KG-FLIP: Knowledge-guided Fashion-domain Language-Image Pre-training for E-commerce

Qinjin Jia[†1*] **Yang Liu**[†✉2] **Shaoyuan Xu**[2] **Huidong Liu**[2]
Daoping Wu[3*] **Jinmiao Fu**[2] **Roland Vollgraf**[2] **Bryan Wang**[2]
[1] North Carolina State University [2] Amazon.com, Inc. [3] Iowa State University
[†]Equal contribution [✉] Corresponding author: {yliuu@amazon.com}

Abstract

Various Vision-Language Pre-training (VLP) models (e.g., CLIP, BLIP) have sprung up and dramatically improved the benchmarks of public general-domain datasets (e.g., COCO, Flickr30k). Such models typically learn the cross-modal alignment from large-scale well-aligned image-text datasets. Adapting these models to downstream applications in specific domains, such as fashion, requires fine-grained in-domain image-text datasets. However, such datasets are usually less semantically aligned and smaller in scale, which requires more efficient pre-training strategies. In this paper, we propose a knowledge-guided fashion-domain language-image pre-training (KG-FLIP) framework that focuses on learning fine-grained representations in the e-commerce domain and utilizes external knowledge (i.e., product attribute schema) to improve the pre-training efficiency. Experimental results demonstrate that KG-FLIP outperforms previous state-of-the-art VLP models on Amazon data and the Fashion-Gen dataset by large margins. KG-FLIP has been successfully deployed in the Amazon catalog system to backfill missing attributes and improve the customer shopping experience.

1 Introduction

Modern e-commerce websites exhibit products with multi-modal information (e.g., product images, product titles, and product bullet points) to inform customers' purchase decisions. The effective exploitation of such multi-modal product information is crucial for product understanding and downstream vision-language (VL) applications, such as product categorization, search, and recommendation. Meanwhile, recent large-scale vision-language pre-training (VLP) models have led to impressive performance improvements on many general-domain VL tasks (Radford et al., 2021; Yu et al., 2022). As a result, there has been a surge of interest in adapting such VLP models to facilitate various applications in e-commerce scenarios.

Unlike the well-aligned coarse-grained language-image datasets in the general domain, the paired data in the e-commerce domain have two characteristics. First, both of the product titles/descriptions and the images contain richly detailed (i.e., fine-grained) product information compared to datasets in the general domain. Second, the product textual information and images usually share only partial information while containing complementary information (i.e., not well-aligned). Thus, an effective pre-training method needs to align the common portion and fuse the distinct facts from each modality in a fine-grained manner. Rather than aligning the entire image and text pair at a global level using contrastive loss as CLIP (Radford et al., 2021) does, we designed our pre-training tasks to focus on a finer level of text tokens and image patches.

In addition, previous VLP methods relied solely on the inductive bias of the model to align cross-modality representations through vast amounts of paired data. Such an approach is data-hungry, inefficient, and disregards the availability of structured product knowledge. Thus, we propose to leverage existing knowledge in the e-commerce catalog to facilitate such alignment. Specifically, for each type of product (e.g., dress), the catalog stores its applicable attributes (e.g., neckline style) and enumerated attribute values (e.g., v-neck, crew-neck). Such attribute knowledge can serve as anchor points to help VLP models efficiently acquire salient semantic relations between modalities.

To address the above challenges, we propose KG-FLIP: a **k**nowledge-**g**uided **F**ashion-domain **L**anguage-**I**mage **P**re-training to improve the VLP models for e-commerce data. The design of KG-FLIP is inspired by the state-of-the-art general-purpose VLP model BLIP [6]. We adapt its design for our use case by 1) replacing the widely-used image-text contrastive (ITC) objective with

[*]Work done during internship at Amazon.

the masked language-image modeling (MLIM) pre-training objective, to facilitate multi-model fusion at the token level instead of cross-model alignment; 2) leveraging the structured knowledge of product attribute schema information to guide the pre-training process, and facilitate the VLP model to learn more fine-grained product representations. These enhancements can be generalized to other real-world applications, where image-text pairs are not well-aligned in semantics and external knowledge can be leveraged to guide the pre-training.

