Conditioned Masked Language and Image Modeling for Image-Text Dense Retrieval

Ziyang Luo[1], **Yadong Xi**[2], **Rongsheng Zhang**[2],
Gongzheng Li[2], **Zeng Zhao**[2], **Jing Ma**[1]*
[1] Department of Computer Science, Hong Kong Baptist University, Hong Kong SAR, China
[2] Fuxi AI Lab, NetEase Inc., Hangzhou, China
cszyluo@comp.hkbu.edu.hk, majing@hkbu.edu.hk
{xiyadong,zhangrongsheng}@corp.netease.com

Abstract

Image-text retrieval is a fundamental cross-modal task that takes image/text as a query to retrieve relevant data of another type. The large-scale two-stream pre-trained models like CLIP have achieved tremendous success in this area. They embed the images and texts into instance representations with two separate encoders, aligning them on the instance-level with contrastive learning. Beyond this, the following works adopt the fine-grained token-level interaction (Masked Language and Image Modeling) to boost performance further. However, the vanilla token-level objectives are not designed to aggregate the image-text alignment information into the instance representations, but the token representations, causing a gap between pre-training and application. To address this issue, we carefully design two novel conditioned token-level pre-training objectives, Conditioned Masked Language and Image Modeling (**ConMLM** and **ConMIM**), forcing models to aggregate the token-level alignment information into the instance representations. Combing with the instance-level contrastive learning, we propose our cross-modal dense retrieval framework, **Con**ditioned **L**anguage-**I**mage **P**re-training (**ConLIP**). Experimental results on two popular cross-modal retrieval benchmarks (MSCOCO and Flickr30k) reveal the effectiveness of our methods.

1 Introduction

Image-text retrieval is an important task in the cross-modal community. Recent years have witnessed the remarkable success of the large-scale language-image pre-trained models in this area. The existing works can be divided into single-stream and two-stream models. The former one as illustrated in Figure 1a relies on the heavy transformer layers (Vaswani et al., 2017) to fuse the cross-modal information (e.g. UNITER (Chen et al., 2020b), and OSCAR (Li et al., 2020b)). The intolerable drawback of these models is the slow inference speed. All possible query-candidate pairs need to be fed into the models to get the retrieval result of a query. For example, the average inference time of UNITER[1] for a text query on MSCOCO (Lin et al., 2014) is more than 30 seconds. Therefore, these models are hard to be applied in real-life industrial applications.

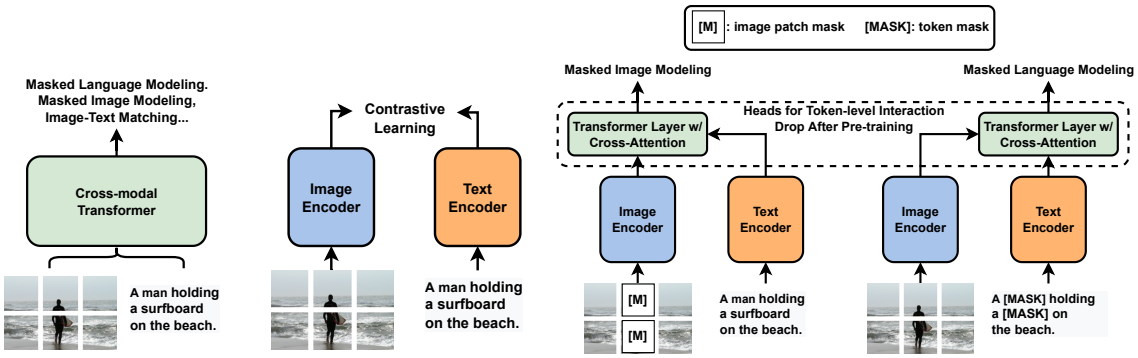
To overcome this limitation, recent works turn to the two-stream models as shown in Figure 1b (e.g., CLIP (Radford et al., 2021)). These models embed images and texts into instance representations (`[CLS]`) with two separate encoders, aligning them on the instance-level with contrastive learning (InfoNCE (van den Oord et al., 2018)) and calculating the retrieval scores with a simple dot-product. Decoupling the correlation of images and texts, the inference speed of these models is much faster. Apart from the instance-level alignment, the fine-grained token-level tasks (Masked Language Modeling, MLM (Devlin et al., 2019) and Masked Image Modeling, MIM (Xie et al., 2022)) are adopted to boost the performance further (Sun et al., 2021; Lu et al., 2022). Nevertheless, the vanilla MLM and MIM as illustrated in Figure 1c are sub-optimal, which are designed to aggregate the token-level alignment information into the token representations, not instance representations. For example, the `[CLS]` representation of the well-known masked language pre-trained model RoBERTa (Liu et al., 2019) performs poorly on the semantic textual similarity tasks without fine-tuning (Reimers and Gurevych, 2019). Therefore, it is necessary for us to come up with more suitable token-level pre-training tasks for image-text dense retrieval.

In this work, we carefully design two novel conditioned token-level objectives, Conditioned Masked Language and Image Modeling

* Corresponding author

[1] 12-layer Transformer model with 110M parameters.

Findings of the Association for Computational Linguistics: EMNLP 2022, pages 130–140
December 7-11, 2022



(a) Single-Stream. (b) Two-Stream. (c) w/ Vanilla Token-level Interaction.

