Abstract—Our goal in this research is to study a more realistic environment in which we can conduct weakly-supervised multi-modal instance-level product retrieval for fine-grained product categories. We first contribute the Product1M datasets and define two real practical instance-level retrieval tasks that enable evaluations on price comparison and personalized recommendations. For both instance-level tasks, accurately identifying the intended product target mentioned in visual-linguistic data and mitigating the impact of irrelevant content are quite challenging. To address this, we devise a more effective cross-modal pretraining model capable of adaptively incorporating key concept information from multi-modal data. This is accomplished by utilizing an entity graph, where nodes represent entities and edges denote the similarity relations between them. Specifically, a novel Entity-Graph Enhanced Cross-Modal Pretraining (EGE-CMP) model is proposed for instance-level commodity retrieval, which explicitly injects entity knowledge in both node-based and subgraph-based ways into the multi-modal networks via a self-supervised hybrid-stream transformer. This could reduce the confusion between different object contents, thereby effectively guiding the network to focus on entities with real semantics. Experimental results sufficiently verify the efficacy and generalizability of our EGE-CMP, outperforming several SOTA cross-modal baselines like CLIP Radford et al. 2021, UNITER Chen et al. 2020 and CAPTURE Zhan et al. 2021.

Index Terms—Multi-modal pre-training, instance-level retrieval, knowledge injection, product dataset.

I. INTRODUCTION

E-COMMERCE applications, as one of the largest multi-modal downstream scenarios, have facilitated person’s lives. These are various product-related tasks within this application, including similar item retrieval [4], online commodity recommendation [5], and identical product match for cross-price comparison [6]. With the wide range of downstream E-commerce applications in the real world, our focus in this work is on the pretraining of multi-modality data of products. In general scenarios, multi-modal vision-language pre-training model [1], [2], [3], [7], [8], [9], [10], [11], [12] such as CLIP [1] and ViLBert [7] have shown superior performances across diverse downstream tasks, such as zero-shot classification [13], cross-modal retrieval [14] and open-world detection [15]. These models learn visual and textual embedding features from a large number of image-text pairs obtained from various sources, exhibiting strong generalization capability and robustness. The fundamental methodology behind these models either involves contrastive semantic alignment of images and texts or specific network architectures designed to facilitate better common space learning. However, the aforementioned models primarily focus on the attention regions they adaptively learn and are easily influenced by the contents that are unrelated to the visual objects of each modality, resulting in ignoring the importance of concept knowledge, i.e., words or phrases with a certain meaning. As an illustration in Fig. 1, we expect the trained multi-modal network to pay more attention to four typical objects (Fair Water, Facial Cream, Cleaner and Clear Lotion) and to ignore extra content (Four piece Set) in a more fine-grained application setting. If the multi-modal network is trained directly by visual-language contrastive learning, the learned attention region may be SK-II, Cream Cleaner, Lotion Four-piece, which is unsuitable and incorrect to achieve the expected objects of attention. In the past few years, there has been increasing interest in designing the knowledge injection for natural language processing (NLP) tasks to enhance the effectiveness of attention in NLP models [16], [17], [18], [19]. However, these models often rely on a well fully annotated knowledge graph, which is time-consuming and labor-intensive to construct for a new task. In this paper, we balance the trade-off between knowledge base construction and knowledge extraction, and propose a simple but effective entity-graph enhanced visual-language pretraining model, which explicitly inject the concept knowledge generated from the caption data into the multi-modal model via a special knowledge graph (entity-graph).
Instance-level product retrieval refers to the retrieval of all single products in a product portfolio image.  

In the paper, we propose a large-scale Entity-Graph Enhanced Cross-Modal Pretraining framework (EGE-CMP) to address the challenges of cross-modal semantic alignment. We are inspired by the effectiveness of knowledge injection [16], [17], [32], [33], [34] into large-scale pretrained models for single modality, and thus we model the cross-modal semantic alignment under the supervision of an entity-graph instead of directly using image and text interaction, which makes the trained model more focused on meaningful entity objects and thus reduce the influence of irrelevant content. Specifically, we extract the noun entity directly from the caption data via named entity recognition (NER) tools and obtain the image proposals from the image data via the bottom-up-attention feature representation model [35]. EGE-CMP is presented to capture the potential synergy of images, entities and title descriptions via entity-graph learning and some task-agnostic masking training. We showcase that some widely used cross-modal pretraining methods [2], [7], [9], [10] might be ineffective under the multi-instance setting without external knowledge guidance. In contrast, EGE-CMP utilizes a hybrid-stream architecture that encodes data of different modalities separately and fuses them in a unified manner, which has been shown to be advantageous for our suggested purposes. Additionally, we introduce the cross-modal contrastive loss to compel EGE-CMP to align images, texts, and entities, so as to avoid the mismatch problem introduced by the unsuitable pretext task. Furthermore, our effective entity-graph enhanced module is adopted to enhance the learning capacity for the whole transformer model by enforcing the model to pay more attention to some entities with real semantic information. In our entity graph, the node is each entity data and the edge is denoted by the semantic relationship between each entity. To encourage the text-encoder to be aware of actual key entities and improve the understanding ability of the transformer model, node-based and subgraph-based ranking losses are also introduced for the model training, requiring it to accurately distinguish the actual positive samples of entities according to the global entity ranking list. As a result, our EGE-CMP model can focus more on the phrases or words with real semantic meaning, and the text-image correlation can be aligned better. Compared with existing knowledge-enhanced pretraining methods based on the knowledge graph, our EGE-CMP model directly utilizes the knowledge directly extracted from the original caption data without any time-consuming and labor-intensive annotated data.

Furthermore, to bridge this gap and facilitate relevant research, we construct a large-scale dataset from the popular shopping website, named Product1M, proposed for the performance verification of multi-modal instance-level retrieval in real-world scenarios. Product1M has about one million image-caption pairings and is divided into two sorts of samples, i.e., single- and multi-product samples. The dataset is one of the biggest multi-modal product image datasets.

1Instance-level product retrieval refers to the retrieval of all single products that existed in a product portfolio image.  

2Image-level product retrieval refers to recognizing a specific product instance in a single-product image.
modal datasets available and is the first to be created exclusively for multimodal instance-level retrieval in real-world settings. Two real-world tasks, multi-product retrieval and identical-product retrieval, are defined in Section III-A and used for the model comparisons. These tasks are more challenging compared with traditional multi-modal retrieval tasks as they focus on more fine-grained feature representations. Multi-product retrieval is designed to match the corresponding single product rather than whole objects in the images, while identical product retrieval is intended to query the best-matched product with the same details of the product’s appearance, shape, brand, color, function, and so on. Hence, both two tasks are more challenging compared with the traditional multi-modal retrieval tasks. In summary, as a large benchmark dataset, our Product1M with ground-truth annotations generated by several human workers. This enables accurate analysis of the performance of the evaluated multi-product and identical-product retrieval algorithms.

The results of experiments on both multi-product retrieval and identical-product retrieval tasks show the superiority of our EGE-CMP over the SOTA cross-modal baselines, such as ViLBERT [7], CLIP [1], UNITER [2], CAPTURE [3] and so on, across all major criteria by a large margin. Moreover, extensive ablation experiments are conducted to demonstrate the generalizability of EGE-CMP and examine various essential factors of our proposed task. Finally, the quantitative and qualitative results provide intuitive evidence that the entity-graph constructed by the caption itself can enhance the representation capability of transformer models.

The main contributions of our paper are threefold.

- We conduct Product1M, one of the largest multi-modal product datasets for the real-world instance-level retrieval task, and contribute a subset extracted from Product1M for identical product retrieval for price comparison. Our Product1M dataset is different from existing e-commerce datasets in that it has real-world testing circumstances and abundant categories, which can help advance future research in high-level difficulties.

- We devise a novel Entity-Graph Enhanced Cross-Modal Pretraining (EGE-CMP) framework to learn instance-level feature representations by injecting entity knowledge with real semantic information into visual-text alignment. We find our entity-graph significantly benefits the text-image cross-modal alignment.

- Experimental results on both our proposed instance-level retrieval tasks, namely multi-product retrieval and same product search demonstrate that the model representation can be effectively augmented by the proposed entity graph. Furthermore, both the quantitative and visualized research results intuitively prove the efficacy of our EGE-CMP framework.

