT | 32.84 | 43.39 | 54.36 | 62.49 | 46.01 | 54.90 | 64.12 | 71.72 | 39.43 | 49.15 | 59.24 | 67.11 | 53.73 |
| base + FTT + Fusion (Ours-Q) | 36.66 | 46.97 | 57.91 | 66.85 | 47.91 | 55.64 | 64.61 | 72.45 | 42.29 | 51.31 | 61.26 | 69.65 | 56.13 |

**Table 4: Ablation experiments using different pretraining setting and scales. Q + P denotes Ours-Q + Pretrain.**

| Methods | R@10 | R@20 | R@50 | R@100 | Mean |
|---|---|---|---|---|---|
| Ours-Q | 42.29 | 51.31 | 61.26 | 69.65 | 56.13 |
| Q + P (w/o EPL) | 42.78 | 52.11 | 62.53 | 70.00 | 56.86 |
| Q + P (w EPL) | 44.18 | 53.19 | 63.68 | 70.71 | 57.94 |
| Q + P (Real20M) | 51.59 | 60.61 | 69.99 | 79.83 | 65.50 |

**Table 5: Ablation experiments using different visual-text fusion models.**

| Methods | R@10 | R@20 | R@50 | R@100 | Mean |
|---|---|---|---|---|---|
| Add | 39.43 | 49.15 | 59.24 | 67.11 | 53.73 |
| Combiner [1] | 40.35 | 49.64 | 59.98 | 68.87 | 54.71 |
| Fusion (Ours) | 42.29 | 51.31 | 61.26 | 69.65 | 56.13 |

to align the features extracted from both domains. *The third set of experiments* introduces our query-driven cross-domain retrieval framework, which incorporates query-driven cross-domain contrastive loss and reconstruction loss to comprehend user intentions and prevent text modality from dominating the retrieval outcomes separately. Through comparing the results obtained from these two sets, we observe that both methods align different domains and establish interconnections between them, leading to a certain level of improvement in performance compared to the first set of experiments. Furthermore, our proposed cross-domain framework can be applied to existing techniques to enhance performance. Notably, the query-driven cross-domain retrieval approach outperforms the content-based method, mainly due to the latter's insufficient understanding of user intent, which poses challenges in capturing the true intent in the micro-video or product domains. In contrast, our proposed cross-domain framework can alleviate this issue, resulting in better average performance increased by 8.02%, 7.41%, 11.92%, 11.87%, and 8.85% for the five methods respectively. *In the fourth set of experiments*, we adopt the ChineseCLIP method as the baseline and add the Frame-level Temporal Transformer (FTT) and Fusion module to obtain fusion features that fully interact with visual and textual information. Additionally, we add the three-level entity-based pre-training method (EPL) to the three commonly used pre-training methods. These two experiments are denoted as Ours-Q and Ours-Q + Pretrain respectively, resulting the improvements of 3.77% and 5.58%, compared to the ChineseCLIP-Q method.

**Qualitative Analyses.** In Figure 5, we showcase the retrieval outcomes of our proposed method on the testing dataset. Specifically, the top two rows demonstrate the retrieval of products via micro-videos, while the bottom two rows illustrate the retrieval of micro-videos via products.

## 5.3 Ablation Studies

To investigate the effectiveness of our approach, we evaluate the key designs in our frameworks on the Real400K dataset, shown in Table 3 and Table 4. In Table 3, we use ChineseCLIP-Q as the base and show the results of the base model without Reconstruction Loss, the base model with Reconstruction Loss, the base model with the FTT model, and the base model with both the FTT and Fusion models. In Table 4, we showcase the performance of our proposed Ours-Q method, along with its variants incorporating VTC, VTM, and MLM pre-training tasks, and further augmented with the EPL pre-training task. By comparing the experimental results from the table, each component adopted in the experiment plays a role in improving the final results. Our proposed query-driven approach is efficient in acquiring massive and diversified data. We also evaluate our methods on the full-scale dataset Real20M and achieve significant improvement compared to Real400K, shown in Table 4, with an average increase of 7.56% in the mean metric.

We also test the effectiveness of typical multimodal fusion methods. Specifically, we adopt the method of directly adding features, represented as Add, the SOTA method on FashionIQ [1], represented as Combiner, and our fusion method, represented as Fusion. As shown in Table 5, our Fusion model achieved better results.

# 6 CONCLUSION

In this paper, we introduce a benchmark for cross-domain retrieval and develop a dataset specifically for this purpose, called PKU Real20M. In contrast to existing e-commerce datasets, PKU Real20M not only contains products and micro-videos, but also both domains are multimodal. Additionally, our dataset is constructed based on search, which helps establish query-driven associations between massive amounts of products and micro-videos. We propose a strong baseline method including a pre-training framework based on three-level entity words prompt learning and a Query-driven Cross-Domain retrieval framework (QCD). We hope that our dataset and baseline will help bridge the gap between research and practical applications in the e-commerce field.

# ACKNOWLEDGMENTS

This work was supported by the grants from the National Key R&D Program of China (2021YFF0901502), the National Natural Science Foundation of China (61925201, 62132001), Beijing Natural Science Foundation (4232005) and Kuaishou.

