mPLUG: Effective and Efficient Vision-Language Learning by Cross-modal Skip-connections

Chenliang Li[*], **Haiyang Xu**[*], **Junfeng Tian, Wei Wang, Ming Yan**[†]**, Bin Bi**[†]**, Jiabo Ye, Hehong Chen, Guohai Xu, Zheng Cao, Ji Zhang, Songfang Huang, Fei Huang, Jingren Zhou, Luo Si**

DAMO Academy, Alibaba Group

{lcl193798, shuofeng.xhy, tjf141457, hebian.ww, ym119608, b.bi, yejiabo.yjb, hehong.chh, guohai.xgh, zhengzhi.cz, zj122146, songfang.hsf, f.huang, jingren.zhou, luo.si}@alibaba-inc.com

Abstract

Large-scale pretrained foundation models have been an emerging paradigm for building artificial intelligence (AI) systems, which can be quickly adapted to a wide range of downstream tasks. This paper presents mPLUG, a new vision-language foundation model for both cross-modal understanding and generation. Most existing pre-trained models suffer from the problems of low computational efficiency and information asymmetry brought by the long visual sequence in cross-modal alignment. To address these problems, mPLUG introduces an effective and efficient vision-language architecture with novel cross-modal skip-connections, which creates inter-layer shortcuts that skip a certain number of layers for time-consuming full self-attention on the vision side.

mPLUG is pre-trained end-to-end on large-scale image-text pairs with both discriminative and generative objectives. It achieves state-of-the-art results on a wide range of vision-language downstream tasks, such as image captioning, image-text retrieval, visual grounding and visual question answering. mPLUG also demonstrates strong zero-shot transferability when directly transferred to multiple video-language tasks.

1 Introduction

Large-scale pre-training of vision-language models have recently received tremendous success on a wide range of cross-modal tasks [1, 2, 3, 4, 5, 6, 7]. Such vision-language models learn cross-modal representations from a quantity of image-text pairs by aligning the visual and linguistic modalities. A great challenge of learning vision-language models is to find a good alignment between the two modalities to close the semantic gap in-between.

To discover a cross-modal alignment, prior studies [4, 8, 9] employ a pre-trained object detector
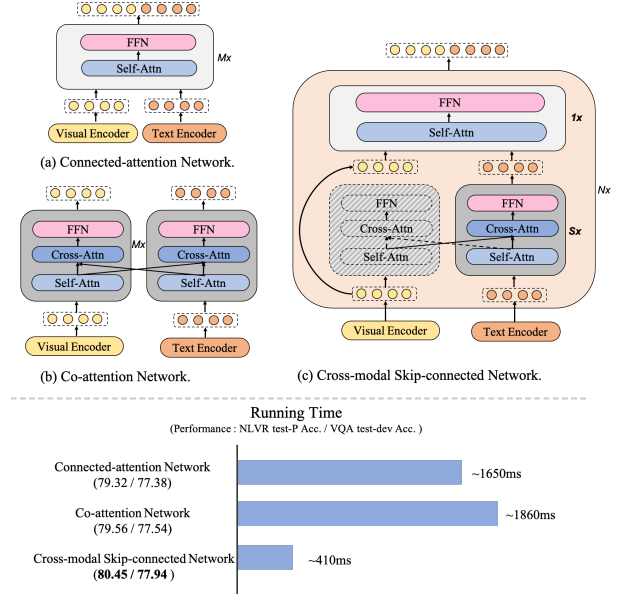

Figure 1: Illustration of two conventional cross-modal fusion networks and our proposed cross-modal skip-connected network. We compare the running time and performance of different fusion networks, where the total fusion layers, image encoder and text encoder are all kept the same. The running time is the total forward time of 100 samples in different fusion networks.

to extract salient regions from images, which are then aligned with language counterparts. Such an architecture, however, is generally limited by the power of the object detector, the pre-defined visual semantics it can represent, and the quantity of annotations available. Besides, it is also computationally expensive to extract region-based visual features from high-resolution (e.g. 600×1000) images. More recent work [3, 7, 6, 10, 11], which scales and performs better on many vision-language tasks, drops the requirement of pre-trained object detection and enables a direct alignment between the image and text representations in an end-to-end manner. These models extract finer-grained visual representation with a long sequence of image patches or grids for good vision understanding [11].

---

[*] Equal contribution
[†] Corresponding authors

However, there exist two significant problems in modeling long visual sequences: 1) *efficiency*: full self-attention on long visual sequences requires much more computation than that on textual sequences, and 2) *information asymmetry*: the caption text in widely-used image-text pre-training data is usually short and highly abstract while more detailed and diverse information can be extracted from the image. This asymmetry presents challenges for effective multi-modal fusion between the modalities.

One straightforward way of multi-modal fusion is the connected-attention network as shown in Figure 1 (a). It adopts a single Transformer [12] network for early fusion of vision and language by simply taking the concatenation of visual and linguistic features as input [13]. This paradigm allows self-attention to discover alignments between the modalities from the bottom level, and requires full self-attention on the concatenation of cross-modal sequences, which is rather time-consuming. Besides, this type of methods process information from both modalities equally, which may suffer from the information asymmetry especially when there is a big difference in information density or sequence lengths between the modalities.

Another line of work keeps separate Transformer networks for both textual and visual features, and uses techniques such as cross-attention to enable cross-modal interaction [11], as shown in Figure 1 (b). This architecture design conducts multi-modal fusion on both modalities independently, which can help alleviate the information asymmetry problem. However, it still suffers from computation inefficiency for full self-attention on long visual sequences, and it is not that parameter-efficient with two separate Transformer networks.

In this work, we propose mPLUG, a unified Multi-modal Pre-training framework for both vision-Language Understanding and Generation. mPLUG performs effective and efficient vision-language learning with novel cross-modal skip-connections to address the fundamental information asymmetry problem. Instead of fusing visual and linguistic representations at the same levels, the cross-modal skip-connections enables the fusion to occur at disparate levels in the abstraction hierarchy across the modalities. It creates inter-layer shortcuts that skip a certain number of layers for visual representations to reflect the semantic richness of language compared to vision. As shown

in Figure 1 (c), in each block of our cross-modal skip-connected network, mPLUG first adopts an asymmetric co-attention architecture at the first few layers for efficiency, by removing the co-attention on vision side. It is then followed by one layer of connected-attention, by concatenating the original visual representation and the co-attention output on the language side as input. In addition to the modeling efficacy due to the asymmetry, the cross-modal skip-connections ease the model training by alleviating vanishing gradients with the inserted shortcuts. Figure 1 shows that the new cross-modal skip-connected network achieves superior performance with at least four times speeding-up than other cross-modal fusion networks.

