GIVL: Improving Geographical Inclusivity of Vision-Language Models with Pre-Training Methods

Da Yin♠   Feng Gao◇   Govind Thattai◇   Michael Johnston◇   Kai-Wei Chang♠,◇

♠ University of California, Los Angeles   ◇ Amazon Alexa AI

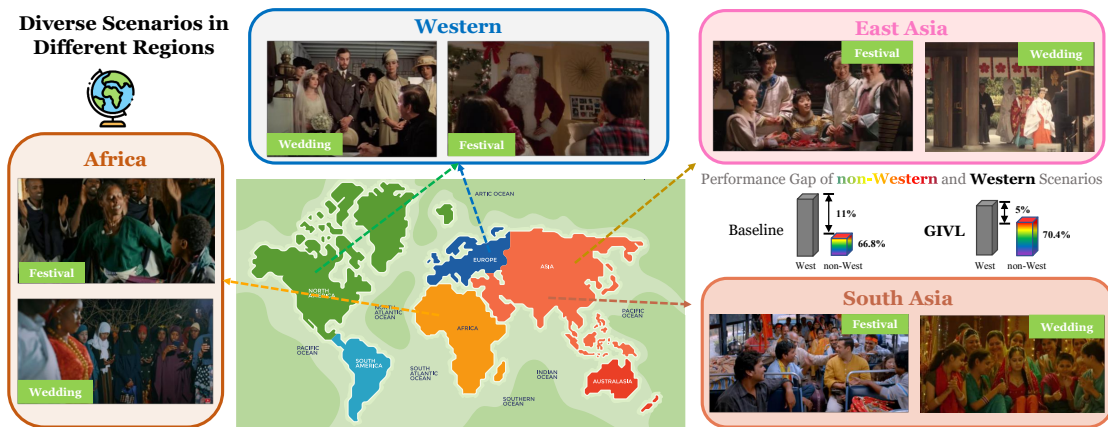{da.yin,kwchang}@cs.ucla.edu, {fenggo,thattg,mjohnstn}@amazon.com

Figure 1. Scenarios around the world including festivals and weddings. Even the same scenarios have distinct visual characteristics across regions (a.k.a. geographically diverse). Compared with prior Vision-Language Pre-trained Models (VLPs), GIVL achieves much better performance on non-Western data in GD-VCR [49]. GIVL can also make the gap between Western and non-Western cases much closer.

Abstract

*A key goal for the advancement of AI is to develop technologies that serve the needs not just of one group but of all communities regardless of their geographical region. In fact, a significant proportion of knowledge is locally shared by people from certain regions but may not apply equally in other regions because of cultural differences. If a model is unaware of regional characteristics, it may lead to performance disparity across regions and result in bias against underrepresented groups. We propose GIVL, a **G**eographically **I**nclusive **V**ision-and-**L**anguage Pre-trained model. There are two attributes of geo-diverse visual concepts which can help to learn geo-diverse knowledge: 1) concepts under similar categories have unique knowledge and visual characteristics, 2) concepts with similar visual features may fall in completely different categories. Motivated by the attributes, we design new pre-training objectives Image-Knowledge Matching (IKM) and Image Edit Checking (IEC) to pre-train GIVL. Compared with similar-size models pre-trained with similar scale of data, GIVL achieves state-of-the-art (SOTA) and more balanced performance on geo-diverse V&L tasks.*

Code and data are released at `https://github.com/WadeYin9712/GIVL`.

1. Introduction

Vision-Language Pre-trained Models (VLPs) [9, 23, 24, 29, 53] have achieved remarkable performance on Vision-Language (V&L) tasks including visual question answering [11, 12, 15], image-text retrieval [22], and image captioning [19, 27]. Pre-trained with large-scale corpora of image-text pairs, *e.g.* COCO [27], OpenImages [21], VLPs are capable of learning multi-modal representations and can be effectively fine-tuned on downstream V&L tasks.

While VLPs can solve a broad range of V&L tasks, to deploy VLPs in real-world applications, it is essential to consider the geographical inclusivity[1] of VLPs. Because of geographic differences, images from different regions embody a large amount of knowledge that is locally shared but cannot be applied in other regions, *i.e.* geographically diverse. For example, in Figure 1, the festivals in different

---

[1]We use regions as a proxy to estimate inclusivity of V&L models. People in the same regions may have different cultures and traditions.

regions look different.

Ideally, a geographically inclusive VLP should be capable of achieving comparable performance over all the images, regardless of their origins. However, current VLPs does not perform equally well on data from different regions. For example, prior works [28, 49] show that on geo-diverse V&L tasks, there is nearly a 20% performance discrepancy between Western and East Asian images when current VLPs are applied. To combat such geographical bias, we aim to design methods to make VLPs achieve more balanced performance across regions.

One solution to mitigating bias is to obtain diverse task-specific annotations for each region and fine-tune VLPs on the new annotations. However, according to [17], most Amazon MTurk annotators are from US and India, and may be unfamiliar with the cultures of other regions. Thus, it is unrealistic to obtain large-scale geo-diverse annotations even in such a popular crowdsourcing platform.

