Multi-queue Momentum Contrast for Microvideo-Product Retrieval*

Yali Du
Nanjing University
duyali2000@gmail.com

Yinwei Wei†
National University of Singapore
weiyinwei@hotmail.com

Wei Ji
National University of Singapore
jiwei@nus.edu.sg

Fan Liu
National University of Singapore
liufancs@gmail.com

Xin Luo
Shandong University
luoxin.lxin@gmail.com

Liqiang Nie
Harbin Institute of Technology (Shenzhen)
nieliqiang@gmail.com

ABSTRACT

The booming development and huge market of micro-videos bring new e-commerce channels for merchants. Currently, more micro-video publishers prefer to embed relevant ads into their micro-videos, which not only provides them with business income but helps the audiences to discover their interesting products. However, due to the micro-video recording by unprofessional equipment, involving various topics and including multiple modalities, it is challenging to locate the products related to micro-videos efficiently, appropriately, and accurately. We formulate the microvideo-product retrieval task, which is the first attempt to explore the retrieval between the multi-modal and multi-modal instances.

A novel approach named Multi-Queue Momentum Contrast (MQMC) network is proposed for bidirectional retrieval, consisting of the uni-modal feature and multi-modal instance representation learning. Moreover, a discriminative selection strategy with a multi-queue is used to distinguish the importance of different negatives based on their categories. We collect two large-scale microvideo-product datasets (MVS and MVS-large) for evaluation and manually construct the hierarchical category ontology, which covers sundry products in daily life. Extensive experiments show that MQMC outperforms the state-of-the-art baselines. Our replication package (including code, dataset, etc.) is publicly available at https://github.com/duyali2000/MQMC.

CCS CONCEPTS

• **Information systems** → **Multimedia and multimodal retrieval**; **Video search**; **Image search**.

KEYWORDS

Datasets, Momentum Contrast, Multi-Modal Retrieval, Microvideo-Product, Multi-Queue

*This research was partially supported by NSFC (62076121)
†Yinwei Wei is the corresponding author.

WSDM '23, February 27-March 3, 2023, Singapore, Singapore

ACM ISBN 978-1-4503-9407-9/23/02…$15.00
https://doi.org/10.1145/3539597.3570405

ACM Reference Format:
Yali Du, Yinwei Wei, Wei Ji, Fan Liu, Xin Luo, and Liqiang Nie. 2023. Multi-queue Momentum Contrast for Microvideo-Product Retrieval. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining (WSDM '23), February 27-March 3, 2023, Singapore, Singapore.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3539597.3570405

1 INTRODUCTION

Micro-video, as a new form of social media, has become an important component in our daily life. Compared with long videos, micro-videos have the instincts of short duration and easy-to-share, making them popular in online sharing platforms. Taking Tiktok as an example, its Monthly Active Users had reached one billion until September 2021[1]. Such a huge market undoubtedly brings new E-commerce chance for merchants. Currently, some publishers start to embed relevant ads into their micro-videos, which not only provides them external income but facilitates the audiences to discover their interested products[17, 19, 21, 23, 25, 26, 43, 45–48].

However, the irrelevant products may harm the audiences' experience and make them disappointed with the micro-video itself. Hence, how to locate the products related to micro-videos challenges micro-video publishers. To remedy this problem, it is of great importance to understand the semantic information of the micro-video and product, so as to measure their affinities. The prior studies hence frame this problem as the video-shop retrieval task. For instance, Zhao *et. al.,* [52] proposed a DPRnet model to discover the keyframe from videos and measure visual similarities with other candidate items. Considering the temporal relation hidden in the video, Cheng *et. al.,* [7] applied the LSTM model to represent the video and developed an AsymNet model to discover the matched items according to their distances to the video. More recently, Godi *et. al.,* [10] constructed a new dataset composed of social videos and their corresponding products' images, proposed SEAM Match-RCNN to extract features from video, and located the visually similar clothes.

Despite their remarkable performance, it is hard to directly adopt these methods to seek the relevant products for micro-videos, due to the following problem:

- The micro-video tends to be shot by amateurs with some non-professional equipment, like a mobile phone and pad. Thereby, the micro-video inevitably contains some environmental noise, resulting in sub-optimal representations of micro-videos.

[1]https://new.qq.com/omn/20210928/20210928A01WW400.html

- Unlike the videos shot by merchants, the micro-video is not designed for the specific product. Hence, the micro-video may involve various topics, which is more challenging than traditional video-shop retrieval.
- The micro-video and product embrace the signal from the multiple modalities, including the visual and textual cues. Therefore, not only cross-modal but intra-modal relations should be considered in microvideo-product matching.

To resolve this problem, we resort to address two technical challenges: (1) *how to distill the informative signal relevant to the product from the content information of micro-video*, and (2) *how to represent the multi-modal micro-video and product to model their similarities*.

Therefore, we develop a new microvideo-product retrieval model, termed Multi-Queue Momentum Contrast (MQMC) method, which consists of the uni-model feature representation learning and multi-modal instance representation learning. To optimize the uni-modal feature encoders, we introduce cross-modal contrastive loss and intra-modal contrastive loss. The former models the consistency between the visual and textual modalities while the latter utilizes the supervision signal from the product information to help the encoders to distinguish the informative signal from the irrelevant content of the micro-video. In multi-modal instance representation learning, we resort to the momentum-based contrastive loss[14] to model the instance-level similarity. Considering different negative micro-videos (products) play various importance to the anchor products (micro-videos), we take a negative selection strategy with multi-queue to distinguish the importance of different negatives by measuring the distance of categories of anchor and negatives.

To evaluate our proposed model, we collect the microvideo-product pairs from the popular micro-video sharing platforms and achieve two datasets: MVS and MVS-large, which contain 13, 165 and 126, 206 microvideo-product pairs respectively. In addition, we manually construct the hierarchical category ontology including 6 upper ontologies, 30 middle ontologies, and 316 lower ontologies. By conducting extensive experiments on these two datasets, MQMC significantly outperforms the state-of-the-art methods, which demonstrates the effectiveness of our proposed model.

In a nutshell, our contributions could be summarized as follows:

- By investigating the prior studies on cross-modal information retrieval, we formulate a new microvideo-product retrieval task. To the best of our knowledge, this is the first attempt to explore the retrieval between the multi-modal and multi-modal instances.
- We propose a novel Multi-Queue Momentum Contrast (MQMC) network consisting of the uni-modal feature and multi-modal instance representation learning, so as to locate the relevant micro-video (product) for the inputted product (micro-video).
- We design a new multi-queue contrastive training strategy, which maintains multiple queues of negative samples and considers the importance of different negatives based on their categories in contrastive loss computation.
- To evaluate our proposed model, we construct two large-scale datasets *i.e.,* MVS and MVS-large and build the hierarchical category ontology of the products. By conducting extensive experiments on the datasets, we demonstrate that our proposed model outperforms the state-of-the-art baselines by a margin.

2 RELATED WORK

2.1 Cross-modal information retrieval

With the explosion of multi-modal data, cross-modal retrieval has attracted wide attention, mainly focusing on image-text, video-text, video-image, etc[12, 16, 20, 22, 24–27, 31, 33, 36, 38, 39, 47, 49, 50]. CLIP[33] used a sufficiently large dataset for pre-training and natural language as a supervisory signal to learn visual representation. ALBEF[20] introduced a contrastive loss to align the image and text representations before fusing them through cross-modal attention. Hit[27] combined hierarchical transformer and momentum contrast method for video-text retrieval. MMT[9] learned an effective representation from different modalities inherent in video over multiple self-attention layers with several video feature extractors. Most of all, the above methods are based on the tasks of single-modal to single-modal, or single-modal to multi-modal, which do not apply to the task of multi-modal to multi-modal retrieval in our paper.