Figure 1: Comparing different categories of Language-Image Pre-training framework for Cross-Modal Retrieval. (a) Single-stream models; (b) Two-stream models; (c) Two-stream models with vanilla token-level interaction.

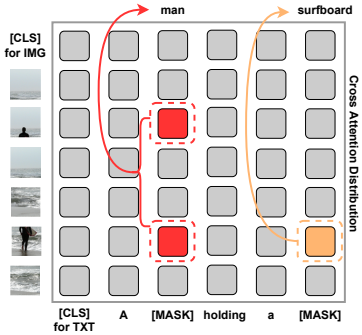


Figure 2: In the vanilla MLM, model can reconstruct the masked text tokens with the token-level alignment, ignoring the instance representation (`[CLS]`).

(**ConMLM** and **ConMIM**) for cross-modal dense retrieval, aggregating the token-level alignment information into the instance representations. For the vanilla MLM in cross-modal pre-training, the image patches representations contain complete semantics information. The model can reconstruct the masked text tokens only with the help of the token-level alignment, as illustrated in Figure 2. In our **ConMLM**, the image patches representations are obtained from the masked image and the image instance representation is obtained from the complete image. As a result, the complete semantics information is only contained in the image instance representation. The text encoder cannot only rely on the token-level alignment to reconstruct the masked text tokens, but need to decode the corresponding information from the image instance representation, which serves as a bridge in our token-level MLM interaction. For our **ConMIM**, it shares a similar idea as **ConMLM** and conditions on the text instance representation.

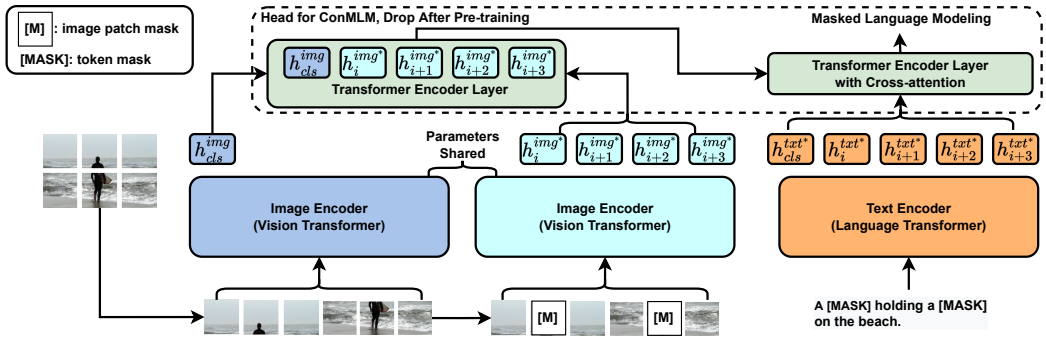
Beyond this, the instance-level interaction is also necessary for cross-modal retrieval. The momentum contrastive learning objective (He et al., 2020) is adopted to align the instance representations of images and texts, decoupling the queue size from the mini-batch size. Combing the instance- and token-level interaction, we proposed our **Con**ditioned **L**anguage-**I**mage **P**re-training (**ConLIP**) framework for image-text dense retrieval. The experimental results on the popular cross-modal retrieval benchmarks (MSCOCO and Flickr30k (Plummer et al., 2015)) reveal that our token-level objects (**ConMLM** and **ConMIM**) are more effective than the vanilla MLM and MIM. In addition, the analysis of the cross attention scores in the token-level pre-training heads corroborates our claim that the instance representations play more active roles in our **ConMLM** and **ConMIM**.

The contributions of our works can be summarized as follows:

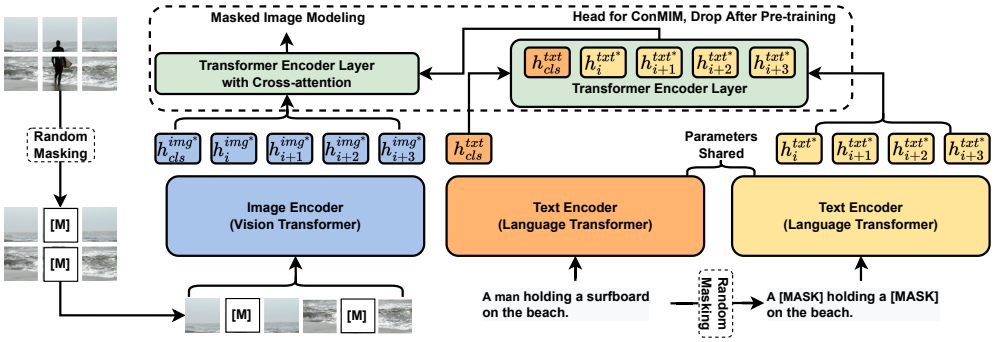
- We design two novel token-level pre-training objectives, **ConMLM** and **ConMIM** for image-text dense retrieval.
- Combining with the instance-level contrastive learning, we introduce an effective conditioned language-image pre-training framework, **ConLIP**.
- Evaluating on the image-text retrieval tasks, our **ConMLM** and **ConMIM** are more effective than the vanilla MLM and MIM.

2 Related Work

Large-scale Pre-trained Models. Since 2019, the large-scale pre-training paradigm has become popular in natural language processing (NLP), computer vision (CV) and cross-modal areas. In NLP,



(a) From Image to Text Conditioned Token-level Interaction (**ConMLM**).



