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, yml19608, b.bi, yejiabo.yjb, hehong.chh, guohai.xgh, zhengzhi.cz, zjl22146, 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.

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-
Information 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) 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 ap-
proach 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 \( \{v_{cls}, v_1, v_2, \ldots, v_j\} \) with an additional [CLS] token. The input text is fed to the text encoder and represented as a sequence of embeddings \( \{l_{cls}, l_1, l_2, \ldots, l_k\} \), 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
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 \( l^{t-1} \) 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 \( l^s \):

\[
l^{s}_{SA} = LN(SA(l^{t-1}) + l^{t-1}) \\
l^{s}_{CA} = LN(CA(l^{s}_{SA}, v^{s-1}) + l^{s}_{SA}) \\
l^s = LN(FFN(l^{s}_{CA}) + l^{s}_{CA})
\]

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 \( v^{n-1} \). The connected image and text feature \( [v^{n-1}; l^{n-1}] \) are fed to the self-attention layer and FFN layer:

\[
[v^{n}_{SA}; l^{n}_{SA}] = LN(SA([v^{n-1}; l^{n-1}]) + [v^{n-1}; l^{n-1}]) \\
[v^{n}; l^{n}] = LN(FFN([v^{n}_{SA}; l^{n}_{SA}]) + [v^{n}_{SA}; l^{n}_{SA}])
\]

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 BERTbase (Devlin et al., 2018) model and
### 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 question 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-
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).

<table>
<thead>
<tr>
<th>Model</th>
<th>RefCOCO val</th>
<th>RefCOCO testA</th>
<th>RefCOCO+ val</th>
<th>RefCOCO+ testA</th>
<th>RefCOCOg val</th>
<th>RefCOCOg test-u</th>
<th>NLVR2 dev</th>
<th>SNLI-VE dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>UNITER</td>
<td>81.41</td>
<td>87.04</td>
<td>74.17</td>
<td>75.90</td>
<td>67.60</td>
<td>74.86</td>
<td>75.77</td>
<td>79.12</td>
</tr>
<tr>
<td>METER</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>82.33</td>
<td>83.05</td>
</tr>
<tr>
<td>ALBETF</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>82.55</td>
<td>83.14</td>
</tr>
<tr>
<td>VILLA</td>
<td>82.39</td>
<td>87.48</td>
<td>74.84</td>
<td>76.17</td>
<td>66.84</td>
<td>76.18</td>
<td>76.71</td>
<td>79.76</td>
</tr>
<tr>
<td>MDETR</td>
<td>86.75</td>
<td>89.58</td>
<td>81.41</td>
<td>79.52</td>
<td>80.09</td>
<td>81.64</td>
<td>80.89</td>
<td>-</td>
</tr>
<tr>
<td>UNICORN</td>
<td>88.29</td>
<td>90.42</td>
<td>83.06</td>
<td>80.30</td>
<td>71.88</td>
<td>83.44</td>
<td>83.93</td>
<td>-</td>
</tr>
<tr>
<td>VLMo</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>85.64</td>
<td>86.86</td>
</tr>
<tr>
<td>SimVLMlarge</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>84.13</td>
<td>84.84</td>
</tr>
<tr>
<td>OFA</td>
<td>90.05</td>
<td>92.93</td>
<td>85.26</td>
<td>84.49</td>
<td>77.77</td>
<td>84.54</td>
<td>85.20</td>
<td>-</td>
</tr>
<tr>
<td>mPLUG</td>
<td>92.40</td>
<td>94.51</td>
<td>88.42</td>
<td>86.02</td>
<td>90.17</td>
<td>78.17</td>
<td>85.88</td>
<td>86.42</td>
</tr>
<tr>
<td>mPLUG large</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>84.58</td>
<td>84.95</td>
</tr>
</tbody>
</table>

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 RefCOCOg 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.

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 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 ViTB for

---

1The SOTA models such as OFA and VLMo both add large-scale text-only and image-only pre-training data for improving the reasoning ability.
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 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-
The tennis racket on the left
red flower things
darkest chair
fridge door that is closed
top with seeds

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.

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.