Our key contributions can be summarized as follows:

- We propose a unified vision-language pre-trained model mPLUG of cross-modal understanding and generation for both effectiveness and efficiency in cross-modal learning.

- We introduce a new asymmetric vision-language architecture with novel cross-modal skip-connections, to address two fundamental problems of information asymmetry and computation inefficiency in multi-modal fusion.

- mPLUG achieves state-of-the-art performance on a wide range of vision-language tasks, including image captioning, image-text retrieval, visual grounding and visual question answering. mPLUG also demonstrates strong zero-shot transferability when directly transferred to a wide range of vision-language and video-language tasks.

2 Related Work

2.1 Vision-Language Pre-training

Vision-Language pre-training (VLP) has recently received tremendous success and achieved state-of-the-art results across a variety of vision-language tasks [14, 15, 16]. In terms of how information from different modalities are aggregated, typical approaches to VLP [1, 2, 3, 5, 6, 17, 18] can be roughly divided into two categories: *dual encoder* and *fusion encoder*. Dual encoder approach utilizes two single-modal encoders to encode images and text separately, and then uses simple functions such as dot product to model the instance-level cross-modal interaction between image and text. The

advantage of dual encoder models like CLIP [17] and ALIGN [18] is that images and text can be pre-computed and cached, which is quite computation-efficient and more appropriate for retrieval tasks. However, they tend to fail in handling more complicated VL understanding tasks that require complex reasoning, such as visual question answering [14]. In contrast, fusion encoder approach uses deep fusion functions such as multi-layer self-attention and cross-attention networks to model the fine-grained cross-modal interaction between image and text sequences. Representative methods of this category include the single-stream architecture such as UNITER [2] and OSCAR [4], and two-stream architecture such as LXMERT [1], ALBEF [6] and ERNIE-ViL [5]. This kind of methods can better capture the underlying association between image and text for vision-language understanding tasks, while it needs to jointly encode all possible image-text pairs, which leads to a relatively slow inference speed.

To improve the inference speed, some recent work such as Pixel-BERT [3], E2E-VLP [19] and ViLT [10] removes the complicated object detector in feature extraction, and conducts end-to-end VL learning with CNN-based grid features and linearly projected patched embeddings, respectively. To combine the benefits of both categories of architectures, VLMo [20] further unifies the dual encoder and fusion encoder modules with shared mixture-of-modality-experts Transformer. In this work, mPLUG introduces a new cross-modal fusion mechanism with cross-modal skip-connections, to enables the fusion to occur at disparate levels in the abstraction hierarchy across the modalities. It achieves superior performances in effectiveness and efficiency across a wide range of VL tasks.

2.2 Skip-connection

Skip-connection is a popular technique to bypass the gradient exploding or vanishing problem for model optimization in deep neural networks, which is widely-used in CV and NLP architectures such as ResNet [21] and Transformer [12]. A variety of skip connection methods have been proposed in recent years [22, 21, 12, 23, 24, 25]. ResNet [21] introduces summed shortcut connections between different layers using simple identity mapping, while highway network [22] designs a transform gating function to control the balance of the input and

the transformed input. DenseNet [23] designs new architectures with concatenated skip-connections, allowing the subsequent layers to re-use all the middle representations of previous layers. Layer Normalization and recursive skip connection are further used in combination with plain skip connection for further stablizing model optimization and better incorporating the transformed input [12, 25]. In this work, mPLUG proposes a new cross-modal skip connection method to address cross-modal fusion problem, and combines the concatenated skip-connection and summed skip-connection for choosing whether to attend to all the concatenated representations of different modalities or just focus on the cross-modal interaction part at each layer.

3 mPLUG

In this section, we will first introduce our new model architecture with the key module of the cross-modal skip-connected network, and then give the details of the pre-training objectives and scalable training infrastructure.

3.1 Model Architecture

As shown in Figure 2, mPLUG consists of two uni-modal encoders for image and text independently, a cross-modal skip-connected network and a decoder for text generation. To better model the inherent modality bias information, we first use two unimodal encoders to encode image and text separately. Following [11, 26], we use a visual transformer [27] directly on the image patches as the visual encoder, which is more computation-friendly than using pre-trained object detectors for visual feature extraction [8, 9]. The visual encoder divides an input image into patches and encodes them as a sequence of embeddings $\{v_{cls}, v_1, v_2, ..., v_M\}$ with an additional $[CLS]$ token. The input text is fed to the text encoder and represented as a sequence of embeddings $\{l_{cls}, l_1, l_2, ..., l_N\}$, where l_{cls} is the embedding of the $[CLS]$ token and used to summarize the input text. Then, the visual and linguistic representations are fed into a cross-modal skip-connected network, which consists of multiple skip-connected fusion blocks. In each skip-connected fusion block, we adopt connected cross-modal fusion to each of S *asymmetric co-attention* layers where S is a fixed stride value. The aim of this network is to take advantage of the effectiveness of the connected cross-modal fusion and the efficiency of the asymmetric co-attention for
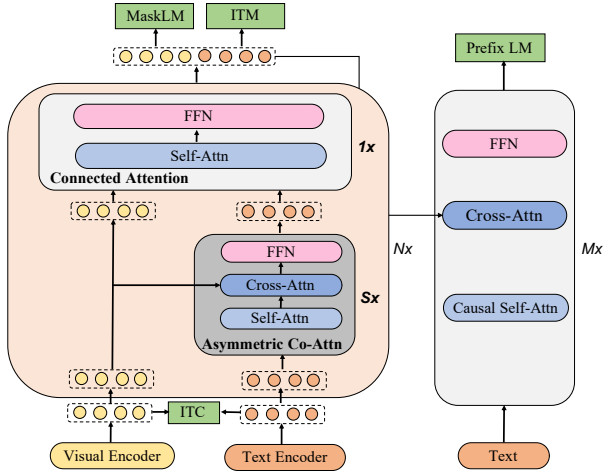
```
Algorithm 1: Pseudocode of Cross-modal Skip-connected Network.

# image, text.ids, text.mask: paired {image, text} pairs.
# image_encoder: vision transformer based encoder.
# text_encoder: language transformer based encoder.
# S: the number of skipped layers in the asymmetric co-attention
# T: total layers of cross-modal skip-connections

def connected_layer(img_feature, txt_feature):
    fusion_feature = concat(img_feature, txt_feature)
    fusion_feature = norm(self_attn(fusion_feature) + fusion_feature)
    fusion_feature = norm(ffn(fusion_feature) + fusion_feature)
    img_feature, txt_feature = split(fusion_feature)
    return img_feature, txt_feature

# asymmetric co-attention architecture
def cross_layer(img_feature, txt_feature):
    txt_feature = norm(self_attn(txt_feature) + txt_feature)
    txt_feature = norm(cross_attn(txt_feature, img_feature)+
        txt_feature)
    txt_feature = norm(ffn(txt_feature) + txt_feature)
    return img_feature, txt_feature

def skip_connected_network(img_feature, txt_feature, S):
    for i in range(1, T+1):
        encoder = connected_layer if (i % (S+1) == 0)
            else cross_layer
        img_feature, txt_feature = encoder(img_feature, txt_feature)
    fusion_feature = concat(img_feature, txt_feature)
    return fusion_feature

img_feature = image_encoder(image)
txt_feature = text_encoder(text.ids, text.mask)
fusion_feature = skip_connected_network(img_feature, txt_feature, S)
```

