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 pre-trained 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 inefficiency and linguistic signal overwhelmed by long visual sequences in cross-modal alignment. To address both problems, mPLUG introduces an effective and efficient vision-language architecture with novel cross-modal skip-connections.

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, including image captioning, image-text retrieval, visual grounding and visual question answering. mPLUG also demonstrates strong zero-shot transferability on vision-language and video-language tasks. The code and pre-trained models are available at https://github.com/alibaba/AliceMind.

1 Introduction

Large-scale pre-training of vision-language models have recently received tremendous success on a wide range of cross-modal tasks (Tan and Bansal, 2019; Chen et al., 2020; Huang et al., 2020; Li et al., 2020b; Yu et al., 2021; Li et al., 2021b; Wang et al., 2021c). Such vision-language models learn cross-modal representations from a quantity of image-text pairs by aligning the visual and linguistic modalities. The key to learning vision-language models is finding a good alignment between the two modalities to close the semantic gap in-between.

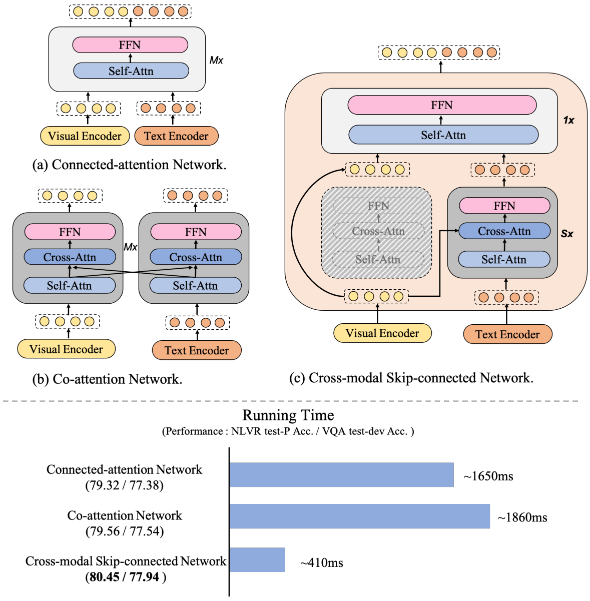


Figure 1: Illustration of two conventional cross-modal fusion networks and our proposed cross-modal skip-connected network. We compare the running time and performance of different fusion networks, where the fusion layers, image encoder and text encoder are all kept the same. The running time is the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

Most recent work (Huang et al., 2020; Wang et al., 2021c; Li et al., 2021b; Kim et al., 2021; Dou et al., 2021) proposes a direct alignment between the image and text representations in an end-to-end manner, which turns out to perform well on many vision-language tasks. These models extract finer-grained visual representation with a long sequence of image patches or grids for good vision understanding (Dou et al., 2021).

However, there are two significant problems in fusing the asymmetric information caused by long visual sequences: 1) *vanishing information*: 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 images as long visual sequences. This infor-

[*] Equal contribution
[†] Corresponding authors

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 7241–7259
December 7-11, 2022 ©2022 Association for Computational Linguistics

mation asymmetry leads to linguistic signal overwhelmed by visual signal when information of the two modalities are directly blended. The linguistic signal can be overlooked, which hinders effective cross-modal fusion, and 2) *inefficiency*: cross-modal fusion focuses so hard on interaction between modalities that heavy and redundant self-attention on long visual sequences takes most computation time. This makes the cross-modal fusion rather inefficient.

One straightforward way of cross-modal fusion is the connected-attention network as shown in Figure 1 (a). It adopts a single Transformer (Vaswani et al., 2017) network for early fusion of vision and language by simply taking the concatenation of visual and linguistic features as input (Li et al., 2019). 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 employs separate Transformer networks for textual and visual features, and uses techniques such as cross-attention to enable cross-modal interaction (Dou et al., 2021), as shown in Figure 1 (b). This architecture design conducts cross-modal fusion on both modalities independently, which helps alleviate the information vanishing. However, it still suffers from computational 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 problem of linguistic signal overwhelmed by visual signal. 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. In one way, it can skip heavy and redundant computation of self-attention between uni-modal visual tokens for efficiency. Besides, we adopt the asymmetric cross-attention from vision to language in certain layers so as to enhance the linguistic representation learning and alleviate information vanishing.

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 which is time-consuming due to long visual sequences. Compared with the connected-attention network, this fusion method keeps more linguistic signal from being overwhelmed by visual signal. 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. This prevents the fused representation from being biased towards linguistic signal and forgetting visual signal. Figure 1 shows that the new cross-modal skip-connected network achieves superior performance with at least four times speed-up than other cross-modal fusion networks. mPLUG pushes the state of the art 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 on a wide range of vision-language and video-language tasks.

2 Related Work

2.1 Vision-Language Pre-training

In terms of how information from different modalities are aggregated, typical approaches to VLP (Tan and Bansal, 2019; Chen et al., 2020; Li et al., 2021b; Radford et al., 2021; Jia et al., 2021; Wang et al., 2022a; Yan et al., 2021; Wang et al., 2022d,c) 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 (Radford et al., 2021) and ALIGN (Jia et al., 2021) is that images and text can be pre-computed and cached, which is quite computation-efficient. However, they tend to fail in handling more complicated VL understanding tasks that require complex reasoning, such as visual question answering (Antol et al., 2015). 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 (Chen et al., 2020) , OSCAR (Li et al., 2020b) and UFO (Wang et al., 2021a), and two-stream architecture such as LXMERT (Tan and Bansal, 2019), ALBEF (Li et al., 2021b), BLIP (Li et al., 2022) and ERNIE-ViL (Yu et al., 2021).

To improve the inference speed, some recent work such as E2E-VLP (Xu et al., 2021a) and ViLT (Kim et al., 2021) 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. 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 (He et al., 2016) and Transformer (Vaswani et al., 2017). A variety of skip connection methods have been proposed in recent years (He et al., 2016; Vaswani et al., 2017; Huang et al., 2017; Liu et al., 2021). ResNet (He et al., 2016) introduces summed shortcut connections between different layers using simple identity mapping, while highway network (Srivastava et al., 2015) designs a transform gating function to control the balance of the input and the transformed input. DenseNet (Huang et al., 2017) designs new architectures with concatenated skip-connections, allowing the subsequent layers to re-use all the middle representations of previous layers. 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.

