

# Query-LIFE: Query-aware Language Image Fusion Embedding for E-Commerce Relevance

Hai Zhu, Yuankai Guo, Ronggang Dou, Kai Liu\*, Xiaoyi Zeng

Alibaba International Digital Commerce

{zhuhai, guoyuankai.gyk, ronggang.drg, baiyang.lk}@alibaba-inc.com  
yuanhan@taobao.com

## Abstract

Relevance module plays a fundamental role in e-commerce search as they are responsible for selecting relevant products from thousands of items based on user queries, thereby enhancing users experience and efficiency. The traditional method calculates the relevance score based on product titles and user queries, but the information in title alone maybe insufficient to describe the product completely. A more general method is to further leverage product image information. In recent years, vision-language pre-training model has achieved impressive results in many scenarios, which leverage contrastive learning to map both textual and visual features into a joint embedding space. In e-commerce, a common practice is to further fine-tune the model using e-commerce data on the basis of pre-trained model. However, the performance is sub-optimal because the vision-language pre-training models lack of alignment specifically designed for queries. In this paper, we propose **Query-aware Language Image Fusion Embedding** to address these challenges (**Query-LIFE**). It utilizes a query-based multimodal fusion to effectively incorporate the image and title based on the product types. Additionally, it employs query-aware modal alignment to enhance the accuracy of the comprehensive representation of products. Furthermore, we design GenFilt, which utilizes the generation capability of large models to filter out false negative samples and further improve the overall performance of the contrastive learning task in the model. Experiments have demonstrated that Query-LIFE outperforms existing baselines. We have conducted ablation studies and human evaluations to validate the effectiveness of each module within Query-LIFE. Moreover, Query-LIFE has been deployed on Miravia Search<sup>1</sup>

\*Corresponding author.

<sup>1</sup>Miravia is a local-to-local e-commerce platform in Spain incubated by Lazada, as one part of Alibaba International Digital Commerce (AIDC) Group. <https://www.miravia.es/>

## 1 Introduction

With the increasing spread of the internet, online shopping has become a convenient option for consumers. Millions of users browse and search for products on e-commerce platforms every day. Consequently, the relevance of the products displayed to users based on their search queries plays a crucial role in the user’s shopping experience and also in the efficiency of the transaction. Therefore, it is crucial for an e-commerce search engine to accurately assess whether the products offered are relevant to the user’s intentions.

Traditional relevance models (Robertson et al., 2009; Huang et al., 2013; Chang et al., 2021; Hu et al., 2014; Yao et al., 2021) have primarily relied on textual information, including user queries and product descriptions (titles, attributes, etc.), to assess relevance between queries and products. However, product information also includes images, which capture a large part of the user’s attention when browsing products. It is therefore becoming increasingly important to include images in relevance modeling. This integration of image and text information has the potential to provide a more comprehensive representation of products and better capture the user’s intent.

In some cases, core information may be omitted from product titles, as shown in Table 1. In such cases, it is difficult to rely on product titles alone to match relevant products with user queries. However, product images can provide additional and valuable information for assessing relevance. Recently, many visual language pre-training (VLP) (Li et al., 2023, 2021; Jia et al., 2021; Wang et al., 2023, 2021) models have been proposed. As shown in Figure 1(a), these VLP models usually consist of both textual and visual encoders and utilize contrastive learning between speech and vision to align representations across different modalities. They have shown impressive performance on vari-