The study builds upon our prior conference presentation, CAPTURE [3]. In the version, we add the following features to the conference version: (1) To be suggested, a new goal, identical-product retrieval, must be adequately specified. (2) We present a novel Cross-Modal Pretraining framework called EGE-CMP to integrate entity information into visual-linguistic alignment. We demonstrate the superiority of our suggested model, EGE-CMP, over many state-of-the-art approaches. Moreover, we show feature embeddings, conducted ablation investigations, and performed generalization analysis.

The rest of the paper is divided into several sections: Section II discusses related works, while Section III describes the definition of multi-product and identical product retrieval tasks and introduces our model, including its network architectures and formulation. Section IV exhibits and explains the experimental findings. Finally, we conclude in Section V.

II. RELATED WORK

A. Visual-Language Datasets

According to the collection source, visual-language pre-training datasets can be categorized into at least two main groups: 1) General Sources. There are several visual-language datasets [36], [37], [38], [39], [40], [41] collected from the social media platforms (e.g. Twitter and Facebook). Twitter100k [36] is presented for cross-media retrieval with a low degree of supervision. It is a collection of 100k image-text pairings crawled from the Twitter website. There are no limitations on the categories of images, and users supply the textual description in a colloquial language. Similarly, this dataset emphasizes text-image retrieval and employs single-modal input as the query, which is incompatible with instance-level multi-modal retrieval. INRIA-Websearch [41] is a cross-modal retrieval dataset that contains 71,478 image-text pairs connected with 353 unique search queries, such as actors, films, and so on. The images are gathered via internet searches, and the textual description is derived from the text that surrounds the images on websites. In this dataset, cross-modal retrieval methods are used to tackle the text-to-image search issue. 2) Product Sources. For the E-commerce datasets, RPC checkout [42] and Dress Retrieval [43] are crawled from the shopping website as two widely used typical datasets. RPC [42] comprises 53,739 photos of individual products and 30,000 photographs of checkout. Three independent levels of annotation are available: category information, point-level annotations, and bounding boxes. There is, however, a gap between RPC and real-world settings, since RPC was gathered in a controlled environment and this dataset is devoid of textual information. Dress Retrieval proposed in [43] collects noisy labeled data scraped from catalogs of E-commerce websites.

As indicated before, our Product1M dataset is an instance-level dataset. Unlike the previous datasets, each picture in the Product1M dataset is connected with a subset of relevant attributes that have been filtered using specified terminology. Typically, each picture comprises a single instance, such as a model wearing a garment, with a relatively uncluttered backdrop. In summary, our Product1M is significantly different from the aforementioned datasets in three ways: (1) it is the first dataset designed specifically for multi-product and identical product instance-level retrieval tasks, both of which have significant promise in the e-commerce business; (2) it enables multi-modal
retrieval in addition to conventional intra- and cross-modal retrieval when both the query and the target include pictures and text; This environment is more practical, as multimodal information is pervasive in real-world scenarios. (3) its weakly annotated examples compel the model to uncover relevant features in the absence of clear labels, endowing the model with a high capacity for generalization to large-scale and noisy data.

B. Intra- and Cross-Modal Retrieval

Intra-modal retrieval [20], [21] has been intensively investigated in the contexts of keyword-based web document retrieval [44] and content-based image retrieval [45]. On the other hand, cross-modal retrieval [24], [25], [26], [27], [28], [29] emerges as a prospective route for efficiently indexing and searching vast amounts of data with multiple modalities, and is extensively utilized in search engines [46], [47] and E-commerce [43], [48], to mention a few. However, these techniques [23], [42], [43], [49], [50] are often limited to unimodal inputs, making them inapplicable to a large number of real-world circumstances in which both the queries and targets include multi-modal information.

C. WSOD: Weakly Supervised Object Detection

By learning from cheaper or publicly accessible data, WSOD [51], [52], [53] decreases its over-reliance on fine-grained labeling. PCL [51] builds proposal clusters in an iterative manner to facilitate the construction of instance classifiers. Moreover, pseudo labels elicited from picture labels [52] and unstructured textual descriptions such as captions [53] are also useful for improving WSOD performance. WSOD, on the other hand, often relies on a fixed-size collection of predefined classes and is thus unsuitable for our proposed task, since class labels are not available and categories may change dynamically.

D. Visual-Linguistic Pretraining

The development of the vision-language pretraining model [54], [55], [56], [57], [58], [59], [60], [61], [62], [63], [64], [65], [66], [67] has advanced tremendously, with joint representations often constructed using a multimodal fusion transformer network topology. Existing pre-trained models for vision-language often learn image-text semantic alignment using a multi-layer Transformer architecture, such as Bert [68], on multi-modal input in a shared cross-modal space. They may be loosely divided into two categories based on their network structural differences: 1) Single-stream models [2], [8], [9], [54], [55], [56], [60] in a unified architecture encode the integrated multi-modal characteristics. 2) Two-stream models [7], [10], [57], [58], [59], [61], [65], [66], [67], on the other hand, use distinct encoders for inputs with varying modalities. However, these approaches are not optimized for instance-level retrieval and may be defective due to network architecture design flaws and incorrect pretext tasks. Furthermore, real-world semantic information cannot be effectively learned during network training, thus limiting performance increases.

E. Knowledge-Enhanced Modeling

In language models, the incorporation of knowledge, concepts, and relations into pretrained models [16], [68] has proven beneficial to language understanding [17]. The existing methods can be coarsely broadly classified into two types: implicit knowledge modeling and explicit knowledge injection. The former attempts to implicitly model knowledge usually consist of entity-level masked modeling [18], [19], [32], entity-based replacement prediction [69], knowledge embedding loss as regularization [34] and universal knowledge-text prediction [33], [70]. In contrast to implicit knowledge modeling, the latter separately maintains a group of parameters for representing structural knowledge. Such methods [18], [19], [32], [33], [69], [70] usually require a heterogeneous information fusion component to fuse multiple types of features obtained from the text and knowledge graphs. In our setting, it is very hard to build a knowledge graph with good annotations from scratch to fit the multi-modal data. In the paper, with the purpose of knowledge modeling and visual-language pretraining, knowledge is represented as an entity parsed from the original caption data and the relationship between different pieces of knowledge is learned via node-based and subgraph-based ranking losses.

Fig. 2 shows five typical vision-and-language models. According to whether external knowledge is utilized, these models can be generally fall into the following categories: 1) models without knowledge supervision that fall under Fig. 2(a) and (b), 2) knowledge-based models that fall under Fig. 2(c), (d) and (e). Our proposed model EGE-CMP belongs to the second type. Unlike the existing models as seen in Fig. 2(c) and (d), our model employs the entity knowledge extracted from the caption, instead of human labeling, (e.g. property and knowledge graph) which extremely save the labor and calculation cost. Our proposed EGE-CMP is the first model of type Fig. 2(e) where the entity knowledge drawn from the caption data is explicitly injected into the model via graph learning.

III. INSTANCE-LEVEL RETRIEVAL ON PRODUCT1M

In this section, we begin by describing the two instance-level tasks, followed by our proposed EGE-CMP model. The definition of the tasks is provided in Section III-A, while the dataset analysis is presented in Section III-B. Our proposed Entity-Graph Enhanced Cross-Modal Pretraining (EGE-CMP) is detailed in Section III-C, covering the model architecture, masked relation prediction, cross-modal contrastive learning, and model inference.

A. Task Definition

1) Multi-Product Retrieval: A product sample \( (i, c) \) is an image-text pair where \( i \) is the product image and \( c \) is the caption. Given a gallery set of single-product samples \( S = \{ S_i | S_i = (I_i, C_i) \} \) and the set of multi-product samples \( P = \{ P_t | P_t = (I_p, C_p) \} \), the task is to retrieve and rank the single-products that appear in the query sample \( P_t \), i.e., to predict a list \( RET^T = [d_{i_1}, d_{i_2}, \ldots, d_{i_{N_t}}] \) \( \forall P_t \in P \), where \( id_{i_k} \) corresponds to a specific single-product sample in \( S \).
Fig. 2. Five categories of visual-language pretraining models. The two models on the left represent typical network structures, while the three models on the right are the knowledge injecting versions. Comparing our structure (e) with (c) and (d), we directly build the entity graph from the caption without the need for additional annotation information. In the figure, CTL, LPM, MRP, ENL, ESL denote the Contrastive Learning, Link Prediction Modeling, Masked Relation Prediction, Entity Node Loss, and Entity Subgraph Loss, respectively.