Pre-training a *unified* VLP with large-scale *unannotated* geo-diverse images and corresponding knowledge could make the VLP a foundation to provide more generalizable representations and help to transfer on comprehending images from various regions easier. In this paper, we propose **GIVL**, a **G**eographically **I**nclusive **V**ision-and-**L**anguage Pre-trained model. We focus on **how to encourage GIVL to better learn geo-diverse knowledge on images from different regions during its pre-training stage**.

We observe two attributes of geo-diverse visual concepts that can contribute to learning geo-diverse knowledge:

A1: Concepts under similar categories have unique knowledge and visual characteristics. For example, traditional Western and Chinese festivals, like *Christmas* and *Chinese New Year* in Figure 1, are held with different rituals and their decoration style differs as well. It is necessary for GIVL to learn the difference between their corresponding knowledge and precisely distinguish these visual concepts. On the other hand, *Christmas* and *Chinese New Year* are both festivals. Learning the commonalities of visual concepts (e.g., both images in Figure 1 belong to the same category "festival") would help model connect Western and non-Western concepts and contribute to more effective transfer on geo-diverse images.

A2: Concepts with similar visual features may lie in completely different categories. In Figure 2, *Chinese paper cuttings* share visual features (e.g., color, shape) with *red frisbee*. Similarly. *sugar cane* and *flute* share visual features. However, these concepts are not related to each other. Since geo-diverse images cover a broader range of visual concepts, differentiating visually similar concepts given visual contexts is also essential.

To this end, besides common objectives Masked Language Modeling (MLM) and Image-Text Matching (ITM) for pre-training VLPs, we propose two additional pre-

Figure 2. Example of Chinese paper cuttings and red frisbee (left), sugar cane and flute (right). Different concepts may be visually similar, but they may have completely different functionalities.

training objectives, **Image-Knowledge Matching (IKM)** and **Image Edit Checking (IEC)**. IKM is used to learn the alignment between images and corresponding textual knowledge in Wikipedia. It requires GIVL to not only judge if the input textual knowledge matches input images, but also identify whether the visual concepts described in input knowledge falls into similar categories of the concepts in input images. This encourages GIVL to learn corresponding relationship between knowledge and images as well as recognize similarity among geo-diverse visual concepts. IEC is proposed to identify whether a visual concept in input image is replaced by another concept that is visually similar but lies in an irrelevant category (see Fig.3 for an example). It enables GIVL to capture nuances between visually similar concepts after the replacement given visual contexts.

Our contributions and empirical results are as follows:

- By considering the attributes of geo-diverse visual concepts, we propose two novel V&L pre-training objectives Image-Knowledge Matching (IKM) and Image Edit Checking (IEC) that can greatly improve the geographical inclusivity of VLPs.

- Compared with similar-size VLPs pre-trained with similar scale of data, GIVL achieves state-of-the-art (SOTA) and more balanced performance over different regions on geo-diverse V&L tasks including MaRVL [28], GD-VCR [49] and WIT Image-Text Retrieval [38]. For geo-diverse zero-shot image classification on Dollar Street dataset[2], GIVL outperforms VinVL [53] 26%.

2. Related Work

Vision-Language Pre-Trained Models (VLPs). VLPs [4, 9, 22–24, 26, 29, 36, 42, 53] are proposed to tackle tasks that require understanding of both images and texts. Following the paradigm of pre-training language models [8, 30, 33], in common practice, VLPs use Transformer [43] as the backbone and pre-train them with large-scale image-caption pairs. The commonly used image-text parallel data are from multiple sources including COCO [27], Flickr30K [50], Conceptual Captions [35] and OpenImages [21] datasets. Currently, VLPs have achieved remarkable performance

---

on various V&L tasks including visual question answering [12, 15], visual reasoning [40], image captioning [27], and image-text retrieval [27, 50]. Most recent works focus on scaling up VLPs; this paper studies an orthogonal but important concerns - how to leverage diverse knowledge to improve inclusivity of VLPs.

Geographical Bias. Geographical bias [7, 31, 34, 44, 48] is a severe problem in that AI applications have. Previous works [28, 49] reveal the fact that on geo-diverse V&L tasks, the performance gap between non-Western and Western images is significant when using VLPs. Similarly, object recognition models' performance greatly drops on non-Western images [7, 34]. Researchers [7, 34, 44] find that one factor of the geographical bias is introduced by an imbalanced distribution of training data with respect to geographical location. They [7] observe that COCO and Open-Images, two widely used pre-trained corpora for VLPs, are amero-centric and euro-centric. We also find that VLPs can understand basic visual information in images from different regions, but are less able to leverage geo-diverse knowledge and reason [49].

