# SyncMask: Synchronized Attentional Masking for Fashion-centric Vision-Language Pretraining

Chull Hwan Song[1*] Taebaek Hwang[1*] Jooyoung Yoon[1] Shunghyun Choi[1] Yeong Hyeon Gu[2†]
[1]Dealicious Inc. [2]Sejong University

## Abstract

*Vision-language models (VLMs) have made significant strides in cross-modal understanding through large-scale paired datasets. However, in fashion domain, datasets often exhibit a disparity between the information conveyed in image and text. This issue stems from datasets containing multiple images of a single fashion item all paired with one text, leading to cases where some textual details are not visible in individual images. This mismatch, particularly when non-co-occurring elements are masked, undermines the training of conventional VLM objectives like Masked Language Modeling and Masked Image Modeling, thereby hindering the model's ability to accurately align fine-grained visual and textual features. Addressing this problem, we propose Synchronized attentional Masking (SyncMask), which generate masks that pinpoint the image patches and word tokens where the information co-occur in both image and text. This synchronization is accomplished by harnessing cross-attentional features obtained from a momentum model, ensuring a precise alignment between the two modalities. Additionally, we enhance grouped batch sampling with semi-hard negatives, effectively mitigating false negative issues in Image-Text Matching and Image-Text Contrastive learning objectives within fashion datasets. Our experiments demonstrate the effectiveness of the proposed approach, outperforming existing methods in three downstream tasks.*

## 1. Introduction

Recently, there has been rapid progress in developing Vision-Language Pretraining (VLP) [32, 38, 6, 39, 19, 18, 42, 23, 36, 20, 27, 3], paving the way to bridge the gap between visual and textual features. These VLP methods, trained on extensive image-text datasets, have enabled a deeper understanding of semantic alignment across different modalities. By fine-tuning these pretrained models for specific tasks, particularly in data-scarce areas like image-text retrieval, notable performance improvements have been observed. In the pretraining phase, various factors such as model architecture, training objectives, and batch sampling techniques play a crucial role in effectively harnessing the joint representation of multi-modal data.

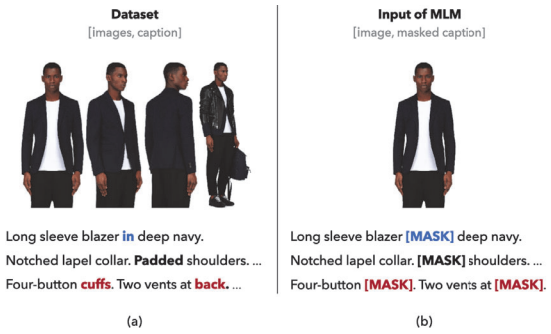


Figure 1. Example of misaligned masks in the MLM task.

However, there are some issues in applying conventional generic VLP approaches to task-specific domains such as fashion. Fashion VLP models [10, 45, 13, 14, 15] typically employ objectives such as Masked Language Modeling (MLM) and Masked Image Modeling (MIM). These methods mask elements like text words or image patches, leveraging surrounding context for prediction or reconstruction. They boost cross-modality by making models infer masked text tokens or image patches from aligned features. However, existing MLM and MIM often suffer from inherent misalignment limitations because the masks are generated randomly, often leading to unmatched elements being masked.

To illustrate these limitations, consider Figure 1 (a) shows a single description associated with four images from the FashionGen [37] dataset. In Figure 1 (b), a random masking scenario is shown where the blue `[MASK]` might lead the model to predict the masked word using only the text context, thereby not incorporating the image information. Similarly, the two red `[MASK]` tokens in (b) lack relevance to the accompanying image, thus hindering the model's ability to connecting between visual and textual features. Fur-

*Equal contribution
†Corresponding author

thermore, in the MIM task, random masking may inadvertently cover parts of the image that contain fashion items not described in the text, leading to mismatches during the training of the alignment of cross-modal features. To solve these problem, we propose SyncMask, selecting masks that represent *synchronously co-occurring* features utilizing the vision-language cross-attention map.

In addition, compared to general domains, the fashion domain often suffers from smaller dataset sizes and less variance in data distribution. This suggests that using standard VLP methods may not adequately distinguish the fine-grained features vital for fashion-related tasks. Thus, we pay attention to a grouped batch sampling technique [3] that similar samples are gradually collected within the batch as the training progresses, impacting the pretraining objectives of Image-Text Contrastive Learning (ITC) and Image-Text Matching (ITM). When similar samples exist within the same batch, it becomes more challenging to differentiate positives and negatives during the training of ITC and ITM compared to using random samples. This encourages the model to more intensely focus on learning fine-grained differences, even with less training.

The existing grouped batch sampling method uses the output features of two uni-modal encoders to find the most similar sample. However, as shown in Figure 1, there are many data that have the same caption for multiple images as for the fashion domain. Therefore, a false negative problem that causes actual positive samples in the same batch to be wrongly labelled as negative when learning ITC and ITM arises if the existing methods are used without changes. To overcome this limitation, we propose a semi-hard negative sampling technique with lower similarities between samples that constitute the batch while removing the false negative.

In summary, the main contributions of this study are:

1. **Synchronized Attentional Masking**: We introduce SyncMask, which replaces random mask in MLM and MIM with targeted mask of co-occurring segments in image-text pairs. By utilizing cross-attention features from a momentum model to generate these masks, this method effectively addresses the problem of mismatched image-text inputs, thereby enhancing fine-grained alignment of cross-modal features.
2. **Refined Grouped Batch Sampling**: Our method incorporates semi-hard negative sampling to tackle data scarcity and distribution disparities in domain-specific datasets, thereby reducing false negatives.

