Using Self-Supervised Dual Constraint Contrastive Learning for Cross-modal Retrieval

Xintong Wang[a;*], **Xiaoyu Li**[bc], **Liang Ding**[d], **Sanyuan Zhao**[b] **and Chris Biemann**[a]

[a]Language Technology Group, Department of Informatics, University of Hamburg
[b]School of Computer Sciences and Technology, Beijing Institute of Technology
[c]Department of Computer Science, Technical University Berlin
[d]JD Explore Academy

Abstract. In this work, we present a self-supervised dual constraint contrastive method for efficiently fine-tuning the vision-language pre-trained (VLP) models that have achieved great success on various cross-modal tasks, since full fine-tune these pre-trained models is computationally expensive and tend to result in catastrophic forgetting restricted by the size and quality of labeled datasets. Our approach freezes the pre-trained VLP models as the fundamental, generalized, and transferable multimodal representation and incorporates lightweight parameters to learn domain and task-specific features without labeled data. We demonstrated that our self-supervised dual contrastive model performs better than previous fine-tuning methods on MS COCO and Flickr 30K datasets on the cross-modal retrieval task, with an even more pronounced improvement in zero-shot performance. Furthermore, experiments on the MOTIF dataset prove that our self-supervised approach remains effective when trained on a small, out-of-domain dataset without overfitting. As a plug-and-play method, our proposed method is agnostic to the underlying models and can be easily integrated with different VLP models, allowing for the potential incorporation of future advancements in VLP models.

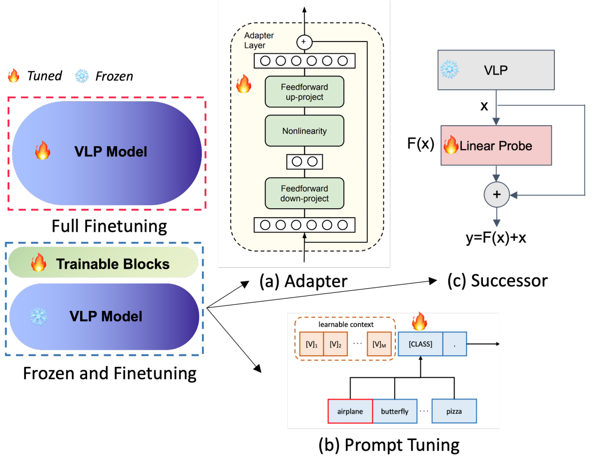

Figure 1. Pre-training and fine-tuning paradigm: full fine-tuning and frozen and fine-tuning.

1 Introduction

With the rapid growth of computational power and extensive large-scale data, increasingly advanced foundation models have been proposed in both the language domain [6, 22, 31] and the vision domain [8, 5]. By leveraging these breakthroughs as the backbone, vision-language pre-trained (VLP) models have made significant strides in a range of cross-modal tasks [19, 35, 2, 32], demonstrating that multimodal representations derived from pre-trained models possess exceptional generalization and transfer capabilities.

In line with the successes of VLP models, recent works [34, 30, 7] have adopted the "pre-training and fine-tuning" paradigm for downstream cross-modal tasks and out-of-domain scenarios. As shown in Figure 1, there are two prevalent fine-tuning strategies. The first, full fine-tuning, involves fine-tuning all parameters, but it carries two notable drawbacks: computational efficiency and catastrophic forgetting [18]. Given the substantial number of parameters in VLP models, considerable memory is required to store these parameters, not to mention train the entire model. For example, the CLIP model [27] utilized 592 V100 GPUs over a span of 18 days. Furthermore, in the absence of high-quality labeled datasets, fully fine-tuning VLP models often results in catastrophic forgetting [18], where the previously learned generalized and transferable multimodal representations from VLP models degrade. The second method, frozen and fine-tuning, offers greater flexibility by freezing VLP model parameters while adding blocks on top to learn out-of-domain and task-specific representations. To achieve state-of-the-art performance on benchmark datasets, these extra blocks tend to be sophisticated and task-specific tricks have been proposed. For instance, in the cross-modal retrieval task, state-of-the-art approaches heavily rely on region feature extraction [10], cross-modal fusion [23], and hard negative sampling [9] during fine-tuning. [28] reveals that while these techniques are crucial for improving performance on benchmark datasets, they come at the cost of increased training time, reduced efficiency, and diminished transferability and utility when applied to different domains.

To address the challenges mentioned earlier, following frozen and fine-tuning, Parameter-Efficient Fine-Tuning (PEFT) [12] has recently gained popularity and attracted significant interest. The core idea of PEFT is to utilize a smaller set of parameters for fine-tuning while retaining the capabilities of pre-trained foundational models

[*] Emails: xintong.wang@uni-hamburg.de, xiaoyu.li@campus.tu-berlin.de, liangding.liam@gmail.com, zhaosanyuan@bit.edu.cn, chris.biemann@uni-hamburg.de. Xiaoyu was a research intern at Language Technology Group.

to improve transferability and adaptability. Among these PEFT approaches, adapters [12] add and update new parameters at the model level, while prompt tuning methods [36] incorporate and train parameters at the input level. Although these techniques have proven effective, they remain inadequate when VLP models are not available for adapter injection, and considerable effort is needed to identify the best prompt templates. In most cases, paired multimodal datasets are not readily accessible. We propose that an optimal solution would involve adding additional parameters at the output level and training the extra layers in a self-supervised manner, without relying on any tailored techniques.

In this paper, we introduce a self-**s**upervised d**u**al **c**onstraint **c**ontrastiv**e** learning for cro**ss**-m**o**dal **r**etrieval task (SUCCESSOR), inheriting the ability of VLP models. In various cross-modal tasks, dual attributes exist [26]. For example, if the primary task in cross-modal retrieval is text retrieval, the dual task would be image retrieval. We construct a dual constraint contrast in the primary modality by back-retrieving negative samples from the dual modality and vice versa, aiming to enhance the alignment of multimodal representations within both intra- and inter-modalities. Specifically, beginning with the primary modality (e.g., vision), we perform forward retrieval (text retrieval) to obtain negative samples from the dual modality (language). We then use these retrieved negative samples to conduct back-retrieval (image retrieval), acquiring candidates in the primary modality. This process allows us to compare the semantic distances between the candidate and original query in the prime modality and vice versa, thereby increasing the alignment and coherence of the multimodal representations.

