Unsupervised Vision-and-Language Pre-training via Retrieval-based Multi-Granular Alignment

Mingyang Zhou[1*]   Licheng Yu[3*]   Amanpreet Singh[3]   Mengjiao Wang[3]   Zhou Yu[2]   Ning Zhang[3]

[1]Uiversity of California, Davis   [2] Columbia University

[3]Meta AI

minzhou@ucdavis.edu, zy2461@columbia.edu, {lichengyu, asg, mengjiaow, ningzhang}@fb.com

Abstract

Vision-and-Language (V+L) pre-training models have achieved tremendous success in recent years on various multi-modal benchmarks. However, the majority of existing models require pre-training on a large set of parallel image-text data, which is costly to collect, compared to image-only or text-only data. In this paper, we explore unsupervised Vision-and-Language pre-training (UVLP) to learn the cross-modal representation from non-parallel image and text datasets. We found two key factors that lead to good unsupervised V+L pre-training without parallel data: (i) joint image-and-text input (ii) overall image-text alignment (even for non-parallel data). Accordingly, we propose a novel unsupervised V+L pre-training curriculum for non-parallel texts and images. We first construct a weakly aligned image-text corpus via a retrieval-based approach, then apply a set of multi-granular alignment pre-training tasks, including region-to-tag, region-to-phrase, and image-to-sentence alignment, to bridge the gap between the two modalities. A comprehensive ablation study shows each granularity is helpful to learn a stronger pre-trained model. We adapt our pre-trained model to a set of V+L downstream tasks, including VQA, NLVR2, Visual Entailment, and RefCOCO+. Our model achieves the state-of-art performance in all these tasks under the unsupervised setting.

1. Introduction

Vision-and-Language pre-trained (VLP) models [8, 15, 16, 18, 23, 24, 27, 31, 32, 37, 39, 41, 43] that learn the joint cross-modal representation have revolutionized the research on various vision-and-language tasks in recent years. However, the success of VLP models relies on the availability of a large-scale aligned image-text corpora. The widely used crowd-sourced pre-training datasets such as MS COCO [7, 29] and Visual Genome [20] require expensive human
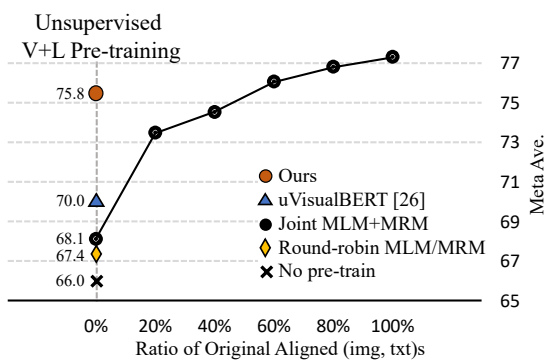
---

*The two authors contribute equally.

Figure 1. Meta average scores of VQA, NLVR2, VE, and RefCOCO+ fine-tuned from different pre-trained models. All pre-training are conducted on Conceptual Captions (CC) with different ratio of parallel data, i.e., a fixed amount of data is originally aligned while the rest is randomly shuffled. 0% refers to the case of unsupervised V+L pre-training. We also plot the performance of our proposed approach against U-VisualBERT [26]. Breakdown of the accuracy of each task is listed in the supplementary file.

annotations which are hard to scale up. Recently, the web crawled image-text datasets like Conceptual Captions 3M [35] and CC12M [5], and SBU Captions [33], *etc.* have dramatically reduced the need for massive human annotation but still require heavy post-cleaning procedures to get aligned image-text pairs. In comparison, the language corpora and image collection are readily available from the web. The convenience of getting a large-scale single-modality data has benefited the self-supervised learning of vision [3, 6, 12] and language [11, 30] domains respectively. This raises a question: Can we take advantage of easily-accessible large-scale single-modality data to perform unsupervised V+L pre-training without parallel text and images (UVLP)?

We define UVLP as follows: given the separately crawled image collection $\mathbf{I} = \{\mathbf{i}_1, \mathbf{i}_2, \ldots, \mathbf{i}_{n^I}\}$ and text corpus $\mathbf{T} = \{\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_{n^T}\}$, we aim to pre-train a multi-

modal model from such data. U-VisualBERT [26] is the first UVLP work, where the authors have trained their model on un-aligned text and image data in a round-robin fashion and simply use object tags as an anchor point to bridge the gap between the two modalities. Their research demonstrates that a shared multi-modal embedding can be learned by just presenting a single modality at a time. This however introduces an input discrepancy between pre-training and fine-tuning stages as each downstream V+L task requires both modalities (image, text) as the input. In this work, we investigate (i) whether presenting a joint image-text data from non-parallel data would improve the learned joint embedding space. Furthermore, (ii) if joint image-text data is fed into the model, how does its latent alignment affect the cross-modal representation learning?