## 2. Related Works

**Vision and Language (VL) Model** Recently, VLMs have focused on enhancing model architecture and designing objectives to integrate visual and textual features effectively. Early studies [32, 38, 39, 6] have used object detectors for extracting visual features as an input for a multi-modal transformer along with textual features. The objective for training models extends vanilla BERT [7] to use MLM, MIM, and ITM losses; however, the object detection module incurs a high computational cost for training and inference. Therefore, there have been attempts to replace it with CNN [19, 18] or linear projection [23]. These studies commonly train models with MLM and ITM, tailoring MIM to their specific architectures. Concurrently, CLIP [36], ALIGN [20] propose models comprising only two unimodal encoders, demonstrating the outstanding representation embedding capabilities of contrastive learning. ALBEF [27] add a contrastive learning objective to the previous multi-modal transformer structure for aligning the two modalities before fusion. Based on this model, GRIT-VLP [3] demonstrate improved learning efficiency when configuring batches with hard negative samples.

| METHOD | MM | MLM | MIM | ATM | AVM | OnUp |
|---|---|---|---|---|---|---|
| DMAE [1] | | | ✓ | | | |
| MaskDistill [34] | | | ✓ | | ✓ | ✓ |
| AttMask [21] | | | ✓ | | ✓ | ✓ |
| ALBEF [27] | ✓ | ✓ | | | | |
| MaskVLM [24] | ✓ | ✓ | ✓ | | | |
| MAMO [44] | ✓ | ✓ | ✓ | | | |
| FashionBERT [10] | ✓ | ✓ | ✓ | | | |
| Kaleido-BERT [45] | ✓ | ✓ | ✓ | ✓ | ✓ | |
| FashionViL [13] | ✓ | ✓ | ✓ | | | |
| FashionSAP [15] | ✓ | ✓ | | | | |
| **Ours** | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

Table 1. Related works *vs*. Ours on Masked Modeling. MM:Multi-Modal. MLM:Masked Language Modeling. MIM:Masked Image Modeling. ATM:Attentional Textual Mask. AVM:Attentional Visual Mask. OnUp:Online Update for Attentional Masking

**FashionVL Model** In recent years, various studies [10, 45, 13, 14, 15] attempted to capture the finer details of images, building upon established models and pre-training objectives from the general VL task. FashionBERT [10] integrates patch-based image features and BERT-based text representations for addressing the limitations of region of interests (RoIs) in capturing fine-grained details. Kaleido-BERT [45] improves fine-grained fashion cross-modality representations through alignment guided masking compared to random masking. FashionViL [13] employs a versatile VLP framework, leveraging two pre-training tasks for capturing the rich fine-grained information of fashion data. FashionSAP [15] employs abstract fashion symbols and an attributes prompt technique for effectively modeling multi-modal fashion attributes. We clarify the importance of the attentional masking technique that employs alignment between MIM and MLM building upon preceding methods. In addition, we underscore a previously unaddressed need for grouped batch sampling within the fashion domain.

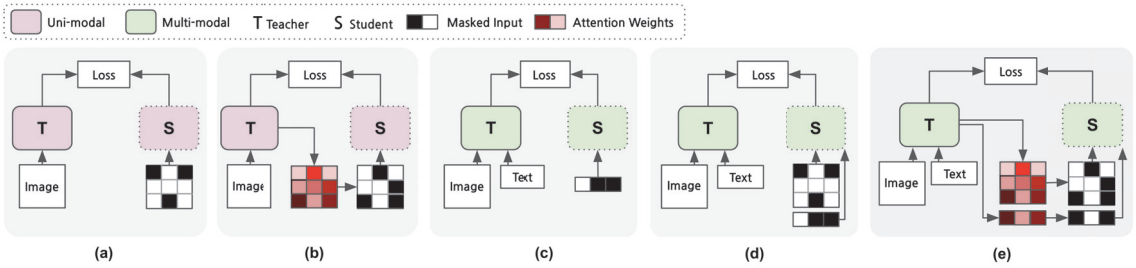


Figure 2. Overview of masking strategies using a teacher-student distillation framework. 1) Uni-modal models: (a) random masking, (b) teacher-guided attentional masking. 2) Multi-modal models: (c) random text masking, (d) random image/text masking, (e) teacher-guided cross-attentional masking (Ours).

**Attention-guided Masked Modeling** MIM and MLM leverage unmasked contextual clues to predict masked visual and textual elements, respectively. As shown in the upper section of Table 1 and Figure 2 (a, b), in the MIM task, prior studies have evolved from random masking [16, 2, 41, 1] to attention-guided masking [29, 34, 21], demonstrating that targeting highly-attended patches with teacher-student distillation framework improves masked modeling outcomes. Masked modeling also has been pivotal in advancing cross-modal alignment within VLMs, spanning both general [38, 32, 27, 24, 44] and fashion-specific [10, 45, 13, 15] domains. This is briefly illustrated in Figure 2 (c, d) and delineated in the middle and lower sections of Table 1. These enhance the model's proficiency aligning co-occurring visual and textual representations. However, challenges emerge when masking non-co-occurring elements, which hinders the accurate pairing of visual and textual features. Kaleido-BERT [45] improve fine-grained cross-modality representations through text-image alignment-guided masking. This requires additional components for an attention-based alignment generator, increasing computational demands and potential model overfitting on fixed text-image mask pairs. Our approach overcomes these limitations by progressively tailoring masks during end-to-end training.

## 3. Methods

We provide an overview of the preliminary aspects, which includes the model architecture and two well-established training objectives for VLP. Subsequently, we present a detailed explanation of the proposed methods, which are synchronized attentional masked modeling and grouped batch sampling with semi-hard negatives.

### 3.1. Preliminaries

For an image-text pair, we refer input sequences as follows: tokenized text embeddings are denoted as $T = [\mathbf{t}_{\texttt{[CLS]}}, \mathbf{t}_1, \ldots, \mathbf{t}_N] \in \mathbb{R}^{(N+1)\times D}$ and visual patch embeddings represented as $V = [\mathbf{v}_{\texttt{[CLS]}}, \mathbf{v}_1, \ldots, \mathbf{v}_{N'}] \in \mathbb{R}^{(N'+1)\times D}$. Here, $D$, $N$, and $N'$ refer to the transformer dimension, the number of text tokens, and the number of image patches, respectively. Additionally, $\mathbf{t}_{\texttt{[CLS]}}$ and $\mathbf{v}_{\texttt{[CLS]}}$ specifically reference the `[CLS]` embeddings.