The recent influx of instructional multi-modal datasets such as Inria Instructional Videos[2], CrossTask[54], YouCook2[53], and HowTo100M[31] has inspired a variety of methods for video-text retrieval, but those are not suitable for the task of microvideo-product retrieval in this paper. AsymNet[7], DPRNet[52], and Fashion Focus[51] built video-to-shop datasets from advertisements in online shopping platforms, but the datasets are not publicly released. Although MovingFashion[11] and Watch and Buy[34] are publicly available datasets of "video-to-shop", they have the disadvantage of a single domain (only clothing).

2.2 Contrastive Learning

Contrastive learning is widely studied in self-supervised and unsupervised learning and has made many remarkable achievements[3–6, 13, 14, 32, 32, 40, 41, 44]. The contrastive learning model is built on the principle that positives are closer to each other in the projection space, while negatives are farther apart. The main challenges are how to choose positives and negatives, how to construct a representation learning model that can follow the above principle, and how to prevent model collapse.

According to the way of choosing negative samples, current contrastive learning methods can be divided into in-batch and out-batch. The end-to-end mechanism uses the anchor's augmented views as positives and considers other samples in the current batch as negatives. SimCLR[4] achieved success in unsupervised visual representation learning, which benefits from large batch size, stronger data augmentation, and the learnable nonlinear projection head. The memory bank mechanism constructs a memory bank to memorize broader negative samples, which has the drawback that the samples in the memory bank are from very different encoders all over the past epoch and they are less consistent. Faced with this problem, MoCo[14] used a momentum-updated key encoder to maintain the consistency of negative representations in the memory bank.

However, increasing the memory size or batch size does not always improve the performance rapidly, because more negatives do not necessarily mean that more difficult negatives are brought. There are many recent improvements in contrastive learning, including loss function, sampling strategy, and data augmentation,

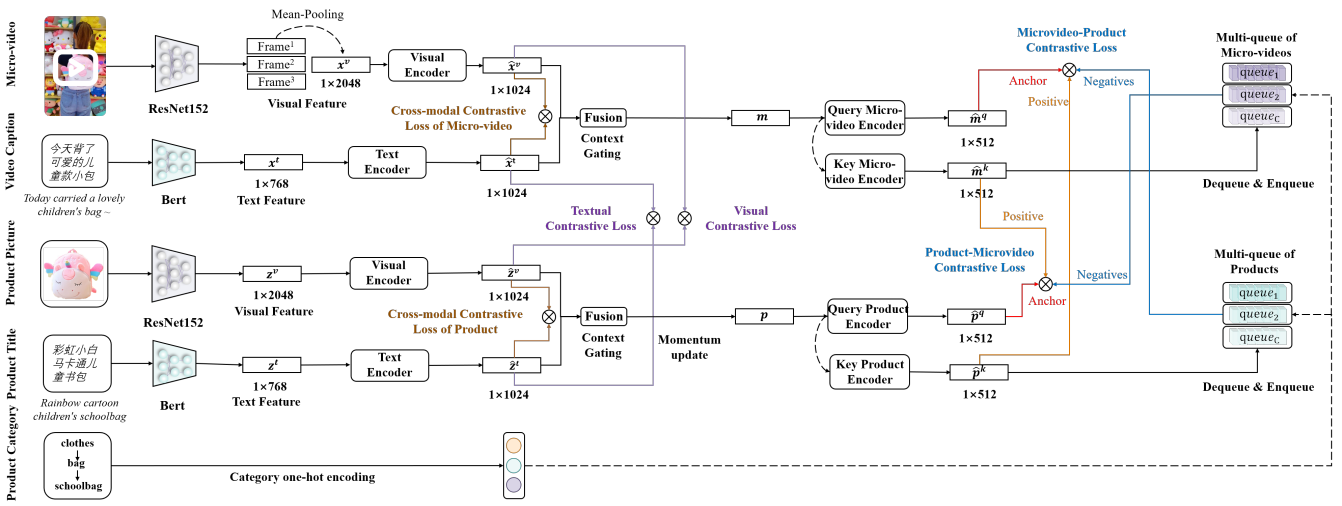


Figure 1: An overview of MQMC for microvideo-product retrieval.

but few relative works on negative samples. The existing works include making use of labels and generating difficult negative samples through Mixup[18, 35].

Our method in multi-modal microvideo-product retrieval benefits from the large-scale negatives using a memory bank and the multi-layer category ontology to distinguish the hard negatives.

3 METHOD

In this section, we first formulate microvideo-product retrieval task, and then detail our proposed model, as shown in Figure 1, consisting of the uni-modal feature representation learning and multi-modal instance representation learning.

3.1 Problem Definition

Given a set $\mathcal{V} = \{v_1, v_2, \ldots, v_M\}$ of M micro-video, the microvideo-product retrieval task aims to discover the most similar product from a set $\mathcal{P} = \{p_1, p_2, \ldots, p_N\}$ of N products. More specifically, for the i-th micro-video, we use a pre-trained ResNet model [15] to extract the visual features from the keyframes. With these features, we perform a mean-pooling operation to obtain the visual feature vector of micro-video $\mathbf{x}_i^v \in \mathbb{R}^{d_v}$, where d_v is the dimension of the visual features[2]. Beyond the visual signal, we consider the caption of i-th micro-video and capture its textual features with the trained BERT model [8], denoted as $\mathbf{x}_i^t \in \mathbb{R}^{d_t}$. Analogously, we obtain the visual and textual features from the images and descriptions of products with the same extractors. Taking j-th product as an example, we denote its visual and textual feature vector as $\mathbf{z}_j^v$ and $\mathbf{z}_j^t$, respectively. Formally, with the obtained features of micro-videos and products, we aim to learn a function to score the similarities of microvideo-product pairs:

$$s_{i,j} = f(\mathbf{x}_i^v, \mathbf{x}_i^t; \mathbf{z}_j^v, \mathbf{z}_j^t), \tag{1}$$

where $f(\cdot)$ is the similarity scoring function and $s_{i,j}$ represents the similarity between the i-th micro-video and j-th product.

[2]Without any particular clarification, all the vectors are in column forms.

3.2 Framework

3.2.1 ***Uni-modal Feature Representation Learning.*** As aforementioned, the visual and textual content of the micro-video contains noise signals, like the surroundings and emotional expression, which are irrelevant to the target product search. Hence, we introduce uni-modal feature encoders for two modalities and design the cross- and intra-modal loss functions to distill the informative signal.

Uni-modal Feature Encoder. With the obtained visual and textual features, we separately adopt the two-layer neural networks equipped with a nonlinear activation function on two modalities:

$$\begin{cases} \hat{\mathbf{x}}_i^v = \boldsymbol{W}_2^v \phi(\boldsymbol{W}_1^v \mathbf{x}_i^v), \\ \hat{\mathbf{x}}_i^t = \boldsymbol{W}_2^t \phi(\boldsymbol{W}_2^t \mathbf{x}_i^t), \end{cases} \tag{2}$$

where $\boldsymbol{W}_{(\cdot)}^v$ and $\boldsymbol{W}_{(\cdot)}^t$ are the trainable parameters in visual and textual modalities, respectively. And, $\phi(\cdot)$ is the *leaky_relu* function [29] in our experiments. $\hat{\mathbf{x}}_i^v \in \mathbb{R}^d$ and $\hat{\mathbf{x}}_i^t \in \mathbb{R}^d$ are the refined vectors in visual and textual modalities, respectively. Wherein, d is the dimension of the mapped spaces. Note that we omit the bias term for briefness.