In terms of our model, we freeze the VLP models to serve as the foundational generalized multimodal representations and add two linear probe layers on top to learn out-of-domain and task-specific representations involving super lightweight parameters for fine-tuning. A skip shortcut is introduced to connect the in-domain representations with the final output of the linear probes, facilitating rapid tuning and model convergence. Our experiments demonstrate that the self-supervised SUCCESSOR model, without relying on region feature extraction or any hard negative sampling techniques, can compete with fine-tuning methods on benchmark datasets such as MS COCO [21] and Flickr 30K [25]. Surprisingly, we discovered that random in-batch negative sampling offers a diverse choice of negative samples, enabling the model to learn fine-grained multimodal semantics, rectify errors from VLP models, and ultimately enhance cross-modal retrieval performance.

Owing to the simplicity of our proposed method, fine-tuning can be completed within hours on an A6000 GPU (48 GB) and can function as a plug-and-play approach, easily integrating with various VLP models without the need for labeled paired data. This adaptability allows for the potential incorporation of future advancements in VLP models. To summarize, our contributions are as follows:

- We introduce a new PEFT approach—a self-supervised dual constraint contrastive method—by adding lightweight, learnable parameters at the output layers. Our method is cost-effective, requiring only a single GPU and a few hours for fine-tuning without the need for labeled datasets, functioning as a plug-and-play solution.
- Our self-supervised method achieves comparable or superior performance to previously fine-tuned state-of-the-art methods on standard benchmark datasets, such as MS COCO and Flickr 30K, without relying on region feature extraction, complex cross-attention fusion, or hard negative sampling strategies.
- By freezing the parameters of VLP models and introducing a skip shortcut, our method yields fast convergence while preserving the generalization and transferability of VLP models. Zero-shot experiments demonstrate that SUCCESSOR further improves cross-modal performance accuracy compared to the VLP backbone, showcasing that SUCCESSOR inherits VLP capabilities.
- A domain adaptation experiment on the education-oriented, small dataset MOTIF [33] reveals that SUCCESSOR performs effectively in domain adaptation without overfitting.

2 Related Work

Vision-language pre-training: We are witnessing an era in which advanced foundational models rapidly evolve in visual and language modalities [8, 5, 6, 22, 3, 31]. In line with the advancements in unimodal foundational models, VLP models have garnered significant research interest. Early models such as ViLBERT [23] employed a dual encoder and cross-attention to learn multimodal representations, while UNITER [4] and OSCAR [20] utilized a fusion encoder with self-attention to learn multimodal alignment. ViLT [16] argued that visual patches from vision transformers are more efficient and enable end-to-end model training. More recently, CLIP [27] adopted large-scale multimodal data from the internet and employed a contrastive method for training, resulting in more powerful multimodal representations and impressive zero-shot performance. Meanwhile, ALBEF [19] demonstrates that image-text contrastive, masked language modeling, and image-text-matching tasks are more efficient than other pre-training tasks. To enhance multimodal generation capabilities, models like BLIP [18], Flamingo [1], and CoCa [35] have been proposed, enabling VLP models to handle both multimodal understanding and generation tasks. Most recently, the VLMo model [2] introduced multiway transformers, unifying the dual encoder and fusion encoder approaches. Building on VLMo, the BEiT-3 model [32] has achieved new state-of-the-art results on cross-modal learning benchmark tasks and even single-modality tasks. We opted for the CLIP model as our VLP model due to its demonstrated efficiency in generalized multimodal feature extraction, moderate parameter size, and the fact that it does not necessitate a pre-trained Fast-RCNN model [10]. Given that our proposed method is a plug-and-play solution, we believe it can be easily applied to other VLP models and even future advancements in the field of VLP.

Parameter-efficient fine-tuning: There are two widely-used fine-tuning approaches: full fine-tuning and frozen fine-tuning. Full fine-tuning presents two drawbacks: computational efficiency and catastrophic forgetting [18]. Given the large number of parameters in VLP models, training the entire model becomes less feasible. Moreover, without high-quality labeled datasets, fully fine-tuning VLP models can lead to catastrophic forgetting [18], where the previously learned generalized and transferable multimodal representations from VLP models deteriorate. In contrast, frozen fine-tuning offers more flexibility and strikes a balance between accuracy and the number of trained task-specific parameters. Recently, parameter-efficient fine-tuning (PEFT) [12] has gained popularity following the frozen fine-tuning fashion. Among PEFT approaches, adapters [12] introduce and update new parameters at the model level, while prompt tuning methods [36, 14] incorporate and train parameters at the input level. Although these techniques have proven effective, they remain inadequate when VLP model training codes are unavailable for adapter injection, and significant effort is required to identify the best prompt templates. We believe that an optimal method involves adding additional parameters at the output level, using VLP models as the fundamental multimodal representation and fine-tuning the extra parameters to learn out-of-domain and task-specific representations.