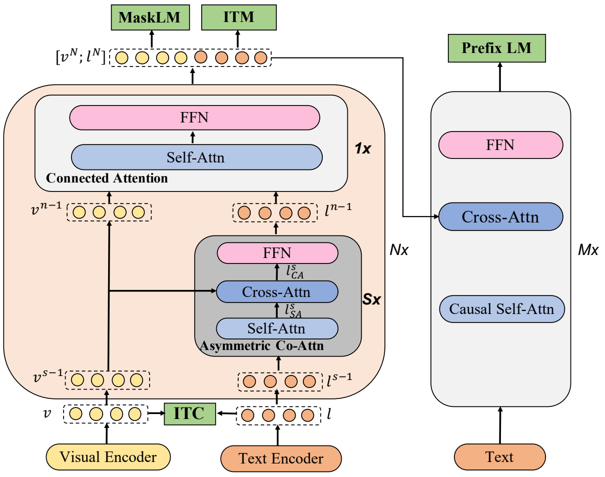
3 mPLUG

3.1 Model Architecture

As shown in Figure 2, mPLUG consists of two unimodal 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 (Dou et al., 2021; Shen et al., 2021), we use a visual transformer (Dosovitskiy et al., 2020) directly on the image patches as the visual encoder, which is more computation-friendly than using pre-trained object detectors for visual feature extraction (Anderson et al., 2018; Zhang et al., 2021). The visual encoder divides an input image into patches and encodes them as a sequence of embeddings $\{v_{cls}, v_1, v_2, ..., v_j\}$ with an additional $[CLS]$ token. The input text is fed to the text encoder and represented as a sequence of embeddings $\{l_{cls}, l_1, l_2, ..., l_k\}$, 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 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 one *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. The asymmetric co-attention layer can retain more linguistic signals so that it is not overwhelmed by visual signals and is more efficient by removing the co-attention on vision side. Then we connect the output text feature and image feature to one connected-attention layer, which



Algorithm 1: Pseudocode of Cross-modal Skip-connected Network.

```
# image, text.ids, text.mask: paired {image, text} pairs.
# vit_encoder: vision transformer based encoder.
# bert_encoder: language transformer based encoder.

def connected_layer(image_features, text_features):
    image_text_features = concat(image_features, text_features)
    attn_features = norm(self_attn(image_text_features))
    ffn_features = norm(ffn(attn_features))
    image_features, text_features = split(ffn_features)
    return image_features, text_features

def cross_layer(image_features, text_features):
    self_attn_features = norm(self_attn(text_features))
    cross_attn_features = norm(cross_attn(self_attn_features,
    image_features))
    text_features = norm(ffn(cross_attn_features))
    return image_features, text_features

def skip_connected_network(image_features, text_features):
    for i in range(M):
        encoder = connected_layer if i % stride == 0 else cross_layer
        image_features, text_features = encoder(image_features,
    text_features)
    image_text_features = concat(image_features, text_features)
    return image_text_features

image_features = vit_encoder(image)
text_features = text_encoder(text.ids, text.mask)
image_text_features = skip_connected_network(image_features, text_features)
```

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.

can prevent the fused representation from being biased towards linguistic signals and forgetting visual signals. 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 feed-forward network (FFN). The input text feature l^{s-1} is first fed to the self-attention layer, and then the visual feature v^{s-1} is injected into the text feature l^s_{SA} by the cross-attention layer which gives l^s_{CA}. The output of cross-attention is added with l^s_{SA} and fed to the FFN layer for the visual-aware text representation l^s:

$$l^s_{SA} = LN(SA(l^{s-1}) + l^{s-1}) \tag{1}$$

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

$$l^s = LN(FFN(l^s_{CA}) + l^s_{CA}) \tag{3}$$

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 and v^{n-1} is equivalent to v^{s-1}. 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 block repeatedly to get the final connected image and text representation. Finally, the connected output $[v^N; l^N]$ is fed into a Transformer decoder (Li et al., 2022, 2021b) for sequence to sequence learning.

3.3 Pre-training Tasks

We perform four standard 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. We provide more details about pre-training tasks in Appendix A.1.

4 Experiments

4.1 Data & Setup

Following the previous work (Li et al., 2021b), we use the same pre-training dataset with 14M images with texts, which includes two in-domain datasets (MS COCO (Lin et al., 2014) and Visual Genome (Krishna et al., 2017)), and three web out-domain datasets (Conceptual Captions (Sharma et al., 2018), Conceptual 12M (Changpinyo et al., 2021), SBU Captions (Ordonez et al., 2011). See Appendix A.2 for more details.

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$_{base}$ (Devlin et al., 2018) model and

Models	# Pretrain Data	VQA		COCO Caption								NoCaps	
				Cross-entropy Optimization				CIDEr Optimization					
		Test-std	Test-dev	B@4	M	C	S	B@4	M	C	S	C	S
E2E-VLP	4M	73.25	73.67	36.2	-	117.3	-	-	-	-	-	-	-
OSCAR	6.5M	73.16	73.44	-	-	-	-	41.7	30.6	140.0	24.5	83.4	11.4
VinVL	5.65M	76.52	76.60	38.5	30.4	130.8	23.4	41.0	31.1	140.9	25.2	97.3	13.8
LEMON$_{large}$	200M	-	-	40.6	30.4	135.7	23.5	42.3	31.2	144.3	25.3	113.4	15.0
UFO	4M	76.76	-	38.7	30.0	131.2	23.3	-	-	-	-	94.3	13.6
METER	4M	77.68	77.64	-	-	-	-	-	-	-	-	-	-
BLIP	129M	78.25	78.32	40.4	-	136.7	-	-	-	-	-	113.2	14.8
VLMo	-	79.94	79.98	-	-	-	-	-	-	-	-	-	-
OFA	18M	79.87	80.02	-	-	-	-	43.5	31.9	149.6	26.1	-	-
SimVLM$_{large}$	1.8B	80.03	80.34	40.3	**33.4**	142.6	**24.7**	-	-	-	-	-	-
Florence	0.9B	80.16	80.36	-	-	-	-	-	-	-	-	-	-
GIT	0.8B	78.81	-	**44.1**	31.5	**144.8**	24.7	44.1	**32.2**	151.1	**26.3**	**125.5**	**16.0**
mPLUG$_{\text{ViT-B}}$	14M	79.89	79.92	41.5	31.1	137.5	23.8	44.9	31.2	150.4	25.2	108.5	13.4
mPLUG$_{\text{ViT-L}}$	14M	**81.27**	**81.26**	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 VQA, COCO Caption "Karpathy" test split and NoCaps validation set. B@4: BLEU@4, M: METEOR, C: CIDEr, S: SPICE. The accuracy of vqa-score is used on VQA. More details about comparison models in Appendix E

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	4M	-	-	-	-	-	-	86.2	97.5	98.92	73.6	92.4	96.0
UNITER	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	4M	70.0	91.1	95.5	54.0	80.8	88.5	-	-	-	-	-	-
VLMo	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	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	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	0.9B	81.8	95.2	-	63.2	85.7	-	97.2	99.9	-	87.9	98.1	-
BLIP	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	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 2: Image-text retrieval results on Flickr30K and COCO datasets.

the skip-connected network is initialized using the last 6 layers of the BERT$_{base}$. Similar to (Li et al., 2022), we explore two variants of ViTs: ViT-B/16 and ViT-L/14 and initialize the visual encoder by CLIP-ViT (Radford et al., 2021). Unless otherwise specified, all results reported in this paper as "mPLUG" uses ViT-L/14. See Appendix A.3 for more details.

