Fusion or Defusion? Flexible Vision-and-Language Pre-Training

Rongyi Sun[1]*, **Ziran Li**[2]*, **Yifeng Ding**[1], **Qifan Wang**[3],
Jingang Wang[2]†, **Hai-Tao Zheng**[1,4]†, **Wei Wu**[2] and **Yunsen Xian**[2]
[1]Tsinghua Shenzhen International Graduate School, Tsinghua University
[2]Meituan [3]Meta AI [4]Peng Cheng Laboratory, Shenzhen, China
{sry20,dingyf20}@mails.tsinghua.edu.cn, wqfcr@fb.com
{liziran02,wangjingang02,xianyunsen}@meituan.com
zheng.haitao@sz.tsinghua.edu.cn, wuwei19850318@gmail.com

Abstract

Existing approaches in the vision-and-language pre-training (VLP) paradigm mainly deploy either fusion-based encoders or dual-encoders, failing to achieve both effectiveness and efficiency in downstream multimodal tasks. In this paper, we build a flexible VLP model by incorporating cross-modal fusions into a dual-encoder architecture, where the introduced fusion modules can be easily decoupled from the dual encoder so as to switch the model to a fusion-free one. To better absorb cross-modal features from the fusion modules, we design a cross-modal knowledge transfer strategy along with other comprehensive pre-training tasks to guide the training process, which can further strengthen both the fusion-based and fusion-free representation learning. Extensive experiments conducted on various downstream vision-language tasks show that our proposed model is well-equipped with effectiveness as well as efficiency, demonstrating a superior performance compared with other strong VLP models.

1 Introduction

With the great development of self-supervised pre-training in both the community of natural language processing (Devlin et al., 2019; Raffel et al., 2020) and computer vision (Dosovitskiy et al., 2021; Bao et al., 2022a), recent researches have also witnessed the success of Vision-and-Language Pre-training (VLP). VLP learns generic multimodal representations from large-scale image-text pairs and can be further finetuned on various downstream Vision-Language (VL) tasks, including image-text retrieval (Lin et al., 2014), visual question answering (Goyal et al., 2017), visual reasoning (Suhr et al., 2019) and visual entailment (Xie et al., 2019).

The core of VLP resides in modeling the interaction between image and text representations. Most of the mainstreams first represent the input image via pre-trained deep feature extractors, then feed the derived visual features along with the text embeddings into multi-layer Transformers(Vaswani et al., 2017), in which cross-modal attention is used to fuse multimodal representations. Despite demonstrating superior performances on downstream VL tasks, the fusion-based methods need to jointly encode image and text representations, significantly degrading the efficiency in retrieval tasks with massive candidates of image-text pairs.

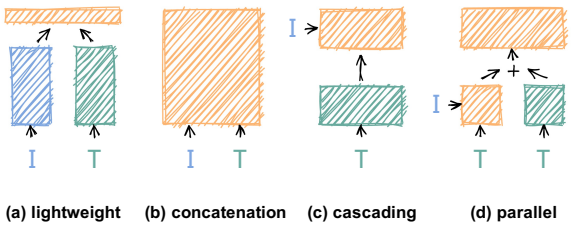

Figure 1: Different designs for vision-language fusions based on multi-head attention, “I" and “T" are short for image and text respectively. (a) Lightweight: only a few or even no parameters are used for VL fusions. (b) Concatenation: multi-head cross-attentions are applied to fuse the concatenation of image and text. (c) Cascading: first uses self-attentions to fully encode the unimodal input, then fuses the encoded features via cross-attentions. (d) Parallel: self-attentions and cross-attentions are independently calculated.

To make VLP models applicable in real-world scenarios, another line of methods independently encode text and image with dual encoders, shown in Fig. 1(a), in which cross-modal fusion is conducted by lightweight modules such as dot production. Thanks to the dual-encoder architecture, encoded features of image and text can be precomputed offline for inference efficiency. Nevertheless, independent encoding with shallow interaction fails to fully exploit the cross-modal interaction, making the performance far from satisfactory in VL classification tasks that require a strong ability of multimodal reasoning.

There are some recent works that attempt to keep

*Equal contribution.
†Corresponding authors.

Findings of the Association for Computational Linguistics: ACL 2023, pages 5105–5119
July 9-14, 2023

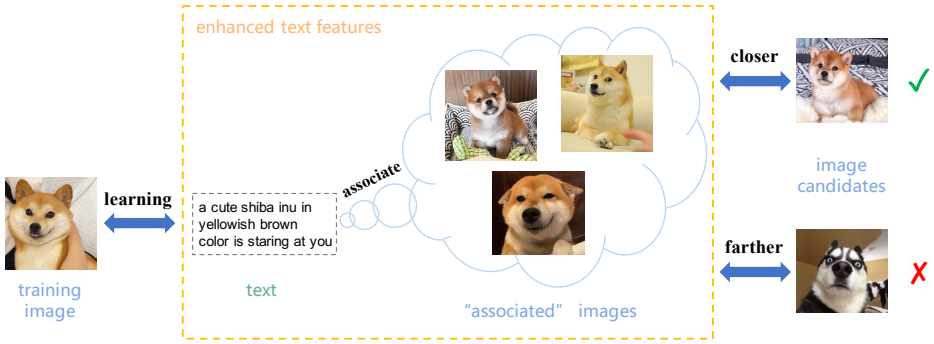

Figure 2: An example intuitively illustrates why a fusion-free text encoder can still work when the cross-modal fusions are removed from it. During training, the cross-modal fusions teach the text feature to "associate" what the related images could be. In inference of image-text retrieval, since a relevant image candidate is naturally closer to other images that are also related to the same text, the text feature along with its "associated" images will become closer to the relevant candidates than the original text feature.

both effectiveness and efficiency in downstream VL tasks. In particular, Wang et al. (2021b) empower a dual-encoder model by distilling knowledge from a fusion-based model. Although the distilled dual-encoder learns useful knowledge from cross-modal fusions while keeping its efficiency, this kind of method needs to pre-train a fusion-based model as a teacher and the performance is severely limited by the ability of the teacher model. VLMo (Bao et al., 2022b) introduces mixture-of-experts to encode various modalities with a modality-agnostic Transformer, which can be used as either a fusion encoder or a dual encoder. However, to fully train such a sparse model with experts towards different modalities, not only the image-text pairs but also massive images and text are required.