# REFERENCES

[1] Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. 2022. Effective conditioned and composed image retrieval combining CLIP-based features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 21466–21474.

[2] Patrick John Chia, Giuseppe Attanasio, Federico Bianchi, Silvia Terragni, Ana Rita Magalhães, Diogo Goncalves, Ciro Greco, and Jacopo Tagliabue. 2022. Fashionclip: Connecting language and images for product representations. *arXiv preprint arXiv:2204.03972* (2022).

[3] Charles Corbière, Hédi Ben-Younes, Alexandre Ramé, and Charles Ollion. 2017. Leveraging Weakly Annotated Data for Fashion Image Retrieval and Label Prediction. In *2017 IEEE International Conference on Computer Vision Workshops, ICCV Workshops 2017, Venice, Italy, October 22-29, 2017.* IEEE Computer Society, 2268–2274. https://doi.org/10.1109/ICCVW.2017.266

[4] Ginger Delmas, Rafael Sampaio de Rezende, Gabriela Csurka, and Diane Larlus. 2022. Artemis: Attention-based retrieval with text-explicit matching and implicit similarity. *arXiv preprint arXiv:2203.08101* (2022).

[5] Xiao Dong, Xunlin Zhan, Yangxin Wu, Yunchao Wei, Michael C. Kampffmeyer, Xiaoyong Wei, Minlong Lu, Yaowei Wang, and Xiaodan Liang. 2022. M5Product: Self-harmonized Contrastive Learning for E-commercial Multi-modal Pretraining. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022.* IEEE, 21220–21230. https://doi.org/10.1109/CVPR52688.2022.02057

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021.* OpenReview.net. https://openreview.net/forum?id=YicbFdNTTy

[7] Xuri Ge, Fuhai Chen, Songpei Xu, Fuxiang Tao, and Joemon M Jose. 2023. Cross-modal Semantic Enhanced Interaction for Image-Sentence Retrieval. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision.* 1022–1031.

[8] Yuying Ge, Ruimao Zhang, Xiaogang Wang, Xiaoou Tang, and Ping Luo. 2019. Deepfashion2: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.* 5337–5345.

[9] Shashank Goel, Hritik Bansal, Sumit Bhatia, Ryan Rossi, Vishwa Vinay, and Aditya Grover. 2022. Cyclip: Cyclic contrastive language-image pretraining. *Advances in Neural Information Processing Systems* 35 (2022), 6704–6719.

[10] Sonam Goenka, Zhaoheng Zheng, Ayush Jaiswal, Rakesh Chada, Yue Wu, Varsha Hedau, and Pradeep Natarajan. 2022. Fashionvlp: Vision language transformer for fashion retrieval with feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 14105–14115.

[11] M Hadi Kiapour, Xufeng Han, Svetlana Lazebnik, Alexander C Berg, and Tamara L Berg. 2015. Where to buy it: Matching street clothing photos in online shops. In *Proceedings of the IEEE international conference on computer vision.* 3343–3351.

[12] Xintong Han, Zuxuan Wu, Phoenix X. Huang, Xiao Zhang, Menglong Zhu, Yuan Li, Yang Zhao, and Larry S. Davis. 2017. Automatic Spatially-Aware Fashion Concept Discovery. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017.* IEEE Computer Society, 1472–1480. https://doi.org/10.1109/ICCV.2017.163

[13] Weixiang Hong, Kaixiang Ji, Jiajia Liu, Jian Wang, Jingdong Chen, and Wei Chu. 2021. Gilbert: Generative vision-language pre-training for image-text retrieval. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval.* 1379–1388.

[14] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[15] Seungmin Lee, Dongwan Kim, and Bohyung Han. 2021. Cosmo: Content-style modulation for image retrieval with text feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 802–812.

[16] Dongxu Li, Junnan Li, Hongdong Li, Juan Carlos Niebles, and Steven CH Hoi. 2022. Align and prompt: Video-and-language pre-training with entity prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 4953–4963.

[17] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. 2020. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 11336–11344.

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

[19] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. 2016. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition.* 1096–1104.

[20] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. 2022. CLIP4Clip: An empirical study of CLIP for end to end video clip retrieval and captioning. *Neurocomputing* 508 (2022), 293–304.

[21] Andrei Neculai, Yanbei Chen, and Zeynep Akata. 2022. Probabilistic Compositional Embeddings for Multimodal Image Retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 4547–4557.

[22] Bolin Ni, Houwen Peng, Minghao Chen, Songyang Zhang, Gaofeng Meng, Jianlong Fu, Shiming Xiang, and Haibin Ling. 2022. Expanding language-image pretrained models for general video recognition. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IV.* Springer, 1–18.

[23] Liqiang Nie, Leigang Qu, Dai Meng, Min Zhang, Qi Tian, and Alberto Del Bimbo. 2022. Search-oriented Micro-video Captioning. In *Proceedings of the 30th ACM International Conference on Multimedia.* 3234–3243.