Geo-Diverse V&L Tasks. GD-VCR [49] studies whether a model can understand commonsense on geo-diverse images. V&L models are required to select the correct answer from four answer choices given textual questions involving geo-diverse commonsense and the corresponding images. MaRVL [28] is another V&L task that requires visual reasoning with cultural knowledge of images from non-Western regions. It is formulated as a binary classification problem in which the model needs to judge whether a sentence correctly describes two images from non-Western regions. WIT image-text retrieval [3, 38] is a standard multimodal retrieval task on geo-diverse Wikipedia images.

3. Methods

In this section, we introduce the pre-training method of GIVL in detail. Section 3.1 provides preliminary of GIVL pre-training method including the definition of visual concept and category. Section 3.2 describes GIVL's pre-training objectives. Section 3.3 and 3.4 illustrate the process of acquiring essential information used to construct input contents for objectives Image-Knowledge Matching (IKM) and Image Edit Checking (IEC). Specifically, Section 3.3 shows how to extract visual concept name from an image caption and its category information from Wikipedia. Section 3.4 shows how to locate visual concept to corresponding detected objects in input image.

3.1. Preliminary

Definition of A Visual Concept and Category. `Visual concept` is an object or scenario that an image mainly involves. For example, Figure 3 shows the visual concept of

Chinese paper cuttings. Each specific visual concept corresponds to one general `category`. Each category covers various visual concepts having particular shared characteristics. For example, the category of visual concept *Chinese paper cuttings* is *art*. The *art* category includes other visual concepts such as *Jewish paper cuttings*. The extraction pipeline for visual concept and its category information will be introduced in Section 3.3.

Pre-Training Corpus. To improve the geographical inclusivity of VLPs, we use Wikipedia Image-Text (WIT) dataset [38] as a primary source of geo-diverse images. WIT contains 2.95M images in total[3]. We also incorporate 0.22M commonly used V&L pre-training images from COCO [27], Flickr30k [50], and GQA. Images in WIT dataset come with the corresponding Wikipedia sections that include the corresponding knowledge of WIT images. This knowledge[4], such as customs and history, is usually culturally related and not explicitly described in the images. Such knowledge plays a crucial role in helping VLPs understand visual concepts in geo-diverse images more comprehensively.

Input for Pre-Training. We organize the input for GIVL pre-training as follows:

$$[\text{CLS}] \; \mathbf{c} \; [\text{SEP}] \; \mathbf{k} \; [\text{SEP}] \; \mathbf{t} \; [\text{SEP}] \; \mathbf{v}, \tag{1}$$

where $\mathbf{c}$ is either an image caption or a GQA question; $\mathbf{k}$ denotes the corresponding knowledge of the visual concept in input image $\mathbf{I}$; $\mathbf{t}$ is either tags of detected objects or a GQA answer; $\mathbf{v}$ is a list of visual embeddings generated from input image $\mathbf{I}$ by a ResNeXt-152 C4 detection model [53]. p_v is the name of the visual concept contained in image $\mathbf{I}$.

3.2. Pre-Training Objectives for GIVL

We pre-train GIVL with four objectives over pre-trained BERT-base model [8]: Masked Language Modeling (MLM), Image-Text Matching (ITM), Image-Knowledge Matching (IKM), Image Edit Checking (IEC). We introduce each pre-training objective as follows.

3.2.1 MLM and ITM Objectives

Masked Language Modeling (MLM) is a learning objective prevalently used in V&L pre-training. Given the context of model inputs, GIVL needs to recover the tokens masked by [MASK]. MLM loss $\mathcal{L}_{MLM}$ is the average of all cross-entropy loss with respect to the probability of predicting the correct masked tokens given a vocabulary.

Image-Text Matching (ITM) is another commonly applied objective that enables GIVL to learn the alignment

---

[3]GIVL focuses on English-only V&L tasks. We only consider images with English captions, which only occupy 30% out of the entire WIT.

[4]The knowledge of COCO, Flickr30K and GQA images is the first sentence in Wikipedia pages of the objects mentioned in captions.
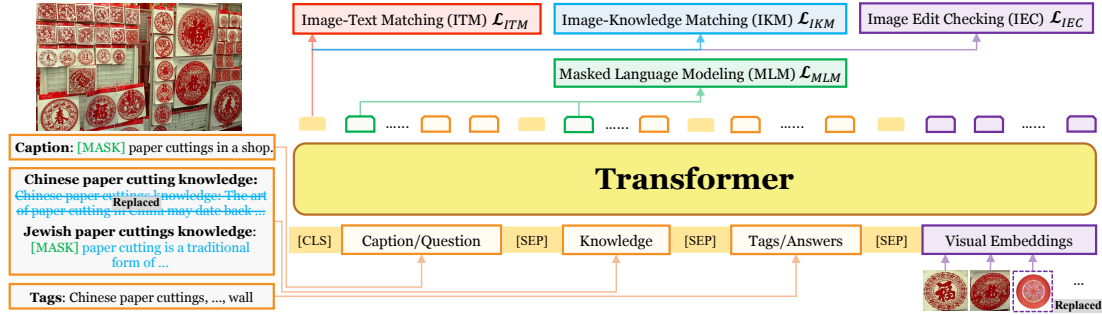
Figure 3. GIVL pre-training method with four pre-training objectives. The input image is about the visual concept *Chinese paper cuttings*. The input knowledge is about *Jewish paper cuttings* rather than *Chinese paper cuttings*, but it is also the knowledge describing a visual concept that shares a similar category with *Chinese paper cuttings*. Hence, for Image-Knowledge Matching (IKM) objective, the input contents belong to Type 3. Also, the visual concept *Chinese paper cuttings* is replaced with a visually similar concept *red frisbee*. Thus, for Image Edit Checking (IEC) objective, the input contents belong to Type 2.