2 Related Work

2.1 Vision-Language Pre-training

The emergence of large-scale pre-training models (e.g., BERT (Devlin et al., 2018), ViT (Kim et al., 2021)) has significantly advanced the state of the art across various uni-modal domains, such as natural language processing (NLP), computer vision (CV), and speech recognition (SR). Recently, researchers have introduced the pre-training and-then fine-tuning paradigm into the vision-language (VL) domain for solving multi-modal tasks, which requires models to comprehend both the input image and text contents (Dou et al., 2022). Existing vision-language pre-training (VLP) models (e.g., CLIP (Radford et al., 2021), ALIGN (Jia et al., 2021), Flamingo (Alayrac et al., 2022)) have proven to be highly effective on various downstream VL tasks, such as image retrieval (IR), text retrieval (TR), and visual question answering (VQA). Consequently, VLP has become the de facto practice to tackle multi-modal problems because of its superior performance (Dou et al., 2022; Chen et al., 2023).

Existing VLP models can be divided into two categories: object-detector (OD)-based VLP models (e.g., LXMERT (Tan and Bansal, 2019), UNITER (Chen et al., 2020), OSCAR (Li et al., 2020)) and end-to-end VLP models (e.g., ALIGN (Jia et al., 2021), ALBEF (Li et al., 2021), METER (Dou et al., 2022)). OD-based VLP models rely on pre-trained object detectors to extract region-based image features, and then utilize a multi-modal encoder to fuse the image features with text tokens. While OD-based VLP models have brought impressive performance, crafting the pre-trained object detectors for them is both annotation-expensive and computation-expensive, because it requires bounding box annotations for pre-training and high-resolution images during inference (Li et al., 2021). On the other hand, end-to-end VLP models directly feed image patch features into a pre-trained ViT model, which eliminates the need for costly annotations and significantly improves inference speed, and have been adopted by the more recent work (Chen et al., 2021; Kim et al., 2021). Thus, we focus on end-to-end VLP models in this work.

2.2 Knowledge-enhanced Vision-Language Pre-training

Recently, there has been a surge of interest in utilizing domain knowledge (e.g., knowledge graph, keywords) to guide VLP in order to reach better performance and improve the pre-training efficiency. For example, Chen et al. (2021) proposed to incorporate knowledge graph (KG) embeddings into VLP models to enhance the learning of semantically aligned and knowledge-aware representations. Although their experimental results demonstrated that KG could benefit VLP, it requires object tags in each image to construct domain-specific KGs. Zhu et al. (2021) presented a knowledge-perceived multi-modal pre-training model in e-commerce that uses product attribute information as the third modality in addition to the visual and linguist modalities. However, this approach requires complete and low-noise product attribute information, and its downstream tasks also require such quality product attribute information to be available as input. This implies increased annotation costs and reduces the scope of the VLP model for use on downstream tasks or data. Considering that product attribute information is usually incomplete and noisy in the real world, we think existing knowledge-enhancement approaches are not optimal, because they either require additional labeling efforts or introduce additional noise to VLP models. Thus, we propose to use attribute information to improve the pre-training efficiency of VLP.

3 Method

This section delineates KG-FLIP. Section 3.1 introduces the architecture of KG-FLIP. Then, Section 3.2 presents pre-training objectives of the model. Finally, Section 3.3 explains how we inject attribute knowledge into KG-FLIP.