Figure 2: The model architecture and objectives of mPLUG, which consists of two unimodal encoders for images and text separately, a cross-modal skip-connected network and a decoder for text generation. An image-text contrastive loss is first applied to align the unimodal representations from the visual encoder and text encoder. Then, we use a novel cross-modal skip-connected network to fuse the visual and linguistic representations effectively and efficiently. We adopt connected cross-modal fusion to every S *asymmetric co-attention* layers, where S is a fixed stride value. Based on the connected representation of the image and prefix sub-sequence, the decoder is trained with a prefix language modeling (Prefix LM) loss by generating the remaining caption.

enhanced cross-modal fusion in a recursive manner. Finally, the output cross-modal representations are fed into a transformer decoder for sequence to sequence learning, which equips mPLUG with both understanding and generation capabilities.

3.2 Cross-modal Skip-connected Network

The cross-modal skip-connected network consists of N skip-connected fusion blocks. In each skip-connected fusion block, we adopt *connected-attention* layer to each of S *asymmetric co-attention* layers where S is a fixed stride value. We first pass the text feature and image feature from unimodal encoders through the S asymmetric co-attention layers, and then connect the output text feature and image feature to one connected-attention layer. We repeat the skip-connected fusion block N times for the final connected image and text representation.

Specifically, the asymmetric co-attention is composed of the self-attention (SA) layer, cross-attention (CA) layer and the feed-forward network (FFN). The input text feature l^{n-1} is first fed to the self-attention layer, and then the visual feature v^{n-1} is injected into the text feature l^n_{SA} by the cross-attention layer which gives l^n_{CA}. The output of self-attention l^n_{SA} and cross-attention l^n_{SA} are

added up and fed to the FFN layer for the visual-aware text representation l^n:

$$l^n_{SA} = LN(SA(l^{n-1}) + l^{n-1}) \tag{1}$$

$$l^n_{CA} = LN(CA(l^n_{SA}, v^{n-1}) + l^n_{SA}) \tag{2}$$

$$l^n = LN(FFN(l^n_{CA}) + l^n_{CA}) \tag{3}$$

where LN is short for layer normalization.

The connected-attention layer is composed of the self-attention (SA) layer and the feed-forward network (FFN). We connect the image feature v^{n-1} and input text feature l^{n-1}, where l^{n-1} is the output of S asymmetric co-attention layers. The connected image and text feature $[v^{n-1}; l^{n-1}]$ are fed to the self-attention layer and FFN layer:

$$[v^n_{SA}; l^n_{SA}] = LN(SA([v^{n-1}; l^{n-1}])+[v^{n-1}; l^{n-1}]) \tag{4}$$

$$[v^n; l^n] = LN(FFN([v^n_{SA}; l^n_{SA}]) + [v^n_{SA}; l^n_{SA}]) \tag{5}$$

Then $[v^n; l^n]$ is fed into the next cross-modal skip-connected network repeatedly to get the final connected image and text representation. Finally, the connected output is fed into a Transformer decoder for sequence to sequence learning.

3.3 Pre-training Tasks

We perform four pre-training tasks including three understanding tasks (Image-Text Contrastive Learning, Image-Text Matching, Masked Language Modeling) and one generation task (Prefix Language Modeling). These pre-training tasks are optimized jointly.

Image-Text Contrastive (ITC): Following [6], we employ the task to align the image features and the text features from the unimodal encoders. Specifically, we calculate the softmax-normalized image-to-text and text-to-image similarity, and take two dynamic memory queues (text, image) to increase the number of negative examples as MoCo [28].

Image-Text Matching (ITM): This task aims to predict whether an image and a sentence match with each other on the cross-modal representation. We also select hard negative image-text pairs based on the contrastive text-image similarity as [6].

Masked Language Modeling (MLM): The task setup is basically the same as in BERT [29], where we randomly mask 15% of tokens in text and the model is asked to predict these masked words with the cross-modal representations.

Prefix Language Modeling (PrefixLM): This task aims to generate the caption given an image and predict the text segment subsequent to the cross-modal context as [30]. It optimizes a cross entropy loss by maximizing the likelihood of text in an autoregressive manner.

4 Distributed Learning on a Large Scale

Training a big model like mPLUG on large-scale datasets faces many efficiency challenges. We increase the throughput from the perspective of reducing memory usage and computation time, thereby accelerating the training of the model.

The memory usage during model training is mainly composed of two aspects: the static memory usage composed of parameters/optimizer states/gradients, etc., and the runtime memory usage caused by intermediate variables like activation values. For static memory overhead, we use the ZeRO [31] technique to partition parameters/optimizer states/-gradients into the entire data-parallel group, so that the static memory overhead of a single GPU can be approximately reduced to $1/N$, where N denotes the number of GPU cards. We use gradient checkpointing [32] for the runtime memory cost, which greatly reduces the runtime memory usage at the

expense of increasing forward time by recomputing part of the activation values during backward pass without keeping them in memory.

To reduce the computation time, we use BF16 precision training. BF16 is a new data type supported by NVIDIA's new Ampere architecture GPU like A100. Compared with the previously widely used mixed-precision training of FP16 and FP32, BF16 has the same representation range as FP32, thereby reducing the risk of numerical overflow and ensuring model convergence stability, and at the same time has the same fast computing speed as FP16.

