De-noised Vision-language Fusion Guided by Visual Cues for E-commerce Product Search

Zhizhang Hu*
University of California, Merced
zhu42@ucmerced.edu

Shasha Li
Amazon Visual Search & AR
shashli@amazon.com

Ming Du
Amazon Visual Search & AR
mingdu@amazon.com

Arnab Dhua
Amazon Visual Search & AR
aduha@amazon.com

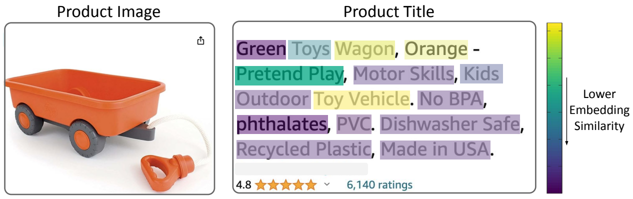
Douglas Gray
Amazon Visual Search & AR
douggray@amazon.com

Abstract

In e-commerce applications, vision-language multimodal transformer models play a pivotal role in product search. The key to successfully training a multimodal model lies in the alignment quality of image-text pairs in the dataset. However, the data in practice is often automatically collected with minimal manual intervention. Hence the alignment of image-text pairs is far from ideal. In e-commerce, this misalignment can stem from noisy and redundant non-visual-descriptive text attributes in the product description. To address this, we introduce the MultiModal alignment-guided Learned Token Pruning (MM-LTP) method. MM-LTP employs token pruning, conventionally used for computational efficiency, to perform online text cleaning during multimodal model training. By enabling the model to discern and discard unimportant tokens, it is able to train with implicitly cleaned image-text pairs. We evaluate MM-LTP using a benchmark multimodal e-commerce dataset comprising over 710,000 unique Amazon products. Our evaluation hinges on visual search, a prevalent e-commerce feature. Through MM-LTP, we demonstrate that refining text tokens enhances the paired image branch's training, which leads to significantly improved visual search performance.

1. Introduction

Multimodal transformer models have been widely adopted in e-commerce product search, including but not limited to caption-to-image search, image-to-image search, and multimodal-to-image search [4, 26, 34, 44, 45]. The success of applying multimodal models in e-commerce product search can be attributed to its strength in understanding vision and language representations of product contents. One of the key factors for training an effective vision-language multimodal model relies on the alignment of image-text pairs in the dataset. In practice, the training dataset is usually collected in an automatic fashion with limited manual cleaning or annotation. As a result, the alignment between text and image is far from ideal.



Green Toys Wagon, Orange - Pretend Play, Motor Skills, Kids Outdoor Toy Vehicle. No BPA, phthalates, PVC. Dishwasher Safe, Recycled Plastic, Made in USA.

Figure 1. Example of a product's image-text pair from an e-commerce website. Phrases in the product title are color-coded by their embedding similarity to the image embedding. Both image and text embeddings are generated by the BLIP-2 [20] model.

This misalignment issue is bi-directional: it could be the case that not all the text content is reflected by the paired image, or the corresponding text does not fully describe the image content. In e-commerce applications, the former issue is common [8, 23] and poses a particularly serious challenge to developing effective multimodal models. In order to promote their listings, sellers are inclined to include as many as product attributes in the product title. However, some of these attributes are functional rather than visual. Therefore, these attribute phrases in the title cannot be reflected in the paired image. Figure 1 shows a sample image-text pair of a product. There are over 20 phrases in the product title. However, most of them are not visual-descriptive – *how can you tell it is "No BPA" by looking at the image*? We further quantify the intra-difference between phrases by calculating their embeddings' cosine similarity to the image embedding, and show the result in Figure 1. It is clear that the non-

[*] This work is done during the internship at Amazon.

visual-descriptive phrases have significantly lower similarity to the image. On the other hand, prior works on e-commerce multimodal models mainly align images to the full title in a brute-force fashion [24, 26, 44]. Therefore, these models are prone to non-optimal image-text alignments and may overfit the noisy text, which eventually affects the model's generalization performance [18]. To meet the challenges of noisy image-text pairs, some multimodal research works propose to improve the scale of the training data and model size [2, 13, 30, 39] or employ specific model designs, e.g., the filter module and captioner module in BLIP [19] model and BLIP-2 model [20]. However, for e-commerce applications, the available data is constrained by the scale of the product catalog, thus the volume is not comparable to open-domain data. In addition, models with specific designs usually have complicated architectures that make training and inference unstable [19].

In this paper, we introduce MultiModal alignment-guided Learned Token Pruning (MM-LTP), a simple yet effective method for training the multimodal transformer model with noisy e-commerce image-text training data. The method leverages the token pruning technique, which was popularly used for improving the model's computational efficiency by discarding unimportant tokens [15, 31], to perform online text cleaning during multimodal model training. The key idea is that given that each phrase has a different importance in describing the image, we can let the model learn to remove unimportant tokens alongside its original multimodal training task. As a result, the model can be trained with implicitly-cleaned image-text pairs. Our method also adopts a differentiable soft binarized mask, which enables the model to learn the decisions about which tokens to be pruned given a layer and task. The learning of the mask is guided by multimodal alignment. We design the MM-LTP to be flexible to work with bi-modal models with alignment loss (e.g., CLIP [30]) and multimodal models with multimodal fusion networks (e.g., ALBEF [18].) In addition, the method is flexible to be used for: (1) fine-tuning a model pre-trained on open-domain datasets with e-commerce datasets, and (2) re-fine-tuning a model that has been previously fine-tuned in a non-pruning fashion on e-commerce-oriented datasets.