Cross-modal retrieval: Cross-modal retrieval, such as image-text retrieval [28, 7] requires accurate alignment and understanding of information from different modalities, making it an ideal task to evaluate the performance of our self-supervised dual constraint contrast method. Past research has focused on various ways to improve results on benchmark datasets like MS COCO and Flickr 30K. Although multiple state-of-the-art methods have been proposed to achieve SOTA results on these datasets, they can be categorized into three main directions. First, for instance, region features [10] are crucial for improving accuracy in the visual modality [28], while BERT [6] features outperform RNN features. However, obtaining region features is less efficient and requires pre-training object detection modules [10]. Visual patch projection [8] is more efficient as it allows for end-to-end model training. Second, fusion encoder [4] use self-attention to learn the interaction between modalities, while dual encoders [23] employ cross-attention to interact with different modalities. Lastly, techniques like in-batch hard negative mining [9] have proven effective in increasing the relevance score between paired data while decreasing the score for non-paired data. [28] reveals that region feature extraction and hard negative mining are essential for achieving the results reported in their paper but also raise reproducibility concerns. Our paper avoids using region features for simplicity, as they rely on an extra module, and we found that hard negative mining is less efficient in terms of training time. Random in-batch negative contrast works quite well for our proposed dual constraint contrast. Importantly, all the works mentioned above are trained in a supervised manner. In many real-world scenarios involving out-of-domain and downstream tasks, labeled paired data may not be available. To the best of our knowledge, we are the first to propose a self-supervised fine-tuning method that does not require labeled data and achieves new state-of-the-art results compared to supervised baselines.

3 Method

In this section, we will first discuss the visual and text embeddings used in our model. Next, to better understand the dual idea, we will explain the prime task, dual task, and cross-modal translation. We will then introduce the architecture of our model and also discuss the skip connection. Lastly, we will discuss the self-supervised dual constraint contrast.

3.1 Multimodal embedding and dual task

Visual and text embedding: We opted for a dual encoder [27] to achieve fast retrieval performance, which encodes images and text separately. The choice of architecture is flexible, allowing for the use of other fusion encoders [4] or multi-way transformer architectures [2] if needed. For visual features, we choose grid features extracted from ResNet [11] and patch projections from vision transformers [8] as two different visual backbones. Although using region features has been proven to achieve better results in the cross-modal retrieval task, we do not use them as they require additional pre-trained object detection modules like Fast-RCNN [10] using the Visual Genome dataset [17], which is less efficient. For text features, like most recent works, we utilize BERT embeddings [6]. Formally:

$$\begin{aligned} \{v_n\}_{n=1}^{N} &= Encoder_{\text{visual}}(v) \\ \{t_m\}_{m=1}^{M} &= Encoder_{\text{text}}(t) \end{aligned} \tag{1}$$

where v and t are the input image and text respectively. Suppose the visual encoder can extract N visual vectors, which can be either grid features or patch projections, in d_1 dimensions. Similarly, the text encoder can extract M token vectors in d_2 dimensions.

After extracting the visual and textual features, we input them into the VLP model to obtain fused representations, which are generalized multimodal representations. Following transformations in $VLP_{vision}(\cdot)$ and $VLP_{text}(\cdot)$, the fused vision features and text features are in the same dimension space, represented as $\mathbb{R}^d$ (i.e. in CLIP $d = 768$).

$$\begin{aligned} \mathbf{v} &= VLP_{\text{vision}}\left(\{\mathbf{v}_n\}_{n=1}^{\text{N}}\right) \\ \mathbf{t} &= VLP_{\text{text}}\left(\{\mathbf{t}_m\}_{m=1}^{\text{M}}\right) \end{aligned} \tag{2}$$

Prime and dual task: Text retrieval (image → text) and image retrieval (text → image) are mutually dual tasks. For simplicity and better explanation, we denote the prime modality as the visual modality, and the prime task as text retrieval. In parallel, the dual modality is the text modality, and the dual task is image retrieval. Cross-modal retrieval relies on accurate multimodal alignment and serves as an ideal task to evaluate the performance of multimodal representation learning in terms of **inter-modality** effectiveness.

Prime task - text retrieval: Given a query from the prime modality (vision), we perform text retrieval by measuring the similarity between the query image and candidate texts (dual modality) in the mini-batch as shown in the equation below:

$$\hat{t} = \underset{(t_i,v)\sim\mathcal{B}}{\operatorname{argmax}} \left(\operatorname{sim}\left(t_i, v\right)\right) \tag{3}$$

where v is the visual feature of the query image, t_i is the text feature of the candidate texts in the mini-batch $\mathcal{B}$. $\hat{t}$ is the text that is most similar to the query image in the semantic space. The $sim(\cdot)$ function can be cosine similarity or cross-entropy.

Dual task – image retrieval: Likewise, given a query text in the dual modality, we conduct image retrieval by measuring the similarity between query text and candidate images in the prime modality within the mini-batch as the equation below:

$$\hat{v} = \underset{(v_i,t)\sim\mathcal{B}}{\operatorname{argmax}} \left(\operatorname{sim}\left(v_i, t\right)\right) \tag{4}$$

where t represents the text feature of the query text, while v_i denotes the visual feature of candidate images within the mini-batch $\mathcal{B}$. $\hat{v}$ corresponds to the image that bears the greatest similarity to the query text within the semantic space.

Cross-modal translation: We introduce a cross-modal translation task, to evaluate **intra-modality** alignment performance. Our preliminary experiments revealed that semantically close instances within unimodal domains, such as visual and language, tend to be separated in the multimodal representation space. This separation can lead to errors like counting mistakes and misclassification of fine-grained features (discussed further in section 5.4). We argue that fused visual and text features should remain close to instances that share similar semantics. To address this, we introduce the cross-modal translation task, which involves using a forward retrieval instance as a candidate to perform a back-retrieval task and then comparing whether the original query in the same modality can still be identified.