5 Experiments

5.1 Data & Setup

Following the previous work [6], we use the same pre-training dataset with 14M images with texts, which includes two in-domain datasets (MS COCO [36] and Visual Genome [37]), and three web out-domain datasets (Conceptual Captions [38], Conceptual 12M [39], SBU Captions [40].

We pretrain the model for 30 epochs with the total batch size of 1024 on 16 NVIDIA A100 GPUs. We use a 6-layer Transformer for both the text encoder and the cross-modal skip-connected network, and a 12-layer Transformer for the decoder. The text encoder is initialized using the first 6 layers of the BERT$_{base}$ [29] model and the skip-connected network is initialized using the last 6 layers of the BERT$_{base}$. We initialize the visual encoder by CLIP-ViT [17] pretrained on 400M noisy image-text pairs. The visual transformer with ViT-B/16 is used as our base architecture, the one with ViT-L/14 as the large architecture. We use the AdamW [41] optimizer with a weight decay of 0.02. The learning rate is warmed-up to 1e-5 (ViT-B/16) and 1e-4 (BERT$_{base}$) for mPLUG$_{ViT-B}$, and 5e-6 (ViT-L/14) and 5e-5 (BERT$_{base}$) for mPLUG$_{ViT-L}$ in the first 1000 iterations, and decayed to 1e-6 following a cosine schedule. During pre-training, we take random image crops of resolution 256×256 (ViT-B/16)/224×224 (ViT-L/14) as input, and also apply RandAugment [42] to improve the generalization of vision encoders. For VQA and image captioning tasks, we do an additional continue pre-training on 4M image-text pairs. We increase the image resolution during finetuning. For image-text contrastive learning, the queue size is set as 65,536 and the momentum coefficient is set as 0.995.

| Models | Data | COCO Caption | | | | | | | | NoCaps | |
| | | Cross-entropy Optimization | | | | CIDEr Optimization | | | | | |
		B@4	M	C	S	B@4	M	C	S	C	S
Encoder-Decoder	CC12M	-	-	110.9	-	-	-	-	-	90.2	12.1
E2E-VLP [19]	4M	36.2	-	117.3	-	-	-	-	-	-	-
VinVL [9]	5.65M	38.5	30.4	130.8	23.4	41.0	31.1	140.9	25.2	97.3	13.8
OSCAR [4]	6.5M	-	-	-	-	41.7	30.6	140.0	24.5	83.4	11.4
SimVLM$_{large}$ [7]	1.8B	40.3	**33.4**	**142.6**	**24.7**	-	-	-	-	-	-
LEMON$_{large}$ [33]	200M	40.6	30.4	135.7	23.5	42.3	31.2	144.3	25.3	113.4	**15.0**
BLIP [34]	129M	40.4	-	136.7	-	-	-	-	-	113.2	14.8
OFA [35]	18M	-	-	-	-	43.5	31.9	149.6	**26.1**	-	-
mPLUG	14M	**43.1**	31.4	141.0	24.2	**46.5**	**32.0**	**155.1**	26.0	**114.8**	14.8

Table 1: Evaluation Results on COCO Caption "Karpathy" test split and NoCaps validation set. B@4: BLEU@4, M: METEOR, C: CIDEr, S: SPICE.

Models	Data	Test-dev	Test-std
Pretrained on COCO, VG, SBU and CC datasets			
VLBERT [43]	4M	71.16	-
E2E-VLP [19]	4M	73.25	73.67
VL-T5 [44]	4M	-	71.30
UNITER[2]	4M	72.70	72.91
OSCAR[4]	4M	73.16	73.44
CLIP-ViL[26]	4M	76.48	76.94
METER[11]	4M	77.68	77.64
ALBEF[6]	4M	74.54	74.70
mPLUG$_{ViT-B}$	4M	**77.94**	**77.96**
Models Pretrained on More Data			
ALBEF [6]	14M	75.84	76.04
BLIP [34]	129M	78.25	78.32
SimVLM [7]	1.8B	80.03	80.34
Florence [45]	0.9B	80.16	80.36
OFA [35]	18M	79.87	80.02
VLMo [20]	-	79.94	79.98
mPLUG$_{ViT-B}$	14M	79.79	79.81
mPLUG$_{ViT-L}$	14M	**81.27**	**81.26**

Table 2: Evaluation Results on VQA test set.

5.2 Evaluation on Vision-Language Tasks

We compare our pre-trained model against other VLP models on the six downstream V+L tasks. We introduce each task and our fine-tuning strategy below. Details of the datasets and fine-tuning hyperparameters are in Appendix.

5.2.1 Visual Question Answering

The VQA task [14] requires the model to answer natural language questions given an image. Most methods [1, 20, 4, 7] deal with visual question answering tasks as multi-label classification on pre-defined answer sets. This strategy achieves strong performance, but it is not suitable for real-world open scenarios. We treat VQA as an answer generation task and directly use unconstrained open-vocab generation during inference, which is different from constrained close-vocab generation models [6, 35]. Following [4, 35], we concatenate the question with the object labels and OCR tokens extracted from image. As shown in Table 2, mPLUG achieves 81.27 on Test-std split and outperforms the SOTA models including SimVLM and Florence, which use $100X$ and $60X$ more pre-training image-text pairs, respectively. Based on the same 4M pre-training data, mPLUG outperforms CLIP-ViL and METER, which also use CLIP [17] as the visual encoder. Besides, under the same settings, mPLUG always significantly outperforms ALBEF and BLIP which only rely on co-attention from images to text for cross-modal fusion. The gain can derive from the network design of cross-modal skip-connections specifically for information asymmetry of the two modalities. Neither ALBEF nor BLIP addresses this problem well, with bias towards the language modality.