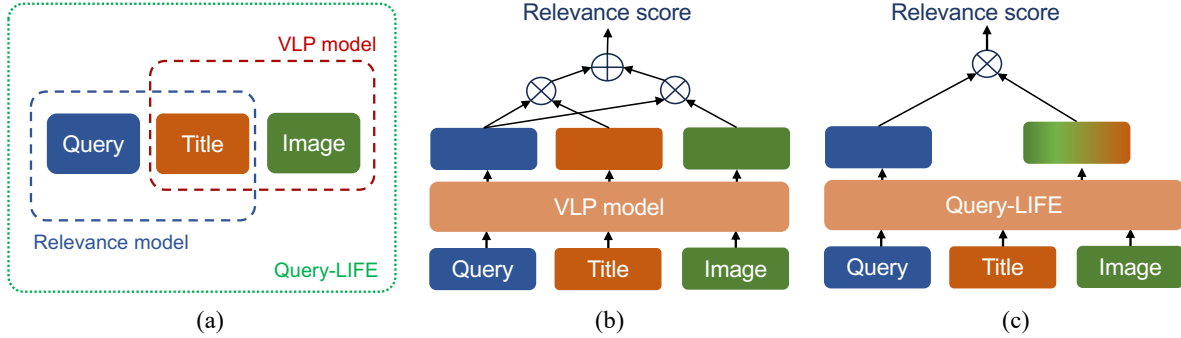


Figure 1: (a) The relationship of relevance model, VLP model and Query-LIFE. (b) VLP model’s divide-and-conquer approach for relevance task. (c) Query-LIFE’s fusion approach for relevance task.




Query	Image	Title
men’s winter coat		Koroshi Jacket in two colors, water-repellent, with hood, for Men
air-conditioning		Split 1x1 MUNDO-CLIMA MUPR12 H11 3027frig R32
golden necklace		Elegant necklace with col-layered pearl gent

Table 1: Both product images and titles can be incorporated together to judge the search relevance with queries.

ous general tasks such as image captioning, visual question answering and text-image retrieval.

In the e-commerce relevance task, these VLP models can extract image features to improve the representation of products with ambiguous titles or correct the representation of products with misleading titles. As shown in Figure 1(b), they encode query, title and image representation separately, and then compute the inner product of query-image and query-title, and then add their inner product as relevance value. However, different product types contain differently weighted information in images and titles, so simple averaging for each modality is not optimal. For example, electronic products often list important parameters in the title, while clothing items tend to feature visual elements such as design, texture, material and color in the images.

In this paper, we propose a general approach called **Query-aware Language Image Fusion Embedding** for relevance modeling in e-commerce (**Query-LIFE**). As shown in Figure 1(a), query, title and image are integrated into the relevance

task. First, we draw random triple data from the logs of online user behavior as training data. Second, as shown in Figure 1(c), in contrast to the divide-and-conquer approach, we use the fusion vector of image and text as the multimodal representation of the product, and then adopt the inner product of query and multimodal representation as the relevance score. Finally, we use supervised contrastive learning to train the model, and utilize the generation ability of both the multimodal large model and the large language model to filter out the false negative samples.

## 2 Related Work

### 2.1 Vision-Language Pre-training

The advent of pre-training models such as BERT (Kenton and Toutanova, 2019), GPT3 (Brown et al., 2020) and ViT (Dosovitskiy et al., 2021) has led to significant advances in NLP and CV tasks, with state-of-the-art results. More recently, researchers have extended the pre-training approach to the vision-language (VL) domain, leading to the development of several impressive VL models (e.g., CLIP (Radford et al., 2021a) and ALIGN (Jia et al., 2021)). These VLP models have shown impressive performance on various multimodal downstream tasks such as image captioning, visual question answering and multimodal retrieval. They achieve this by utilizing large image-text pairs and then employing contrastive learning to align images and text in the joint embedding space. These VLP models are divided into two categories: Object Detector (OD)-based VLP models (e.g., UNITER (Chen et al., 2020), OSCAR (Li et al., 2020)) and end-to-end VLP models (e.g., ALBEF (Li et al., 2021), BLIP (Li et al., 2023)). OD-based VLP

models rely on bounding box annotations during pre-training and require high-resolution images for inference, making them both annotation-intensive and computationally expensive. In contrast, end-to-end VLP models directly use the features of image patches as input to a pre-trained ViT model. This eliminates the need for costly annotations and significantly improves the speed of inference. As a result, end-to-end VLP models have gained prominence in recent research (Chen et al., 2021; Kim et al., 2021). This is why we also use the end-to-end VLP model in this paper.