### TABLE I

| Dataset | |samples| |categories| |instances| |#obj/im| |weak supervision| |multi-modal| |instance-level retrieval| |multi-task|
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Twitter10K [36] | 100,000 | | | | | | | | | | | | |
| Visual Genome [37] | 108,000 | | | | | | | | | | | | |
| Flickr30K [38] | 31,000 | | | | | | | | | | | | |
| SBU [39] | 589,000 | | | | | | | | | | | | |
| NUS-WIDE [40] | 269,648 | 81 | | | | | | | | | | | |
| INRIA-Webscrape [41] | 71,478 | 353 | | | | | | | | | | | |
| RPC checkout [42] | 30,000 | 200 | 367,935 | 12.26 | | | | | | | | | |
| Dress Retrieval [43] | 20,200 | | | | | | | | | | | | |
| Product1M (Ours) | 1,182,083 | 458 | 92,200 | 2.85 | | | | | | | | | |

*Indicates inapplicable. The #instances and #obj/im of product 1M are in italics since there are no instance labels for the train set and we only count the instances in the val and test set. Product1M is one of the largest multi-modal datasets and the first to be specifically designed for instance-level retrieval in real-world applications.

#### 2) Identical-Product Retrieval:

In the task, we do not differentiate between single-products and multi-products. Similar to the multi-product retrieval task, given the **gallery** set of product samples \( S = \{ S_i \} | S_i = (I_i, C_i) \} \) and a set of product samples \( P = \{ P_i \} | P_i = (I_i, C_i) \} \), the task is to retrieve and rank the products that appear in the query sample \( P_i \), i.e., to predict a list \( RETR_i = [id_1, id_2, ..., id_k, ..., id_N] \) where \( id_k \) corresponds to a same product sample with different angles, views, colors and backgrounds in \( S \). This task emphasizes a higher level of fine-grained matching compared to traditional fine-grained retrieval tasks.

#### B. Dataset Analysis

1) **Dataset Statistics**: We collect several product samples from 49 different brands on E-commerce websites. These image-text samples are then manually sorted into single- and multi-product categories based on the associated product information.

For the **multi-product retrieval problem**, Product1M is divided into the **train**, **val**, **test**, and **gallery** sets. The **train** set has 1,132,830 samples, including both single- and multi-product samples, while the **val** and **test** sets contain only multi-product examples, totaling 2,673 and 6,547 samples, respectively. The **gallery** set contains 40,033 single-product samples in 458 categories, 392 of which occur in the **val** and **test** sets, while the others serve as interference objects for validating the robustness of a retrieval technique. It is important to note that the **gallery**, **val**, and **test** sets are annotated with class labels solely for evaluation purposes; they are not used in the training process, and the **train** set samples are not annotated. Table I and Fig. 3 summarize statistics of the data for Product1M.

As for the **identical-product retrieval task**, we extract a subset from our Product1M to construct **train**, **test**, and **gallery** sets. This subset totally contains 208,228 samples with 819 unique products including mixed single-product and multi-product samples. The **train**, **gallery** and **test** sets contain 192,164, 14,628, and 1,436 samples, respectively. All test samples are annotated by at least five crowdsourced workers, with each worker labeling a pair of data - one query data and one gallery data - to determine whether both samples are the same product. As depicted in Fig. 4, identical product retrieval is more fine-grained compared to traditional similarity retrieval tasks.

2) **Dataset Annotation**: We first select a query set from the whole dataset. To construct a reliable gallery set, we use a ResNet50 [71] and Bert-Base [68] to extract features and generate the query candidate pool from all the data that is not contained in the training subset. By concatenating the features and calculating the cosine similarity to all other instances in the dataset, we produce a pre-ranked candidate list to minimize the labeling cost. During the crowd-sourced annotation process, human workers review both images and captions in the candidate list to determine which samples match the query instance. The annotation system is identical to that described in [72].

3) **Dataset Characteristics**: Multi-product nature and complex combinations: Multi-product photos are prevalent
Fig. 3. Characteristics and statistics of Product1M: (1a) Complex combinations of single products; (1b) Weak supervision and fuzzy correspondence; (1c) Difficulties in real-world settings; (2) Long-tailed category distribution of Product1M. The line represents the sample size of each category in decreasing order. Product1M comprises a diverse set of categories and the long-tailed class distribution corresponds to real-world cases.

Fig. 4. Several comparison demos of identical-product retrieval and similarity product retrieval. Accurate/Inaccurate matched results are framed in green/red boxes.

on E-commerce websites and serve as the query images for product retrieval at the instance level. As shown in Fig. 3(1a), products may be structured in various ways and layouts, with a high number of instances. Due to the abundance and variety of fine-grained single-product samples, complex combinations appear in various portfolio photos.

Weak Supervision and Fuzzy Correspondence: We investigate retrieving data in two common modalities, namely pictures and text. Compared to datasets with precise class names, the oversight provided by commodity captions is inadequate and often uninformative. In Fig. 3(1b), we illustrate many sorts of difficult samples. Various samples’ captions include abbreviations, i.e., a reduced version of several goods. However, an acronym such as ‘eight-piece set’ has no information about the items. The second sort of sample contains irrelevant data, with the commodities specified in the title appearing in the picture but not in the title, or vice versa. The vast dispersion of imprecise connection between pictures and titles complicates instance-level retrieval even more.

Consistency With Real-World Scenarios: In Fig. 3(1c), we provide several difficult examples. They may feature a complicated backdrop with irrelevant elements, amorphous watermarks, or substantial clutter that obscures the product information. Certain items from various categories may seem practically identical save for the packaging description, e.g., day cream versus night cream. As seen in Fig. 3(2a, 2b), the long-tailed distribution of Product1M fits nicely with real-world settings.

C. Methodology

Fig. 5 depicts the general architecture of our end-to-end visual-language model (EGE-CMP). The network architecture of our model comprises three types of transformer modules, namely single-modal transformer (SMT), cross-modal transformer (CMT), and co-modal transformer (COMT), arranged from bottom to top. In addition, our model includes three modules based on function: Masked Relation Learning, Cross-modal Contrastive Learning and Entity-graph Enhanced Learning. In this section, we elaborate on the architectural design of EGE-CMP in Section III-C1, including detailed functions of three kinds of transformer modules and the direction of data flow. Then we describe three main function modules that enable the self-supervised learning of EGE-CMP in Section III-C2, C3 and C4. Finally, in Section III-C5, we present the inference procedure for instance-level retrieval.

1) Architectural Design of EGE-CMP: As previously mentioned, our EGE-CMP comprises three types of transformer
Fig. 5. Overview of the Entity-Graph Enhanced Cross-Modal Pretraining (EGE-CMP) model. The EGE-CMP model takes three types of data as input: Entity data is initially processed using NER tools (bottom-middle). A cross-modal contrastive loss is then employed to encourage closer alignment between modality data and their positive samples, while increasing separation from negative samples (right). Following the cross-transformer and CO-transformer stages, an entity-graph is constructed using the AdaGAE [73] graph network. To enhance the model’s attention towards meaningful entities and reduce ambiguity, node-based, and subgraph-based ranking losses are proposed (left).

Fig. 6. Entity data are directly extracted from caption data. Image, caption, and entity data are denoted via triples and fed into the model.

structures: SMT, CMT, and COMT. These structures respectively achieve single-modality feature extraction, cross-modal information exchange, and multi-modal information fusion. EGE-CMP takes an image-text pair \((I, C)\) as input where \(I\) represents the product image and \(C\) represents the caption data. Bottom-up-attention model [35] is used to process Image \(I\) and obtain the proposal features \(p = \{p_1, p_2, ..., p_m\}\) corresponding to the most likely detected objects and their position information \(q = \{q_1, q_2, ..., q_n\}\). Caption \(C\) is parsed by NER analysis tools to obtain the noun entities \(EC = \{E_1, E_2, ...\}\). In Fig. 6, we provide a demo of Image-Caption-Entity for intuitive understanding. Based on the above preparation, SMT extracts the embeddings \(v, t, e\) for three different inputs (image, caption, and entity). Then CMT performs cross-modal fusion respectively on \(v\) and \(t\) by exchanging key-value pairs to generate the new representations \(h_v\) and \(h_t\). Similarly, we use the same operation to produce new representations \(h_e\) and \(h_c\). To ensure representation consistency, we take the average of \(h_t\) and \(h_c\) to obtain the new representations of text \(h_t\). Finally, \(\{h_v, h_t\}\) and \(\{h_e, h_c\}\) are concatenated separately in the feature dimension and fed into two COMTs to produce joint image-text-entity representations \(f\) for multi-product retrieval and identical-product retrieval tasks. We give detailed explanations of each transformer structure below.