Moreover, we also implement a two-layer network to map the visual (textual) features of the product into the same space of the refined visual (textual) vector, *i.e.,* $\hat{\mathbf{z}}_i^v \in \mathbb{R}^d$ and $\hat{\mathbf{z}}_i^t \in \mathbb{R}^d$. These mapping operations are conducted for the following two contrastive loss functions.

Cross-modal Contrastive Loss. To optimize the feature encoders, we introduce a cross-modal contrastive loss to explicitly model the consistency between the visual and textual. Specifically, it is implemented by identifying the positive pair of cross-modal vectors, like $<\hat{\mathbf{x}}_i^v, \hat{\mathbf{x}}_i^t>$, learned from the same instance (*i.e.,* micro-video or product) from multiple negative pairs of the vectors from different instances. To be more specific, we treat $\hat{\mathbf{x}}_i^v$ as the anchor and $\hat{\mathbf{x}}_i^t$ as the positive vector. And, we randomly sample multiple textual vectors of other micro-videos as the negative vectors. Similarity, we establish the cross-modal negative pairs for the product, formally,

$$\{\hat{\mathbf{x}}_{i,1}^t, \hat{\mathbf{x}}_{i,2}^t, \ldots, \hat{\mathbf{x}}_{i,K}^t\}; \quad \{\hat{\mathbf{z}}_{i,1}^t, \hat{\mathbf{z}}_{i,2}^t, \ldots, \hat{\mathbf{z}}_{i,K}^t\}, \tag{3}$$

where $\hat{\mathbf{x}}^t_{i,\cdot}$ and $\hat{\mathbf{z}}^t_{j,\cdot}$ are the negative vectors of i-th micro-video and j-th product. And, K is the pre-defined number of negative vectors.

With the anchor, positive, and negative vectors, we opt for the InfoNCE[32] loss, formally,

$$\begin{aligned}\mathcal{L}_1 = &-\ln \frac{\exp(\frac{(\hat{\mathbf{x}}^v_i)^\top \hat{\mathbf{x}}^t_i}{||\hat{\mathbf{x}}^v_i||\cdot||\hat{\mathbf{x}}^t_i||}\cdot\frac{1}{\tau})}{\exp(\frac{(\hat{\mathbf{x}}^v_i)^\top \hat{\mathbf{x}}^t_i}{||\hat{\mathbf{x}}^v_i||\cdot||\hat{\mathbf{x}}^t_i||}\cdot\frac{1}{\tau}) + \sum_{k=1}^{K}\exp(\frac{(\hat{\mathbf{x}}^v_i)^\top \hat{\mathbf{x}}^t_{i,k}}{||\hat{\mathbf{x}}^v_i||\cdot||\hat{\mathbf{x}}^t_{i,k}||}\cdot\frac{1}{\tau})} \\ &-\ln \frac{\exp(\frac{(\hat{\mathbf{z}}^v_i)^\top \hat{\mathbf{z}}^t_i}{||\hat{\mathbf{z}}^v_i||\cdot||\hat{\mathbf{z}}^t_i||}\cdot\frac{1}{\tau})}{\exp(\frac{(\hat{\mathbf{z}}^v_i)^\top \hat{\mathbf{z}}^t_i}{||\hat{\mathbf{z}}^v_i||\cdot||\hat{\mathbf{z}}^t_i||}\cdot\frac{1}{\tau}) + \sum_{k=1}^{K}\exp(\frac{(\hat{\mathbf{z}}^v_i)^\top \hat{\mathbf{z}}^t_{i,k}}{||\hat{\mathbf{z}}^v_i||\cdot||\hat{\mathbf{z}}^t_{i,k}||}\cdot\frac{1}{\tau})},\end{aligned} \tag{4}$$

where τ is a temperature parameter.

Intra-modal Contrastive Loss. Beyond the consistency between different modalities, we further leverage the supervision signal from the product information to optimize the feature encoders of the micro-video. It can help the encoders to distinguish the informative signal from the irrelevant content of the micro-video. Following similar operations, we construct the contrastive pairs of the micro-video and product in each modality. In particular, we group the i-th micro-video and corresponding product as the positive pair, *i.e.,* $< \hat{\mathbf{x}}^v_i, \hat{\mathbf{z}}^v_i >$ and $< \hat{\mathbf{x}}^t_i, \hat{\mathbf{z}}^t_i >$. And then, we randomly sample multiple products as the negative samples:

$$\{\hat{\mathbf{z}}^v_{i,1}, \hat{\mathbf{z}}^v_{i,2}, \ldots, \hat{\mathbf{z}}^v_{i,K}\}, \quad \{\hat{\mathbf{z}}^t_{i,1}, \hat{\mathbf{z}}^t_{i,2}, \ldots, \hat{\mathbf{z}}^t_{i,K}\}. \tag{5}$$

Accordingly, we conduct the contrastive loss, formally,

$$\begin{aligned}\mathcal{L}_2 = &-\ln \frac{\exp(\frac{(\hat{\mathbf{x}}^v_i)^\top \hat{\mathbf{z}}^v_i}{||\hat{\mathbf{x}}^v_i||\cdot||\hat{\mathbf{z}}^v_i||}\cdot\frac{1}{\tau})}{\exp(\frac{(\hat{\mathbf{x}}^v_i)^\top \hat{\mathbf{z}}^v_i}{||\hat{\mathbf{x}}^v_i||\cdot||\hat{\mathbf{z}}^v_i||}\cdot\frac{1}{\tau}) + \sum_{k=1}^{K}\exp(\frac{(\hat{\mathbf{x}}^v_i)^\top \hat{\mathbf{z}}^v_{i,k}}{||\hat{\mathbf{x}}^v_i||\cdot||\hat{\mathbf{z}}^v_{i,k}||}\cdot\frac{1}{\tau})} \\ &-\ln \frac{\exp(\frac{(\hat{\mathbf{x}}^t_i)^\top \hat{\mathbf{z}}^t_i}{||\hat{\mathbf{x}}^t_i||\cdot||\hat{\mathbf{z}}^t_i||}\cdot\frac{1}{\tau})}{\exp(\frac{(\hat{\mathbf{x}}^t_i)^\top \hat{\mathbf{z}}^t_i}{||\hat{\mathbf{x}}^t_i||\cdot||\hat{\mathbf{z}}^t_i||}\cdot\frac{1}{\tau}) + \sum_{k=1}^{K}\exp(\frac{(\hat{\mathbf{x}}^t_i)^\top \hat{\mathbf{z}}^t_{i,k}}{||\hat{\mathbf{x}}^t_i||\cdot||\hat{\mathbf{z}}^t_{i,k}||}\cdot\frac{1}{\tau})}.\end{aligned} \tag{6}$$

3.2.2 ***Multi-modal Instance Representation Learning***. After obtaining the refined features of the micro-video and product, we aim to fuse the multi-modal feature to capture multi-modal instance representations. In this part, we elaborate on the fusion model and cross-instance contrastive loss to optimize it.

Multi-modal Fusion. We apply the context gating mechanism [30] to fuse the visual and textual features of instances, including micro-videos and products. It is defined as

$$\begin{cases}\boldsymbol{m}_i = (\boldsymbol{W}^m_2 \hat{\mathbf{x}}^v_i + \boldsymbol{W}^m_1 \hat{\mathbf{x}}^t_i) \circ \sigma(\boldsymbol{W}^m_3(\boldsymbol{W}^m_2 \hat{\mathbf{x}}^v_i + \boldsymbol{W}^m_1 \hat{\mathbf{x}}^t_i)), \\ \boldsymbol{p}_j = (\boldsymbol{W}^p_2 \hat{\mathbf{z}}^v_j + \boldsymbol{W}^p_1 \hat{\mathbf{z}}^t_j) \circ \sigma(\boldsymbol{W}^p_3(\boldsymbol{W}^p_2 \hat{\mathbf{z}}^v_j + \boldsymbol{W}^p_1 \hat{\mathbf{z}}^t_j)),\end{cases} \tag{7}$$

where $\boldsymbol{m}_i \in \mathbb{R}^{d'}$ and $\boldsymbol{p}_j \in \mathbb{R}^{d'}$ are the multi-model feature vectors of i-th micro-video and j-th product, respectively. d' is the dimension of the multi-model feature vector. Moreover, $\boldsymbol{W}^m_\cdot$ and $\boldsymbol{W}^p_\cdot$ are the trainable weight matrices in the micro-video and product fusion models. $\circ$ denotes element-wise multiplication and σ is an element-wise sigmoid activation.