In this paper, we propose a unified and flexible VLP model named FOD, which incorporates cross-modal fusions into a dual-encoder architecture for achieving both efficacy and efficiency in multimodal scenarios. Specifically, we adopt a dual architecture with one image encoder and one text encoder, in which cross-modal fusions are placed in the text encoder side. Considering that conventional fusions are based on either concatenation (Kim et al., 2021; Singh et al., 2022) or cascading (Li et al., 2021a; Dou et al., 2022) that can't be directly decoupled from the boarding encoder, we employ a parallel-style fusion module to model cross-modal interactions, shown in Fig. 1. In this way, FOD can explicitly capture the complex interaction between modalities during training while switching the fusion-based text encoder to a fusion-free one by removing the fusion module.

In order to retain more cross-modal knowledge in FOD when the fusion modules are removed, we further design a cross-modal knowledge transfer strategy that forces both the unimodal features of image and text to approximate the multimodal representation produced by the fusion-based encoder. Intuitively, since paired image and text describe the same object in different views, we can naturally associate a set of relevant images when given a caption (and vice versa). Thus, if the text feature learns to "associate" its related images and absorbs them to enhance itself, the enhanced text feature can become closer to the relevant image candidates (and also farther to the unrelated ones) in inference. A concrete example illustrating this intuition is shown in Fig. 2.

We evaluate our model on both image-text retrieval tasks and vision-language understanding tasks. Experimental results show that our model outperforms other VLP methods on all downstream VL tasks, and even performs competitively with models that use a larger order of magnitude of data for pre-training. Thanks to the detachable fusion module and the strategy of knowledge transfer, our model can be flexibly switched to a fusion-free pattern to enjoy a much faster inference speed of retrieval while retaining most of the performance.

2 Related Work

Without considering the ways of visual feature extraction, the approaches of vision-language pre-training can be divided into two categories based on the interaction form between image and text. The first category, fusion-based model, explicitly utilizes deep fusion layers with cross-modal attention to model the interaction of images and texts (Tan and Bansal, 2019; Lu et al., 2019; Su et al., 2019; Li et al., 2019; Chen et al., 2020; Li et al.,

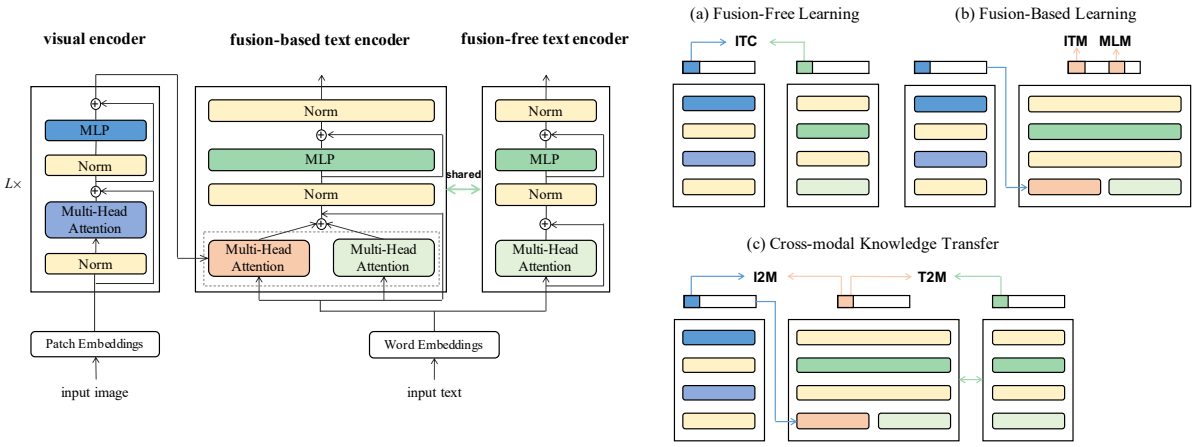

Figure 3: Overview of FOD, which consists of an image encoder and a flexible text encoder. The text encoder can be switched between fusion-based pattern and fusion-free pattern. We utilize three complementary learning strategies to jointly train FOD: Fusion-free learning with Image-Text Contrastive learning (ITC), Fusion-based learning with Image-Text Matching (ITM) and Mask Language Modeling (MLM), and Cross-modal Knowledge Transfer with Image-to-Multimodal (I2M) and Text-to-Multimodal (T2M) learning.

2020, 2021b; Gan et al., 2020; Zhang et al., 2021; Huang et al., 2020, 2021; Kim et al., 2021; Li et al., 2021a; Wang et al., 2021c; Li et al., 2022; Zeng et al., 2022; Wang et al., 2022). These models perform well on vision-language understanding tasks due to the ability of capturing deep cross-modal features. However, for vision-language retrieval tasks, the fusion-based methods need to encode all the possible image-text pairs to find the most relevant candidate, resulting in extremely high time cost.

The second category, dual-based model, utilizes a visual encoder and a text encoder to separately encode images and text, while the interaction between images and text is modeled by cosine similarity or linear projection (Radford et al., 2021; Jia et al., 2021; Yao et al., 2021). Although dual-based models are effective for retrieval tasks since features can be pre-computed and cached offline, the shallow interaction is insufficient to tackle the vision-language understanding tasks that require complex VL reasoning. Besides, training a dual-based model often necessitates a large number of image-text pairs (e.g. 300M for Filip (Yao et al., 2021) and 1.8 B for ALIGN (Jia et al., 2021)).

Recently, some researchers have devoted themselves to investigating a unified model that is well-performed on vision-language understanding tasks while maintaining the efficiency towards retrieval tasks (Wang et al., 2021b; Liu et al., 2021; Wang et al., 2021a; Bao et al., 2022b; Dou et al., 2022). To achieve this, one line of the works leverage knowledge distillation, in which a fusion-encoder model is pre-trained as a teacher model to guide the training of a dual-encoder model (Wang et al., 2021b), but the performance is inevitably limited by the teacher model. Other efforts attempt to train a modality-agnostic encoder with shared parameters, which can be used as either a fusion encoder or a dual encoder (Wang et al., 2021a; Bao et al., 2022b). Despite the benefits of modeling all the modalities into a single encoder, it is hard to fully train such a huge model and a large number of training samples in different modalities are required. Different from these methods, we incorporate a detachable cross-modal fusion module into a dual-encoder architecture, which can easily remove the fusion module in inference and switch to a fusion-free model. More importantly, our model does not rely on teacher models or massive data in other modalities.

3 Model Architecture

As shown in Fig. 3, FOD is in a transformer-based dual-encoder architecture that includes a visual encoder and a text encoder. The text encoder can be flexibly switched between a fusion-based pattern and a fusion-free pattern. For the fusion-based pattern, cross-modal fusions are incorporated into the text encoder to model multimodal interactions. For the fusion-free pattern, the fusion module is decoupled from the text encoder so as to get rid of the cross-modal calculation. During training, both fusion-based and fusion-free patterns are involved in the learning process, while in inference, the text encoder will be switched to one of the two patterns

according to the type of downstream tasks. In the following sections, we introduce the visual encoder and the two patterns of the text encoder, followed by the pre-training strategies.