**Single-Modal Transformer (SMT):** For image input \(I\), the position information \(q\) is aligned with the same dimension as the proposed feature \(p\) using a linear projection layer. The aligned vectors are then added together as the input to obtain the preliminary image embeddings \(u_v\), which are obtained through the following equation:

\[
u_v = \text{SMT}(p + \text{LinearProj}(q)).
\]

For caption input \(C\) or entity input \(E\), WordPiece is used to tokenize the text into word tokens \(\{w_t^l\}_{l=1}^L\), where \(L\) is the length of tokens in caption \(C\) or entity \(E\). Our SMT embeds the discrete word token \(w_t\) into a continue word embedding \(u_t\) and \(u_e\):

\[
u_t = \text{SMT}(w_t), \quad u_e = \text{SMT}(w_e),
\]

where \(w_t\) and \(w_e\) are word tokens of caption input \(C\) and entity input \(E\), respectively.

**Cross-Modal Transformer (CMT):** By exchanging key-value pairs in the multi-headed attention mechanism, the Cross-Transformer is utilized to construct the inter-modal relations between the two distinct modalities. In detail, we use the query of one modality, the key, and the value from another modality for achieving cross attention between any two modalities. We
separately learn the new embeddings $h_{t_1}$ and $h_u$ between image and text modalities. Similarly, the new embeddings $h_{t_2}$ and $h_e$ are generated by CMT. Since the downstream tasks only require image and caption data, we select the $h_{t_1}$ as the output $h_t$ for the COMT module. The detailed process is shown as follows:

$$
\begin{aligned}
    h_{t_1}, h_u &= \text{CMT}(u_t, u_c), \\
    h_{t_2}, h_e &= \text{CMT}(u_t, u_c), \\
    h_t &= (h_{t_1} + h_{t_2})/2. \\
\end{aligned}
$$

(3)

**CO-Modal Transformer (COMT):** Given image features $h_v$, caption features $h_{t_1}$, and concatenated entity features $h_e$, COMT performs cross-modal fusion over all $\{h_v\}_{i=1}^L$, $\{h_{t_1}\}_{i=1}^L$, and $\{h_e\}_{i=1}^L$ for joint visual-language-entity learning. Because the caption and entity data could express the semantic information at different levels, we construct two branches of co-transformer to solve the problem. One branch is to learn the common representation between $h_v$ and $h_{t_1}$. We add a common learnable position embedding $q_{e1}$ to caption features $h_{t_1}$ and image features $h_v$, to incorporate sequence ordering and learn the low-level semantic information. Then, the image and caption representations after position embedding $q_{e1}$ and following a special [CLS] token are concatenated as the input to the branch of COMT. We repeat the same operation for $h_{t_1}$ and $h_e$, resulting in two groups of fused presentations: $[f_v, f_{t_1}]$ and $[f_e, f_{t_2}]$. To learn the fused text representation, we concatenate $f_{t_1}$ and $f_{t_2}$, and apply a linear projection layer to generate the final output $f_t$. The joint features $f$ are learned as follows:

$$
\begin{aligned}
    f_v, f_{t_1} &= \text{COMT}_1([\text{CLS}], [u_t, u_c] + q_{e1}), \\
    f_e, f_{t_2} &= \text{COMT}_2([\text{CLS}], [u_t, u_c] + q_{e1}), \\
    f_t &= \text{LinearProj}([f_{t_1}, f_{t_2}]), \\
    f &= [f_v, f_{t_1}, f_e].
\end{aligned}
$$

(4)

2) **EGE-CMP by Masked Multi-Modal Learning:** We employ multiple pretext tasks to facilitate the self-supervised learning of EGE-CMP, leveraging large-scale multi-modal data. For modality-specific feature learning, we incorporate three masked multi-modal modeling tasks, namely Masked Language Modeling task (MLM), Masked Region Prediction task (MRP), and Masked Entity Modeling (MEM) following the standard BERT [68] and VisualBERT [9]. MLM reconstructs the masked token information, enhancing language understanding with the support of visual and knowledge representations. MRP recovers the masked visual features, improving visual object perception. MEM refines the semantic knowledge understanding, enabling models to comprehend common concepts with real relevance.

**Masked Language Modeling (MLM):** In the task, approximately 15% of the text tokens are masked, and the remaining entity inputs are used to reconstruct the masked entity $x_e$ from the joint features $h_e$ learned through COMT module. The masked tokens $h_e^m$ are processed by a fully-connected (FC) layer ($FC_{MLM}$) and projected to a pre-defined discrete word space for classification:

$$
    x'_i = FC_{MLM}(h_e^m),
$$

$$
\mathcal{L}_{\text{MLM}} = -\mathbb{E} \left[ \frac{1}{|\mathcal{M}_{\text{MLM}}|} \sum_{i \in \mathcal{M}_{\text{MLM}}} \log P(x_i | x'_i) \right],
$$

(5)

where $\mathcal{M}_{\text{MLM}}$ represents the index set of masked text tokens.

**Masked Region Prediction (MRP):** Similar to MLM, we randomly mask approximately 15% of the proposal inputs $p^m_i$. Unlike the discrete work tokens, in the task, we directly regress the masked features $p_i$, which are supervised by the features extracted from the bottom-up attention model [35] with an MSE loss. The masked proposal inputs $h_i^m$ are fed into a fully-connected (FC) layer ($FC_{MRP}$) and project to a binary space for determining whether the input is masked:

$$
    p'_i = FC_{MRP}(p_i^m),
$$

$$
\mathcal{L}_{\text{MRP}} = \frac{1}{2 |\mathcal{M}_{\text{MRP}}|} \sum_{i \in \mathcal{M}_{\text{MRP}}} |p'_i - p_i^m|^2,
$$

(6)

where $\mathcal{M}_{\text{MRP}}$ denotes the index set of masked proposal inputs, and $p_i$ and $p'_i$ represent the predicted results and ground-truth, respectively.

**Masked Entity Modeling (MEM):** In MEM, approximately 15% of entity tokens are masked, and the remaining entity inputs are used to reconstruct the masked entity $x_e$ from the joint features $h_e$ learned through COMT module. The masked tokens $h_e^m$ are fed into a fully-connected (FC) layer ($FC_{MEM}$) to project them into a discrete pre-defined word space for classification:

$$
    x'_e = FC_{MEM}(h_e^m),
$$

$$
\mathcal{L}_{\text{MEM}} = -\mathbb{E} \left[ \frac{1}{|\mathcal{M}_{\text{MEM}}|} \sum_{i \in \mathcal{M}_{\text{MEM}}} \log P(x_e_i | x'_e) \right],
$$

(7)

where $\mathcal{M}_{\text{MEM}}$ is the index set of masked entity tokens.

3) **EGE-CMP by Cross-Modal Contrastive Loss:** In addition to intra-modal feature learning, our EGE-CMP is capable of constructing coherent representations of multi-modal inputs and learning their alignment.

To achieve this, we employ inter-modality contrastive learning [1], [74] to align the visual and textual modalities. There are $2N$ data points in total for a minibatch of $N$ image-text samples. We consider the related image-text pairings to be $N$ positive pairs and the remaining $2(N - 1)$ mismatched pairs to be negative pairs. Given an image-text pair $(x_i, x_j)$ and their encoded features $(\tilde{x}_i, \tilde{x}_j)$, the cross-modal contrastive loss for this positive pair is calculated as:

$$
\mathcal{L}_{\text{VT}}(x_i, x_j) = -\log \frac{\exp(\text{sim}(\tilde{x}_i, \tilde{x}_j) / \tau)}{\sum_{k=1}^{2N} \text{1}_{[k \neq i]} \exp(\text{sim}(\tilde{x}_i, \tilde{x}_k) / \tau)},
$$

(8)

where $\text{sim}(u, v) = u^\top v / ||u|| ||v||$ computes the cosine similarity of $(u, v)$ pairs, $\tau$ denotes the temperature parameter, and $\text{1}_{[k \neq i]}$ is a binary indicator function that returns 1 if $k \neq i$.

