FashionKLIP: Enhancing E-Commerce Image-Text Retrieval with Fashion Multi-Modal Conceptual Knowledge Graph

Xiaodan Wang[1], **Chengyu Wang**[2], **Lei Li**[3], **Zhixu Li**[1*], **Ben Chen**[2], **Linbo Jin**[2], **Jun Huang**[2], **Yanghua Xiao**[1*], **Ming Gao**[3]
[1] Shanghai Key Laboratory of Data Science, School of Computer Science, Fudan University
[2] Alibaba Group, Hangzhou, China [3] East China Normal University, Shanghai, China
{xiaodanwang20,zhixuli,shawyh}@fudan.edu.cn
{chengyu.wcy,chenben.cb,yuyi.jlb,huangjun.hj}@alibaba-inc.com
leili@stu.ecnu.edu.cn, mgao@dase.ecnu.edu.cn

Abstract

Image-text retrieval is a core task in the multi-modal domain, which arises a lot of attention from both research and industry communities. Recently, the booming of vision-language pre-trained (VLP) models has greatly enhanced the performance of cross-modal retrieval. However, the fine-grained interactions between objects from different modalities are far from well-established. This issue becomes more severe in the e-commerce domain, which lacks sufficient training data and fine-grained cross-modal knowledge. To alleviate the problem, this paper proposes a novel e-commerce knowledge-enhanced VLP model FashionKLIP. We first automatically establish a multi-modal conceptual knowledge graph from large-scale e-commerce image-text data, and then inject the prior knowledge into the VLP model to align across modalities at the conceptual level. The experiments conducted on a public benchmark dataset demonstrate that FashionKLIP effectively enhances the performance of e-commerce image-text retrieval upon state-of-the-art VLP models by a large margin. The application of the method in real industrial scenarios also proves the feasibility and efficiency of FashionKLIP. [1]

1 Introduction

The explosive growth of multi-modal content on the Web has promoted the research of various cross-modal tasks. Image-text retrieval, which finds correlated texts (or images) for a given image (or text) (Karpathy and Fei-Fei, 2015; Faghri et al., 2017), is a popular cross-modal task with strong practical values in a wide range of industrial applications. Recently, the booming of vision-language pre-trained (VLP) models (Yao et al., 2021; Zeng et al., 2021; Li et al., 2020c) has greatly improved the representation learning across data of different modalities, leading to significant performance improvement.

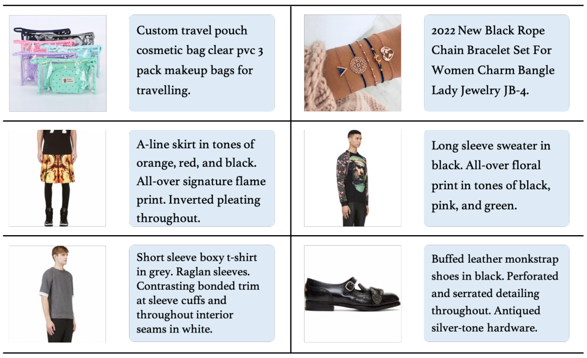

Figure 1: Examples of image-text pairs in e-commerce.

However, in the field of e-commerce, the image-text retrieval task has its own challenges. Here, we suggest that image-text pairs of products have unique characteristics that are different from the general domain (such as MS-COCO (Lin et al., 2014), Flickr30k (Young et al., 2014) and Conceptual Captions (Sharma et al., 2018)), with examples shown in Figure 1. 1) While most texts in the general domain contain descriptions with complete sentence structures, descriptions or queries in e-commerce are usually composed of multiple phrases, describing product details such as materials or styles. 2) Images in the general domain usually have rich backgrounds; in contrast, a product image mainly consists of a large commodity figure in the center without a lot of background objects. These unique domain characteristics make general-domain models difficult to be directly adopted to the image-text retrieval tasks in e-commerce.

Recently, several domain-specific VLP models including FashionBERT (Gao et al., 2020), Kalei-

[1]All the codes and model checkpoints have been released to public in the EasyNLP framework (Wang et al., 2022). URL: `https://github.com/alibaba/EasyNLP`.

*Corresponding author.

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 5: Industry Track, pages 149–158
July 10-12, 2023 ©2023 Association for Computational Linguistics

doBERT (Zhuge et al., 2021), CommerceMM (Yu et al., 2022), EI-CLIP (Ma et al., 2022) and Fashion-ViL (Han et al., 2022) are proposed based on e-commence image-text pairs, which greatly improve the performance of e-commerce image-text retrieval. Despite the success, the fine-grained cross-modal alignment issue remains unsolved, which may result in the inaccurate matching of details between images and texts. Although some e-commerce VLP models use fine-grained information from either image perspectives (Han et al., 2022) or patch-based image classification (Gao et al., 2020; Yu et al., 2022), they are short of semantic-level alignments across modalities. Some other work (Ma et al., 2022; Zhu et al., 2021) focuses on entities in text modalities, but rarely considers cross-modal interactions. In the general domain, fine-grained interactions could be achieved with object detection (Li et al., 2020c; Tan and Bansal, 2019), scene graph parsing (Cui et al., 2021), or semantic analysis (Yu et al., 2021; Li et al., 2020b). Unfortunately, these tools lose their effectiveness in the e-commerce domain.