Given the scarcity of publicly available e-commerce datasets, in order to evaluate our MM-LTP method, we establish a benchmark multimodal e-commerce dataset based on the uni-modal Amazon ESCI dataset [32] with over 710,000 unique products sold on Amazon.com. Similar to the approach adopted by the prior work [17], our work also leverages the strength of multi-modal learning while focusing on the vision encoder for evaluation. This is because in e-commerce, customers predominantly use images as visual cues to search for products, rather than performing image-to-text or text-to-image product search [3, 35, 37, 41, 43]. By retaining only the most salient text tokens, our method ensures clear, concise linguistic cues guide the image branch during training. Focusing on tightly coupled textual concepts improves the image model's ability to recognize and respond to visual patterns. Our text pruning leverages this cross-modal regularization effect to increase the accuracy and efficiency of the image encoder for visual search. With extensive experiments with both ALBEF [18]-like and CLIP [30]-like experiments, we demonstrate the effectiveness of the proposed MM-LTP method. The MM-LTP can boost the model to gain over **5** percentage point on Recall@1, compared with models trained without MM-LTP method. Our main contributions can be summarized as:

- We present the MultiModal alignment-guided Learned Token Pruning (MM-LTP) method, which uses token pruning to enhance on-the-fly text cleaning when training multimodal transformer models. It addresses misalignment challenges on e-commerce datasets.
- The proposed multimodal soft token pruning method is flexible to be integrated with both self-attention and cross-attention mechanisms, and is adaptable to models with either explicit or implicit multimodal fusion.

2. Related Work

Vision-language Pre-training The success of large-scale transformer-based pre-training in the field of Natural Language Processing [7] has boosted research works in vision-language pre-training. These models are trained on large-scale image-text pairs and learn a joint vision-language embedding space for various downstream tasks. CLIP model [30] leverages a broader source of supervision from text to train a predictive model that aligns text with image, resulting in a task-agnostic model comparable to task-specific supervised models. ALIGN [13] scales up the CLIP model with a noisy dataset without expensive filtering or post-processing steps that cover more than one billion image alt-text pairs. CLIP and ALIGN show promising results in vision-based downstream tasks, however, they ignore the interaction between two modalities and vision-language downstream tasks. Recent studies propose to learn joint embeddings of image contents and natural language during pre-training, like OSCAR [22], UNIMO [21] and UNITER [6]. These works use an object detector backbone to capture vision features first, then a transformer-based model is applied to the concatenated vision and text features to learn joint embeddings. ViLT [16] further breaks through the regional feature from convolutional networks and adopts vision transformer [9] to fuse the whole global image feature with natural languages. ALBEF [18] and TCL [38] further exploit contrastive loss functions to align image and text features before modeling their joint embeddings, increasing the interaction between two modalities and achieving a state-of-the-art performance (SOTA).

Multimodal Models for E-commerce Initial works such as FashionBERT [10] and Kaleido-BERT [46] utilized a transformer-based model along with a custom masking strategy for pre-training, aiming to generate more detailed features for clothing retrieval. Following this, CAPTURE [42] introduced a method to generate distinctive instance features through masked multi-modal learning and cross-modal contrastive pre-training, which resulted in impressive performance in instance-level product retrieval tasks. K3M [45] took a step further by incorporating the knowledge modality into multi-modal pre-training to mitigate noise and supplement missing information in the image and text modalities. SCALE [8] put forth a self-harmonized contrastive learning framework capable of integrating six different modalities into a single model. More recently, CommerceMM [40] used a contrastive and MLM-based pre-training that can be applied to 14 different tasks.

3. Methodology

In a nutshell, our method masks text tokens given each token's importance derived from the attention score matrix. The overview of MM-LTP is illustrated in Figure 2.

3.1. Token Importance Quantification

The first step in text pruning is to quantify the importance of each text token in relation to the image data. We focus on quantifying the importance of the two most common fusion approaches. The first is explicit fusion with cross-attention, where token importance is directly reflected in the pairwise attention scores between text and image patches. In this paper, we refer to the paradigm of cross-attention in the ALBEF model [18]. The second one is implicit fusion like in CLIP [30], which uses a contrastive loss between two modalities' representations. Though the text tokens do not explicitly attend to image patches in this case, we hypothesize that analyzing the self-attention patterns within the text encoder similarly reveals fine-grained textual dependencies and importance for grounding in the visual content. The attention score matrix in both cross-attention and self-attention provides model-agnostic insights into how individual tokens are weighted during multimodal alignment, applicable across architectures. Therefore, we propose to use the attention score matrix as guidance for quantifying the importance of text tokens.

Given an input query sequence $\mathrm{x} \in \mathbb{R}^{m\times n}$ with m tokens, and input key sequence $\mathrm{z} \in \mathbb{R}^{k\times l}$ with k tokens, the attention score matrix is calculated as:

$$\mathrm{Attn}(\mathrm{x}, \mathrm{z}) = \frac{\mathrm{x}\mathrm{W}_q\mathrm{W}_k^T\mathrm{z}^T}{\sqrt{d}}, \tag{1}$$

where $\mathrm{W}_q \in \mathbb{R}^{n\times d}$ and $\mathrm{W}_k \in \mathbb{R}^{l\times d}$ are trainable weight matrices. For self-attention, we have $m = k$ and $n = l$. This attention score matrix measures each input query token's pairwise importance on every key token. Given the cross-attention's key tokens are from the image, to make MM-LTP flexible, we focus on quantifying the average importance of query tokens to guide further token pruning. Hence, following [11, 14, 15], we can define the importance score of i-th query token (x_i) in a multi-head attention as:

$$s(\mathrm{x}_i) = \frac{1}{H}\frac{1}{k}\sum_{h=1}^{H}\sum_{j=1}^{k}\mathrm{Attn}(\mathrm{x}_i, \mathrm{z}_j), \tag{2}$$