4.2 Evaluation on Vision-Language Tasks

We compare our pre-trained model against other VLP models on the six downstream V+L tasks including visual quesion answering on VQAv2(Antol et al., 2015) , image captioning on MS COCO Caption(Chen et al., 2015a) and NoCaps(Agrawal et al., 2018), image-text retrieval on COCO(Lin et al., 2014) and Flickr30K(Plummer et al., 2015), visual grounding on RefCOCO/RefCOCO+/RefCOCOg(Yu et al., 2016; Mao et al., 2016), visual entailment on SNLI-VE(Xie et al., 2019), and visual reasoning on NLVR2(Suhr et al., 2018). Details of the datasets and fine-tuning hyperparameters are in Appendix A.4. Details of the comparison methods are in Appendix E

Visual Question Answering. 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 (Li et al., 2021b; Wang et al., 2022b). As shown in Table 1, 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. Besides, under the same pre-training data, 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 de-

Model	RefCOCO			RefCOCO+			RefCOCOg		NLVR2		SNLI-VE	
	val	testA	testB	val	testA	testB	val-u	test-u	dev	test-P	dev	test
UNITER	81.41	87.04	74.17	75.90	81.45	66.70	74.86	75.77	79.12	79.98	79.39	79.38
METER	-	-	-	-	-	-	-	-	82.33	83.05	80.86	81.19
ALBEF	-	-	-	-	-	-	-	-	82.55	83.14	80.80	80.91
VILLA	82.39	87.48	74.84	76.17	81.54	66.84	76.18	76.71	79.76	81.47	80.18	80.02
MDETR	86.75	89.58	81.41	79.52	84.09	70.62	81.64	80.89	-	-	-	-
UNICORN	88.29	90.42	83.06	80.30	85.05	71.88	83.44	83.93	-	-	-	-
VLMo	-	-	-	-	-	-	-	-	**85.64**	**86.86**	-	-
SimVLM$_{large}$	-	-	-	-	-	-	-	-	84.13	84.84	85.68	85.62
OFA	90.05	92.93	85.26	84.49	90.10	77.77	84.54	85.20	-	-	**90.30**	**90.20**
mPLUG	**92.40**	**94.51**	**88.42**	**86.02**	**90.17**	**78.17**	**85.88**	**86.42**	84.58	84.95	89.45	89.29

Table 3: Evaluation results on Visual grounding (ReferCOCO, ReferCOCO+, and ReferCOCOg), NLVR2 and SNLI-VE. We use the accuracy of IOU 0.5 on visual grounding (a prediction is right if the IoU between the grounding-truth box and the predicted bounding box is larger than 0.5)

sign 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. We also present cases of VQA on out-of-answer-list or out-of-domain images in Appendix 5.2.

Image Captioning. Following (Li et al., 2020b; Wang et al., 2022b), we first fine-tune mPLUG with cross-entropy loss and then with CIDEr optimization (Rennie et al., 2017) 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. The two models use $10X$ and $100X$ pre-training data more than mPLUG . mPLUG performs the best on CIDEr evaluation and surpasses the SOTA model by a large margin of 5.5 on COCO Caption Karpathy test set and 1.4 on NoCaps validation set.

Image-Text Retrieval. As shown in Table 2, 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.

Visual Grounding. Table 3 shows that mPLUG outperforms all the SOTA methods. We observe that in RefCOCO testB the images often contain arbitrary objects and in RecCOCOg 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. See Appendix B for more qualitative examples.

NLVR2 & SNLI-VE. As shown in Table 3, mPLUG can obtain competitive performances to the SOTA models [1] in both NLVR2 and SNLI-VE tasks, and even outperform SimVLM (Wang et al., 2021c) and BLIP (Li et al., 2022), which use far more pre-training data.

4.3 Effectiveness and Efficiency

4.3.1 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 different 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$_{\text{ViT-B}}$ 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$_{\text{ViT-L}}$ for

[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.

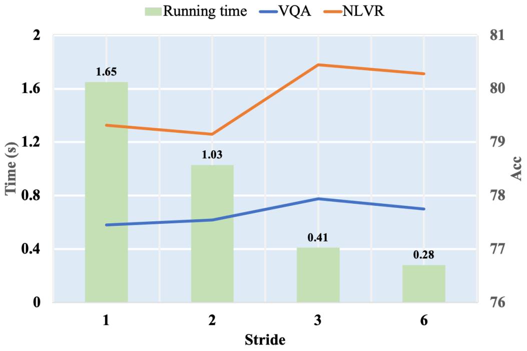


Figure 3: Results w.r.t different stride values in cross-modal skip-connected network on running time and performance of VQA test-dev and NLVR2 test-P, where the running time is the total forward time of 100 samples.

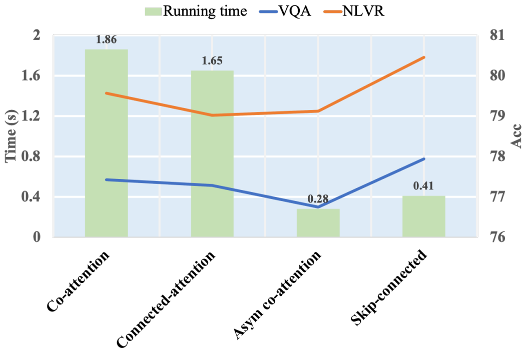


Figure 4: Results w.r.t different cross-modal fusions on running time and performance on VQA test-dev and NLVR2 test-P, where the running time is the total forward time of 100 samples.

faster pre-training.