(b) From Text to Image Conditioned Token-level Interaction (**ConMIM**).

Figure 3: An illustration of our conditioned token-level interaction.

the BERT-like pre-trained language models (Devlin et al., 2019; Liu et al., 2019; Lan et al., 2020; Raffel et al., 2020; He et al., 2021) show remarkable language understanding ability. In CV, the pre-trained vision transformers (Dosovitskiy et al., 2021; Xie et al., 2022; Bao et al., 2022) show superior image recognition ability. In the cross-modal area, the pre-trained transformer-based models (Tan and Bansal, 2019; Wang et al., 2022; Dou et al., 2022) also succeed in many cross-modal tasks, like visual question answering (Antol et al., 2015) and visual entailment (Xie et al., 2019). In our work, we align the pre-trained vision transformer (ViT) and language transformer (BERT) with our token- and instance-level cross-modal pre-training for image-text dense retrieval.

Image-Text Retrieval. The goal of this task is to retrieve the relevant image/text with the query from another modality (Chen et al., 2020a). The early works majorly focus on the two-stream models (Kiros et al., 2014; Faghri et al., 2018; Wang et al., 2019; Bastan et al., 2020), embedding the image and text with two separate encoders. Later, the single-stream models with heavy cross/merge-attention layers encode images and texts within the same model and achieve much better performance (Chen et al., 2020b; Li et al., 2020b; Gan et al., 2020; Kim et al., 2021). As we mentioned in the introduction, these models' inference speed is too slow compared with the two-stream models. Recent works turn back to the two-stream style (Jia et al., 2021; Sun et al., 2021; Radford et al., 2021; Wen et al., 2021; Lu et al., 2022). Pre-training with the large-scale paired image-text data, the performance of these models becomes closer to the single-stream counterparts.

Apart from the instance-level alignment, recent works also introduce the Masked Language and Image Modeling as token-level tasks to boost models' performance (Sun et al., 2021; Lu et al., 2022). Inspired by the single-modal dense text retrieval works (Gao and Callan, 2021, 2022; Chuang et al., 2022), the vanilla MLM and MIM are sub-optimal for dense retrieval. Our works introduce two novel token-level objections conditioned on the instance representations (**ConMLM** and **ConMIM**) to fill in this gap.

3 Methodology

3.1 Overview

The cross-modal dense retrieval aims to learn two separate encoders to embed images and texts with instance representations. If the image and text share the same semantic meaning, their representations

will have a high similarity score (cosine similarity). Following the previous works (Radford et al., 2021; Jia et al., 2021), the images and texts are encoded by a Vision Transformer (Dosovitskiy et al., 2021) and a Language Transformer (Devlin et al., 2019). Formally, given a text $x = [x_1, x_2, ...]$ and all patches of an image $y = [y_1, y_2, ...]$, we can write:

$$\begin{aligned}[h_{cls}^{txt}, h_1^{txt}, ...] &= \text{TRF}^{txt}([\texttt{CLS}^{txt}; x]), \\ [h_{cls}^{img}, h_1^{img}, ...] &= \text{TRF}^{img}([\texttt{CLS}^{img}; y]),\end{aligned} \tag{1}$$

where TRF is a transformer model. The special token `[CLS]` is concatenated and encoded with the rest of the text tokens or image patches. Its hidden state in the final layer serves as the instance representation. The similarity score is calculated as:

$$\begin{aligned}sim(x, y) &= sim(h_{cls}^{txt}, h_{cls}^{img}) \\ &= \frac{(h_{cls}^{txt})^T (h_{cls}^{img})}{\|h_{cls}^{txt}\|_2 \|h_{cls}^{img}\|_2}\end{aligned} \tag{2}$$

If we add the random masks to the input, we can write:

$$\begin{aligned}[h_{cls}^{txt^*}, h_1^{txt^*}, ...] &= \text{TRF}^{txt}([\texttt{CLS}^{txt}; x^*]), \\ [h_{cls}^{img^*}, h_1^{img^*}, ...] &= \text{TRF}^{img}([\texttt{CLS}^{img}; y^*]),\end{aligned} \tag{3}$$

where $(\cdot)^*$ indicates that some tokens/patches are masked in the input.

During pre-training, the large-scale language-image paired data are required to align the two encoders with instance- and token-level interaction. For the instance-level interaction, most of the previous works align the instance representations with the contrastive learning, $\mathcal{L}_{inst}$, maximizing the similarity scores ($sim(x, y)$) between paired samples.

For the token-level interaction, models are followed by two pre-training heads (one-layer transformer with cross-attention). In the vanilla MLM, the masked text tokens ($[h_{cls}^{txt^*}, h_1^{txt^*}, ...]$) align with the complete image patches tokens ($[h_{cls}^{img}, h_1^{img}, ...]$) to reconstruct the masked sentence. In the vanilla MIM, the masked image patches tokens ($[h_{cls}^{img^*}, h_1^{img^*}, ...]$) are reconstructed based on the relevant information from the complete text tokens ($[h_{cls}^{txt}, h_1^{txt}, ...]$). After pre-training, these two heads are dropped.