where H is the number of independently parameterized heads in the multi-head attention. This importance score of i-th query token can be interpreted as the average of all key tokens' attention from all heads. However, not all Key tokens necessarily contribute equally valuable information for determining the query token's importance. Uniformly averaging over all Keys may dilute the useful signal. Therefore, we propose calculating the importance score based on the attention received by the [CLS] token of the Key tokens. In cross-attention, the image [CLS] token encodes aggregated visual concepts. In self-attention, the text [CLS] represents the linguistic context. Attending to these consolidated representations provides a less noisy measure of query relevance than individual Keys. Specifically, the importance score of the i-th query token is computed as the average of the Key [CLS] token's attention from all heads. Assuming the [CLS] token is in the first (0-th) position of the sequence, the updated importance score of i-th query token (x_i) in multi-head attention can be defined as:

$$s(\mathrm{x}_i) = \frac{1}{H}\sum_{h=1}^{H}\mathrm{Attn}(\mathrm{x}_i, \mathrm{z}_0), \tag{3}$$

With this refined importance metric, MM-LTP condenses the diverse key set into compact unified representations, enabling a robust quantification of the query token's importance.

3.2. Pruning with Learned Threshold

Given each query token's importance, MM-LTP prunes unimportant tokens by comparing the score with a threshold τ. This process allows the model to discard noisy tokens that contribute less to multimodal alignment and fusion. However, setting the value of τ' s is a nontrivial task. The appropriate threshold may differ between tasks and datasets. The threshold may also vary across transformer layers, as deeper layers capture higher-level concepts where fewer tokens may be relevant. Hence manually setting a static heuristic threshold is impractical. Therefore, we model τ as a learnable parameter, allowing it to adapt to the specific requirements of each task, data, and layer.

A combination learns the τ of two components, namely differentiable pruning mask and token pruning loss. The dif-

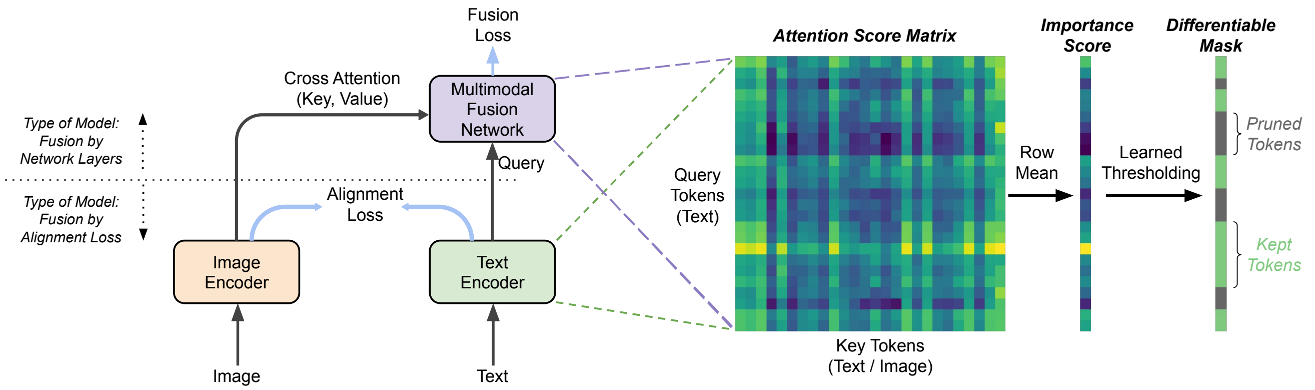


Figure 2. Overview of the MM-LTP method. It is flexible to work with the self-attention matrix in the text encoder or cross-attention matrix if the model has fusion network layers. It takes the attention score matrix to calculate the importance score for each query text token. The unimportant tokens are masked following a learnable thresholding mechanism.

ferentiable pruning mask approximates a non-differentiable binary mask with the learnable τ during back propagation. Inspired by Tempered Sigmoid Activations [28], for the l-th layer in the model, the differentiable pruning mask for the i-th query token (x_i) is defined as:

$$M_l(\mathrm{x}_i) = \sigma(\frac{s_l(\mathrm{x}_i) - \tau_l}{T}), \tag{4}$$

where T is the temperature parameter. When T is sufficiently small, the output of tokens whose importance score is greater than τ will be close to one, and vice versa. The mask $M_l(\mathrm{x}_i)$ is then multiplied with i-th query token's output at layer l. For tokens whose importance scores are smaller than the threshold, their layer output is close to zero and hence they will not become major information sources in succeeding layers, which has an equivalent effect of suppressing these tokens. Analytically, the gradient $\frac{dM_l(\mathrm{x}_i)}{d\tau_l}$ achieves its maximum magnitude when the threshold is close enough to the importance score. This implies that threshold training can be focused specifically on tokens that are on the verge of being pruned or retained, rather than on all tokens indiscriminately [15].

To encourage the model for pruning, we adopt pruning loss as an additional training objective, which is commonly found in prior works [15]. We propose an L1 loss-based method:

$$\mathcal{L}_{Prune} = \frac{1}{N} \sum_{l=1}^{L} \frac{\|M_l(\mathrm{x})\|_1}{d_l^Q}, \tag{5}$$

where d_l^Q is the sequence length of the Query at layer l. The scaling factor d_l^Q is designed for models with dynamic Query length, which is helpful for normalizing the mask's L1 norm to a unified scale. Intuitively, when more tokens are situated close to the threshold, the gradient $\frac{\mathcal{L}_{Prune}}{d\tau_l}$ becomes larger. Consequently, this causes an increase in the threshold value, resulting in the pruning of a greater number of tokens that are proximate to the threshold boundary. Generally, for models with original training objectives $\mathcal{L}_{Model}$, the updated training objective is:

$$\mathcal{L} = \mathcal{L}_{Model} + \lambda \cdot \mathcal{L}_{Prune}, \tag{6}$$

where λ is the regularization parameter to control the aggressiveness of pruning.

