EI-CLIP: Entity-aware Interventional Contrastive Learning for E-commerce Cross-modal Retrieval

Haoyu Ma[1] *, Handong Zhao[2] †, Zhe Lin[2], Ajinkya Kale[2], Zhangyang Wang[3],
Tong Yu [2], Jiuxiang Gu [2], Sunav Choudhary [2], Xiaohui Xie [1]
[1]University of California, Irvine, [2]Adobe Research, [3]University of Texas at Austin
{haoyum3,xhx}@uci.edu, {hazhao,zlin,akale,tyu,jigu,schoudha}@adobe.com, atlaswang@utexas.edu

Abstract

Cross language-image modality retrieval in E-commerce is a fundamental problem for product search, recommendation, and marketing services. Extensive efforts have been made to conquer the cross-modal retrieval problem in the general domain. When it comes to E-commerce, a common practice is to adopt the pretrained model and finetune on E-commerce data. Despite its simplicity, the performance is sub-optimal due to overlooking the uniqueness of E-commerce multimodal data. A few recent efforts [10, 72] have shown significant improvements over generic methods with customized designs for handling product images. Unfortunately, to the best of our knowledge, no existing method has addressed the unique challenges in the e-commerce language. This work studies the outstanding one, where it has a large collection of special meaning entities, e.g., "Dissel (brand)","Top (category)", "relaxed (fit)" in the fashion clothing business. By formulating such out-of-distribution finetuning process in the Causal Inference paradigm, we view the erroneous semantics of these special entities as confounders to cause the retrieval failure. To rectify these semantics for aligning with e-commerce domain knowledge, we propose an intervention-based entity-aware contrastive learning framework with two modules, i.e., the Confounding Entity Selection Module and Entity-Aware Learning Module. Our method achieves competitive performance on the E-commerce benchmark Fashion-Gen. Particularly, in top-1 accuracy (R@1), we observe 10.3% and 10.5% relative improvements over the closest baseline in image-to-text and text-to-image retrievals, respectively.

1. Introduction

Cross visual and linguistic retrieval, as a fundamental component in the multimodal searching system, has been extensively studied [13, 18, 24, 27, 32, 38, 41, 43, 69, 70]. It takes linguistic data as the query and retrieves the corresponding visual data, or verse vice. One key challenge in this area is how to align the visual and textual data semantically.

Figure 1. Illustration of domain shift between general domain and e-commerce domain. In e-commerce domain, a collection of tag entities with strong domain semantics are associated with a title/description and image.

In the cross-modal retrieval of e-commerce products, there are many unique characteristics in both e-commerce image and language. As shown in Fig. 1, an e-commerce product image usually only contains a simple scene with one or two foreground objects and a plain background. Meanwhile, an e-commerce language is usually composed of a set of metadata (tag entities) [15, 39], including product title/description, brand, category, composition, etc. Previous works such as FashionBERT [10] and KaleidoBERT [72] suggest that cross-modal retrieval in fashion domains requires more fine-grained features (*e.g.* short sleeve and crewneck). However, the popular Region of Interest (RoI) [11] based methods detect unsatisfactory region proposals with either repeated object regions or irrelevant sub-regions to the product. To this end, these works focus on fine-grained representation learning of images through the patch-based method. Despite the great successes, they only focus on the challenges of images, while the language part

*Work done during the author's internship at Adobe Research.
†Corresponding author.

still follows the vanilla BERT [5].

In this work, we improve cross-modality product retrieval from the language part. Specifically, we design our model with the following two motivations about the unique language in e-commerce. **Motivation-1:** the word tokens often come up with special meanings in e-commerce, while the pretrained language model part in [10, 38, 72] is biased despite of the large-scale pretraining corpus. For instance, entity "`diesel`" in pretrained CLIP model is strongly associated with the concept "fuel", while in e-commerce fashion domain, "`diesel`" is tagged as a brand entity. Other examples include "`canada goose` (brand)", "`golden goose` (brand)", "`top` (category)", to name a few. Such *out-of-distribution* problem in multimodal finetuning is recently studied from the causal inference viewpoint [67]. Zhang *et al.* formulate this undesirable spurious correlations between image and language as "confounders" learned from the pretrained dataset. By modeling with structural causal model (SCM) graph [36], the authors perform hard intervention to remove the dataset bias via backdoor intervention [36]. However, when modeling the confounding variables, Zhang *et al.* follow the traditional BERT token vocabulary, treating each entity as a group of (sub)word tokens as others [10, 72]. This overlooks a large collection of special meaning entities in e-commerce, such as "`Dissel` (brand)","`top` (category)", "`relexed` (fit)". Moreover, this will inevitably intertwine different entities with the shared confounding (sub)word tokens, such as "`Canada Goose`" and "`Golden Goose`". To this end, the language part should be entity-aware [31, 47, 71] and disentangled from the conventional meanings of special entities encoded in the pretrained language model.

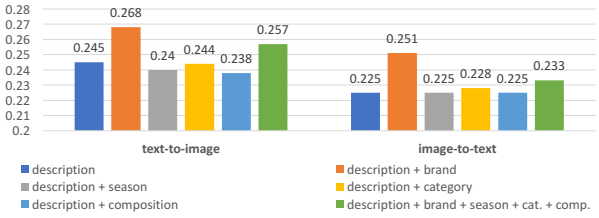

Figure 2. Empirical analysis of image-to-text and text-to-image tasks on Fashion-Gen. We finetune the pretrained CLIP model by concatenating different textual meta data. Results on top-1 accuracy are reported.

Meanwhile, the varieties of meta data leads to our **Motivation-2:** meta data contribute unevenly to the cross-modality retrieval. Specifically, previous methods usually concatenate all the metadata together to form a long sentence [10, 24, 38, 41, 43, 72]. However, this straightforward solution treats each meta information equally. In practice, for different image/text pairs, metadata (tag entity) may contribute differently. Some metadata can even be harmful to retrieval. To support the claim, we conduct an empirical study on Fashion-Gen dataset using a simple yet effective CLIP model [38]. We finetune the pretrained CLIP model given different meta entity concatenations on Fashion-Gen dataset. From Fig. 2, it is observed that given the product description (dark blue), "brand" (orange) is the only helpful metadata. Adding "category" (yellow), "season" (grey), or "composition" (light blue) can contribute little or even harm the performance. More importantly, if we concatenate all the meta data (green), both performances are dropped compared to only appending "brand" in text-to-image and image-to-text tasks. To this end, it is thus important to identify the beneficial metadata while discarding the others.

