

Supplementary material for “SyncMask: Synchronized Attentional Masking for Fashion-centric Vision-Language Pretraining”

Chull Hwan Song^{1*} Taebaek Hwang^{1*} Jooyoung Yoon¹ Shunghyun Choi¹ Yeong Hyeon Gu^{2†}

¹Dealicious Inc. ²Sejong University

A. Objectives for Vision-Language Pretraining

In the realm of Vision-Language Pretraining (VLP), a variety of objectives are employed to effectively integrate vision-language features. This appendix delves into two such objectives, Image-Text Contrastive Learning (ITC) and Image-Text Matching (ITM), which were not extensively covered in the main text. For consistency, we use the same notation as employed in the main text.

Image-Text Contrastive Learning At the beginning of the multi-modal encoder, ITC aligns the shared latent space of the textual encoder and visual encoder. It employs a similarity function, denoted as $\sigma = g_v(z_v)^\top g_t(z_t)$ where z_v is the [CLS] embedding of the vision encoder features $f_\theta^I(V)$ and z_t is derived from the text encoder features $f_\theta^T(T)$. Then $g_v(\cdot)$ and $g_t(\cdot)$ mean linear projections which are mapping the [CLS] embeddings into normalized lower dimensional representations. We adopt the momentum contrastive learning [17, 27, 15] and leverage two queues that store U recent uni-modal features obtained from the momentum model, denoted as $g'_v(z'_v)$ and $g'_t(z'_t)$ for their normalized features. Using this features, we formulate similarity as:

$$\sigma(V, T) = g_v(z_v)^\top g'_t(z'_t) \quad (\text{A14})$$

$$\sigma(T, V) = g_t(z_t)^\top g'_v(z'_v) \quad (\text{A15})$$

For every matching set of text and image, the corresponding similarity measures are expressed as:

$$h_k^{v2t}(V) = \frac{\exp(\sigma(V, T_k)/\tau)}{\sum_{u=1}^U \exp(\sigma(V, T_u)/\tau)}, \quad (\text{A16})$$

$$h_k^{t2v}(T) = \frac{\exp(\sigma(T, V_k)/\tau)}{\sum_{u=1}^U \exp(\sigma(T, V_u)/\tau)} \quad (\text{A17})$$

where τ denotes a learnable scaling factor. We define the ground-truth similarity using one-hot label as $\mathbf{y}^{v2t}(V)$

*Equal contribution

†Corresponding author

and $\mathbf{y}^{t2v}(T)$. Thus, ITC is constructed based on the cross-entropy, $\text{CE}(\cdot, \cdot)$, between \mathbf{h} and \mathbf{y} :

$$\mathcal{L}_{\text{ITC}} = \frac{1}{2} \mathbb{E}_{(V, T) \sim D} [\text{CE}(\mathbf{y}^{v2t}(V), \mathbf{h}^{v2t}(V)) + \text{CE}(\mathbf{y}^{t2v}(T), \mathbf{h}^{t2v}(T))] \quad (\text{A18})$$

Image-Text Matching For the ITM loss, the model categorizes image-text pairs as either corresponding (positive) or non-corresponding (negative) by utilizing a shared representation derived from the output embedding of the [CLS] token from the multi-modal encoder $f_\theta^M(f_\theta^T(T), f_\theta^I(V))$. This vector is fed through a fully connected (FC) layer and softmax for binary classification, producing the prediction probability h^{itm} . Similar to previous works [27, 3], we exploit hard negatives in the ITM task, identifying pairs that exhibit akin meanings but vary in intricate specifics, employing in-batch contrastive similarity based on ITC. The ITM objective is expressed as:

$$\mathcal{L}_{\text{ITM}} = \mathbb{E}_{(V, T) \sim D} [\text{CE}(\mathbf{y}^{itm}, \mathbf{h}^{itm}(V, T))], \quad (\text{A19})$$

B. Semi-hard Negatives for Grouped Batch

Grouped Mini-batch Sampling GRIT-VLP [3] employs an adaptive sampling strategy to gather similar samples within mini-batches, enhancing the effectiveness of mining hard negatives for both ITC and ITM. GRouped mIni-baTch sampling (GRIT) strategy consists of four phases: 1) Collecting, where [CLS] features z_v and z_t are stored in two queues, which are larger than a mini-batch size; 2) Example-level shuffling, which ensures randomness among the samples; 3) Grouping, where similar examples are grouped based on similarity scores; and 4) Mini-batch-level shuffling, which shuffles mini-batches for improving the model’s ability to generalize. We tailor the 3) *Grouping* phase to suit the characteristics of fashion data.