Cross-instance Contrastive Loss. Facing a vast number of candidate micro-videos and products in practice, we resort to the momentum-based contrastive loss [14] to model the instance-level similarity. Typically, a queue is maintained to store the encoded representations of samples. With the momentum update mechanism for parameters and the memory-bank update mechanism of enqueuing the new mini-batch samples and dequeuing the oldest ones, we can obtain large and consistent instances to conduct contrastive learning stably and smoothly.

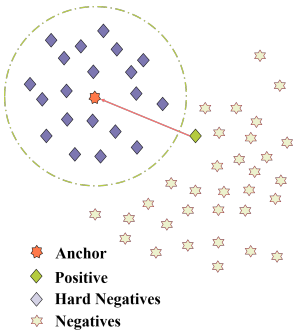


Figure 2: Importance of negatives in different categories.

Towards this end, we build the query and key encoders for the micro-video and product and feed the representations of instances into them, formally

$$\begin{cases}\hat{\boldsymbol{m}}^k_i = h(\boldsymbol{m}_i; \theta^k_m), \\ \hat{\boldsymbol{m}}^q_i = h(\boldsymbol{m}_i; \theta^q_m), \\ \hat{\boldsymbol{p}}^k_j = h(\boldsymbol{p}_j; \theta^k_p), \\ \hat{\boldsymbol{p}}^q_j = h(\boldsymbol{p}_j; \theta^q_p).\end{cases} \tag{8}$$

Wherein, k and q are used to indicate the key and query encoders, respectively. $\hat{\boldsymbol{m}}^{(\cdot)}_i \in \mathbb{R}^{d'}$ and $\hat{\boldsymbol{p}}^{(\cdot)}_j \in \mathbb{R}^{d'}$ separately denote the encoded vectors of micro-video and product. $h(\cdot)$ denotes the encoder, in which $\theta^{(\cdot)}_m$ and $\theta^{(\cdot)}_p$ are the parameters to be trained.

Based on the momentum contrastive learning mechanism, the parameters θ^q_m and θ^q_p are updated by back-propagation, while θ^k_m and θ^k_p are momentum updated as follows:

$$\begin{cases}\theta^k_m \leftarrow m\theta^k_m + (1-m)\theta^q_m, \\ \theta^k_p \leftarrow m\theta^k_p + (1-m)\theta^q_p,\end{cases} \tag{9}$$

where $m \in [0, 1)$ is a pre-defined momentum coefficient.

However, this method seldom considers the difference between the samples in the queue and treats them as the negative samples equally in the contrastive loss computation. In fact, different negative micro-videos (products) play various importance to the anchor products (micro-videos). To illustrate this, we take one micro-video and its corresponding product as an anchor and positive instances, and randomly sample multiple products as negative ones. We scatter these instances with t-SNE algorithm [42], as illustrated in Figure 2. Observing the distances between the anchor and other instances, we find that a portion of negative samples, namely the 'hard' negative sample, are closer to the anchor than the positive one. They make more contributions to optimize the representation learning of instances, whereas the other negative samples hardly help the optimization. As the prior study [35] mentioned, the effective negative in contrastive loss should satisfy the two principles:

- *Principle 1. The labels of true negatives should differ from that of the anchor x.*

- *Principle 2. The most useful negative samples are ones that the embedding currently believes to be similar to the anchor.*

Therefore, we design a multi-queue momentum contrast training strategy by which we organize the mini-batch instances according to their categories (*e.g.,* bag, toy, and fruit) and enqueue them into several category-aware queues. When the anchor and positive instances emerge in the training phase, we dequeue the oldest samples from the queues as the negative samples to perform the momentum contrastive learning. Formally, we build the multi-queue structure as,

$$\begin{cases} queue_1 = [q_1^1, q_2^1, \ldots, q_T^1], \\ queue_2 = [q_1^2, q_2^2, \ldots, q_T^2], \\ \cdots, \\ queue_C = [q_1^C, q_2^C, \ldots, q_T^C], \end{cases} \tag{10}$$

where $queue_t$ denotes t-th queue corresponding to t-th category. C and T are the numbers of queues and the length of each queue, respectively.

When a mini-batch of microvideo-product pairs is fed, we collect the negative samples from the multiple queues according to the categories of the positive instances. For instance, given four products that belong to 1-th, 1-th, 2-th, and 3-th categories, we dequeue two negative samples from $queue_1$, one sample from $queue_2$, and one sample from $queue_3$. Reorganizing these dequeued instances as a negative batch, we can guarantee the 'hard' negative sample of each anchor in the contrastive training phase.

Moreover, considering the various categories of instances in the negative batch, we score the importance of anchor and negative sample pairs to formulate a multi-queue contrastive loss. For this purpose, we perform a mean-pooling operation on all instances of the same category as the representation of the category. Then, we can score the importance of instance pairs by measuring the distance of representation of their categories, formally,

$$e_{i,j} = 1 - \sum_{l=1}^{L} \zeta \cdot \exp(norm(d(\boldsymbol{c}_i^1, \boldsymbol{c}_j^1))) - \zeta \cdot \exp(norm(d(\boldsymbol{c}_i^2, \boldsymbol{c}_j^2))), \tag{11}$$

where $e_{i,j}$ denotes the importance score for i-th and j-th instances. $norm(\cdot)$ and $d(\cdot)$ represent the normalization and Euclidean distance functions, respectively. In addition, ζ is the hyper-parameter controlling the degree of the score. It is worth noting that since we construct a hierarchical ontology of categories, each instance can be classified into two categories in the fine-grained and coarse-grained levels, *e.g.,* $\mathbf{c}_i^1, \mathbf{c}_j^1 \in \mathbb{R}^{d'}$ and $\mathbf{c}_i^2, \mathbf{c}_j^2 \in \mathbb{R}^{d'}$. Hence, we compute the similarities of category pairs at two levels.

$$\begin{aligned} \mathcal{L}_3 = &- \ln \frac{\exp(\frac{(\hat{\mathbf{m}}_i^k)^\top \hat{\mathbf{p}}_i^q}{||\hat{\mathbf{m}}_i^k|| \cdot ||\hat{\mathbf{p}}_i^q||} \cdot \frac{1}{\tau})}{\exp(\frac{(\hat{\mathbf{m}}_i^k)^\top \hat{\mathbf{p}}_i^q}{||\hat{\mathbf{m}}_i^k|| \cdot ||\hat{\mathbf{p}}_i^q||} \cdot \frac{1}{\tau}) + \sum_{k=1}^{K} e_{i,i_k} \exp(\frac{(\hat{\mathbf{m}}_i^k)^\top \hat{\mathbf{p}}_{i,k}^q}{||\hat{\mathbf{m}}_i^k|| \cdot ||\hat{\mathbf{p}}_{i,k}^q||} \cdot \frac{1}{\tau})} \\ &- \ln \frac{\exp(\frac{(\hat{\mathbf{p}}_i^k)^\top \hat{\mathbf{m}}_i^q}{||\hat{\mathbf{p}}_i^k|| \cdot ||\hat{\mathbf{m}}_i^q||} \cdot \frac{1}{\tau})}{\exp(\frac{(\hat{\mathbf{p}}_i^k)^\top \hat{\mathbf{m}}_i^q}{||\hat{\mathbf{p}}_i^k|| \cdot ||\hat{\mathbf{m}}_i^q||} \cdot \frac{1}{\tau}) + \sum_{k=1}^{K} e_{i,i_k} \exp(\frac{(\hat{\mathbf{p}}_i^k)^\top \hat{\mathbf{m}}_{i,k}^q}{||\hat{\mathbf{p}}_i^k|| \cdot ||\hat{\mathbf{m}}_{i,k}^q||} \cdot \frac{1}{\tau})}. \end{aligned} \tag{12}$$