In our work, we indicate that these vanilla MLM and MIM are sub-optimal. We design two novel conditioned masked language and image modeling objectives $\mathcal{L}_{token}^{Con}$ to aggregate the token-level alignment information into the instance representations. The overall pre-training objective is as follows:

$$\mathcal{L}_{all} = \mathcal{L}_{token}^{Con} + \mathcal{L}_{inst}. \tag{4}$$

3.2 Conditioned Token-level Interaction

The vanilla MLM and MIM objectives in cross-modal pre-training are sub-optimal for the image-text retrieval task. The model can easily reconstruct the masked token with the token-level alignment, which is different from our goal to learn good instance-level representations. For example, in Figure 2, the model can ignore the instance representations and reconstruct the masked token "man" based on the image patches of "man".

To fill in this gap, we carefully design two novel conditioned token-level objectives, Conditioned Masked Language Modeling (**ConMLM**) and Conditioned Masked Image Modeling (**ConMIM**) as illustrated in Figure 3. The motivation for our design is to increase the importance of the instance representations during token-level pre-training, enhancing the retrieval performance.

ConMLM. As illustrated in Figure 3a, our method requires two passes. In the first pass, we input a complete image into image encoder to extract the image instance representation h_{cls}^{img} as equation 1. In the second pass, we input an image with random masks to extract the image patch representations ($[h_1^{img^*}, h_2^{img^*}, ...]$) as equation 3. Then we concatenate these two kinds of representations ($[h_{cls}^{img}; h_1^{img^*}, h_2^{img^*}, ...]$) and input them into the extra token-level pre-training head for masked text tokens reconstruction.

Since a part of image tokens and text tokens are masked, only the image instance representation h_{cls}^{img} contains complete semantic information. Image and text encoders need to aggregate the token-alignment information into the instance representations to reconstruct the masked text token. For example, in Figure 3a, both the image patches and text tokens of "man" and "surfboard" are masked. In order to reconstruct them, the text encoder is forced to decode the corresponding information from the image instance representation h_{cls}^{img}. The objective function of **ConMLM** is similar to the vanilla MLM, but conditioned on the image instance representation:

$$\mathcal{L}_{ConMLM}^{I2T} = \sum_i \log P(x_i^* | x^*, y^*, h_{cls}^{img}), \tag{5}$$

where x_i^* is the masked token in the randomly masked text x^*.

ConMIM. As illustrated in Figure 3b, such objective shares the similar idea as **ConMLM**. Only the text instance representation h_{cls}^{txt} contains complete semantic information. Models are asked to reconstruct the masked image patches conditioned on h_{cls}^{txt}. Following the state-of-the-art pre-trained self-supervised Vision Transformer, SimMIM (Xie et al., 2022), we directly require our models to reconstruct the image patch with L-1 norm,

$$\mathcal{L}_{ConMIM}^{T2I} = \sum_i \|y_i - z_i\|_1, \tag{6}$$

where $y_i, z_i \in R^{3HW \times 1}$ are the true RGB values and the predicted values conditioned on the text information (x^* and h_{cls}^{txt}), respectively.

The final conditioned token-level objective is as follows:

$$\mathcal{L}_{token}^{Con} = \mathcal{L}_{ConMIM}^{T2I} + \mathcal{L}_{ConMLM}^{I2T}. \tag{7}$$

3.3 Instance-level Interaction

Apart from our conditioned token-level interaction, instance-level interaction is also necessary for image-text retrieval pre-training. To align the cross-modal information, we adopted the momentum contrastive learning to cache the negative samples with an image queue Q^{img} and a text queue Q^{txt}, so that we can decouple the queue size from the mini-batch size. To maintain the queues dynamically, we still need two momentum encoders for images and texts that share the same structure as the original ones. Following the works of MoCo (He et al., 2020), the momentum encoders are updated by:

$$\begin{aligned} \theta_m^{txt} &= m\theta_m^{txt} + (1-m)\theta_o^{txt}, \\ \theta_m^{img} &= m\theta_m^{img} + (1-m)\theta_o^{img}, \end{aligned} \tag{8}$$

where θ_m denotes the parameters of the momentum encoders, θ_o denotes the parameters of the original encoders and m is the momentum hyperparameter.

Traditionally, the momentum contrastive learning loss from text to image representations is calculated as:

$$\mathcal{L}_{CL}^{T2I} = -\log \frac{e^{sim(h_{cls}^{txt}, h_{cls}^{img})/\tau}}{e^{sim(h_{cls}^{txt}, h_{cls}^{img})/\tau} + \sum_j e^{sim(h_{cls}^{txt}, q_j)/\tau}}, \tag{9}$$

where τ is the temperature hyperparameter and $q_j \in Q^{img}$.

In our works, we follow the decoupled contrastive learning (DCL (Yeh et al., 2021)) to tackle the negative-positive-coupling (NPC) effect to remove the positive samples in the denominator:

$$\begin{aligned} \mathcal{L}_{CL}^{T2I} = &-sim(h_{cls}^{txt}, h_{cls}^{img})/\tau \\ &+ \log \sum_j e^{sim(h_{cls}^{txt}, q_j)/\tau}. \end{aligned} \tag{10}$$

Our ablation studies reveal that DCL leads to better performance.