To improve the fine-grained alignment between images and texts in e-commerce, this paper proposes an e-commerce knowledge-enhanced VLP model - **FashionKLIP**. Particularly, we first propose a data-driven strategy to construct a multi-modal conceptual knowledge graph in e-commerce (called **FashionMMKG**) from a large-scale e-commerce image-text corpus, where the fashion concepts are automatically extracted and organized in the form of a semantic hierarchy, each associated with its representative images. The FashionMMKG is later incorporated as the prior cross-modal fashion knowledge in training a CLIP-style model to support e-commerce image-text retrieval. For model training, we learn the representation alignment of image-text pairs across the two modalities by contrastive learning, and further optimize the alignment at the conceptual level. The conceptual alignment is further obtained by matching the text representations with the visual prototype representations of the fashion concepts in FashionMMKG.

Our contributions can be summarized as follows:

- We innovatively propose a data-driven approach to construct a multi-modal conceptual knowledge graph in the e-commerce domain named FashionMMKG without human intervention.
- We construct an e-commerce knowledge-enhanced VLP model called FashionKLIP, which learns conceptual-level alignments based on the prior knowledge in FashionMMKG.
- We conduct experiments on a popular fashion benchmark dataset FashionGen (Rostamzadeh et al., 2018) and show that FashionKLIP outperforms state-of-the-art VLP models in the e-commerce domain.
- We also apply the method to real industrial scenarios and observe significant improvements in image/text-to-product retrieval tasks.

2 Related Work

Vision-Language Pre-training. VLP models can be categorized into single-stream models (Chen et al., 2020; Li et al., 2020a; Gan et al., 2020), which first concatenate multi-modal inputs for interactions, and dual-stream models (Jia et al., 2021; Radford et al., 2021; Yao et al., 2021; Li et al., 2020b), which obtain the representations of the image and text respectively and learn the alignment afterwards. Although single-stream models may lead to high retrieval accuracy due to the early fusion of images and texts, the inference efficiency is sacrificed to a certain extend. Recently, to focus more on fine-grained semantic level interactions of images and texts, some works improve the similarity strategy by calculating between the image patch and the text token (Yao et al., 2021) or leverage fine-grained image information through object detectors (Li et al., 2020c,b; Gan et al., 2020; Zeng et al., 2021). Others introduces structured scene graphs for semantic knowledge (Yu et al., 2021). Despite their success in general domain, such methods are hard to be adopted to e-commerce data.

Fashion-based Retrieval. FashionBERT (Gao et al., 2020) first adopts pre-training tasks such as masking strategy to e-commerce images and texts. KaleidoBERT (Zhuge et al., 2021) extracts a series of multi-grained image patches for augmentation to guide masking strategy for fine-grained matching. CommerceMM (Yu et al., 2022) proposes pre-training tasks to align uni-modal with multi-modal features for more consistent alignment. EI-CLIP (Ma et al., 2022) defines the entity-aware retrieval task from the linguistic perspective by introducing a causal model to concatenate different meta-data as e-commerce entities. Lately,

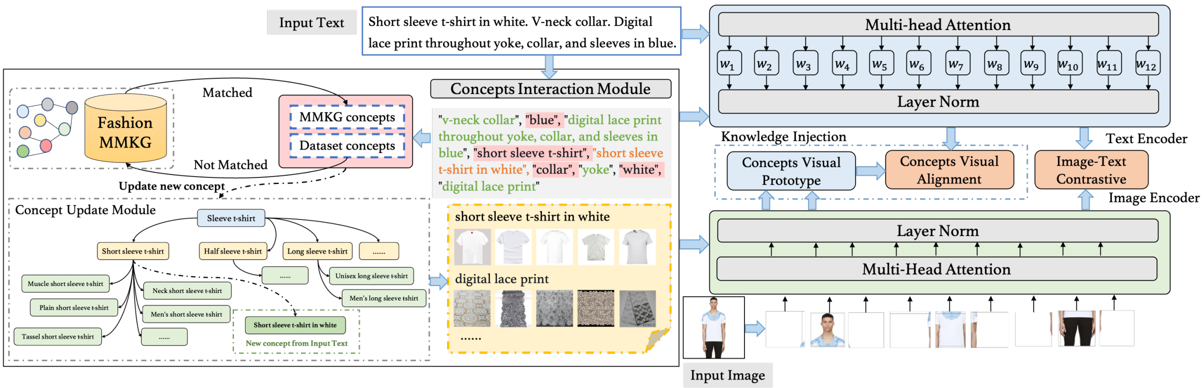

Figure 2: Model architecture of FashionKLIP with fashion images and texts as inputs.