5.2.2 Image Captioning

The image captioning task requires a model to generate an appropriate and fluent caption for a given image. We evaluate image captioning on two datasets COCO Caption [47] and NoCaps [48]. mPLUG finetuned with training data of COCO Caption is tested on both of the datasets. We train

Models	# Pretrain data	MSCOCO (5K test set)						Flickr30K (1K test set)					
		TR			IR			TR			IR		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
E2E-VLP [19]	4M	-	-	-	-	-	-	86.2	97.5	98.92	73.6	92.4	96.0
UNITER [2]	4M	65.7	88.6	93.8	52.9	79.9	88.0	87.3	98.0	99.2	75.6	94.1	96.8
OSCAR [4]	4M	70.0	91.1	95.5	54.0	80.8	88.5	-	-	-	-	-	-
UNIMO [46]	4M	-	-	-	-	-	-	89.4	98.9	99.8	78.0	94.2	97.1
VLMo [20]	4M	78.2	94.4	97.4	60.6	84.4	91.0	95.3	99.9	100.0	84.5	97.3	98.6
ALIGN [18]	1.8B	77.0	93.5	96.9	59.9	83.3	89.8	95.3	99.8	100.0	84.9	97.4	98.6
ALBEF [6]	14M	77.6	94.3	97.2	60.7	84.3	90.5	95.9	99.8	100.0	85.6	97.5	98.9
Florence [45]	0.9B	81.8	95.2	-	63.2	85.7	-	97.2	99.9	-	87.9	98.1	-
BLIP [34]	14M	80.6	95.2	97.6	63.1	85.3	91.1	96.6	99.8	100.0	87.2	97.5	98.8
BLIP [34]	129M	82.4	95.4	97.9	65.1	86.3	91.8	97.4	99.8	99.9	87.6	97.7	99.0
mPLUG	14M	**82.8**	**96.1**	**98.3**	**65.8**	**87.3**	**92.6**	**97.6**	**100.0**	**100.0**	**88.4**	**97.9**	**99.1**

Table 3: Image-text retrieval results on Flickr30K and COCO datasets.

mPLUG on the MS COCO Caption and test on the same Karpathy split [4, 7] and NoCaps validation set. Following [4, 35], we first fine-tune mPLUG with cross-entropy loss and then with CIDEr optimization [49] for extra 5 epochs. As shown in Table 1, mPLUG with only 14M pre-training images can outperform the SOTA models including LEMON and SimVLM on both COCO Caption and Nocaps datasets, which uses more than $10X$ and $100X$ pre-training data, respectively. For the COCO Caption, mPLUG performs the best on CIDEr evaluation and surpasses the SOTA model by a large margin of 5.5 on Karpathy test set. We use the best checkpoint on COCO Caption and predict on the Nocaps validation set directly.

5.2.3 Image-Text Retrieval

We conduct experiments for both image-to-text retrieval (TR) and text-to-image retrieval (IR) on COCO [36] and Flickr30K [53] datasets. Following [6, 34], we jointly optimize the ITC loss and the ITM loss during fine-tuning. During inference, we first select top-k candidates by computing the dot-product similarity between the image and text encoder features, and then rerank the selected candidates based on their ITM scores. We set $k = 256$ for COCO and $k = 128$ for Flickr30K. As shown in Table 3, mPLUG outperforms all existing methods on both datasets. Using 14M images, mPLUG achieves better performance than BLIP with 129M and Florence with 0.9B pre-training data. Using the same 14M pre-training images, mPLUG substantially outperforms the previous best model BLIP by +2.7% in TR recall@1 on COCO and +1.0 % in TR recall@1 on Flickr30K.

5.2.4 Visual Grounding

Given a query in plain text and an image, visual grounding requires models to localize the referred object in the image. Instead of regressing the bounding boxes directly, we concatenate visual features and attended textual features and feed them into the decoder to predict the coordinates. Table 4 shows that mPLUG outperforms all the SOTA methods. We observe that in RefCOCO testB the images often contain arbitrary objects and in Rec-COCOg test-u the expressions are longer than other datasets. Compared with the previous best model OFA, mPLUG achieves 3.16% absolute improvement on RefCOCO testB and 1.22% absolute improvement on RefCOCOg test-u. It demonstrates that mPLUG learns better multi-modal interaction from cross-modal skip-connections and is better at handling complex images and long queries.

5.2.5 Visual Reasoning

We consider two datasets for visual reasoning: NLVR2 [54] and SNLI-VE [55]. The NLVR2 [54] task requires the model to predict whether a sentence describes a pair of images. Following [34], we use two cross-attention layers to process the two input images, and their outputs are merged and fed to the FFN. An MLP classifier is then applied on the output embedding of the language [CLS] token. The SNLI-VE [55] task requires the model to evaluate how the given image and text are semantically correlated, i.e., entailment, neutral, or contradiction. Following [35], the image premise, text premise and text hypothesis are fed to the encoder. While we remove the decoder, and only use the encoder modules for three-way classification, which can save nearly half of the total computation cost. We predict the class probabilities using

Model	RefCOCO			RefCOCO+			RefCOCOg	
	val	testA	testB	val	testA	testB	val-u	test-u
VLBERT [43]	-	-	-	72.59	78/57	62.30	-	-
UNITER [2]	81.41	87.04	74.17	75.90	81.45	66.70	74.86	75.77
VILLA [50]	82.39	87.48	74.84	76.17	81.54	66.84	76.18	76.71
MDETR [51]	86.75	89.58	81.41	79.52	84.09	70.62	81.64	80.89
UNICORN [52]	88.29	90.42	83.06	80.30	85.05	71.88	83.44	83.93
OFA [35]	90.05	92.93	85.26	84.49	90.10	77.77	84.54	85.20
mPLUG	**92.40**	**94.51**	**88.42**	**86.02**	**90.17**	**78.17**	**85.88**	**86.42**

Table 4: Visual grounding results (Acc@0.5) on ReferCOCO, ReferCOCO+, and ReferCOCOg.

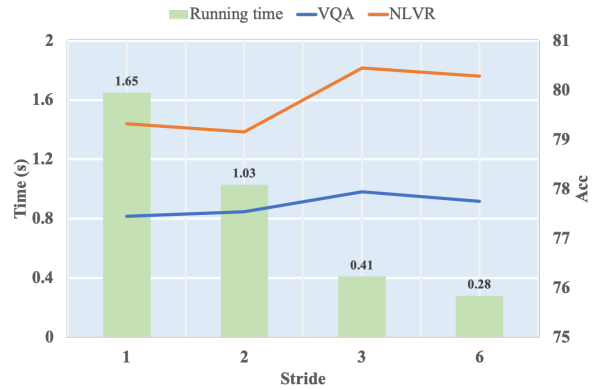| Model	NLVR2		SNLI-VE	
	dev	test-P	dev	test
LXMERT[1]	74.90	74.50	-	-
VL-T5[44]	-	73.6	-	-
UNITER[2]	79.12	79.98	79.39	79.38
CLIP-ViL[26]	-	-	80.61	80.20
METER[11]	82.33	83.05	80.86	81.19
UNIMO[46]	-	-	81.11	80.63
ALBEF[6]	82.55	83.14	80.80	80.91
BLIP[34]	82.67	82.30	-	-
SimVLM$_{large}$[7]	84.13	84.84	85.68	85.62
VLMo[20]	**85.64**	**86.86**	-	-
OFA[35]	-	-	**90.30**	**90.20**
mPLUG	84.58	84.95	89.45	89.29

Table 5: Evaluation Results on NLVR2 and SNLI-VE.