3.1 Visual Encoder

We utilize Vision Transformer (Dosovitskiy et al., 2021) to build the visual encoder. Given a 2D image $I \in \mathbb{R}^{C\times H\times W}$, we first reshape I into a sequence of 2D image patches $V^p \in \mathbb{R}^{N\times(P^2\cdot C)}$, where (H, W) is the original image resolution, C is the number of channels, (P, P) is the patch resolution, and $N = HW/P^2$ is the number of patches.

$$V^p = [v_1^p; \cdots ; v_N^p]. \tag{1}$$

Then we flatten the patches and embed them to $V^e \in \mathbb{R}^{N\times D}$ with a trainable linear projection $\omega \in \mathbb{R}^{(P^2\cdot C)\times D}$, where D is the hidden size.

$$V^e = [v_1^p\omega; \cdots ; v_N^p\omega]. \tag{2}$$

We also prepend a learnable embedding $V_{cls}^e \in \mathbb{R}^D$ to the patch embeddings V^e. Besides, positional information is also important for path representations. Therefore, the embedded patches $\bar{V}$ are obtained by summing $[V_{cls}^e; V^e]$ and learnable 1D position embeddings $V_{pos} \in \mathbb{R}^{(N+1)\times D}$. Finally, we obtained visual features V by encoding $\bar{V}$ with the visual encoder VE.

$$\begin{aligned} \bar{V} &= [V_{cls}^e; v_1^e; \cdots ; v_N^e] + V_{pos}, \\ V &= \text{VE}(\bar{V}). \end{aligned} \tag{3}$$

3.2 Text Encoder

As mentioned before, there are two patterns of the text encoder: fusion-free text encoder and fusion-based text encoder. These two patterns are both based on Transformers (Vaswani et al., 2017) and share all the fusion-free parameters except the output linear projection in the last encoding layer.

Given the input text $t = \{t_{cls}; w_1; \cdots ; w_S\}$, we first embed t to $T^0 \in \mathbb{R}^{S\times D}$ via a word embedding matrix and a position embedding matrix. Then the text embedding T^0 can be fed into different patterns of the text encoder to produce different output features.

3.2.1 Fusion-free Text Encoder

In this pattern, the text encoder skips the cross-modal fusions and outputs text-only features. The text encoder is a L-layer Transformer, and the output of the l-th layer T^l is computed as follows:

$$\begin{aligned} T_s^l &= \text{MSA}(T^{l-1}, T^{l-1}, T^{l-1}), \\ \hat{T}^l &= \text{LN}(T_s^l + T^{l-1}), \\ T^l &= \text{LN}(\text{MLP}(\hat{T}^l) + \hat{T}^l), \\ T &= T^L, \end{aligned} \tag{4}$$

where MSA, LN and MLP are shot for Multi-Head Self-Attention, layer normalization and multi-layer perceptron respectively, T is the final features of the fusion-free text encoder.

3.2.2 Fusion-based Text Encoder

To fully capture vision-and-language interactions, both self-attention and cross-modal attention are considered in the fusion-based encoder. Specifically, in the l-th layer, we separately compute the fusion-free self-attention and the image-fused cross-attention, and then sum them up to produce the multimodal features. The detailed process is shown as follows:

$$\begin{aligned} M^0 &= T^0, \\ M_s^l &= \text{MSA}(M^{l-1}, M^{l-1}, M^{l-1}), \\ M_c^l &= \text{MCA}(M^{l-1}, V, V), \\ \tilde{M}^l &= \frac{1}{2} \times (M_s^l + M_c^l), \\ \hat{M}^l &= \text{LN}(\tilde{M}^l + M^{l-1}), \\ M^l &= \text{LN}(\text{MLP}(\hat{M}^l) + \hat{M}^l), \\ M &= M^L, \end{aligned} \tag{5}$$

where MCA is Multi-Head Cross Attention, V is the final visual features produced by the visual encoder. The MCA, LN and MLP modules are reused from the fusion-free text encoder. Notably, the cross-modal attention is introduced in a parallel manner, which is parameter-efficient and can be easily decoupled from the encoder.

In addition, the cross-modal fusions can also be placed in the visual side to build a fusion-based visual encoder, or be placed in both sides for deeper interaction. We will discuss this in the experiment section.

4 Pre-training Strategies

FOD is jointly trained with three different strategies, namely fusion-free learning, fusion-based learning and cross-modal knowledge transfer, which are complementary to each other.

4.1 Fusion-free Learning

For this strategy, we utilize image-text contrastive learning to train the dual architecture with the ability of unimodal encoding, which is not only beneficial to other cross-modal learning strategies, but also the basis for applying the model to downstream retrieval tasks.

4.1.1 Image-Text Contrast

We select V_{cls} and T_{cls} produced by visual encoder and fusion-free text encoder to compute the loss of contrastive learning. In order to have more negative examples here, we maintain two queues to store the most recent K image and text representations computed by momentum encoders like MoCo (He et al., 2020). For convenience, we denote these representations in queues as V_{cls}^k and T_{cls}^k, where $k \in \{1, \cdots, K\}$. For each image representation V_{cls}^j and text representation T_{cls}^j in the current batch, the image-to-text similarities p_j^{i2t} and text-to-image similarities p_j^{t2i} are computed by:

$$\begin{aligned} s_{j,k}^{i2t} &= g(f_v(V_{cls}^j))^\top g(f_t(T_{cls}^k)), \\ s_{j,k}^{t2i} &= g(f_t(T_{cls}^j))^\top g(f_v(V_{cls}^k)), \end{aligned} \tag{6}$$

$$p_j^{i2t} = \frac{\exp(s_{j,j}^{i2t}/\sigma)}{\sum_{k=1}^K \exp(s_{j,k}^{i2t}/\sigma)}, \; p_j^{t2i} = \frac{\exp(s_{j,j}^{t2i}/\sigma)}{\sum_{k=1}^K \exp(s_{j,k}^{t2i}/\sigma)}, \tag{7}$$

where f_v and f_t are linear projections, g is L2 normalization, and σ is a learnable temperature parameter. Let $\mathbf{y}^{i2t}$ and $\mathbf{y}^{t2i}$ denote the ground-truth ont-hot similarity, where positive pairs have a probability of 1 and negative pairs have a probability of 0. The image-text contrastive loss $\mathcal{L}_{\text{itc}}$ is defined as the cross-entropy $\mathcal{H}$ between p and y:

$$\mathcal{L}_{\text{itc}} = \frac{1}{2} \times \left[\mathcal{H}(y^{\text{i2t}}, p^{\text{i2t}}) + \mathcal{H}(y^{\text{t2i}}, p^{\text{t2i}})\right]. \tag{8}$$

4.2 Fusion-based learning

For this strategy, we apply image-text matching (ITM) and mask language modeling (MLM) to the fusion-based text encoder for learning both coarse-grained and fine-grained cross-modal fusions.