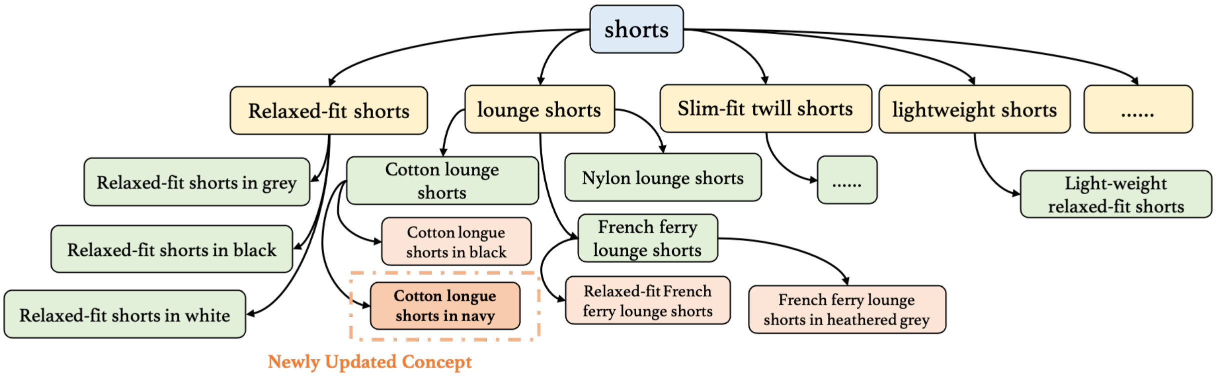

Figure 3: The sub-tree structure with root concept "shorts". The tree can be dynamically updated by inserting new concepts, such as "cotton lounge shorts in navy".

Fashion-ViL (Han et al., 2022) designs a flexible architecture for various downstream tasks. However, current methods still suffer from insufficient fine-grained semantic alignment, which may diminish the cross-modal understanding capability of models at semantic level.

3 Methodology

This section introduces how FashionMMKG is constructed and how FashionKLIP incorporates conceptual-level interactions of cross-modal fashion knowledge from FashionMMKG.

3.1 FashionMMKG Construction

Textual Modality. Instead of building an ontology-based knowledge graph (Deng et al., 2022), we automatically construct FashionMMKG to alleviate the gap with real-world user queries. The construction procedures include first determining the concept set through mining massive fashion texts and then matching each concept with its corresponding images. Given a fashion dataset $D\{T, I\}$ containing N image-text pairs, we first extract all the texts T. We use the NLP tool spacy[2] for sentence components analysis and part-of-speech tagging.We obtain multi-grained concept phrases by concatenating adjective modifiers with the key word. For an input text "Heathered cotton lounge shorts in navy. Elasticized waistband with drawstring closure", we extract root concepts such as "navy", "waistband", "closure" and "heathered", as well as more detailed phrases: "cotton lounge shorts", "cotton lounge shorts in navy", "heathered cotton lounge shorts in navy", etc. Based on different conceptual hierarchical granularities of extracted results, we build up hypernym-hyponym ("is-a") relationships between concepts in the form of relation triplets by judging whether two concepts are contained by each other, such as <*"cotton lounge shorts in navy", is-a, "cotton lounge shorts"*>.

After all the relation triplets are extracted, we organize these fashion concepts in a hierarchical structure. A sub-tree with the root node "shorts" is shown in Figure 3. The construction process of the hierarchical structure can be further implemented