Figure 3: Results w.r.t different stride values in cross-modal skip-connected network on running time and performance of VQA test-dev and NLVR2 test-P, where the running time is the total forward time of 100 samples.

the multimodal encoder's output representation of the language [CLS] token. As shown in Table 5, mPLUG can obtain competitive performances to the SOTA models [1] in both visual reasoning tasks, and even outperform SimVLM [7] and BLIP [34], which use far more pre-training data.

5.3 Effectiveness and Efficiency

To validate the effectiveness and efficiency of our proposed cross-modal skip-connected network, we conduct in-depth analysis on different stride values and various cross-modal fusion methods.

5.3.1 Analysis of Stride for Skip

The stride S is the key factor to control the effectiveness and efficiency tradeoff. Therefore, we further compare the running time and performance of dif-

ferent stride value S in cross-modal skip-connected network on VQA and NLVR2 tasks. Specifically, we test four different stride values, which can be divisible by the total number of cross-modal fusion layers. The model is chosen as mPLUG$_{ViT-B}$ and all the other experiment settings are kept the same. As shown in Figure 3, we can see that the larger S is, the more efficient cross-modal fusion is, where the running time can be largely reduced from skipping the vision co-attention layers by $5X$ times from $S = 1$ to $S = 6$. The performances of mPLUG on both datasets gradually increases when $S = 3$, and slightly decreases later on. Compared with $S = 3$, mPLUG can achieve comparable performance at $S = 6$, while speeding up by nearly 30%. Therefore, we set $S = 6$ on mPLUG$_{ViT-L}$ for faster pre-training.

5.3.2 Analysis of Cross-modal Fusion

We compare the effectiveness and efficiency of different cross-modal fusion variants in terms of run-

---

[1]The SOTA models such as OFA and VLMo both add large-scale text-only and image-only pre-training data for improving the reasoning ability.
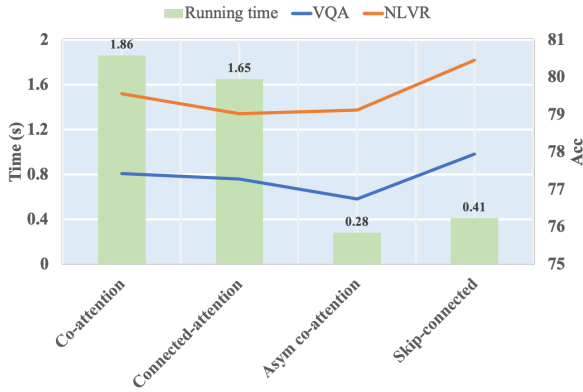
Figure 4: Results w.r.t different cross-modal fusions on running time and performance on VQA test-dev and NLVR2 test-P, where the running time is the total forward time of 100 samples.

Model	Throughput (Samples/S)
baseline	124.0
+ BFloat16	182.7
+ Gradient Checkpoint	238.2
+ ZeRO	**422.5**

Table 6: Training Throughput

ning time and performance on VQA and NLVR2 tasks. Specifically, we pre-train mPLUG with different cross-modal fusion network based on the same image encoder and text encoder. All the pre-training settings and the number of fusion layers are kept the same as in the original mPLUG pre-training. As shown in Figure 4, the fusion methods of co-attention and connected-attention both requires much more running time due to long visual sequence. Compared with the two fusion methods, our proposed skip-connected network is $4X$ faster and obtain better performance on both datasets. We also compare it with the asymmetric co-attention used in BLIP [6, 34] which only relies on the co-attention layers from images to text. Despite running slightly faster than the skip-connected network does, the asymmetric co-attention performs worse in accuracy on both datasets. The performance degradation is attributed to the information asymmetry and bias towards language, as shown in Section 5.2.1.

5.3.3 Large-scale Training

Combining the techniques introduced in Section 4 has dramatically increased the training throughput. With the utilization of memory saving and acceler-

Model	In	Near	Out	Overall
SimVLM$_{base}$[7]	83.2	84.1	82.5	83.5
SimVLM$_{huge}$[7]	101.2	100.4	102.3	101.4
Oscar†[4]	85.4	84.0	80.3	83.4
VinVL†[9]	103.7	95.6	83.8	94.3
SimVLM$_{huge}$†[7]	113.7	110.9	115.2	112.2
mPLUG	86.34	81.5	90.49	84.02
mPLUG†	**116.7**	**113.75**	**117.0**	**114.8**

Table 7: Image captioning results on NoCaps validation split (zero-shot and finetuned), and {In, Near, Out} refer to in-domain, near-domain and out-of-domain respectively. † denotes the models finetuned on COCO Caption dataset.

Model	TR		IR	
	R@1	R@5	R@1	R@5
Zero-Shot				
CLIP [17]	88.0	98.7	68.7	90.6
ALIGN [18]	88.6	98.7	75.7	93.8
FLIP [56]	89.8	99.2	75.0	93.4
Florence [45]	90.9	99.1	76.7	93.6
ALBEF† [6]	94.1	99.5	82.8	96.3
BLIP† [34]	94.8	99.7	84.9	96.7
mPLUG	**93.0**	**99.5**	**82.2**	**95.8**
mPLUG†	**95.8**	**99.8**	**86.4**	**97.6**

Table 8: Zero-shot image-text retrieval results on Flickr30K. † denotes the models finetuned on COCO.

ated training techniques, the throughput of mPLUG improves $3X$ more from 124 samples per second to 422 samples per second, as shown in Table 6.

5.4 Zero-shot Transferability

In this section, we examine the generalization of mPLUG and compare the zero-shot result on two Vision-Language and three Video-Language tasks.

5.4.1 Zero-shot Vision-Language Tasks

The pretraining of mPLUG adopts image-text contrastive and prefix language modeling tasks on large-scale image-text pairs. Thus, mPLUG has zero-shot generalization ability in image-text retrieval and image captioning. **Image Caption**: First, we take the pretrained mPLUG model and directly decode on NoCaps validation set without further finetuning. Following[7, 34], we feed a prefix prompt *"A picture of"* into the text encoder to improve the quality of decoded captions.