The model consists of two components: a teacher model, referred to as the momentum model, and a student model. The student model, denoted as $f_\theta$, is parameterized by $\theta$, which includes a textual encoder $f^T_\theta(T) \in \mathbb{R}^{(N+1)\times D}$, visual encoder $f^I_\theta(V) \in \mathbb{R}^{(N'+1)\times D}$, and multi-modal encoder $f^M_\theta(f^T_\theta(T), f^I_\theta(V)) \in \mathbb{R}^{(N+1)\times D}$. For the momentum model $f_{\theta'}$, the training parameters are updated by the exponential moving average method, $\theta' \leftarrow \beta\theta' + (1-\beta)\theta$, where $\beta$ represents a hyperparameter.

**Image-Text Contrastive Learning (ITC)** At the front of the multi-modal encoder, ITC pre-aligns the joint latent space of the textual encoder and visual encoder. This objective have proved its effectiveness in VLMs [36, 27, 26, 3, 13, 15]. We also adopt the ITC loss framework proposed by [27, 26, 15], which incorporates a momentum encoder for utilizing soft labels as ITC training targets, thereby addressing potential positive instances within negative pairs.

**Image-Text Matching (ITM)** For the ITM loss, the model classifies image-text pairs as either matched (positive) or not matched (negative) using a joint representation obtained from the `[CLS]` token output embedding of the multi-modal encoder. This vector is passed through an FC layer and softmax for binary prediction. Like ALBEF [27], we exploit hard negatives in the ITM task, identifying pairs that share similar semantics but differ in fine-grained details, using in-batch contrastive similarity from ITC.

### 3.2. Synchronized Attentional Masked Modeling

We extend the use of momentum model, a self-supervised tool for momentum distillation outlined in MoCo [17] and ALBEF [27], beyond its conventional role of generating pseudo-labels. We employ its multi-modal encoder, which calculate the cross-attention map to identify patches and tokens where image and text features strongly correlate. These elements, indicated by heightened attention weights, are then

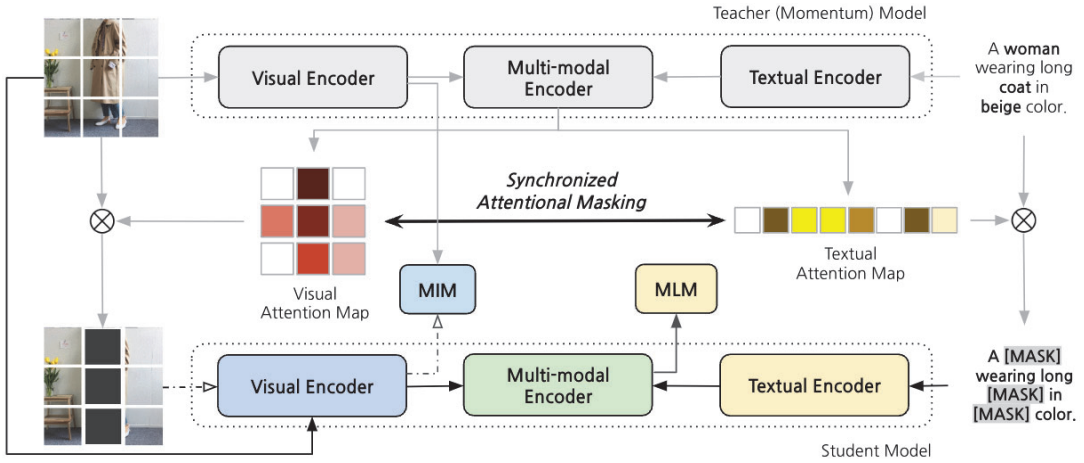


Figure 3. A schematic overview of the SyncMask process: Leveraging cross-attention features from the teacher (momentum) model to generate informative masks for both MIM and MLM tasks. It is important to note that the input for MLM consists of unmasked image paired with masked text.

selectively masked. This attentional masking, targeting *synchronously co-occurring* features, enhances cross-modality over traditional random masking methods in the MLM and MIM phases. Moreover, the momentum model's output features provide enhanced labels for these masked regions, offering a depth of information beyond conventional discrete labels. We will explore this method further to understand its full potential in capturing intricate multi-modal interactions.

**Vision-Language Synchronized Attentionl Masking** MLM and MIM elevate the alignment of visual and textual representations in models. However, when non-co-occurring elements are masked, these techniques face limitations that restrict the model's ability to accurately match multi-modal features. To address this issue, we propose a Synchronized attentional Masking (SyncMask) strategy into masked multi-modal modeling objectives. As depicted in Figure 3, we extracted two sets of synchronized attention weights from the cross-attention module of the multi-modal encoder's last layer in the momentum model. The module enables the model to fuse image-text features using a cross-attention mechanism that processes a query ($Q^T$), key ($K^I$), and value ($V^I$), as represented by the following equation:

$$\texttt{Attention}(Q_i^T, K_i^I, V_i^I) = \alpha(Q_i^T, K_i^I) \odot V_i^I \quad (1)$$

where $1 \leq i \leq H$, with $H$ denoting the number of heads in the MHA, and $\odot$ representing the Hadamard product. In this context, $\alpha$ refers the cross-attention function, which can be expressed as:

$$\alpha(Q_i^T, K_i^I) = \texttt{softmax}(\frac{Q_i^T (K_i^I)^\top}{\sqrt{d}}) \in \mathbb{R}^{(N'+1)\times(N+1)} \quad (2)$$