As motivated, we propose an Entity-aware Intervention-based contrastive learning framework, termed **EI-CLIP**, for e-commerce product retrieval problem with two specific module designs in the causal learning paradigm, *i.e., Entity-Aware Learning Module* (**EA-learner**) for **motivation-1** and *Confounding Entity Selection Module* (**CE-selector**) for **motivation-2**. It is worth clarifying that we do not propose a new causality method, but rather formulate the entity-aware e-commerce cross-modal retrieval problem in the casual view. Specifically, the EA-learner learns an individual representation for each informative confounding entity for better mitigating the out-of-distribution problem. Then the CE-selector aims to automatically select the most informative group of meta data (*e.g.,* "brand" in Fig. 2) from the abundant textual meta data.

We summarize our main contributions as follows:

- To the best of our knowledge, this is the pioneering work to tackle the challenges introduced by e-commerce special entities in language modality. Previous cross-modal retrieval works only focus on images.
- We are the first to formulate the entity-aware retrieval task in causal view. We argue that the erroneous semantics of e-commerce special entities learned in the general domain are the *confounders* causing the retrieval failures.
- Equipped with backdoor adjustment [36] in causal inference, we propose an Entity-aware Intervention-based contrastive learning framework (**EI-CLIP**), with two new components, *i.e.,* CE-selector and EA-learner.
- **EI-CLIP** achieves competitive performance on e-commerce benchmark dataset Fashion-Gen. In particular, in top-1 accuracy (R@1), we observe 10.3% and 10.5% relative improvements over the closest baseline in image-to-text and text-to-image, respectively.

2. Related Work

Image-Text Matching Visual-linguistic representation learning has many downstream applications including im-

age caption, visual question answering, cross-modal retrieval (image-text matching), etc. Our work is closely related to image-text matching, where the key problem is how to semantically align the image and text. The early works start from exploiting the shallow models to project the entire image and sentence into the latent subspace, then align two modalities in image/sentence level [13, 19]. In the recent decade, deep models (*e.g.,* convolutions neural network for image and long short-term memory network [16] for sentence) have been widely applied to extract better representation, then make image/sentence level alignment via canonical correlation analysis [41, 55], ranking loss [9, 20], hard example mining [3, 8], etc. To achieve a fine-grained level alignment, attention mechanism has been incorporated to align word/region tokens with different levels of granularity, such as word level [18, 22, 52], phrase/relation level [24, 49], etc. Recently, with the great success of Transformer-based pretraining [5, 48], many visual linguistic pretraining methods have been proposed, such as VL-BERT [43], ViLBERT [32], VideoBERT [44], LXMERT [45], Unicoder-VL [23], OSCAR [25], etc. Most recently, with the development of vision transformers [7, 29, 33, 46, 56, 62], Radford *et al.* [38] introduce a simple yet powerful multimodal pretraining framework (CLIP) based on contrastive learning [4, 12, 12, 14, 59–61] on a 400-million image-text paired training corpus. Although no word/region level alignment mechanism is specifically designed, it has shown a superior capability on word/region token level alignment to other methods. Our work follows the CLIP framework.

Fashion-Based Cross-Modality Retrieval Compared with the general vision-language domain, the fashion-based task requires paying more attention to the task-specific knowledge, such as the fine-grained information [6, 10, 64, 72]. FashionBERT [10] was the first vision-language model in the fashion domain. It proposed a patch-based method to retain more raw pixel-level information. Then the split non-repeated patches together with query word tokens are fed to the cross-modal BERT model for joint learning. Later on, Kaleido-BERT [72] further applies several self-supervised tasks at different scales to focus more on image-text coherence. However, all of these methods only focus on the visual part, while ignoring the uniqueness of e-commerce language. Our work aims to solve challenges from the language modality.

Causality in Multimodal Learning Causal inference has been successfully explored in a number of vision and language applications, such as image classification [2, 30, 58, 63], semantic segmentation [65], video action localization [26, 28, 57] in vision, and text classification [53], text question answering [42], named entity recognition [68] in language. This work focuses on multimodal learning, where a few existing works have touched upon. Wang *et al.* [50, 51] propose a visual commonsense region-based convolutional neural network (VC R-CNN) to deal with the spurious correlation within objects in image. Despite that the de-confounded VC R-CNN shows encouraging results on many multimodal applications, the causal intervention is only considered for the visual domain. Zhang *et al.* [67] study the spurious correlations in multimodal pre-trained models when applied to an out-of-distribution finetuning task. The core idea of the proposed DeVLBert is to employ hard intervention to back-door adjust [36] the conditional probability of an object token (in visual) given a word token (in language), or verse vice. In this work, we are also interested in a similar problem motivated by the practical challenge of adapting a pre-trained generic multimodal model to an out-of-distribution downstream e-commerce dataset. While apart from the task difference, *i.e.* generic multimodal representation learning (DeVLBert) vs. specific cross-modality retrieval (ours), our work aims to mitigate the bias semantics of special entities, while previous works focus on the correlations among objects.

3. Methodology

3.1. Revisiting CLIP

Radford *et al*. [38] suggest that the predetermined object categories provide limited supervision to the computer vision networks. Instead, directly learning from the raw text description about an image is an effective way that leverages rich supervision information. [38] proposes the CLIP (Contrastive Language-Image Pre-training) model, which applies contrastive learning to learn visual representations from scratch on a dataset of 400 million image-text pairs. Specifically, given a batch of image-text pairs $\{(I_i, T_i)\}_{i=1}^{N}$, where N is the batch size, the image encoder $h^I(\cdot)$ and the text encoder $h^T(\cdot)$ firstly encode the image and text into vectors on the multi-modal embedding space $\mathcal{R}^d$, where d is the dimension of the embedding. Denote the image embedding and text embedding as $E_i^I = h^I(I_i)$ and $E_i^T = h^T(T_i)$, respectively. As shown in Fig. 3 (a), during the training, the CLIP model calculates the cosine similarity $E_i^T \odot E_j^I$ $(i, j \in \{1, 2, ..., N\})$ of all $N \times N$ possible pairs. To jointly train the image and text encoders, CLIP maximizes the similarity for N matched pairs while minimizing the similarity for all other $N^2 - N$ unmatched pairs. In practice, CLIP optimizes a symmetric cross-entropy loss over the $N \times N$ similarity scores matrix.