between texts and images. Following VinVL [53], given an input image $\mathbf{I}$, we construct three types of input contents for $\mathbf{c}$ and $\mathbf{t}$. It is formulated as a 3-way classification task, $y^{c,t} \in \{0, 1, 2\}$, where 0 represents that $\mathbf{c}$ and $\mathbf{t}$ both match the input image $\mathbf{I}$; 1 means when $\mathbf{t}$ matches image $\mathbf{I}$ whereas $\mathbf{c}$ mismatches the image $\mathbf{I}$; 2 indicates $\mathbf{c}$ matches $\mathbf{I}$ but $\mathbf{t}$ mismatches $\mathbf{I}$. ITM loss is the cross-entropy loss with respect to the probability of predicting the type of input contents upon [CLS] representation.

3.2.2 Image-Knowledge Matching (IKM)

We propose Image-Knowledge Matching (IKM) to assist GIVL in learning knowledge of geo-diverse visual concepts. With the help of IKM, we encourage GIVL to learn the corresponding knowledge of the visual concepts and discover connections between geo-diverse visual concepts.

Although the visual characteristics of the geo-diverse visual concepts in GIVL's pre-training corpus may be poles apart, they could be clustered in similar categories. For example, in Figure 1, the visual characteristics of traditional Western and non-Western festivals are different, but these scenarios all belong to the same category *festival*. Learning to identify category similarity can connect diverse visual concepts under similar categories and generalize to understanding more relevant concepts across regions more simply. On the other hand, each of the visual concepts in similar categories associates with unique knowledge. Therefore, it is also crucial for GIVL to precisely distinguish if input knowledge aligns with the input image.

To this end, we construct the three types of input contents and formulate IKM as a 3-way classification task to enforce GIVL to identify the input type:

- Type 1: $\mathbf{k}$ matches input image $\mathbf{I}$;

- Type 2: $\mathbf{k}$ mismatches input image $\mathbf{I}$ and the visual concept described by $\mathbf{k}$ does **NOT** fall into a similar

category of the visual concept p_v in $\mathbf{I}$;

- Type 3: $\mathbf{k}$ mismatches input image $\mathbf{I}$ but the visual concept described by $\mathbf{k}$ falls into a similar category of the visual concept p_v in $\mathbf{I}$.

To select knowledge $\mathbf{k}$ for Type 3 input in IKM, we need to conduct two steps (i) extracting the name of visual concept p_v of input image $\mathbf{I}$ from its caption (for GQA data, see supplementary) and (ii) looking for visual concepts under similar categories. More details of extracting p_v from image caption and its category information will be introduced in Section 3.3. After the visual concept name p_v is extracted from the caption, to find a visual concept which falls in the most relevant categories to p_v, we randomly sample 200 visual concepts as candidates from rest WIT training corpora. Then we select the candidate concept that has the most semantically similar category with p_v's category. Specifically, the sampled candidates are ranked by cosine similarity between text embeddings[5] of their category names and the embedding of p_v's category name. The process of selecting the most relevant visual concept p^* is illustrated in Eq. (2),

$$p^* = \arg\max_{p_i} \mathbf{CosineSim}(z_{p_i}, z_{p_v}), \qquad (2)$$

where z_{p_i} is the embedding of i-th sampled visual concept p_i's category, z_{p_v} is the embedding of p_v's category, **CosineSim** denotes the function quantifying cosine similarity between two embeddings. The corresponding knowledge of p^* can be regarded as $\mathbf{k}$ in Type 3 input.

For preparing $\mathbf{k}$ in Type 2 input content, we first set up a threshold for the cosine similarity between the embeddings of category names ($\tau = 0.3$) to filter out the visual concepts relevant with p_v. Then we randomly pick one of the retained visual concepts. The selected visual concept indicates the

---

[5]We utilize FastText [2] embeddings pre-trained on Wikipedia. Phrase embeddings are mean pooled embeddings of the words in the phrases.
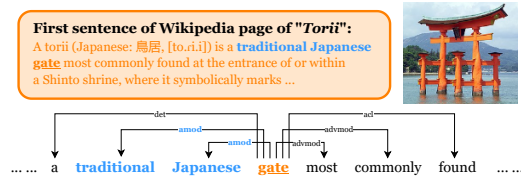
Figure 4. Steps to mine category information of visual concepts. The composition of the head noun ("*gate*", root of parse tree) and its modifiers ("*traditional Japanese*", words with "amod" relation with "*gate*") can be treated as the category of *torii* ("*traditional Japanese gate*").

one that has irrelevant category information with p_v. Its corresponding knowledge can be used as $\mathbf{k}$ in Type 2 input.