The function $\alpha(Q_i^T, K_i^I)$ computes the attention weights for the image from the perspective of the text. Similarly, by altering the query ($Q$) and key ($K$), we calculate the text attention weight from the image perspective, as represented by the following equation:

$$\alpha(Q_i^I, K_i^T) = \texttt{softmax}(\frac{Q_i^I (K_i^T)^\top}{\sqrt{d}}) \in \mathbb{R}^{(N+1)\times(N'+1)} \quad (3)$$

Utilizing Equation 2 and Equation 3, we derive two synchronized textual-visual cross-attention weights. These weights, $\mathbf{o}^T \in \mathbb{R}^N, \mathbf{o}^I \in \mathbb{R}^{N'}$, are obtained by averaging the patch tokens of the last layer, excluding the `[CLS]` token. This process allows us to map each word in the sentence sequence to its corresponding attention in $\mathbf{o}^T$. Further, $\mathbf{o}^I$ can be reshaped to $\mathbb{R}^{P\times P}$, which aligns with the image patches.

$$\texttt{Idx}^T = \texttt{shuffle}(\texttt{sort}^{desc}(\mathbf{o}^T)[\geq L])[\geq K] \quad (4)$$

$$\texttt{Idx}^I = \texttt{shuffle}(\texttt{sort}^{desc}(\mathbf{o}^I)[\geq L'])[\geq K'] \quad (5)$$

In Equation 4 and Equation 5, we first sort the attention weights $\mathbf{o}^T$ and $\mathbf{o}^I$ in descending order ($\texttt{sort}^{desc}$), and then extract their indices ($\texttt{Idx}$). Equation 4 focuses on indices corresponding to the top $L$ values of $\mathbf{o}^T$, where $L$ is a threshold greater than the actual mask size $K$. These indices are randomly shuffled ($\texttt{shuffle}$) to introduce randomness, and ultimately, only those indices that satisfy the condition $K \leq L$ are retained. A similar approach is applied to $K'$ and $L'$ in Equation 5. The parameters $K, K', L$, and $L'$ are defined based on a mask ratio $r \in [0, 1]$.

The final attention masks for textual and visual components are represented by the vectors $\mathbf{m}^T$ and $\mathbf{m}^I$, respectively. The computation of these masks utilizes the indices

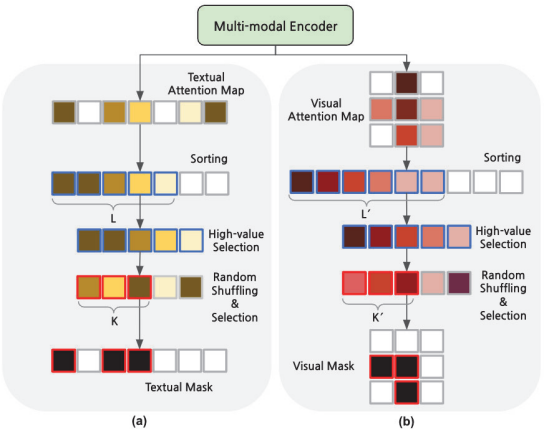


Figure 4. Selection phase of the SyncMask

$\texttt{Idx}^T$ for text and $\texttt{Idx}^I$ for images, as formulated in the equation below:

$$\mathbf{m}^T[\texttt{Idx}^T] \leftarrow 1, \quad \mathbf{m}^I[\texttt{Idx}^I] \leftarrow 1 \tag{6}$$

Initially, the vectors $\mathbf{m}^T$ and $\mathbf{m}^I$ are zero-initialized with dimensions $N$ (for text) and $N'$ (for images). They are then updated to binary attention masks, which are integral in capturing the synchronized interaction between textual and visual elements, as illustrated in Figure 4. Subsequently, these masks are employed in the masked modeling processes for text and images (detailed in Equation 7 and Equation 8).

Building upon this foundation, the text mask vector $\mathbf{m}^T$ consists of elements $[m_1^t, \ldots, m_N^t] \in \{0,1\}^N$, while the image mask vector $\mathbf{m}^I$ is composed of $[m_1^i, \ldots, m_{N'}^i] \in \{0,1\}^{N'}$. For each tokenized text vector, the masked version $\tilde{\mathbf{t}}_i$ is determined by:

$$\tilde{\mathbf{t}}_j = (1 - m_j^t) \cdot \mathbf{t}_j + m_j^t \cdot \mathbf{t}_{\text{mask}} \tag{7}$$

In this equation, $1 \leq j \leq N$, and $\mathbf{t}_{\text{mask}}$ denotes the special token used for textual masking. In the context of Masked Image Modeling (MIM), each image patch is processed with a learnable mask, resulting in the masked image vector $\tilde{\mathbf{v}}_k$:

$$\tilde{\mathbf{v}}_k = (1 - m_k^i) \cdot \mathbf{v}_k + m_k^i \cdot \mathbf{v}_{\text{mask}} \tag{8}$$

Here, $1 \leq k \leq N'$, and $\mathbf{v}_{\text{mask}}$ represents the learnable mask embedding [4]. The masked tokenized inputs are thus represented as $\tilde{T} = [\mathbf{t}_{\text{cls}}; \tilde{\mathbf{t}}_1; \ldots; \tilde{\mathbf{t}}_N]$ for text and $\tilde{V} = [\mathbf{v}_{\text{cls}}; \tilde{\mathbf{v}}_1; \ldots; \tilde{\mathbf{v}}_{N'}]$ for image. These masked inputs are processed using the proposed SyncMask $\mathbf{m}^T$ and $\mathbf{m}^I$.

**Synchronized Attentional Masked Language Modeling** MLM predicts masked words based on the surrounding contextual text and image. Many existing VLMs applied the MLM method proposed in BERT [7], randomly masking words with a probability of 15%. However, this approach may not be suitable for vision-language datasets with short caption lengths, especially for nonstandard datasets such as fashion, demanding a thorough understanding of fine-grained attributes. We leverage previous works that addressed these issues by increasing masking probabilities [3] and employing attribute prompts [15]. Building upon these methods, we employ masks generated by SyncMask which is contextually attuned to the corresponding image.