## 2.2 E-commerce VLP Model

There are also some VLP models that are specifically geared towards e-commerce scenarios. FashionBERT (Gao et al., 2020) was the first vision-language pre-training model that utilizes mask language loss and contrastive learning of cover images. Later, Kaleido-BERT (Zhuge et al., 2021) adopted multiple self-supervised tasks at different scales to focus more on the coherence between title and image. EI-CLIP (Ma et al., 2022) proposed an intervention-based framework for contrastive learning with entities. KG-FLIP (Jia et al., 2023) proposes a knowledge-guided fashion-domain language-image pre-training framework and utilizes external knowledge to improve the efficiency of pre-training.

## 3 Method

### 3.1 Model Architecture

In this section, we will present our model architecture in detail. As shown in Figure 2, the entire model training is divided into an internal alignment and an external alignment. The internal alignment is used to match the features of product titles and images. The external alignment is used to match the relevance between user queries and products. The model architecture consists of three modules: an image pre-processing backbone, a universal modal encoder and GenFilt. The image preprocessing backbone is the Visual Transformer (ViT) (Dosovitskiy et al., 2021), which divides the image into patches and encodes them as a sequence of embeddings with an additional  $[CLS]$  token to represent the global image features. The universal modal encoder is shared weight and includes self-attention layer, cross-attention layer and feed-forward layer. GenFilt is designed to filter out false-negative samples during in-batch sampling.

### 3.2 Vision-Language Pre-training

The VLP model uses Image-Text Contrastive (ITC) loss to match image features and text features, resulting in positive image-text pairs having similar representations and reducing the similarity between negative pairs (Radford et al., 2021b). ITC loss has been shown to be an effective target for improving image and speech representation, even in the absence of labeled data. The formula is as follows:

$$\mathcal{L}_{ITC} = -\frac{1}{N} \sum_{i=1}^N \log \frac{\exp(Z_{T_i} \cdot Z_{I_i} / \tau)}{\sum_{j=1}^N \exp(Z_{T_j} \cdot Z_{I_j} / \tau)}. \quad (1)$$

where  $Z_T$  and  $Z_I$  are normalized text and image embeddings,  $Z_{I_i}$  is the  $i$ -th positive image sample in the batch.  $N$  and  $\tau$  are batch size and temperature parameter respectively.

### 3.3 Query-based Modal Alignment

In e-commerce search scenarios, the relevance of products depends heavily on user queries. However, users' search queries are short and concise. Calculating relevance based on query and title alone can easily lead to relevance score errors. To mitigate the impact of the above problem on the relevance score, we introduce image information to improve product representation. We also introduce the concept of title-image fusion representation for products (referred to as multimodal representation or  $\mathcal{M}$  representation). the  $\mathcal{M}$  representation is defined as the interaction between the product title and the image. In contrast to the divide-and-conquer approach, we use the inner product to compute the relevance between the  $\mathcal{M}$  representation and the query. To further match the  $\mathcal{M}$  representation with the user queries, we use the query-multi contrastive (QMC) loss. Additionally, we use the query-title contrastive loss (QTC) to match the query with the title. At the same time, the query-image contrast loss (QIC) is used to further align the query with the images. These loss functions play a crucial role in matching user queries and different product modalities and improve the relevance score.

In the e-commerce relevance task, the same query often generates positive pairs with different products. In addition, there are many labeled negative examples in the dataset. Therefore, supervised contrastive learning is introduced, which is more suitable for the relevance task. The formula