4.2.1 Image-Text Matching

ITM focuses on coarse-grained multimodal learning, which aims to predict whether a pair of image and text is matched or not. Since the image-text pairs in a batch are all positive, we sample global hard negative image-text pairs from all input batches on all the GPUs based on the similarity scores calculated in Eq. 7. Then we feed the final hidden vector of the fusion-based encoder M_{cls} into a binary classifier to predict a two-class probability p^{itm}. Given the ground-truth label $y^{\text{itm}} \in \{0, 1\}$, the image-text matching loss $\mathcal{L}_{\text{itm}}$ is defined as the cross-entropy $\mathcal{H}$ between y^{itm} and p^{itm}:

$$\mathcal{L}_{\text{itm}} = \mathcal{H}(y^{\text{itm}}, p^{\text{itm}}). \tag{9}$$

4.2.2 Masked Language Modeling

MLM predicts masked tokens on the image-fused text features, which serves as the fine-grained cross-modal learning. Formally, we randomly mask 15% of the tokens in the text sequence t with a whole word masking strategy (Cui et al., 2021) and denote the input embedding of the masked text as $\bar{T}^0$. Then the model is trained to predict the masked tokens based on the final outputs $\bar{M}$ by feeding $\bar{T}^0$ into the fusion-based encoder. The detailed process is similar to Eq. 5. Let y^{mask} denote the ground-truth label of the masked tokens, and p^{mask} denote the models' prediction for the masked tokens, then the masked language modeling loss is defined as the cross-entropy $\mathcal{H}$ between y^{mask} and p^{mask}:

$$\mathcal{L}_{\text{mlm}} = \mathcal{H}(y^{\text{mask}}, p^{\text{mask}}). \tag{10}$$

4.3 Cross-modal Knowledge Transfer

In our preliminary experiments, we observe that if the ITM loss is removed from the training process, the performance in retrieval tasks would dramatically degrade. From the perspective of feature distributions, we believe that ITM can better close the spatial distance between the unimodal features of image and text, which encourages us to explicitly utilize ITM to enhance unimodal representations.

To achieve this, we further design the strategy of cross-modal knowledge transfer (CKT). Given an image-text pair, we can first extract its image V_{cls}, text T_{cls} and multimodal representations M_{cls}. Obviously, M_{cls} is the most comprehensive feature that describes the image-text pair among them, but only V_{cls} and T_{cls} are used to compute similarity score in retrieval tasks. In this case, if we enhance the text feature to actively associate its related images by transferring knowledge from M_{cls} to T_{cls}, it will be easier to find the relevant image candidates based on the enhanced text feature in inference (and similar for V_{cls}). Thus, we force both V_{cls} and T_{cls} to approximate M_{cls} via mean-squared loss in the last layer, which are calculated as follows:

$$\begin{aligned} \mathcal{L}_{\text{I2M}} &= MSE(f_v(V_{cls}), f_t(M_{cls})), \\ \mathcal{L}_{\text{T2M}} &= MSE(f_t(T_{cls}), f_t(M_{cls})), \end{aligned} \tag{11}$$

Model	# Pretrain Images	MSCOCO (5K) Text Retrieval R@1	R@5	R@10	Image Retrieval R@1	R@5	R@10	Flickr30k (1K) Text Retrieval R@1	R@5	R@10	Image Retrieval R@1	R@5	R@10
UNITER-B	4M	64.4	87.4	93.1	50.3	78.5	87.2	85.9	97.1	98.8	72.5	92.4	96.1
OSCAR-B	4M	70.0	91.1	95.5	54.0	80.8	88.5	-	-	-	-	-	-
ViLT-B	4M	61.5	86.3	92.7	42.7	72.9	83.1	83.5	96.7	98.6	64.4	88.7	93.8
Inference based on "Dual" setting													
ALIGN	1.2B	77.0	93.5	96.9	59.9	83.3	89.8	95.3	99.8	100.0	84.9	97.4	98.6
Distill	4M	-	-	-	-	-	-	82.2	96.7	98.5	68.2	89.8	94.2
ALBEF	4M	65.9	88.5	93.8	49.1	76.4	84.9	89.7	98.5	99.7	74.5	93.2	96.3
X-VLM	4M	71.4	91.9	96.4	54.5	81.6	88.9	90.2	99.1	99.7	78.4	95.2	97.8
VLMo-B	4M	74.8	93.1	**96.9**	57.2	82.6	89.8	92.3	99.4	99.9	79.3	95.7	97.8
Ours	3M	**77.3**	**94.3**	**96.9**	**58.9**	**83.2**	**90.0**	**94.6**	**99.7**	**99.9**	**83.5**	**96.4**	**98.1**
Inference based on "Re-Rank" setting													
BLIP†	129M	81.2	95.7	97.9	64.1	85.8	91.6	97.2	99.9	100.0	87.5	97.7	98.9
ALBEF†	4M	73.1	91.4	96.0	56.8	81.5	89.2	94.3	99.4	99.8	82.8	96.7	98.4
X-VLM†	4M	80.4	95.5	**98.2**	63.1	85.7	91.6	96.8	99.8	**100.0**	86.1	97.4	98.7
Ours†	3M	**82.2**	**95.8**	97.9	**65.2**	**86.4**	**91.9**	**97.4**	**100.0**	**100.0**	**87.3**	**97.7**	**98.9**

Table 1: Fine-tuned image-text retrieval results on MSCOCO (5K test set) and Flickr30K (1K test set). † inference is based on the "Re-Rank" setting. The bold numbers denote the best results of methods that are pre-trained with the standard 4M data.