CLIP calculates the similarities only based on the global embedding of images and texts. Therefore, it only learns the correspondence between word tokens and detailed image features implicitly. To this end, a sufficiently large dataset is required to learn this fine-grained correspondence during the pre-training process. The CLIP constructs a dataset with 400 million image-text pairs on the Internet. However, the model is easily biased towards the '*commonsense*' knowl-

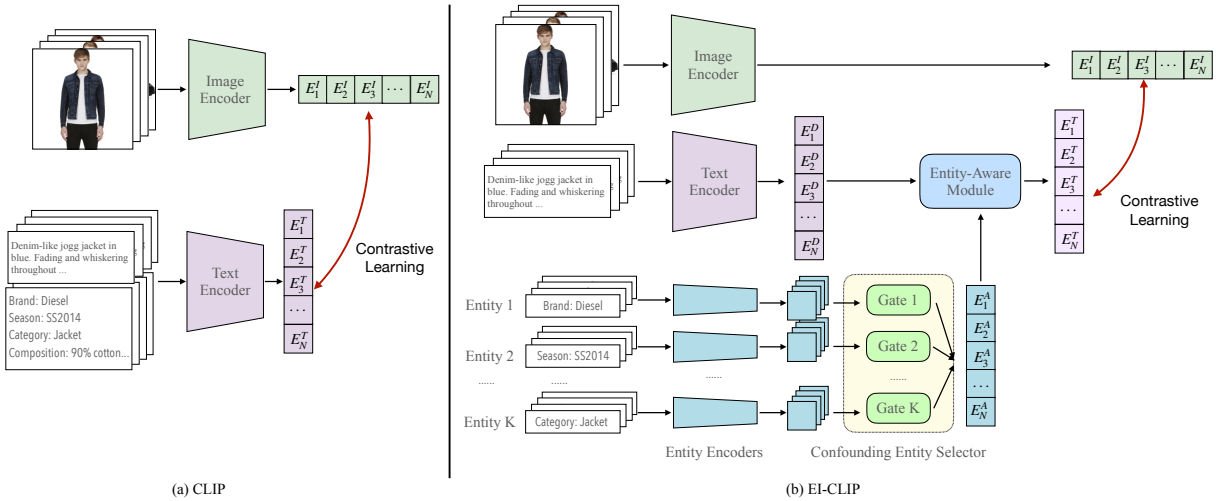

Figure 3. Comparison of CLIP (a) and our EI-CLIP (b) during the training. The CLIP (a) directly appends all entities to the text description, while our EI-CLIP encodes them individually. We further apply the CE-Selector to automatically select the significant confounding entities, and the EA-Learner to introduce the entity information into the description. Please refer to Section 3.3 for more details.

edge[1] when learning from this large dataset in the general domain. Typically, the bias towards the common domain is beneficial. However, when adapting it to other domains with contrastive learning, it is challenging for the model to learn all the domain-specific knowledge well, such as the knowledge of the e-commerce domain. For example, the word "`diesel`" typically refers to the "diesel fuel" in the commonsense. However, it is a brand of clothes in the fashion domain. Another example is the brand "`golden goose`". In commonsense, we treat them as two separate words "`golden`" and "`goose`", and refer them to the color and the animal. Considering the limited number of fine-tuning examples, it is difficult for the model to learn that these special words refer to brands. Thus, the model still maintains its commonsense knowledge about these words. Consequently, a method to mitigate the erroneous semantics of these unique words in CLIP is required.

3.2. CLIP in the Causal View

In the causal view, we regard the given text T_i as X and image I_i as Y. For CLIP, the goal of contrastive learning is to learn both function $P(Y|X)$ and function $P(X|Y)$. We use the calculation of $P(Y|X)$ as an example to illustrate the causal view. We consider the semantics of these special entities as confounders Z, which may affect either X or Y. Specifically, we define $z = g(a, b)$, which means entity a takes the semantics b. The entity a usually maintains several semantics and is part of the text X. For example, $g(\texttt{golden goose}, \textit{"animal"})$ means the word '`golden goose`" refers to an animal, while $g(\texttt{golden goose}, \textit{"brand"})$ refers to the brand. The confounders may introduce spurious correlations in the model when only learning from $P(Y|X)$. Formally, by the Bayes Rule, the likelihood can be written as [36]:

$$P(Y|X) = \sum_z P(Y, z|X) = \sum_z P(Y|X, z)\underline{P(z|X)}, \tag{1}$$

where the confounder z introduces the bias of the training set via $P(z|X)$. As the CLIP is trained in the general domain, it is easily biased towards the commonsense. Given the text X = *"A T-shirt of golden goose"*, most of the likelihood sum in Eq. 1 will be assigned to $P(Y|X, z = g(\texttt{golden goose}, \textit{"animal"}))$, since $P(z = g(\texttt{golden goose}, \textit{"animal"})|X)$ is large in the general domain. Thus, when adapting to the fashion domain, the function $P(Y|X)$ tends to retrieve an image with goose or golden color, rather than retrieve the clothes of the corresponding brand.

To adjust the influence of confounder Z in other domain, we intervene X with the *do*-calculus [50, 67]. Specifically, we cut off the dependency between X and Z. By the definition of *do*-calculus, we have

$$P(Y|do(X)) = \sum_z P(Y|X, z)\underline{P(z)}. \tag{2}$$

Compared with Eq. 1, z is no longer affected by X. The prediction of Y is subject to the prior $P(z)$ of the training set, which can be easily pre-calculated [50, 67]. In fashion domain, the prior $P(z = g(\texttt{golden goose}, \textit{"brand"})$ dominates the likelihood. Thus, the bias towards commonsense in general domain can be mitigated.

[1] Commonsense can be biased. For example, "banana is yellow" is commonsense, which is not necessary. Instead, bananas can be red or green. Other biased cases in CLIP are discussed in [1].

3.3. EI-CLIP: Implementation

In the e-commerce domain, the text T_i consists of two components: one is the text description T_i^D, which describes the details of product. The other is the entity set $T_i^A = \{a_i^k\}_{k=1}^K$, where K is the total number of entities, and a_i^k is the k-th entity. There are usually some meta data (tag entities) about the product, such as brand and category, which represents domain-specific knowledge. To tackle these challenging entities, we propose the *EI-CLIP*, as shown in Fig. 3(b). Specifically, we design two modules to implement $P(Y|do(X))$. One is *Entity-Aware Learning Module* (EA-Learner), and the other is *Confounding Entity Selection Module* (CE-Selector).