4.3.2 Analysis of Cross-modal Fusion

We compare the effectiveness and efficiency of different cross-modal fusion variants in terms of running 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 (Li et al., 2021b, 2022) which only relies on the co-attention layers from

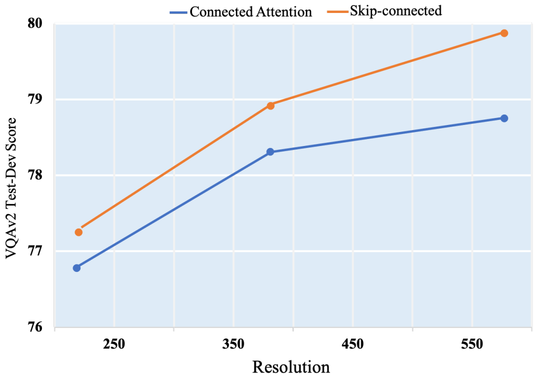


Figure 5: Results w.r.t two cross-modal fusions on VQA test-dev of different image resolution during finetuning.

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 bias towards language and partial visual information forgetting.

In addition, we investigate the effectiveness of our model in tackling the problem of linguistic signal overwhelmed caused by long visual sequences in cross-modal alignment. As shown in Figure 5, we lengthen visual sequences by increasing the image resolution, and test the performance of different fusion models on VQA. When the image resolution is low, our skip-connected fusion model gives marginal improvement compared to the connected-attention fusion model. The improvement becomes significant as the image resolution increases. When the image resolution reaches 576, a visual sequence is 60 times the length of a text sequence. This shows the power of our model in tackling information vanishing. A few cases produced by the mPLUG with different image resolutions are available in Appendix B.

5 Case Study

5.1 Visual Grounding

We present cases with different resolutions in Figure 6 and performance comparisons in Figure 9. We notice that low resolution can lead to failure in multimodal understanding and alignment. Even though the model attends to the correct areas, its predicted bounding box can also be inaccurate because of the low resolution. Thanks to the skip-connection structure, by increasing the resolution of the images, our mPLUG can alleviate the linguistic signal overwhelmed problem and can ade-

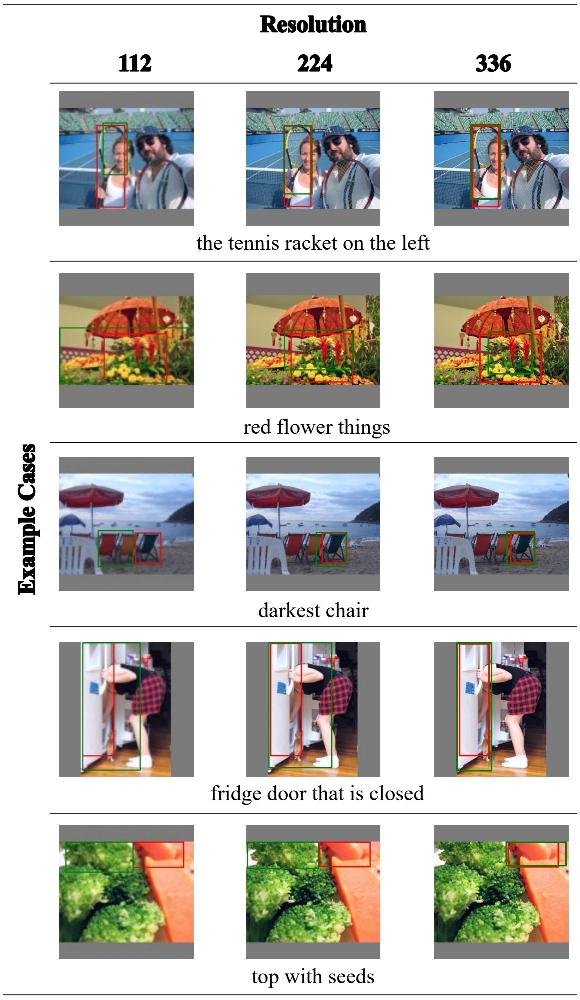


Figure 6: Visual grounding results from mPLUG with different resolution on RefCOCO dataset. **Red** denotes the ground truth bounding box. **Green** denotes the predicted bounding box.

quately utilize the visual and linguistic signal and correct the prediction.

5.2 VQA

As shown in Figure 7, We present cases of VQA on out-of-answer-list and out-of-domain images to demonstrate the capability of transferring to unseen domains and generating the answer that not in the answer list. The first row of images are all from vqa test set and the corresponding answers are not in the 3,129 answer list, which is always used in the previous models. We treat VQA as an answer generation task and directly use unconstrained open-vocab generation during inference. Therefore, our model can answer the question correctly. The second row of images are generated by dall-2 (Ramesh et al., 2022), which are out-of-domain sampls. mPLUG also achieves good performance for the out-of-domain images.

Model	In	Near	Out	Overall
SimVLM$_{base}$	83.2	84.1	82.5	83.5
SimVLM$_{huge}$	101.2	100.4	102.3	101.4
Oscar†	85.4	84.0	80.3	83.4
VinVL†	103.7	95.6	83.8	94.3
SimVLM$_{huge}$†	113.7	110.9	115.2	112.2
mPLUG	86.3	81.5	90.5	84.0
mPLUG†	**116.7**	**113.7**	**117.0**	**114.8**

Table 4: 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	88.0	98.7	68.7	90.6
ALIGN	88.6	98.7	75.7	93.8
FILIP	89.8	99.2	75.0	93.4
Florence	90.9	99.1	76.7	93.6
ALBEF†	94.1	99.5	82.8	96.3
BLIP†	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 5: Zero-shot image-text retrieval results on Flickr30K. † denotes the models finetuned on COCO.