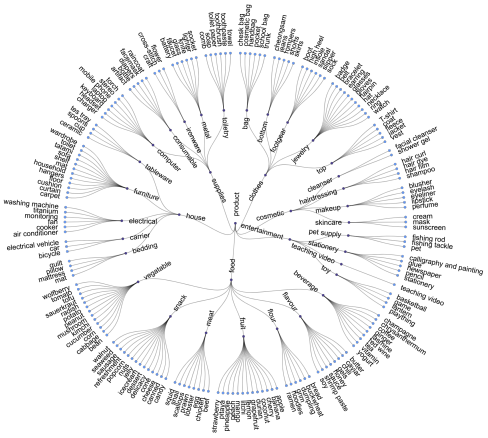


Figure 3: The category ontology.

Thus, the overall objective function is $\mathcal{L}$:

$$\mathcal{L} = \alpha \mathcal{L}_1 + \beta \mathcal{L}_2 + \delta \mathcal{L}_3, \tag{13}$$

where α, β, and δ are three hyper-parameters to balance the cross-modal, cross-instance, and cross-category contrasts, restricted in the range of zero to one. We test different compositions of α, β and δ in our experiments.

4 EXPERIMENTS

To evaluate our proposed model, we conduct extensive experiments on two constructed datasets. In this section, we elaborate on the datasets, metrics, and baselines for evaluation followed by the description of implementation details.

4.1 Datasets

To test our proposed model, we collect two large-scale datasets, MVS and MVS-large, from the micro-video sharing platform. They consist of some microvideo-product pairs, which are embedded by their publishers manually. In particular, we capture the micro-video content information, including the micro-videos and their captions. As for the products, we crawl both their pictures associated with textual descriptions. Beyond the content information, we manually construct a hierarchical category ontology according to the categories of the products.

Finally, we obtain the MVS dataset with 13,165 microvideo-product pairs and MVS-large with 126,206 pairs. Moreover, we achieve a three-layer category ontology, ranging from clothes to furniture, consisting of 6 coarse-grained categories, 30 middle-grained categories, and 316 fine-grained categories[3]. As shown in Figure 3, we present the statistics of partial data of the hierarchical category ontology.

4.2 Evaluation Metrics

For each dataset, we used the ratio 3:1:1 to randomly split the microvideo-product pairs to constitute the train set, validation set, and test set. The train set, the validation set, and the test set are used

[3]Due to insufficient instances of fine-grained categories, we conduct our experiments with other two-level categories.

Table 1: The experimental results of start-of-the-art on MVS and MVS-large.

Datasets	Methods	Microvideo-Product Retrieval					Product-Microvideo Retrieval				
		R@1	R@5	R@10	MedR	Rsum	R@1	R@5	R@10	MedR	Rsum
MVS	Base	0.0	0.2	0.4	1319	0.6	0.0	0.2	0.4	1314	0.6
MVS	HowTo100M[31]	24.9	30.1	32.5	147	87.5	21.8	25.7	29.5	128	77.0
MVS	AVLnet[36]	18.5	31.6	38.3	44	88.4	15.3	31.1	38.1	45	84.5
MVS	MoCo[14]	35.3	41.1	42.9	45	119.3	37.0	40.8	41.1	40	118.9
MVS	CLIP[33]	35.5	42.3	43.7	77	121.5	33.5	41.5	43.4	68	118.4
MVS	Hit[27]	40.0	42.8	44.7	35	127.5	42.7	43.9	44.9	32	131.5
MVS	MQMC	**44.7**	**48.7**	**50.2**	**10**	**143.6**	**44.9**	**48.7**	**50.5**	**9**	**144.1**
MVS	*Improv%*	11.75%	13.79%	12.30%	71.43%	12.63%	5.15%	10.93%	12.47%	71.88%	9.58%
MVS-large	Base	0.0	0.0	0.0	12643	0.0	0.0	0.0	0.0	12734	0.0
MVS-large	HowTo100M[31]	7.7	20.1	27.1	83	54.9	7.5	19.8	26.7	107	54.0
MVS-large	AVLnet[36]	16.8	38.0	46.6	14	101.4	17.0	37.6	46.2	14	100.8
MVS-large	CLIP[33]	20.8	43.6	51.0	10	115.4	18.4	38.9	47.0	14	104.3
MVS-large	MoCo[14]	21.2	44.0	51.3	9	116.5	19.1	39.5	47.7	13	106.3
MVS-large	Hit[27]	24.5	44.1	50.4	10	119.0	20.9	41.2	49.5	11	111.6
MVS-large	MQMC	**27.3**	**47.7**	**54.2**	**7**	**129.2**	**25.1**	**46.6**	**53.9**	**7**	**125.5**
MVS-large	*Improv%*	11.43%	8.16%	5.65%	22.22%	8.57%	20.10%	13.11%	8.89%	36.36%	12.46%

to optimize parameters, tune the hyper-parameters and evaluate the performance in the experiments, respectively.

We test the performance with several metrics widely used in information retrieval, including Recall at K (*i.e.,* R@K and K=1, 5, 10), Median Rank (MedR), and Rsum. Specifically, R@K is the percentage of test queries that the relevant item found among the top-K retrieved results. The MedR measures the median rank of correct items in the retrieved ranking list, where a lower score indicates a better model. We also take the sum of all R@K as Rsum to reflect the overall retrieval performance.

4.3 baselines

We compare our proposed model with state-of-the-art models, including

- **Base** The benchmark method concatenates visual and textual features of an instance to measure similarity.
- **HowTo100M**[31] This method learns a joint text-video embedding from the paired videos and captions and uses the max-margin ranking loss with a negative sampling strategy.
- **AVLnet**[36] This method introduces a self-supervised network that learns a shared audio-visual embedding space from raw audio, video, and text for audio-video retrieval.
- **MoCo**[14] This method introduces momentum contrast for unsupervised visual embedding. This method keeps the dictionary keys as consistent as possible despite its evolution under the hypothesis that good representation benefits from a large dictionary containing a rich set of negative samples.
- **CLIP**[33] This method learns a multi-modal embedding space of image and text by contrastive pre-training on large-scale datasets. This method maximizes the cosine similarity of the image and text embeddings of the N real pairs in the batch and minimizes the cosine similarity of the other $N^2 - N$ incorrect pairings.
- **Hit**[27] This method learns hierarchical embeddings with a hierarchical transformer for video-text retrieval, which performs hierarchical cross-modal contrastive matching at both feature-level and semantic-level.