This type of contrastive loss encourages similarity between the encoded features of positive pairings from various modalities while distinguishing the encoded features of negative pairs.
To enrich the semantic information of the text encoder in an implicit manner, we stack single entity data as input and align it with corresponding caption data. We use cross-modal contrastive learning to train a pair of entity-caption data \((e_i, x_j)\) in a mini-batch of \(N\) entity caption pairs, which is represented as:

\[
\mathcal{L}_{ET}(e_i, x_j) = - \log \frac{\exp \left( \frac{\cos(e_i, \tilde{x}_j)}{\tau} \right)}{\sum_{k=1}^{N} \delta_{k \neq i} \exp \left( \frac{\cos(e_i, \tilde{x}_k)}{\tau} \right)}. \tag{9}
\]

4) EGE-CMP by Entity-Graph Enhanced Learning: To enhance our framework’s generalization capability and entity discrimination, we develop an effective entity-graph enhanced (EGE) module that explicitly incorporates entity knowledge into the model training process. Given a caption \(C\) with a set of entities \(E_C = \{E_1, E_2, \ldots\}\) extracted directly from the caption data, the EGE module aims to learn a semantic graph connecting each entity for more accurate instance-level retrieval. In the EGE module, a unique entity queue \(l\) is initialized, and each entity within the queue is encoded by our COMT module to obtain the joint embeddings \(f_e\). These embeddings are then fed to the AdaGAE network [73], an effective graph clustering network that learns better graph semantic relationships \(f_g\) between entities. We propose node-based ranking loss and subgraph-based ranking loss to ensure each entity pays more attention to its semantically close neighbors. The module’s process is illustrated in Fig. 7.

Drawing inspiration from the memory bank of contrastive learning, we maintain a queue \(l\) during our training process for more accurate negative sample selection. The queue is initialized with an entity dictionary containing all entity data, with its length equal to the number of unique entities in our dataset.

To update the queue \(l\), all entities are fed into the COMT module to obtain feature embeddings \(l_f\). During the graph learning process, each entity is considered a node, and the similarities between different nodes form edges. To learn a better graph representation, we input the entity embedding \(l_f\) into the AdaGAE [73] model to learn the graph relationship representations \(l_g\) between different entities as prior knowledge. For better entity correlation in our EGE-CMP model, cosine similarity \(S_g\) of \(l_g\) is used as the entity graph to determine which entity is a negative sample. During network training, negative sample selection is performed for a minibatch of data based on similarity \(S_g\). For any entity \(h_{e_i}\), in a minibatch of data with \(l_{m}\) length, one of the last \(k\) entity samples \(h_{e_k}\) is randomly selected as the negative sample. The node-based ranking loss is represented as:

\[
\mathcal{L}_{node}(h_{e_i}, h_{e_k}, y) = \max(0, -y * (h_{e_i} - h_{e_k}) + \text{margin}), \tag{10}
\]

where \(y\) equals 1 and margin is to set 0.

Additionally, we propose a sub-graph-based loss for preserving the relationships of nearest neighbor graphs for both positive and negative data points. In our approach, the feature embedding of any sub-graph is obtained through global-add pooling, which aggregates information from its \(k\) nearest neighbor nodes. For a positive and negative nearest neighbor subgraph denoted as \(h_{g_i}\) and \(h_{g_k}\) respectively, with respect to node \(h_{e_i}\), we define the following formulation:

\[
\mathcal{L}_{subgraph}(h_{g_i}, h_{g_k}, y) = \max(0, -y * (h_{g_i} - h_{g_k}) + \text{margin}), \tag{11}
\]

where \(y\) equals 1 and margin is to set 0.

We jointly optimize our model by a weighted loss:

\[
\mathcal{L}_{EGE-CMP} = \mathcal{L}_{ET} + \mathcal{L}_{VT} + \mathcal{L}_{MLM} + \mathcal{L}_{MRM} + \mathcal{L}_{MEM} + \mathcal{L}_{node} + \mathcal{L}_{subgraph}. \tag{12}
\]

5) Inference for Instance-Level Retrieval: For both the single- and multi-product samples, the proposal-level features are extracted via the bottom-up-attention network [35] pre-trained on the Visual Genome dataset [37]. Both captions and entities serve as input to be fed into EGE-CMP. During inference, the COMT layer produces \(f_e\) and \(f_t\) as the feature embeddings of the visual and textual inputs, respectively, during inference. These embeddings are concatenated and normalized to obtain the joint representations of the instance. Additionally, since the Text/Visual and Text/Entity Transformers are supervised using cross-modal contrastive loss, we could believe that concatenating the feature embeddings of COMTs for retrieval is useful. Our retrieval model then uses the generated features as input. For both multi-product and identical-product retrieval tasks, after generating the cosine similarity matrix between an instance and the gallery set samples, the single-product samples with the greatest similarity are selected and returned for each query.

IV. EXPERIMENTS

A. Implementation Details

We utilize the bottom-up-attention model pre-trained [35] on the VG dataset [37] to extract proposal-wise features for visual data. Additionally, we use the BERT [68] model to initialize the linguistic transformer of our EGE-CMP. We limit the number of input areas to between 10 and 36 by picking regions with predicted confidence greater than a predefined threshold as in [7]. The region’s features are then flattened and input into EGE-CMP using RoAlign. For the entity extraction, We use the NER API provided by Alibaba company. The NER contains 40 detailed annotations including Common words, Material, Style, Style elements, Color, Brand, Functional efficacy, Dimensions, Quality and fineness, Scenes, Crowd, Set, Time season,
Model, New product, Series, Marketing service, Location, Person’s name, Entertainment, Institutional entity, Movie name, Game name, Number, Unit, Category, New word, Modifier, Proper noun, Category modifier, Symbol, Prefix, Suffix, Gift, Negative, Acting. However, in our paper, we only focus on five NER annotations: Quality and fineness, Brand, Category modifier, Set, Modifier and New word as the selected entities to construct our entity set. We set the number of transformer layers to 12, with Text/Visual/Entity Transformer, Text/Visual and Text/entity Cross-Transformer, and Co-Transformer set to $L = 4$, $K = 4$, and $H = 4$, respectively. To ensure a fair comparison, we set the hidden state size of EGE-CMP and other baselines to 768. The transformer blocks in EGE-CMP have a hidden state size of 768 with 8 attention heads. Furthermore, we separately attach a 512 dimensional fully connected layer after Co-Transformer and Text/Visual/Entity Transformer for task-agnostic masked learning and cross-modal contrastive learning. By concatenating the features from the final Co-transformer layer, a 1024 dimensional feature vector is obtained for retrieval, which is also the same for other baselines. The maximum sequence length for the sentence is set to 36. During training, we use a total batch size of 128 for 10 epochs on 8 RTX 3090 GPUs. Additionally, we use an Adam [75] optimizer with an initial learning rate of 1e-4 and a linear learning rate decay schedule in the EGE-CMP training process. The entity numbers are limited to 12,000 for faster training speed. During inference, EGE-CMP generates instance features from texts and proposal-wise features. We normalize the instance features of the image/text and concatenate them as the overall representations before retrieval. For the retrieval task, we compute cosine similarities between each box query with single-product samples using the overall representations mentioned above and score all retrieval single-product results based on their similarity. The top-$N$ samples are returned as the retrieval results. We include Figs. 8, 9 and 10 to showcase the visualization results by EGE-CMP on multi-product retrieval and identical product retrieval tasks. To ensure a fair comparison with other baselines, we conduct all experiments using the same training settings and assessment procedures.
### B. Evaluation Metrics

We use two commonly used measures \cite{78,79} Precision (Prec@N), mean Average Precision (mAP@N) and a new metric mean Average Recall (mAR@N) to evaluate the instance-level retrieval performances. To avoid the unnecessary and impractical retrieval of each and every product, we report mAP@N, mAR@N, and Prec@N, where N = 10, 50, 100. Prec@N is widely used in the retrieval literature that evaluates the average accuracy of the top-N predictions per image. Specifically, Prec@N(q) is defined as follows:

\[
\text{Prec}@N(q) = \frac{1}{N} \sum_{i=1}^{N} \text{acc}_q(i),
\]

where acc_q(i) is a binary indicator function that returns 1 when the i-th prediction is correct for the q-th query and 0 otherwise. mAP@N is computed as the average AP@N per image, where AP@N(q) is computed as follows:

\[
\text{AP}@N(q) = \frac{1}{\text{min}(m_q, N)} \sum_{k=1}^{N} P_q(k) \text{rel}_q(k),
\]

where m_q is the total number of ground truth images, i.e., corresponding single-product images that appear in the q-th query image. \(P_q(k)\) is the precision at rank k for the q-th query, and \(\text{rel}_q(k)\) is a binary indicator function that returns 1 when the k-th prediction is correct for the q-th query and 0 otherwise.