EA-Learner This module aims to explicitly capture each individual entity information without worrying about ambiguous entity semantics between general and e-commerce domain or intertwined entity representation because of shared (sub)word tokens (**motivation-1** in Section 1). The contrastive learning is formed as a classification task within a mini-batch. We denote $j \in \{1, 2, ..., N\}$ as the index in the mini-batch. Therefore, the prediction $P(Y|X, z)$ in Eq. 2 can be regarded as a classifier: $P(Y|X, z) = \text{Softmax} f_j(X, z)$, where $f_j(X, z)$ denotes the classification head of intervention. Similar to [50, 67], with the approximation of NGSM (Normalized Weighted Geometric Mean) [54], Eq. 2 can be implemented as:

$$P(Y|do(X)) \approx \text{Softmax}[\mathbb{E}_z(f_j(X, z))]. \tag{3}$$

By definition, we have $z = g(a, b)$. Thus, $\mathbb{E}_z(f_j(X, z))$ in Eq. 3 can be written as:

$$\begin{aligned}\mathbb{E}_z(f_j(X, z)) &= \sum_z f_j(X, z)P(z) \\ &= \sum_a \sum_b f_j(X, z = g(a, b))P(z = g(a, b)).\end{aligned} \tag{4}$$

In practice, $P(z = g(a, b))$ can be approximated by counting the frequency of all semantics b for a given entity a in the training set. For simplicity, we assume that there is only one special entity a_i in text description T_i (*i.e.* variable X) in the fashion domain. The entity a_i maintains multiple semantics $b_{i,m} \in \mathbb{B}_i^1 \cup \mathbb{B}_i^2$, where $\mathbb{B}_i^1$ contains all semantics of a_i in the general domain, $\mathbb{B}_i^2$ contains the special semantics of a_i in the fashion domain, and m is the index of semantics in the set $\mathbb{B}_1 \cup \mathbb{B}_2$. When $b_{i,m} \in \mathbb{B}_1$, a_i refers to the general semantics. However, note that in our fashion retrieval problem, a_i is already marked as special entity and assigned to one unique semantic (*e.g.* `golden goose` as "brand") in meta data. Thus, the probability $P(z = g(a_i, b_{i,m}))$ is 0 when $b_{i,m} \in \mathbb{B}_1$. To this end, we only need to consider the case of $b_{i,m} \in \mathbb{B}_2$. As the semantic is unique, we train an entity encoder $h^A(\cdot)$ from scratch to learn entity embedding: $E_i^A = h^A(a_i) \in \mathcal{R}^d$. a_i is processed as a whole, instead of multiple (sub)word tokens. Meanwhile, we apply the text encoder $h^T(\cdot)$ to obtain the embedding of T_i^D: $E_i^D = h^T(T_i^D)$. As concluded in [17, 38], there are linear relationships [34] within the multi-modal embeddings. In this respect, we obtain the global embedding of T_i through $E_i^T = E_i^D + E_i^A$. In this case, $f_j(X, z)$ is parameterized by $E_j \odot (E_i^D + E_i^A)$. Thus, Eq. 3 can be rewritten as:

$$P(Y|do(X)) \approx \text{Softmax}\left[E_j^I \odot \left(E_i^D + E_i^A\right)\right]. \tag{5}$$

With this design, the language part of CLIP is aware of the unique semantics of these entities and disentangled from their general semantics encoded in the pretraining process.

CE-Selector As there are K entities with special semantics, a common practice is concatenating all entities with the text description T_i^D at the raw string level. However, as shown in Fig. 2, this naive approach does not generalize well, since not all groups of confounding entities are informative, and some confounders are even harmful. With **motivation-2** in Section 1, the CE-selector aims to select the important entities, whose semantics are unique and informative in retrieving images in fashion domain.

As K entities belong to different groups, such as brand and category, we learn K separate entity encoders $h_k^A(\cdot)$. Once obtained the embedding of all entities $\{h_k^A(a_i^k)\}_{k=1}^K$, we follow the gating mechanism [66] and design a gate network $G_k(\cdot)$ to determine the importance of each groups of entities and select useful confounders. Specifically, the selection factor w_k can be defined as $w_k = G_k(h_k^A(a_i^k))$. We implement $G_k(\cdot)$ with an MLP layer and a sigmoid function to ensure the value of w_k is in the range of $(0, 1)$. We further fuse them together into one global entity embedding $\hat{E_i^A}$ by $\hat{E_i^A} = \sum_k w_k \cdot h_k^A(a_i^k)$. Thus, with multiple entities in $T_i^A = \{a_i^k\}_{k=1}^K$, Eq. 3 is implemented by:

$$P(Y|do(X)) \approx \text{Softmax}[E_j^I \odot (E_i^D + \sum_k w_k \cdot h_k^A(a_i^k))]. \tag{6}$$

Training To avoid the commonsense bias affecting the learning of $h_k^A(\cdot)$, we disentangle $h_k^A(\cdot)$ and the pre-trained $h^T(\cdot)$ during the training. In detail, besides the contrastive loss between E_j^I and E_i^D + $\hat{E_i^A}$, we also calculate the contrastive loss between E_j^I and E_i^D, and the contrastive loss between E_j^I and $\hat{E_i^A}$ simutaneously.

4. Experiments

4.1. Settings

Datasets Following FashionBERT [10] and Kaleido-BERT [72], we evaluate our method on the Fashion-Gen dataset [39]. There are $67,666$ fashion products. Each product holds one text description and one to six images from different angles. In detail, $260,480$ and $35,528$ image-text pairs

are used for training and testing, respectively. There are 4 groups of entities for each product. Specifically, they are `Brand`, `Sub-Category`, `Season`, and `Composition`. We use `B`, `C`, `S`, and `P` to denote them. In total, there are 570 brands, 122 sub-categories, 10 seasons, and 16,844 types of compositions among all products. Besides, we create a new dataset upon the subset of Amazon reviews [35]. It contains 20,507 image-text pairs in the fashion domain. We use 14,354 pairs for training and 6,153 for testing. We only use the 184 brands as the special entities. The text description of this dataset is conciser and more ambiguous than Fashion-Gen, which makes it more challenging.