[2] https://spacy.io/usage/linguistic-features

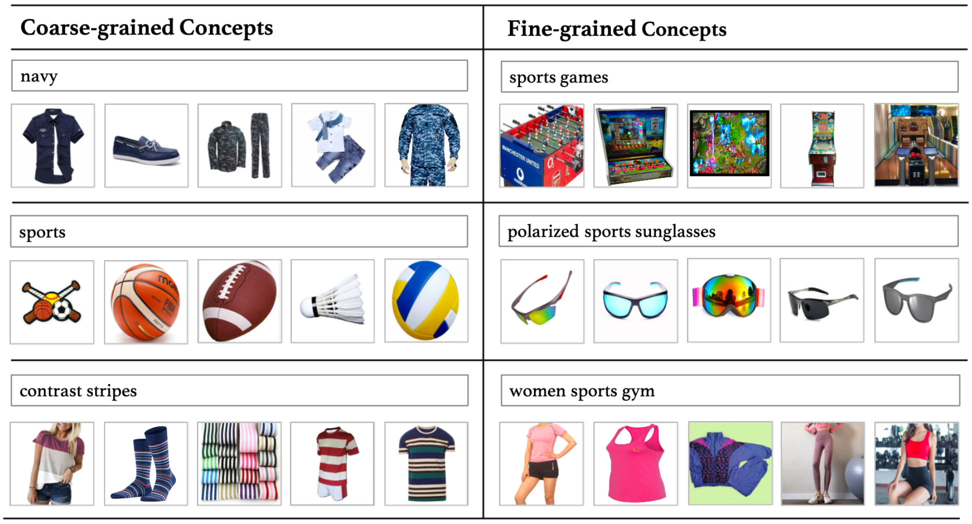

Figure 4: Coarse-grained and fine-grained concepts with their matched images from FashionMMKG.

in a dynamic process. When previously unseen concepts appear, we can add these new concepts into existing hierarchical trees, as the newly updated concept "short sleeve t-shirt in white" in Figure 2.

Visual Modality. For the visual modality, we adopt a prompt-based image retrieval method for each concept, and iteratively update the procedure in the subsequent visual-linguistic training process. Utilizing the generalization ability of a pre-trained CLIP-style model, we retrieve product images from the image set I, with the query formulated as *"A photo of* $\{concept\}$*"* as in (Radford et al., 2021; Yao et al., 2021; Gu et al., 2022). Based on the cosine distance of the image and text features, a naive approach is to select the top k images with the highest similarities as the concept visual prototype.

The retrieval results of some concepts are shown in Figure 4. We can see that the top k images of coarse-grained concepts are usually visually diverse, while images tend to be more semantically consistent when it comes to more specific concepts. To ensure that both similarity and diversity of visual representations for each concept are considered, we slightly expand the range of image candidates (using a larger k), and employ the MMR algorithm (Carbonell and Goldstein, 1998) to improve the diversity of the selected images. It runs in an iterative process until a sufficient number of images are selected from the k candidates. Denote C as the candidate image set and S as the collection of images that have been selected for concept c. Each time, we choose an image v_i by:

$$\begin{aligned} MMR(v_i) = & \underset{v_i \in C \setminus S}{argmax}[\lambda Sim(c, v_i) \\ & - (1-\lambda) \underset{v_j \in S}{max}\, Sim(v_i, v_j)] \end{aligned} \tag{1}$$

where $Sim(\cdot, \cdot)$ is the cosine similarity between the corresponding text/image features, and λ is the coefficient to adjust the relevance and diversity of results. Here, we set $\lambda = 0.8$ by default.

3.2 FashionKLIP Training

During the model training, as shown in Figure 2, we first extract concepts from the texts. If there are new concepts, FashionMMKG is automatically expanded. For parameter optimization, FashionKLIP consists of two tasks: image-text contrastive learning (ITC) for matching images and texts globally, and concept-visual alignment learning (CVA) for conceptual-level cross-modal alignment.

ITC. We train a CLIP-style model to learn the global representations of image-text pairs. For b image-text pairs in each training batch, denote L_k^I and L_k^T as the contrastive image-to-text and text-to-image matching loss, respectively. The ITC loss function can be expressed as $L_{ITC} = \frac{1}{2}\sum_{k=1}^{b}(L_k^I + L_k^T)$, with L_k^T to be defined as:

$$L_k^T(x_k^T, \{x_j^I\}_{j=1}^b) = -log\frac{exp(s_{k,k}^T)}{\sum_j exp(s_{j,k}^T)} \tag{2}$$

where the corresponding text of an image x_k^I is x_k^T, and $s_{j,k}^T$ is the cosine similarity between the image/text features of x_j^I and x_k^T. L_k^I is defined symmetrically to L_k^T.

CVA. We further align concepts and visual prototypes from the FashionMMKG. For an input text x_k^T with image x_k^I, we obtain a multi-grained concept set $Con(x_k^T)$, where hypernym concepts from the tree are also introduced to avoid paying much attention to fine-grained concepts but ignoring the cross-modal understanding of high-level concepts. For a concept $c_i \in Con(x_k^T)$, we denote $S(c_i)$ to be the collection of the selected similar yet diverse images to represent the visual characteristics of the concept (as described previously in Section 3.1). We select q images with the highest scores with image x_k^I in $S(c_i)$ for each $c_i \in Con(x_k^T)$, for the model to learn conceptual alignments. We compute the weighted contrastive loss between each c_i and any conceptual image $x_{\tilde{k}}^I \in S(c_i)$, together with conceptual images generated from other texts concepts within the same training batch:

$$L_k^{CT}(Con(x_k^T), \{S(x_j^T)\}_{j=1}^b) = -\frac{1}{q}\sum_{c_i}\sum_{x_{\tilde{j}}^I \in S(c_i)} w(x_{\tilde{k}}^I, x_k^I) log \frac{exp(s_{\tilde{k},k}^T)}{\sum_j exp(s_{\tilde{j},k}^T)} \tag{3}$$