	Train	Test	Total
Products	637,511	80,000	717,511
Pairs	858,231	186,084	1,044,315

Table 1. Dataset statistics. The pair in the training set stands for the image-text pairs, while the pair in the testing set means the pair of the main image and one auxiliary image.

Category	Consumable	Hardline	Softline	Others
Ratio	16.31%	61.22%	10.99%	11.48%

Table 2. Distribution of product categories of the dataset.

4. Experiments

4.1. Dataset

A dataset consisting of image-text pairs of general-purposed e-commerce product types is necessary for evaluating the proposed MM-LTP method. However, the dataset used in prior works can hardly fulfill this requirement. For example, the Fashion-Gen Data [46], Fashion 200k data [12], Shopping100 data [1], and FashionIQ data [36] all focus on the fashion domain. The M5 Product Data [8] and Product 1M Data [42] are in the form of Chinese product titles, as

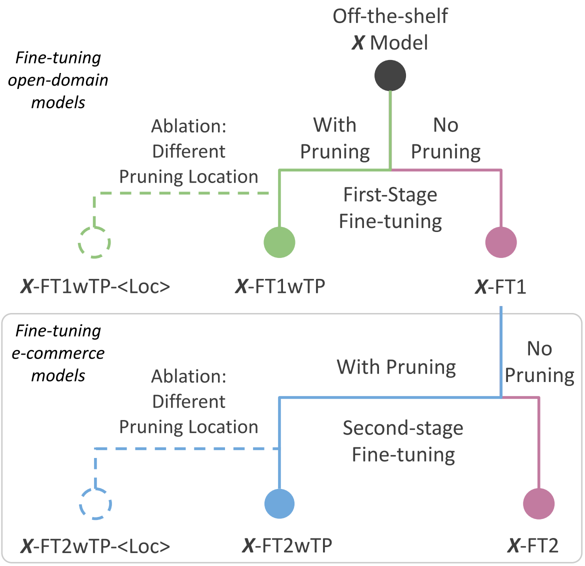


Figure 3. Illustration of the experiment and naming rules. The MM-LTP is designed to be flexible to work with both fine-tuning the model pre-trained with open-domain datasets, as well as e-commerce domain models.

the unique characteristics of the Chinese language and its tokenizing effect on the proposed MM-LTP is out of the scope of this work. Also, there are related works that collect data from online marketplaces without releasing the dataset [3, 35, 37, 40, 41, 43]. Therefore, we establish a benchmark multimodal dataset[1] in the general e-commerce domain with English captions. The dataset is built upon the Amazon ECSI dataset [32], which is a uni-modal dataset for product shopping queries. After removing products that are no longer available or have less than two images, the dataset covers over 710,000 products sold on Amazon.com. Each product's data contains a product title, a main image, and multiple (0 to 10) auxiliary images. The statistics of this dataset can be found in Table 1. This dataset covers most common product categories, including but not limited to *Hardlines* (e.g., electronics, furniture, ...), *Softlines* (apparel, shoes,..), *Consumables* (personal care, pantry, ...), etc. The distribution of product categories can be found in Table 2.

4.2. Experiment Setting

We evaluate MM-LTP in a self-contained fashion with a two-stage process. The experiment setting is depicted in Figure 3. Designed with flexibility in mind, MM-LTP can be applied to both open-domain pre-trained models and models already fine-tuned for the e-commerce field. In the first stage, we utilize the pre-trained weights of the open-domain model to assess MM-LTP's effectiveness in fine-tuning open-domain models. This stage demonstrates how MM-LTP can improve general-purpose models by selectively pruning tokens. In the second stage, we use the model fine-tuned in a common, non-pruning fashion on the target domain as a starting point, to evaluate MM-LTP's capability in further improving the fine-tuned e-commerce model. This stage illustrates MM-LTP's adaptability and effectiveness in a specialized domain, where precise alignment between visual and textual information is crucial. For both stages, we establish a baseline using the model without token pruning, ensuring a fair and comprehensive comparison. Additionally, we include an ablation study to explore the impact of the layer of pruning, considering models equipped with both self-attention and cross-attention. In this paper, we select CLIP [30] and ALBEF [18] as exemplar multimodal models for all evaluations mentioned above. For ALBEF, the default pruning layer is the fusion encoder with cross-attention. We also evaluate the performance of applying the pruning to both the fusion encoder and text encoder's self-attention, on top of the default setting.

4.3. Implementation Details and Metric

All experiments were conducted using 8 NVIDIA A100 GPUs, utilizing the PyTorch deep learning framework [29] and the Ray distributed computing framework [27]. Both the CLIP and ALBEF models employ a standard ViT-B/16 [9] vision encoder with 12 layers and 86M parameters. CLIP's text encoder is a 12-layer transformer with 63M parameters, while ALBEF's text and fusion encoders are built on a 6-layer transformer, totaling 124M parameters. Both models use pre-trained weights provided by their authors. For the CLIP model, training is conducted for over 100 epochs with a batch size of 1360, using the AdamW optimizer [25] with a weight decay of 0.02. The learning rate was initialized at $5e^{-6}$, warmed up to $2e^{-5}$ after 10 epochs, and then decreased to $5e^{-6}$ using the cosine decay strategy. For ALBEF, the original work's pre-training configuration was used for the first-stage fine-tuning experiments, and the retrieval training configuration was applied for the second-stage fine-tuning experiments. The batch size is adjusted to 320 from the original configurations. In token pruning, layer-wise thresholds are initialized with linearly rising values, ending with a fixed threshold of 0.01 at the final layer. The temperature parameter T is set at $1e^{-4}$. From empirical exploration, a pruning loss's regularization parameter λ of 0.1 is found suitable for all experiments.

