End-to-End Unsupervised Vision-and-Language Pre-training with Referring Expression Matching

Chi Chen[1,2,5], **Peng Li**[*][3], **Maosong Sun**[1,2,5], **Yang Liu**[*][1,2,3,4,5,6]
[1]Dept. of Comp. Sci. & Tech., Institute for AI, Tsinghua University, Beijing, China
[2]Beijing National Research Center for Information Science and Technology
[3]Institute for AI Industry Research, Tsinghua University, Beijing, China
[4]Beijing Academy of Artificial Intelligence, Beijing, China
[5]International Innovation Center of Tsinghua University, Shanghai, China
[6]Quan Cheng Laboratory

Abstract

Recently there has been an emerging interest in unsupervised vision-and-language pre-training (VLP) that learns multimodal representations without parallel image-caption data. These pioneering works significantly reduce the cost of VLP on data collection and achieve promising results compared to supervised VLP. However, existing unsupervised VLP methods take as input pre-extracted region-based visual features from external object detectors, which both limits its flexibility and reduces computational efficiency. In this paper, we explore end-to-end unsupervised VLP with a vision encoder to directly encode images. The vision encoder is pre-trained on image-only data and jointly optimized during multimodal pre-training. To further enhance the learned cross-modal features, we propose a novel pre-training task that predicts which patches contain an object referred to in natural language from the encoded visual features. Extensive experiments on four vision-and-language tasks show that our approach outperforms previous unsupervised VLP methods and obtains new state-of-the-art results[1].

1 Introduction

Vision-and-language pre-training (VLP) (Lu et al., 2019; Li et al., 2019; Chen et al., 2020; Kim et al., 2021; Li et al., 2021c; Zhang et al., 2021; Radford et al., 2021; Ramesh et al., 2021) has achieved great success on a wide range of vision-and-language tasks, e.g., visual question answering (Zhang et al., 2021), image-text retrieval (Radford et al., 2021) and text-to-image generation (Ramesh et al., 2021). The major challenge for VLP is how to bridge the gap between the representations of vision and language modalities, which is typically addressed by training on *large-scale parallel image-text datasets* (Lin et al., 2014; Krishna et al., 2017; Sharma et al., 2018; Ordonez et al., 2011) with specially designed pre-training tasks. However, these datasets require either extensive human annotations or massive data cleaning efforts, making them difficult to collect, especially when compared to the large amount of unimodal data.

To alleviate this problem, there has recently emerged some works exploring unsupervised vision-and-language pre-training (UVLP), where only *non-parallel* image and text data is leveraged (Li et al., 2021b; Zhou et al., 2022). Specifically, Li et al. (2021b) propose to use image region features and their detected object tags produced by an object detector as pesudo-parallel pairs to bridge the gap between the two modalities. Zhou et al. (2022) further enrich the training data with retrieved text pieces based on object tags and pre-train their model with multi-granular alignment tasks. These works achieve competitive results compared to several supervised VLP models, demonstrating the potential of UVLP.

However, current research on UVLP adopts a two-step training strategy that first extracts region-based image features with an external object detector and then builds a multimodal model based on the region features. This is considered to have several limitations for VLP. First, region features may be sub-optimal for VLP because they are designed for object detection tasks rather than general cross-modal understanding and are fixed in the pre-training process (Xu et al., 2021; Huang et al., 2021). Second, the process of extracting region features is time-consuming, which significantly reduces the inference efficiency (Kim et al., 2021). Finally, this two-step training strategy hinders the use of vision pre-trained models (V-PTMs) such as ViT (Dosovitskiy et al., 2020) and Swin Transformer (Liu et al., 2021c), which are not off-

[*]Corresponding authors: Peng Li (lipeng@air.tsinghua.edu.cn) and Yang Liu (liuyang2011@tsinghua.edu.cn)

[1]Code is available at `https://github.com/THUNLP-MT/E2E-UVLP`

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 10799–10810
December 7-11, 2022 ©2022 Association for Computational Linguistics

the-shelf object detectors but achieve promising performance on general vision tasks. Therefore, *how to perform UVLP in an end-to-end manner, i.e., using raw images instead of region features, is still a valuable open question.*

To explore the question, we propose an end-to-end UVLP framework named E2E-UVLP. The framework consists of a vision encoder and a pre-trained language model (PLM), both of which are pre-trained on *unimodal* data and they are connected by a linear projection layer. Taking image patches as input, our framework is capable of leveraging a wide range of V-PTMs. Without using object tags in inference, the computational cost introduced by external object detectors is eliminated. Inspired by previous works (Li et al., 2021b; Zhou et al., 2022; Liu et al., 2021b; Li et al., 2020), we derive a *masked tag prediction (MTP)* pre-training task which predicts the masked object tags given a *raw image* and the other object tags detected from it. Combining MTP with the widely used *masked language modeling (MLM)* task (Devlin et al., 2018), we successfully make E2E-UVLP achieve comparable or better results than existing UVLP methods, justifying *end-to-end UVLP is feasible*.