3.1 FLIP Architecture

We use the BLIP (Li et al., 2022) architecture as our backbone model, which is now one of the state-of-the-art general-purpose VLP models. We choose BLIP for the following reasons: 1) instead of using

a pre-trained object detector as the image encoder, BLIP uses ViT (Dosovitskiy et al., 2020), which is more computing-friendly and eliminates the need for bounding box annotations; 2) BLIP has a specially added text decoder – thus can be utilized for both VL understanding (e.g., multi-modal attribute classification) and VL generation (e.g., image captioning) downstream tasks in e-commerce; 3) training a VLP model from scratch is time-consuming and expensive. Reusing a pre-trained checkpoint, which has been empirically demonstrated to be very effective, can conspicuously reduce the R&D time and expenses of our proposed KG-FLIP model.

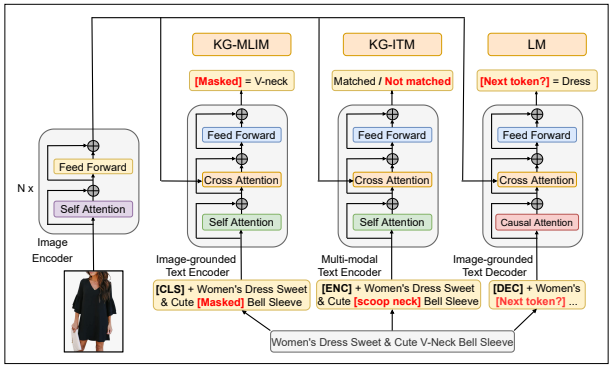

Figure 1: The architecture of KG-FLIP. It consists of an image encoder, an image-grounded text encoder, a multi-modal text encoder, and an image-grounded text decoder. Three pre-training tasks at the top are: knowledge-guided masked language-image modeling (KG-MLIM); knowledge-guided image-text matching (KG-ITM); and language modeling (LM). Components with the same color use the hard-parameter sharing.

As illustrated in Figure 1, KG-FLIP contains an image encoder, an image-grounded text encoder, a multi-modal text encoder, and an image-grounded text decoder. Similar to BLIP, we craft the image encoder using the visual transformer (ViT-B/16) (Dosovitskiy et al., 2020). The text encoders are built upon BERT (Devlin et al., 2018), but we insert an additional cross-attention layer, which helps to fuse visual and linguistic information, between the self-attention layer and the feed-forward layer of each block. The text decoder is similar to the text encoder, except that we replace the self-attention layers with the causal self-attention layers to auto-regressively predict next tokens. In the following, we describe each of the components mentioned.

Image encoder: encodes input images and maps them into visual information representations. Concretely, each input image is first segmented into patches, and then ViT takes these patches as input and encodes them into a sequence of embeddings. The embeddings carry all visual information perceived from the image, and are finally mapped into the key matrix (K) and value matrix (V) for computing the cross-attention scores with the text.

Image-grounded text encoder: fuses the visual and linguistic information through the cross-attention layer of each transformer block. Specifically, the output of the self-attention layer in each block, which carries linguistic information obtained from the text input, is mapped into the query matrix (Q). Then, in each cross-attention layer, we use the query matrix (Q) together with the key matrix (K) and the value matrix (V), both coming from the ViT, to produce the output.

Multi-modal text encoder: has the same structure as the image-grounded text encoder. A special token is prepended to the beginning of the input text, and its output is used as the global representation of the fused visual and linguistic information.

Image-grounded text decoder: which is employed for performing VL generation downstream tasks (e.g., captioning). The causal self-attention layers enable the decoder to generate text in an auto-regressive manner. Specifically, a special token [DEC] is used as the start signal, and then the module iteratively generates the next token based on generated or supervised tokens in previous steps, until it reaches the end-of-sequence token.