[24] Manish Pathak and Aditya Jain. 2021. Solving Fashion Recommendation–The Farfetch Challenge. *arXiv preprint arXiv:2108.01314* (2021).

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

[26] Negar Rostamzadeh, Seyedarian Hosseini, Thomas Boquet, Wojciech Stokowiec, Ying Zhang, Christian Jauvin, and Chris Pal. 2018. Fashion-gen: The generative fashion dataset and challenge. *arXiv preprint arXiv:1806.08317* (2018).

[27] Rohan Sarkar, Navaneeth Bodla, Mariya Vasileva, Yen-Liang Lin, Anurag Beniwal, Alan Lu, and Gerard Medioni. 2022. Outfittransformer: Outfit representations for fashion recommendation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 2263–2267.

[28] Lei Shi, Kai Shuang, Shijie Geng, Peng Su, Zhengkai Jiang, Peng Gao, Zuohui Fu, Gerard de Melo, and Sen Su. 2020. Contrastive visual-linguistic pretraining. *arXiv preprint arXiv:2007.13135* (2020).

[29] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2019. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530* (2019).

[30] Yuxin Tian, Shawn Newsam, and Kofi Boakye. 2023. Fashion Image Retrieval With Text Feedback by Additive Attention Compositional Learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision.* 1011–1021.

[31] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (Eds.). 5998–6008. https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[32] Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogério Feris. 2021. Fashion IQ: A New Dataset Towards Retrieving Images by Natural Language Feedback. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021.* Computer Vision Foundation / IEEE, 11307–11317. https://doi.org/10.1109/CVPR46437.2021.01115

[33] An Yang, Junshu Pan, Junyang Lin, Rui Men, Yichang Zhang, Jingren Zhou, and Chang Zhou. 2022. Chinese CLIP: Contrastive Vision-Language Pretraining in Chinese. *arXiv preprint arXiv:2211.01335* (2022).

[34] Xunlin Zhan, Yangxin Wu, Xiao Dong, Yunchao Wei, Minlong Lu, Yichi Zhang, Hang Xu, and Xiaodan Liang. 2021. Product1M: Towards Weakly Supervised Instance-Level Product Retrieval via Cross-Modal Pretraining. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021.* IEEE, 11762–11771. https://doi.org/10.1109/ICCV48922.2021.01157

[35] Hongguang Zhu, Yunchao Wei, Yao Zhao, Chunjie Zhang, and Shujuan Huang. 2023. AMC: Adaptive Multi-expert Collaborative Network for Text-guided Image Retrieval. *ACM Transactions on Multimedia Computing, Communications and Applications* (2023).

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

# A SUPPLEMENTARY MATERIAL

## A.1 Comparing Query-based and Content-based

Due to the possibility of micro-videos containing various products in real-world scenarios, and the difficulty of using video-related text to accurately identify the main object in the video [23], content-based methods struggle to accurately associate micro-videos with the target product. Our query-driven approach can guide the model

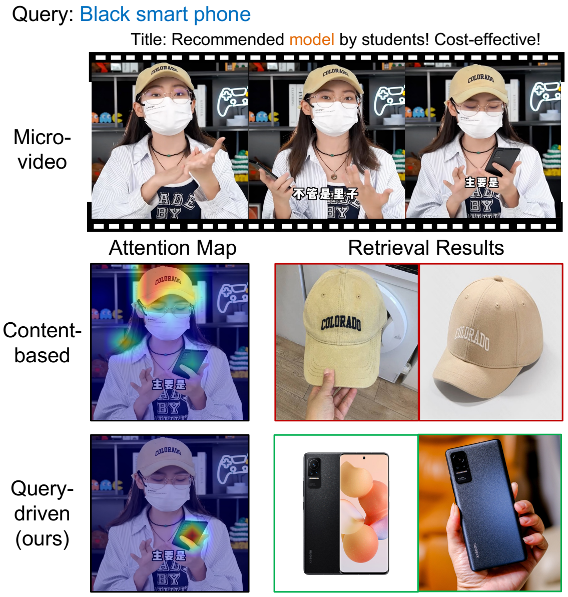

Figure 6: The comparison between our proposed query-based and content-based retrieval results.

Table 6: Ablation experiments using different video modeling models.

| Methods | R@10 | R@20 | R@50 | R@100 | Mean |
|---|---|---|---|---|---|
| Add | 39.22 | 47.78 | 57.57 | 64.87 | 52.36 |
| Mean Pooling | 38.71 | 48.11 | 58.13 | 65.89 | 52.71 |
| LSTM | 39.01 | 47.80 | 57.82 | 65.11 | 52.44 |
| FTT (Ours) | 39.43 | 49.15 | 59.24 | 67.11 | 53.73 |

to learn the intention of the micro-video through user queries, enabling accurate retrieval of the target product, as shown by the attention map and top two retrieval results in Figure 6.

## A.2 Comparing different video modeling models

Additionally, we also evaluate the results of different fusion methods for videos based on CLIP [20, 22], as shown in Table 6. From the results, FTT performs better than the other methods.