To explore these two questions, we simply use the images and captions from Conceptual Captions (CC) dataset [35] as independently collected uni-modal corpus and perform the following analysis. First, we compare the pre-trained model's performance between the two data input strategies: one is presenting one image or text at a time (round-robin) and the other is presenting a concatenation of a pair of randomly sampled image and text (0% alignment ratio). Second, we prepare five sets of image-text pairs from Conceptual Captions with different levels of pairwise alignment by controlling the ratio of original aligned image-text data from 20% to 100% (while the remaining is randomly sampled from each modality). A single-stream transformer is used for all experiments with the standard pre-training objectives: masked language modeling (MLM) on language input and masked region modeling (MRM) on vision input. After pre-training, we adapt the model to a series of four downstream V+L tasks, including VQA [2], NLVR2 [38], Visual Entailment (VE) [42], and RefCOCO+ [44]. The performance is measured as the meta average of all tasks after fine-tuning. The results are summarized in Fig. 1. From Fig. 1, it is clear that joint MLM+MRM learning outperforms round-robin MLM/MRM. Such gains show that **joint image-and-text input is necessary for UVLP** even when the input is un-aligned. We also observe a strong positive correlation between the alignment of image-text pairs and the meta average of the fine-tuned downstream tasks of the resulting model. This conveys a seemingly intuitive but quite important message that **the more aligned the image-text data is the better the pre-trained model performs**.

Inspired by these analyses, we propose Unsupervised Vision-and-Language Pre-training via Retrieval-based Multi-Granular Alignment (μ-VLA), which uses our novel unsupervised V+L pre-training curriculum for non-parallel data. We first construct a weakly-aligned image-text dataset via retrieval. Given an image, we take its detected object tags as the reference sentence and retrieve the closest sentences from the text corpus via sentence BERT

embedding [34] similarity. Though the constructed pairs are noisy, the mere weak alignment of concepts is key to learning the latent alignment. We propose to let the model gradually learn a multi-granular alignment, i.e., region-to-object tag level, region-to-noun phrase level, and image-to-sentence level to more effectively bridge the gap between the two modalities. We show how each granularity learned from the weakly-aligned pairs contributes to the final pre-trained model's performance. Experiments show our approach achieves the state-of-art performance (in Fig. 1), with a clear gain over [26] on the 4 downstream tasks.

Towards practical applications, we also validate the effectiveness of our approach under a more realistic setting, where the images are from CC and the captions are from BookCorpus (BC) [49]. Similar performance gains are achieved in this harder setting, showing the robustness of our approach.

To summarize, our contributions are three-fold: (i) We analyze what leads to a good unsupervised V+L pre-training and found two key factors: joint image-and-text input, and overall alignment between image-text pairs. (ii) Accordingly, we propose a novel retrieval-based pre-training curriculum, which applies multi-granular alignment pre-training tasks between weakly aligned image-text pairs to bridge the gap between the two modalities. (iii) We provide comprehensive experiments and analyses showing the robustness of our approach when compared to SOTA supervised and unsupervised V+L pre-training methods.

2. Related Work

Vision-and-Language Pre-training Inspired by the success of natural language processing [4, 11], there is a recent surge of interest in pre-training for vision and language. For example, there are different architectures (*e.g.* two-stream models [17, 24, 31, 32, 39, 43] vs. single-stream models [8, 23, 25, 27, 37]), features (*e.g.* regions [1] *vs.* grids [16]), backbones (*e.g.* ConvNets [16] *vs.* Transformers [18]) *etc*. All these works aim to exploit the large-scale aligned image-text corpora [17, 20, 29, 33, 35] to pre-train a powerful multi-modal model, which is then adapted to various downstream V+L tasks, such as VQA [2], NLVR2 [38], Visual Entailment (VE) [42], Referring Expression Comprehension [44], and Image-Text Retrieval.

Various pre-training tasks have been introduced to achieve this, including the most notable Masked Language Modeling (MLM), Masked Region Modeling (MRM), and Image-Text Matching (ITM). Several other variants have also been explored, such as predicting the object tags [14, 28], sequence generation [41, 47], word-region alignment [8]. In this paper, we propose learning a multi-granular alignment between word and region, phrase and region, and image and sentence to better bridge the gap between vision and language.
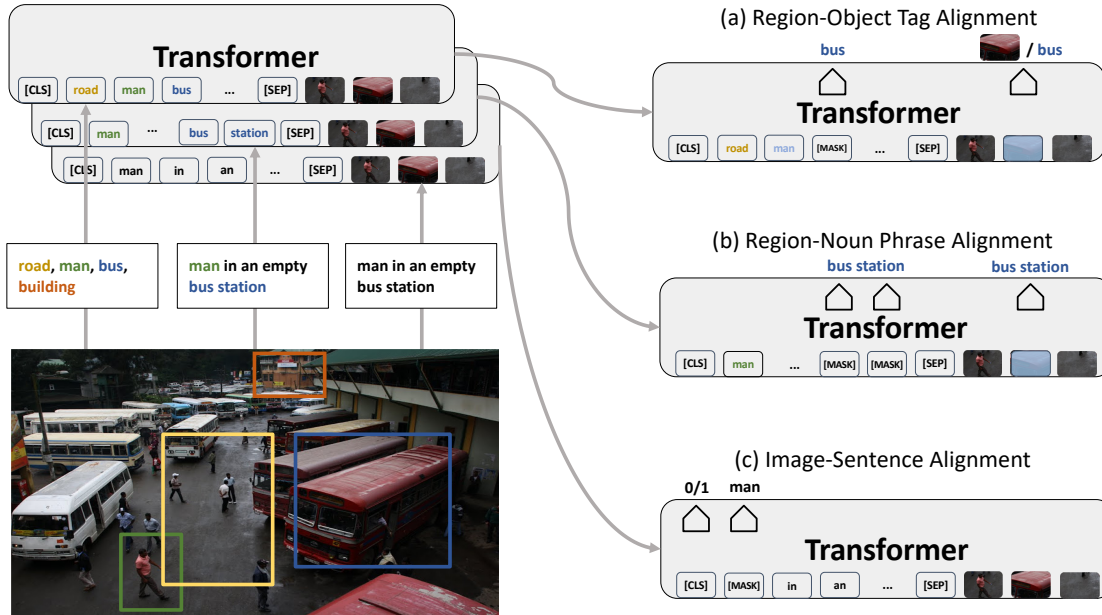
Figure 2. Overview of our method. On the left we form three types of image-text pairs as input data to learn cross-modal alignment on three different granularities: region-tag alignment, region-phrase alignment, and image-text alignment. The models is iteratively pre-trained on each granularity and the model parameters are shared. On the right-hand side, we demonstrate the details of the pre-training objectives for each granularity.