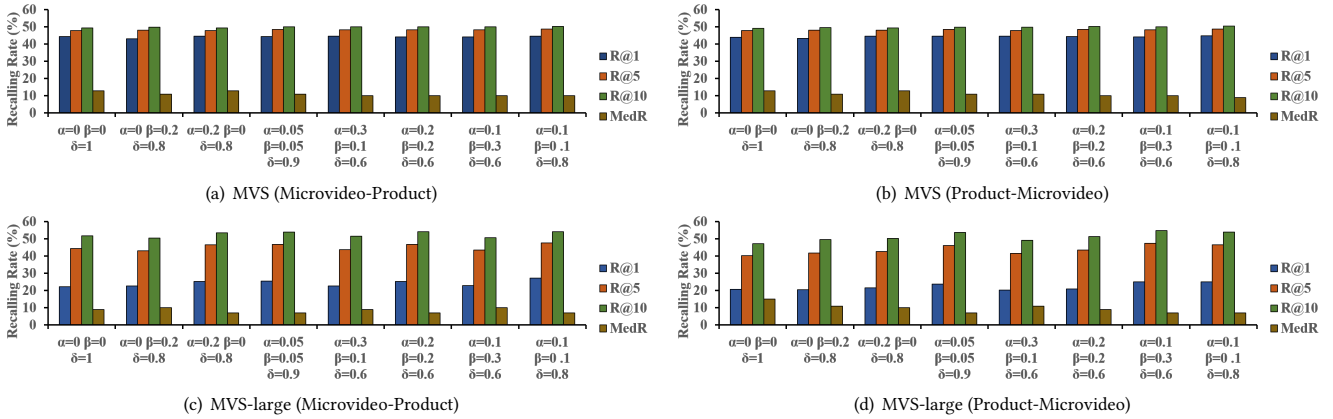
4.4 Implementation Details

We adopt pre-trained feature extractors in different modalities. In particular, we extract 2,048-dimensional visual features with Resnet152 model [15] pre-trained on ImageNet and 768-dimensional textual features with BERT-base-uncased model [8] pre-trained on the wiki. The visual features of multiple frames are pooled as the feature of one micro-video. For AVLnet [36], we also extracted the 128-dimensional audio features from VGGish model [37] pre-trained on YT8M [1].

We set the hidden size of visual and textual projectors to 1,024. The dimensions of micro-video and product encoding are both set to 512. The leaky ReLU[29] is used as the activation function and BatchNorm is appended to hidden layers. The initial learning rate is set to $1e-4$ and the network is optimized by Adam [28] optimizer. The weight decay is set to $1e-3$ and cosine decay is used for scheduling the learning rate. The momentum of updating the key encoder is set to 0.999 and τ is set to 0.07. The length of multi-queue, *i.e.,* T, should vary with the batch size. we set the batch size as 64 and 256 on MVS and MVS-large datasets, respectively. And, T of MVS and MVS-large are set as 192 and 2, 048, respectively. For the baselines, we do the same options and follow the designs in their articles to achieve the best performance.

4.5 Overall Performance Comparison

To demonstrate the effectiveness of our proposed model, we start by doing a comparison between our proposed model and the baselines on MSV and MSV-large datasets, respectively. Specifically, we list their results *w.r.t* recall and MedR in Table 1, where *Improv.%* represents the relative improvement of the best-performing method (bolded) over the strongest baselines (underlined).



(a) MVS (Microvideo-Product) (b) MVS (Product-Microvideo)

(c) MVS-large (Microvideo-Product) (d) MVS-large (Product-Microvideo)

Figure 4: Ablation study on different contrasts.

Table 2: Ablation study on different modalities.

Datasets	Modalities	Microvideo-Product Retrieval				
		R@1	R@5	R@10	MedR	Rsum
MVS	Visual only	36.9	43.5	44.3	20	124.7
MVS	Text only	26.1	37.0	41.8	31	104.9
MVS	**All**	**44.7**	**48.7**	**50.2**	**10**	**143.6**
MVS-large	Visual only	17.8	36.7	44.5	18	98.9
MVS-large	Text only	10.9	28.6	36.7	35	76.3
MVS-large	**All**	**27.3**	**47.7**	**54.2**	**7**	**129.2**

- Without any doubt, our proposed model outperforms all comparison methods by a clear margin. In particular, our proposed model improves over the strongest baselines *w.r.t.* R@0 by 3.79% and 6.80% in MVS and MVS-large datasets, respectively. It demonstrates the effectiveness of our proposed model.
- Comparing with the other contrastive-based model (*i.e.,* MoCo, Clip), we find that our proposed model achieves better performance. We attribute such an improvement to our multi-queue contrastive training strategy.
- Beyond the visual and textual information, MQMC considers the acoustic features in the microvideo-product retrieval yet unexpectedly performances are poor in most cases. The reason might be that the content of the micro-video contains some noise information that negatively affects the performance. It verifies our arguments and the reasonability of our proposed model again.

4.6 Ablation Study

To evaluate the designs in our proposed model, we conduct ablation studies on the two datasets. We test the effectiveness of uni-modal features and multi-modal instance representation learning. In addition, we evaluate the hyper-parameters α, β, and δ designed for the balance between the representation learning parts. Next, we dive into the multi-queue contrastive training strategy to further test its effectiveness and robustness.

4.6.1 ***Uni-modal and multi-modal Representation Learning.*** In order to test the uni-modal and multi-modal representation learning, we first explore the effects of different modalities and compare the results over the two datasets, as listed in Table 2. From the results, we observe that:

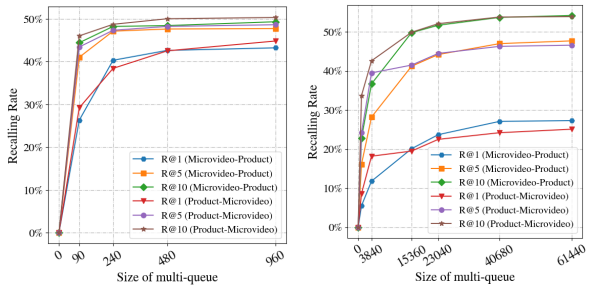
Table 3: The impacts of Multi-queue for retrieval performance. Sing. and Multi. denote the single and multiple queues, respectively. w/o S. represents the queues without important scores.

Datasets	Queue	Microvideo-Product Retrieval				
		R@1	R@5	R@10	MedR	Rsum
MVS	Sing. & w/o S.	42.5	47.1	48.3	17	137.9
MVS	Multi. & w/o S.	43.8	47.4	48.8	18	140.0
MVS	Multi.	**44.7**	**48.7**	**50.2**	**10**	**143.6**
MVS-large	Sing.& w/o S.	18.5	38.5	46.3	15	103.3
MVS-large	Multi.& w/o S.	21.4	41.7	49.0	12	112.1
MVS-large	Multi.	**27.3**	**47.7**	**54.2**	**7**	**129.2**

- As expected, the performance with multi-modal representation learning significantly outperforms that with uni-modal ones, including visual and textual modalities, on MVS and MVS-large datasets. It demonstrates the effectiveness of the multi-modal instance representation learning, including the multi-modal fusion method and cross-instance contrastive loss.
- Jointly analyzing the results of the baselines model shown in Table 1, we find that the performance with uni-modal representation learning is even better than that of the state-of-the-art baselines in some cases. We attribute this phenomenon to our designed uni-modal representation learning, which distills the informative signal from the noise information caused by the complex and chaotic background.

Next, we perform the experiments by varying the value of hyper-parameters α, β, and δ in the range of 0 to 1. Observing the results illustrated in Figure 4, we have the following findings the cross-instance contrast plays a vital role in the microvideo-product retrieval task. This might be that the supervision signal from the cross-instance similarity makes more contribution to optimize the uni- and multi-modal representation learning. Nevertheless, the cross- and intra-modal contrastive losses also cannot be ignored, which verifies our uni-modal feature representation learning again. Overall, from the results, we find that our proposed model gains the best performance when we set $\alpha = 0.1$, $\beta = 0.1$, and $\delta = 0.8$, respectively.