Although the MTP task is effective, further investigation reveals that the obtained model is less effective when dealing with complex attributes of objects, e.g., locating objects or determining the relationship between objects in an image. We argue it is due to two pitfalls of the MTP objective: (1) Discrepancy between training and inference: An object is referred to by its tag and numerically encoded position in training, while referred to only in natural language in inference. Similar discrepancies have been shown to hurt performance significantly in PLM studies (Brown et al., 2020; Liu et al., 2021a). (2) Natural language expression insensitivity: As both the tag and the position of an object have been given in training, the model does not need to locate or distinguish objects by itself, not to say grounding natural language expressions on visual concepts. To alleviate the problems, we propose a novel pre-training task named *referring expression matching (REM)*. Given an image split into patches and an object tag, we convert the tag into a referring expression heuristically (e.g., "man on the right") and predict which patches contain the referred object. By using the synthetic referring expressions in training, the discrepancy has been reduced. Moreover, thanks to the promising language processing ability of PLMs, the obtained model generalizes well from the limited heuristically selected expressions to unseen ones in training, resulting in better downstream task performance.

In summary, our contributions are three-fold:

- A novel framework E2E-UVLP is proposed to perform end-to-end unsupervised vision-and-language pre-training without using costly and sub-optimal region features relied heavily by previous works.
- A referring expression matching pre-training task is proposed to reduce training-inference discrepancy and improve generalization to richer natural language expressions.
- Extensive experiments on four representative vision-and-language tasks show the superiority of our proposed framework over strong unsupervised baselines.

2 Method

2.1 Model Architecture

As shown in Figure 1, our proposed E2E-UVLP consists of a vision encoder and a pre-trained language model acting as a multimodal encoder. The vision encoder can be a vision pre-trained model such as ViT (Dosovitskiy et al., 2020) and Swin Transformer (Liu et al., 2021c). Each image I is encoded by the vision encoder into a sequence of patch features. These patch features are then linearly projected and added with corresponding modal-type embeddings to form the vision representations $V = \{\mathbf{v}_1, \ldots, \mathbf{v}_N\}$ where N is the number of patches. For text input L, it is tokenized into a sequence of word tokens. Each token's representation $\mathbf{t}_i$ is the sum of its word embeddings, its position or location embeddings (depending on the type of text input), and its modal-type embeddings. The resulting text representations $T = \{\mathbf{t}_1, \ldots, \mathbf{t}_M\}$ are then concatenated with V and fed into the multimodal encoder to get multimodal representations. We use BERT (Devlin et al., 2018) to initialize the multimodal encoder and word embeddings. Finally, the multimodal representations are used for different kinds of pre-training and fine-tuning tasks.

2.2 Pesudo-Parallel Data Synthesis

Since there is no parallel image-text data available in unsupervised vision-and-language pre-training (UVLP), it is important to find other ways

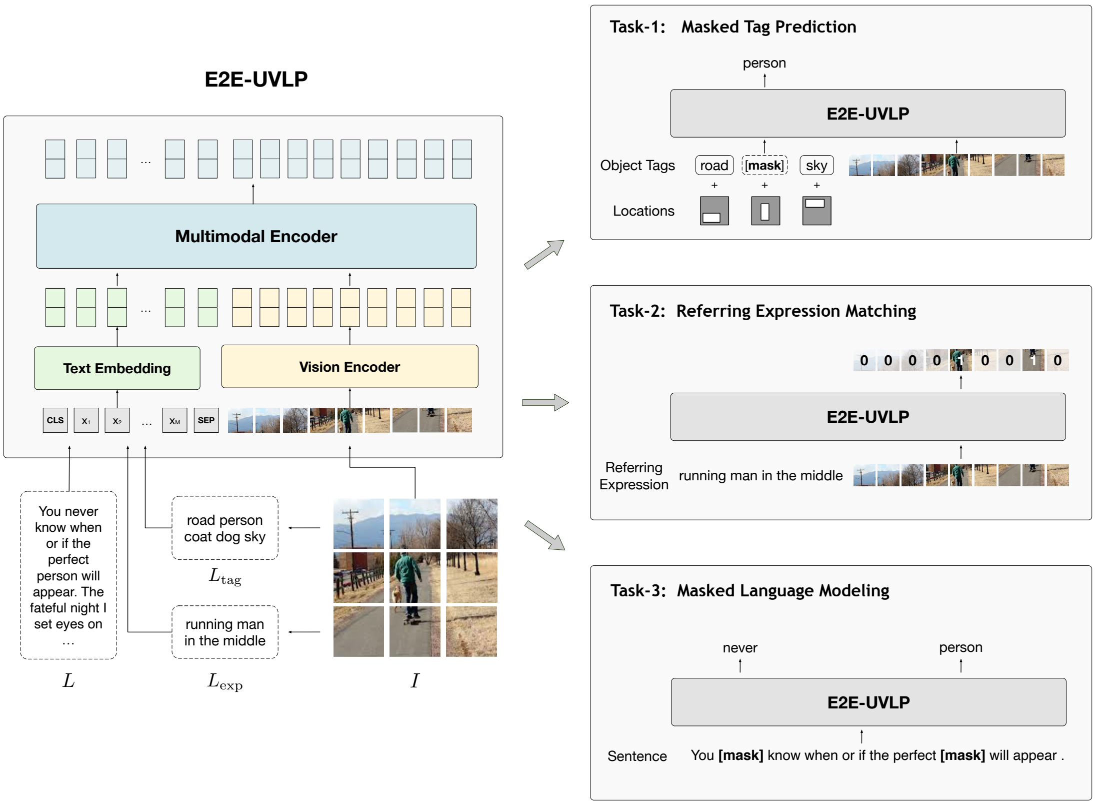

E2E-UVLP
Multimodal Encoder
Text Embedding
Vision Encoder
CLS
X1
X2
...
XM
SEP
You never know when or if the perfect person will appear. The fateful night I set eyes on …
road person coat dog sky
running man in the middle
L
L_{tag}
L_{exp}
I
Task-1: Masked Tag Prediction
person
E2E-UVLP
Object Tags
road
[mask]
sky
Locations
Task-2: Referring Expression Matching
0 0 0 0 1 0 0 1 0
E2E-UVLP
Referring Expression
running man in the middle
Task-3: Masked Language Modeling
never
person
E2E-UVLP
Sentence
You [mask] know when or if the perfect [mask] will appear .