Unsupervised Vision-and-Language Pre-training without Parallel Data Inspired by the works on multi-lingual contextual language modeling [9, 10, 21, 22], U-VisualBERT [26] first propose the *unsupervised* vision-and-language pre-training without parallel data (UVLP). U-VisualBERT [26] conducts the masked prediction on the text-only and image-only corpora and introduce the object tags as anchor points to bridge the two modalities. The authors treat the tags as a sentence when performing MLM, where tags provide alignment with the regions in a picture and implicitly learn a tag-region-level alignment. However, the anchor tags are still quite different from the text input, missing the sentence completeness and naturalness. Besides, the latent cross-modal alignment is shown to be important in our analysis (from Fig. 1). As comparison, our pre-training involves a retrieval-based weakly aligned V+L data construction and learns a more comprehensive multi-granular cross-modal alignment. With same data as U-VisualBERT, our approach achieves a clear and consistent gain across all the downstream tasks in our experiments.

3. Method

In this section, we introduce the two core components of our μ-VLA's architecture for unsupervised V+L pre-training without parallel data: (1) construct a weakly aligned image-text corpus from independent vision and lan-

guage data sources; (2) our novel pre-training curriculum to enable the model to capture the cross-modal alignment on three granularity including region-to-tag level alignment (RT), region-to-noun phrase level alignment (RN), and image-to-sentence level alignment (IS).

3.1. Model Overview

We use the well-known single-stream model architecture for our experiments as [8,23,25,27,37]. As shown in Fig. 2, our main backbone is a single transformer, where we feed the concatenation of visual embeddings of an image and the tokens of a caption as its input. Given an image $\mathbf{i}$, we first use an off-the-shelf Faster R-CNN (VinVL [45]) to detect the objects $\mathbf{v} = \{v_1, ..., v_{k^v}\}$. The visual embedding of each region is then encoded as the sum of its regional feature, its location embedding[1], and the modality embedding. For a given caption $\mathbf{t}$, we denote its tokenized sequence as $\mathbf{t} = \{t_1, ..., t_{k^t}\}$. After multiple layers of self-attention, the two modalities are fused together and the output hidden vectors can be used for various pre-training tasks.

---

[1]The 5-dimensional vector $[\frac{x_1}{W}, \frac{y_1}{H}, \frac{x_2}{W}, \frac{y_2}{H}, \frac{(y_2-y_1)(x_2-x_1)}{W \cdot H}]$ is projected to the visual embedding space. $(x_1, y_1), (x_2, y_2)$ are the coordinates of the top left and bottom right point of the detected region, and W, H are the image width and height.

3.2. Weakly-aligned Image-Text Corpus

As in the analysis of Sec 1, we observe a strong correlation between the degree of image-text alignment in the training data and the performance of the pre-trained model. Inspired by this finding, we believe it important to initialize some weak semantic alignment between the two modalities as the input source. Specifically, we retrieve k sentences that are semantically closed to a given I_i. Previous work [40] shows the visually grounded caption covers a good ratio of words that are naturally related to specific visual contents, *e.g.* concrete nouns. Thus, we utilize the semantic association between the objects that appear in the image and a candidate sentence as the indicator to measure the alignment degree.

Specifically, we take the object tags $\mathbf{o} = \{o_1, ..., o_{k^o}\}$ from the above detected $\mathbf{v}$ and feed the sequence into an off-the-shelf sentence BERT embedding model [34] to obtain the query embedding $\mathbf{e_o}$. Similarly, we feed each candidate sentence into the same model getting the candidate embedding $\mathbf{e_t}$. We retrieve the top K candidates with the highest cosine similarity score to form an initial weakly-aligned image-text pairs for a given image $\mathbf{i}$. We denote the retrieved captions as $\{\mathbf{t}^r(\mathbf{i})\}_{r=1}^K$ and the overall weakly aligned corpus as $\mathbf{R}$.

3.3. Pre-training Tasks

In this subsection, we introduce a set of pre-training objectives that we designed to facilitate the model to capture the different levels of vision and language alignment. Fig. 2 shows the overview of our model and its pre-training tasks.

3.3.1 Region-Tag Alignment Learning

We first propose to align the object tags onto the image regions. As shown in Fig. 2(a), We concatenate the object tags detected from each image with its source image to form an input pair $[\mathbf{o}, \mathbf{v}]$ fed into the model. We denote the mask indices as $\mathbf{m} \in \mathbb{N}^{M\,2}$. We randomly mask out the object tags and regions, and apply masked language modeling (MLM) and masked region modeling (MRM) for the pre-training.