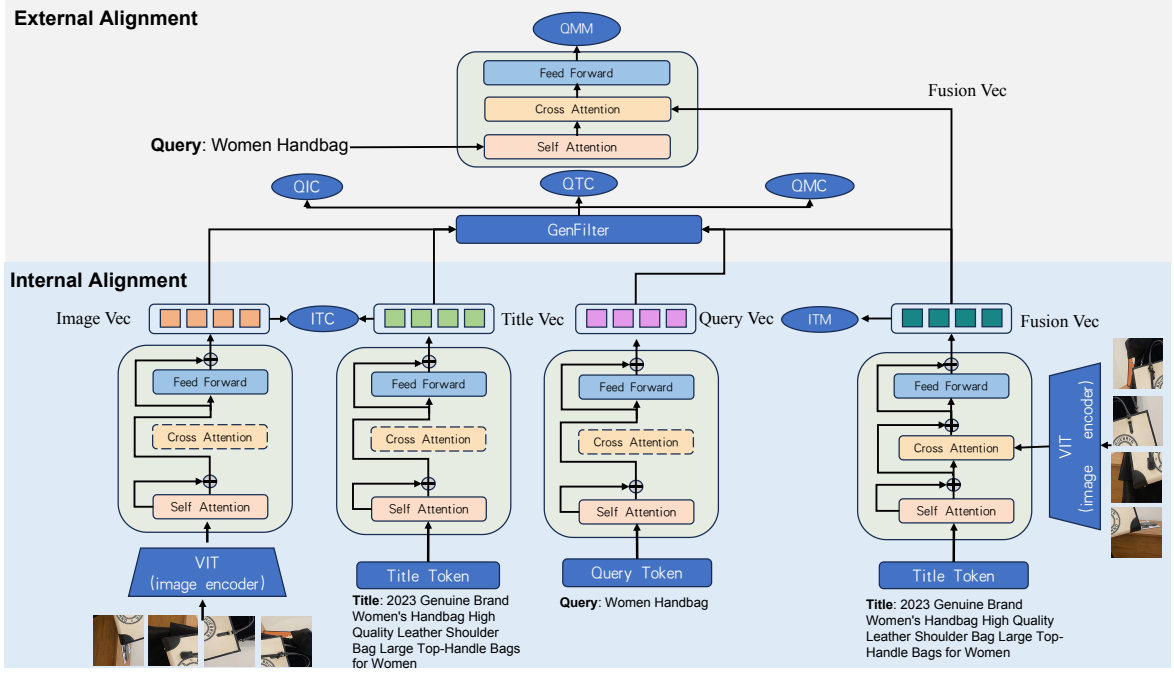


Figure 2: Overview of Query-LIFE. The overall training process is divided into internal alignment and external alignment. The model architecture consists of three modules: an image preprocessing backbone, a universal modal encoder (light-color block) and GenFilt.

is defined as follows:

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^N \left\{ \frac{1}{|P(i)|} \sum_{p \in P(i)} \left[ \log \frac{\exp(Q_i \cdot Z_p^x / \tau)}{\sum_{j=1}^N \exp(Q_i \cdot Z_j^x / \tau)} \right] \right\}. \quad (2)$$

where  $P(i)$  is all positive samples in the  $i$  batch,  $Q$  is normalized query embedding,  $Z_p^x, x \in [I, T, M], p \in P(i)$ , are normalized image/text/multi modal embedding in positive samples.  $\tau$  and  $N$  are temperature parameter and batch size. Bringing in different modal by  $x \in [I, T, M]$ , this loss function can represent QIC, QTC and QMC loss respectively.

### 3.4 Query-based Modal Fusion

We use image-text matching (ITM) to learn the  $\mathcal{M}$  representation of the product. The goal of ITM is to learn an image-title fusion that captures the matching between the image and text modalities. In ITM, we view the task as a binary classification problem where the model predicts whether an image-text pair is positive or negative. We use a hard negative mining strategy (Jia et al., 2021). In the hard negative mining strategy, negative pairs with higher similarity are selected within a group. The ITM loss can be expressed as follows:

$$\mathcal{L}_{ITM} = -E_{(I,T) \sim P} [\log\{P(y_{(I,T)})|(I,T)\}] \quad (3)$$

where  $P$  is a distribution of in-batch samples,  $y_{(I,T)} \in (0, 1)$  represents whether the image  $I$  and the text  $T$  are matched, and  $P(y_{(I,T)}|(I, T))$  is the output of the multimodal embedding followed by a two-class linear classifier.

We are aware that an image-text comparison alone may not be sufficient, as different product types contain different amounts of information in their images and titles. For example, electronic products often list important parameters in the title, while garments tend to have visual attributes such as material, color and size in the images. To enable the model to learn a more effective fusion representation, we introduce Query- $\mathcal{M}$  matching (QMM). For the cross-attention layer in external matching, the inputs of  $Q$  are the user’s query, the inputs of  $KV$  are the  $\mathcal{M}$  representation. In this way, the model can generate fused representations with a query-oriented alignment. QMM not only allows the model to extract features from both the images and the titles, but also to assign different weights to each modality based on the user queries. QMM and ITM have the same loss function listed in equation 3.