Model	# Pretrain data	MSRVTT-Retrieval R@1 R@5 R@10
Zero-Shot		
MIL-NCE [57]	How100M	9.9 24.0 32.4
VideoCLIP [58]	How100M	10.4 22.2 30.0
VATT [59]	How100M, AudSet	- - 29.7
ALPRO [60]	W2M, C3M	24.1 44.7 55.4
VIOLET [61]	Y180M, W2M, C3M	25.9 49.5 59.7
CLIP [17]	WIT400M	26.0 49.4 60.7
Florence [45]	FLD900M	37.6 63.8 72.6
BLIP † [34]	129M	43.3 65.6 74.7
mPLUG	14M	38.1 59.2 68.2
mPLUG †	14M	**44.3 66.4 75.4**
Fine-Tuning		
VideoCLIP [58]	How100M	30.9 55.4 66.8
ALPRO [60]	C3M, W2M	33.9 60.7 73.2
VIOLET [61]	Y180M, C3M, W2M	34.5 63.0 73.4

Table 9: Zero-shot video-language results on text-to-video retrieval on the 1k test split of the MSRVTT dataset. † denotes the models finetuned on COCO. Video datasets include HowTo100M [62], WebVid-2M(W2M) [63], YT-Temporal-180M(Y180M) [64]. Image datasets include CC3M(C3M) [38], FLD900M [45], WIT400M [17]. Audio datasets include AudioSet(AudSet) [65].

As shown in Table 7, the zero-shot performance of mPLUG is competitive with fully supervised baselines such like Oscar and VinVL. With further finetuning on MSCOCO dataset, mPLUG outperforms the SimVLM$_{huge}$, which use more pretraining image-text pairs and has larger model parameters. **Image-text Retrieval**: We perform zero-shot retrieval on Flickr30K. The result is shown in Table 8, where zero-shot mPLUG outperforms models (CLIP, ALIGN, Florence) pretrained with more image-text pairs. Following [34], we also evaluate zero-shot retrieval by the model finetuned on MSCOCO dataset. Table 8 shows that mPLUG achieves better performance than the previous SOTA models.

5.4.2 Zero-shot Transfer to Video-Language Tasks

To evaluate the generalization ability of mPLUG to Video-Language Tasks, we conduct zero-shot experiments on Video-text Retrieval, Video Caption and Video Question Answering. Following [34], we uniformly sample n frames for each video ($n = 8$ for Retrieval, $n = 16$ for QA, $n = 8$ for Caption), and concatenate the frame features into a single sequence. **Video-text Retrieval**: We evaluate the mPLUG models pretrained and further finetuned on the COCO-retrieval image-text dataset

Model	MSRVTT-QA Acc	MSVD-QA Acc	VATEX-Cap CIDEr
Zero-Shot			
VQA-T [66]	2.9	7.5	-
BLIP [34]	19.2	35.2	37.4
mPLUG	**21.1**	**37.2**	**42.0**

Table 10: Zero-shot video-language results on Question-Answer and Caption tasks.

without any video pre-training or supervision. Table 9 shows that zero-shot mPLUG can outperform the SOTA models pretrained on far more pretraining data (e.g., Florence, BLIP), and can even outperform models finetuned on the supervised video dataset without using temporal information (e.g., VideoCLIP, VIOLET); **Video Question Answering**: Following BLIP [34], We treat Video QA as an answer generation task and perform evaluation based on models finetuned on VQA. As shown in Table 10, the zero-shot mPLUG outperforms BLIP pretrained with more image-text pairs; **Video Caption**: We use a prefix prompt *"A video of"* to improve the quality of decoded captions. Table 10 shows that zero-shot mPLUG also achieves better performance than BLIP.

6 Conclusion

This paper presents mPLUG, an effective and efficient VLP framework for both cross-modal understanding and generation. mPLUG introduces a new asymmetric vision-language architecture with novel cross-modal skip-connections, to address two fundamental problems of information asymmetry and computation efficiency in cross-modal alignment. Pretrained on large-scale image-text pairs, mPLUG achieves state-of-the-art performance on a wide range of vision-language tasks. mPLUG also demonstrates strong zero-shot transfer ability when directly applied to multiple video-language tasks. Our work explores the cross-modal alignment with a newly-designed VLP architecture and we hope it can help promote future research on image-text foundation models.

References

[1] Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*.

[2] 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.

[3] 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*.

[4] 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.

[5] Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2021. Ernie-vil: Knowledge enhanced vision-language representations through scene graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 3208–3216.

[6] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in Neural Information Processing Systems*, 34.

[7] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. 2021. Simvlm: Simple visual language model pretraining with weak supervision. *CoRR*, abs/2108.10904.

[8] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086.

[9] 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.

[10] Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. *arXiv preprint arXiv:2102.03334*.

[11] Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Zicheng Liu, Michael Zeng, et al. 2021. An empirical study of training end-to-end vision-and-language transformers. *arXiv preprint arXiv:2111.02387*.

[12] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

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

[14] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

[15] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

[16] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. 2016. Modeling context in referring expressions. In *European Conference on Computer Vision*, pages 69–85. Springer.

[17] 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. *arXiv preprint arXiv:2103.00020*.

[18] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. *arXiv preprint arXiv:2102.05918*.

[19] Haiyang Xu, Ming Yan, Chenliang Li, Bin Bi, Songfang Huang, Wenming Xiao, and Fei Huang. 2021. E2e-vlp: End-to-end vision-language pre-training enhanced by visual learning. *arXiv preprint arXiv:2106.01804*.

[20] Wenhui Wang, Hangbo Bao, Li Dong, and Furu Wei. 2021. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *arXiv preprint arXiv:2111.02358*.

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

[22] Rupesh Kumar Srivastava, Klaus Greff, and Jürgen Schmidhuber. 2015. Highway networks. *arXiv preprint arXiv:1505.00387*.

[23] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708.

[24] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. 2017. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-first AAAI conference on artificial intelligence*.

[25] Fenglin Liu, Xuancheng Ren, Zhiyuan Zhang, Xu Sun, and Yuexian Zou. 2021. Rethinking skip connection with layer normalization in transformers and resnets. *arXiv preprint arXiv:2105.07205*.

[26] Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, and Kurt Keutzer. 2021. How much can clip benefit vision-and-language tasks? *arXiv preprint arXiv:2107.06383*.

[27] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

[28] 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.