Note that $w(x_{\tilde{k}}^I, x_k^I)$ is the cosine similarity between concept image $x_{\tilde{k}}^I$ and input image x_k^I, used as the weight for loss calculation. This forces the representation of a concept c_i similar to its conceptual images $S(c_i)$, but dis-similar to those of conceptual images from other texts. Similarly, by changing the loss function from text-to-image to image-to-text, we have the symmetric loss L_k^{CI}. Thus, the loss function of CVA is expressed as:

$$L_{CVA} = \frac{1}{2}\sum_{k=1}^{b}(L_k^{CI} + L_k^{CT}) \tag{4}$$

Overall Loss. The total loss function is formulated as: $L = \frac{1}{2}(L_{ITC} + L_{CVA})$. In addition, as the representations of images are continuously updated during model training, at the end of each epoch, we leverage Faiss (Johnson et al., 2019) to retrieve top-k images to update the visual prototype representations of the matched concepts.

4 Experiments

We conduct comprehensive evaluations on FashionGen (Rostamzadeh et al., 2018) to show that FashionKLIP outperforms SOTA methods.

4.1 Implementation Details

We first construct FashionMMKG with details shown in Appendix A.1.

Model Training. The specific settings of models are described in Appendix A.2. For training, we conduct both domain-specific pre-training and fine-tuning for base and large versions of FashionKLIP. We initialize FashionKLIP from CLIP pre-trained weights and continually pre-train the model based on our in-house dataset for MMKG construction (as described previously), only using the contrastive learning process over image-text pairs. Specially, the continual pre-training process is conducted with the parameters of the image encoder fixed. Overall, we have four models: FashionKLIP-S (small), FashionKLIP-M (medium), FashionKLIP-B (base) and FashionKLIP-L (large).

Benchmark Dataset. We use a widely-used benchmark dataset (i.e., FashionGen (Rostamzadeh et al., 2018)) for model evaluation. It contains 67,666 fashion items of 293,008 image-text pairs in 121 sub-categories, with 260,480 pairs for training and 32,528 for validation.

Evaluation. For image-text retrieval tasks, based on a text query, we consider two settings for evaluation. 1) Strictly following (Gao et al., 2020; Zhuge et al., 2021; Ma et al., 2022; Yu et al., 2022), the model is required to pick the matched image in 101 samples, including 1 ground-truth image with 100 randomly selected images within the same product sub-category (denoted as "Sample"). 2) As some recently published works (Ma et al., 2022) also consider large-scale candidates on the entire set, each query is compared with every item in the full dataset (denoted as "Full"). The settings for image-to-text matching are likewise. Recall@1/5/10 is regarded as evaluation metrics as previous works (Gao et al., 2020; Zhuge et al., 2021; Yu et al., 2022).

Method	Image-to-Text			Text-to-Image		
	R@1	R@5	R@10	R@1	R@5	R@10
FashionBERT	23.96	46.31	52.12	26.75	46.48	55.74
KaleidoBERT	28.00	60.10	68.40	33.90	60.50	68.60
CommerceMM	41.60	64.00	72.80	39.60	61.50	72.70
CLIP	36.11	67.81	80.00	35.32	65.98	77.84
EI-CLIP	38.70	72.20	84.25	40.06	71.99	82.90
FashionKLIP-B	**60.79**	**85.67**	**91.95**	**54.00**	**78.49**	**86.28**

Table 1: Retrieval results on FashionGen (Sample).

Model	Image-to-Text			Text-to-Image		
	R@1	R@5	R@10	R@1	R@5	R@10
CLIP	22.50	49.50	62.00	24.50	51.10	63.60
EI-CLIP	25.70	54.50	66.80	28.40	57.10	69.40
FashionKLIP-B	**37.01**	**59.78**	**67.39**	**43.70**	**63.74**	**72.67**

Table 2: Retrieval results on FashionGen (Full).

Model	Image-to-Text			Text-to-Image		
	R@1	R@5	R@10	R@1	R@5	R@10
FashionKLIP-S	14.58	34.28	44.14	17.59	36.74	47.20
FashionKLIP-M	23.21	45.45	54.98	28.42	49.95	59.74
FashionKLIP-B	37.01	59.78	67.39	43.70	63.74	72.67
FashionKLIP-L	47.16	69.27	75.39	54.60	75.06	81.39

Table 3: Retrieval results on FashionGen (Full) of FashionKLIP under different model sizes.