Model	# Pretrain Images	Flickr30k (1K) Text Retrieval R@1	R@5	R@10	Image Retrieval R@1	R@5	R@10
UNITER	4M	83.6	95.7	97.7	68.7	89.2	93.9
ViLT	4M	69.7	91.0	96.0	51.3	79.9	87.9
Inference based on "Dual" setting							
CLIP	400M	88.0	98.7	99.4	68.7	90.6	95.2
ALIGN	1.2B	88.6	98.7	99.7	75.7	93.8	96.8
ALBEF	4M	81.3	96.4	98.3	67.9	89.2	93.8
X-VLM	4M	84.7	97.9	99.3	72.5	92.7	96.3
VLMo	4M	88.2	97.9	99.4	73.4	92.9	96.6
Ours	3M	**89.8**	**98.8**	**99.7**	**77.5**	**94.5**	**97.1**
Inference based on "Re-Rank" setting							
ALBEF†	4M	90.5	98.8	99.7	76.8	93.7	96.7
X-VLM†	4M	94.1	99.3	99.9	82.3	96.1	98.0
Ours†	3M	**95.5**	**99.8**	**100.0**	**84.5**	**96.2**	**98.2**

Table 2: Zero-Shot image-text retrieval results on Flickr30K (1K test set). † inference is based on the "Re-Rank" setting.

Model	# Pretrain Images	VQAv2 test-dev	test-std	NLVR2 dev	test-P
SimVLM	1.8B	77.87	78.14	81.72	81.77
BLIP	129M	78.25	78.32	82.15	82.24
UNITER	4M	72.70	72.91	77.18	77.85
OSCAR	4M	73.16	73.44	78.07	78.36
ViLT	4M	71.26	-	75.70	76.13
Distill	4M	68.05	-	74.16	74.30
ALBEF	4M	74.54	74.70	80.24	80.50
VLMo	4M	76.64	76.89	82.77	83.34
X-VLM	4M	78.07	78.09	84.16	84.21
Ours	3M	**78.91**	**78.91**	**84.75**	**85.29**

Table 3: Results on vision-language understanding tasks, including visual question answering (VQAv2) and visual reasoning (NLVR2).

where f_v and f_t are the linear projections used in Eq. 6. We do not freeze M_{cls} in knowledge transfer so that multimodal and unimodal features can be jointly trained.

5 Experiment

5.1 Pre-training Settings

5.1.1 Datasets

Following previous works (Chen et al., 2020; Kim et al., 2021), we use four well-known image captioning datasets for pre-training: SBU Captions (Ordonez et al., 2011), Microsoft COCO (Lin et al., 2014), Visual Genome (Krishna et al., 2017) and Google Conceptual Captions (GCC) (Sharma et al., 2018). Since images in GCC and SBU are provided in url format and some of them are inaccessible, we only collected **3.4M** images, which is around **600K less** than the original settings. In the experiments, we term the setting of 3.4M images as **3M**.

5.1.2 Implementation Details

For model settings, the visual encoder adopts the same architecture as ViT-Base (Dosovitskiy et al., 2021) and we initialize it with pre-trained weights of Beit (Bao et al., 2022b). The text encoder is modified on Bert-Base (Devlin et al., 2019) by adding a multi-head cross attention and we initialize it with pre-trained weights of uncased-bert-base. For hyper-parameter settings during pre-training, the resolution of input images is 256×256 and the patch size is 16×16. RandAugment (Cubuk et al., 2020) is applied to the input images. We use AdamW optimizer (Loshchilov and Hutter, 2017) with weight decay of 1e-2 and the learning rate is

Fusion Methods	MSCOCO (5K)		Flickr30k (1K)	
	TR@1	IR@1	TR@1	IR@1
Concatenation	72.5	54.2	92.6	80.5
Cascading	73.0	54.5	91.7	81.2
Parallel	**73.5**	**55.4**	**93.1**	**81.6**

Table 4: Ablation study of different fusion methods. Results are obtained by only pre-training models with 50K steps.

Objectives		MSCOCO (5K)		VQAv2	NLVR2
I2M	T2M	TR	IR	test-dev	test-P
✗	✗	86.3	73.3	77.56	**83.45**
✗	✓	86.6	74.1	-	-
✓	✗	86.8	73.2	-	-
✓	✓	**87.2**	**74.3**	**77.57**	83.37

Table 5: Ablation study of cross-modal knowledge transfer. For retrieval tasks, we report the average of R@1, R@5 and R@10.

warmed up to 1e-4 over the first 1k steps. We pre-train for 300K steps on 32 NVIDIA A100 GPUs with a batch size of 2048.

5.2 Downstream Vision-Language Tasks

5.2.1 Image-Text Retrieval Tasks

The vision-language retrieval tasks include image-to-text retrieval and text-to-image retrieval. We evaluate our model on the Karpathy and Fei-Fei (2015) split of MSCOCO (Lin et al., 2014) and Flickr30K. During fine-tuning, we preserve the loss of image-text contrastive learning, image-text matching and cross-modal knowledge transfer. For a better comparison with various methods, we have two settings in the inference phase, namely "Dual" and "Re-Rank".

For the "Dual" setting, we use Eq. 6 to pre-compute images and text representations separately, and compute the similarity scores of all possible image-text pairs by dot production. For the "Re-Rank" setting, we first utilize the similarity scores derived from Eq. 6 to select the top-k candidates, and then predict the final results by calculating their ITM scores (p^{itm}).

5.2.2 Visual Question Answering

The VQAv2 (Goyal et al., 2017) task requires to predict answers based on the given pair of an image and a question. Following Cho et al. (2021) and Li et al. (2021a), we treat VQA as an answer generation problem. In order to compare fairly with other methods, we restrict the answer generation space

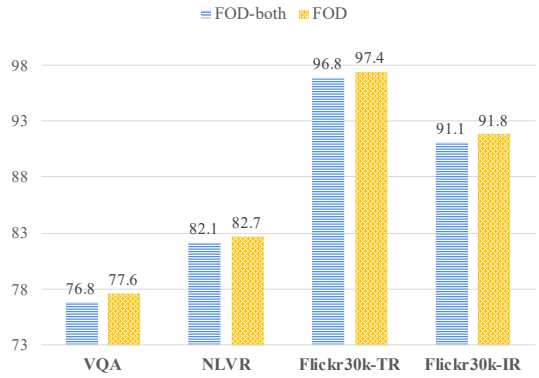

Figure 4: Ablation study of placing fusions in both text and visual encoders. FOD-both: fusions are added on both sides.

to the same candidate set (Kim et al., 2021; Bao et al., 2022b) during inference.

5.2.3 Natural Language for Visual Reasoning

The NLVR2 (Suhr et al., 2019) task asks the model to predict whether a text correctly describes a pair of images. We follow previous work (Li et al., 2021a; Zeng et al., 2022) to extend the fusion-based encoder to enable reasoning over image pairs and feed the encoded vector of the input pair into a classification layer to predict answer.