Specifically, MLM on the object tags is formulated as

$$\mathcal{L}_{\text{MLM}}^{\text{R-T}} = -\mathbb{E}_{(\mathbf{o},\mathbf{v})\sim\mathbf{I}} \log P(\mathbf{o_m}|\mathbf{o_{\backslash m}}, \mathbf{v}),$$

where the goal is to predict the masked object tags based on the observation of their surrounding tags $\mathbf{o_{\backslash m}}$ and image regions $\mathbf{v}$. On the vision side, MRM includes both masked region classification loss (MRC) and masked region feature regression loss (MRFR):

$$\mathcal{L}_{\text{MRM}}^{\text{R-T}} = \mathbb{E}_{(\mathbf{o},\mathbf{v})\sim\mathbf{I}}[f_{\text{MRC}}(\mathbf{v_m}|\mathbf{v_{\backslash m}}, \mathbf{o}) + f_{\text{MRFR}}(\mathbf{v_m}|\mathbf{v_{\backslash m}}, \mathbf{o})].$$

---

[2]$\mathbb{N}$ is the natural numbers, M is the vocabulary size, and $\mathbf{m}$ is the set of masked indices.

Between the two, MRC learns to predict the object semantic class for each masked region $c(\mathbf{v_m})$. We feed the last hidden output of the masked region $\mathbf{v_m}$ into a FC layer and softmax function to predict the classification probabilities $g_\theta(\mathbf{v_m})$. The objective is to minimize the cross-entropy of $f_{\text{MRC}}(\mathbf{v_m}|\mathbf{v_{\backslash m}}, \mathbf{o}) = \text{CE}(c(\mathbf{v_m}), g_\theta(\mathbf{v_m}))$. MRFR learns to regress the transformer output of each masked region $\mathbf{v_m}$ to its visual features. We apply a FC layer to convert its hidden output to a vector $h_\theta(\mathbf{v_m})$ of the same dimension as the input regional feature $r(\mathbf{v_m})$. We apply L2 regression: $f_{\text{MRFR}}(\mathbf{v_m}|\mathbf{v_{\backslash m}}, \mathbf{o}) = ||h_\theta(\mathbf{v_m}) - r(\mathbf{v_m})||_2^2$.

For region-tag alignment learning, we have our pretraining objective function as

$$\mathcal{L}^{\text{R-T}} = \mathcal{L}_{\text{MLM}}^{\text{R-T}} + \mathcal{L}_{\text{MRM}}^{\text{R-T}}$$

3.3.2 Region-Noun Phrase Alignment Learning

Due to the small vocabulary size of object tags, the region-tag alignment learning can only capture a limited amount of localized concepts. To increase the diversity of concepts, we propose to align the noun phrases from the retrieved sentences to the corresponding regions as well. As in Fig. 2(b), given an image $\mathbf{i}$ and its retrieved weakly aligned caption $\mathbf{t}^r(\mathbf{i})$, we first detect the noun phrases from the caption using spacy [13]. Note the detected noun phrases sometimes contain the attribute words, which further benefits this pretraining task. We link the noun phrase to its closest visual region by computing the word2vec similarity between the phrase and object tag (associated to each region). The pretraining still consists of MLM and MRM but are performed with different masking strategy and supervision signal.

Specifically, for both MRM and MLM, we only mask the linked noun phrases from the caption or the linked object regions. We make the masking probability proportional to the linked similarity score. Each time we only mask out one modality (phrase or region) to encourage it to be recovered by its linked content. The region-to-phrase MLM is then formulated as $\mathcal{L}_{\text{MLM}}^{\text{R-P}} = -\mathbb{E}_{(\mathbf{v},\mathbf{t}^r)\sim\mathbf{R}} \log P(\mathbf{t_m^r}|\mathbf{t_{\backslash m}^r}, \mathbf{v})$.

On the vision side, we propose using the phrase-guided masked region-to-token classification (p-MRTC) on the masked regions:

$$\mathcal{L}_{\text{MRM}}^{\text{R-P}} = \mathbb{E}_{(\mathbf{v},\mathbf{t}^r)\sim\mathbf{R}} f_{\text{p-MRTC}}(\mathbf{v_m}|\mathbf{v_{\backslash m}}, \mathbf{t}^r),$$

where we directly classify the masked region to its linked noun phrase (sub-word tokens) in BERT vocabulary. Enlarging the vocabulary has shown to be beneficial to MRM [48]. Our proposed p-MRTC leverages the additional noun-phrase to encourage more diverse local region to language alignment.

For region-noun phrase alignment learning, we have our pretraining objective function as

$$\mathcal{L}^{\text{R-P}} = \mathcal{L}_{\text{MLM}}^{\text{R-P}} + \mathcal{L}_{\text{MRM}}^{\text{R-P}}$$