We adopt the standard evaluation metric in retrieval, i.e., Recall@K (denoted as R@K), which is defined as the proportion of test queries for which the correct targets are successfully identified within the top-K retrieved samples [5]. Unless specified, the unit in tables of retrieval performance is the percentage (%.)

[1] We plan to release this dataset.

	R@1	R@5	R@10
Off-the-shelf CLIP	42.56	51.29	56.62
CLIP-FT1	53.59	63.24	69.03
CLIP-FT1wTP	**55.21**	**65.13**	**70.97**
↑	1.62	1.89	1.94

Table 3. Retrieval performance of CLIP model with different first-stage fine-tuning settings. The best results are in bold font. ↑ indicates the relative improvement with MM-LTP. Note that the CLIP-FT1 model also achieves a significant improvement, which underscores the domain disparity between the off-the-shelf model's pre-training data and our e-commerce dataset. The ↑ indicates the improvement from models further fine-tuned by the e-commerce datasets. (Unless specifically mentioned, this applies to all tabular results).

4.4. First-Stage Retrieval Performance

The primary objective of the first-stage retrieval is to assess the capability of MM-LTP in improving the fine-tuning of models that are pre-trained on open-domain data, specifically for e-commerce product retrievals. The results for CLIP and ALBEF, when integrated with MM-LTP, are presented in Table 3 and Table 4, respectively. These results highlight that MM-LTP effectively boosts the retrieval performance across models with varying fusion strategies.

CLIP As illustrated in Table 3, both CLIP-FT1 and CLIP-FT1wTP register an improvement of over 10 percentage points compared to the baseline CLIP. This significant enhancement underscores the domain disparity between CLIP's original pre-training data and our specialized e-commerce dataset. Notably, CLIP-FT1wTP outperforms CLIP-FT1 by over 1.6 percentage points across all metrics. Such a performance boost suggests that MM-LTP is particularly adept at refining models that rely on self-attention-based text encoders. Given that models like CLIP solely utilize self-attention and are guided by contrastive loss, MM-LTP's ability to prune redundant and noisy tokens from product captions is especially desirable. Pruning not only reduces noise in the text encoder but also provides cleaner and more focused textual cues to the vision backbone. When the vision model is trained with these refined textual cues, it can form a better association between visual and textual features. The improved alignment between the two modalities ensures that the vision backbone can identify and retrieve relevant products more accurately, based on the denoised textual information. In essence, by enhancing the text representation, MM-LTP indirectly strengthens the vision backbone, leading to improved retrieval performance in e-commerce settings.

ALBEF Table 4 presents the retrieval performance of the ALBEF model under various configurations. The results reveal that the integration of MM-LTP with cross-attention leads to over three percentage points retrieval performance boost. Remarkably, when pruning is extended to both self-attention and cross-attention, there's an uplift of over five percentage points. This gain surpasses the improvement observed in the CLIP model. Such outcomes suggest that MM-LTP is more adept at pruning noisy text tokens when there's an explicit multimodal interaction, such as cross-attention, compared to an implicit one.

	R@1	R@5	R@10
Off-the-shelf ALBEF	38.60	47.40	53.03
ALBEF-FT1	51.68	62.27	68.68
ALBEF-FT1wTP	54.59	65.09	71.36
ALBEF-FT1wTP-All	**57.06**	**67.54**	**73.74**
↑	5.38	5.27	5.06

Table 4. Retrieval performance of ALBEF model with different first-stage fine-tuning settings. ↑ indicates the highest relative improvement with MM-LTP.

Furthermore, the intricate design of multiple optimization objectives in ALBEF complements MM-LTP's functionality. The contrastive alignment loss and self-attention mechanisms in ALBEF's text encoder are similar to those in CLIP. They effectively serve as a preprocessing step for explicit multimodal fusion. By the time the multimodal fusion occurs, the text tokens deemed noisy and redundant have already undergone soft-pruning. This ensures that the vision embeddings are more attuned to the remaining informative text tokens, thereby enhancing their alignment and retrieval accuracy. This synergy between MM-LTP and ALBEF's design is particularly beneficial for training a robust vision encoder. The improved alignment is crucial for image-to-image retrieval tasks. In such tasks, the model relies heavily on the vision encoder to extract and compare visual features, and vision encoders trained from denoised textual information have better capabilities of recognizing subtle visual patterns and nuances.

The evaluation of CLIP and ALBEF demonstrates the effectiveness of MM-LTP in the first-stage fine-tuning, particularly for models pre-trained with open-domain data being adapted for e-commerce applications.

4.5. Second-Stage Retrieval Performance

Table 5 and Table 6 present the results of the second-stage fine-tuning for both models, simulating the application of MM-LTP for fine-tuning an e-commerce model. For CLIP, the integration of MM-LTP results in an over 2.3 percentage points boost in retrieval performance. In the case of ALBEF, MM-LTP contributes to an improvement of approximately five percentage points across all metrics.

These outcomes show the potential of MM-LTP in refining e-commerce models to improve performance. When compared to the baseline model CLIP-FT2 which undergoes

	R@1	R@5	R@10
Off-the-shelf CLIP	42.56	51.29	56.62
CLIP-FT1	53.59	63.24	69.03
CLIP-FT2	54.55	64.29	70.08
CLIP-FT2wTP	**55.95**	**65.81**	**71.57**
↑	2.36	2.57	2.54

Table 5. Retrieval performance of CLIP model with different fine-tuning settings at different stages. ↑ indicates the improvement with MM-LTP in the second-stage fine-tuning.