<table>
<thead>
<tr>
<th>Model</th>
<th>In</th>
<th>Near</th>
<th>Out</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>SimVLMbase</td>
<td>83.2</td>
<td>84.1</td>
<td>82.5</td>
<td>83.5</td>
</tr>
<tr>
<td>SimVLMhuge</td>
<td>101.2</td>
<td>100.4</td>
<td>102.3</td>
<td>101.4</td>
</tr>
<tr>
<td>Oscar†</td>
<td>85.4</td>
<td>84.0</td>
<td>80.3</td>
<td>83.4</td>
</tr>
<tr>
<td>VinVL†</td>
<td>103.7</td>
<td>95.6</td>
<td>83.8</td>
<td>94.3</td>
</tr>
<tr>
<td>SimVLMhuge†</td>
<td>113.7</td>
<td>110.9</td>
<td>115.2</td>
<td>112.2</td>
</tr>
<tr>
<td>mPLUG</td>
<td>86.3</td>
<td>81.5</td>
<td>90.5</td>
<td>84.0</td>
</tr>
<tr>
<td>mPLUG†</td>
<td>116.7</td>
<td>113.7</td>
<td>117.0</td>
<td>114.8</td>
</tr>
</tbody>
</table>

Table 5: Zero-shot image-text retrieval results on Flickr30K. † denotes the models finetuned on COCO.

<table>
<thead>
<tr>
<th>Model</th>
<th>TR R@1</th>
<th>TR R@5</th>
<th>IR R@1</th>
<th>IR R@5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zero-Shot</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CLIP</td>
<td>88.0</td>
<td>98.7</td>
<td>68.7</td>
<td>90.6</td>
</tr>
<tr>
<td>ALIGN</td>
<td>88.6</td>
<td>98.7</td>
<td>75.7</td>
<td>93.8</td>
</tr>
<tr>
<td>FILIP</td>
<td>89.8</td>
<td>99.2</td>
<td>75.0</td>
<td>93.4</td>
</tr>
<tr>
<td>Florence</td>
<td>90.9</td>
<td>99.1</td>
<td>76.7</td>
<td>93.6</td>
</tr>
<tr>
<td>ALBEF†</td>
<td>94.1</td>
<td>99.5</td>
<td>82.8</td>
<td>96.3</td>
</tr>
<tr>
<td>BLIP†</td>
<td>94.8</td>
<td>99.7</td>
<td>84.9</td>
<td>96.7</td>
</tr>
<tr>
<td>mPLUG</td>
<td>93.0</td>
<td>99.5</td>
<td>82.2</td>
<td>95.8</td>
</tr>
<tr>
<td>mPLUG†</td>
<td>95.8</td>
<td>99.8</td>
<td>86.4</td>
<td>97.6</td>
</tr>
</tbody>
</table>

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 samples. mPLUG also achieves good performance for the out-of-domain images.

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 SimVLMhuge, 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
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

<table>
<thead>
<tr>
<th>Model</th>
<th># Pretrain data</th>
<th>MSRVTT-Retrieval R@1</th>
<th>MSRVTT-Retrieval R@5</th>
<th>MSRVTT-Retrieval R@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zero-Shot</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MIL-NCE</td>
<td>How100M</td>
<td>9.9</td>
<td>24.0</td>
<td>32.4</td>
</tr>
<tr>
<td>VideoCLIP</td>
<td>How100M</td>
<td>10.4</td>
<td>22.2</td>
<td>30.0</td>
</tr>
<tr>
<td>VATF</td>
<td>W2M, C3M</td>
<td>24.1</td>
<td>44.7</td>
<td>55.4</td>
</tr>
<tr>
<td>ALPRO</td>
<td>Y180M, W2M, C3M</td>
<td>25.9</td>
<td>49.5</td>
<td>59.7</td>
</tr>
<tr>
<td>VIOLET</td>
<td>W1T400M</td>
<td>26.0</td>
<td>49.4</td>
<td>60.7</td>
</tr>
<tr>
<td>CLIP</td>
<td>FDL900M</td>
<td>37.6</td>
<td>63.8</td>
<td>72.6</td>
</tr>
<tr>
<td>Florence</td>
<td>129M</td>
<td>43.3</td>
<td>65.6</td>
<td>74.7</td>
</tr>
<tr>
<td>BLIP †</td>
<td>14M</td>
<td>38.1</td>
<td>59.2</td>
<td>68.2</td>
</tr>
<tr>
<td>mPLUG †</td>
<td>14M</td>
<td>44.3</td>
<td>66.4</td>
<td>75.4</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Model</th>
<th>MSRVTT-QA Acc</th>
<th>MSVD-QA Acc</th>
<th>VATEX-Cap CIDEr</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zero-Shot</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VQA-T</td>
<td>2.9</td>
<td>7.5</td>
<td>-</td>
</tr>
<tr>
<td>BLIP</td>
<td>19.2</td>
<td>35.2</td>
<td>37.4</td>
</tr>
<tr>
<td>mPLUG</td>
<td>21.1</td>
<td>37.2</td>
<td>42.0</td>
</tr>
</tbody>
</table>

Table 6 shows that zero-shot mPLUG can outperform the SOTA models pre-trained on more 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.