Figure 1: Overview of our proposed E2E-UVLP framework. The model learns cross-modal representations from non-parallel text and image data in an end-to-end fashion (Section 2.1). To bridge the gap between the two modalities, we generate pesudo-parallel text L_{tag} and L_{exp} for each image I (Section 2.2). On the right side, we illustrate how to conduct pre-training tasks using different types of data (Section 2.3).

to bridge the gap between the two modalities. Inspired by Li et al. (2021b), we use detected object tags of each image to synthesize pesudo-parallel image-text data, which are used in the pre-training tasks for E2E-UVLP. Specifically, for each image I, we use an external object detector to generate its object proposals $\{(o_i, b_i)\}_{i=1}^{K}$, where o_i is the object tag and $b_i \in \mathcal{R}^4$ denotes the bounding box location. We derive two kinds of pesudo-parallel image-text data for each image and its corresponding object proposals.

Image-Tag Pair Because $\{o_i\}_{i=1}^{K}$ are essentially a bag of text words that describe the objects detected in the image, they can be viewed as text data weakly aligned with the image. We concatenate the object tags to form a text input $L_{\text{tag}} = o_1, \ldots, o_K$, and compute the location embeddings for o_i as a linear projection of b_i. Each token in the tokenized L_{tag} is embedded as the sum of its word embeddings and location embeddings. The reason for using location embeddings instead of position embeddings as for natural text is to distinguish between objects of the same category in the same image. We name this kind of synthetic data $(I, L_{\text{tag}}) \in \mathcal{D}_{\text{tag}}$ as *image-tag* paired data.

Image-Expression Pair The text input in image-tag paired data differs from natural text in two ways. First, it is composed of only noun words and does not conform to the grammar of natural language. Second, the location annotations do not exist in real text. We will discuss later the limitations of models trained on such data in Section 2.3. Here we describe another kind of synthetic pesudo-parallel data called *image-expression* paired data, which is more similar to real text. The idea is to generate a referring expression for one object in the image that can distinguish it from other objects (e.g., "smaller white sheep on the right" in Figure 2).

Specifically, for an image I and its detected object proposals $\{(o_i, b_i)\}_{i=1}^{K}$, we first use non-maximum suppression (NMS) to remove the redundant proposals with high overlap and filter out proposals with low prediction confidence. Then we randomly select one object (o_k, b_k) and find all proposals of the same category $\{(o_i, b_i)\}_{i=1 \ldots K}^{o_i = o_k}$. Based on these object proposals, we identify a

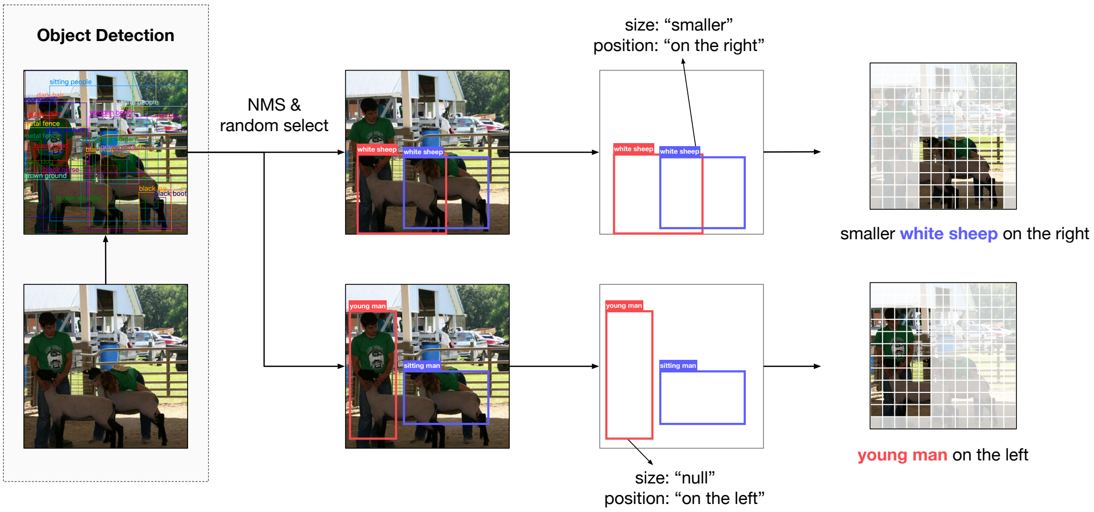


Figure 2: An illustration of the process of synthesizing referring expressions. We randomly select one target object and heuristically generate discriminative descriptions based on the bounding boxes of all objects of the same class.

group of cues that can discriminate the target object (o_k, b_k) from other objects and heuristically convert them into a referring expression. Inspired by Kazakos et al. (2021), we consider attributes generated from the object detectors for these objects, as well as the the relative size and position between the target object and other objects of the same object class. For the first example in Figure 2, as the area of the bounding box of the right object is smaller than that of the left one, we add a size description “smaller” to the expression. The final referring expression L_{exp} is the combination of size, attribute, object and position descriptions, guaranteed to refer to the target object in the image. L_{exp} and I together form the image-expression paired data $(I, L_{\text{exp}}) \in \mathcal{D}_{\text{exp}}$.