4.6.2 ***Multi-queue Contrastive Training Strategy.*** To evaluate the effectiveness of the multi-queue, we conduct the experiments



(a) MVS (b) MVS-large

Figure 5: Ablation study on two datasets to investigate the contributions of different sizes of Multi-queue.

of our proposed model without the multi-queue architecture and compare its performance with that of our proposed model equipped with the multi-queue. Moreover, we discard the difference of multiple queues by setting the weight (*i.e.,* $e_{i,j}$) of each anchor-negative pair to 1. We list the results in Table 3 and find that: the results *w.r.t.* R@1, R@5, R@10, and MedR are increased when we utilize the multi-queue architecture. It justifies our designed multi-queue contrastive loss. Furthermore, we compare the performance with the importance scores and without the scores. The performance *w.r.t.* R@1 is improved from 43.8% to 44.7% on MVS and from 21.4 % to 27.3% on MVS-large. We suggest that the proposed model is benefited from the utilization of the scores to distinguish the importance of different negatives.

Further, we explore the influences of the multi-queue size in microvideo-product retrieval. For this goal, we conduct the experiments on the two datasets by varying the sizes from 90 to 960 and from 960 to 61440, respectively. Observing the experimental results shown in Figure 5, we find that the introduction of large-scale negatives for similarity learning indeed achieves considerable performance improvements. We attribute it to broader negative interactions for obtaining more precise and discriminative representations. In fact, due to the unbalanced distribution of categories and the existence of the long-tail problem, the size of the multi-queue is limited by a few categories with a small amount of data. When the actual training samples cannot fill the queue, the negatives in the queue are not single and independent, so the repeated positive cases in the queue are mistakenly divided into negative cases. Moreover, by reason of the category sensitivity of multi-queue, the above errors will be magnified, affecting the learning effect of the retrieval modal.

4.7 Case Study

To visualize our proposed model, we randomly sample four micro-videos from two datasets and conduct the strongest baseline (Hit) and our proposed models on them. In particular, we collect the Top-3 results of two models 6 and separately mark the correct predictions with green circles and others with yellow circles at the top left corners, as shown in Figure 6.

According to the results, in the first two cases, our proposed model not only matches the visually similar items in micro-videos and pictures of products but also explores the relationships between textual and visual modalities sufficiently to obtain identical ones more accurately. In the upper left case, with the assistance of textual

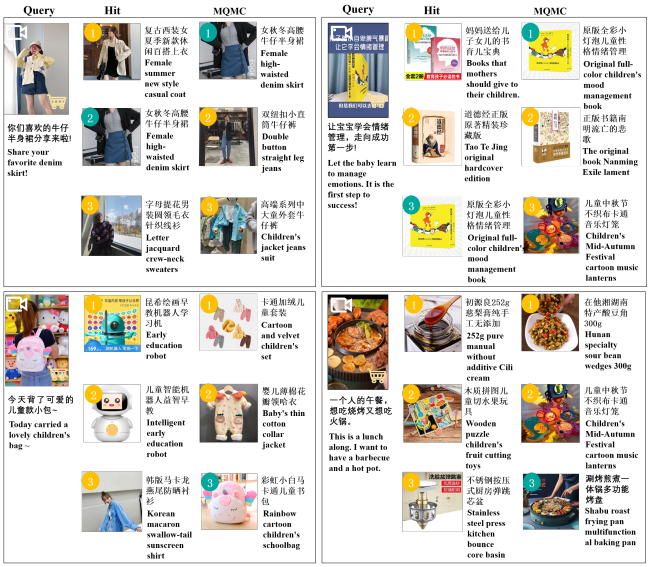


Figure 6: Four examples with top-3 microvideo-product retrieval results of MQMC and Hit.

modality, confusing visual modality can focus more attention on the skirt that the micro-video wants to display, rather than the coat that occupies more space in the frame. In the upper right case, our proposed model can effectively remove the interference of large areas of the blue background and white characters in videos, and perform accurate retrieval for book categories with fewer data. However, in the last two cases, due to the obvious visual differences between some micro-videos and products, our proposed model can not achieve the best match. In the lower-left case, the video background is very cluttered, with multiple indistinguishable objects that are very similar to each other. In the lower right case, due to a large amount of variant visual information between the frame and picture, as well as the incomplete outline of the item limited by the shooting angle of the micro-video, the retrieval faces great difficulty.

5 CONCLUSION

This paper is the first to formulate the microvideo-product retrieval task between multi-modal and multi-modal instances. We propose the Multi-Queue Momentum Contrastive learning network (MQMC) for bidirectional microvideo-product multimodal-to-multimodal retrieval, which consists of the uni-model feature representation learning and multi-modal instance representation learning, integrating cross-modal contrast, intra-modal contrast and cross-instance contrast organically. In addition, we present a discriminative negative selection strategy with a multi-queue to distinguish the importance of different negatives with their categories. Two large-scale multi-modal datasets are built for microvideo-product retrieval, and we construct a category ontology and manually annotate the multi-layer ontologies of all products of the datasets. Extensive experiments prove the validity of the proposed model. In the future, MQMC can be extended as a general method for multimodal-to-multimodal retrieval. And more technological innovations can be researched such as multi-modal integrated approaches, hard negative selection strategy, etc.

REFERENCES

[1] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. 2016. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675* (2016).

[2] Jean-Baptiste Alayrac, Piotr Bojanowski, Nishant Agrawal, Josef Sivic, Ivan Laptev, and Simon Lacoste-Julien. 2016. Unsupervised learning from narrated instruction videos. In *CVPR*. 4575–4583.

[3] Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. 2021. Conditioned Image Retrieval for Fashion using Contrastive Learning and CLIP-based Features. In *ACM MM Asia*. 1–5.

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *ICML*. PMLR, 1597–1607.

[5] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. 2020. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297* (2020).

[6] Xinlei Chen and Kaiming He. 2021. Exploring simple siamese representation learning. In *CVPR*. 15750–15758.

[7] Zhi-Qi Cheng, Xiao Wu, Yang Liu, and Xian-Sheng Hua. 2017. Video2shop: Exact matching clothes in videos to online shopping images. In *CVPR*. 4048–4056.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[9] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. 2020. Multi-modal transformer for video retrieval. In *ECCV*. Springer, 214–229.

[10] Marco Godi, Christian Joppi, Geri Skenderi, and Marco Cristani. 2022. MovingFashion: a Benchmark for the Video-to-Shop Challenge. In *WACV*. 1678–1686.

[11] Marco Godi, Christian Joppi, Geri Skenderi, and Marco Cristani. 2022. MovingFashion: a Benchmark for the Video-to-Shop Challenge. In *WACV*. 1678–1686.

[12] Yongshun Gong, Jinfeng Yi, Dong-Dong Chen, Jian Zhang, Jiayu Zhou, and Zhihua Zhou. 2021. Inferring the Importance of Product Appearance with Semi-supervised Multi-modal Enhancement: A Step Towards the Screenless Retailing. In *ACM MM*. 1120–1128.

[13] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. 2020. Bootstrap your own latent-a new approach to self-supervised learning. *NIPS* 33 (2020), 21271–21284.

[14] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *CVPR*. 9729–9738.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *CVPR*. 770–778.

[16] Yupeng Hu, Liqiang Nie, Meng Liu, Kun Wang, Yinglong Wang, and Xian-Sheng Hua. 2021. Coarse-to-fine semantic alignment for cross-modal moment localization. *TIP* 30 (2021), 5933–5943.

[17] Hao Jiang, Wenjie Wang, Yinwei Wei, Zan Gao, Yinglong Wang, and Liqiang Nie. 2020. What aspect do you like: Multi-scale time-aware user interest modeling for micro-video recommendation. In *ACM MM*. 3487–3495.