Implementation Details We start from the released pre-trained CLIP model [38], which applies ViT-B [7] as the visual encoder $f_I(\cdot)$ and the Transformer [48] as the text encoder $f_T(\cdot)$. The input image is resized to 224×224, and the input text description is represented by the lower-cased byte pair encoding (BPE) [40] with a 49,152 vocab size. The entity encoder $h_k^A(\cdot)$ is implemented by one embedding layer and one MLP layer. The embedding dimension d is set to 512. Following [10, 72], the Adam optimizer with weight decay $1e-4$ is applied to finetune the pre-trained CLIP model. The total number of finetuning epochs is set to 20. The initial learning rate is set to $5e-5$, and the cosine annealing learning rate decay scheduler is applied. We also adopt a warming-up strategy for the first 1K steps.

Evaluation We evaluate our method on image-to-text (I2T) retrieval and text-to-image (T2I) retrieval in e-commerce. Given a query in one modality, this task requires retrieving the matched item in the other modality from the candidate rank set. Given a text description (or image), the positive candidate is the associated ground-truth image (or text description) from the same product. For the negative candidates, we consider two kinds of settings. 1) Following [10, 72], we randomly sample 100 images (or text description) from other products within the same sub-category. We denote this sampling strategy as "Sample 100". 2) We also consider the entire product set as our candidate set (denote as "Full candidate"), which is a more challenging setting. It is more in line with actual product retrieval scenarios and has been widely adopted in the product recommendation field [21]. We use Rank@1 (Top-1 accuracy), Rank@5, Rank@10 to evaluate the performance of both retrieve tasks. Following [72], $SumR$=(Rank@1+Rank@5+Rank@10)*100 is regarded as an overall metric of the model.

4.2. Effectiveness of EI-CLIP

We consider several baselines to verify the effectiveness of EI-CLIP. In detail, these models are: ① The pre-trained CLIP released by [38]; ② We finetune the CLIP with image I_i and only description T_i^D of each product; ③ We finetune the CLIP with image I_i and the combination of description and all entities at the raw string level; ④ EI-CLIP, which sets the weight w_k of all entities equally; ⑤ EI-CLIP.

Table 1. Retrieval performances (Full candidate) on Fashion-Gen. ↑ means the relative improvement.

	Image-to-text			Text-to-image			$SumR$
	R@1	R@5	R@10	R@1	R@5	R@10	
①	9.4	24.5	33.5	10.7	26.8	35.8	141
②	22.5	49.5	62.0	24.5	51.1	63.6	273
③	23.3	51.5	64.6	25.7	53.9	66.5	285
④	25.2	52.6	64.8	28.2	56.6	68.4	296
⑤	**25.7**	**54.5**	**66.8**	**28.4**	**57.1**	**69.4**	**302**
↑	10.3%	5.8%	3.1%	10.5%	5.9%	4.4%	6.0%

Quantitative Results Results are presented in Table 1. Firstly, the pre-trained CLIP model ① does not generalize well on the fashion domain. Thus, fine-tuning is necessary to mitigate this gap. Secondly, the improvement from ② to ③ suggests that the information of entities is beneficial for retrieval. Thirdly, ③ performs much better than ④. It implies that our EA-learner can learn the semantics of special entities better (w.r.t. **motivation-1**). Furthermore, we can obtain a further improvement at all levels of recalls of the retrieval with the selector. As the results of ⑤, the relative improvement is 10.3% on R@1 of I2T (from 23.3% to 25.7%), and 10.5% on R@1 of T2I (from 25.7% to 28.4%). As retrieving the correct product from the full candidate is much more challenging, our improvement is enormous. This supports the success of CE-selector (w.r.t. **motivation-2**). For the gating mechanism, we find that the learned w_k of `Season` is around as small as 0.05. This also supports that our method can automatically select important confounders and remove harmful confounders.

Qualitative Results We present examples of cross-modal retrieval for model ③ (baseline) ⑤ (ours) in Fig. 4. In example (a) and (b), brands consist of multiple words. Since the baseline split them into several tokens, it is hard to understand their semantics. Our method can recognize these words and retrieve the image correctly. In example (c), the results from baseline are similar to the unit of a diesel worker, but the query requires a "black" jacket instead. In example (d), the color of the top-1 result of baseline is golden write, but the query requires the heel collar to be "golden-tone". These cases suggest that the fine-tuned model still maintains the general semantics of these words and cannot learn the domain knowledge. However, ours can solve them correctly and learn these words as the brand. Besides, as in examples (e) and (f), our model still performs better on brands like "`Gucci`" and "`Burberry`", which are not biased towards commonsense. We hypothesize that the limited amount of these words in the general domain make it difficult to learn them well. Our method can miti-