	R@1	R@5	R@10
Off-the-shelf ALBEF	38.60	47.40	53.03
ALBEF-FT1	51.68	62.27	68.68
ALBEF-FT2	53.83	64.44	70.83
ALBEF-FT2wTP	54.77	65.37	71.78
ALBEF-FT2wTP-All	**56.75**	**67.39**	**73.65**
↑	5.07	5.12	4.97

Table 6. Retrieval performance of ALBEF model with different fine-tuning settings at different stages. ↑ indicates the highest improvement with MM-LTP in the second-stage fine-tuning.

	R@1	R@5	R@10
CLIP-FT1wTP-Setup 1	53.99	64.06	70.10
CLIP-FT1wTP-Setup 2	53.38	63.41	69.26
CLIP-FT1wTP-Setup 3	54.92	64.75	70.65
CLIP-FT1wTP	**55.21**	**65.13**	**70.97**

Table 7. Retrieval performance of CLIP model with different importance score's calculation setup in the first-stage fine-tuning.

fine-tuning without token pruning, the CLIP-FT2wTP model outperforms it by roughly one percentage point. Similarly, ALBEF's top-performing second-stage fine-tuned model, ALBEF-FT2wTP-All, exceeds the baseline ALBEF-FT2 by about three percentage points. The trajectory of these improvements mirrors the trends observed during the first-stage fine-tuning.

Delving deeper into the comparative gains between the two stages, it becomes evident that MM-LTP achieves more significant improvement during the first stage. One plausible explanation for this observation is that, at the first stage, the model is more malleable, allowing MM-LTP to more effectively prune and refine the textual cues.

4.6. Ablation Study

	R@1	R@5	R@10
CLIP-FT2wTP-Setup 1	54.03	63.86	69.82
CLIP-FT2wTP-Setup 2	53.69	63.35	69.18
CLIP-FT2wTP-Setup 3	55.68	65.55	71.31
CLIP-FT2wTP	**55.95**	**65.81**	**71.57**

Table 8. Retrieval performance of CLIP model with different importance score's calculation setup in the second-stage fine-tuning.

	R@1	R@5	R@10
ALBEF-FT1wTP-Setup 1	50.42	60.75	67.12
ALBEF-FT1wTP-Setup 2	50.72	61.13	67.52
ALBEF-FT1wTP-Setup 3	54.26	64.73	71.01
ALEBF-FT1wTP	54.59	65.09	71.36
ALBEF-FT1wTP-All-Setup 1	56.15	66.67	72.94
ALBEF-FT1wTP-All-Setup 2	56.55	67.15	73.40
ALBEF-FT1wTP-All-Setup 3	56.82	67.46	73.70
ALBEF-FT1wTP-All	**57.06**	**67.54**	**73.74**

Table 9. Retrieval performance of ALBEF model with different importance score's calculation setup in the first-stage fine-tuning..

To analyze the effectiveness of calculating the importance score using the Key's [CLS] token with the Query's non-[CLS] token, we carried out an ablation study with distinct setups for importance score computation: *Setup 1*: The computation involves all Key tokens and all Query tokens. *Setup 2*: The computation uses the Key's [CLS] token and all Query tokens. *Setup 3*: The computation incorporates all Key tokens and the Query's non-[CLS] token.

The retrieval results for both CLIP and ALBEF models under these configurations are detailed in Table 7, Table 8, and Table 9. A comprehensive analysis of these results reveals that models employing our proposed calculation approach consistently yield the topmost retrieval performance, irrespective of the fine-tuning stages and token pruning layers. Among the three configurations, Setup 3 has the performance closest to that of MM-LTP. For the CLIP model, Setup 1 slightly outperforms Setup 2. However, for ALBEF, Setup 1 lags behind Setup 2 by a narrow margin.

These findings confirm the value of harnessing the aggregated information present in the Key's [CLS] token. This approach acts as an additional denoising step that refines the importance scores. Particularly for models like ALBEF, which employ cross-attention between text (Query) and image (Key), potential redundancy and noise exist in both modalities. By emphasizing the [CLS] token's consolidated information, we mitigate these challenges, ensuring a more accurate alignment between text and image representations. Furthermore, by focusing on cleaner and more concise textual cues, the vision encoder is trained to recognize and prioritize salient visual features more effectively.

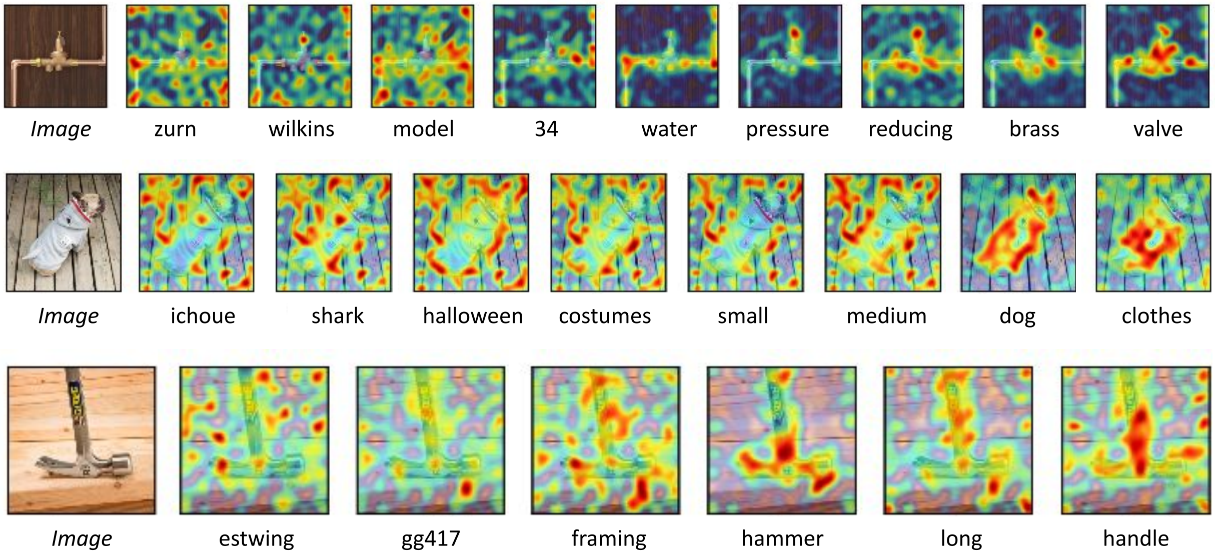


Figure 4. Grad-CAM visualizations on the cross-attention maps of the ALBEF-FT1wTP model, corresponding to individual words in the product title.