To evaluate the recall of instance-level retrieval results, we propose metric mAR@N, which can be computed as the average AR@N per image. AR@N(q) is re-defined in our paper as follows:

\[
\text{AR}@N(q) = \frac{1}{C_q} \sum_{c=1}^{C} \text{Ret}^c(q) \min \left( \frac{\text{Ret}^c_r}{\min(\text{Ret}^c_r, N_q, C_q)}, 1 \right),
\]

where C is the total number of single-product categories in the gallery set, C_q equals to the number of existing categories in the q-th query, \(\text{Ret}^c(q)\) is a binary indicator function that returns 1 when class c exists in the q-th query, \(\text{Ret}^c_r\) is the number of retrieved products belonging to class c for the q-th query, \(N_q\) is the number of ground truths belonging to class c in the gallery set, \(r_q^c\) is the instance ratio of category c in the q-th query, and \([\cdot]\) is rounding operation. As per the equation, AR@N(q) takes the category distribution into account, i.e., the inclusion of instance ratio is informative for evaluating both the correctness and diversity of a retrieval algorithm and guarantees that some trivial results are not overestimated.\(^4\)

### C. Weakly-Supervised Instance-Level Retrieval

Tables II and III show the retrieval performances compared with several intra- and cross-modal models for both multi-product and identical product retrieval tasks.

#### 1) Multi-Product Retrieval Performance: Intra-Modal Schemes:

Our EGE-CMP model is evaluated against two intra-modal schemes: Image-based and Text-based schemes. To perform image-based retrieval, we use the Image SMT layer as depicted in Section III-C1 and apply the same image input and pretext tasks as those used for EGE-CMP, namely MRP. Similarly, for text-based retrieval, we employ the Text SMT layer with only textual input and MLM pretext task. To ensure a fair comparison, the depth of these two models is increased to 24 layers in order to maintain the same number of parameters as EGE-CMP. In addition to two uni-modal baseline methods, we compare our methods with two unsupervised image retrieval models SPQ \cite{76} and MeCoQ \cite{77}. We replace their CNN model backbone (VGG \cite{80} and Resnet \cite{71}) with vision transformer \cite{81} for more fair verification. According to the experimental results shown in Tables II and III, we can observe that these two schemes trail significantly behind since they are restricted to unimodal data, indicating the importance of modeling the relationship between multimodality data. In Section IV-E, we provide additional experiment findings to substantiate this conclusion.

#### Cross-Modal Schemes: In Table II, we present a comparison of EGE-CMP against several commonly used self-supervised cross-modal pretraining approaches. These include state-of-art

| Method            | mAP@10 | mAP@50 | mAP@100 | mAR@10 | mAR@50 | mAR@100 | Prec@10 | Prec@50 | Prec@100 |
|-------------------|--------|--------|---------|--------|--------|---------|---------|---------|----------|
| Image-based       | 47.46  | 39.79  | 37.06   | 20.81  | 16.12  | 14.63   | 36.28   | 31.20   | 29.20    |
| Text-based        | 68.85  | 62.21  | 60.56   | 24.01  | 17.48  | 16.44   | 62.48   | 59.26   | 54.86    |
| SPQ \cite{75}     | 62.67  | 55.49  | 52.29   | 21.99  | 15.71  | 14.94   | 52.64   | 45.68   | 41.83    |
| MeCoQ \cite{76}   | 65.33  | 59.72  | 55.76   | 22.18  | 16.85  | 15.98   | 56.73   | 47.58   | 42.41    |

\(^4\)For instance, for a 2A+3B query image, \(r_q^c = 0.4\) and \(r_q^d = 0.6\).

\(^5\)For a 2A+3B query image and \(N = 100\), AR@100(q) returns 0.51 and 1.0 for retrieval results 1A+99B and 40A+60B, respectively.
single- and two-stream Vision-language models, as well as a state-of-art zero-shot classification model, i.e., CLIP [1]. CLIP* refers to a CLIP-like architecture that encodes images and text separately and is trained using contrastive learning. Notably, EGE-CMP outperforms all of these benchmarks on the majority of three major retrieval criteria. Two-stream models, such as LXMERT [10], ViLBERT [7], and CLIP*, do not perform better than single-stream models, indicating that the manner of fusion of cross-modal information is an important component. It is important to note that EGE-CMP outperforms CAPTURE [3] when employing the hybrid transformer architecture by using entity knowledge. We attribute EGE-CMP’s superior performance to its entity knowledge injection. Finally, we investigate the impact of various layer types in Section IV-E.

2) Identical Product Retrieval Performance: With the same settings as the multi-product retrieval task, we also perform identical-product retrieval on all the compared methods and our EGE-CMP model. Table III presents the experimental results. We observe a slight improvement in the performance of all methods due to the limited number of the gallery set. Meanwhile, compared with multi-product retrieval, the performances of the multi-modal pretraining models, i.e., ViLBert [7], UNITER [2], CAPTURE [3] and so on, are much closer, indicating that identical product retrieval is a more challenging task. In both mAP and Prec metrics, our EGE-CMP still achieves the best performance. This outcome provides strong evidence that the use of commonsense words or phrases existing in caption data is beneficial to performance improvement.

D. Impact of Entity Pretext Tasks and Graph Ranking Loss

Table IV provides the experimental results of our EGE-CMP under different entity loss settings and different feature representations, including Masked Entity Modeling, Entity Node Loss, and Entity Subgraph Loss. We leverage Masked Entity Modeling (E-Masked) at the Entity/Text and Text/Visual COMT layer to enhance the representation of tokens related to the real contexts, which benefits the learning of modality combination. In terms of mAP metrics, E-Masked Loss improves our model’s retrieval performance by 2.8, 7.0, and 5.2 respectively (#1 vs #2) under the different feature representations. We also observe that this

| Method          | mAP@1 | mAP@5 | mAP@10 | Prec@1 | Prec@5 | Prec@10 |
|-----------------|-------|-------|--------|--------|--------|---------|
| Image-based     | 52.90 | 44.65 | 41.51  | 52.90  | 35.22  | 33.75   |
| Text-based      | 70.21 | 66.21 | 63.88  | 70.21  | 63.87  | 60.54   |
| SPQ [75]        | 64.72 | 60.43 | 58.34  | 64.72  | 59.74  | 56.45   |
| MeCoQ [76]      | 66.56 | 61.37 | 58.64  | 66.56  | 60.22  | 57.95   |
| ViLBERT [7]     | 81.17 | 83.45 | 81.88  | 81.74  | 63.07  | 56.99   |
| LXMIERT [10]    | 81.09 | 82.02 | 81.48  | 81.09  | 64.76  | 58.78   |
| CLIP* [1]       | 81.09 | 82.02 | 81.11  | 81.00  | 62.43  | 57.98   |
| Vl-BERT [8]     | 81.39 | 83.02 | 81.44  | 81.23  | 64.05  | 58.41   |
| VisualBERT[9]   | 79.90 | 81.91 | 80.32  | 79.90  | 62.39  | 55.87   |
| UNITER [2]      | 81.74 | 83.45 | 81.88  | 81.74  | 63.07  | 56.99   |
| CAPTURE [3]     | 82.03 | 83.60 | 82.04  | 82.03  | 65.22  | 59.03   |
| EGE-CMP (Ours)  | 84.66 | 86.37 | 84.79  | 84.66  | 67.76  | 62.00   |

The bold values indicate the best performance.

D. Impact of Entity Pretext Tasks and Graph Ranking Loss

Table IV provides the experimental results of our EGE-CMP under different entity loss settings and different feature representations, including Masked Entity Modeling, Entity Node Loss, and Entity Subgraph Loss. We leverage Masked Entity Modeling (E-Masked) at the Entity/Text and Text/Visual COMT layer to enhance the representation of tokens related to the real contexts, which benefits the learning of modality combination. In terms of mAP metrics, E-Masked Loss improves our model’s retrieval performance by 2.8, 7.0, and 5.2 respectively (#1 vs #2) under the different feature representations. We also observe that this
improvement is particularly pronounced in the deeper layers of the model. Moreover, after applying Entity Node Loss to the features from the Text/Visual Transformer and those from the Co-Transformer for retrieval, it further improves mAP metrics, respectively (#2 vs #3). We also verify the effectiveness of Entity Subgraph Loss and observe that it is quite useful to further improve the model performance, suggesting that the improvement mainly comes from the Entity learning.