We follow BLIP's design of parameter sharing between three branches to reduce model size with demonstrated performance gain. (Li et al., 2022)

3.2 Pre-training Objectives

KG-FLIP jointly optimizes three pre-training objectives: knowledge-guided masked language-image modeling (KG-MLIM), knowledge-guided image-text matching (KG-ITM), and language modeling (LM). Similar to BLIP, we use two understanding-based pre-training objectives and one generation-based pre-training objective. These three objectives activate different functionalities while contributing to each other through hard-parameter sharing. We first describe the three pre-training objectives of the model without the knowledge guidance (KG):

Masked Language-Image Modeling Loss (MLIM): is similar to MLM in pre-training language models (e.g., BERT), but it utilizes both the image and the contextual text to predict the masked tokens (Chen et al., 2022), which helps the model to learn cross-modal alignment at the token level

instead of instance level as in ITC. Formally, the MLIM loss can be represented by,

$$\mathcal{L}_{mlim} = - \mathop{\mathbb{E}}_{(I,T)\in\mathcal{D}} \mathop{\mathbb{E}}_{\mathcal{M}\subset T} \left[\sum_{t_i\in\mathcal{M}} \log p(t_i|I,\hat{T}) \right],$$

where one uniformly samples an image I and its corresponding text T from the dataset $\mathcal{D}$, masks a random token subset $\mathcal{M}$ from T, and predicts it given the image and the masked text $\hat{T}$.

Image-Text Matching Loss (ITM): aims to learn joint VL embeddings that effectively fuse the information from input image-text pairs. ITM facilitates the model to produce more effective and fine-grained VL representations by using these representations to judge whether image-text pairs are matched (positive pairs) or not matched (negative pairs). The ITM loss can be expressed as:

$$\mathcal{L}_{itm} = - \mathop{\mathbb{E}}_{(I,T)\sim p_{samp}(I,T|\mathcal{D})} \left[\log p(y_{I,T}|I,T)\right],$$

where p_{samp} is a distribution that samples positive and negative training examples, $y_{I,T} \in \{0,1\}$ represents whether the image I and the text T are matched, and $\log p(y_{i,T}|I,T)$ is the output of the [ENC] token in multi-modal text encoder followed by a classification layer.

Language Modeling Loss (LM): aims to auto-regressively generate desired textual information given an image (e.g., for captioning) or an image-text pair (e.g., for Visual Question Answering). It optimizes the loss,

$$\mathcal{L}_{lm} = - \mathop{\mathbb{E}}_{(I,T)\in\mathcal{D}} \left[\sum_{t_i\in T} \log p(t_i|I,T_{<i}) \right],$$

where each token t_i is predicted given the image I and all text tokens in T before position i.

3.3 Knowledge Guidance

To facilitate KG-FLIP to fuse two modalities more effectively, we utilize attribute knowledge to guide MLIM and ITM objectives, as described below:

Knowledge-guided MLIM (KG-MLIM): utilizes attribute information to guide MLIM by ameliorating the masking policy, as illustrated in Figure 2. The original policy of BERT (Devlin et al., 2018) uniformly chooses 15% of input tokens, of which 80% are replaced with a special masked token [MASK], 10% are replaced with a random textual token, and 10% remain unchanged. Rather than treating all tokens the same, masking product attribute words allows the VLP model to focus on learning salient product information and provide anchor points to align both modalities, thus producing more effective VL representations than the original 15% random masking policy.

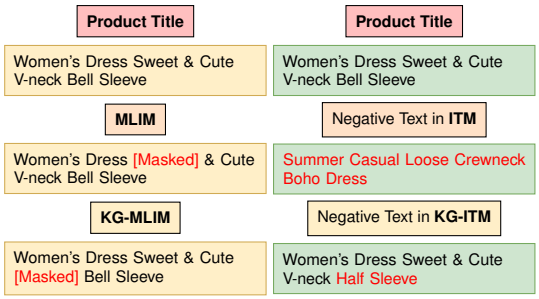

Figure 2: KG-MLIM vs. MLM, KG-ITM vs. ITM. (Left) Comparing to MLM which randomly selects 15% of words to mask, KG-MLIM prioritizes masking attribute words (e.g., crew neck, sleeveless). (Right) General ITM forms a negative pair by replacing the paired text with another text sample in the batch. By contrast, KG-ITM synthesizes a "harder" negative example by replacing the attribute word in the paired text with another value of the same attribute.