4.2 Experimental Results

Overall Retrieval Results. We conduct both “full” and “sample” evaluation of FashionKLIP-B against existing SOTA models. In addition, we report the results of different FashionKLIP models on FashionGen using the full evaluation criteria, as shown in Table 3. As the main experimental results shown in Table 1, we can see that FashionKLIP model significantly outperforms the existing SOTA models by a large margin. In particular, on the R@1 metric, FashionKLIP-B even greatly surpasses the methods with multi-modal fusion encoders for more unified representation learning such as CommerceMM (Yu et al., 2022). On full evaluation results in Table 2, FashionKLIP-B shows a remarkable increase of 11-15% compared to EI-CLIP (Ma et al., 2022). For smaller settings such as FashionKLIP-M, the retrieval performance is also competitive and closer to CLIP. As the “full” setting is closer to real-world retrieval scenarios and more challenging as it aims to select from a large candidate set, the performance of FashionKLIP is significant, further proving that the framework can be generalized to wider application scenarios. Based on the experimental results on either setting, we can conclude the effects brought by fashion knowledge, and confirm that more attention to cross-modal conceptual-level interactions leads to an increase in e-commerce image-text matching. [3]

Ablation Studies. To further analyze the importance of conceptual-level fashion image-text alignment, we present different variants of FashionKLIP in Table 4 for two evaluation settings. We can see from the results that both CVA and the FDP contribute to performance improvement. Although the retrieval results decrease slightly when not using FDP, the removal of CVA will harm the retrieval performance more heavily. Besides, the introduction of FDP and CVA at the same time boosts the performance as “Full Implement.” shows, proving the necessity to utilize fashion data for pre-training, which helps establish a better mapping between concepts and images as prior knowledge. More importantly, the focus on fashion knowledge better guides conceptual-level interactions and brings a rise to the alignment between images and texts.

[3]Note that a few works (e.g., Fashion-ViL (Han et al., 2022)) employ additional multi-modal fusion encoders and uniform representation learning (that may be not suitable for fast vector retrieval in real-world applications) and evaluate their models on randomly sampled subsets of FashionGen. Hence, their works are not directly comparable.

Method	Eval.	Image-to-Text			Text-to-Image		
		R@1	R@5	R@10	R@1	R@5	R@10
Full Implement.	Sample	**60.79**	**85.67**	**91.95**	**54.00**	**78.49**	**86.28**
w/o. CVA	Sample	56.70	84.53	91.65	51.43	77.44	85.36
w/o. FDP	Sample	58.90	84.87	91.35	52.57	77.14	84.87
Full Implement.	Full	**37.01**	**59.78**	**67.39**	**43.70**	**63.74**	**72.67**
w/o. CVA	Full	35.41	57.92	65.97	40.63	61.73	69.40
w/o. FDP	Full	36.10	58.32	66.07	42.05	61.66	69.65

Table 4: Ablation studies on FashionKLIP-B, where FDP represents fashion-domain pre-training.

5 Industrial Application

In this section, we verify the effectiveness of FashionKLIP on our Alibaba global e-commerce platform. Specifically, we apply it to product search with two specific retrieval tasks including image-to-product (I2P) and text-to-product (T2P) retrieval, as shown in Figure 5.

Model	Parameters	RT	QPS
CLIP	151M	61.26ms	16.32
FashionKLIP-B	151M	60.45ms	16.54
FashionKLIP-M	91M	42.69ms	23.43

Table 5: Average inference speed over 1,000 samples in terms of Response Rime (RT) and the Query Per Second (QPS) on a single GPU (NVIDIA V100).

For T2P, we employ a weighted scoring function to compute the similarity score between a query text and a product (with a title and an image) as follows: $Score_{t2p} = \alpha * Score_{t2t} + (1-\alpha) * Score_{t2i}$, where $0 < \alpha < 1$, $Score_{t2t}$ and $Score_{t2i}$ refer to the embedding similarity score between the query text and the product title, together with the query text and the product image. Similarly, for I2P, we have $Score_{i2p} = \alpha * Score_{i2t} + (1-\alpha) * Score_{i2i}$.

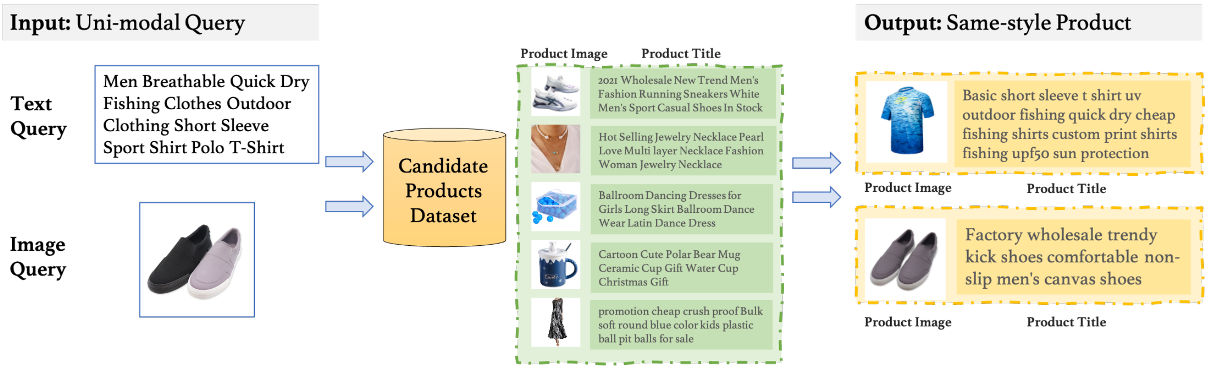

Figure 5: Example on image-to-product and text-to-product retrieval for e-commerce product search.