5.3 Main Results

5.3.1 Image-Text Retrieval Results

Table 1 and Table 2 show the results of fine-tuned and zero-shot image-text retrieval on MSCOCO and Flickr30K. For a fair comparison, only base-size models pre-trained on the standard 4M data are selected as the compared models. In this setting, our model achieves state-of-the-art performance on both datasets, and even performs competitively with CLIP, ALIGN and BLIP that are pre-trained on a larger order of magnitude of data. Furthermore, thanks to the designed parallel-style fusions and cross-modal knowledge transfer strategy, more cross-modal knowledge is retrained when the fusion module is decoupled in inference, narrowing the gap between "Dual" and "Re-Rank" settings. Detailed analysis of performances between "Dual" and "Re-Rank" settings are given in Appendix.

5.3.2 Vision-Language Understanding Results

The VQA2 and NLVR2 are categorized as understanding tasks since they both require the ability of VL reasoning, and the results are shown in Table 3. Our model achieves the best performances on both tasks among all the competitors that are also

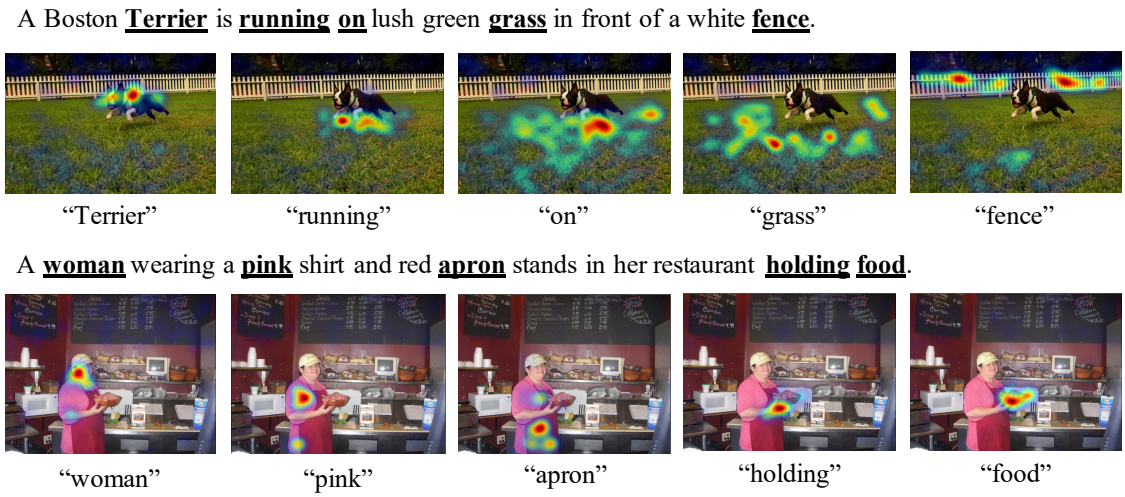

Figure 5: Grad-CAM visualizations of cross-modal attention maps according to different words on VL retrieval.

in base-size and pre-trained with the standard 4M data, and even outperforms models pre-trained on more data like SimVLM and BLIP, demonstrating the effectiveness and efficiency of our model.

5.4 Ablation Studies

5.4.1 Different Designs of Fusions

We incorporate cross-modal fusions into a dual architecture to better model vision-language interactions. Conventional fusions are based on two kinds of methods, namely concatenation and cascading. (1) Concatenation jointly encodes the concatenation of image and text, which is quadratic in time complexity and twice in memory consumption. (2) Cascading first uses self-attention to encode the text input and then fuses it with image via cross-attentions, which has a strong dependency between cross-attention and self-attention. Table 4 reports the ablation results of different fusions, our design that incorporates cross-modal fusions in a parallel manner outperforms other methods on retrieval tasks, showing that parallel-style fusion can switch our model into the "Dual" setting more flexibly.

5.4.2 Cross-modal Knowledge Transfer

We conduct the ablation experiments towards the strategy of cross-modal knowledge transfer, which is shown in Table 5. The objectives of I2M and T2M are defined in Eq. 11. From the results we can observe that: (1) I2M specifically improves the performance on image-text retrieval (TR) while T2M is beneficial for the text-image (IR) side, which are consistent with their intuitions; (2) I2M and T2M are complementary to each other. Adding both I2M and T2M during training can further bring improvements for retrieval tasks while keeping the performances on VL understanding tasks.

5.4.3 Fusions on Both Sides

Intuitively, in addition to placing cross-modal fusions in the text encoder, we can also add the fusion modules into the visual side in a similar way. In this setting, ITM and downstream classifications are based on the concatenation of the multimodal features produced by both text and visual encoders. Fig. 4 shows the results of placing fusions in different sides, from which we find that when fusions are placed on both sides, the performance unexpectedly drops on all downstream tasks. We analyze that one possible reason comes to the difference between text and vision in self-supervised learning. It is obvious that BERT naturally works better in self-supervision than ViT, and thus we can utilize the MLM task from BERT to learn fine-grained cross-modal interaction. When it comes to the visual side, self-supervised tasks are much more complex than MLM, inevitably making it more difficult to train such a VLP model.

6 Qualitative Analysis

We further provide a qualitative analysis by using Grad-CAM (Selvaraju et al., 2017) to illustrate the per-word visualizations of the cross-modal attention maps of the fusion-based encoder. As shown in Fig. 5, from the visualizations we observe that when conducting image-text matching tasks, our model can focus on specific regions in an image according to different words in each sentence, including objects, actions, attributes and background. More examples are given in Appendix.

7 Conclusion

In this work, we propose a flexible VLP model that incorporates cross-modal fusions into a dual-encoder architecture. The fusion module can be easily decoupled in inference, enabling the model to be switched between fusion-based and a fusion-free patterns according to different scenarios. Extensive experiments conducted on both image-text retrieval and vision-language understanding tasks show that our model is well-equipped with effectiveness and efficiency compared with existing VLP models.

Limitations

The findings of this study have to be seen in light of some limitations. (1) It is non-trivial to extend our model for generation tasks. Since the main focus of this work is to improve both effectiveness and efficiency of the dual-encoders, text-decoder is not considered in model design. In the future, autoregressive mechanisms will be consider to applied in model architecture so that the model can be directly used for generation tasks like image captioning. (2) There may be disadvantages of the model in region-level VL tasks such as Object Detection. The reason is that these tasks require images in high resolution and fine-grained annotations of bounding boxes, which are non-trivial in generic VLP settings. To solve this problem, exploring different levels of granularity between image-text pairs is a promising direction and will be considered as the future work.