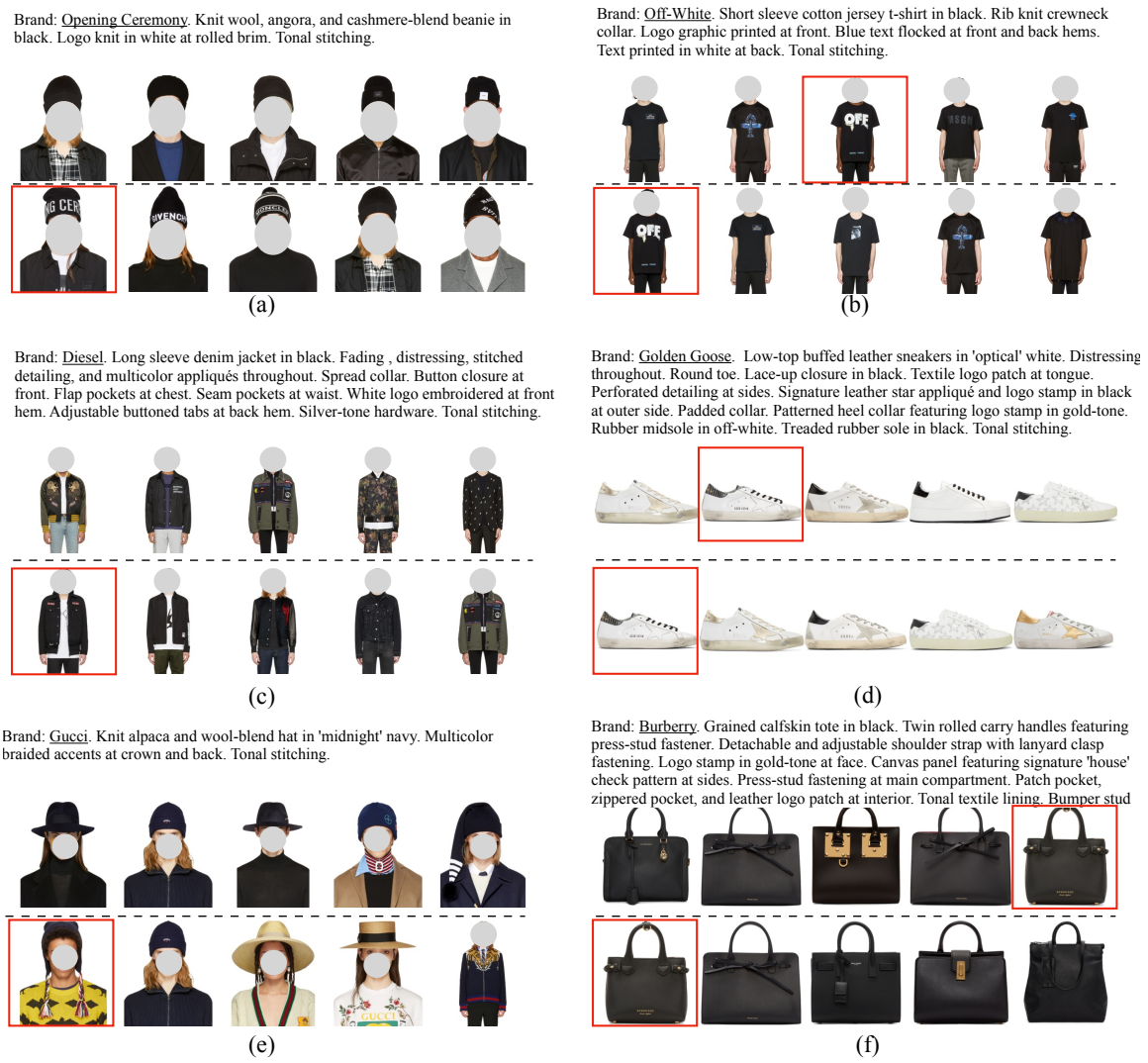

Figure 4. Examples of text-to-image results. For each example, the first row is the query text, the second row is the top-5 retrieved results of fine-tuning CLIP (model ③), and the third row is the top-5 results from EI-CLIP. The correct answers are boxed in red.

gate the distribution gap yet. More examples of I2T and T2I are shown in Supplementary.

4.3. Compare with state-of-the-art methods

We then compare our method with previous state-of-the-art works on Fashion-Gen. As in [72], FashionBERT [10] and Kaleido-BERT [72] already beat all previous multi-modal learning networks including ImageBERT [37], OSCAR [25], VLBEERT [43], and ViLBERT [32] by a large margin. Thus, we only focus on the comparison of our work with FashionBERT and Kaleido-BERT. We follow the "Sample 100" strategy [10, 72] to obtain the candidate set for a fair comparison. As all candidates belong to the same sub-category, we discard the category entity in evaluation. Besides, we also reproduce previous works with entities. We still evaluate model ②, ③, and ⑤ in Section 4.2, but use different candidate sets. The results are shown in Table 2. Firstly, the fine-tuned vanilla CLIP achieves a clear improvement compared with previous methods, whether with or without the entities. We believe that contrastive learning helps the model smoothly learn the ability to identify an input one modality with a bunch of inputs in the other modality. Secondly, EI-CLIP still brings some further improvement, although it is relatively marginal. Note that, in this evaluation, all 100 negative samples belong to the same category, making the category entity useless to distinguish between ground-truth and negative candidates. Besides, this is an easier evaluation as the candidate size is small.

4.4. Results on Amazon-Review

We further evaluate our method on Amazon-Review. As only the brand entity serves as the confounder, we no

Table 2. Cross-modal retrieval performances (Sample 100) on Fashion-Gen. The reported $SumR$ of Fashion-BERT [10] and Kaleido-BERT [72] are 251.36 and 319.52, respectively. Methods marked with '*' are results from our reproduction. FBERT stands for Fashion-BERT.

Methods		FBERT*	FBERT*	CLIP	CLIP	EI-CLIP
With Entities?		No	Yes	No	Yes	Yes
I2T	R@1	31.37	35.30	36.11	**39.17**	38.70
	R@5	62.97	68.44	67.81	71.26	**72.20**
	R@10	75.20	82.34	80.00	83.69	**84.25**
T2I	R@1	24.09	31.06	35.32	38.61	**40.06**
	R@5	54.73	63.95	65.98	69.69	**71.99**
	R@10	69.44	78.68	77.84	82.23	**82.90**
	$SumR$	317.8	359.8	363.1	384.7	**390.1**

longer need the CE-selector. We conduct the full candidate retrieval and the results are shown in Table 3. Our method still outperforms CLIP on the challenging Amazon dataset, showing that EI-CLIP generalizes well on another e-commerce scenario.

Table 3. Performances on Amazon-Review (Full-candidate).

	Image-to-text			Text-to-image			$SumR$
	R@1	R@5	R@10	R@1	R@5	R@10	
CLIP	22.2	49.9	61.6	23.5	48.9	61.5	267
EI-CLIP	25.9	54.2	65.3	23.7	49.4	61.6	280

4.5. Ablation Studies

Entity set As different entities play different roles, we then explore the contribution of each attribute on Fashion-Gen. As shown in Fig. 2 and Table 4, at raw string level, different attributes bring various improvement. The brand can bring a remarkable improvement, while season, sub-category, and compositions can only bring slight improvement or even hurt the performance. Besides, the performance of the combination of all attributes (Experiment (c)) is even worse than only using the brand attribute individually. This suggests that the naive strategy cannot fully utilize entities. We hypothesize that other appending all attributes together may introduce much noise in the raw text, and thus disturb the learning of self-attention modules.

Batch Size We also explore the influence of batch size, as it heavily affects the performance of contrastive learning [4, 38]. Typically, a larger batch size brings better performance, but it requires a larger GPU memory. We vary the batch size from 16 to 128, and plot the results of R@1 in Fig. 5. At all levels of batch size, our EI-CLIP consistently beats the baseline fine-tuning CLIP. Moreover, the improvement is more obvious on small batch size settings, thus it is more beneficial for users with limited GPU memory.

Table 4. Ablation studies of each type of entity.