IKM loss is a cross-entropy loss with respect to the probability of predicting the type of relationship between the input image and knowledge upon [CLS] representation,

$$\mathcal{L}_{IKM} = -\frac{1}{|\mathcal{D}|} \sum_{i=1}^{|\mathcal{D}|} \log p(y_i^k|\mathbf{c}, \mathbf{k}, \mathbf{t}, \mathbf{v}), \qquad (3)$$

where $\mathcal{D}$ indicates the entire pre-training corpus[6] and y_i^k is the label for the input type in IKM.

3.2.3 Image Edit Checking (IEC)

To better differentiate visually similar but irrelevant concepts, we propose another pre-training objective Image Edit Checking (IEC). In geo-diverse setting, it is highly likely that visual concepts share similar visual characteristics but fall into completely different categories. For example, in Figure 2, *Chinese paper cuttings* are red and circular, which aligns with the visual characteristics of *red frisbee*. IEC is designed to identify whether a specific visual concept p_v in input image $\mathbf{I}$ is replaced by another visually similar one in an irrelevant category.

We consider two types of input contents for IEC:

- Type 1: Input image $\mathbf{I}$ remains the same;

- Type 2: The visual embedding of the visual concept p_v in input image $\mathbf{I}$ is replaced with the embedding of another concept that is visually similar but falls into an irrelevant category with p_v.

In Figure 3, since the visual concept *Chinese paper cuttings* is replaced with *red frisbee*, the input type is Type 2[7].

To prepare input contents of Type 2 data, we need to accomplish two steps (i) seeking the corresponding detected objects of the visual concept p_v in input image $\mathbf{I}$ from its caption (ii) looking for visually similar concepts for replacement. The pipeline of locating visual concept p_v is

introduced in Section 3.4. After the visual concept p_v is located, to select the proper visual concept for replacement in Type 2 input, we randomly sample 20 images, and then collect the visual embeddings and tag names of all the detected objects in the sampled images as candidates. The visual concept for replacement is selected according to two criteria: (i) its category is dissimilar[8] with the category information of concept p_v and (ii) its visual embedding is closest to p_v's visual embedding. We select irrelevant visual concepts with p_v to guarantee that the replacement is unreasonable given the image context.

IEC loss is a binary cross-entropy loss with respect to the probability of predicting whether the input image is modified upon the [CLS] representation,

$$\mathcal{L}_{IEC} = -\frac{1}{|\mathcal{D}|} \sum_{i=1}^{|\mathcal{D}|} \log p(y_i^v|\mathbf{c}, \mathbf{k}, \mathbf{t}, \mathbf{v}), \qquad (4)$$

where y_i^v is the label for input type in IEC. The final loss L is the sum of all losses mentioned above:

$$\mathcal{L} = \mathcal{L}_{MLM} + \mathcal{L}_{ITM} + \mathcal{L}_{IKM} + \mathcal{L}_{IEC}. \qquad (5)$$

3.3. Acquiring Categories of Visual Concepts

Acquiring the categories of visual concepts is a prerequisite step to construct GIVL inputs for IKM and IEC. We first need to extract the visual concept name p_v in input image $\mathbf{I}$ from its image caption. We achieve this by parsing the caption with [32]. p_v is the composition of the head noun and its modifiers in the parse tree. For example, given a caption "*Chinese paper cuttings in a shop*", p_v is *Chinese paper cuttings*, which is composed of the head noun "*cuttings*" and its modifiers "*Chinese paper*" in its parse tree.

To acquire p_v's category, we then search for Wikipedia with keyword p_v. If p_v is an entry of Wikipedia, we find that the category information can be usually found in the first sentence of Wikipedia introduction paragraph. As shown in Figure 4, the category of *torii* (i.e., *traditional Japanese gate*) is present in the first sentence "*A torii ... is a traditional Japanese gate most commonly ...*" Then we notice that the category name is the phrase consisting of the head noun and its modifiers in the first sentence. In this example, the head noun of the first sentence is "*gate*" and its modifier words are "*traditional*" and "*Japanese*". The final concatenation, "*traditional Japanese gate*", is the category of *torii*. Though the category information mined with these simple heuristics is imperfect, the extraction method is easy to implement and efficient in acquiring categories of large quantities of visual concepts.

---

[8]We use the cosine similarity between embeddings of the candidate visual concept's category and p_v's category. Any candidate concepts with a similarity lower than 0.3 are treated as dissimilar ones.
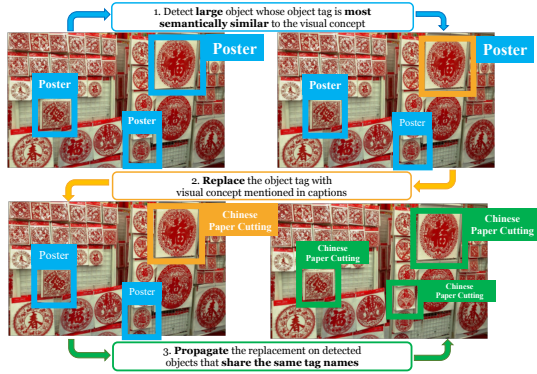
Figure 5. Steps to locate novel visual concepts in input images.