2.3 Pre-training Tasks

In this section, we introduce the pre-training tasks that enable our proposed E2E-UVLP to learn effective multimodal representations using only non-parallel image and text data *without region features*.

Masked Tag Prediction (MTP) This task aims to learn the alignment of the object concepts in two modalities using the image-tag paired data. Inspired by Li et al. (2021b), we randomly mask out the tags in L_{tag} and predict the masked tags conditioned on the raw image and other tags. Note that region features are conditioned on instead of the raw image in (Li et al., 2021b).

Specifically, the objective for MTP is computed as follows:

$$\mathcal{L}_{\text{MTP}} = -\mathbb{E}_{(I, L_{\text{tag}}) \sim \mathcal{D}_{\text{tag}}} \log P\left(T_{\mathbf{m}} \mid T_{\setminus \mathbf{m}}, V\right), \tag{1}$$

where $T_{\mathbf{m}}$ and $T_{\setminus \mathbf{m}}$ denote the masked tags and observed tags, respectively. For the masked tags, we keep the original bounding box locations and replace only the object tags with the special mask tokens. Different from previous works, we do not apply masked vision modeling (MVM) for the image-tag paired data, as it has been shown to cause performance degradation for end-to-end VLP (Dou et al., 2022). We also study object-guided masked vision modeling (Liu et al., 2021b), which aims to predict region features from grid image features, and observe no improvement compared to using MTP alone. We assume this is because L_{tag} already carries the object information, thus making such a task trivial.

Although MTP is effective, we find that the models trained with it are less effective in dealing with complex attributes of objects. For example in Figure 3, the model pre-trained with MTP provides an incorrect answer probably because it fails to focus on the correct patches of the object required in the question. We assume that this is due to two pitfalls of the MTP objective: (1) Discrepancy between training and inference. During pre-training, the text input L_{tag} is composed of objects with their tags and bounding box locations, unlike natural language sentences used in inference, which use positional encoding. In PLM studies, similar discrepancies have been shown to significantly impair performance (Brown et al., 2020; Liu et al.,

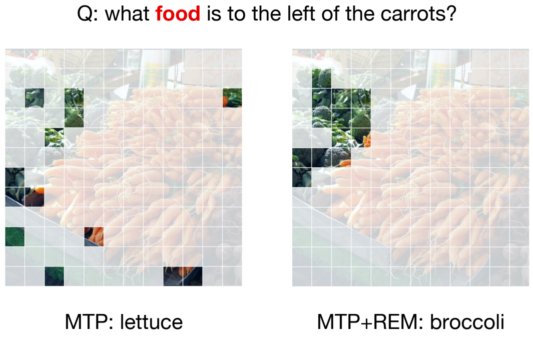


Figure 3: Comparison between VQA models fine-tuned from pre-trained models with different pre-training tasks. We visualize the most relevant patches for the keyword "food" in the given question. The model pre-trained with only MTP fails to identify the corresponding patches, which leads to a wrong answer.

2021a). (2) Natural language expression insensitivity. When trained with MTP, the model only needs to predict the tags of masked objects at given locations, rather than identifying target objects corresponding to natural language expressions from raw images. This problem is more pronounced in end-to-end UVLP because the images are encoded without using any object information.

Referring Expression Matching (REM) To alleviate the deficiencies of MTP, we design this novel pre-training task based on the image-expression paired data described in Section 2.2. The task of REM is to predict the position of the object referred to by the synthetic referring expression L_{exp}. We use the bounding box of the referred object as ground truth and convert it to a binary mask $R \in \{0, 1\}^N$ of the same size as the patch features, where the values corresponding to the inside of the bounding box are set to 1 and the others are set to 0. Then, given the model's prediction $\hat{R} = f(I, L_{\text{exp}}) \in [0, 1]^N$, we define the REM objective as

$$\mathcal{L}_{\text{REM}} = \mathbb{E}_{(I,L_{exp})\sim\mathcal{D}_{exp}}\text{DL}(R, \hat{R}) + \text{BCE}(R, \hat{R}), \quad (2)$$

where DL is the soft dice loss

$$\text{DL}(R, \hat{R}) = 1 - \frac{2\sum_{i=1}^{N} r^i \cdot \hat{r}^i}{\sum_{i=1}^{N} r^i + \sum_{i=1}^{N} \hat{r}^i}, \quad (3)$$

and BCE is the binary cross entropy loss

$$\text{BCE}(R, \hat{R}) = -\sum_{i=1}^{N} \big((1 - r^i)\log(1 - \hat{r}^i) + r^i \log(\hat{r}^i)\big) . \quad (4)$$

We use these two losses because they have been proven to be effective in image segmentation (Isensee et al., 2018), which aims to classify each pixel in an image into a certain object class.

We assume that using REM as a pre-training task can complement MTP in two ways. First, it will alleviate the discrepancy between training and inference as the model observes the text input in the form of natural language. Second, it explicitly enforces the model to localize the referred object from patch features, which strengthens the alignment between the learned visual concepts and the related linguistic expressions. As shown in Figure 3, the model pre-trained with MTP and REM successfully locates the corresponding image patches and gives the correct answer.