More specifically, let's assume the forward-retrieval task as text retrieval. Using Eq. (5), we first identify the text instance in the mini-batch that has the maximum similarity score with the query image. Next, we use this retrieved text candidate to perform a back-retrieval

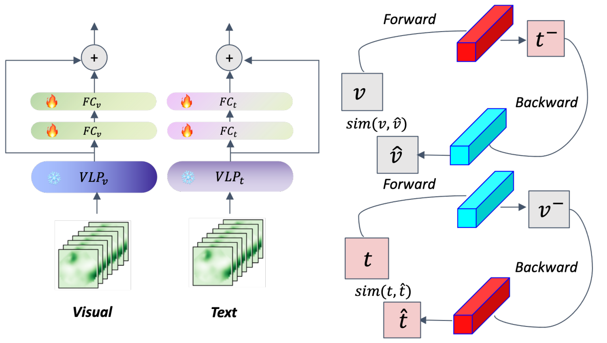

Figure 2. Illustration of (left) our framework (Sec. 3.2) and dual constraint contrast (right) (Sec. 3.3).

task, image retrieval, finding the image in the same mini-batch with the highest similarity score to the candidate text. Finally, we use a $sim(\cdot)$ function to measure the similarity between the back-retrieved image and the original query image.

$$\mathrm{sim}(v, \hat{v}) = sim(v, \underset{(v_i, \hat{t}) \sim \mathcal{B}}{\mathrm{argmax}} \left(\mathrm{sim}\left(v_i, \hat{t}\right)\right)) \tag{5}$$

where v represents the visual feature of the query image, while $\hat{v}$ denotes the image obtained from the back-retrieval task. $\hat{t}$ refers to the forward retrieved text candidate, as shown in Eq. (3), and v_i represents the image feature in the same mini-batch $\mathcal{B}$.

Similarly, we can initiate the process with the language modality, where the forward-retrieval task is image retrieval and the back-retrieval task is text retrieval, as shown in Eq. (6).

$$\mathrm{sim}(t, \hat{t}) = sim(t, \underset{(t_i, \hat{v}) \sim \mathcal{B}}{\mathrm{argmax}} \left(\mathrm{sim}\left(t_i, \hat{v}\right)\right)) \tag{6}$$

where t denotes the text feature of the query text, while $\hat{t}$ represents the text obtained from the back-retrieval task. $\hat{v}$ refers to the forward retrieved image candidate, as shown in Eq. (4), and t_i signifies the text feature in the same mini-batch $\mathcal{B}$.

It is important to note that, since we utilize multimodal representations from VLP models as the backbone, the forward-retrieved instances are likely to be similar to the original query. This likelihood is due to the consideration of relevant pairs in the data construction using VLP. Given that VLP multimodal representations are generalized and transferable, and the candidate from the forward retrieval serves as a bridge, we hypothesize that we can leverage this dual process to form a dual constraint contrast loss. This approach would allow the model to be trainable without labeled paired dataset by only adding extra parameters at the output level to learn out-of-domain and task-specific representations. To realize this hypothesis, we introduce a skip connection and self-supervised dual constraint contrast, which will be discussed in the next section.

3.2 *Framework and skip connection*

Given the remarkable generalization and transfer capabilities of VLP models, we use VLP as the backbone and freeze the VLP parameters for parameter-efficient fine-tuning to obtain the fundamental in-domain multimodal representations. For simplicity, we add two linear probe layers at the output level of the VLP backbone to learn out-of-domain and task-specific multimodal representations, as illustrated in Figure 2. For visual and text modalities, following Eq. (1) and (2), the fused visual and text representations can be expressed as follows:

$$\begin{aligned} v &= FC_v \left(VLP_{\text{vision}} \left(\{v_n\}_{n=1}^N \right) \right) \\ t &= FC_t \left(VLP_{\text{text}} \left(\{t_m\}_{m=1}^M \right) \right) \end{aligned} \tag{7}$$

where $VLP_{\text{vision}} \left(\{v_n\}_{n=1}^N \right)$ denotes the fused visual feature from VLP models, and $VLP_{\text{text}} \left(\{t_m\}_{m=1}^M \right)$ denotes the fused text feature from VLP models. FC represents linear probe layers. v and t indicate the fused visual and text features after the linear probe layers.

However, since our method is based on self-supervised dual contrast, the model needs to have a basic ability to retrieve candidates as shown in Eq. (5) and Eq. (6); otherwise, the model will collapse. Therefore, we introduce a skip connection [11], linking the representation from VLP to the final output prediction of our model as Eq. (8). In practice, the skip connection is crucial for robust fine-tuning and fast convergence. This is because, in the beginning, the VLP model will contribute more to retrieving the candidate and allow the method to compare the query and candidate in the same modality.

$$\begin{aligned} v &= \alpha_1 \cdot VLP_{\text{vision}} \left(\{v_n\}_{n=1}^N \right) + \alpha_2 \cdot FC_v \left(VLP_{\text{vision}} \left(\{v_n\}_{n=1}^N \right) \right) \\ t &= \gamma_1 \cdot VLP_{\text{text}} \left(\{t_m\}_{m=1}^M \right) + \gamma_1 \cdot FC_t \left(VLP_{\text{text}} \left(\{t_m\}_{m=1}^M \right) \right) \end{aligned} \tag{8}$$

where α_1, α_2, γ_1, and γ_2 are hyperparameters to balance the importance of the fused multimodal representations from VLP and the representations after linear probe layers. In experiments, we find that these hyperparameters are not sensitive and are set to 1.0.

Note that our model adds extra parameters at the output level. This approach is more flexible compared to Adapter and Prompt tuning methods. Adapter methods [12] inject extra parameters at the model level, which require the training code of VLP models, while prompt tuning methods [36] incorporate and train parameters at the input level, necessitating considerable effort to find the best template. Our model consists of two linear probe layers and a skip connection shortcut. We do not use any complex fusion layers or cross-attention-based methods. As seen in Table 2, our methods surpass fine-tuning-based methods and outperform fine-tuned VLP backbones by a large margin. In the following section, we will introduce how to leverage our model to form a dual constraint contrast loss.

3.3 *Self-supervised dual constraint contrast*

As discussed in section 3.1, our dual constraint contrast is formed by the forward retrieval and back retrieval as a loop. More specifically, we first obtain fused image and text features for each mini-batch using Eq. (8). Assuming the forward-retrieval task is text retrieval, we take each image instance v as the query to find the most similar negative text sample t^- in the batch by computing the similarity scores as per Eq. (3). In our experiments, we employ the cosine similarity function for measuring the similarity. We refer to the retrieved text as the negative sample since we do not know if it is the anchor in a self-supervised method. We use the forward-retrieved text t^- to conduct the back-retrieval task-image retrieval. Similarly, we compute the similarity scores between the text t^- and all the images in the mini-batch to obtain a similarity vector. We then normalize the similarity vector using the Softmax function. Finally, we form the loss as the cross-entropy loss using the normalized similarity vector with the pseudo-label vector where the original query image is one, and

Table 1. Comparison of our proposed method with five state-of-the-art VLP methods and one plug-and-play method on the image-text retrieval task. For grid features $*$, PixelBERT used ResNet-50 features and CLIP as well as our models used two variants, ResNet-50 and ViT-L patches.