3.3.3 Image-Sentence Alignment Learning

We apply image-text matching (ITM) objective as the previous supervised V+L pre-training research [8, 23] to learn the cross-modal sentence-level alignment. As in Fig. 2(c), given an input pair $[\mathbf{v}, \mathbf{t}^r]$, the final hidden vector of the special token [CLS] is fed through a FC layer to output a single score $\mathbf{s}_\theta(\mathbf{v}, \mathbf{t}^r)$, which predicts if the given image-text input is a semantically matched pair or not. We use the label $y \in \{0, 1\}$ to indicate if a retrieved pair is a match. The training objective for the ITM task is to minimize the binary cross-entropy loss: $\mathcal{L}_{\text{ITM}} = \text{CE}(y, s_\theta(\mathbf{v}, \mathbf{t}^r))$. On the language side, we also apply standard MLM to help the model learn to align other language tokens besides noun phrases and object tags to the visual context. The objective function is then formulated as $\mathcal{L}_{\text{MLM}}^{\text{I-S}} = -\mathbb{E}_{(\mathbf{v}, \mathbf{t}^r) \sim \mathbf{R}} \log P(\mathbf{t}_{\mathbf{m}}^r | \mathbf{t}_{\backslash \mathbf{m}}^r, \mathbf{v})$. The image-sentence level alignment pretraining objective function is

$$\mathcal{L}^{\text{I-S}} = \mathcal{L}_{\text{MLM}}^{\text{I-S}} + \mathcal{L}_{\text{ITM}}$$

3.4. Multi-Granular Pre-training Curriculum

We propose a multi-granular curriculum to iteratively pre-train the model on the region-to-tag, region-to-noun phrase, and image-to-sentence level. According to our findings in Sec. 1, learning from image-text pairs with higher degree of cross-modal alignment is beneficial to the performance of unsupervised V+L pre-trained model. Therefore, we propose using an estimated image-text alignment score to guide our multi-granular pre-training. Specifically, we have an ITM header defined in Sec. 3.3.3 to learn the image-text alignment. We also use it to predict matching score as a weight to modulate the input data for each of our retrieval-based pre-training tasks. This allows us to provide more importance to relatively more aligned image-text pairs over time to help our model to learn better cross-modal alignment on multiple granularities.

To train the alignment model's ITM classifier, we use our retrieved corpus $\mathbf{R}$ as positive samples and randomly shuffled pairs as negative samples in the first m epochs. This warms up the models to make reasonable estimations on the alignment of image-text input pairs. After m epochs, we start to incorporate the alignment prediction score w_{ITM} in our training objective. To summarize, our multi-granular pre-training loss is

$$\mathcal{L} = \begin{cases} \mathcal{L}^{\text{R-T}} + \mathcal{L}^{\text{R-P}} + \mathcal{L}^{\text{I-S}} & \text{if epoch} < m \\ \mathcal{L}^{\text{R-T}} + w_{\text{ITM}}(\mathcal{L}^{\text{R-P}} + \mathcal{L}^{\text{I-S}}) & \text{if epoch} \geq m, \end{cases}$$

where $\mathcal{L}^{\text{R-T}}$, $\mathcal{L}^{\text{R-P}}$, and $\mathcal{L}^{\text{I-S}}$ are the loss functions for region-tag alignment pre-training, region-noun phrase alignment pre-training, and image-sentense alignment pre-training. We set m as 1 in our final implementation.

4. Experiments

In this section, we provide the detailed experimental set up to evaluate our proposed μ-VLA against previous supervised and unsupervised VLP models. More specifically, we introduce our pre-training dataset, baselines, and our pre-training setting.

4.1. Pre-training Datasets

We prepare the un-aligned data under two different settings: (1) We use images and text separately from Conceptual Captions (CC) [35] ignoring the alignment information; (2) We use images from Conceptual Captions (CC) [35] and text from BookCorpus (BC) [49]. Setting (1) sets up a fair comparison with previous supervised methods by keeping the domain and the quality of training data consistent. Our proposed model trained in this setting is called μ-VLA$_{CC}$. Setting (2) mimics a more realistic challenge where we have large-scale images and text data from different domains, in particular the text sources are not similar to captions of the images. μ-VLA$_{BC}$ has been trained in this setting.

As introduced in section 3.2, for each image we retrieve 5 text data points (captions from CC or sentences from BC) from the text corpus that are semantically similar to the detected objects in the image. This creates weakly-aligned image-text pairs for our pre-training models.

4.2. Baselines

We compare the performance of our proposed μ-VLA to the following baselines:
Base Model VisualBERT that is initialized from BERT. It does not undergo any pre-training but is directly fine-tuned on the downstream tasks.
Supervised Pre-trained Models Supervised pre-trained VLP models that are trained only on CC, including VIL-BERT [31], VL-BERT [37], and UNITER [8]. We also report the numbers on the Supervised VisualBERT implemented in U-VisualBERT [26] that is trained on CC and an additional 2.5 Million text segments from BC. For fair comparison with our proposed method, we also introduce the aligned vision-language pre-training model (Aligned VLP) that is pre-trained on the 3M (image, caption) pairs from CC and 3M (image, object tag) pairs.
Unsupervised Pre-trained Models U-VisualBERT is pre-trained on individual image or text corpus in a round-robin fashion and captures the cross-modal alignment by using detected object tags as the anchor point. For fair comparison, we re-implemented this method to pre-train with the VinVL object features [46] and BC.