Similarly, the DCL loss from image to text representations is:

$$\begin{aligned} \mathcal{L}_{CL}^{I2T} = &-sim(h_{cls}^{txt}, h_{cls}^{img})/\tau \\ &+ \log \sum_j e^{sim(p_j, h_{cls}^{img})/\tau}. \end{aligned} \tag{11}$$

where $p_j \in Q^{txt}$. The total loss of the instance-level interaction is defined as:

$$\mathcal{L}_{inst} = \mathcal{L}_{DCL}^{I2T} + \mathcal{L}_{DCL}^{T2I}. \tag{12}$$

4 Experiments

4.1 Experiment Setup

Model Configuration. Our framework adopts the pre-trained vision transformer, ViT-B/16[2] of Dosovitskiy et al. (2021) as our image encoder and BERT-base-uncased[3] of Devlin et al. (2019) as our language encoder. All parameters in the token-level pre-training heads are initialized randomly.

Pre-training Datasets. Our pre-training data come from Conceptual Caption-3M[4] (Sharma et al., 2018) and -12M (Changpinyo et al., 2021). We combine them and randomly collect 200k, 5.3M and 9.5M image-text pairs for our experiments. Notably, these datasets are harvested from the web and contain much noise. In real-life applications, obtaining a large amount of high-quality annotated image-text data is hard. We make our pre-training settings more similar to reality.

Evaluation Datasets and Metrics. We adopt two popular image-text retrieval benchmarks (MSCOCO (Lin et al., 2014) and Flickr30k (Plummer et al., 2015)) to evaluate our models. Each

[2] https://huggingface.co/google/vit-base-patch16-224-in21k

[3] https://huggingface.co/bert-base-uncased

[4] https://ai.google.com/research/ConceptualCaptions/download

Model	TC	#I-T	Flickr30k Test (1K Images)						MSCOCO Test (5K Images)					
			T2I Retrieval			I2T Retrieval			T2I Retrieval			I2T Retrieval		
			R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
Cutting-edge single-stream models														
PixelBERT	$O(n^2)$	5.6M	71.5	92.1	95.8	87.0	98.9	99.5	50.1	77.6	86.2	63.6	87.5	93.6
Unicoder-VL	$O(n^2)$	3.8M	71.5	91.2	95.2	86.2	96.3	99.0	48.4	76.7	85.9	62.3	87.1	92.8
UNITER	$O(n^2)$	9.5M	72.5	92.4	96.1	85.9	97.1	98.8	50.3	78.5	87.2	64.4	87.4	93.1
ViLT	$O(n^2)$	9.9M	64.4	88.7	93.8	83.5	96.7	98.6	42.7	72.9	83.1	61.5	86.3	92.7
UNIMO	$O(n^2)$	9.5M	74.7	93.4	96.1	89.7	98.4	99.1	-	-	-	-	-	-
VILLA	$O(n^2)$	9.5M	74.7	92.9	95.8	86.6	97.9	99.2	-	-	-	-	-	-
Cutting-edge two-stream models														
ALIGN(zs)	$O(n)$	1.8B	75.7	93.8	96.8	88.6	98.7	99.7	45.6	69.8	78.6	58.6	83.0	89.7
CLIP-ViT(zs)	$O(n)$	400M	62.2	85.7	91.9	82.1	96.6	99.0	33.0	58.4	69.0	52.5	76.7	84.6
Frozen in time	$O(n)$	5.5M	61.0	87.5	92.7	-	-	-	-	-	-	-	-	-
LightningDOT	$O(n)$	9.5M	69.9	91.1	95.2	83.9	97.2	98.6	45.8	74.6	83.8	60.1	85.1	91.8
COOKIE	$O(n)$	5.9M	68.3	91.1	95.2	84.7	96.9	98.3	46.6	75.2	84.1	61.7	86.7	92.3
Our Baseline two-stream models														
CL(zs)	$O(n)$	5.3M	51.5	78.6	86.6	67.7	88.5	94.6	27.0	51.5	62.7	39.8	64.5	74.6
Vanilla*(zs)	$O(n)$	5.3M	51.0	79.0	86.6	66.5	87.9	92.7	27.0	51.7	62.9	39.4	64.5	74.6
CL	$O(n)$	5.3M	70.7	90.9	94.4	86.9	97.2	98.8	46.8	74.7	83.6	62.7	85.8	92.4
Vanilla*	$O(n)$	5.3M	70.8	90.8	94.8	86.4	97.4	**99.6**	46.3	74.6	83.6	62.3	86.5	92.4
Our two-stream models														
ConLIP(zs)	$O(n)$	5.3M	52.8	79.8	87.0	68.8	90.0	94.7	28.0	52.6	63.7	40.3	65.2	75.8
ConLIP(zs)	$O(n)$	9.5M	55.2	80.8	87.7	69.6	90.1	94.7	31.4	56.5	67.4	42.7	67.7	77.2
ConLIP	$O(n)$	5.3M	71.8	91.3	95.0	87.2	97.3	98.9	47.5	75.0	83.9	63.4	86.9	92.9
ConLIP	$O(n)$	9.5M	**74.1**	**91.8**	**95.5**	**89.1**	**97.8**	98.7	**49.1**	**76.1**	**84.6**	**64.7**	**87.9**	**93.4**

*(CL: only pre-trained with contrastive learning. *: replacing our ConMLM and ConMIM with the vanilla MLM and MIM. zs: zero-shot performance.)*

Table 1: The image-text retrieval results on the MSCOCO and Flickr30k test sets. **Bold** indicates the best results of the two-stream models. #I-T corresponds to the number of image-text pairs during cross-modal pre-training. TC is the time complexity during inference.

image corresponds to 5 different captions. We follow the Karpathy's split (Karpathy and Li, 2015) to split 113.2k/5k/5k (MSCOCO) and 29.8k/1k/1k (Flickr30k) images for train/val/test, respectively. For the evaluation metrics, we adopt the widely-used R@k (k=1,5,10) to evaluate our models on the image-text retrieval test sets.