Acknowledgements

This research is supported by National Natural Science Foundation of China (Grant No.62276154), Research Center for Computer Network (Shenzhen) Ministry of Education, Beijing Academy of Artificial Intelligence (BAAI), the Natural Science Foundation of Guangdong Province (Grant No. 2023A1515012914), Basic Research Fund of Shenzhen City (Grant No.JCYJ20210324120012033 and JSGG20210802154402007), the Major Key Project of PCL for Experiments and Applications (PCL2021A06), and Overseas Cooperation Research Fund of Tsinghua Shenzhen International Graduate School (HW2021008). Jingang Wang is funded by Beijing Nova Program (Grant No.20220484098).

References

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. 2022a. BEiT: BERT pre-training of image transformers. In *International Conference on Learning Representations*.

Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, Songhao Piao, and Furu Wei. 2022b. VLMo: Unified vision-language pre-training with mixture-of-modality-experts. In *Advances in Neural Information Processing Systems*.

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

Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. 2021. Unifying vision-and-language tasks via text generation. In *International Conference on Machine Learning*, pages 1931–1942. PMLR.

Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. 2020. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 702–703.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. 2021. Pre-training with whole word masking for chinese bert. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3504–3514.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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.

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. *ICLR*.

Zi-Yi Dou, Aishwarya Kamath, Zhe Gan, Pengchuan Zhang, Jianfeng Wang, Linjie Li, Zicheng Liu, Ce Liu, Yann LeCun, Nanyun Peng, Jianfeng Gao, and Lijuan Wang. 2022. Coarse-to-fine vision-language pre-training with fusion in the backbone. In *Advances in Neural Information Processing Systems*.

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. *Advances in Neural Information Processing Systems*, 33:6616–6628.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.

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

Zhicheng Huang, Zhaoyang Zeng, Yupan Huang, Bei Liu, Dongmei Fu, and Jianlong Fu. 2021. 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.

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 preprint arXiv:2004.00849*.

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

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5583–5594. PMLR.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73.

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

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021a. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: Asimple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. 2021b. Unimo: Towards unified-modal understanding and generation via cross-modal contrastive learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2592–2607.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer.

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

Haoliang Liu, Tan Yu, and Ping Li. 2021. Inflate and shrink: Enriching and reducing interactions for fast text-image retrieval. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9796–9809.

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

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

Vicente Ordonez, Girish Kulkarni, and Tamara Berg. 2011. Im2text: Describing images using 1 million captioned photographs. In *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 618–626. IEEE Computer Society.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565.

Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. 2022. FLAVA: A foundational language and vision alignment model. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 15617–15629. IEEE.

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

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. 2019. A corpus for reasoning about natural language grounded in photographs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6418–6428.

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Jianfeng Wang, Xiaowei Hu, Zhe Gan, Zhengyuan Yang, Xiyang Dai, Zicheng Liu, Yumao Lu, and Lijuan Wang. 2021a. Ufo: A unified transformer for vision-language representation learning. *arXiv preprint arXiv:2111.10023*.

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022. OFA: unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 23318–23340. PMLR.

Zekun Wang, Wenhui Wang, Haichao Zhu, Ming Liu, Bing Qin, and Furu Wei. 2021b. Distilled dual-encoder model for vision-language understanding. *arXiv preprint arXiv:2112.08723*.

Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. 2021c. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*.

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

Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. 2021. Filip: Fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783*.

Yan Zeng, Xinsong Zhang, and Hang Li. 2022. Multi-grained vision language pre-training: Aligning texts with visual concepts. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 25994–26009. PMLR.

Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5579–5588.

A Appendix

A.1 Statistics of Pre-training Datasets

In the expriments, we use four widely-used datasets for pre-training: SBU Captions (Ordonez et al., 2011), Microsoft COCO (Lin et al., 2014), Visual Genome (Krishna et al., 2017) and Google Conceptual Captions (Sharma et al., 2018). Due to the inaccessible problem of url, we can only collected 3.4M images, which is around 600K less than the original settings (Kim et al., 2021; Li et al., 2021a; Bao et al., 2022b). Details are shown in Table 6. Intuitively, if we can access to the full 4M data, our model could perform better.

	MSCOCO	VG	SBU	GCC	Sum
Original	113K	108K	867K	3.01M	4M
Ours	113K	108K	853K	2.36M	3.4M

Table 6: Comparison of # Images used in Pre-training between official settings and ours.

A.2 Implementation Details

Image-Text Retrieval. Different from pre-training, we set the resolution of images to 384×384 and use the tasks of ITC, ITM and CKT in image-text retrieval. The batch size is 256 and the initial learning rate is 1e-5. For Flickr30K and MSCOCO, we finetune 10 epochs and 5 epochs respectively. For the "Re-Rank" setting that selects top-k candidates, k is set to 128 for Flickr30K and 256 for MSCOCO following Li et al. (2021a) and Zeng et al. (2022).

Visual Question Answering. For visual question answering, most methods convert VQAv2 to a classification task by preserving the most frequent 3192 answers in datasets. However, this will prevent some data from being used for fine-tuning because their answers are not in the candidate set. Thus, we follow previous work (Cho et al., 2021; Li et al., 2021a; Zeng et al., 2022) and treat VQA as an answer generation problem. More specifically, we predict the probability distribution on the vocabulary of the first token, and select the top-k candidates with the highest probability from the distribution. Finally, we use language-modeling loss to predict the final answer from the top-k candidates. For a fair comparison, we restrict the answer generation space to the same candidate set (Kim et al., 2021; Bao et al., 2022b) during inference. We finetune our model for 8 epochs with 256 batch size and the learning rate is 2e-5. The resolution of images is set to 576×576 (Dou et al., 2022) and k is set to 128.

Natural Language for Visual Reasoning. For NLVR2, we follow previous work (Li et al., 2021a; Zeng et al., 2022) and extend the fusion-based encoder to enable reasoning over image pairs, in which an additional pre-training step is applied for training model to reason the relations among text and images. Then, we fine-tune the model for 15 epochs. The batch size is 128, learning rate is 2e-5 and the resolution of the input image is set to 384×384.

A.3 Performance Retaining

For VL retrieval tasks that involve massive candidates of image-text pairs, it is crucial for a VLP model to have the ability of acting as a dual-encoder for efficient inference. Table 7 reports the comparisons between "Re-Rank" and "Dual" settings on retrieval tasks. Our model performs best in terms of performance retraining when switched from "Re-Rank" to "Dual" setting, showing the effectiveness of the designed parallel-style fusions and cross-modal knowledge transfer strategy.