In total, the collected dataset contains 58,463 products (with images and titles) and 3,021 queries.

Model	Image-to-Product				Text-to-Product			
	R@1	R@5	R@10	R@20	R@1	R@5	R@10	R@20
CLIP	82.93	93.07	95.40	96.59	49.43	75.46	84.27	89.41
FashionKLIP-M	84.81	93.22	95.15	96.44	48.00	75.56	84.96	90.85
FashionKLIP-B	**87.48**	**95.94**	**97.97**	**98.91**	**52.10**	**79.96**	**89.02**	**93.77**

Table 6: Retrieval results on e-commerce image-to-product and text-to-product retrieval.

We conduct zero-shot experiments for T2P and I2P on FashionKLIP-B and FashionKLIP-M and compare it with the baseline CLIP (Radford et al., 2021), as shown in Table 6. For models of the same size, we can see that FashionKLIP-B greatly outperforms CLIP on Recall@1-20 and particularly achieves an improvement of 3~5% on both tasks for R@1. For our model in a smaller size, FashionKLIP-M is still comparable, which mainly reflects on the R@1 and R@5 results of I2P task and the R@5 to R@20 results of T2P. However, the inference of FashionKLIP-M is faster. In Table 5, taking text-to-product as an example, we report the Response Time (RT) and Query Per Second (QPS) using different text encoders to encode user queries on a single GPU (NVIDIA V100). We can see that with similar performance (CLIP and FashionKLIP-M), our model has much lower RT and higher QPS. Hence, we confirm FashionKLIP's feasibility on multi-modal tasks in the industrial applications.

6 Conclusion and Future Work

This paper proposes a novel data-driven approach to construct a multi-modal conceptual knowledge graph in e-commerce namely FashionMMKG. An e-commerce knowledge-enhanced VLP model namely FashionKLIP is then constructed by learning the conceptual-level alignments from the prior knowledge in FashionMMKG. Our empirical study shows that FashionKLIP outperforms state-of-the-art VLP models in the e-commerce domain. We conduct experiments under industrial scenarios and verify its practical value in real-world applications and confirm the efficiency of FashionKLIP. In the future, we will apply the knowledge-enhanced strategy for general large-scale pre-training and bring benefit to more multi-modal tasks.

7 Acknowledgement

This work is funded by National Key Research and Development Project (No.2020AAA0109302), National Natural Science Foundation of China (No.62072323), Shanghai Science and Technology Innovation Action Plan (No. 22511104700, 22511105902), Shanghai Municipal Science and Technology Major Project (No.2021SHZDZX0103), Science and Technology Commission of Shanghai Municipality Grant (No. 22511105902) and supported by Alibaba Group through Alibaba Innovative Research Program.

References

Jaime Carbonell and Jade Goldstein. 1998. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 335–336.

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

Yuhao Cui, Zhou Yu, Chunqi Wang, Zhongzhou Zhao, Ji Zhang, Meng Wang, and Jun Yu. 2021. Rosita: Enhancing vision-and-language semantic alignments via cross-and intra-modal knowledge integration. In

Proceedings of the 29th ACM International Conference on Multimedia, pages 797–806.

Shumin Deng, Hui Chen, Zhoubo Li, Feiyu Xiong, Qiang Chen, Mosha Chen, Xiangwen Liu, Jiaoyan Chen, Jeff Z Pan, Huajun Chen, et al. 2022. Construction and applications of open business knowledge graph. *arXiv preprint arXiv:2209.15214*.

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

Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2017. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612*.

Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. 2020. Large-scale adversarial training for vision-and-language representation learning. *Advances in Neural Information Processing Systems*, 33:6616–6628.

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

Jiaxi Gu, Xiaojun Meng, Guansong Lu, Lu Hou, Minzhe Niu, Hang Xu, Xiaodan Liang, Wei Zhang, Xin Jiang, and Chunjing Xu. 2022. Wukong: 100 million large-scale chinese cross-modal pre-training dataset and a foundation framework. *arXiv preprint arXiv:2202.06767*.

Xiao Han, Licheng Yu, Xiatian Zhu, Li Zhang, Yi-Zhe Song, and Tao Xiang. 2022. Fashionvil: Fashion-focused vision-and-language representation learning. *arXiv preprint arXiv:2207.08150*.