Implementation Details. We pre-train and fine-tune our models with the AdamW optimization algorithm (Loshchilov and Hutter, 2019), 4096 batch size, mixed-precision training, FP16 and 30 epochs. We also adopt the linear learning rate decay warm-up strategy. The warm-up step is set to 10% of the total training steps. Each image is resized into the size of 224x224 with center-crop. The masking ratios for images and texts are 50% and 15%. The other hyper-parameter values and implementation details are listed in the Appendix A.

Baseline Systems. To examine whether our **ConMLM** and **ConMIM** are effective, we pre-train two extra models with the same setting as our **ConLIP**: (1) **CL**: this model is trained with only the instance-level contrastive learning (DCL); (2) **Vanilla**: the token-level objectives are replaced by the vanilla MLM and MIM. In addition, we also include the results of the cutting-edge systems to reveal the effective of our **ConLIP**. For the single-stream models, we include 6 models: PixelBERT (Huang et al., 2020), Unicoder-VL (Li et al., 2020a), UNITER (Chen et al., 2020b), ViLT (Kim et al., 2021), UNIMO (Li et al., 2021) and VILLA (Gan et al., 2020). For the two-stream models, we include 5 models: ALIGN (Jia et al., 2021), CLIP (Radford et al., 2021), Frozen in Time (Bain et al., 2021), LightningDOT (Sun et al., 2021) and COOKIE (Wen et al., 2021).

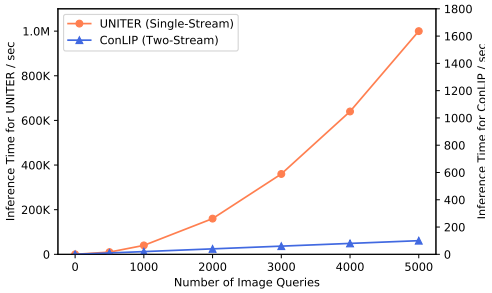


Figure 4: Comparing the inference time between single-stream and two-stream models on MSCOCO test set.

4.2 Image-Text Retrieval Results

Table 1 compare the performance of our **ConLIP** and the baseline systems on two popular image-text retrieval benchmarks, MSCOCO and Flickr30k. The experimental results indicate that our **ConMLM** and **ConMIM** are more effective than the vanilla MLM and MIM.

ConLIP and Vanilla Baseline. Table 1 shows that our baselines are strong models with compara-

Figure 5: Retrieval examples of our **ConLIP** for the text query "a man wears an orange hat and glasses" in the Flickr30k test set.

ble performance with the cutting-edge two-stream models. Replacing the vanilla MLM and MIM with our novel **ConMLM** and **ConMIM**, our **ConLIP** achieves better image-text retrieval performance on both MSCOCO and Flickr30k. Especially, the zero-shot R@1 performance of our **ConLIP** is around 2 points higher than the vanilla baseline on the Flickr30k test set. These experimental results reveal that our **ConMLM** and **ConMIM** are two more suitable token-level pre-training tasks for image-text dense retrieval.

ConLIP and Cutting-edge Models. Apart from our vanilla baselines, we also compare our **ConLIP** with the cutting-edge single- and two-stream models. Pre-training with the same amount of image-text pairs, our **ConLIP** achieves comparable performance with the cutting-edge single-stream models. In addition, Figure 4 shows that the inference time of our **ConLIP** is extremely faster than the single-stream model. Our **ConLIP** has $O(n)$ inference time complexity, while the single-stream models have $O(n^2)$. Compared with the cutting-edge two-stream models, our **ConLIP** also performs better. Notably, our models are only pre-trained with the noisy image-text data from the web, while most of the cutting-edge models also include the human-annotated image-text data in their pre-training. These results indicate that our **ConLIP** is an effective framework for image-text dense retrieval.

Qualitative Examples. Figure 5 shows a retrieval example of our **ConLIP** for the text query "a man wears an orange hat and glasses" in the Flickr30k test set. Our **ConLIP** successfully retrieves the ground truth image as the top-1 result. Though the second image shares the same keywords ("man", "orange", "hat" and "glasses") as the ground truth image, our model still can detect that this image is mismatched and assign a lower score. Some keywords are mismatched for the remaining three images, so our model assigns much lower scores to them.

Model	TXT $\rightarrow h_{cls}^{img}$	IMG $\rightarrow h_{cls}^{txt}$
Mean		
Vanilla*	0.06	0.00
ConLIP	**0.17**	**0.18**
Standard Deviation		
Vanilla*	0.09	0.00
ConLIP	**0.13**	**0.09**

Table 2: The mean and standard deviation of the cross attention scores from the text tokens to the h_{cls}^{img} and the image patch tokens to the h_{cls}^{txt} in the token-level pre-training heads. We average the scores over the samples in the Flickr30k validation set. (*: Replacing our ConMLM and ConMIM with the vanilla MLM and MIM.)