[29] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[30] Bin Bi, Chenliang Li, Chen Wu, Ming Yan, Wei Wang, Songfang Huang, Fei Huang, and Luo Si. 2020. Palm: Pre-training an autoencoding&autoregressive language model for context-conditioned generation. *arXiv preprint arXiv:2004.07159*.

[31] Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. Zero: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–16. IEEE.

[32] Tianqi Chen, Bing Xu, Chiyuan Zhang, and Carlos Guestrin. 2016. Training deep nets with sublinear memory cost. *arXiv preprint arXiv:1604.06174*.

[33] Xiaowei Hu, Zhe Gan, Jianfeng Wang, Zhengyuan Yang, Zicheng Liu, Yumao Lu, and Lijuan Wang. 2021. Scaling up vision-language pre-training for image captioning. *CoRR*, abs/2111.12233.

[34] 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*.

[35] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022. Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. *arXiv preprint arXiv:2202.03052*.

[36] 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.

[37] 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.

[38] 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.

[39] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. 2021. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3558–3568.

[40] Vicente Ordonez, Girish Kulkarni, and Tamara L Berg. 2011. Im2text: Describing images using 1 million captioned photographs. In *Advances in neural information processing systems*, pages 1143–1151.

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

[42] 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.

[43] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, pages 13–23.

[44] Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. 2021. Unifying vision-and-language tasks via text generation. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 1931–1942. PMLR.

[45] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. 2021. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*.

[46] Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. 2020. Unimo: Towards unified-modal understanding and generation via cross-modal contrastive learning. *arXiv preprint arXiv:2012.15409*.

[47] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. 2015. Microsoft COCO captions: Data collection and evaluation server. *CoRR*, abs/1504.00325.

[48] Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. 2018. nocaps: novel object captioning at scale. *CoRR*, abs/1812.08658.

[49] Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. 2017. Self-critical sequence training for image captioning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1179–1195.

[50] Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. 2020. Large-scale adversarial training for vision-and-language representation learning. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

[51] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. 2021. MDETR - modulated detection for end-to-end multi-modal understanding. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 1760–1770. IEEE.

[52] Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Faisal Ahmed, Zicheng Liu, Yumao Lu, and Lijuan Wang. 2021. Crossing the format boundary of text and boxes: Towards unified vision-language modeling. *CoRR*, abs/2111.12085.

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

[54] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. 2018. A corpus for reasoning about natural language grounded in photographs. *arXiv preprint arXiv:1811.00491*.

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

[56] 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*.

[57] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. 2020. End-to-end learning of visual representations from uncurated instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9879–9889.

[58] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. 2021. Videoclip: Contrastive pre-training for zero-shot video-text understanding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6787–6800.

[59] Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. 2021. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *Advances in Neural Information Processing Systems*, 34.

[60] Dongxu Li, Junnan Li, Hongdong Li, Juan Carlos Niebles, and Steven CH Hoi. 2021. Align and prompt: Video-and-language pre-training with entity prompts. *arXiv preprint arXiv:2112.09583*.

[61] Tsu-Jui Fu, Linjie Li, Zhe Gan, Kevin Lin, William Yang Wang, Lijuan Wang, and Zicheng Liu. 2021. Violet: End-to-end video-language transformers with masked visual-token modeling. *arXiv preprint arXiv:2111.12681*.

[62] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2630–2640.

[63] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. 2021. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1728–1738.

[64] Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. 2021. Merlot: Multimodal neural script knowledge models. *Advances in Neural Information Processing Systems*, 34.

[65] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780. IEEE.

[66] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. 2021. Just ask: Learning to answer questions from millions of narrated videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1686–1697.

[67] 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.

[68] 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.

[69] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. 2016. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20.

7 More Experiments Details

7.1 Downstream Task Details

We evaluate mPLUG on the six downstream vision-language tasks. The hyperparameters that we use for finetuning on the downstream tasks are listed in Table 11. Following [6], all tasks adopt RandAugment, AdamW optimizer with a weight decay of 0.05 and a cosine learning rate schedule. We use an image resolution of 336×336, except for VQA where we use 504×504 images. For VQA and image captioning tasks, we also do an additional continue pre-training on 4M image-text pairs, which can bring about 0.2+ accuracy improvement. Next we introduce the dataset settings in detail.

VQA. We conduct experiment on the VQA2.0 dataset [67], which contains 83k/41k/81k images for training/validation/test. Following [6], we use both training and validation splits for training, and incorporate additional training data from Visual Genome [37].

Task	LR (ViT-L/BERT$_{base}$)	batch size	epochs
VQA	2e-5/5e-6	1024	8
Captioning†	1e-5&8e-7	256	5
Retrieval	1e-5/2e-6	256	5
Visual Grounding	2e-5/2e-6	512	120
NLVR2	5e-5/5e-6	256	15
SNLI-VE	2e-5	64	5

Table 11: Finetuning hyperparameters for downstream tasks. † denotes two stages fine-tuning.

Image Captioning. We finetune on COCO's Karpathy train split, and evaluate on COCO's Karpathy test split and No-Caps validation split. Following [4, 35], we first fine-tune mPLUG with cross-entropy loss for 5 epochs with a learning rate of 1e-5 and a batch size of 256. Based on the fine-tuned model, we the fine-tune it with CIDEr optimization [49] for extra 5 epochs with a smaller learning rate of 8e-7. During inference, we use beam search with a beam size of 10, and set the maximum generation length as 20.

Image-Text Retrieval. We adopt the widely-used Karpathy split [68] for both COCO and Flickr30K. COCO contains 113/5k/5k images for train/validation/test, and Flickr30K contains 29k/1k/1k images for train/validation/test.

Visual Grounding. We evaluate our method on three referring expression grounding datasets: RefCOCO, RefCOCO+ [16] and RefCOCOg [69]. The RefCOCO and RefCOCO+ datasets share 19K images and contain 142/141K queries. The RefCOCOg dataset contains 25K images and 95K queries. To fully use training data, we first train the model with a mixed dataset with a learning rate of 2e-5. Then we continue fine-tuning the model on each dataset with a learning rate of 2e-6.

NLVR2 & SNLI-VE. We conduct experiment both on the official split [54, 55].

7.2 Pre-training Dataset Details

Table 12 shows the statistics of the 14M pre-training images with texts.

	COCO	VG	SBU	CC3M	CC12M
image	113K	100K	860K	3M	10M
text	567K	769K	860K	3M	10M

Table 12: Statistics of the pre-training datasets.