Method	Params	Architecture	Fine-tuning	Visual Tokens	Pre-trained Datasets	BS	Self-supervised	Loss
ViLBERT	221M	fusion encoder	full fine-tune	Region	CC	64	✗	cross-entropy
PixelBERT	124M	fusion encoder	full fine-tune	Grid*	VG, MSCOCO	512	✗	cross-entropy
UNITER	110M	fusion encoder	full fine-tune	Region	CC, SBU, VG, MSCOCO	64	✗	cross-entropy
ViLT	111M	fusion encoder	full fine-tune	Region	CC, SBU, VG, MSCOCO	256	✗	cross-entropy
CLIP	2.3M	dual encoder	frozen fine-tune	Grid*	WIT	128	✗	contrastive loss
BCAR	2.2M	fusion encoder	frozen fine-tune	Region	VG	128	✗	ranking loss
Successor	2.3M	dual encoder	frozen fine-tune	Grid*	✗	128	✓	cross-entropy

the other images are zero. Likewise, we can begin with the forward retrieval task as the image retrieval task and the back retrieval task as the text retrieval task. We train our model using cross-entropy loss:

$$\mathcal{L}\left(\theta^V,\theta^L\right) = -\frac{1}{|\mathcal{B}|}\left(\sum_{i=1}^{M} y_i^V \log\left(f^{V\to L\to V}_{\left(\theta^V,\theta^L\right)}\left(v,t^-,\hat{v}\right)_i\right) + \sum_{i=1}^{M} y_i^L \log\left(f^{L\to V\to L}_{\left(\theta^V,\theta^L\right)}\left(t,v^-,\hat{t}\right)_i\right)\right) \tag{9}$$

where (θ^V,θ^L) are the trainable parameters in the two layers of the linear probe, $|\mathcal{B}|$ is the batch size, and M is the number of instances in the batch. y^V and y^L represent the pseudo-labels where the query image or text is one and other images and text are set to zero. $f^{V\to L\to V}$ represents the loop from text retrieval to image retrieval, and $f^{L\to V\to L}$ is the reverse dual loop, going from image retrieval to text retrieval.

4 Experiment Setup

Datasets. We evaluate our method on two widely used benchmark datasets for cross-modal retrieval, MS COCO and Flickr 30K, and one distinct dataset called MOTIF, with more complex text. In more detail, MS COCO contains 123,287 images, each with five sentences describing the image's content. Flickr 30K has 31,783 images; like MS COCO, it is also paired with five corresponding sentences. Following the typical approach to split datasets in most of the literature, we use the Karpathy split [15] method for MS COCO and Flickr 30K datasets. We use MOTIF, a language-oriented multimodal dataset, to test the domain transfer effect. MOTIF has 1,125 sentences with at least three complex words, and the structure is more complex than the sentences in MS COCO and Flickr 30K. We randomly split the dataset into training and test datasets as 900/225 images. In the implementation within a self-supervised setting, we employ pre-trained VLP models (without exposing any datasets) to conduct cross-modal retrieval. The goal is to find relevant pairs, which may or may not be correct. During training, shuffling is also performed to increase the diversity of negative samples within the mini-batch.

Evaluation metrics. We evaluate the cross-modal retrieval performance and cross-modal translation performance using $recall@K$ as the evaluation metric. In our experiments, we report R@1, R@5 and R@10. To provide a clearer description of the tasks in our experiments, we use the following abbreviations: "IR" for Image Retrieval, "TR" for Text Retrieval, "ITI" for Image-Text-Image Translation, and "TIT" for Text-Image-Text Translation.

Implementation details. We implement our model using the PyTorch framework and utilize the CLIP model as the backbone. The dimension of the fused visual and text features from CLIP is 768. Our model is trained on an Nvidia RTX A6000 GPU, but we only utilize 14GB of its 48GB memory. For optimization, we employ the Adam optimizer with a learning rate of 1e-5 and a weight decay of 1e-5. The balance hyperparameters, α_1, α_2, γ_1, and γ_2, show minimal sensitivity in our experiments and are all set to 1.0.

We use two CLIP architectures as backbones in our experiments. For the text encoder, we employ BERT as the backbone. In the visual encoder, we implement two versions: one using ResNet-50 [11] and the other using ViT-L/14@336px [8]. To be clear, we refer to the dual contrastive model using ResNet-50 as Successor@RN50 and the one using ViT-L/14@336px as Successor@ViT-L. Both linear probe layers have a dimension of 768 and employ the non-linear activation function, ReLU. We train the Successor@RN50 model for 25 epochs and the Successor@ViT-L model for 20 epochs.

5 Results and Analysis

To demonstrate the effectiveness of our proposed method presented in Section 3, we compare it with six baselines. First, we compare our method with five pre-trained state-of-the-art methods: ViLBERT [23], PixelBERT [13], UNITER [4], ViLT [16], and CLIP [27]. Due to reproducibility issues, we cite the results from [28] and report and compare them in our paper. Since CLIP did not use MS COCO and Flickr 30K for pre-training, we adopted the standard linear probing method to fine-tune CLIP. Next, we compare our method with the most recent work on plug-and-play method, BCAR [7], in the cross-modal retrieval task. Detailed settings and comparisons can be found in Table 1.