4.3 Cross Attention Analysis

In the introduction and Figure 2, we claim that the instance representations are ignored in the vanilla MLM and MIM for cross-modal pre-training. Our retrieval experimental results in Table 1 also reveal that our **ConMLM** and **ConMIM** can lead to better instance representations. In this section, we conduct an in-depth analysis of the cross-attention pattern in the two token-level pre-training heads to understand the influence of our designs.

We first consider the instance representation of image h_{cls}^{img}. We use the mean of the cross-attention scores from the text tokens to h_{cls}^{img} as the measurement of importance during token-level interaction (**ConMLM** or vanilla MLM). Table 2 shows that the mean score is close to zero in the vanilla MLM, revealing that h_{cls}^{img} is almost ignored by the text tokens. For our **ConMLM**, the score is around three times higher, revealing that h_{cls}^{img} acts as a more important role in our token-level interaction.

Objective	MSCOCO T2I R@1	MSCOCO I2T R@1
Instance-level		
(1) InfoNCE	16.83	21.68
(2) DCL	17.09	22.24
Token-level		
(2) + ConMLM	17.26	22.44
(2) + ConMIM	**17.37**	22.60
(2) + ConMLM + ConMIM	17.25	**22.80**

Table 3: Ablation study for different pre-training objectives. All models are pre-trained on our CC200k dataset. The scores are the zero-shot retrieval results.

#Layer	MSCOCO T2I R@1	MSCOCO I2T R@1
3	16.73	**24.42**
2	16.73	23.94
2*	17.24	22.50
1	**17.25**	22.80

Table 4: Ablation study for different number of token-level interaction layers. * indicates that the parameters of the two layers are shared. All models are pre-trained on our CC200k dataset. The scores are the zero-shot retrieval results.

In addition, we also analyze the standard deviation of the cross-attention scores. A higher standard deviation indicates a wider range of scores. In Table 2, we can find that the score in our **ConLIP** has a higher standard deviation, revealing that the cross attention scores from the text tokens to h_{cls}^{img} spread out over a wider range.

Beyond this, we consider the instance representation of text h_{cls}^{txt}. The cross-attention scores from the image patch tokens to it share the similar pattern as h_{cls}^{img}. This analysis corroborates that our **ConMLM** and **ConMIM** are more suitable than the vanilla MLM and MIM for image-text dense retrieval.

4.4 Ablation Studies

In this section, we conduct ablation studies to compare different settings for our models. Since large-scale pre-training is time-consuming, we choose to pre-train our models with our CC200k dataset and evaluate the zero-shot retrieval performance on the MSCOCO. The experiments contain three perspectives: (1) different pre-training objectives; (2) different number of token-level pre-training layers; (3) different masking ratios for images and texts.

Pre-training Objectives. Table 3 compares different pre-training objectives for our models. For the instance-level pre-training, the decoupled contrastive learning (DCL) is a more effective loss than the traditional InfoNCE. For the token-level pre-training, both **ConMLM** and **ConMIM** can lead to better image-text retrieval performance.

Ratio (Image)	Ratio (Text)	MSCOCO T2I R@1	MSCOCO I2T R@1
50%	15%	17.25	22.80
	25%	17.23	22.80
	30%	17.17	22.72
	40%	17.23	22.76
40%	15%	17.03	22.82
60%		17.14	22.80
70%		17.17	22.78

Table 5: Ablation study for different masking ratios for images and texts. All models are pre-trained on our CC200k dataset. The scores are the zero-shot retrieval results.

Token-level Pre-training Heads Designs. In our **ConLIP**, we adopt one-layer transformers as our token-level pre-training heads. We wonder how the different number of layers affects our models' performance. Table 4 compares four different designs. First, increasing the number of layers can boost the image-to-text retrieval performance, but degenerate text-to-image scores. In addition, sharing the parameters of two token-level layers cannot lead to better performance.

Masking Ratios. In our **ConMLM** and **ConMIM**, the masking ratios are an important hyper-parameter during pre-training. He et al. (2022) indicate that the higher masking ratio can lead to a better self-supervised pre-trained Vision Transformer. Wettig et al. (2022) argue that 15% is not the perfect masking ratio for BERT. We compare several different masking ratios for our models. Table 5 shows that different masking ratios do not have much effect on the retrieval performance. Therefore, we choose to use the default setting (50% for image and 15% for text).

5 Conclusion

In this work, we design two novel token-level pre-training tasks, **ConMLM** and **ConMIM** for image-text dense retrieval. Combining with the instance-level objective, we propose our language-image pre-training framework **ConLIP**. The experimental results and the cross-attention analysis reveal the effectiveness of our methods.

6 Limitations

The major limitation of our work is scalability. In our settings, we pre-train our models with 5.3M or 9.5M image-text pairs, much smaller than the 400M pairs of CLIP (Radford et al., 2021). Therefore, it is unclear how the model performance would be if we scale up the size of pre-training datasets. However, these experiments require enormous GPU resources (256-592 V100 GPUs for CLIP), which are unaffordable.

7 Acknowledgments

We would like to thank the anonymous reviewers for their excellent feedback. This work is partially supported by the Key Research and Development Program of Zhejiang Province (No. 2022C01011), HKBU One-off Tier 2 Start-up Grant (Ref. RCOFSGT2/20-21/SCI/004) and Hong Kong RGC ECS (22200722).

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: visual question answering. In *ICCV*, pages 2425–2433.

Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. 2021. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *ICCV*, pages 1728–1738.

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. 2022. BEit: BERT pre-training of image transformers. In *ICLR*.