Let $\boldsymbol{h}(V, \tilde{T})$ denote the model's predicted probability for a masked token. $\tilde{\boldsymbol{y}}$ denote a one-hot vocabulary distribution in which the ground-truth token is assigned a probability of 1. MLM minimize cross-entropy loss, described as follows:

$$\mathcal{L}_{\text{MLM}} = \mathbb{E}_{(V,\tilde{T})\sim D}[\text{CE}(\tilde{\boldsymbol{y}}, \boldsymbol{h}(V, \tilde{T})], \tag{9}$$

where $\text{CE}(\cdot,\cdot)$ refers the cross-entropy between two vectors.

**Synchronized Attentional Masked Image Modeling** In the distillation-based MIM [21, 29, 34], a teacher encoder sees the full image, whereas the student encoder, seeing the masked image, tackles the reconstruction objective. Our method adopts a similar objective framework, but with a key difference: our masks, generated through SyncMask, are designed to reflect text-informed elements in masked image. To calculate the MIM loss, the following approach is used:

$$\begin{aligned}&\textsc{Dist}(\boldsymbol{f}_{\theta'}^I(V),\ \boldsymbol{f}_{\theta}^I(\tilde{V})) \\ &= \frac{1}{\Omega(\mathbf{m}^I)} \sum_{k=1}^{N'} \mathbf{m}_k^I \cdot \ell_1^{Smooth}(\boldsymbol{f}_{\theta'}^I(V)_k, \boldsymbol{f}_{\theta}^I(\tilde{V})_k)\end{aligned} \tag{10}$$

where $\boldsymbol{f}_{\theta'}^I(\cdot)_k$ and $\boldsymbol{f}_{\theta}^I(\cdot)_k$ represent the output feature of teacher and student model for the $k$-th image patch, respectively. $\Omega(\cdot)$ means the number of elements with a value of 1 in a vector.

$$\ell_1^{Smooth}(a,b) = \begin{cases} 0.5(a-b)^2 & \text{if } |a-b| < \gamma \\ |a-b| - 0.5 & \text{otherwise,} \end{cases} \tag{11}$$

where $\ell_1^{Smooth}$ [11] represents a robust L1 loss less sensitive to outliers than the L2 loss and $\gamma$ is a hyperparameter set to 1. Conclusively, the training objective of MIM can be formulated as:

$$\mathcal{L}_{\text{MIM}} = \mathbb{E}_{(V,\tilde{V})\sim D}[\textsc{Dist}(\boldsymbol{f}_{\theta'}^I(V),\ \boldsymbol{f}_{\theta}^I(\tilde{V}))] \tag{12}$$

The final loss ($\mathcal{L}$) is given as:

$$\mathcal{L} = \mathcal{L}_{\text{MIM}} + \mathcal{L}_{\text{MLM}} + \mathcal{L}_{\text{ITC}} + \mathcal{L}_{\text{ITM}} \tag{13}$$

where $\mathcal{L}_{\text{ITC}}$ and $\mathcal{L}_{\text{ITM}}$ denote ITC and ITM, respectively. Due to space constraints, detailed formulations of these two losses are elaborated in the Appendix.

| METHODS | I2T⋆ | T2I⋆ | MEAN⋆ | I2T | T2I | MEAN |
|---|---|---|---|---|---|---|
| | R@1 | R@1 | | R@1 | R@1 | |
| VSE++ [9] | 4.59 | 4.60 | 4.60 | – | – | – |
| VL-BERT [38] | 19.26 | 22.63 | 20.95 | – | – | – |
| ViLBERT [32] | 20.97 | 21.12 | 21.05 | – | – | – |
| Image-BERT [35] | 22.76 | 24.78 | 23.77 | – | – | – |
| OSCAR [28] | 23.39 | 25.10 | 24.25 | – | – | – |
| FashionBERT [10] | 23.96 | 26.75 | 25.36 | – | – | – |
| Kaleido-BERT [45] | 27.99 | 33.88 | 30.94 | – | – | – |
| CommerceMM [43] | 41.60 | 39.60 | 62.75 | – | – | – |
| EI-CLIP [33] | 38.70 | 40.06 | 39.38 | 25.70 | 28.40 | 27.05 |
| ALBEF [27] | 63.97 | 60.52 | 62.20 | 41.68 | 50.95 | 46.32 |
| FashionViL [13] | 65.54 | 61.88 | 63.71 | 42.88 | 51.34 | 47.11 |
| FashionSAP [15] | 73.14 | 70.12 | 71.63 | 54.43 | 62.82 | 58.63 |
| Ours | **75.00** | **71.00** | **73.00** | **55.39** | **64.06** | **59.73** |

Table 2. Cross-modal retrieval result on FashionGen [37] in the sub/full set of evaluation following previous work. ⋆:sub set.

### 3.3. Grouped Batch with Semi-hard Negatives

In the generic domain, [3] proposed the GRIT strategy for enhancing training effectiveness by forming mini-batches with similar examples. In this strategy, the *grouping based on similarity* phase plays a crucial role. During this phase, similarity calculations are performed in both directions, utilizing the `[CLS]` outputs from the unimodal encoders. For each example, the algorithm iteratively identifies the index with the highest similarity, alternating between the image-to-text and text-to-image directions in a sequential manner. These highly similar indices are grouped together within the mini-batch, ensuring that each mini-batch consists of examples exhibiting the highest possible similarity.

However, in fashion datasets, this strategy leads to a false-negative problem in ITM and ITC, where true positives are mislabeled as negatives within mini-batches. To address this problem, we opt to group semi-hard negatives with relatively lower similarity (the $s^{th}$ highest pairwise similarity) instead of the highest ($1^{st}$) during the *grouping based on similarity* phase, where $s$ represents a predefined hyperparameter that is greater than 1. In addition, we prevent true positive samples from grouping by considering the item indices of the samples. Through this approach, the model is trained with negatives that are similar but exhibit meaningful differences, thereby enabling the effective learning of fine-grained distinctions with a limited dataset. More details are in the Appendix.