It is worth noting that all the baselines are fine-tuned in a supervised manner, and ViLBERT, UNITER, ViLT, and BCAR use region features, which have been proven to improve results but increasing training time as the involvement extra detector modules. These baselines also employed extra datasets like CC [29], VG [17], and SBU [24] datasets, as well as other techniques to enhance performance. However, our proposed self-supervised dual contrastive method aims to create a more flexible and adaptable approach for cross-modal retrieval tasks that do not rely on specific tricks, region features, or a pre-trained object detection module like Fast-RCNN.

By comparing our method with the supervised baselines, we aim to demonstrate the effectiveness of our approach in scenarios where labeled paired datasets are unavailable, and out-of-domain cases. In doing so, we highlight the advantages of our method, which involves lightweight trainable parameters, making it a more practical choice for a wider range of real-world applications.

Table 2. Cross-modal retrieval results on MS COCO and Flickr 30K datasets. The top half of the table displays the performance of fine-tuned VLP models using supervised methods, while the bottom half showcases the results of our proposed approaches, including the ResNet-50 variant and the ViT variant. The best results are highlighted in blue and red.

Model	**Flickr 30K 1K Test**						**MS COCO 5K Test**					
	IR@1	IR@5	IR@10	TR@1	TR@5	TR@10	IR@1	IR@5	IR@10	TR@1	TR@5	TR@10
Supervised VLP Performance												
ViLBERT	58.2	84.9	91.5	76.8	93.7	97.6	38.6	68.2	79.0	53.5	79.7	87.9
PixelBERT	59.8	85.5	91.6	75.7	94.7	97.1	41.1	69.7	80.5	53.4	80.4	88.5
UNITER	62.9	87.2	92.7	78.3	93.3	96.5	37.8	67.3	78.0	52.8	79.7	87.8
ViLT	62.2	87.6	93.2	83.7	97.2	98.1	42.6	72.8	**83.4**	62.9	**87.1**	**92.7**
CLIP@RN50	**68.5**	**91.6**	**95.6**	**84.7**	**97.3**	**99.1**	43.1	70.8	80.9	59.7	83.8	90.6
CLIP@ViT-L	**73.7**	**93.2**	**96.3**	**88.3**	**98.7**	**99.5**	**46.5**	**73.4**	82.7	**63.6**	86.2	92.5
BCAR	62.6	85.8	91.1	82.3	96.0	98.4	44.3	73.2	83.2	61.3	86.1	92.6
Dual Contrast Performance (Ours)												
Successor@RN50	**71.3** ↑	**92.2**↑	**96.0** ↑	**87.6** ↑	**98.5** ↑	**99.3** ↑	43.8	71.4	81.1	60.5	85.1	91.3
Successor@ViT-L	**74.9** ↑	**94.1** ↑	**96.8** ↑	**89.1** ↑	**98.7** ↑	**99.5** ↑	**46.8** ↑	**74.1** ↑	**83.2** -	**64.7** ↑	**86.5** -	**92.7** ↑

5.1 Comparison to state-of-the-art methods

Table 2 presents a comprehensive comparison with state-of-the-art VLP models and one plug-and-play method on Flickr 30K and MS COCO datasets as mentioned above. The top half of the table displays the performance of fine-tuned VLP models trained in a supervised manner. The bottom half showcases the results of our proposed approaches, including the ResNet-50 variant and the ViT variant. The best results are highlighted in blue for the top-performing results among the baseline methods, while red represents the best results achieved by our methods, surpassing the best baseline results.

Our self-supervised ViT variant outperforms all the baseline results on both Flickr 30K and MS COCO datasets. Additionally, both the ResNet-50 variant and ViT variant surpass all baselines on the Flickr 30K dataset. This demonstrates the effectiveness of our dual constraint contrast methods. In particular, Successor@RN50 achieves a 1.5% (536.8 → 544.9) relative gain, and Successor@ViT-L achieves a 0.62% (549.7 → 553.1) relative gain on the Flickr 30K dataset compared with the best baselines, CLIP@RN50 and CLIP@ViT-L, in supervised VLP performance. Successor@ViT-L also obtains a 1.5% (441.5 → 448) relative gain on MS COCO compared with the best results of the ViLT model. These results highlight the effectiveness of our dual constraint contrast methods and their stability, as our model is simple and omits any tricks for simplicity.

Regarding parameter-efficient fine-tuning performance, we observe that the frozen and fine-tuning paradigm works better than fully fine-tuning the model. From the perspective of trainable parameters, CLIP, BCAR, and our Successor model have 98% fewer parameters than PixelBERT, UNITER, and ViLT, and 99% fewer parameters than ViLBERT. Nevertheless, CLIP and BCAR achieve the best results among baseline methods on the Flickr 30K dataset and the image retrieval task on the MS COCO dataset, which demonstrates that Parameter-Efficient Fine-Tuning (PEFT) methods are more efficient for fine-tuning while maintaining high accuracy. ViLT attains the best results in MS COCO as the dataset is much larger, and learning from supervised labels helps improve accuracy. Importantly, our methods outperform all the PEFT baselines and achieve better or comparable results to all the baselines and even the ViLT model on MS COCO.

Lastly, compared with fine-tuned CLIP, our Successor shares the same architecture but achieves similar or even better results on both datasets in a self-supervised manner. This supports our hypothesis that the VLP model possesses exceptional generalization and capabilities. We can fine-tune the model by adding extra layers at the output level to inherit the abilities of VLP and learn out-of-domain and task-specific representations without labeled data.

The improved performance of our method demonstrates that the learned out-of-domain and task-specific multimodal representations possess strong inter-modality effectiveness. As discussed in Section 3.1, cross-modal translation tasks can evaluate the intra-modality alignment of multimodal representations. In Table 3, we conduct cross-modal translation and compare Successor@RN50 and Successor@ViT-L with the baseline CLIP@RN50 and CLIP@ViT-L on Flickr 30K and MS COCO datasets. Both variants consistently achieve better results than the baseline, illustrating the effectiveness of our method. As we opted for CLIP as the VLP backbone, the improved results indicate that closely related semantic instances within a unimodal representation maintain their proximity in the multimodal representation space compared with VLP models. This highlights the successful intra-modality alignment achieved by our dual constraint contrast methods.