### E. Impact of Layer Configuration

We examine how the arrangement of transformer layers may affect the performance of our model in Table V. In the Config column, the triplet represents the number of Text/Visual Transformer, Cross-Transformer, and Co-Transformer layers, respectively. For a fair comparison, we initially eliminate layers of a certain type while keeping the depth of the resulting network at the same level as EGE-CMP’s, i.e., 12 layers. The terms ‘w/o-Cross’, ‘w/o-Co’, and ‘w/o-T/V/E’ denote the models obtained by removing the EGE-CMP Cross-Transformer, Co-Transformer, and Text/Visual/Entity Transformer layers. As can be seen, these three models perform worse than EGE-CMP, demonstrating the efficacy of its hybrid-stream design. Furthermore, we find that a single model transformer (SMT) is crucial for our model, indicating that a finely-tuned transformer is more beneficial for the overall performance of the model. Additionally, we display the embeddings created by UNITER, CAPTURE, and EGE-CMP in Fig. 11 using t-SNE [82]. It turns out that the EGE-CMP characteristics are more discriminative, which is advantageous for the retrieval job. Meanwhile, we also show the generated subgraphs via our EGE-CMP model in Fig. 12. From the figure, we can observe that our entity graph could learn the product correlations not only from the semantic similarity but also from the product functions.

### G. Comparisons on Single-Product Retrieval

It is worth noting that EGE-CMP can be employed for both single\(^6\) and multi-product retrieval. Indeed, it excels at these two tasks and surpasses other baselines when dealing with single-product retrieval. Specifically, we select a single-product sample from the gallery set as a query and execute single-product retrieval from the remaining samples of the gallery set. The performance of four models, namely UNITER-single, LXMERT-single, CAPTURE-single, and EGE-CMP-single, is compared in Table VII. As can be seen, single-product retrieval performs considerably better than multi-product retrieval, as the challenge is significantly reduced when there is only one instance/entity in the image/text. Furthermore, we observe that the performance of ‘EGE-CMP- single’ surpasses that of the updated and collecting vast volumes of clean labels is too expensive. In contrast to detection methods, our retrieval approach does not depend on a fixed-size collection of predefined classes or on fine-grained box annotations. To demonstrate this point, we perform zero-shot retrieval studies and provide the results in Table VI. We manually remove 5/10/20 brands from the train set and train EGE-CMP on the remaining samples to avoid discarding the deleted categories during training. Subsequently, we evaluate EGE-CMP on the classes of these unseen brands. Furthermore, we compare our model with the LXMERT two-stream model and the UNITER single-stream model. As can be observed, EGE-CMP outperforms LXMERT and UNITER across all three evaluation criteria, thus highlighting its generalizability. Additionally, we display the embeddings created by UNITER, CAPTURE, and EGE-CMP in Fig. 11 using t-SNE [82]. It turns out that the EGE-CMP characteristics are more discriminative, which is advantageous for the retrieval job. Meanwhile, we also show the generated subgraphs via our EGE-CMP model in Fig. 12. From the figure, we can observe that our entity graph could learn the product correlations not only from the semantic similarity but also from the product functions.

\(^6\)The single-product retrieval is a specific case of identical product search. There is only one instance in all product samples.
other two baselines, highlighting the superiority of EGE-CMP in single-product retrieval.

### H. Complexity Analysis

Table VIII presents the FLOPS values for various models. From the table, it becomes apparent that certain well-known approaches, such as ViLBERT [7], VisualBERT [9], and UNITER [2], which lack cross-modal attention with key-value exchange, exhibit lower computational complexity. In the CAPTURE [3] method, an additional cross-modal attention module and a multi-layer joint attention module are incorporated into the network, resulting in increased algorithmic complexity. In our methods, we have two kinds of cross-modal attention modules (Visual-Text and Entity-Text) to fully utilize the cross-modality information and an extra entity queue to facilitate entity relationship learning. Hence, compared with other models, our model has higher computational complexity.

### V. Conclusion

In this study, we have introduced a pioneering effort to extend the scope of canonical intra-/cross-modal retrieval towards a more generalized scenario known as weakly-supervised multi-modal instance-level product retrieval. To facilitate multi-modal
pre-training, we introduce the Product1M, one of the largest available visual-language retrieval datasets, and the only one designed specifically for instance-level retrieval in real-world contexts. Additionally, we present an entity-graph enhanced hybrid-stream transformer, termed EGE-CMP, which excels in capturing potential synergy and understanding common relevant data across diverse modalities. We incorporate authoritative semantic information into the entity knowledge, concentrating cross-modal contrastive learning on related regions or words within multi-modal characteristics. Extensive experimental evidence confirms that our EGE-CMP outperforms several state-of-the-art multi-modal pretraining models across a broad array of metrics. We believe that the proposed EGE-CMP model, Product1M dataset, and established baselines will stimulate further research, forming a more trustworthy and adaptable retrieval search foundation. In the future, we want to make the model adaptively learn the instance areas rather than prepared object proposals for lower computational complexity.

ACKNOWLEDGMENT

We thank MindSpore for the partial support of this work, which is a new deep learning computing framework.

REFERENCES

[1] A. Radford et al., “Learning transferable visual models from natural language supervision,” in Proc. 35th Int. Conf. Mach. Learn., M. Meila and T. Zhang, Eds., 2018, pp. 153–161.

[2] Y.-Y. Chen et al., “UNITER: Universal image-text representation learning,” in Proc. Eur. Conf. Comput. Vis., Springer, 2020, pp. 740–756.

[3] H. Luo et al., “CLIP4Clip: An empirical study of CLIP for end to end video clip retrieval,” 2021, arXiv:2104.08860.

[4] https://www.mindspore.cn/
[42] X.-S. Wei, Q. Cui, L. Yang, P. Wang, and L. Liu, “RPC: A large-scale retail product checkout dataset,” 2019, arXiv:1901.07249.

[43] C. Corbierre, H. Ben-Younes, A. Ramé, and C. Ollion, “Leveraging weakly annotated data for fashion image retrieval and label prediction,” in Proc. Int. Conf. Comput. Vis. Workshops, 2017, pp. 2268–2274.

[44] F. Ensan and E. Bagheri, “Document retrieval model through semantic linking,” in Proc. Int. Conf. Web Search Data Mining, 2017, pp. 181–190.

[45] H. Noh, A. Araujo, J. Sim, T. Weyand, and B. Han, “Large-scale image retrieval with attentive deep local features,” in Proc. Int. Conf. Comput. Vis., 2017, pp. 3456–3465.

[46] S. Butcher, C. L. Clarke, and G. V. Cormack, Information Retrieval: Implementing and Evaluating Search Engines. Cambridge, MA, USA: MIT Press, 2016.

[47] D. Harman et al., “Information retrieval: The early years,” Found. Trends Inf. Retrieval, vol. 13, no. 5, pp. 425–577, 2019.

[48] X. Ji, W. Yang, M. Zhang, and Y. Yang, “Cross-domain image retrieval with attention modeling,” in Proc. ACM Int. Conf. Multimedia, 2017, pp. 1654–1662.

[49] Z. Lin, Z. Yang, F. Huang, and J. Chen, “Regional maximum activations of convolutions with attention for cross-domain beauty and personal care product retrieval,” in Proc. ACM Int. Conf. Multimedia, 2018, pp. 2073–2077.

[50] W. Wang, X. Yang, B. C. Ooi, D. Zhang, and Y. Zhuang, “Effective deep learning-based multi-modal retrieval: VLDJ,” vol. 25, no. 1, pp. 79–101, 2016.

[51] P. Tang et al., “PCL: Proposal cluster learning for weakly supervised object detection.” IEEE Trans. Pattern Anal. Mach. Intell., vol. 42, no. 1, pp. 176–191, Jan. 2020.

[52] Y. Shen, R. Ji, S. Zhang, W. Zuo, and Y. Wang, “Generative adversarial learning towards fast weakly supervised detection,” in Proc. IEEE/CVF Comput. Vis. Pattern Recognit., 2018, pp. 5764–5773.