3.4. Locating Visual Concepts in Images

With a limited amount of object class labels, it is difficult for current object detectors to detect a geo-diverse visual concept p_v. Therefore, we introduce a simple approach to efficiently locate the corresponding object given a visual concept p_v. We find that a visual concept p_v is commonly (i) classified as a tag name that has similar semantics with p_v's category, and (ii) its image patch occupies a large portion of the image. To this end, we design heuristics to locate novel visual concepts according to our empirical findings. First, only the top-10 large detected objects from each image will be considered. Second, we calculate the similarity between their object tags and p_v's category. The one with the highest similarity score will be treated as the object corresponding to p_v. We take Figure 5 as an example. The visual concept p_v to be located is *Chinese paper cuttings*. Suppose that one of the *Chinese paper cuttings* (the object in top right corner) is among top-10 large detected objects. Besides, its original detected object tag is *poster*, which is the most semantically similar to *Chinese paper cuttings*'s category. Hence, we can replace its original object tag with *Chinese paper cuttings* as it is the corresponding object we are looking for.

The method above only locates one visual concept per image. However, it is possible that one image may contain multiple identical visual concepts. For example, in Figure 5, there are a couple of *Chinese paper cuttings*. To solve this problem, we simply propagate the visual concept name of *Chinese paper cuttings* to other objects that share the same original detection labels.

4. Experiments

We conduct two sets of experiments to evaluate GIVL. We first evaluate GIVL on multiple geo-diverse V&L tasks including zero-shot image classification, V&L reasoning and image-text retrieval. It helps us to verify the effectiveness of GIVL under geo-diverse settings. On the other hand, experiments on common V&L tasks are conducted to prove the generalizability of GIVL's pre-training method.

4.1. Baselines for Ablation Study

Five baselines are described below. For fair comparison, pre-training corpus, number of pre-training steps, and hyper-parameters are all identical to GIVL[9]. Since V&L pre-training is extremely time consuming, all baselines are pre-trained with 500K steps in ablation study.

GIVL w/o $\mathcal{L}_{IKM}$ & GIVL w/o $\mathcal{L}_{IEC}$. GIVL w/o $\mathcal{L}_{IKM}$ and GIVL w/o $\mathcal{L}_{IEC}$ is the model pre-trained without Image-Knowledge Matching (IKM) objective and Image Edit Checking (IEC) objective, respectively. We demonstrate the effectiveness of our proposed pre-training objectives with these two baselines.

VinVL*. VinVL* is pre-trained only with MLM and ITM objectives as VinVL [53]. It also shares the same pre-training corpus with GIVL. The only difference between GIVL and VinVL* is objectives. GIVL is pre-trained with Image-Knowledge Matching (IKM) and Image Edit Checking (IEC) but VinVL* is not. Comparing GIVL and VinVL* can manifest the improvement by introducing IKM and IEC objectives on geo-diverse V&L tasks. The comparison is also fair for the pre-training methods of GIVL and VinVL on common V&L tasks.

GIVL w/ CLIP. Some recent VLPs utilize CLIP [33] as the vision encoder. We replace object-level visual encoder in GIVL with CLIP to check if it can further improve performance. CLIP provides grid-level visual representation instead of object-level's. Therefore, IEC objective is removed because it involves object-level replacements.

GIVL-B. The only difference between GIVL and GIVL-B is that the IKM objective of GIVL-B is a binary classification objective instead of 3-way classification. For IKM, it requires GIVL-B to identify whether the input knowledge matches the image contents. GIVL-B doesn't need to judge whether the input knowledge describes a visual concept that shares similar category with the concept in input image. The comparison between GIVL and GIVL-B is able to demonstrate the effect of incorporating category information for learning the knowledge of geo-diverse visual concepts.

4.2. Results on Geo-Diverse Benchmarks

Geo-Diverse Zero-Shot Image Classification. Geo-diverse zero-shot image classification is a downstream geo-diverse V&L task that directly evaluates the effectiveness of the pre-training methods. We evaluate models on Dollar Street dataset[10]. It is labeled with 127 classes, each of which contains images around the world. For classification on one image, we compose 127 inputs, each of which is the concatenation of one class name, the class's corresponding

---

[9]Details of experimental setups are described in Appendix A.

[10]Images in Dollar Street are labeled with country information. The proxy to categorize Western and non-Western countries is based on [16].