5.2 Zero-shot performance

The multimodal representations obtained from VLP models have demonstrated remarkable generalization capabilities. Our proposed method involves freezing the VLP as the foundation and adding linear probe layers at the output level. Thus, training the model with extra data should act as the incremental of the foundation knowledge of VLP. We hypothesize that this approach should yield better zero-shot performance compared to the original VLP. To test this hypothesis, we first train our proposed model described in Section 3 using the dual constraint contrast method on the Flickr 30K dataset and evaluate the model on the MS COCO 5K test set. Similarly, we train our proposed model on the MS COCO dataset and test it on the Flickr 30K dataset. As demonstrated in Table 4 and Table 5, our fine-tuned model with the dual contrast method achieves better or comparable zero-shot performance compared to the fine-tuned CLIP model with the same backbone on both datasets. We attribute these improvements to the frozen and fine-tuning paradigm. In comparison to the full fine-tuning approach, full fine-tuned models run the risk

Table 3. Cross-modal translation results on MS COCO and Flickr 30K datasets. The **bold** number represents the best results achieved by models.

Model	Flickr 30K 1K Test						MS COCO 5K Test					
	ITI@1	ITI@5	ITI@10	TIT@1	TIT@5	TIT@10	ITI@1	ITI@5	ITI@10	TIT@1	TIT@5	TIT@10
CLIP@RN50	87.9	99.7	100.0	65.0	82.9	93.6	71.2	98.0	99.6	40.0	67.3	82.2
CLIP@ViT-L	91.0	99.9	100.0	70.9	**86.4**	94.8	73.1	97.8	99.7	43.2	68.1	81.9
Successor@RN50	91.2	**100.0**	100.0	68.5	84.4	92.7	72.6	97.8	99.7	40.3	67.6	82.2
Successor@ViT-L	**92.0**	99.8	**100.0**	**71.8**	86.2	**95.1**	**74.3**	**98.3**	**99.8**	**43.8**	**68.8**	**83.1**

Table 4. Zero-shot performance results on the Flickr 30K dataset. The **bold** number represents the best results achieved by models.

Model	Flickr30K-IR			Flickr30K-TR		
	R@1	R@5	R@10	R@1	R@5	R@10
CLIP@RN50	61.5	84.7	90.0	81.8	95.9	98.1
CLIP@ViT-L	64.0	86.6	91.6	84.8	**97.9**	**99.1**
Successor@RN50	66.9	89.2	93.2	82.7	97.0	98.6
Successor@ViT-L	**70.6**	**91.7**	**95.1**	**86.5**	97.4	98.9

Table 5. Zero-shot performance results on the MS COCO dataset. The **bold** number represents the best results achieved by models.

Model	MSCOCO5K-IR			MSCOCO5K-TR		
	R@1	R@5	R@10	R@1	R@5	R@10
CLIP@RN50	35.1	59.7	69.9	54.8	78.8	86.4
CLIP@ViT-L	36.8	61.4	71.3	57.5	81.1	87.7
Successor@RN50	38.2	64.0	74.0	55.6	78.9	86.5
Successor@ViT-L	**42.7**	**68.1**	**77.6**	**60.2**	**82.0**	**89.4**

of causing catastrophic forgetting [18], where the previously learned generalized and transferable multimodal representations from VLP models degrade. By using the frozen and fine-tuned paradigm, we can avoid this issue and maintain the quality of multimodal representations, leading to improved zero-shot performance.

5.3 Domain adaptation performance

To better investigate the domain adaptation performance of our proposed model, we train and compare our method with the baseline CLIP model on the MOTIF dataset. The MOTIF dataset is an education-oriented multimodal dataset, where the sentence structure and vocabulary are more complex than those in the Flickr 30K or MS COCO datasets. Table 6 demonstrates that our proposed method performs well in acquiring out-of-domain multimodal representations compared to the supervised method on CLIP. In addition to its self-supervised attributes, our proposed model can effectively train and transfer knowledge on a small dataset without overfitting. This characteristic makes it particularly useful for domain adaptation tasks, where it is essential to leverage and adapt existing knowledge to new, complex domains with limited labeled data available.

Table 6. Domain adaptation performance results on the MOTIF dataset. The **bold** number represents the best results achieved by models.

Model	MOTIF-IR			MOTIF-TR		
	R@1	R@5	R@10	R@1	R@5	R@10
CLIP@RN50	48.0	96.8	**100.0**	48.0	81.6	90.4
Successor@RN50	**50.4**	**97.6**	99.2	**48.0**	**82.4**	**92.8**

5.4 Error analysis and ablation study

In addition to quantitative analysis, we also examine the quality of retrieved results through a qualitative investigation. We observe that VLP models sometimes exhibit fine-grained errors, primarily related to counting mistakes, nuanced color and pattern understanding, and complex noun and verb comprehension. For instance, the term "runners" is the plural form of "runner"; however, the CLIP model overlooks this vital information while retrieving images. Our proposed model is capable of capturing fine-grained multimodal representations. Meanwhile, we conduct ablation studies using two variants on the Flickr 30K dataset. For each model, we remove either the dual constraint loss from $V \rightarrow L \rightarrow V$ or $L \rightarrow V \rightarrow L$. The results in Table 7 show that when the loss from $V \rightarrow L \rightarrow V$ is removed, the image retrieval performance degrades, while removing the loss from $L \rightarrow V \rightarrow L$ causes the text retrieval performance to degrade.

Table 7. Ablation study results on Flickr 30K dataset. The underlined number represents a degradation in performance.