## 4. Experiments

### 4.1. Implementation Details

The foundational architecture of the proposed model is aligned with prior works for demonstrating the effectiveness of the proposed approach [27, 3, 15]. The image encoder adopts the architecture of ViT-B16 [8], whereas the text encoder comprises the first six blocks of the BERT-bas3 [7]. The multi-modal encoder extends the self-attention layers of the last six blocks of BERT with the cross-attention layers. The proposed model is initialized with pre-trained ALBEF [27] same as the Fashion-SAP to ensure a fair comparison [15]. In addition, we employ the same data augmentation and prompt input strategies as FashionSAP [15]. During pre-training, we conduct experiments using 8 RTX 3090 GPUs each with a batch size of 8 for 30 epochs. We adopt a momentum queue size of 48,000 to facilitate grouped batch sampling, which is consistent with GRIT-VLP [3]. The input image size is set to 256 $\times$ 256. We apply the AdamW [31] optimizer with a learning rate of 6e-5.

### 4.2. Datasets

**FashionGen [37]** FashionGen comprises 320K text-image pairs and 40K unique fashion items, each represented by multiple images from different angles. For pre-training, we employ the FashionGen train set, which contains approximately 260.5K text-image pairs. In addition, FashionGen supports various downstream tasks, including text-to-image retrieval, image-to-text retrieval, category recognition, and subcategory recognition.

**FashionIQ [40]** FashionIQ encompasses 77K unique fashion items and includes 18K training triplets (i.e., query image, modified text, target image) and 6K validation datasets for a text-guided image retrieval task. It contains three different categories: Dress, Toptee, and Shirt.

### 4.3. Downstream Tasks

**Cross-modal Retrieval** We evaluate a cross-modal retrieval task that includes image-text retrieval (ITR) and text-

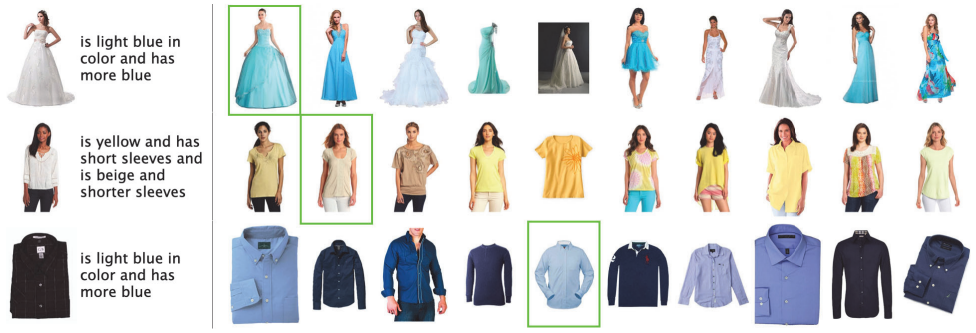


Figure 5. The top-10 TGIR results of the SyncMask model on the FashionIQ dataset. On the left, the reference images paired with their guided descriptions are shown, while the right side presents the model's predicted images ranked by descending scores. Ground truth images are distinctly outlined with a green bounding box. It is worth mentioning that the set of predictions includes other images that also qualify as suitable matches.

| METHODS | DRESS | | TOPTEE | | SHIRT | | MEAN | |
|---|---|---|---|---|---|---|---|---|
| | R@10 | R@50 | R@10 | R@50 | R@10 | R@50 | R@10 | R@50 |
| CIRR [30] | 17.45 | 40.41 | 21.64 | 45.38 | 17.53 | 38.81 | 18.87 | 41.53 |
| VAL [5] | 22.53 | 44.00 | 27.53 | 51.68 | 22.38 | 44.15 | 24.15 | 46.61 |
| CosMo [25] | 25.64 | 50.30 | 29.21 | 57.46 | 24.90 | 49.18 | 26.58 | 52.31 |
| DCNet [22] | 28.95 | 56.7 | 30.44 | 58.29 | 23.95 | 47.3 | 27.78 | 54.10 |
| FashionVLP [12] | 32.42 | 60.29 | 38.51 | 68.79 | 31.89 | 58.44 | 34.27 | 62.51 |
| FashionViL [13] | 33.47 | 59.94 | 34.98 | 60.79 | 25.17 | 50.39 | 31.21 | 57.04 |
| FashionSAP [15] | 33.71 | 60.43 | 41.91 | 70.93 | 33.17 | 61.33 | 36.26 | 64.23 |
| Ours | **33.76** | **61.23** | **44.82** | **72.06** | **35.82** | **62.12** | **38.13** | **65.14** |

Table 3. Text-guided image retrieval performance in FashionIQ [40]

image retrieval (TIR). Further, ITR focuses on finding relevant textual descriptions for a given image query. TIR, the inverse task, retrieves pertinent images based on a textual query. These tasks assess the effectiveness of the model in capturing cross-modal relationships between text and images within retrieval scenarios. Following the previous works [13, 15], we evaluate cross-modal retrieval not only on the subset with 1K retrievals but also on the full dataset of FashionGen [37]. The results, including R@1 scores for both subset and the full set, are presented in Table Table 2, demonstrating an improved performance compared to that of the previous results.