### 3.5 GenFilt

Most VLP models use in-batch sampling to generate negative image-title pairs. However, in

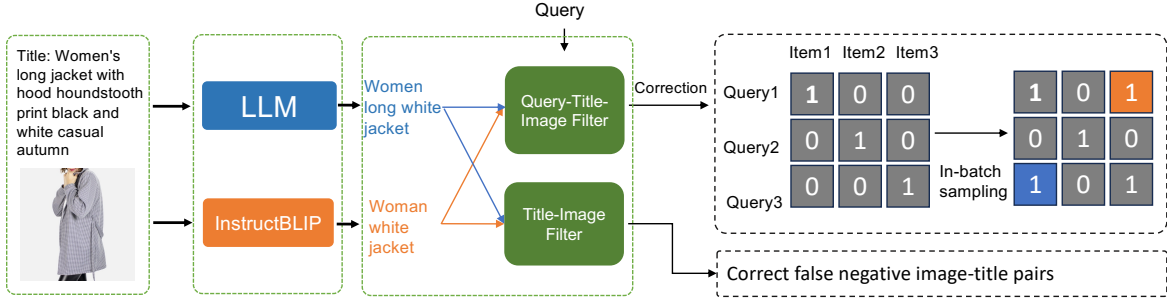


Figure 3: Overview of GenFit. GenFit adopts LLM and InstructBLIP to extract brief text description. Then compare the similarity of query-product pairs and correct the false negative query-product pairs. In addition, GenFit can also calculate the similarity of image-title pairs and correct false negative title-image pairs.

the triplet data  $\langle \text{query}, \text{title}, \text{image} \rangle$ , multiple user queries may be relevant for the different products. In-batch sampling leads to false negative samples. These similar or even identical search queries are incorrectly treated as negative examples and thus affect the relevance score.

Inspired by CapFilt (Li et al., 2023), we propose a method called Generating and Filtering (GenFit) to address the impact of false negative sampling on the training process. It improves the quality of the training data by enabling extensive model generation. As shown in Figure 3, GenFit consists of two modules. The first module is generation. We use a large language model (LLM) and a multimodal model (InstructBLIP) (Dai et al., 2023) to extract important text features from the product title and image, respectively. The second module is filtering. We calculate the similarity between image feature and text feature (I-T), the similarity between query feature and image feature (Q-I) and the similarity between query feature and text feature (Q-T). Finally, we set a threshold  $\sigma$  based on these similarities, and the similarity of query-product pairs (Q-I or Q-T) and image-text pairs (I-T) that are above the threshold are also corrected as positive patterns.

## 4 Experiments

### 4.1 Baselines and Datasets

**Large-scale Industrial Datasets.** We selected 1.3 million  $\langle \text{query}, \text{title}, \text{image} \rangle$  pairs from Miravia Search’s online click log. In addition, 200,000 labeled data are selected as an evaluation set, with a 1:1 ratio of positive to negative data.

**Baselines.** In our experiments, we compare Query-LIFE with several baselines, includ-

ing BERT (Kenton and Toutanova, 2019), ALBEF (Li et al., 2021), CLIP (Radford et al., 2021a), BLIP2 (Li et al., 2023) and CommerceMM (Yu et al., 2022). We used 16 A10 16G GPUs for training.

### 4.2 Evaluation Metrics

**Offline Evaluation Metrics.** Area Under Curve (AUC) and Recall@K (R@K) are used as metrics. We calculate the similarity between the query  $\rightarrow$  title, query  $\rightarrow$  image, and query  $\rightarrow \mathcal{M}$  and sort the set of candidates based on these similarities. Recall@K measures the percentage of matches that appear in the list with the highest K-rank (Gao et al., 2020).

**Online Evaluation Metrics.** We use the number of orders (Order\_cnt), the average number of buyers (Order\_uv) and the GMV (Gross Merchandise Volume) as online evaluation metrics. These metrics reflect the changes in user orders.

**Human Evaluation.** We performed a sample of 1,000 queries and selected the top 10 query-item pairs of the exposure page for each query to perform a human relevance score. The relevance of a query item can be categorised into three types: excellent, fair, and bad.