Masked Language Modeling (MLM) We also apply MLM on the text-only input to predict the masked tokens based on the surrounding text context. Given text input L from the text-only corpus $\mathcal{D}_L$, we formulate the MLM objective as

$$\mathcal{L}_{\text{MLM}} = -\mathbb{E}_{L\sim\mathcal{D}_L} \log P\left(T_{\mathbf{m}} \mid T_{\backslash\mathbf{m}}\right). \quad (5)$$

Note that no aligned images are observed in the computation of the MLM.

3 Experiments

3.1 Datasets

Following previous UVLP works (Li et al., 2021b; Zhou et al., 2022), we take images and captions from Conceptual Captions (CC) (Sharma et al., 2018) *without the alignment information* to construct the unsupervised image and text datasets. We also try a more realistic setting to use images from CC and sentences from BookCorpus (Zhu et al., 2015) where images and text are collected seperately from different domains. Similar to (Li et al., 2021b), we downsample the BookCorpus dataset to ensure the number of sentences in each training epoch is the same as the number of images.

3.2 Baselines

We compare our E2E-UVLP with both supervised and unsupervised vision-language pre-trained models. For supervised vision-language pre-trained models, we compare with models using different kinds of image features including **region features** (VisualBERT (Li et al., 2019), UNITER (Chen et al., 2020) and VinVL (Zhang et al., 2021)), **grid features** (E2E-VLP (Xu et al., 2021) and

Model	Visual Embed	VQA2	NLVR2	VE	Flickr30k		
		Test-Dev	Test-P	Test	R@1	R@5	R@10
Supervised (w/ Paired Image-Text Data)							
VisualBERT (Li et al., 2019)	Region	70.9	73.9	-	61.2	86.3	91.9
UNITER (Chen et al., 2020)	Region	72.7	77.9	78.3	72.5	92.4	96.1
VinVL (Zhang et al., 2021)	Region	76.0	**83.1**	-	-	-	-
E2E-VLP (Xu et al., 2021)	Grid	72.4	75.2	-	-	-	-
SOHO (Huang et al., 2021)	Grid	73.3	77.3	-	72.5	92.7	96.1
ViLT (Kim et al., 2021)	Patch	71.3	76.1	-	66.4	88.7	93.8
Visual Parsing (Xue et al., 2021)	Patch	74.0	78.1	-	73.5	93.1	96.4
ALBEF (Li et al., 2021a)	Patch	74.5	80.5	80.3	**82.8**	**96.7**	**98.4**
METER-CLIP-ViT$_{\text{BASE}}$ (Dou et al., 2022)	Patch	**77.7**	83.0	**81.2**	82.2	96.3	98.3
Unsupervised (w/o Paired Image-Text Data)							
U-VisualBERT (Li et al., 2021b)	Region	70.7	71.0	-	55.4	82.9	89.8
U-VisualBERT$_{\text{VinVL}}$ (Zhou et al., 2022)	Region	71.8	53.2	76.8	-	-	-
μ-VLA (Zhou et al., 2022)	Region	72.1	73.4	77.3	-	-	-
E2E-UVLP	Patch	**73.3**	**74.6**	**78.2**	**66.4**	**89.7**	**94.1**

Table 1: Evaluation results on four V+L downstream tasks. All unsupervised models are pre-trained on non-parallel images and text from CC. Our proposed E2E-UVLP outperforms previous UVLP methods, and achieves comparable performance to some supervised VLP models.

Method	Visual Embed	VQA2	NLVR2	VE	Flickr30k		
		Test-Dev	Test-P	Test	R@1	R@5	R@10
U-VisualBERT	Region	70.5	71.2	-	54.4	82.2	89.2
μ-VLA	Region	71.2	67.1	77.1	-	-	-
E2E-UVLP	Patch	**73.5**	**73.7**	**77.9**	**65.6**	**90.3**	**94.7**

Table 2: Experimental results of pre-training with images from CC and text from BookCorpus.

SOHO (Huang et al., 2021)) and **patch features** (ViLT (Kim et al., 2021), Visual Parsing (Xue et al., 2021), ALBEF (Li et al., 2021a) and METER-CLIP-ViT$_{\text{BASE}}$(Dou et al., 2022)). Note that in addition to the CC dataset we use for pre-training, these models typically use other parallel data sources such as MSCOCO (Lin et al., 2014), VG (Krishna et al., 2017) and SBU (Ordonez et al., 2011).

For unsupervised vision-language pre-trained models, we compare with U-VisualBERT (Li et al., 2021b), U-VisualBERT$_{\text{VinVL}}$ which is a version of U-VisualBERT with VinVL object features re-implemented by Zhou et al. (2022), and μ-VLA (Zhou et al., 2022). All of these models use region-based image features.