**Text-guided Image Retrieval** This task aims to retrieve target images by considering query pairs that reference image and modified descriptions; this is more challenging than traditional retrievals. Therefore, we need to select the target image for identifying minor differences in the changes in description while maintaining the characteristics of the reference image. For a fair comparison, we adopt a similar fine-tuning as outlined in the FashionSAP [15]. In the actual dataset, there are many images that match with the query pairs in addition to the target image referred to as the real correct answer, as shown in Figure 5. Thus, a qualitative evaluation can be conducted in that the model finds similar images well in addition to the actual correct answer. As shown in Table 3, the proposed model surpasses previous models and demonstrates state-of-the-art performance.

| METHODS | CR | | SCR | |
|---|---|---|---|---|
| | Acc | Macro-F | Acc | Macro-F |
| F-BERT [10] | 91.25 | 70.50 | 85.27 | 62.00 |
| K-BERT [45] | 95.07 | 71.40 | 88.07 | 63.60 |
| F-ViL [13] | 97.48 | 88.60 | 92.23 | 83.02 |
| FashionSAP [15] | 98.34 | 89.84 | **94.33** | 87.67 |
| Ours | **98.41** | **90.31** | 94.21 | **87.83** |

Table 4. CR and SCR results on FashionGen [37].

**Category / Subcategory Recognition** In this downstream task, the objective is to classify the category and subcategory of fashion items using the textual and visual information provided. In line with earlier studies [10, 45, 13, 15], we simply attach a linear layer to the `[CLS]` token, which serves as the fusion feature, for task label prediction. As indicated in Table 4, our proposed model exhibits competitive performance compared to existing models.

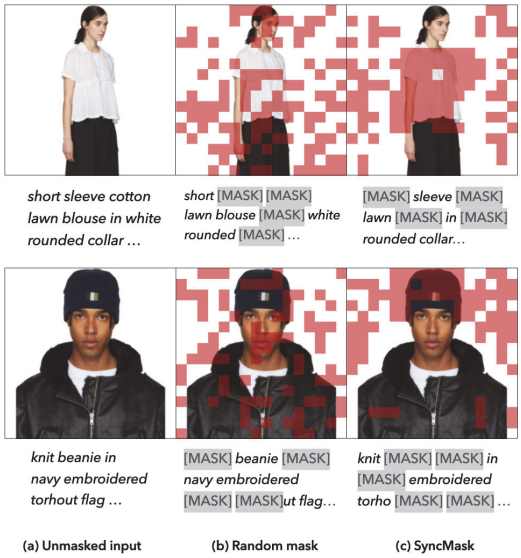


Figure 6. Examples of random masking (b) and SyncMask (c) for MIM and MLM. The latter applies masks to pertinent features in both the input image and text (a), offering a more context-sensitive selection than the former.

### 4.4. Ablation Study

**Random *vs*. Attentional Masking** We ablate the SyncMask with other non-synchronized masking methods. Table 5 lists the comparison results for the combination of random and attentional masking. In this experiment, we evaluated five experiments for I2T (R@1), T2I (R@1), CR (Macro-F), SCR (Macro-F), and TMIR (R@10), which were evaluated in the previous experiments. In these experiments, the results of the three experiments with at least one attentional masking were higher than those with no attentional masking (all random). This can be considered an indication of the efficacy of attentional masking. Further, the synergistic performance of the proposed synchronized textual-visual attentional masking (+2.14) is greater than that of only one attentional masking (+0.4, +0.66), which shows the superiority of the SyncMask approach. Figure 6 displays the distinction between applying random masks (b) and SyncMask (c) to unmasked image-text pairs (a) from the FashionGen [37]. SyncMask strategically places masks on image patches that correlate with the text and applies masks to the text informed by the image details, unlike random masking.

**Random *vs*. Hardest *vs*. Semi-hard Negative Sampling** Table 6 compare the effectiveness of various grouped mini-batch sampling strategies across five downstream tasks. The methods evaluated encompassed four scenarios: random grouping, hardest grouping (+1.90), hardest grouping with exclusion of false negatives (+1.98), and semi-hard grouping

| $Rt$ | $Rv$ | $At$ | $Av$ | Mean | Gain | I2T R@1 | T2I R@1 | CR Macro-F | SCR Macro-F | TMIR R@10 |
|---|---|---|---|---|---|---|---|---|---|---|
| ✓ | ✓ | | | 70.31 | | 73.30 | 69.80 | 86.21 | 86.16 | 36.08 |
| | ✓ | ✓ | | 70.71 | +0.40 | 74.80 | 70.10 | 86.49 | 86.31 | 35.84 |
| ✓ | | | ✓ | 70.97 | +0.66 | 74.30 | 70.80 | 85.12 | 86.78 | 37.87 |
| | | ✓ | ✓ | **72.45** | +2.14 | **75.00** | **71.00** | **90.31** | **87.83** | **38.13** |

Table 5. Ablation study results for *Random vs. Attention* Masked Modeling on five downstream tasks. $Rt$:Random Textual masking. $Rv$:Random Visual masking. $At$:Attentional Textual masking. $Av$:Attentional Visual masking.

| Group | Efn | Mean | Gain | I2T R@1 | T2I R@1 | CR Macro-F | SCR Macro-F | TMIR R@10 |
|---|---|---|---|---|---|---|---|---|
| Random | | 69.99 | | 72.30 | 69.00 | 86.21 | 85.40 | 37.06 |
| Hardest | | 71.89 | +1.90 | 74.70 | 70.50 | 90.20 | 86.64 | 37.40 |
| Hardest | ✓ | 71.97 | +1.98 | 74.30 | 71.00 | 90.10 | 87.15 | 37.28 |
| Semi-hard | ✓ | **72.45** | +2.46 | **75.00** | **71.00** | **90.31** | **87.83** | **38.13** |

Table 6. Ablation study results comparing the *Grouped Batch Sampling Strategy* across five downstream tasks. Group: Strategy for grouping phase of GRIT. Efn: Exclude False Negative in a grouping phase using index.

also excluding false negatives (+2.46). These findings suggest that grouping similar samples in a mini-batch is more beneficial for learning than composing batches with random samples. However, given that the fashion dataset often has multiple captions per image or vice versa, performance gains were observed when systematically preventing the grouping of false negatives by using indexing. Nevertheless, due to the prevalence of inherently similar samples that cannot be systematically excluded, we opted for grouping semi-hard negatives instead of the hardest ones, which resulted in a significant performance boost. This highlights the importance of further research from a data perspective, not just in terms of model architecture or loss functions.