4.3. Training Setup

Our transformer architecture consists of 12 layers of transformer blocks, where each block has 768 hidden units

Model	VQA2 Test-Dev	NLVR2 Test-P	VE Test	RefCOCO+			Meta-Ave
				Dev	TestA	TestB	
ViLBERT [31]	70.6	-	-	72.3	78.5	62.6	-
VL-BERT [37]	71.2	-	-	71.6	77.7	61.0	-
UNITER$_{CC}$ [8]	71.2	-	-	72.5	79.4	63.7	-
VisualBERT [25, 26]	70.9	73.9	-	73.7	79.5	64.5	-
Aligned VLP	**72.5**	**75.9**	**78.7**	**82.1**	**86.6**	**75.0**	**77.3**
Base	70.1	51.2	73.2	69.4	74.8	60.3	65.9
U-VisualBERT [26]	71.8	53.2	76.8	78.2	83.6	69.9	70.0
μ-VLA$_{CC}$	**72.1**	**73.4**	**77.3**	**80.3**	**85.5**	**73.7**	**75.8**
μ-VLA$_{BC}$	71.2	67.1	77.1	79.7	85.0	72.7	73.8

Table 1. Evaluation results on four V+L downstream tasks. Our model trained with un-aligned data (μ-VLA$_{CC}$, μ-VLA$_{BC}$) achieves comparable performance with the supervised model trained with aligned data (Aligned VLP). μ-VLA$_{CC}$ and μ-VLA$_{BC}$ also outperform U-VisualBERT on nearly all tasks.

V+L Alignment	VQA Test-Dev	NLVR2 Test-P	VE Test	RefCOCO+			Meta-Ave
				Dev	TestA	TestB	
μ-VLA$_{CC}$ (R-T)	71.7	52.0	75.6	78.7	83.3	70.0	69.5
μ-VLA$_{CC}$ (R-N)	71.4	69.4	76.5	77.4	81.5	68.7	73.7
μ-VLA$_{CC}$ (I-S)	71.6	71.5	76.8	75.7	80.3	67.9	73.9
μ-VLA$_{CC}$ (R-T + R-N)	71.9	72.4	76.4	79.3	84.5	71.7	75.0
μ-VLA$_{CC}$ (R-T + R-N + I-S)	**72.1**	**73.4**	**77.3**	**80.3**	**85.0**	**73.7**	**75.8**

Table 2. Effect of cross-modal alignment on the three types of granularities: region-tag alignment(R-T), region-noun phrase alignment(R-N), and image-sentence alignment(I-S)

and 12 self-attention heads. We initialize the model from BERT$_{base}$ and pre-train for 20 epochs on their respective pre-training datasets with a batch size of 480. The region features for images are obtained from the pre-trained VinVL object detectors [46]. We use Adam optimizer [19] with a linear warm-up for the first 10% of training steps, and set the peak learning rate as 6e-5. After warm up, a linear-decayed learning-rate scheduler gradually drops the learning rate for the rest of training steps. All models were trained on 4 NVIDIA A100 GPUs, with 40GB of memory per GPU using MMF [36]. The pre-training takes 3 days. We evaluate our pre-trained models on four downstream tasks: Visual Question Answering (VQA 2.0) [1], Natural Language for Visual reasoning [38] (NLVR^2), Visual Entailment [42] (VE), and Referring Expression [44] (RefCOCO+). Detailed training settings for each task can be found in our supplementary material.

4.4. Experimental Results

We first compare μ-VLA to various supervised models that are pre-trained on CC and to the state-of-the-art unsupervised V+L pre-training method, U-VisualBERT on the four downstream tasks. Besides reporting scores for each individual task, we also compute the meta-average score to reflect the overall performance across all tasks. The results

are summarized in Table 1.

Compared to Base. It is clear from Table 1 that both μ-VLA$_{CC}$ and μ-VLA$_{BC}$ outperform the Base model by a large margin on all benchmarks.

Compared to Aligned VLP. It also achieves better performance than existing supervised models like VilBERT [31], which is potentially due to the usage of better visual regional features of VinVL [45]. When compared to Aligned VLP, which is trained with the same architecture and visual features, our model is only slightly worse. This shows the effectiveness of our proposed pre-training curriculum which can learn comparable universal representation across vision and language as the supervised models without any parallel image-text corpus.

Compared to UVLP. Our μ-VLA also achieves consistently better performance than the previous UVLP method: U-VisualBERT. This improvement shows how our proposed cross-modal alignment pre-training curriculum effectively bridges the gap across the two modalities. In particular, our model outperforms U-VisualBERT in the task of NLVR2 by more than 20%. As NLVR2 is known to benefit more from image-sentence cross-modal alignment from previous supervised V+L pre-training research [8], this observation indicates that our model is able to capture the instance-level cross-modal alignment without parallel data.

When μ-VLA is trained on BC text and CC images *i.e.* μ-VLA$_{BC}$, it still achieves comparable or better performance than U-VisualBERT except for VQA. The slight advantage U-VisualBERT has over μ-VLA$_{BC}$ in VQA is potentially due the similar style between the VQA text and the pre-trained CC captions. However, this does not overshadow the overall better performance of μ-VLA. It shows that our proposed method is more robust than U-VisualBERT training on the uni-modal datasets collected from separate domains, which makes it more useful in practical settings.
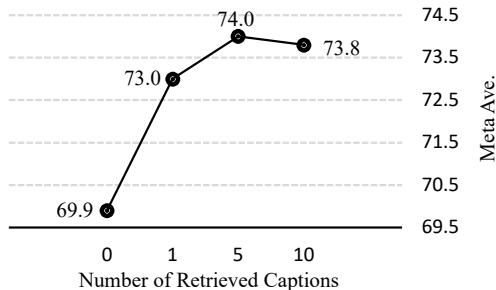
Figure 3. Meta average scores of non-parallel V+L pre-training with different number of retrieved candidate sentences.