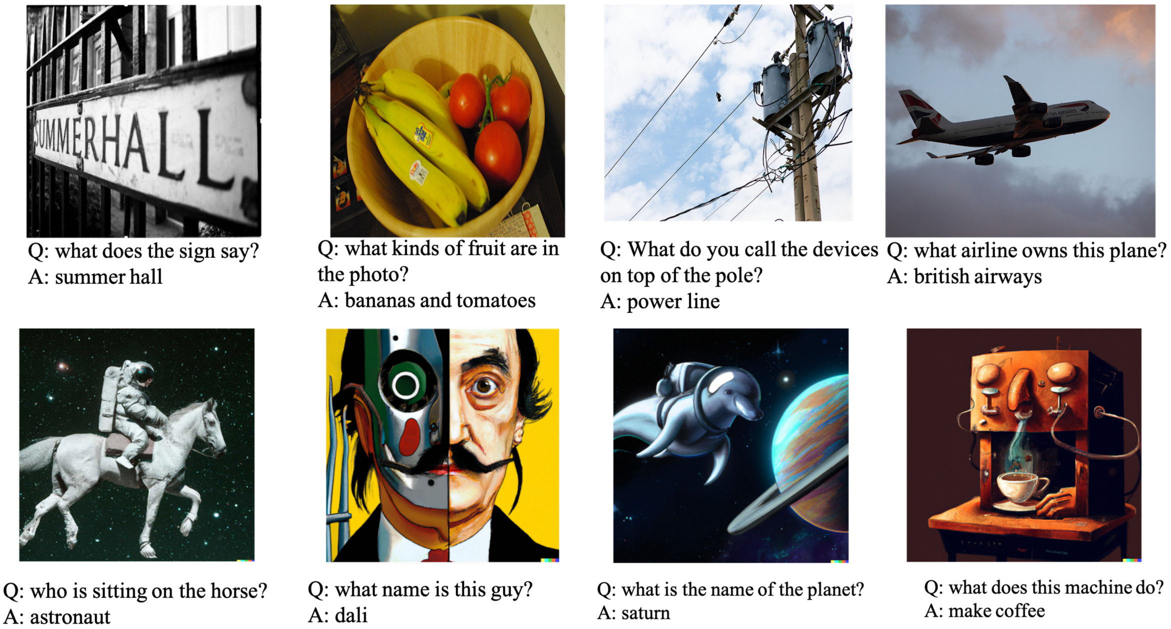
5.3 Zero-shot Transferability

We examine the generalization of mPLUG and compare the zero-shot result on two Vision-Language and three Video-Language tasks.

Image Caption. We take the pretrained mPLUG model and directly decode on NoCaps validation set without further finetuning. As shown in Table 4, 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 pre-training image-text pairs and has larger model parameters.

Image-text Retrieval. We perform zero-shot retrieval on Flickr30K. The result is shown in Table 5, where zero-shot mPLUG outperforms models (CLIP, ALIGN, Florence) pretrained with more image-text pairs. Table 5 shows that mPLUG achieves better performance than the previous SOTA models.

Video-text Retrieval. We evaluate the mPLUG models pretrained and further finetuned on the COCO-retrieval image-text dataset without any



Q: what does the sign say?
A: summer hall

Q: what kinds of fruit are in the photo?
A: bananas and tomatoes

Q: What do you call the devices on top of the pole?
A: power line

Q: what airline owns this plane?
A: british airways

Q: who is sitting on the horse?
A: astronaut

Q: what name is this guy?
A: dali

Q: what is the name of the planet?
A: saturn

Q: what does this machine do?
A: make coffee

Figure 7: The first row of samples are from vqa test set, and the corresponding gold answers are out-of-answer-list. The second row of images are generated by dall-2 (Ramesh et al., 2022), which are out-of-domain samples.

Model	# Pretrain data	MSRVTT-Retrieval		
		R@1	R@5	R@10
Zero-Shot				
MIL-NCE	How100M	9.9	24.0	32.4
VideoCLIP	How100M	10.4	22.2	30.0
VATT	How100M, AudSet	-	-	29.7
ALPRO	W2M, C3M	24.1	44.7	55.4
VIOLET	Y180M, W2M, C3M	25.9	49.5	59.7
CLIP	WIT400M	26.0	49.4	60.7
Florence	FLD900M	37.6	63.8	72.6
BLIP †	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	How100M	30.9	55.4	66.8
ALPRO	C3M, W2M	33.9	60.7	73.2
VIOLET	Y180M, C3M, W2M	34.5	63.0	73.4

Table 6: 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. More details about pretrain data in Appendix A.5

video pre-training or supervision. Table 6 shows that zero-shot mPLUG can outperform the SOTA models pre-trained on more pre-training data (e.g., Florence, BLIP), and can even outperform models finetuned on the supervised video dataset (e.g., VideoCLIP, VIOLET).

Video Question Answering. Following BLIP (Li et al., 2022), We treat Video QA as an answer generation task and perform evaluation based on models finetuned on VQA. As shown in Table 7, the zero-shot mPLUG outperforms BLIP.

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

Table 7: Zero-shot video-language results on Question-Answer and Caption tasks.

Video Caption. Table 7 shows that zero-shot mPLUG also outperforms BLIP for the superior cross-modal generation ability.

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.

7 Limitations

Despite the effectiveness and efficiency of mPLUG across a wide range of downstream image-text tasks, our model still has several limitations:

Scalability. In our current settings, we pre-train mPLUG with only 14M image-text pairs on the largest size of 12+12 layer Transformer encoder-decoder, and it is not clear how well the performance will be if we pre-train a bigger mPLUG on a larger pre-training dataset with other types of available data such as text-only, image-only data as well as some labeled data.

Vision Encoder. The vision encoder within VLP architecture plays important roles on both the effectivenss and efficiency, so far we only experiment on the public CLIP vision encoder, and it is worth investigating how to efficiently extract more semantic-related visual features with a self-designed visual encoder on large-scale image-text and image only data.

Information Vanishing. Due to the presence of remaining connected attention layers, the information vanishing caused by long visual sequences still exists when there is an extremely large gap in information density (e.g., very short text). We think this is a significant problem in cross-modal research, which deserves more in-depth investigation and attention.

References

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.

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.

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.

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.

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.

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*.

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.

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

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

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

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

Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. 2020. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 702–703.

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*.

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*.

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*.

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*.

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

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.

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

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

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.

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.

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.

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

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc 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*.

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1780–1790.

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

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

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.

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

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

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

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*.

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

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

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

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*.

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

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

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.

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.

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.

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.

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.

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*.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*.

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.

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.

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*.

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

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2019. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*.

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*.

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

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.

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

Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. 2022a. Git: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100*.

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

Wei Wang, Bin Bi, Ming Yan, Chen Wu, Zuyi Bao, Jiangnan Xia, Liwei Peng, and Luo Si. 2019. Structbert: incorporating language structures into pre-training for deep language understanding. *arXiv preprint arXiv:1908.04577*.

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

Xinyu Wang, Jiong Cai, Yong Jiang, Pengjun Xie, Kewei Tu, and Wei Lu. 2022c. Named entity and relation extraction with multi-modal retrieval. In *Proceedings of EMNLP*.