## 5. Conclusion

We introduced *Synchronized attentional Masking* for enhanced masked modeling in fashion-centric VLMs. Leveraging cross-attention features of a momentum model, our method tailors the random mask into a targeted mask for synchronously co-occurring segments in image-text pairs in MLM and MIM objectives. This approach effectively resolves misaligned image-text input issues and improving fine-grained cross-modal representation. Additionally, we addressed data scarcity and distribution challenges in fashion datasets, refining grouped batch sampling with semi-hard negatives for ITM and ITC losses. The experimental results showed our methods outperformed established benchmarks in multiple downstream tasks.

## Acknowledgment

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) under the metaverse support program to nurture the best talents (IITP-2024-RS-2023-00254529) grant funded by the Korea government (MSIT).

## References


[1] Yutong Bai, Zeyu Wang, Junfei Xiao, Chen Wei, Huiyu Wang, Alan L Yuille, Yuyin Zhou, and Cihang Xie. Masked autoencoders enable efficient knowledge distillers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24256–24265, 2023.

[2] Hangbo Bao, Li Dong, and Furu Wei. BEiT: BERT pre-training of image transformers. In *ICLR*, 2022.

[3] Jaeseok Byun, Taebaek Hwang, Jianlong Fu, and Taesup Moon. Grit-vlp: Grouped mini-batch sampling for efficient vision and language pre-training. In *European Conference on Computer Vision*, pages 395–412. Springer, 2022.

[4] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021.

[5] Yanbei Chen, Shaogang Gong, and Loris Bazzani. Image search with text feedback by visiolinguistic attention learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3001–3011, 2020.

[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.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019.

[8] 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.

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

[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.

[11] Ross Girshick. Fast R-CNN. In *ICCV*, 2015.

[12] Sonam Goenka, Zhaoheng Zheng, Ayush Jaiswal, Rakesh Chada, Yue Wu, Varsha Hedau, and Pradeep Natarajan. Fashionvlp: Vision language transformer for fashion retrieval with feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14105–14115, 2022.

[13] Xiao Han, Licheng Yu, Xiatian Zhu, Li Zhang, Yi-Zhe Song, and Tao Xiang. Fashionvil: Fashion-focused vision-and-language representation learning. In *ECCV*. Springer, 2022.

[14] Xiao Han, Xiatian Zhu, Licheng Yu, Li Zhang, Yi-Zhe Song, and Tao Xiang. Fame-vil: Multi-tasking vision-language model for heterogeneous fashion tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2669–2680, 2023.

[15] Yunpeng Han, Lisai Zhang, Qingcai Chen, Zhijian Chen, Zhonghua Li, Jianxin Yang, and Zhao Cao. Fashionsap: Symbols and attributes prompt for fine-grained fashion vision-language pre-training. In *CVPR*, 2023.

[16] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *arXiv:2111.06377*, 2021.

[17] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[18] Zhicheng Huang, Zhaoyang Zeng, Yupan Huang, Bei Liu, Dongmei Fu, and Jianlong Fu. Seeing out of the box: End-to-end pre-training for vision-language representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12976–12985, 2021.

[19] Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. 2020.

[20] 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.

[21] Ioannis Kakogeorgiou, Spyros Gidaris, Bill Psomas, Yannis Avrithis, Andrei Bursuc, Konstantinos Karantzalos, and Nikos Komodakis. What to hide from your students: Attention-guided masked image modeling. In *ECCV*, 2022.

[22] Jongseok Kim, Youngjae Yu, Hoeseong Kim, and Gunhee Kim. Dual compositional learning in interactive image retrieval. In *AAAI*, pages 1771–1779, 2021.

[23] 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.

[24] Gukyeong Kwon, Zhaowei Cai, Avinash Ravichandran, Erhan Bas, Rahul Bhotika, and Stefano Soatto. Masked vision and language modeling for multi-modal representation learning. In *ICLR*, 2023.

[25] Seungmin Lee, Dongwan Kim, and Bohyung Han. Cosmo: Content-style modulation for image retrieval with text feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 802–812, 2021.

[26] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint arXiv:2201.12086*, 2022.

[27] 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.

[28] 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.

[29] Zhaowen Li, Zhiyang Chen, Fan Yang, Wei Li, Yousong Zhu, Chaoyang Zhao, Rui Deng, Liwei Wu, Rui Zhao, Ming Tang, et al. MST: Masked self-supervised transformer for visual representation. In *NeurIPS*, volume 34, 2021.

[30] Zheyuan Liu, Cristian Rodriguez-Opazo, Damien Teney, and Stephen Gould. Image retrieval on real-life images with pre-trained vision-and-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2125–2134, 2021.

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

[32] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019.

[33] 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.

[34] Zhiliang Peng, Li Dong, Hangbo Bao, Qixiang Ye, and Furu Wei. A unified view of masked image modeling. In *arXiv preprint arXiv:2210.10615*, 2022.

[35] 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.

[36] 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.

[37] 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.

[38] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. In *International Conference on Learning Representations*, 2019.

[39] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. 2019.

[40] 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.

[41] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9653–9663, 2022.

[42] Hongwei Xue, Yupan Huang, Bei Liu, Houwen Peng, Jianlong Fu, Houqiang Li, and Jiebo Luo. Probing inter-modality: Visual parsing with self-attention for vision-and-language pre-training. volume 34, pages 4514–4528, 2021.

[43] 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.

[44] Zijia Zhao, Longteng Guo, Xingjian He, Shuai Shao, Zehuan Yuan, and Jing Liu. Mamo: Fine-grained vision-language representations learning with masked multimodal modeling. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1528–1538, 2023.

[45] 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 Computer Vision and Pattern Recognition*, pages 12647–12657, 2021.