EXPER.	Image-to-text			Text-to-image			$SumR$
	R@1	R@5	R@10	R@1	R@5	R@10	
②	22.5	49.5	62.0	24.5	51.1	63.6	273
③	23.3	51.5	64.6	25.7	53.9	66.5	285
Str-B	25.1	53.0	65.8	26.8	54.8	67.4	293
Str-C	22.8	50.0	62.5	24.4	51.9	64.1	276
Str-S	22.5	49.4	61.8	24.0	50.9	63.3	272
Str-P	22.5	48.6	61.3	23.8	50.0	63.1	269
⑤	25.7	54.5	66.8	28.4	57.1	69.4	302
Emb-B	25.6	53.0	65.5	27.8	55.0	67.2	294
Emb-C	23.3	50.3	63.0	24.9	51.4	64.3	277
Emb-S	20.4	46.5	59.0	24.7	50.7	63.5	264
Emb-P	22.2	49.4	61.8	24.5	51.1	63.6	273

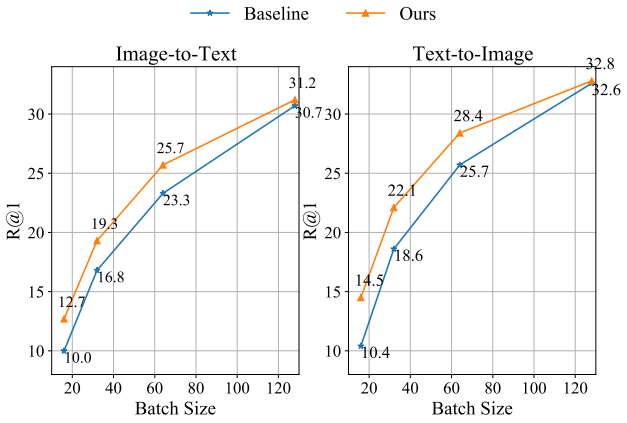

Figure 5. Comparison of ours (EI-CLIP) and baseline (Experiment (c) in Section 4.2) with different batch sizes.

5. Limitations

One potential limitation is that we only consider the semantics of entities from the meta data of products as the confounders. This assumption simplifies the design of the network and clearly demonstrates the benefits of our network. However, in practice, any hidden variables can be the confounders to affect the learning of $P(Y|X)$. Besides, not all e-commerce products contain clean meta information as Fashion-Gen. Handling a set of noisy meta data is out of this paper's scope, but could be a potential challenge.

6. Conclusion

In this paper, we first point out that the bias of common knowledge limits the generalization ability of the CLIP model when fine-tuning on the e-commerce domain. To alleviate this issue, we follow the theory of causal intervention and propose EI-CLIP. Specifically, we consider the entities of products from the meta data as the confounder and encode them separately with independent networks. Extensive experiments demonstrate that our method achieves better performance and focuses more on the semantics of special entities in the e-commerce domain.

References

[1] Sandhini Agarwal, Gretchen Krueger, Jack Clark, Alec Radford, Jong Wook Kim, and Miles Brundage. Evaluating clip: towards characterization of broader capabilities and downstream implications. *arXiv preprint arXiv:2108.02818*, 2021. 4

[2] Krzysztof Chalupka, Pietro Perona, and Frederick Eberhardt. Visual causal feature learning. In Marina Meila and Tom Heskes, editors, *UAI*, pages 181–190. AUAI Press, 2015. 3

[3] Tianlang Chen, Jiajun Deng, and Jiebo Luo. Adaptive offline quintuplet loss for image-text matching. In *ECCV*, pages 549–565, 2020. 3

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, pages 1597–1607, 2020. 3, 8

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *ACL*, 2019. 2, 3

[6] Eric Dodds, Jack Culpepper, Simao Herdade, Yang Zhang, and Kofi Boakye. Modality-agnostic attention fusion for visual search with text feedback. *arXiv preprint arXiv:2007.00145*, 2020. 3

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 3, 6

[8] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. *BMVC*, 2017. 3

[9] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. *NIPS*, 2013. 3

[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 *SIGIR*, pages 2251–2260, 2020. 1, 2, 3, 5, 6, 7, 8

[11] Ross Girshick. Fast r-cnn. In *ICCV*, pages 1440–1448, 2015. 1

[12] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *NeurIPS*, pages 21271–21284, 2020. 3

[13] David R Hardoon, Sandor Szedmak, and John Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural computation*, 16(12):2639–2664, 2004. 1, 3

[14] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, pages 9729–9738, 2020. 3

[15] Ruining He and Julian McAuley. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *WWW*, pages 507–517, 2016. 1

[16] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 3

[17] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. *ICML*, 2021. 5

[18] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, pages 3128–3137, 2015. 1, 3

[19] Tae-Kyun Kim, Josef Kittler, and Roberto Cipolla. Discriminative learning and recognition of image set classes using canonical correlations. *TPAMI*, 29(6):1005–1018, 2007. 3

[20] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*, 2014. 3

[21] Walid Krichene and Steffen Rendle. On sampled metrics for item recommendation. In *KDD*, pages 1748–1757, 2020. 6

[22] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *ECCV*, pages 201–216, 2018. 3

[23] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *AAAI*, pages 11336–11344, 2020. 3

[24] Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. Visual semantic reasoning for image-text matching. In *ICCV*, pages 4653–4661. IEEE, 2019. 1, 2, 3

[25] 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. 3, 7

[26] Yicong Li, Xun Yang, Xindi Shang, and Tat-Seng Chua. Interventional video relation detection. In *ACMMM*, pages 4091–4099, 2021. 3

[27] Fenglin Liu, Xian Wu, Chenyu You, Shen Ge, Yuexian Zou, and Xu Sun. Aligning source visual and target language domains for unpaired video captioning. *TPAMI*, 2021. 1

[28] Yuan Liu, Jingyuan Chen, Zhenfang Chen, Bing Deng, Jianqiang Huang, and Hanwang Zhang. The blessings of unlabeled background in untrimmed videos. In *CVPR*, pages 6176–6185, 2021. 3

[29] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, pages 10012–10022, 2021. 3

[30] David Lopez-Paz, Robert Nishihara, Soumith Chintala, Bernhard Scholkopf, and Léon Bottou. Discovering causal signals in images. In *CVPR*, pages 6979–6987, 2017. 3

[31] Di Lu, Spencer Whitehead, Lifu Huang, Heng Ji, and Shih-Fu Chang. Entity-aware image caption generation. *EMNLP*, 2018. 2

[32] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*, 2019. 1, 3, 7

[33] Haoyu Ma, Liangjian Chen, Deying Kong, Zhe Wang, Xingwei Liu, Hao Tang, Xiangyi Yan, Yusheng Xie, Shih-Yao Lin, and Xiaohui Xie. Transfusion: Cross-view fusion with transformer for 3d human pose estimation. *BMVC*, 2021. 3