Model	Flickr30k			MSCOCO		
	R	D	drop↓	R	D	drop↓
ALBEF	95.2	92.0	3.2 (3.4%)	81.3	76.4	4.9 (6.0%)
X-VLM	96.5	93.4	3.1 (3.2%)	85.8	80.8	5.0 (5.8%)
Ours	**96.9**	**95.4**	**1.5 (1.5%)**	**86.6**	**83.4**	**3.2 (3.7%)**

Table 7: Results of different retrieval settings on MSCOCO (5K) and Flickr30k (1K). "R" and "D" are short for "Re-Rank" and "Dual" settings. We report the average of TR and IR.

A.4 Inference Speed

We further evaluate the inference time of our models and other compared methods on MSCOCO dataset. All the models are evaluated on a single A100 GPU. From the results reported in Table 8, we can observe that our model is well-equipped both efficacy and efficiency in retrieval tasks. Notably, when our model is switched to the fusion-free (dual) pattern, it can still achieve a comparable performance compared with other methods while enjoy a much faster inference speed.

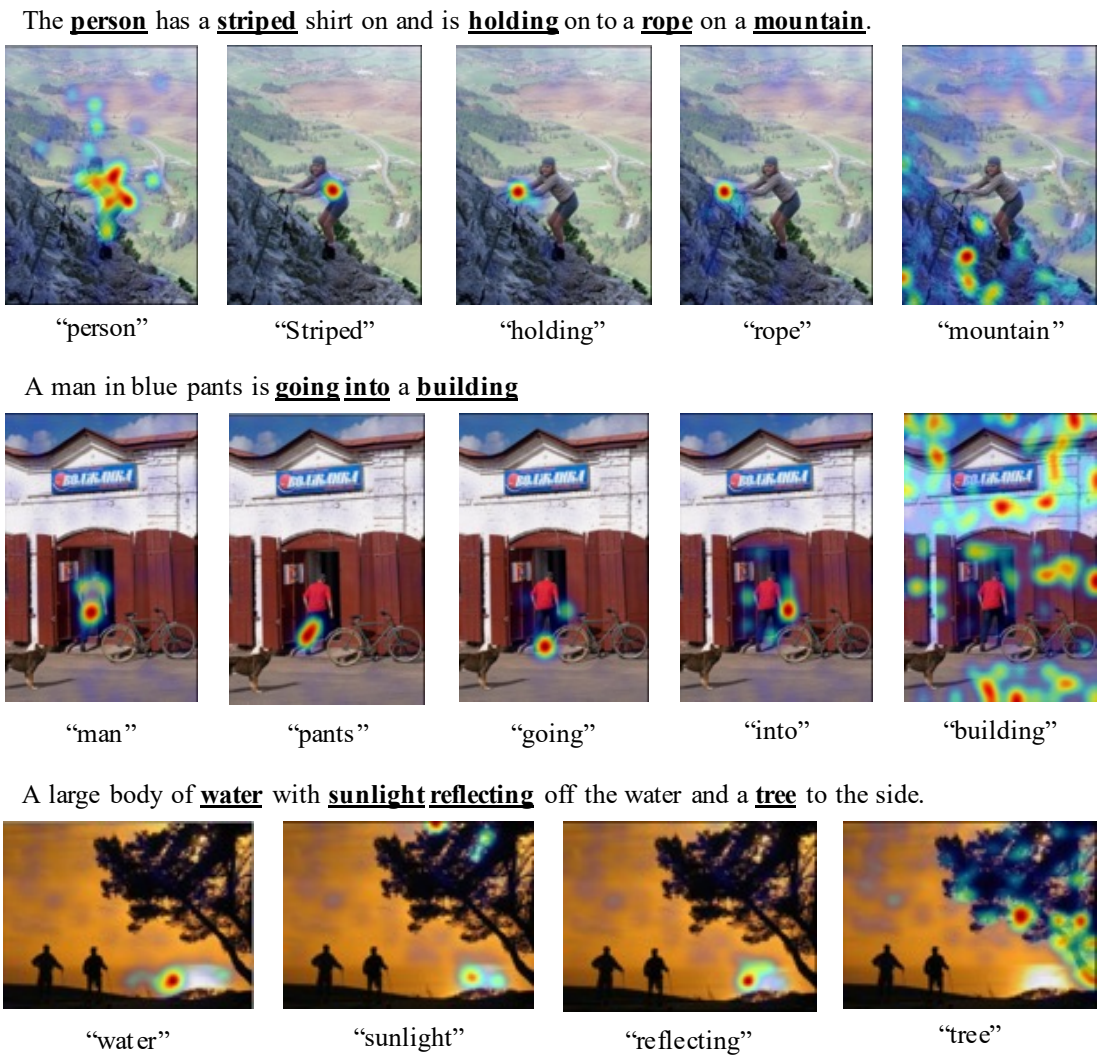

Figure 6: More examples of visualizations of cross-modal attention maps according to different words on VL retrieval.

Model	Inference Time	Speedup	MSCOCO TR@1	MSCOCO IR@1
OSCAR-B‡	-	$\ll 1.0\times$	70.0	54.0
ViLT-B	$\sim$ 10h	$1.0\times$	61.5	42.7
ALBEF†	$\sim$ 900s	$40\times$	73.1	56.8
VLMo-B	$\sim$ 30s	$1,200\times$	74.8	57.2
Ours†	$\sim$ 900s	$40\times$	**82.2**	**65.2**
Ours	$\sim$ 30s	$1,200\times$	77.3	58.9

Table 8: Results of inference speed on MSCOCO (5K). ‡ relies on a heavy object detector. † inference is based on the "Re-Rank" method.

A.5 Visualization

We provide more examples of per-word visualizations of our fusion-based encoder finetuned on VL retrieval tasks, as shown in Fig. 6. The visualizations suggest that our model can focus on specific regions of the image according to different words in text when conducts image-text matching.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
The 'limitation' section on page 9.

☒ A2. Did you discuss any potential risks of your work?
This study is a fondamental work based on public datasets, and does not involve privacy, ethics and other dangerous related content.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract and section 1.

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

5.2/5.2/Appendix A.1

☐ B1. Did you cite the creators of artifacts you used?
Not applicable. Left blank.

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Not applicable. Left blank.

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Not applicable. Left blank.

☒ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
The datasets used in this paper are widely used in corresponding fields.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
5.2/5.2/Appendix A.1

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
5.2/5.2/Appendix A.1

C ☑ Did you run computational experiments?

5.3/5.4 Appendix A.3/A.4/A.5

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
5.1.2 / Appendix A.2

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
5.2 / Appendix A.2

☒ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
We report the average results of many experiments

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
5.1.2/ 5.2 / Appendix A.2

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.