### 4.3 Offline Experiments

The previous models calculate the cosine similarity between query and title (query  $\rightarrow$  title) or image (query  $\rightarrow$  image). In Query-LIFE, we introduce another method that calculates the cosine similarity between the embedding of the query and the multimodal (query  $\rightarrow \mathcal{M}$ ). As shown in the Table 2, The AUC for the query  $\rightarrow \mathcal{M}$  proposed by Query-LIFE is higher than that of the baselines. It

	Model Training Para	Query $\rightarrow$ Title				Query $\rightarrow$ Image				Query $\rightarrow \mathcal{M}$			
		R@5	R@10	R@20	AUC	R@5	R@10	R@20	AUC	R@5	R@10	R@20	AUC
BERT	110M	<b>0.142</b>	0.186	0.340	0.865	-	-	-	-	-	-	-	-
ALBEF	233M	0.060	0.124	0.223	0.652	0.054	0.116	0.212	0.706	-	-	-	-
CLIP	151M	0.068	0.125	0.272	0.542	0.068	0.147	0.272	0.554	-	-	-	-
BLIP2	188M	0.113	0.170	0.272	0.752	0.056	0.159	0.316	0.771	-	-	-	-
CommerceMM	270M	0.093	0.153	0.312	0.671	<b>0.094</b>	0.179	0.302	0.668	-	-	-	-
Query-LIFE	188M	0.125	<b>0.215</b>	<b>0.351</b>	<b>0.871</b>	0.079	<b>0.204</b>	<b>0.329</b>	<b>0.871</b>	<b>0.113</b>	<b>0.215</b>	<b>0.386</b>	<b>0.891</b>
Query-LIFE w/o QMA	188M	0.068	0.170	0.318	0.741	0.079	0.147	0.306	0.805	0.068	0.193	0.329	0.784
Query-LIFE w/o QMF	188M	0.136	0.207	0.318	0.856	0.090	0.147	0.306	0.863	0.110	0.193	0.295	0.877
Query-LIFE w/o QMM	188M	0.124	0.211	0.335	0.856	0.079	0.201	0.306	0.866	0.079	0.205	0.314	0.879
Query-LIFE w/o ITM	188M	0.128	0.211	0.323	0.861	0.081	0.162	0.311	0.869	0.108	0.212	0.336	0.881
Query-LIFE w/o GenFilt	188M	0.102	0.147	0.261	0.816	0.056	0.147	0.321	0.835	0.090	0.136	0.295	0.849
Query-LIFE on short query	188M	0.031	0.081	0.167	0.855	0.023	0.092	0.142	0.858	0.023	0.092	0.156	0.887
Query-LIFE on long query	188M	0.228	0.357	0.592	0.871	0.121	0.313	0.576	0.886	0.174	0.366	0.622	0.902

Table 2: Offline results compared with different baselines.

Model	R@5	R@10	R@20	AUC	
$\frac{Q \rightarrow T + Q \rightarrow I}{2}$	0.090	0.181	0.329	0.781	
Query-LIFE	0.079	<b>0.215</b>	0.318	0.882	
Query $\rightarrow \mathcal{M}$	Query-LIFE	<b>0.102</b>	<b>0.215</b>	<b>0.386</b>	<b>0.891</b>

Table 3: The R@K and AUC of divide-and-conquer approach and query  $\rightarrow \mathcal{M}$ .

can be seen that the relevance score is effectively improved by introducing image information and external alignment of query-product.

At R@10 and R@20, the query  $\rightarrow \mathcal{M}$  of Query-LIFE is also better than the baselines. Furthermore, we compare the performance of Query-LIFE and the divide-and-conquer approach. As shown in Table 3, Query  $\rightarrow \mathcal{M}$  outperforms the divide-and-conquer approach in all metrics. This clearly shows the advantage of query  $\rightarrow \mathcal{M}$ .