[34] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013. 5

[35] Jianmo Ni, Jiacheng Li, and Julian McAuley. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *EMNLP*, pages 188–197, 2019. 6

[36] Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge, UK, 2009. 2, 3, 4

[37] Di Qi, Lin Su, Jia Song, Edward Cui, Taroon Bharti, and Arun Sacheti. Imagebert: Cross-modal pre-training with large-scale weak-supervised image-text data. *arXiv preprint arXiv:2001.07966*, 2020. 7

[38] 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 *ICML*, pages 8748–8763, 2021. 1, 2, 3, 5, 6, 8

[39] Negar Rostamzadeh, Seyedarian Hosseini, Thomas Boquet, Wojciech Stokowiec, Ying Zhang, Christian Jauvin, and Chris Pal. Fashion-gen: The generative fashion dataset and challenge. *arXiv preprint arXiv:1806.08317*, 2018. 1, 5

[40] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. *ACL*, 2016. 6

[41] Jie Shao, Leiquan Wang, Zhicheng Zhao, Fei Su, and Anni Cai. Deep canonical correlation analysis with progressive and hypergraph learning for cross-modal retrieval. *Neurocomputing*, 214:618–628, 2016. 1, 2, 3

[42] Rebecca Sharp, Mihai Surdeanu, Peter Jansen, Peter Clark, and Michael Hammond. Creating causal embeddings for question answering with minimal supervision. *EMNLP*, pages 138–148, 2016. 3

[43] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. *ICLR*, 2020. 1, 2, 3, 7

[44] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *ICCV*, 2019. 3

[45] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *EMNLP*, 2019. 3

[46] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *ICML*, pages 10347–10357, 2021. 3

[47] Alasdair Tran, Alexander Mathews, and Lexing Xie. Transform and tell: Entity-aware news image captioning. In *CVPR*, pages 13035–13045, 2020. 2

[48] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, pages 5998–6008, 2017. 3, 6

[49] Sijin Wang, Ruiping Wang, Ziwei Yao, Shiguang Shan, and Xilin Chen. Cross-modal scene graph matching for relationship-aware image-text retrieval. In *WACV*, pages 1508–1517, 2020. 3

[50] Tan Wang, Jianqiang Huang, Hanwang Zhang, and Qianru Sun. Visual commonsense r-cnn. In *CVPR*, pages 10760–10770, 2020. 3, 4, 5

[51] Tan Wang, Jianqiang Huang, Hanwang Zhang, and Qianru Sun. Visual commonsense representation learning via causal inference. In *CVPR Workshops*, pages 378–379, 2020. 3

[52] Yaxiong Wang, Hao Yang, Xueming Qian, Lin Ma, Jing Lu, Biao Li, and Xin Fan. Position focused attention network for image-text matching. *IJCAI*, 2019. 3

[53] Zach Wood-Doughty, Ilya Shpitser, and Mark Dredze. Challenges of using text classifiers for causal inference. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *ACL*, pages 4586–4598, 2018. 3

[54] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, pages 2048–2057, 2015. 5

[55] Fei Yan and Krystian Mikolajczyk. Deep correlation for matching images and text. In *CVPR*, pages 3441–3450, 2015. 3

[56] Xiangyi Yan, Hao Tang, Shanlin Sun, Haoyu Ma, Deying Kong, and Xiaohui Xie. After-unet: Axial fusion transformer unet for medical image segmentation. In *WACV*, pages 3971–3981, 2022. 3

[57] Xun Yang, Fuli Feng, Wei Ji, Meng Wang, and Tat-Seng Chua. Deconfounded video moment retrieval with causal intervention. In *SIGIR*, pages 1–10, 2021. 3

[58] Xu Yang, Hanwang Zhang, Guojun Qi, and Jianfei Cai. Causal attention for vision-language tasks. In *CVPR*, pages 9847–9857, 2021. 3

[59] Chenyu You, Nuo Chen, and Yuexian Zou. Self-supervised contrastive cross-modality representation learning for spoken question answering. In *Findings of the Association for Computational Linguistics: EMNLP*, 2021. 3

[60] Chenyu You, Yuan Zhou, Ruihan Zhao, Lawrence Staib, and James S Duncan. Simcvd: Simple contrastive voxel-wise representation distillation for semi-supervised medical image segmentation. *IEEE Transactions on Medical Imaging*, 2022. 3

[61] Yuning You, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen. Graph contrastive learning with augmentations. *NeurIPS*, 33:5812–5823, 2020. 3

[62] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zi-Hang Jiang, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *ICCV*, pages 558–567, 2021. 3

[63] Zhongqi Yue, Hanwang Zhang, Qianru Sun, and Xian-Sheng Hua. Interventional few-shot learning. In *NeurIPS*, 2020. 3

[64] 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 *ICCV*, pages 11782–11791, 2021. 3

[65] Dong Zhang, Hanwang Zhang, Jinhui Tang, Xian-Sheng Hua, and Qianru Sun. Causal intervention for weakly-supervised semantic segmentation. *NeurIPS*, pages 655–666, 2020. 3

[66] He Zhang, Sebastian Starke, Taku Komura, and Jun Saito. Mode-adaptive neural networks for quadruped motion control. *ACM Transactions on Graphics (TOG)*, 37(4):1–11, 2018. 5

[67] Shengyu Zhang, Tan Jiang, Tan Wang, Kun Kuang, Zhou Zhao, Jianke Zhu, Jin Yu, Hongxia Yang, and Fei Wu. Devlbert: Learning deconfounded visio-linguistic representations. In *ACMMM*, pages 4373–4382, 2020. 2, 3, 4, 5

[68] Wenkai Zhang, Hongyu Lin, Xianpei Han, and Le Sun. Debiasing distantly supervised named entity recognition via causal intervention. In *ACL*, 2021. 3

[69] Handong Zhao, Zhengming Ding, and Yun Fu. Multi-view clustering via deep matrix factorization. In *AAAI*, pages 2921–2927, 2017. 1

[70] Handong Zhao, Hongfu Liu, and Yun Fu. Incomplete multi-modal visual data grouping. In *IJCAI*, pages 2392–2398, 2016. 1

[71] Wentian Zhao, Yao Hu, Heda Wang, Xinxiao Wu, and Jiebo Luo. Boosting entity-aware image captioning with multi-modal knowledge graph. *arXiv preprint arXiv:2107.11970*, 2021. 2

[72] 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 *CVPR*, pages 12647–12657, 2021. 1, 2, 3, 5, 6, 7, 8