4.7. Grad-CAM Visualization

In Figure 4, We provide a Grad-CAM [33] visualization of the ALBEF-FT1wTP model's cross-attention maps corresponding to each word in the product title. Following the implementation in ALBEF's paper, we pick the multimodal fusion encoder's third layer for the visualization.

The attention maps reveal distinct patterns of focus. Words that are visually descriptive, such as ”valve,” ”dog,” and ”handle,” exhibit concentrated attention areas. This suggests that the model emphasizes regions in the image that correspond to these descriptive terms. In contrast, brand names or words that lack a direct visual counterpart in the image, like ”zurn,” ”ichoue,” and ”estwing,” show diffused and scattered attention patterns.

The difference in attention distribution demonstrates the model's ability to discern between text tokens. The model appears to diminish its attention toward text tokens that are potentially noisy or less relevant while honing in on tokens that provide meaningful visual cues. Such behavior aligns with our fundamental hypothesis and motivation: to prioritize informative text tokens and reduce the influence of extraneous ones. This selective attention mechanism not only highlights the model's capability to differentiate between visually grounded and non-grounded textual information but also provides a rationale for our token pruning approach.

5. Discussion

The MM-LTP method has demonstrated its effectiveness through our evaluation. To ensure its robustness and wide applicability, future investigations will focus on evaluating MM-LTP's adaptability and efficiency across a spectrum of real-world scenarios, including datasets with diverse levels of image-text misalignment, varying sizes, and different balances between visual and non-visual attributes.

Exploring MM-LTP's performance across different backbone model sizes and specialized product categories will also be crucial in uncovering its limits and potential. By conducting these investigations and assessing its applicability to a wide range of e-commerce platforms, we aim to establish MM-LTP as a robust and versatile solution for improving e-commerce product search experiences.

6. Conclusion

In this paper, we address the challenges of noisy image-text pair alignments in e-commerce datasets and propose the MM-LTP method as a solution. Leveraging token pruning, MM-LTP facilitates training multimodal transformer models with cleaner image-text pairings. By pruning redundant and noisy text tokens, MM-LTP denoises the text branch and strengthens the vision encoder, leading to a more efficient multimodal model for e-commerce applications. Our evaluation with a large-scale e-commerce dataset has demonstrated MM-LTP's effectiveness in improving visual search performance. Also, the proposed method is flexible and compatible with models like CLIP that rely on alignment loss and those like ALBEF with fusion networks.

References

[1] Kenan E Ak, Joo Hwee Lim, Jo Yew Tham, and Ashraf A Kassim. Efficient multi-attribute similarity learning towards

attribute-based fashion search. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1671–1679. IEEE, 2018. 4

[2] Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, Songhao Piao, and Furu Wei. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *Advances in Neural Information Processing Systems*, 35:32897–32912, 2022. 2

[3] Sean Bell, Yiqun Liu, Sami Alsheikh, Yina Tang, Edward Pizzi, M Henning, Karun Singh, Omkar Parkhi, and Fedor Borisyuk. Groknet: Unified computer vision model trunk and embeddings for commerce. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2608–2616, 2020. 2, 5

[4] Ben Chen, Linbo Jin, Xinxin Wang, Dehong Gao, Wen Jiang, and Wei Ning. Unified vision-language representation modeling for e-commerce same-style products retrieval. *arXiv preprint arXiv:2302.05093*, 2023. 1

[5] Yanbei Chen and Loris Bazzani. Learning joint visual semantic matching embeddings for language-guided retrieval. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16*, pages 136–152. Springer, 2020. 5

[6] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020. 2

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2

[8] Xiao Dong, Xunlin Zhan, Yangxin Wu, Yunchao Wei, Michael C Kampffmeyer, Xiaoyong Wei, Minlong Lu, Yaowei Wang, and Xiaodan Liang. M5product: Self-harmonized contrastive learning for e-commercial multi-modal pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21252–21262, 2022. 1, 3, 4

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2, 5

[10] Dehong Gao, Linbo Jin, Ben Chen, Minghui Qiu, Peng Li, Yi Wei, Yi Hu, and Hao Wang. 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, 2020. 3

[11] Saurabh Goyal, Anamitra Roy Choudhury, Saurabh Raje, Venkatesan Chakaravarthy, Yogish Sabharwal, and Ashish Verma. Power-bert: Accelerating bert inference via progressive word-vector elimination. In *International Conference on Machine Learning*, pages 3690–3699. PMLR, 2020. 3

[12] Xintong Han, Zuxuan Wu, Phoenix X Huang, Xiao Zhang, Menglong Zhu, Yuan Li, Yang Zhao, and Larry S Davis. Automatic spatially-aware fashion concept discovery. In *Proceedings of the IEEE international conference on computer vision*, pages 1463–1471, 2017. 4

[13] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021. 2

[14] Gyuwan Kim and Kyunghyun Cho. Length-adaptive transformer: Train once with length drop, use anytime with search. *arXiv preprint arXiv:2010.07003*, 2020. 3