To this end, we propose to use knowledge (i.e., product attribute schema) to guide MLIM to mask significant attribute tokens rather than random-selected tokens. Concretely, we use the enumerated attribute values (e.g., "v-neck", "sleeveless") from the catalog system to identify significant words in the text that match our attribute value names. After that, we maintain an overall masking ratio of 15%, and if the number of detected significant attribute words exceeds 15%, we randomly select a subset of them to be masked. Otherwise, we randomly mask other tokens to fill up to 15%. In this way, we implement KG-MLIM, which enables VLP to focus on noteworthy attribute words.

Knowledge-guided ITM (KG-ITM): leverages attribute knowledge to synthesize "harder" negative image-text pairs, letting KG-FLIP determine whether the image-text pairs are matched or not matched. Specifically, in the standard ITM objective, p_{samp} typically utilizes the input image-text pairs in each batch as positive samples (Chen et al., 2022), and creates negative ones by replacing the image or text in each paired sample with randomly selected from other samples. The next step is to predict whether each image-text pair is matched. However, since images or text of different products are typically disparate, the negative samples are

usually too facile to train the model effectively.

Hence, we propose to leverage attribute knowledge to synthesize "harder" negative image-text pairs for the ITM loss. Similar to KG-MLIM, we use attribute values to search for salient attribute words in the text. If any attribute word in the text is detected, we synthesize a negative text string by replacing each identified word with another random attribute word from the same attribute class (e.g., "blue" → "red", "v-neck" → "crew neck"). Otherwise, if we do not spot any attribute word, we select a random text to construct the negative sample. Thus, these "more difficult" synthesized negative samples force KG-FLIP to produce more effective VL representations that capture subtle (i.e., fine-grained) distinctions between samples.

4 Experiments

4.1 Experimental Setup

We initialized all parameters with a BLIP checkpoint (Li et al., 2022), and then pre-trained KG-FLIP using a dataset of 1.9M pairs of Amazon product images and product texts (title and bullet points) in the fashion domain (viz., dresses and shoes). To investigate the potential promise of KG-FLIP, we tested KG-FLIP on two most common VL downstream tasks in e-commerce: we perform product attribute extraction on the Amazon product attribute dataset and product categorization on the Fashion-Gen dataset (Rostamzadeh et al., 2018), which we describe in detail below.

The Amazon product attribute dataset: contains a sample of products in our pre-training datasets that also have corresponding attribute values in the catalog. We further annotated another 600 image-title pairs as the validation and test set, which are used for hyper-parameter tuning and performance evaluation, respectively.

The Fashion-Gen dataset*: incorporates 293,008 fashion data pairs. The dataset contains 48 main categories (e.g., "Dresses", "Jeans") and 121 sub-categories (e.g., "Short Dresses", "Leather Jackets"). We tested KG-FLIP by performing the sub-category classification based on visual and linguistic modalities. In our experiments, we use the same training and testing data as used in Kaleido-BERT (Zhuge et al., 2021) and CMA-CLIP (Liu et al., 2021). The numbers of training and testing samples are 260,480 and 32,528, respectively.

4.2 Results

Product attribute extraction: The attribute-extraction task aims to automatically infer product attribute information (e.g., color, neck style) from product images and textual information such as title and description. Following (Liu et al., 2021), we formulate this problem as a multi-task classification task. We add a multi-layer perception (MLP) head for each attribute in Table 1 on top of the [ENC] output embedding from the multi-modal text encoder and fine-tune them simultaneously. We compare the results with CMA-CLIP, BLIP, and an unguided version of KG-FLIP, which was pre-trained with standard MLIM and ITM without knowledge guidance. All models are pre-trained and fine-tuned on the same datasets. Table 1 below shows the recall at 90% precision (R@90P) on the test set.