Xinyu Wang, Min Gui, Yong Jiang, Zixia Jia, Nguyen Bach, Tao Wang, Zhongqiang Huang, and Kewei Tu. 2022d. ITA: Image-text alignments for multi-modal named entity recognition. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3176–3189, Seattle, United States. Association for Computational Linguistics.

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

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

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

Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. 2021b. 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.

Ming Yan, Haiyang Xu, Chenliang Li, Junfeng Tian, Bin Bi, Wei Wang, Weihua Chen, Xianzhe Xu, Fan Wang, Zheng Cao, Zhicheng Zhang, Qiyu Zhang, Ji Zhang, Songfang Huang, Fei Huang, Luo Si, and Rong Jin. 2021. Achieving human parity on visual question answering. *CoRR*, abs/2111.08896.

Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. 2021a. 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.

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

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*.

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.

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.

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*.

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.

Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021. Vinvl: Making visual representations matter in vision-language models. *CoRR*, abs/2101.00529.

A Implementation Details

A.1 Pre-training Tasks

We perform four pre-training tasks including Image-Text Contrastive Learning (ITC), Image-Text Matching (ITM), Masked Language Modeling (MLM) and Prefix Language Modeling (PrefixLM). The ITC task is first applied to align the unimodal representations from the visual encoder and text encoder. Then, the ITM and MLM are used on the visual and linguistic representations. 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.

Image-Text Contrast (ITC). Following (Li et al., 2021b), 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 (He et al., 2020).

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 (Li et al., 2021b).

Masked Language Modeling (MLM). The task setup is basically the same as in pre-train language models (Devlin et al., 2018; Wang et al., 2019), 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 (Bi et al., 2020). It optimizes a cross entropy loss by maximizing the likelihood of text in an autoregressive manner.

A.2 Pre-training Dataset

Table 8 shows the statistics of the 14M images with texts used in the pre-training stage.

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

Table 8: Statistics of the pre-training datasets.

A.3 Pre-training Details

We use the AdamW (Loshchilov and Hutter, 2017) 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$_{\text{ViT-B}}$, and 5e-6 (ViT-L/14) and 5e-5 (BERT$_{base}$) for mPLUG$_{\text{ViT-L}}$ in the first 1000 iterations, and decayed to 1e-6 following a cosine schedule. mPLUG$_{\text{ViT-B}}$ and mPLUG$_{\text{ViT-L}}$ are pretrained with 16*A100-80G GPUs on the 14M pre-training dataset for 107 hours and 145 hours, respectively.

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 (Cubuk et al., 2020) 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.

A.4 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 9. Following (Li et al., 2021b), 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. The VQA task (Antol et al., 2015) requires the model to answer natural language questions given an image. Most methods (Tan and Bansal, 2019; Wang et al., 2021b; Li et al., 2020b; Wang et al., 2021c) 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 conduct experiment on the VQA2.0 dataset (Goyal et al., 2017), which contains 83k/41k/81k images for training/validation/test. Following (Li et al., 2021b), we use both training and validation splits for training, and incorporate additional training data from Visual Genome (Krishna et al., 2017). Following (Li et al.,

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

2020b; Wang et al., 2022b), we concatenate the question with the object labels and OCR tokens extracted from image.

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 (Chen et al., 2015b) and NoCaps (Agrawal et al., 2018). mPLUG finetuned with training data of COCO Caption is tested on both of the datasets. We train mPLUG on the MS COCO Caption and test on the same Karpathy split (Li et al., 2020b; Wang et al., 2021c) and NoCaps validation set. Following (Li et al., 2020b; Wang et al., 2022b), we first finetune 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 (Rennie et al., 2017) for extra 5 epochs with a smaller learning rate of 8e-7. We use the best checkpoint on COCO Caption and predict on the Nocaps validation set directly. During inference, we use beam search with a beam size of 10, and set the maximum generation length as 20.

Image-Text Retrieval. We conduct experiments for both image-to-text retrieval (TR) and text-to-image retrieval (IR) on COCO (Lin et al., 2014) and Flickr30K (Plummer et al., 2015) datasets. We adopt the widely-used Karpathy split (Karpathy and Fei-Fei, 2015) for both COCO and Flickr30K. COCO contains 113k/5k/5k images for train/validation/test, and Flickr30K contains 29k/1k/1k images for train/validation/test. Following (Li et al., 2021b, 2022), 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.

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. We evaluate our method on three referring expression grounding datasets: RefCOCO, RefCOCO+ (Yu et al., 2016) and RefCOCOg (Mao et al., 2016). 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 consider two datasets for visual reasoning: NLVR2 (Suhr et al., 2018) and SNLI-VE (Xie et al., 2019). The NLVR2 (Suhr et al., 2018) task requires the model to predict whether a sentence describes a pair of images. Following (Li et al., 2022), 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 (Xie et al., 2019) task requires the model to evaluate how the given image and text are semantically correlated, i.e., entailment, neutral, or contradiction. Following (Wang et al., 2022b), 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 the multimodal encoder's output representation of the language [CLS] token.

Zero-shot Vision-Language Tasks. The pre-training 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. Following(Wang et al., 2021c; Li et al., 2022), we feed a prefix prompt *"A picture of"* into the text encoder to improve the quality of decoded captions.

Zero-shot Video-Language Tasks. To evaluate the generalization ability of mPLUG to Video-Language Tasks, we conduct zero-shot experiments