3.3 Implementation Details

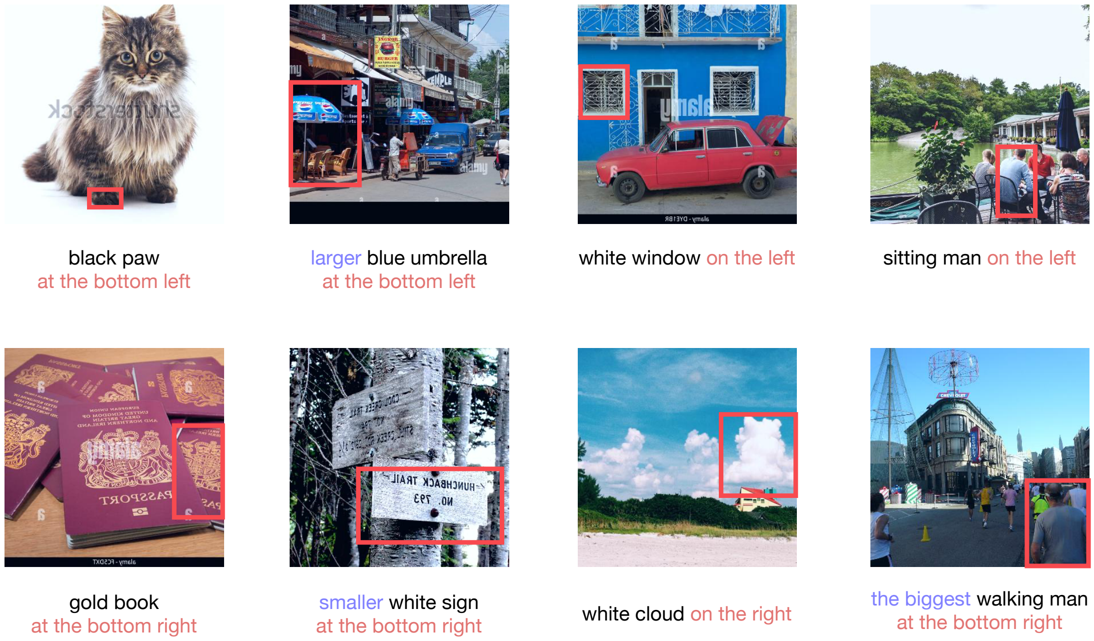
For the model architecture of E2E-UVLP, we use a 12-layer Swin-Transformer as the image encoder and a 12-layer Transformer acting as the multimodal encoder, which are initialized with pre-trained weights of Swin-B/32 and BERT-base, respectively. We utilize the widely-used object detector BUTD (Anderson et al., 2018) to extract object proposals for the images as in other region-based VLP methods. We resize each image to the size of of 384×384 with center-cropping for both pre-training and fine-tuning.

For the pre-training of E2E-UVLP, we set the total training iterations to 100k with a batch size of 512. We use AdamW with a peak learning rate of 3×10^{-5}. The learning rate is warmed-up to the peak value in the first 10% of the iterations, and then linearly decayed to 0. All the pre-training experiments are conducted on 16 NVIDIA V100s with 32GB memory per GPU.

We evaluate our model on four typical downstream tasks: Visual Question Answering (VQA) (Goyal et al., 2017), Natural Language for Visual Reasoning (NLVR2) (Suhr et al., 2018), Visual Entailment (VE) (Xie et al., 2019) and Image Retrieval on Flickr30k (Flickr30k) (Plummer et al., 2015). For details of the downstream tasks, please refer to the appendix.

3.4 Results

Table 1 shows the main results of E2E-UVLP on four downstream tasks. For each model, we list the type of image features used for pre-training. From the table we can see that E2E-UVLP con-



black paw at the bottom left

larger blue umbrella at the bottom left

white window on the left

sitting man on the left

gold book at the bottom right

smaller white sign at the bottom right

white cloud on the right

the biggest walking man at the bottom right

Figure 4: Examples of the synthetic image-expression paired data. We mark the objects referred to with red bounding boxes. The generated expressions are able to distinguish the target object from other objects in the image by heuristically adding size and position descriptions to the detected object tags and attributes.

sistently outperforms previous UVLP methods on all downstream tasks, which demonstrates that our end-to-end approach can learn a better cross-modal representation than the approaches using region features. When compared to supervised VLP models, our model achieves competitive results. Specifically, our model achieves a VQA score of 73.3% on the test-dev split, which is even higher than the performance of some supervised models. Finally, note that in the supervised VLP, the best model using region features (VinVL) performs better than the models using other types of image features, while in the unsupervised setting our approach using patch features outperforms the methods based on the same region features of VinVL. We attribute this to the use of the pre-training task that is more suitable for patch features in unsupervised VLP, i.e., REM (Section 2.3).

We also investigate pre-training using images from CC and text from BookCorpus, and the results are shown in Table 2. Previous works suggest that experimental results in this setting decline notably compared to pre-training with the in-domain CC captions (Li et al., 2021b; Zhou et al., 2022). In our experiments, however, we observe comparable or only slightly degraded performance on three of the four downstream tasks. On the VQA task, the model trained on BookCorpus is even slightly better than the model trained on CC by 0.2%. These results demonstrate that our model is robust to the sources of text and image data, which makes it more practical in realistic scenarios.

Pre-training Tasks	VQA2 Test-Dev	NLVR2 Test-P
None	70.1	51.2
MLM	69.9	50.3
MTP	71.7	67.4
REM	70.7	70.4
MTP + REM	72.8	72.8
MLM + MTP	72.6	74.1
MLM + REM	73.2	74.5
MTP + MLM + REM	**73.6**	**74.6**

Table 3: Ablation study of different pre-training tasks. All models are pre-trained with non-parallel images and text from MSCOCO.