**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 effectiveness 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


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.


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 (PrefixLM) 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.

<table>
<thead>
<tr>
<th></th>
<th>COCO</th>
<th>VG</th>
<th>SBU</th>
<th>CC3M</th>
<th>CC12M</th>
</tr>
</thead>
<tbody>
<tr>
<td>image</td>
<td>113K</td>
<td>100K</td>
<td>860K</td>
<td>3M</td>
<td>10M</td>
</tr>
<tr>
<td>text</td>
<td>567K</td>
<td>769K</td>
<td>860K</td>
<td>3M</td>
<td>10M</td>
</tr>
</tbody>
</table>

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\textsubscript{base}) for mPLUG\textsubscript{ViT-B}, and 5e-6 (ViT-L/14) and 5e-5 (BERT\textsubscript{base}) for mPLUG\textsubscript{ViT-L} in the first 1000 iterations, and decayed to 1e-6 following a cosine schedule. mPLUG\textsubscript{ViT-B} and mPLUG\textsubscript{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.,
Table 9: Finetuning hyperparameters for downstream tasks. † denotes two stages fine-tuning.

<table>
<thead>
<tr>
<th>Task</th>
<th>LR (ViT-L/BERT base)</th>
<th>batch size</th>
<th>epochs</th>
</tr>
</thead>
<tbody>
<tr>
<td>VQA</td>
<td>2e-5/5e-6</td>
<td>1024</td>
<td>8</td>
</tr>
<tr>
<td>Captioning†</td>
<td>1e-5 &amp; 8e-7</td>
<td>256</td>
<td>5</td>
</tr>
<tr>
<td>Retrieval</td>
<td>1e-5/2e-6</td>
<td>256</td>
<td>5</td>
</tr>
<tr>
<td>Visual Grounding</td>
<td>2e-5/2e-6</td>
<td>512</td>
<td>120</td>
</tr>
<tr>
<td>NLVR2</td>
<td>5e-5/5e-6</td>
<td>256</td>
<td>15</td>
</tr>
<tr>
<td>SNLI-VE</td>
<td>2e-5</td>
<td>64</td>
<td>5</td>
</tr>
</tbody>
</table>

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 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 (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

<table>
<thead>
<tr>
<th>Query Length</th>
<th>Example Cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>1~3</td>
<td>child, batter, red bear, center arm, top left donut</td>
</tr>
<tr>
<td>4~7</td>
<td>front right blue bike, oven on the right, all the way right, far right middle row, woman corner lower right</td>
</tr>
<tr>
<td>8~16</td>
<td>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</td>
</tr>
<tr>
<td>16+</td>
<td>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 and 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</td>
</tr>
</tbody>
</table>

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 8: Visual Grounding Examples](image.png)

### 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.

![Figure 9: Resolution vs Accuracy](image.png)

Figure 9: Visual grounding performance comparison between different resolution on RefCOCO dataset.
<table>
<thead>
<tr>
<th>Model</th>
<th>Visual Encoder</th>
<th>Text Encoder</th>
<th>Skip-connected Network</th>
<th>Text Decoder</th>
<th>Total Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>mPLUGViT-B (S=1)</td>
<td>2.95s</td>
<td>0.14s</td>
<td>1.65s</td>
<td>0.32s</td>
<td>5.06s</td>
</tr>
<tr>
<td>mPLUGViT-B (S=6)</td>
<td>2.95s</td>
<td>0.14s</td>
<td>0.28s</td>
<td>0.32s</td>
<td>3.69s</td>
</tr>
<tr>
<td>mPLUGViT-S (S=1)</td>
<td>0.97s</td>
<td>0.14s</td>
<td>1.65s</td>
<td>0.32s</td>
<td>3.08s</td>
</tr>
<tr>
<td>mPLUGViT-S (S=6)</td>
<td>0.97s</td>
<td>0.14s</td>
<td>0.28s</td>
<td>0.32s</td>
<td>1.71s</td>
</tr>
</tbody>
</table>

Table 10: The running time of different modules on 100 samples. mPLUGViT-S has 4 transformer layers and other parameters are consistent with mPLUGViT-B.

<table>
<thead>
<tr>
<th>VQA test-dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>mPLUGViT-B</td>
</tr>
<tr>
<td>w/o ITC</td>
</tr>
<tr>
<td>w/o PrefixLM</td>
</tr>
<tr>
<td>w/o ITM</td>
</tr>
<tr>
<td>w/o MLM</td>
</tr>
</tbody>
</table>

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 outperform 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 multimodal 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 misalignments 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.