[15] Sehoon Kim, Sheng Shen, David Thorsley, Amir Gholami, Woosuk Kwon, Joseph Hassoun, and Kurt Keutzer. Learned token pruning for transformers. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 784–794, 2022. 2, 3, 4

[16] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR, 2021. 2

[17] Kushal Kumar, Tarik Arici, Tal Neiman, Jinyu Yang, Shioulin Sam, Yi Xu, Hakan Ferhatosmanoglu, and Ismail Tutar. Unsupervised multi-modal representation learning for high quality retrieval of similar products at e-commerce scale. *arXiv preprint arXiv:2008.10726*, 2023. 2

[18] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021. 2, 3, 5

[19] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022. 2

[20] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 1, 2

[21] Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. Unimo: Towards unified-modal understanding and generation via cross-modal contrastive learning. *arXiv preprint arXiv:2012.15409*, 2020. 2

[22] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer, 2020. 2

[23] Fan Liu, Delong Chen, Xiaoyu Du, Ruizhuo Gao, and Feng Xu. Mep-3m: A large-scale multi-modal e-commerce product dataset. *Pattern Recognition*, 140:109519, 2023. 1

[24] Shilei Liu, Lin Li, Jun Song, Yonghua Yang, and Xiaoyi Zeng. Multimodal pre-training with self-distillation for product understanding in e-commerce. In *Proceedings of the Sixteenth*

ACM International Conference on Web Search and Data Mining, pages 1039–1047, 2023. 2

[25] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[26] Haoyu Ma, Handong Zhao, Zhe Lin, Ajinkya Kale, Zhangyang Wang, Tong Yu, Jiuxiang Gu, Sunav Choudhary, and Xiaohui Xie. 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, 2022. 1, 2

[27] Philipp Moritz, Robert Nishihara, Stephanie Wang, Alexey Tumanov, Richard Liaw, Eric Liang, Melih Elibol, Zongheng Yang, William Paul, Michael I Jordan, et al. Ray: A distributed framework for emerging {AI} applications. In *13th USENIX symposium on operating systems design and implementation (OSDI 18)*, pages 561–577, 2018. 5

[28] Nicolas Papernot, Abhradeep Thakurta, Shuang Song, Steve Chien, and Úlfar Erlingsson. Tempered sigmoid activations for deep learning with differential privacy. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9312–9321, 2021. 4

[29] Adam Paszke, Sam Gross, Soumith Chintala, Yang Wei, Zhe Wang, Joseph Turner, Alban Desmaison, Luca Antiga, and Jeff Donahu. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32, 2019. 5

[30] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 3, 5

[31] Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. Dynamicvit: Efficient vision transformers with dynamic token sparsification. *Advances in neural information processing systems*, 34:13937–13949, 2021. 2

[32] Chandan K. Reddy, Lluís Màrquez, Fran Valero, Nikhil Rao, Hugo Zaragoza, Sambaran Bandyopadhyay, Arnab Biswas, Anlu Xing, and Karthik Subbian. Shopping queries dataset: A large-scale ESCI benchmark for improving product search. 2022. 2, 5

[33] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 8

[34] Youhua Tang, Xiong Xiong, Siyang Sun, Baoliang Cui, Yun Zheng, and Haihong Tang. Tmml: Text-guided mulimodal product location for alleviating retrieval inconsistency in e-commerce. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3275–3279, 2023. 1

[35] Son Tran, Ming Du, Sampath Chanda, R Manmatha, and Cj Taylor. Searching for apparel products from images in the wild. *arXiv preprint arXiv:1907.02244*, 2019. 2, 5

[36] Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris. Fashion iq: A new dataset towards retrieving images by natural language feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11307–11317, 2021. 4

[37] Fan Yang, Ajinkya Kale, Yury Bubnov, Leon Stein, Qiaosong Wang, Hadi Kiapour, and Robinson Piramuthu. Visual search at ebay. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 2101–2110, 2017. 2, 5

[38] Jinyu Yang, Jiali Duan, Son Tran, Yi Xu, Sampath Chanda, Liqun Chen, Belinda Zeng, Trishul Chilimbi, and Junzhou Huang. Vision-language pre-training with triple contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15671–15680, 2022. 2

[39] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022. 2

[40] Licheng Yu, Jun Chen, Animesh Sinha, Mengjiao Wang, Yu Chen, Tamara L Berg, and Ning Zhang. 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, 2022. 3, 5

[41] Andrew Zhai, Hao-Yu Wu, Eric Tzeng, Dong Huk Park, and Charles Rosenberg. Learning a unified embedding for visual search at pinterest. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2412–2420, 2019. 2, 5

[42] Xunlin Zhan, Yangxin Wu, Xiao Dong, Yunchao Wei, Minlong Lu, Yichi Zhang, Hang Xu, and Xiaodan Liang. Product1m: Towards weakly supervised instance-level product retrieval via cross-modal pretraining. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11782–11791, 2021. 3, 4

[43] Yanhao Zhang, Pan Pan, Yun Zheng, Kang Zhao, Yingya Zhang, Xiaofeng Ren, and Rong Jin. Visual search at alibaba. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 993–1001, 2018. 2, 5

[44] Xiaoyang Zheng, Zilong Wang, Sen Li, Ke Xu, Tao Zhuang, Qingwen Liu, and Xiaoyi Zeng. Make: Vision-language pre-training based product retrieval in taobao search. In *Companion Proceedings of the ACM Web Conference 2023*, pages 356–360, 2023. 1, 2

[45] Yushan Zhu, Huaixiao Zhao, Wen Zhang, Ganqiang Ye, Hui Chen, Ningyu Zhang, and Huajun Chen. Knowledge perceived multi-modal pretraining in e-commerce. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 2744–2752, 2021. 1, 3

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

puter Vision and Pattern Recognition, pages 12647–12657, 2021. 3, 4