4.4.1 Ablation Study on Multi-Granular Alignment

We conduct ablation study to verify the effectiveness of the three types of visual-language alignment for unsupervised V+L pre-training, namely region-tag alignment (R-T), region-noun Phrase alignment (R-N), and image-sentence alignment (I-S). We first evaluate each individual type of alignment to measure its usefulness for different downstream tasks. Then, we gradually add each type of alignment into the UVLP. For this ablation study we pre-train μ-VLA on CC images and text, and the results are summarized in Table 2.

From Table 2, we can see that aligning local regions to object tags (R-T) and noun phrases (R-N) are especially helpful for the task of RefCOCO+, which requires the model to understand specific objects that natural expressions describe. Meanwhile, aligning the image and sentence at instance-level (I-S) benefits NLVR2 and VE. Especially on NLVR2, the model that captures the global vision and language alignment μ-VLA$_{CC}$ (I-S) obtains 19.5% gain over the model that only learns the local alignments between regions and object tags μ-VLA$_{CC}$ (R-T). This observation is consistent with previous research [8], where the performance of model on NLVR2 is boosted after introducing pre-training objectives that capture the cross-modal alignment in the image-text pairs. Our results demonstrate that even with just weakly-aligned sentences, we can still

effectively learn the instance-level cross-modal alignment. Combining the region-tag and region-noun phrase alignment (R-T+R-N) for UVLP, we observe that these two types of grounding and matching compensate each other. μ-VLA$_{CC}$ (R-T+R-N) shows a marginal but consistent improvement over models that only learn a single type of local region-to-language alignment (R-T, R-N). After adding object-phrase level alignment we can further improve the performance on NLVR2 and VE, which gives us our best performing model μ-VLA$_{CC}$ (R-T + R-N + I-S).

4.4.2 Ablation Study on Number of Retrieved Candidates

We conduct experiments to verify the impact the number of retrieved candidate text for each image has on the performance. We create three variants of pre-training corpus, where the number of retrieved candidate are 1, 5, and 10 based on the rank of the similarity of each candidate text to the query image's detected object tags. The candidate text is sampled from CC. We pre-train our μ-VLA model with only the pre-training objectives to capture the sentence-image alignment (I-S). For each variant of pre-training corpus, we train the model for the same number of steps. We compute the meta average score for the three resulting pre-trained models and visualize them in Fig. 3.

Fig. 3 shows that retrieving more than one candidate text for an image greatly benefits the pre-trained model to learn a better joint representation between vision and language, demonstrated by stronger performance in the downstream tasks. We suspect this is because the closeness between the candidate caption and the detected object tags in language embedding space does not always mean a better alignment between the candidate caption and the image. A better and more semantically similar caption candidate for the image could be found in the other caption candidates. However, when we increase the number of candidate captions to 10, we observe a slight drop on the overall performance compared to the model that is pre-trained on corpus with 5 candidate captions. This indicates that having too many candidate captions to form the weakly-aligned pairs with the query image for V+L pre-training may also introduce unnecessary noise. Hence, we set the number of retrieved captions in our experiments to 5.

4.4.3 Visualization

To get a sense of the quality of the retrieved sentences, we show some examples of retrieved text from both CC and BC in Fig. 4. The first row demonstrates a positive case of retrieved captions from CC, where we observe a good coverage of the objects in the image such as "young woman", "sofa", and "couch" in the top retrieved sentences. Similarly, our retrieval method can also retrieve good candidates
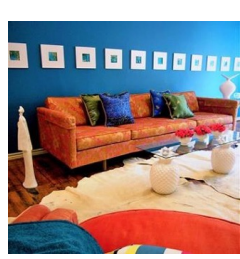
Figure 4. Examples of retrieved text from both CC and BC. The covered grounded noun phrases in retrieved sentences are highlighted in green bar for positive examples.

from BC that describe many visual objects from the image as depicted in row 2. This observation demonstrate the effectiveness of picking candidates based on their closeness to the object list in the language embedding space.

We also compare the text-to-image attention between the pre-trained U-VisualBERT and μ-VLA without task-specific fine-tuning as [8, 48]. As shown in Fig. 5, we feed into the models an aligned pair whose caption is "young woman seated on the beach", we visualize the local cross-modality alignment between regions and tokens. we found our full model μ-VLA can better attend on the described regions, showing higher-quality alignment is learned through the proposed pre-training. More visualizations are in the supplementary file.

5. Conclusion

We propose an unsupervised vision-and-language pre-training approach via retrieval-based multi-granular alignment to learn strong vision and language joint representations with un-aligned text and image sources. We introduce two core designs of our proposed approach: (1) construct a retrieval-based weakly-aligned image-text corpus. (2) multi-granular pre-training objectives to enable the model to capture the cross-modal alignment at different granularity levels. Our experiments show that our model can consistently outperform the previous state-of-the-art unsupervised pre-trained models and can achieve similar performance as the fully-aligned pre-trained models.