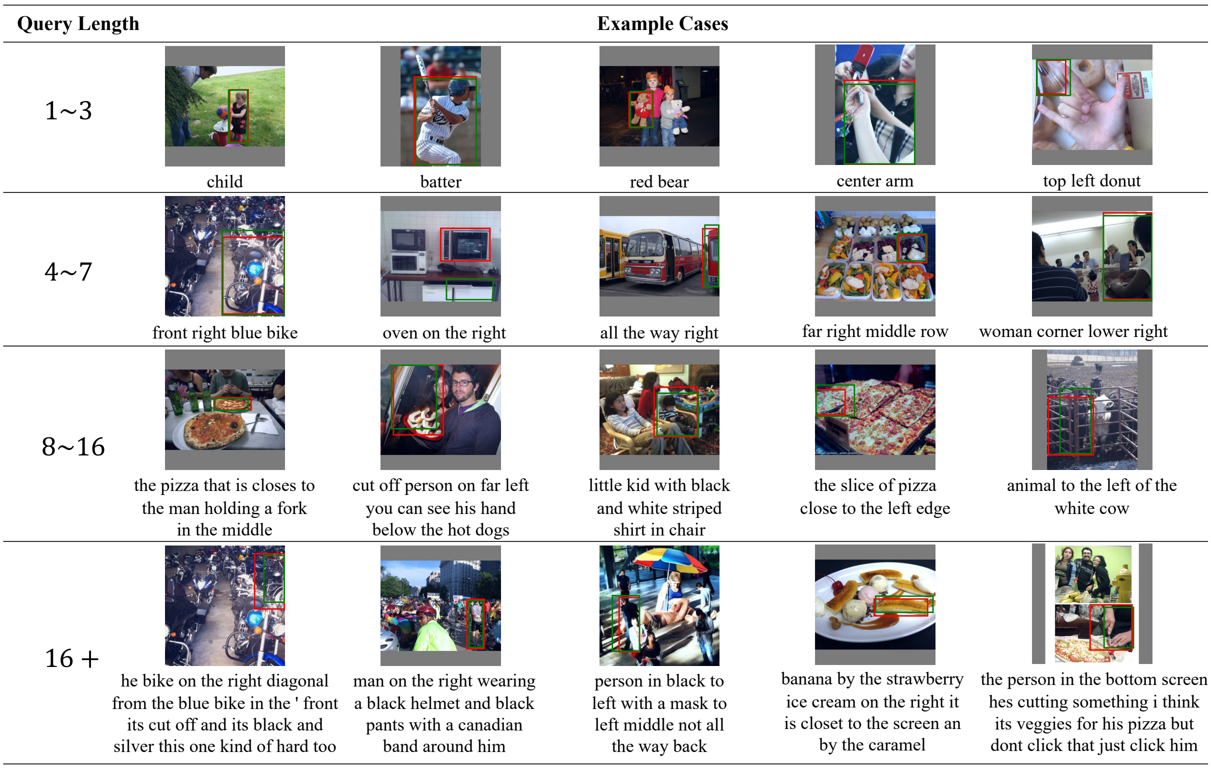

Query Length
Example Cases
1~3
child
batter
red bear
center arm
top left donut
4~7
front right blue bike
oven on the right
all the way right
far right middle row
woman corner lower right
8~16
the pizza that is closes to the man holding a fork in the middle
cut off person on far left you can see his hand below the hot dogs
little kid with black and white striped shirt in chair
the slice of pizza close to the left edge
animal to the left of the white cow
16 +
he bike on the right diagonal from the blue bike in the ' front its cut off and its black and silver this one kind of hard too
man on the right wearing a black helmet and black pants with a canadian band around him
person in black to left with a mask to left middle not all the way back
banana by the strawberry ice cream on the right it is closet to the screen an by the caramel
the person in the bottom screen hes cutting something i think its veggies for his pizza but dont click that just click him

Figure 8: Random sampled visual grounding cases with different query length on RefCOCO dataset. **Red** denotes the ground truth bounding box. **Green** denotes the predicted bounding box.

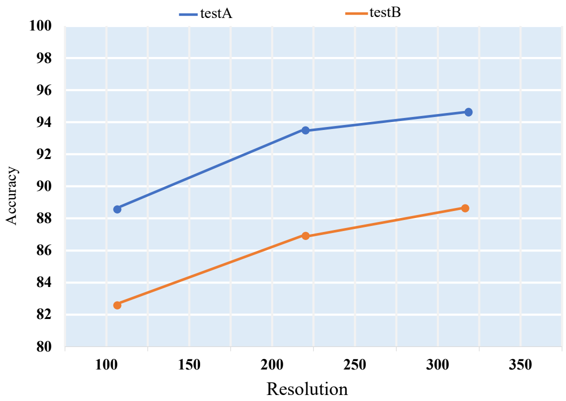


Figure 9: Visual grounding performance comparison between different resolution on RefCOCO dataset.

on Video-text Retrieval, Video Caption and Video Question Answering. Following (Li et al., 2022), 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. For the video caption task, we use a prefix prompt *"A video of"* to improve the quality of decoded captions.

A.5 Video Pre-training Data

The video-language pre-training models always use video datasets and image datasets. The Video datasets include HowTo100M (Miech et al., 2019), WebVid-2M(W2M) (Bain et al., 2021), YT-Temporal-180M(Y180M) (Zellers et al., 2021). Image datasets include CC3M(C3M) (Sharma et al., 2018), FLD900M (Yuan et al., 2021), WIT400M (Radford et al., 2021). Audio datasets include AudioSet(AudSet) (Gemmeke et al., 2017).

B Visualization of Visual Grounding

We group the test data with different query lengths from the RefCOCO dataset and random sample five examples from each group. The results are presented on Figure 8. The results show that our mPLUG can handle both simple and complex queries well. We also notice that the model gives a wrong prediction at the second example in the second row. It means the ability to discriminate against similar objects can be further improved. We also find that when the target object is partially obscured, the model tends to frame the visible part instead of the whole object, which results in inaccurate prediction.

Model	Visual Encoder	Text Encoder	Skip-connected Network	Text Decoder	Total Time
mPLUG$_{\text{ViT-B}}$ (S=1)	2.95s	0.14s	1.65s	0.32s	5.06s
mPLUG$_{\text{ViT-B}}$ (S=6)	2.95s	0.14s	0.28s	0.32s	3.69s
mPLUG$_{\text{ViT-S}}$ (S=1)	0.97s	0.14s	1.65s	0.32s	3.08s
mPLUG$_{\text{ViT-S}}$ (S=6)	0.97s	0.14s	0.28s	0.32s	1.71s

Table 10: The running time of different modules on 100 samples. mPLUG$_{\text{ViT-S}}$ has 4 transformer layers and other parameters are consistent with mPLUG$_{\text{ViT-B}}$

	VQA test-dev
mPLUG$_{\text{ViT-B}}$	**79.89**
w/o ITC	78.17
w/o PrefixLM	78.45
w/o ITM	79.36
w/o MLM	79.54

Table 11: Ablation tests on pre-training tasks of mPLUG on the VQA test-dev set.

C Ablation Study and Time-consuming

Figure 3 calculates and compares the running time of the cross-modal skip-connected module, which is one of the most important modules. In Table 10, we run the whole model on 1*V100-32G GPU to calculate the running time of the total forward time on 100 samples. The speedup is still significant for the whole model when S=6 compared with S=1.

We have conducted the ablation study of pre-training tasks on mPLUG with ViT-B vision encoder and tested the performance on VQA test-dev set. As shown in Table 11, the ITC task and PrefixLM task are the most effective and beneficial.