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

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137.

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

Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. 2020b. Unimo: Towards unified-modal understanding and generation via cross-modal contrastive learning. *arXiv preprint arXiv:2012.15409*.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020c. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Haoyu Ma, Handong Zhao, Zhe Lin, Ajinkya Kale, Zhangyang Wang, Tong Yu, Jiuxiang Gu, Sunav Choudhary, and Xiaohui Xie. 2022. Ei-clip: Entity-aware interventional contrastive learning for e-commerce cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18051–18061.

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

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

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

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565.

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

Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Well-read students learn better: On the importance of pre-training compact models. *arXiv preprint arXiv:1908.08962*.

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

Chengyu Wang, Minghui Qiu, Taolin Zhang, Tingting Liu, Lei Li, Jianing Wang, Ming Wang, Jun Huang, and Wei Lin. 2022. Easynlp: A comprehensive and easy-to-use toolkit for natural language processing.

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

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2021. Ernie-vil: Knowledge enhanced vision-language representations through scene graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 3208–3216.

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

Yan Zeng, Xinsong Zhang, and Hang Li. 2021. Multi-grained vision language pre-training: Aligning texts with visual concepts. *arXiv preprint arXiv:2111.08276*.

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

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

A Appendix

A.1 FashionMMKG

Full statistics of our FashionMMKG are shown in Table 8, where we give both the total numbers (cnt) of items such as the number image-text pairs and concepts, and the average of some attributes (avg) such as occurrence and concept length. As for the data source, we extract fashion concepts from titles of 900,000 product image-text pairs collected from our global e-commerce platform. [4]

A.2 Model Settings

We release models with various parameter sizes for industrial applications. The specific hyperparameters of different FashionKLIP models are shown in Table 7.

Image Encoder We follow Vision Transformer (ViT) (Dosovitskiy et al., 2020) closely as the image encoder and the modifications of different models lie in the number of layer normalization and the width of attention heads. The size of non-overlapping image patches are also set to be different. FashionKLIP-L adopts the ViT-L/14 as the image encoder with 24 layers, while FashionKLIP-M uses ViT with 12-layer 512 wide in 88M parameter and the patch size is 32.

Text Encoder We adopts a Transformer (Vaswani et al., 2017), utilizing the same architecture as described in (Radford et al., 2019) as the text encoder. For models in different sizes, we refer to (Turc et al., 2019) to set the attention width and number of attention heads of the text encoder.

Model Input Images are cropped uniformly to 224×224 pixels before entering the model. We limit the maximum input length of the text to 77, with a vocabulary of 49,408.

For a fair comparision, we utilize FashionKLIP-B model to compare against other baseline models, which uses ViT-B/32 (Dosovitskiy et al., 2020) as the image encoder, and adopts a 12-layer 512 wide Text Transformer as the text encoder as (Radford et al., 2021), in 63M parameter with 8 attention heads each layer.

A.3 Model Training

The batch size of pre-training is 1,024 per GPU with 8 A100 GPUs (80G), for 20 epoches in total. The learning rate is 5e-5. During dataset-specific model fine-tuning, we retrieve top-20 images for each concept in FashionMMKG and then select 5 images as the visual prototype based on the proposed criteria. The batch size of fine-tuning is 32 per GPU, with a learning rate of 1e-5 on two A100 GPUs. As smaller pre-trained CLIP weights

[4] https://www.alibaba.com/

Model	Embedding dimension	Input resolution	Vision Transformer				Text Transformer			
			parameters	layers	width	patch size	parameters	layers	width	heads
FashionKLIP-L	768	224	303M	24	1024	14	124M	12	768	12
FashionKLIP-B	512	224	88M	12	768	32	63M	12	512	8
FashionKLIP-M	512	224	40M	12	512	32	51M	8	512	8
FashionKLIP-S	384	224	22M	12	384	16	33M	8	384	6

Table 7: Hyperparemters of FashionKLIP in different model settings.

are not available, we initialize FashionKLIP-M and FashionKLIP-S models from the pre-trained FashionKLIP-B model by truncating the weights of FashionKLIP-B to the size based on the settings of smaller models. After that, we utilize the contrastive learning process for continually pre-training on the e-commerce in-house data. The batch size during pre-training for FashionKLIP-M and FashionKLIP-S is 256 per GPU on 8 GPUs and the learning rate is 5e-5.

Item Name	Statistics
Image-text pairs (cnt)	900,000
Root-concepts (cnt)	5,135
All concepts (cnt)	99,076
Nodes per tree (avg)	213.8 (1∼25600)
Concept length (avg)	3.4 (1∼21)
Occurrence (avg)	17.1 (1∼77250)
Images per concept (avg)	20
All images (cnt)	76,964

Table 8: Statistics of FashionMMKG.