[53] K. Ye, M. Zhang, A. Kovashka, W. Li, D. Qin, and J. Berent, “Cap2Det: Learning to amplify weak caption supervision for object detection,” 2019, pp. 9685–9694.

[54] X. Li et al., “Oscar: Object-semantics aligned pre-training for vision-language tasks,” in Proc. Eur. Conf. Comput. Vis., Springer, 2020, pp. 121–137.

[55] D. Gao et al., “FashionBERT: Text and image matching with adaptive loss for cross-modal retrieval,” in Proc. Int. ACM SIGIR Conf. Res. Develop. Informat. Retrieval, 2020, pp. 2251–2260.

[56] W. Li et al., “UNIMO: Towards unified-modal understanding and generation via cross-modal contrastive learning.” 2020, arXiv:2012.15409.

[57] T. Fu et al., “VIOLET: End-to-end video-language transformers with masked visual-token modeling,” 2021, arXiv:2111.12681.

[58] J. Yang, Y. Bisk, and J. Gao, “TACO: Token-aware cascade contrastive learning for video-text alignment,” in Proc. Int. Conf. Comput. Vis., 2021, pp. 11 542–11 552.

[59] L. Yao et al., “FILIP: Fine-grained interactive language-image pre-training,” in Proc. Int. Conf. Learn. Representations, 2022. [Online]. Available: https://openreview.net/forum?id=spDhcsEDC2

[60] W. Kim, B. Son, and I. Kim, “ViLT: Vision-and-language transformer without convolution or region supervision,” in Proc. 38th Int. Conf. Mach. Learn., M. Meila and T. Zhang, Eds., Jul. 18–24, 2021, vol. 139, pp. 5583–5594.

[61] L. Yuan et al., “Florence: A new foundation model for computer vision,” 2021, arXiv:2111.11432.

[62] J. Lin et al., “M6: A Chinese multimodal pretrainer,” 2021, arXiv:2103.00823.

[63] Z. Wang, J. Yu, A. W. Yu, Z. Dai, Y. Tsvetkov, and Y. Cao, “SimViLM: Simple visual language model pretraining with weak supervision,” 2021, arXiv:2108.10904.

[64] N. Mu, A. Kirillov, D. A. Wagner, and S. Xie, “SLIP: Self-supervision meets language-image pre-training,” in Proc. 17th Eur. Conf. Comput. Vis. (ECCV), S. Avidan, G. J. Brostow, M. Cissé, G. M. Farinella, and Tal Hassner, Eds., Tel Aviv, Israel, Oct. 23–27, 2022, vol. 13686, pp. 529–544.

[65] C. Jia et al., “Scaling up visual and vision-language representation learning with noisy text supervision,” in Proc. Int. Conf. Mach. Learn., ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., PMLR, 2021, pp. 4904–4916.

[66] X. Zhu et al., “Uni-perceiver: Pre-training unified architecture for generic perception for zero-shot and few-shot tasks,” 2021, arXiv:2112.01522.

[67] Y. Huo et al., “WenLan: Bridging vision and language by large-scale multi-modal pre-training,” 2021, arXiv:2103.06561.

[68] J. Devlin, M. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics - Hous. Lang. Technol., J. Burstein, C. Doran, and T. Solorio, Eds., 2019, pp. 4171–4186.

[69] W. Xiong, J. Du, W. Y. Wang, and V. Stoyanov, “PrereRaied encyclopedia: Weakly supervised knowledge-pretrained language model,” in Proc. Int. Conf. Learn. Representations, 2020.

[70] L. He, S. Zheng, T. Yang, and F. Zhang, “Klimo: Knowledge graph enhanced pretrained language model with fine-grained relationships,” in Proc. Conf. Empir. Methods Natural Lang. Process., 2021, pp. 4536–4542.

[71] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE/CVF Comput. Vis. Pattern Recognit., 2016, pp. 770–778.

[72] X. Dong et al., “M5Product: Self-harmonized contrastive learning for E-commerce multi-modal pretraining,” in Proc. IEEE/CVF Comput. Vis. Pattern Recognit., 2022, pp. 21–21 230.

[73] X. Li, H. Zhang, and R. Zhang, “Adaptive graph auto-encoder for general data clustering,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 44, no. 12, pp. 9725–9732, Dec. 2022.

[74] T. Chen, S. Komblith, M. Norouzi, and G. Hinton, “A simple framework for contrastive learning of visual representations,” in Proc. Int. Conf. Mach. Learn., PMLR, 2020, pp. 1597–1607.

[75] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” 2014, arXiv:1412.6980.

[76] Y. K. Jang and N. I. Cho, “Self-supervised product quantization for deep unsupervised image retrieval,” in Proc. Int. Conf. Comput. Vis., 2021, pp. 12065–12074.

[77] J. Wang, Z. Zeng, B. Chen, T. Dai, and S.-T. Xia, “Contrastive quantization with code memory for unsupervised image retrieval,” in Proc. AAAI Conf. Artif. Intell., 2022, pp. 2468–2476.

[78] T. Weyand, A. Araujo, B. Cao, and J. Sim, “Google landmarks dataset v2 - A large-scale benchmark for instance-level recognition and retrieval,” in Proc. IEEE/CVF Comput. Vis. Pattern Recognit., 2020, pp. 2572–2581.

[79] D. Cai, X. Gu, and C. Wang, “A revisit of deep hashings for large-scale content based image retrieval,” 2017, arXiv:1711.06016.

[80] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in Proc. Int. Conf. Learn. Representations, Y. Bengio and Y. LeCun, Eds., 2015.

[81] A. Dosovitskiy et al., “An image is worth 16x16 words: Transformers for image recognition at scale,” in Proc. Int. Conf. Learn. Representations, 2021.

[82] L. Van der Maaten and G. Hinton, “Visualizing data using t-SNE,” J. Mach. Learn. Res., vol. 9, no. 11, pp. 2579–2605, 2008.
Yunchao Wei received the PhD degree from Beijing Jiaotong University in 2016. He is a full professor of Beijing Jiaotong University. He was a senior lecturer of Australian Artificial Intelligence Institute with the University of Technology Sydney from 2019 to 2021. He was a postdoctoral researcher in Beckman Institute with UIUC from 2017 to 2019, and a postdoctoral researcher of National University of Singapore from 2016 to 2017. His current research interest focuses on applying deep learning to computer vision tasks.

Xiaoyong Wei is a professor and the head of the Department of Computer Science, Sichuan University of China since 2010, and is a visiting professor of the Department of Computing, The Hong Kong Polytechnic University. His research interests include multimedia computing, health computing, machine learning, and large-scale data mining.

Yaowei Wang (Senior Member, IEEE) received the PhD degree in computer science from the Graduate University of Chinese Academy of Sciences, Beijing, China, in 2005. He was a professor with the National Engineering Laboratory for Video Technology (NELVT), Peking University Shenzhen Graduate School, Shenzhen, China, in 2019. From 2014 to 2015, he worked as an academic visitor with the Vision Laboratory, Queen Mary University of London, London, U.K. He worked with the Department of Electronics Engineering, Beijing Institute of Technology, Beijing, from 2005 to 2019. He is currently an associate professor with the Peng Cheng Laboratory, Shenzhen.

Minlong Lu received the BEng degree from Zhejiang University in 2011 and the PhD degree in computer science from Simon Fraser University, Canada, and Zhejiang University, China, in 2018. He was a postdoctoral research assistant with the Vision and Media Lab, Simon Fraser University, advised by Prof. Z.-N. Li. He is currently a research engineer with Alibaba and Ant Group. His research interests are in the area of multimedia content analysis.

Xiaochun Cao received the BE and ME degrees both in computer science from Beihang University (BUAA), China, and the PhD degree in computer science from the University of Central Florida, USA, with his dissertation nominated for the university level Outstanding Dissertation Award. He is a professor with the School of Cyber Science and Technology, Sun Yat-Sen University. After graduation, he spent about three years at ObjectVideo Inc. as a research scientist. From 2008 to 2012, he was a professor with Tianjin University.

Xiaodan Liang (Member, IEEE) received the PhD degree from Sun Yat-sen University in 2016. She is currently an associate professor with Sun Yat-sen University. She was a postdoc researcher with the machine learning department, Carnegie Mellon University, working with Prof. Eric Xing, from 2016 to 2018. She has published several cutting-edge projects on human-related analysis, including human parsing, pedestrian detection, and instance segmentation, 2D/3D human pose estimation, and activity recognition.