Limitations: As we only consider the detected object list to retrieve the candidate sentences, the retrieved sentences often do not cover other visually grounded information compared to the ground truth captions. Besides, the detected object tags are often those general concepts lacking
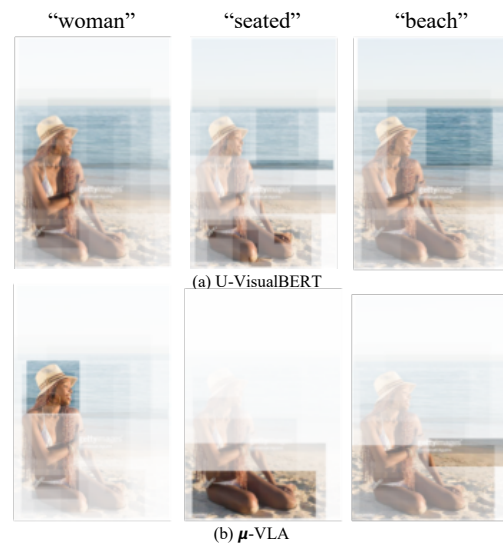

Figure 5. Text-to-image attention given the aligned pair whose caption is "young woman seated on the beach".

diversity. Our retrieval results and in turn our pre-trained models could be affected by such limitations. We hope to address the issue by learning a Siamese network between visual concepts and sentence for better retrieval and exploiting even larger uni-modal datasets to increase the diversity in the future research.

Societal Impact: The models are trained on the public datasets widely used in the community. However, these datasets are known with biases, which may in turn affect our model predictions. We do not recommend relying on the models to make real-world decisions.

References

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, 2018. 2, 6

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *ICCV*, 2015. 2

[3] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021. 1

[4] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020. 2

[5] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*, 2021. 1

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *preprint arXiv:2002.05709*, 2020. 1

[7] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 1

[8] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *ECCV*, 2020. 1, 2, 3, 5, 6, 7, 8

[9] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*, 2019. 3

[10] Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data. In *ICLR*, 2018. 3

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2018. 1, 2

[12] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 1

[13] Matthew Honnibal and Ines Montani. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear, 2017. 4

[14] Xiaowei Hu, Xi Yin, Kevin Lin, Lijuan Wang, Lei Zhang, Jianfeng Gao, and Zicheng Liu. Vivo: Surpassing human performance in novel object captioning with visual vocabulary pre-training. In *AAAI*, 2021. 2

[15] Zhicheng Huang, Zhaoyang Zeng, Yupan Huang, Bei Liu, Dongmei Fu, and Jianlong Fu. Seeing out of the box: End-to-end pre-training for vision-language representation learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 12976–12985. Computer Vision Foundation / IEEE, 2021. 1

[16] Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. *arXiv preprint arXiv:2004.00849*, 2020. 1, 2

[17] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *ICCV*, 2021. 2

[18] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *ICML*, 2021. 1, 2

[19] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 6

[20] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vis.*, 123(1):32–73, 2017. 1, 2

[21] Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. Unsupervised machine translation using monolingual corpora only. In *ICLR*, 2018. 3

[22] Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. Phrase-based & neural unsupervised machine translation. In *EMNLP*, 2018. 3

[23] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *AAAI*, 2020. 1, 2, 3, 5

[24] Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq R. Joty, Caiming Xiong, and Steven C. H. Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *CoRR*, abs/2107.07651, 2021. 1, 2

[25] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019. 2, 3, 6

[26] Liunian Harold Li, Haoxuan You, Zhecan Wang, Alireza Zareian, Shih-Fu Chang, and Kai-Wei Chang. Unsupervised vision-and-language pre-training without parallel images and captions. In *NACCL*, 2021. 1, 2, 3, 5, 6

[27] Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. Unimo: Towards unified-modal understanding and generation via cross-modal contrastive learning. *arXiv preprint arXiv:2012.15409*, 2020. 1, 2, 3

[28] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *ECCV*, 2020. 2

[29] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proceedings of ECCV*, pages 740–755. Springer, 2014. 1, 2

[30] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019. 1

[31] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *NeurIPS*, 2019. 1, 2, 5, 6

[32] Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 12-in-1: Multi-task vision and language representation learning. In *CVPR*, 2020. 1, 2

[33] Vicente Ordonez, Girish Kulkarni, and Tamara L. Berg. Im2text: Describing images using 1 million captioned photographs. In *NIPS*, 2011. 1, 2

[34] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *EMNLP*, 2019. 2, 4

[35] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, 2018. 1, 2, 5

[36] Amanpreet Singh, Vedanuj Goswami, Vivek Natarajan, Yu Jiang, Xinlei Chen, Meet Shah, Marcus Rohrbach, Dhruv Batra, and Devi Parikh. Mmf: A multimodal framework for vision and language research. https://github.com/facebookresearch/mmf, 2020. 6

[37] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*, 2019. 1, 2, 3, 5, 6

[38] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. *arXiv preprint arXiv:1811.00491*, 2018. 2, 6

[39] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *EMNLP*, 2019. 1, 2

[40] Hao Tan and Mohit Bansal. Vokenization: Improving language understanding with contextualized, visual-grounded supervision. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 2020. 4

[41] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*, 2021. 1, 2

[42] Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. Visual entailment: A novel task for fine-grained image understanding. *arXiv preprint arXiv:1901.06706*, 2019. 2, 6

[43] Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. Ernie-vil: Knowledge enhanced vision-language representations through scene graph. *arXiv preprint arXiv:2006.16934*, 2020. 1, 2

[44] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *ECCV*, 2016. 2, 6

[45] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5579–5588, June 2021. 3, 6

[46] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *CVPR*, 2021. 5, 6

[47] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and vqa. In *AAAI*, 2020. 2

[48] Mingyang Zhou, Luowei Zhou, Shuohang Wang, Yu Cheng, Linjie Li, Zhou Yu, and Jingjing Liu. Uc2: Universal cross-lingual cross-modal vision-and-language pre-training. In *CVPR*, 2021. 4, 8

[49] Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015. 2, 5