Model	#Param	Acc.	Western/non-Western
Prior VLPs			
VinVL*	112M	1.21	1.77/1.01
VinVL [53]	112M	1.29	1.25/1.30
Ours			
GIVL w/o $\mathcal{L}_{IKM}$	112M	21.37	25.31/20.37
GIVL w/o $\mathcal{L}_{IEC}$	112M	12.96	12.71/13.02
GIVL w/ CLIP	199M	18.04	22.89/16.82
GIVL-B	112M	20.35	23.93/19.45
GIVL	112M	**27.25**	**31.65/26.15**

Table 1. Results on geo-diverse zero-shot image classification on Dollar Street dataset. We also show the respective performance on Western and non-Western images.

Model	Data/Steps	#Param	Acc.	NLVR2	Δ
Prior VLPs					
ViLBERT [29]	3.3M/-	274M	66.53	77.40	10.87
VinVL [53]	5.65M/2M	112M	72.48	81.03	8.55
VinVL*	3.17M/500K	112M	69.66	77.93	8.27
X-VLM† [52]	16M/-	216M	73.02	84.41	11.39
ALBEF† [23]	14M/-	210M	73.17	82.54	9.37
METER† [9]	4M/-	352M	73.47	82.33	8.86
Ours					
GIVL w/o $\mathcal{L}_{IKM}$	3.17M/500K	112M	72.11	-	-
GIVL w/o $\mathcal{L}_{IEC}$	3.17M/500K	112M	68.58	-	-
GIVL w/ CLIP	3.17M/500K	199M	71.78	-	-
GIVL-B	3.17M/500K	112M	70.26	-	-
GIVL	3.17M/500K	112M	72.50	79.06	**6.56**
GIVL	3.17M/900K	112M	**72.70**	79.87	7.17

Table 2. Results on MaRVL testing set. We also show the performance discrepancy Δ between NLVR2 and MaRVL. † denotes the results reported in [55]. Bolded numbers indicate highest performance and smallest performance gap on MaRVL other than models colored with gray.

knowledge[11], tag names and visual embeddings of the detected objects. We compare the probability of predicting that each class name matches the input image via ITM objective for all the 127 classes. The class with the highest probability is treated as the final classification result.

As shown in Table 1, GIVL outperforms both VinVL and VinVL* by a significant margin around 26%. GIVL achieves 6%-20% improvement in ablation studies, demonstrating the effectiveness of the proposed IKM and IEC objectives. We also find that GIVL outperforms GIVL w/ CLIP which involves a strong vision encoder. It further demonstrates that object-level visual representations and object-level pre-training objective IEC are effective for learning geo-diverse visual concepts.

Multicultural Visual Reasoning (MaRVL). Following NLVR2 [40], MaRVL [28] is a V&L task that requires models to identify whether a sentence correctly describes the contents of two input images. MaRVL images involve diverse visual concepts in non-Western regions. Since MaRVL[12] is merely a testing set, following [28], we fine-tune models on NLVR2 training set and then select the best

---

[11]The knowledge of each class sources from Wikipedia and Wordhoard.
[12]We use the translated English version of MaRVL dataset in [55].

Model	#Param	Acc.	Non-West Acc.	Δ
Prior VLPs				
VisualBERT [29]	135M	53.95	-	10.42
ViLBERT [29]	274M	59.99	-	7.28
VinVL*	112M	69.07	66.45	8.46
VinVL [53]	112M	70.20	66.78	11.04
Ours				
GIVL w/o $\mathcal{L}_{IKM}$	112M	69.56	65.96	-
GIVL w/o $\mathcal{L}_{IEC}$	112M	69.89	66.92	-
GIVL w/ CLIP	199M	70.43	67.25	10.20
GIVL-B	112M	69.56	65.96	-
GIVL	112M	70.32	68.41	6.14
GIVL (1M)	112M	**72.01**	**70.4**	**4.97**

Table 3. Results on GD-VCR. We also show the results on all the non-Western images in GD-VCR and discrepancy Δ between Western and non-Western images.

checkpoint on the dev set of NLVR2 to evaluate on MaRVL.

From Table 2, we observe that GIVL outperforms the ablated baselines pre-trained without our proposed objectives IKM and IEC, respectively. Also, similar to the observations on Dollar Street dataset, compared with VinVL* pre-trained with the same corpus as GIVL, GIVL achieves higher performance. It further demonstrates that the pre-training objectives of GIVL can help VLPs learn geo-diverse visual concepts better than VinVL.

We also compare GIVL with VLPs (i) $3\times$ larger model (METER) and (ii) pre-trained with $2-5\times$ larger corpus (VinVL, X-VLM and ALBEF). GIVL achieves competitive performance with much less data and smaller model size. Additionally, we attach importance to performance gap on NLVR2 and MaRVL. It is shown that the visual concepts in NLVR2 dataset are Western-centric [28]. A smaller performance gap between NLVR2 and MaRVL means less bias against non-Western regions. We observe that GIVL can achieve more balanced performance on both datasets, while other VLPs including METER, X-VLM and ALBEF have larger performance discrepancy.