3.5 Ablation Study

In this section, we conduct an ablation study on the pre-training tasks. To save experimental cost, we pre-train the models with non-parallel images and text from MSCOCO and only report results on the VQA and NLVR2 tasks. Table 3 shows the results. We use “None” to represent the results of training directly on the downstream tasks without pre-training. From the table, we can see that: (1) Cross-modal inputs (MTP/REM) and MLM are two key factors for the success of E2E-UVLP, as removing either of them will bring a significant performance degradation. Specifically, the combination of MTP and MLM can achieve decent results for E2E-UVLP, which illustrates the feasibility of end-to-end pre-training. (2) It is better to use both MTP and REM for pre-training than to use only one of them.. This verifies our assumption in Section 2.3 that REM can complement MTP and facilitate the model to learn better cross-modal representations. (3) Replacing MTP with REM can improve the performance on both of the downstream tasks, indicating that REM is possibly a more effective pre-training task for UVLP.

3.6 Visualization

In Figure 4, we provide some examples of the generated image-expression pairs as described in Section 2.2. As we can see, the generated referrinng expressions are able to distinguish the target object from other objects in the image by heuristically adding discriminative size and position descriptions. For example, in the second image, there are two umbrellas both with the attribute of the color blue. Since the target object has a larger bounding box, a size description "larger" is added. Similarly, by taking into account the relative positions of the two bounding boxes, a position description of "at the bottom left" is added. The resulting expression will be able to distinguish between the two objects.

4 Related Work

Vision-and-Language Pre-training Current research on visual-and-language pre-training (VLP) can be generally divided into two categories: the two-step training strategy and the end-to-end training strategy. Most works (Lu et al., 2019; Li et al., 2019; Chen et al., 2020; Zhang et al., 2021) fall into the first category where they first use external object detectors such as BUTD (Anderson et al., 2018) to extract region features for the images and then use them together with text embeddings to generate multimodal representations. However, the region features may be sub-optimal for VLP because they are designed for object detection tasks and are fixed during the pre-training process (Xu et al., 2021; Huang et al., 2021). Recently, some works integrate the encoding of images into the pre-training process, taking the raw images as input to learn the vision-and-language representations in an end-to-end fashion. These approach can be further categorized into the ones using grid features encoded with CNNs such as E2E-VLP (Xu et al., 2021) and SOHO (Huang et al., 2021), and the ones using patch features encoded with ViTs such as ViLT (Kim et al., 2021), Visual Parsing (Xue et al., 2021) and ALBEF (Li et al., 2021a). In this work, we apply a similar end-to-end approach to unsupervised visual-and-language pre-training with image patch features.

All of these works on VLP require access to large-scale parallel image-text datasets (Lin et al., 2014; Krishna et al., 2017; Sharma et al., 2018; Ordonez et al., 2011), which are difficult to collect due to the large amount of annotations or data cleaning efforts required. To alleviate this problem, some recent works explore unsupervised vision-and-language pre-training has emerged, in which only non-parallel image and text data are utilized. Our work also belongs to this category.

Unsupervised Vision-and-Language Pre-training Li et al. (2021b) first propose the idea of unsupervised vision-and-language pre-training (UVLP) without using paired image-text data. Their model, U-VisualBERT, is alternately pre-trained on both image-only and text-only data. In addition, they utilize object tags as anchor points for cross-modal alignment to compensate for the absence of aligned data and achieve similar performance to supervised models. Zhou et al. (2022) suggest that using tags alone is not sufficient and propose pre-training tasks for multi-granular alignment learning with a retrieved weakly aligned image-text corpus for UVLP.

The most important difference between our work and previous UVLP works is that we use an end-to-end training approach to implement UVLP, which is the first to the best of our knowledge. Besides, we identify the limitations of models trained with tags and propose a novel pre-training task, REM, to address these deficiencies. As a result, our approach significantly outperforms previous region-based UVLP works on all downstream tasks.

Referring Expression Comprehension and Generation The task of REM is inspired by the research lines of referring expression generation (REG) and comprehension (REC). REG is generally treat as a special case of image captioning to generate referring expressions from visual features with RNNs (Liu et al., 2017; Zarrieß and Schlangen, 2018), while Kazakos et al. (2021) generate synthetic referring expressions heuristically from object annotations. We apply a similar generation strategy but on the detected object proposals. The task of REC is to localize an object from candidate objects given a referring expression (Mao et al., 2016; Liu et al., 2019), while our proposed REM directly predicts the referred object from patch features without object candidates.

5 Conclusion

We propose a novel framework that performs end-to-end unsupervised vision-and-language pre-training without using costly and sub-optimal region features. To reduce the training-inference discrepancy, we propose a new pre-training task that predicts the locations of objects with synthetic referring expressions that are more similar to real text. Experiments show that our approach consistently outperforms existing unsupervised vision-and-language pre-training methods, and achieves competitive results compared to supervised vision-and-language pre-trained models.

Limitations

Although our approach eliminates the dependence on object detection during inference, it still requires object proposals for pre-training, which would damage the efficiency of pre-training. In addition, our method is limited by the finite number of object tags and the lack of diversity of heuristically generated referring expressions. We hope to address this limitation by jointly training a generator with unsupervised vision-language pre-training that automatically generates referring expressions or other type of psudo-parallel text for the images.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 61925601), the National Social Science Fund of China (No. 62236011), Beijing Academy of Artificial Intelligence (BAAI), and a grant from the Guoqiang Institute, Tsinghua University. We thank all anonymous reviewers for their valuable comments and suggestions on this work. We also thank Xinyu Ma for his help in the experiments.