Table 1: Recall at 90% precision on the Amazon product attribute dataset. (attribute names are anonymized for compliance reasons)

Attribute	CMA-CLIP	BLIP	unguided KG-FLIP	KG-FLIP
dress attribute 1	29.1	53.1	52.1	**57.3**
dress attribute 2	42.3	41.0	**52.6**	48.7
dress attribute 3	57.3	61.1	65.9	**67.9**
dress attribute 4	33	36.7	**44.1**	42.1
dress attribute 5	71.5	65.1	71.8	**74.1**
shoe attribute 1	89.2	**94.6**	92.7	94.1
shoe attribute 2	90.0	**92.0**	91.0	**92.0**
shoe attribute 3	78.5	85.2	82.8	**85.6**
shoe attribute 4	98.7	**99.0**	98.7	**99.0**
Average	65.51	69.75	72.32	**73.42**

Product categorization: The task of product categorization is to automatically determine the sub-category for each product given its image-text pair. Similarly, we also formulate this problem as a classification task and stack an MLP head on top. Each VLP model in Table 2 was fine-tuned on the Fashion-Gen training set, and we then reported the accuracy of the categorization on the test set.

Overall, the results in Table 1 and Table 2 show that KG-FLIP outperforms all other VLP models on both datasets. For the Amazon product attribute dataset, KG-FLIP and unguided-FLIP offer performance gains of 3.67% and 2.57% in terms of R@90P, respectively, over BLIP. For the Fashion-Gen dataset, KG-FLIP can outperform the current benchmark (i.e., FashionViL) and BLIP in terms of accuracy by 2.1% and 0.36%, respectively. In

*Note that the Fashion-Gen dataset was only used to benchmark and illustrate the advance we made. It was not involved in building or optimizing our deployed model.

Table 2: Accuracy (%) on the Fashion-Gen dataset.

Method	Accuracy
ImageBERT (Qi et al., 2020)	80.11
FashionBERT (Gao et al., 2020)	85.27
OSCAR (Li et al., 2020)	84.23
KaleidoBERT (Zhuge et al., 2021)	88.07
CMA-CLIP (Liu et al., 2021)	93.60
FashionViL (Han et al., 2022)	92.23
BLIP (Li et al., 2022)	93.96
unguided KG-FLIP	94.15
KG-FLIP	**94.32**

summary, KG-FLIP has demonstrated its eminent performance, which makes it a compelling VLP solution for partially semantically aligned real-world VL data in e-commerce scenarios.

5 Model Deployment

Currently, we have successfully deployed our KG-FLIP model in a real-world application to backfill missing product attributes in the e-commerce catalog. E-commerce websites curate their product information in their catalog system. In addition to unstructured information (e.g., product titles and descriptions), structured product attributes (e.g., color and size) play an essential role in various downstream applications, including search and recommendation. For example, customers can filter search results by product attribute values and quickly identify their desired products. However, missing product attribute values are common, given the large number of products offered on e-commerce websites. Improving the coverage of product attributes with high accuracy is critical to improving the customer experience and maintaining customer trust. In addition, complete and accurate product attribute information can also improve the performance of various downstream applications (e.g., alternative product recommendations).

Compared to previous image-only and text-only models, KG-FLIP can infer product attributes from both modalities and increase precision and recall by large margins. Another advantage is that it can predict thousands of product attributes in a single model, which implies that model development and maintenance efforts are significantly reduced compared to single attribute models. However, training thousands of attributes in one model makes single-machine training infeasible because of the massive size of the training data. To overcome this challenge, we have developed our own distributed training infrastructure to support large-scale model training. Our infrastructure leverages the power of AWS Batch[†] Multi-Node Parallel, and the DeepSpeed framework, which allows us to automatically launch, configure, and manage a cluster of GPU instances, and train our model on 100 million image-text pairs for 10 epochs within a week with twenty p3.16xlarge instances. We also automated the process of launching a distributed job with just one command, which enables any individual to conduct distributed training tasks on their own and accelerates the experiment speed by reducing 90% of manual efforts. The model deployment is through AWS SageMaker[‡]. We leveraged AWS Batch to perform large-scale batch mode inference to backfill billions of product-attribute pairs with high accuracy since mid-2022.