Model	IR@1	IR@5	IR@10	TR@1	TR@5	TR@10
Successor@RN50	**71.3**	**92.2**	**96.0**	**87.6**	**98.5**	**99.3**
- $V \rightarrow L \rightarrow V$	63.5	87.3	92.6	85.5	97.6	99.1
- $L \rightarrow V \rightarrow L$	70.1	91.3	95.5	79.5	94.7	97.9
Successor@ViT-L	**74.9**	**94.1**	**96.8**	**89.1**	**98.7**	**99.5**
- $V \rightarrow L \rightarrow V$	65.0	87.4	93.1	88.0	98.2	99.5
- $L \rightarrow V \rightarrow L$	74.1	93.4	96.4	84.6	97.0	99.1

6 Conclusion

In this work, we present a self-supervised dual constraint contrast method designed to efficiently fine-tune VLP models using a "frozen and fine-tuning" paradigm. By incorporating additional linear probe layers at the output level and incorporating a skip shortcut, we achieve fast convergence. As our approach only updates lightweight parameters (2.3M), the training cost is significantly lower compared to other full fine-tuning methods. Consequently, we can train our model using a single GPU, achieving convergence within hours while maintaining comparable or superior performance to existing fine-tuned VLP and PEFT methods on two benchmark datasets. Furthermore, our method demonstrates strong domain transfer capabilities. With its simplicity and feasibility, our approach is agnostic to the underling models and has the potential to harness the power of more advanced VLP models in the future.

Acknowledgements

This research was funded by the German Research Foundation DFG Transregio SFB 169: Crossmodal Learning: Adaptivity, Prediction, and Interaction.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al., 'Flamingo: a visual language model for few-shot learning', *Advances in Neural Information Processing Systems*, **35**, 23716–23736, (2022).

[2] Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, Songhao Piao, and Furu Wei, 'Vlmo: Unified vision-language pre-training with mixture-of-modality-experts', *Advances in Neural Information Processing Systems*, **35**, 32897–32912, (2022).

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, , and Amanda Askell, 'Language models are few-shot learners', *Advances in neural information processing systems*, 1877–1901, (2020).

[4] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu, 'Uniter: Universal image-text representation learning', in *16th European Conference Computer Vision*, pp. 104–120, (2020).

[5] Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, et al., 'Scaling vision transformers to 22 billion parameters', *arXiv preprint arXiv:2302.05442*, (2023).

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, 'BERT: Pre-training of deep bidirectional transformers for language understanding', in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 4171–4186, (2019).

[7] Haiwen Diao, Ying Zhang, Wei Liu, Xiang Ruan, and Huchuan Lu, 'Plug-and-play regulators for image-text matching', *IEEE Transactions on Image Processing*, (2023).

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al., 'An image is worth 16x16 words: Transformers for image recognition at scale', *arXiv preprint arXiv:2010.11929*, (2020).

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

[10] Ross Girshick, 'Fast r-cnn', in *Proceedings of the IEEE international conference on computer vision*, pp. 1440–1448, (2015).

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, 'Deep residual learning for image recognition', in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, (2016).

[12] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly, 'Parameter-efficient transfer learning for nlp', in *International Conference on Machine Learning*, pp. 2790–2799, (2019).

[13] Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu, 'Pixel-bert: Aligning image pixels with text by deep multi-modal transformers', *arXiv preprint arXiv:2004.00849*, (2020).

[14] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim, 'Visual prompt tuning', in *17th European Conference Computer Vision*, pp. 709–727, (2022).

[15] Andrej Karpathy and Li Fei-Fei, 'Deep visual-semantic alignments for generating image descriptions', in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3128–3137, (2015).

[16] Wonjae Kim, Bokyung Son, and Ildoo Kim, 'Vilt: Vision-and-language transformer without convolution or region supervision', in *International Conference on Machine Learning*, pp. 5583–5594, (2021).

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

[18] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi, 'Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation', in *International Conference on Machine Learning*, pp. 12888–12900, (2022).

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

[20] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al., 'Oscar: Object-semantics aligned pre-training for vision-language tasks', in *16th European Conference Computer Vision*, pp. 121–137, (2020).

[21] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick, 'Microsoft coco: Common objects in context', in *13th European Conference Computer Vision*, pp. 740–755. Springer, (2014).

[22] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov, 'Roberta: A robustly optimized bert pretraining approach', *arXiv preprint arXiv:1907.11692*, (2019).

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

[24] Vicente Ordonez, Girish Kulkarni, and Tamara Berg, 'Im2text: Describing images using 1 million captioned photographs', *Advances in neural information processing systems*, **24**, (2011).

[25] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik, 'Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models', in *Proceedings of the IEEE international conference on computer vision*, pp. 2641–2649, (2015).

[26] Tao Qin, *Dual Learning*, Springer, 2020.

[27] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al., 'Learning transferable visual models from natural language supervision', in *International conference on machine learning*, pp. 8748–8763, (2021).

[28] Jun Rao, Fei Wang, Liang Ding, Shuhan Qi, Yibing Zhan, Weifeng Liu, and Dacheng Tao, 'Where does the performance improvement come from? -a reproducibility concern about image-text retrieval', in *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2727–2737, (2022).

[29] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut, '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*, pp. 2556–2565, (2018).

[30] Yi-Lin Sung, Jaemin Cho, and Mohit Bansal, 'Vl-adapter: Parameter-efficient transfer learning for vision-and-language tasks', in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5227–5237, (2022).

[31] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al., 'Llama: Open and efficient foundation language models', *arXiv:2302.13971*, (2023).

[32] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al., 'Image as a foreign language: Beit pretraining for all vision and vision-language tasks', *arXiv preprint arXiv:2208.10442*, (2022).

[33] Xintong Wang, Florian Schneider, Özge Alacam, Prateek Chaudhury, and Chris Biemann, 'MOTIF: Contextualized images for complex words to improve human reading', in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 2468–2477, (2022).

[34] Taojiannan Yang, Yi Zhu, Yusheng Xie, Aston Zhang, Chen Chen, and Mu Li, 'Aim: Adapting image models for efficient video action recognition', *arXiv preprint arXiv:2302.03024*, (2023).

[35] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu, 'Coca: Contrastive captioners are image-text foundation models', *arXiv preprint arXiv:2205.01917*, (2022).

[36] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu, 'Learning to prompt for vision-language models', *International Journal of Computer Vision*, **130**(9), 2337–2348, (2022).