[18] Yannis Kalantidis, Mert Bulent Sariyildiz, Noe Pion, Philippe Weinzaepfel, and Diane Larlus. 2020. Hard negative mixing for contrastive learning. *NIPS* 33 (2020), 21798–21809.

[19] Chenyi Lei, Yong Liu, Lingzi Zhang, Guoxin Wang, Haihong Tang, Houqiang Li, and Chunyan Miao. 2021. SEMI: A Sequential Multi-Modal Information Transfer Network for E-Commerce Micro-Video Recommendations. In *ACM SIGKDD*.

[20] 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. *NIPS* 34 (2021).

[21] Mengmeng Li, Tian Gan, Meng Liu, Zhiyong Cheng, Jianhua Yin, and Liqiang Nie. 2019. Long-tail hashtag recommendation for micro-videos with graph convolutional network. In *CIKM*. 509–518.

[22] Ming Li, Xiao-Bing Xue, and Zhi-Hua Zhou. 2009. Exploiting multi-modal interactions: A unified framework. In *IJCAI*. Citeseer.

[23] Yongqi Li, Meng Liu, Jianhua Yin, Chaoran Cui, Xin-Shun Xu, and Liqiang Nie. 2019. Routing micro-videos via a temporal graph-guided recommendation system. In *ACM MM*. 1464–1472.

[24] Fan Liu, Zhiyong Cheng, Huilin Chen, Yinwei Wei, Liqiang Nie, and Mohan Kankanhalli. 2022. Privacy-Preserving Synthetic Data Generation for Recommendation Systems. In *SIGIR*. 1379–1389.

[25] Meng Liu, Liqiang Nie, Meng Wang, and Baoquan Chen. 2017. Towards micro-video understanding by joint sequential-sparse modeling. In *ACM MM*. 970–978.

[26] Meng Liu, Liqiang Nie, Xiang Wang, Qi Tian, and Baoquan Chen. 2018. Online data organizer: micro-video categorization by structure-guided multimodal dictionary learning. *TIP* 28, 3 (2018), 1235–1247.

[27] Song Liu, Haoqi Fan, Shengsheng Qian, Yiru Chen, Wenkui Ding, and Zhongyuan Wang. 2021. Hit: Hierarchical transformer with momentum contrast for video-text retrieval. In *ICCV*. 11915–11925.

[28] Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* (2017).

[29] Andrew L Maas, Awni Y Hannun, Andrew Y Ng, et al. 2013. Rectifier nonlinearities improve neural network acoustic models. In *ICML*. Citeseer, 3.

[30] Antoine Miech, Ivan Laptev, and Josef Sivic. 2017. Learnable pooling with context gating for video classification. *arXiv preprint arXiv:1706.06905* (2017).

[31] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *ICCV*. 2630–2640.

[32] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* (2018).

[33] 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 *ICML*. PMLR, 8748–8763.

[34] Jun Rao, Yue Cao, Shuhan Qi, Zeyu Dong, Tao Qian, and Xuan Wang. 2021. Watch and Buy: A Practical Solution for Real-time Fashion Product Identification in Live Stream. In *WAB*. 23–31.

[35] Joshua David Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. 2021. Contrastive Learning with Hard Negative Samples. In *ICLR*. https://openreview.net/forum?id=CR1XOQ0UTh-

[36] Andrew Rouditchenko, Angie Boggust, David Harwath, Brian Chen, Dhiraj Joshi, Samuel Thomas, Kartik Audhkhasi, Hilde Kuehne, Rameswar Panda, Rogerio Feris, Brian Kingsbury, Michael Picheny, Antonio Torralba, and James Glass. 2021. AVLnet: Learning Audio-Visual Language Representations from Instructional Videos. In *Proc. Interspeech 2021*. 1584–1588.

[37] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[38] Teng Sun, Chun Wang, Xuemeng Song, Fuli Feng, and Liqiang Nie. 2022. Response Generation by Jointly Modeling Personalized Linguistic Styles and Emotions. *TOMM* 18, 2 (2022), 52:1–52:20.

[39] Teng Sun, Wenjie Wang, Liqiang Jing, Yiran Cui, Xuemeng Song, and Liqiang Nie. 2022. Counterfactual Reasoning for Out-of-distribution Multimodal Sentiment Analysis. In *ACM MM*. ACM, 15–23.

[40] Yonglong Tian, Dilip Krishnan, and Phillip Isola. 2020. Contrastive multiview coding. In *ECCV*. Springer, 776–794.

[41] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. 2020. What Makes for Good Views for Contrastive Learning?. In *NIPS*. Curran Associates, Inc., 6827–6839.

[42] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *JMLR* 9, 11 (2008).

[43] Xiao Wang, Tian Gan, Yinwei Wei, Jianlong Wu, Dai Meng, and Liqiang Nie. 2022. Micro-video Tagging via Jointly Modeling Social Influence and Tag Relation. In *ACM MM*. 4478–4486.

[44] Wei Wei, Chao Huang, Lianghao Xia, Yong Xu, Jiashu Zhao, and Dawei Yin. 2022. Contrastive meta learning with behavior multiplicity for recommendation. In *WSDM*. 1120–1128.

[45] Yinwei Wei, Zhiyong Cheng, Xuzheng Yu, Zhou Zhao, Lei Zhu, and Liqiang Nie. 2019. Personalized hashtag recommendation for micro-videos. In *ACM MM*. 1446–1454.

[46] Yinwei Wei, Xiang Wang, Weili Guan, Liqiang Nie, Zhouchen Lin, and Baoquan Chen. 2019. Neural multimodal cooperative learning toward micro-video understanding. *TIP* 29 (2019), 1–14.

[47] Yinwei Wei, Xiang Wang, Liqiang Nie, Xiangnan He, Richang Hong, and Tat-Seng Chua. 2019. MMGCN: Multi-modal graph convolution network for personalized recommendation of micro-video. In *ACM MM*. 1437–1445.

[48] Xin Xin, Jiyuan Yang, Hanbing Wang, Jun Ma, Pengjie Ren, Hengliang Luo, Xinlei Shi, Zhumin Chen, and Zhaochun Ren. 2022. On the User Behavior Leakage from Recommender System Exposure. *TOIS* (2022).

[49] Yang Yang, Jia-Qi Yang, Ran Bao, De-Chuan Zhan, Hengshu Zhu, Xiao-Ru Gao, Hui Xiong, and Jian Yang. 2021. Corporate Relative Valuation using Heterogeneous Multi-Modal Graph Neural Network. *IEEE Trans Knowl Data Eng* (2021).

[50] Han-Jia Ye, De-Chuan Zhan, Xiaolin Li, Zhen-Chuan Huang, and Yuan Jiang. 2016. College student scholarships and subsidies granting: A multi-modal multi-label approach. In *ICDM*. IEEE, 559–568.

[51] Yanhao Zhang, Qiang Wang, Pan Pan, Yun Zheng, Cheng Da, Siyang Sun, and Yinghui Xu. 2021. Fashion Focus: Multi-modal Retrieval System for Video Commodity Localization in E-commerce. *arXiv preprint arXiv:2102.04727* (2021).

[52] Hongrui Zhao, Jin Yu, Yanan Li, Donghui Wang, Jie Liu, Hongxia Yang, and Fei Wu. 2021. Dress like an internet celebrity: Fashion retrieval in videos. In *IJCAI*. 1054–1060.

[53] Luowei Zhou, Chenliang Xu, and Jason J Corso. 2018. Towards automatic learning of procedures from web instructional videos. In *AAAI*.

[54] Dimitri Zhukov, Jean-Baptiste Alayrac, Ramazan Gokberk Cinbis, David Fouhey, Ivan Laptev, and Josef Sivic. 2019. Cross-task weakly supervised learning from instructional videos. In *CVPR*. 3537–3545.