6 Conclusion

In this paper, we introduced a knowledge-guided fashion-domain language-image pre-training framework for e-commerece, dubbed KG-FLIP. By utilizing the product attribute knowledge to guide MLIM and ITM pre-training objectives, our KG-FLIP model facilitates the vision-language pre-training and enhances the product representation learning for e-commerce data that are partially aligned while also containing complementary information. The evaluation results have demonstrated its prominent performance against other state-of-the-art benchmarks on both Amazon and Fashion-Gen datasets. The KG-FLIP model has been deployed in a real-world application and improved the customer shopping experience.

7 Limitations

There are two main limitations to this study. First, because of the lack of downstream datasets, we did not evaluate KG-FLIP on other downstream VL tasks in e-commerce (e.g., substitute recommendation). Therefore, the robustness of the KG-FLIP model on other downstream tasks requires further investigation. Second, the experimental results empirically show that the proposed knowledge-guided pre-training objectives are more effective in producing VL representations that capture subtle distinctions between samples than the standard objectives. However, a theoretical analysis of the effectiveness of our knowledge-guidance strategies is lacking.

[†] https://aws.amazon.com/batch/
[‡] https://aws.amazon.com/sagemaker/

8 Ethics Statement

We discuss ethical issues from these aspects:

Intended Use. If the technology is functioning as intended, both sellers and customers of e-commence platforms could benefit from the KG-FLIP model. KG-FLIP could help customers to quickly identify their desired products (e.g., by filtering search results by product attribute values). It could also help sellers by reducing their manual efforts when listing new products (e.g, the platforms can automatically recommend the attribute values).

Failure modes. In case of failure, KG-FLIP might output inaccurate product attribute information. Such non-factual information may harm customers' shopping experience. For example, the substitute recommendation system, which may use the incorrect product information provided by KG-FLIP, may recommend a non-desired product to our customers and hurt their shopping experience.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*.

Fei-Long Chen, Du-Zhen Zhang, Ming-Lun Han, Xiu-Yi Chen, Jing Shi, Shuang Xu, and Bo Xu. 2023. Vlp: A survey on vision-language pre-training. *Machine Intelligence Research*, 20(1):38–56.

Feilong Chen, Duzhen Zhang, Minglun Han, Xiuyi Chen, Jing Shi, Shuang Xu, and Bo Xu. 2022. Vlp: A survey on vision-language pre-training. *arXiv preprint arXiv:2202.09061*.

Kezhen Chen, Qiuyuan Huang, Yonatan Bisk, Daniel McDuff, and Jianfeng Gao. 2021. Kb-vlp: Knowledge based vision and language pretraining. In *Proceedings of the 38 th International Conference on Machine Learning, PMLR 139, 2021. ICML, workshop, 2021*.

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 *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX*, pages 104–120. Springer.

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

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*.

Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Pengchuan Zhang, Lu Yuan, Nanyun Peng, et al. 2022. An empirical study of training end-to-end vision-and-language transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18166–18176.

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.

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.

Wonjae Kim, Bokyung 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, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint arXiv:2201.12086*.

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-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer.

Huidong Liu, Shaoyuan Xu, Jinmiao Fu, Yang Liu, Ning Xie, Chien-Chih Wang, Bryan Wang, and Yi Sun. 2021. Cma-clip: Cross-modality attention clip for image-text classification. *arXiv preprint arXiv:2112.03562*.

Di Qi, Lin Su, Jia Song, Edward Cui, Taroon Bharti, and Arun Sacheti. 2020. Imagebert: Cross-modal pre-training with large-scale weak-supervised image-text data. *arXiv preprint arXiv:2001.07966*.

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.

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*.

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

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

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.