References

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.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

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 *ECCV*.

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. *arXiv preprint arXiv:2010.11929*.

Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Pengchuan Zhang, Lu Yuan, Nanyun Peng, et al. 2022. An empirical study of training end-to-end vision-and-language transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18166–18176.

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 *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhicheng Huang, Zhaoyang Zeng, Yupan Huang, Bei Liu, Dongmei Fu, and Jianlong Fu. 2021. Seeing out of the box: End-to-end pre-training for vision-language representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12976–12985.

Fabian Isensee, Jens Petersen, Andre Klein, David Zimmerer, Paul F Jaeger, Simon Kohl, Jakob Wasserthal, Gregor Koehler, Tobias Norajitra, Sebastian Wirkert, et al. 2018. nnU-Net: Self-adapting framework for U-Net-based medical image segmentation. *arXiv preprint arXiv:1809.10486*.

Ioannis Kazakos, Carles Ventura, Miriam Bellver, Carina Silberer, and Xavier Giró-i Nieto. 2021. SynthRef: Generation of synthetic referring expressions for object segmentation. *arXiv preprint arXiv:2106.04403*.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. ViLT: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR.

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

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

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

Liunian Harold Li, Haoxuan You, Zhecan Wang, Alireza Zareian, Shih-Fu Chang, and Kai-Wei Chang. 2021b. Unsupervised vision-and-language pre-training without parallel images and captions. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5339–5350.

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

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Proceedings of ECCV 2020*, pages 121–137, Cham. Springer International Publishing.

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.

Jingyu Liu, Liang Wang, and Ming-Hsuan Yang. 2017. Referring expression generation and comprehension via attributes. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4856–4864.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021a. Pretrain, prompt, and predict: A systematic survey of prompting methods in natural language processing. *CoRR*, abs/2107.13586.

Xihui Liu, Zihao Wang, Jing Shao, Xiaogang Wang, and Hongsheng Li. 2019. Improving referring expression grounding with cross-modal attention-guided erasing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1950–1959.

Yongfei Liu, Chenfei Wu, Shao-yen Tseng, Vasudev Lal, Xuming He, and Nan Duan. 2021b. KD-VLP: Improving end-to-end vision-and-language pretraining with object knowledge distillation. *arXiv preprint arXiv:2109.10504*.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021c. Swin Transformer: Hierarchical vision Transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022.

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

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.

Vicente Ordonez, Girish Kulkarni, and Tamara Berg. 2011. Im2Text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems*, 24.

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, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8821–8831. PMLR.

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.

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*.

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

Haiyang Xu, Ming Yan, Chenliang Li, Bin Bi, Songfang Huang, Wenming Xiao, and Fei Huang. 2021. E2E-VLP: End-to-end vision-language pre-training enhanced by visual learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 503–513.

Hongwei Xue, Yupan Huang, Bei Liu, Houwen Peng, Jianlong Fu, Houqiang Li, and Jiebo Luo. 2021. Probing inter-modality: Visual parsing with self-attention for vision-and-language pre-training. *Advances in Neural Information Processing Systems*, 34:4514–4528.

Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. 2019. Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6281–6290.

Sina Zarrieß and David Schlangen. 2018. Decoding strategies for neural referring expression generation. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 503–512.

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

Mingyang Zhou, Licheng Yu, Amanpreet Singh, Mengjiao Wang, Zhou Yu, and Ning Zhang. 2022. Unsupervised vision-and-language pre-training via retrieval-based multi-granular alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16485–16494.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27.

A Details of Downstream Tasks

Most of our settings on downstream tasks follow the setup of ViLT (Kim et al., 2021) with small adjustments, as detailed below.

VQA The VQA task involves answering the question according to the given image, which requires an understanding of both vision and language. Following Kim et al. (2021), we fine-tune the model on the train and validation sets, and $1,000$ validation image-question pairs are reserved for internal validation. We use the $3,129$ most frequent answers as answer candidates following Yu et al. (2019). We set the batch size to 256 and the peak learning rate to 5×10^{-5}. The model is fine-tuned for 10 epochs.

NLVR2 The task of NLVR2 is to determine whether a natural language description is true given a pair of images. Following Kim et al. (2021), we reformulate the input to two image-caption pairs, concatenating the two representations as the input of a classification head. Following Li et al. (2021b), we perform task-specific pre-training before fine-tuning using mask-and-predict objective for 10 epochs. The batch size is 256 and the peak learning rate is 1×10^{-5}. During fine-tuning, we use a batch size of 128 and set the peak learning rate to 1×10^{-5}. The model is fine-tuned for 10 epochs.

VE The VE task is derived from Flickr30K (Plummer et al., 2015) images and Stanford Natural Language Inference (SNLI) (Bowman et al., 2015) dataset. Given an image premise P and text hypothesis H, the task aims to determine whether P implies H. This task is a 3-way classification problem to output *entailment*, *neutral*, or *contradiction* based on the relation inferred from the input image-text pair. The batch size is set as 256 and we set the peak learning rate as 7×10^{-5} to train for 5 epochs.

Image Retrieval Given a caption, the image retrieval task is to find the corresponding image from a collection of images. Following UNITER (Chen et al., 2020), we sample 31 negative image-text pairs along with a positive sample to construct a mini-batch for each GPU. The model is fine-tuned for 10 epochs with a peak learning rate of 5×10^{-5}.