In the Grouping phase, similarity scores for image-text pairs within sub-queues are calculated using methods similar to Equation A14 and Equation A15. Each sub-queue has a













Anchor	Positive	Negative		
		Random	Hardest	Semi-hard
<i>Long sleeve quilted bomber jacket ...</i>				
<i>Slim-fit trousers in navy...</i>				
<i>Rib knit cotton-blend beanie in black ...</i>				

Figure A7. Illustration of the differences in negative samples selected within a mini-batch using various *Grouping* methods: Random, Hardest and Semi-hard (Ours).

size of S , smaller than the queue size of Collecting phase but larger than the mini-batch size. These scores denoted as $q^{v2t}(V) \in \mathbb{R}^S$ and $q^{t2v}(T) \in \mathbb{R}^S$, where V and T mean each image-text pair, are used to find similar examples. The process starts by randomly selecting an initial pair from the sub-queue. The algorithm then iteratively finds and stores the index of the example with the highest similarity score, one by one, in the index queue I . Instead of relying on a one-way similarity metric, the algorithm alternates between using image-to-text and text-to-image similarity scores. Thus, half of the pairs are selected based on their image-to-text similarity, while the remaining half are chosen based on their text-to-image similarity. By alternating between the two directions, the method ensures a balanced grouping, effectively capturing the relational nuances between images and texts in both directions.

For the $(i+1)^{th}$ iteration, when considering a specific pair (V_j, T_j) with the index j , the selection of I_{i+1} is determined as follows:

$$I_{i+1} = \begin{cases} \operatorname{argmax}_{k \notin I} q_k^{t2v}(T_j) & \text{if } I_i \text{ is chosen with } q^{v2t} \\ \operatorname{argmax}_{k \notin I} q_k^{v2t}(V_j) & \text{if } I_i \text{ is chosen with } q^{t2v} \end{cases} \quad (\text{A20})$$

Refined Grouping for Semi-hard Negatives In the context of ITM and ITC tasks, the model is trained using both positive and negative images relative to a text anchor within a mini-batch. As illustrated in Figure A7, conventional VLP methods have relied on randomly selected negative images for training. This approach enables the model to identify negatives without necessarily learning fine-grained details. In contrast, the GRIT [3] strategy constructs batches with

hard negatives in a generic domain, facilitating more efficient learning by focusing on fine-grained information.

However, in the fashion datasets, multiple images of the same item, differing only in angles, are often paired with a single text. Consequently, forming batches with the most similar samples can lead to the issue of false negatives. To address this, we have adjusted the grouping phase to include semi-hard negatives, which share similar features but are not identical. This adjustment ensures a more nuanced and effective training process, particularly suited to the unique challenges presented in the fashion domain. In our proposed modification in the Grouping phase, the algorithm shifts its focus from selecting the highest similarity score sample to choosing the s^{th} most similar sample, where s is a hyperparameter. We pre-train on the FashionGen [37] dataset, featuring three angle-specific images and one full-body image per text. Thus, observing reduced false negatives when s is set to 3, we adopted this value for s . This method can be expressed as:

$$I_{i+1} = \begin{cases} \operatorname{argsort}_{k \notin I} (q_k^{t2v}(T_j))_s & \text{if } I_i \text{ is chosen with } q^{v2t} \\ \operatorname{argsort}_{k \notin I} (q_k^{v2t}(V_j))_s & \text{if } I_i \text{ is chosen with } q^{t2v} \end{cases} \quad (\text{A21})$$

In this equation, the `argsort` function sorts the elements $q_k^{t2v}(T_j)$ or $q_k^{v2t}(V_j)$ in descending order, excluding indices already present in the index queue I , and I_{i+1} is then assigned the s^{th} index from this sorted list. This change aims to accumulate semi-hard negatives within the same mini-batch, thereby minimizing the false negatives in both ITC and ITM. To demonstrate the effectiveness of our method, we set all hyperparameters, including queue and sub-queue sizes, identical to those in GRIT-VLP [3], except for s .