Finally, we tested the performance of the model on long queries (length > 4) and short queries (length < 2) separately. AUC and R@K for different query lengths are listed in Table 2. Long queries contain more information, so that both AUC and R@K are significantly higher than for short queries. In addition, the Query  $\rightarrow \mathcal{M}$  task is still better than the Query  $\rightarrow$  title and Query  $\rightarrow$  image tasks, which further emphasises the robustness of our model. Additionally, we list the t-test in the Appendix.

#### 4.4 Online Experiment

Furthermore, we carry out online A/B experiments for one month. As shown in Table 5, all the efficiency metrics are increased. The results verified that Query-LIFE can attracts higher conversions for our platform. Query-LIFE has been deployed online and brings stable conversion improvements for Miravia Search. in addition, annotators are invited to evaluate whether the relevance is improved by the Query-LIFE. The results are shown in Table 4. Compared to the baseline, the main improve-

ment is that the score for "Excellent" increased by 4.42% and the score for "Poor" decreased by 2.79%. Further ablation experiments are listed in the Appendix.

	Excellent	Fair	Bad
Query-LIFE	+4.42%	+2.17%	-2.79%

Table 4: Results of human evaluation.

	Order_cnt	Order_uv	GMV
Query-LIFE	+4.11%	+3.06%	+3.19%

Table 5: Online A/B tests of Query-LIFE.

## 5 Conclusion

In this paper, we propose a novel approach for learning the multimodal representation of products in e-commerce search relevance. We design a query-based multimodal fusion module that effectively generates dynamic fusion representations that incorporate product image and text based on product types. We propose a query-based modal matching module that utilizes supervised contrastive learning to match the multimodal representation of products based on the search query. In addition, we propose the GenFilt module that utilizes the LLM (Large Language Model) and the ability to generate information from image and text to solve the false negative sampling problem in contrastive learning. The experimental results show that Query-LIFE performs better than the existing baseline solutions in both relevance tasks. In addition, Query-LIFE was successfully used in Miravia search, leading to improvements in both search relevance and conversion rate.

## References

- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- Wei-Cheng Chang, Daniel Jiang, Hsiang-Fu Yu, Choon Hui Teo, Jiong Zhang, Kai Zhong, Kedarnath Kolluri, Qie Hu, Nikhil Shandilya, Vyacheslav Ievgrafov, et al. 2021. Extreme multi-label learning for semantic matching in product search. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 2643–2651.
- Kezhen Chen, Qiuyuan Huang, Yonatan Bisk, Daniel McDuff, and Jianfeng Gao. 2021. Kb-vlp: Knowledge based vision and language pretraining. In *ICML workshop*.
- Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholi, 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.
- Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. [Instructblip: Towards general-purpose vision-language models with instruction tuning](#).
- Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*.
- 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.
- Baotian Hu, Zhengdong Lu, Hang Li, and Qingcai Chen. 2014. Convolutional neural network architectures for matching natural language sentences. *Advances in neural information processing systems*, 27.
- Po-Sen Huang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Acero, and Larry Heck. 2013. Learning deep structured semantic models for web search using clickthrough data. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, pages 2333–2338.
- 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.
- Qinjin Jia, Yang Liu, Daoping Wu, Shaoyuan Xu, Huidong Liu, Jinmiao Fu, Roland Vollgraf, and Bryan Wang. 2023. Kg-flip: Knowledge-guided fashion-domain language-image pre-training for e-commerce. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 81–88.
- Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.
- Wonjae Kim, Bokyoung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.
- Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705.
- Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020. Oscar: Object-semantic aligned pre-training for vision-language tasks. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pages 121–137. 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. 2021a. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021b. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.
- Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

- Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. 2023. Image as a foreign language: Beit pretraining for vision and vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19175–19186.
- Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. 2021. Simvlm: Simple visual language model pretraining with weak supervision. In *International Conference on Learning Representations*.
- Shaowei Yao, Jiwei Tan, Xi Chen, Keping Yang, Rong Xiao, Hongbo Deng, and Xiaojun Wan. 2021. Learning a product relevance model from click-through data in e-commerce. In *Proceedings of the Web Conference 2021*, pages 2890–2899.
- Licheng Yu, Jun Chen, Animesh Sinha, Mengjiao Wang, Yu Chen, Tamara L Berg, and Ning Zhang. 2022. Commercem: 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.
- 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.