D Differences from BLIP/ALBEF

Below we give more detailed differences between our mPLUG and BLIP/ALBEF technically. We introduce a new asymmetric vision-language architecture with novel cross-modal skip-connections, to address the problem of linguistic signal overwhelmed by visual signal. and computation inefficiency in multi-modal fusion. We first adopts an asymmetric co-attention architecture at the first few layers for efficiency, by removing the co-attention on vision which is time-consuming due to long visual sequences. To keep the fused representation from being biased towards linguistic signal and forgetting visual signal, we then add one layer of connected-attention, by concatenating the original visual representation and the co-attention output on the language side as input. In contrast, both BLIP and ALBEF employ the asym co-attention architecture, which leads to the fused representation biased towards linguistic signal. Moreover, BLIP only transmits the fused text representation to the decoder, which lacks visual information. It is difficult for the text sequence representation to simultaneously represent the long visual sequence and the text sequence.

E Comparison Methods

- **LXMERT** (Tan and Bansal, 2019): is the pioneering work to pre-train a two-stream multimodal Transformer, which consists of an object relationship encoder, a language encoder and a cross-modality encoder. It is widely used as a baseline method for VLP models.

- **E2E-VLP** (Xu et al., 2021a): proposes the first end-to-end VLP method for both V+L understanding and generation, with a unified Transformer encoder-decoder architecture.

- **VinVL** (Zhang et al., 2021): pre-trains a large-scale object-attribute detection model with much larger amounts of supervised data on four public object detection datasets for extracting better region-based visual feature.

- **OSCAR** (Li et al., 2020b): proposes to use object tags detected in images as anchor points to ease the learning of cross-modal alignments, where the input to the Transformer is a combination of image, text and object tags.

- **LEMON** (Hu et al., 2021): provides the first empirical study on the scaling behavior of VLP for image captioning, and achieves new state of the arts on several major image captioning benchmarks.

- **METER** (Dou et al., 2021): systematically investigates how to design and pre-train a fully transformer-based VL model in an end-to-end manner.

- **VLMo** (Wang et al., 2021b): presents a unified vision-language pretrained model that jointly learns a dual encoder and a fusion encoder with a modular Transformer network.

- **BLIP** (Li et al., 2022): proposes a new VLP framework which transfers flexibly to both vision-language understanding and generation tasks. It effectively utilizes the noisy web data by bootstrapping the captions.

- **OFA** (Wang et al., 2022b): proposes a unified multimodal pretrained model that unifies modalities and tasks based on the encoder-decoder architecture.

- **SimVLM** (Wang et al., 2021c): different from previous VLP methods that only use limited (4M-10M) image-text pairs for pre-training, it proposes a simple VLP model with a single prefix language modeling objective, which pre-trains on a extremely large aligned cross-modal data of about 1.8B noisy image-text pairs. This is also a latest state-of-the-art method on image captioning.

- **Florence** (Yuan et al., 2021): introduces a new computer vision foundation model, which expands the representations from coarse (scene) to fine (object), from static (images) to dynamic (videos), and from RGB to multiple modalities (caption, depth).

- **GIT** (Wang et al., 2022a): proposes a generative image-to-text Transformer, to unify vision-language tasks such as image/video captioning and question answering.

- **ALBEF** (Li et al., 2021b): introduces a contrastive loss to align the image and text representations before fusing them through cross-modal attention, which enables more grounded vision and language representation learning.

- **UNITER** (Chen et al., 2020): proposes an improved single-stream VLP method, by designing two new pre-training strategies: 1) it uses conditional masking on pre-training tasks instead of random masking strategy, 2) it designs a new word-region alignment pre-training task via the use of optimal transport to explicitly encourage fine-grained alignment between words and image regions.

- **ALIGN** (Jia et al., 2021): leverages a noisy dataset of over one billion image alt-text pairs, obtained without expensive filtering or post-processing steps in the Conceptual Captions dataset.

- **VLBERT** (Su et al., 2019): is a pioneering work to pre-train a single-stream multi-modal Transformer, which jointly trains both the Transformer-based cross-modal fusion and Fast R-CNN image feature extractor in both pre-training and fine-tuning phases. It is widely used as a baseline method for VLP models.

- **VL-T5** (Cho et al., 2021): proposes a unified framework that learns different tasks in a single architecture with the same language modeling objective.

- **VILLA** (Gan et al., 2020): is the first known effort on large-scale adversarial training for vision-and-language (V+L) representation learning.

- **MDETR** (Kamath et al., 2021): proposes an end-to-end modulated detector that detects objects in an image conditioned on a raw text query, like a caption or a question.

- **UNICORN** (Yang et al., 2021b): proposes a vision-language (VL) model that unifies text generation and bounding box prediction into a single architecture.

- **CLIP** (Radford et al., 2021): demonstrates that the simple pre-training task of predicting which caption goes with which image is an efficient and scalable way to obtain the image representation.

- **CLIP-ViL** (Shen et al., 2021): proposes to use CLIP as the visual encoder in various VL models, which can significantly outperforms widely-used visual encoders trained with in-domain annotated data.

- **UNIMO** (Li et al., 2020a): proposes a unified-modal pre-training architecture with cross-modal contrastive learning, which can effectively adapt to both single-modal and multi-modal understanding and generation tasks. Except for limited image-text pairs, it utilizes large amounts of single-modal data such as text or image for pre-training.

- **ViLBERT** (Lu et al., 2019): proposes one of the first work that extend the BERT architecture to a multi-modal two-stream VLP model, which processes both visual and textual inputs in separate streams that interact through co-attentional transformer layers.

- **MIL-NCE** (Miech et al., 2020): is capable of addressing mis alignments inherent in narrated videos.
- **ALPRO** (Li et al., 2021a): proposes a new visually-grounded pre-training task, prompting entity modeling. It aims to learn fine-grained region-entity alignment.
- **VATT** (Akbari et al., 2021): presents a framework for learning multi-modal representations from unlabeled data using convolution-free Transformer architectures.
- **VIOLET** (Fu et al., 2021): proposes a fully end-to-end video-language transformer, which adopts a video transformer to explicitly model the temporal dynamics of video inputs.
- **VideoCLIP** (Xu et al., 2021b): proposes a contrastive approach to pre-train a unified model for zero-shot video and text understanding.
- **VQA-T** (Yang et al., 2021a): proposes to avoid manual annotation and generate a large-scale training dataset for video question answering making use of automatic cross-modal supervision.
- **FILIP** (Yao et al., 2021): achieves finer-level alignment through a cross-modal late interaction mechanism. It uses a token-wise maximum similarity between visual and textual tokens to guide the contrastive objective.