Muhammet Bastan, Arnau Ramisa, and Mehmet Tek. 2020. T-vse: Transformer-based visual semantic embedding. *ArXiv*, abs/2005.08399.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. 2021. Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*, pages 3558–3568.

Jianan Chen, Lu Zhang, Cong Bai, and Kidiyo Kpalma. 2020a. Review of recent deep learning based methods for image-text retrieval. In *MIPR*.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020b. Uniter: Universal image-text representation learning. In *ECCV*, pages 104–120.

Yung-Sung Chuang, Rumen Dangovski, Hongyin Luo, Yang Zhang, Shiyu Chang, Marin Soljavci'c, Shang-Wen Li, Wen tau Yih, Yoon Kim, and James R. Glass. 2022. Diffcse: Difference-based contrastive learning for sentence embeddings. In *NAACL*, pages 4207–4218.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, pages 4171–4186.

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. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*.

Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Pengchuan Zhang, Lu Yuan, Nanyun Peng, Zicheng Liu, and Michael Zeng. 2022. An empirical study of training end-to-end vision-and-language transformers. In *CVPR*, pages 18166–18176.

Fartash Faghri, David J. Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2018. VSE++: improving visual-semantic embeddings with hard negatives. In *BMVC*, page 12.

Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. 2020. Large-scale adversarial training for vision-and-language representation learning. In *NeurIPS*, page 13.

Luyu Gao and Jamie Callan. 2021. Condenser: a pre-training architecture for dense retrieval. In *EMNLP*, pages 981–993.

Luyu Gao and Jamie Callan. 2022. Unsupervised corpus aware language model pre-training for dense passage retrieval. In *ACL*, pages 2843–2853.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Doll'ar, and Ross B. Girshick. 2022. Masked autoencoders are scalable vision learners. In *CVPR*, pages 16000–16009.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *CVPR*, pages 9726–9735.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *ICLR*.

Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. 2020. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. *ArXiv*, abs/2004.00849.

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

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

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *ICML*, pages 5583–5594.

Ryan Kiros, Ruslan Salakhutdinov, and Richard S. Zemel. 2014. Unifying visual-semantic embeddings with multimodal neural language models. *ArXiv*, abs/1411.2539.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *ICLR*.

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

Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. 2021. UNIMO: Towards unified-modal understanding and generation via cross-modal contrastive learning. In *ACL*, pages 2592–2607.

Xiujun Li, Xi Yin, Chunyuan Li, Xiaowei Hu, Pengchuan Zhang, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020b. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *ECCV*, pages 121–137.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *ICLR*.

Haoyu Lu, Nanyi Fei, Yuqi Huo, Yizhao Gao, Zhiwu Lu, and Jiaxin Wen. 2022. Cots: Collaborative two-stream vision-language pre-training model for cross-modal retrieval. *ArXiv*, abs/2204.07441.

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *ICCV*, pages 2641–2649.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR*, 21(140).

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *EMNLP*, pages 3982–3992.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, pages 2556–2565.

Siqi Sun, Yen-Chun Chen, Linjie Li, Shuohang Wang, Yuwei Fang, and Jingjing Liu. 2021. Lightning-DOT: Pre-training visual-semantic embeddings for real-time image-text retrieval. In *NAACL*, pages 982–997.

Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. In *EMNLP*, pages 5100–5111.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *ArXiv*, abs/1807.03748.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NeurIPS*, pages 5998–6008.

Liwei Wang, Yin Li, and Svetlana Lazebnik. 2019. Learning two-branch neural networks for image-text matching tasks. In *TPAMI*, page 394–407.

Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. 2022. SimVLM: Simple visual language model pretraining with weak supervision. In *ICLR*.

Keyu Wen, Jinchao Xia, Yuanyuan Huang, Linyang Li, Jiayan Xu, and Jie Shao. 2021. Cookie: Contrastive cross-modal knowledge sharing pre-training for vision-language representation. In *ICCV*, pages 2208–2217.

Alexander Wettig, Tianyu Gao, Zexuan Zhong, and Danqi Chen. 2022. Should you mask 15% in masked language modeling? *ArXiv*, abs/2202.08005.

Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. 2019. Visual entailment: A novel task for fine-grained image understanding. *ArXiv*, abs/1901.06706.

Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. 2022. Simmim: A simple framework for masked image modeling. In *CVPR*, pages 9653–9663.

Chun-Hsiao Yeh, Cheng-Yao Hong, Yen-Chi Hsu, Tyng-Luh Liu, Yubei Chen, and Yann LeCun. 2021. Decoupled contrastive learning. *ArXiv*, abs/2110.06848.

Hyperparameters	Pretraining & Fine-tuning
Epochs	30
Batch Size	4096
Queue Size	12288
m	0.99
τ	0.05
LR	5e-5
LR Decay	Linear
Warmup Steps	10%
Max Text Length	50
Weight Decay	0.01
Dropout Rate	0.1
Image Size	224(PT), 336(FT)

Table 6: The details of our hyperparameters.

Models	#Params
ViT-B/16	86.4M
BERT-base	109M

Table 7: Number of parameters.

A Hyperparameters and Implementation Details

Table 6 lists the hyperparameters in our experiments. Table 7 lists the number of parameters in ViT and BERT. All our experiments are conducted on 8 A100 GPUs. The average pre-training time for each model is about 60 hours (5.3M) or 100 hours (9.5M).