Geo-Diverse Visual Commonsense Reasoning (GD-VCR). GD-VCR is a testing set to evaluate multi-modal models' ability to understand geo-diverse commonsense knowledge in images. It is a multiple-choice QA task which requires geo-diverse commonsense reasoning. We fine-tune models on VCR [51] training set and select the best checkpoint on VCR's dev set to evaluate on GD-VCR.

As shown in Table 3, GIVL outperforms all prior similar-size VLPs models trained with similar number of images. GIVL also outperforms all ablated baselines except for GIVL w/ CLIP, which uses a much stronger visual encoder and only achieves 0.1% subtle improvements. Besides, we highlight the performance gap between Western and non-Western data in GD-VCR. GIVL has significantly smaller gap than any of the ablated baselines. While GIVL w/ CLIP has a marginal improvement over GIVL, the performance gap of GIVL is 4.06% smaller than GIVL w/ CLIP.
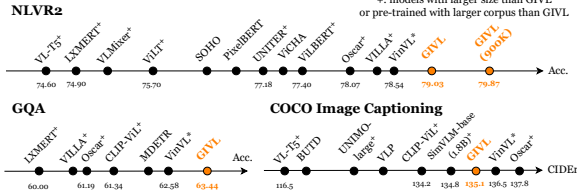
Figure 6. GIVL performance on common V&L tasks. Complete results are shown in Appendix B.

Model	Data	#Param	I/R	T/R
Prior VLPs				
LXMERT [42]	-	240M	14.28	14.86
VisualBERT [24]	-	135M	15.36	15.75
UNITER [4]	-	-	15.43	16.01
VL-BERT [39]	3.3M	-	15.11	16.09
ViLBERT [29]	3.3M	274M	15.40	16.93
VinVL [53]	5.65M	112M	27.78	28.65
VinVL*	3.17M	112M	25.44	25.50
Ours				
GIVL w/o $\mathcal{L}_{IKM}$	3.17M	112M	26.21	26.97
GIVL w/o $\mathcal{L}_{IEC}$	3.17M	112M	28.08	28.18
GIVL w/ CLIP	3.17M	199M	27.94	28.17
GIVL-B	3.17M	112M	29.97	29.86
GIVL	3.17M	112M	28.00	28.79
GIVL (1M)	3.17M	112M	**29.98**	**30.79**

Table 4. Results on WIT image-text retrieval task. I/R and T/R denote image retrieval and text retrieval. The evaluation metric is Recall@1. 1M denotes the number of pre-training steps.

Wikipedia Image-Text Retrieval (WIT). WIT image-text retrieval is a standard retrieval task on geo-diverse Wikipedia images[13]. Table 4 shows that GIVL achieves superior performance comparing to baselines except GIVL-B. Pre-trained with 1M steps, GIVL obtains SOTA performance on WIT image-text retrieval task.

4.3. Results on Common V&L Benchmarks

Besides testing GIVL on geo-diverse V&L tasks, we benchmark GIVL on common V&L task to investigate whether the pre-training method of GIVL is competitive among existing VLPs. We don't expect GIVL to perform the best among SOTA VLPs on these V&L benchmarks, because they are annotated with Western-centric data and SOTAs are trained with much larger similar data as well. We aim to answer two questions. Q1: *Is GIVL able to obtain comparable performance with VLPs pre-trained with similar scale of data?* Q2: *Can GIVL perform as strongly as SOTA VLPs pre-trained with the same corpus?*

To answer Q1, we evaluate GIVL on common V&L benchmarks including NLVR2, GQA and COCO captioning. As shown in Figure 6, for NLVR2, GIVL is able to beat 11 VLPs with much more parameters and pre-trained with more data. For GQA, GIVL performs better than most of the VLPs. For COCO image captioning, it can even obtain close performance with SimVLM-base, a VLP pre-trained

---

Figure 7. GIVL and VinVL's performance on non-Western and Western images related to geo-diverse categories.

with 1.8B images. Overall, even if GIVL is pre-trained with the corpus whose domain is not similar with common V&L tasks, it can still achieve competitive results.

For Q2, we target on VinVL, a strong VLP that once swept leaderboards of multiple V&L tasks. For fair comparison, we reproduce the pre-training process of VinVL with GIVL pre-training corpus. As mentioned in Section 4.1, we denote the reproduced pre-training as VinVL*. On above three V&L datasets covered in Figure 6, the performance difference between GIVL and VinVL* is subtle. We argue that GIVL could perform as well as VinVL on common V&L benchmarks if it was pre-trained with VinVL corpus.

4.4. Qualitative Study on Geo-Diverse Categories

We showcase examples from GD-VCR and Dollar Street datasets to better demonstrate GIVL's advantages. In Figure 7, non-Western *festivals*, *servants* and *religions* are quite different from those in Western regions. We find that GIVL's performance gap on the images involving these categories is significantly smaller than VinVL on GD-VCR. Moreover, GIVL's performance on non-Western images is 5-8% higher than VinVL. For Dollar Street, while the overall performance of GIVL is around 30%, it can achieve above 55% accuracy when recognizing *vegetables* and *drying clothes* which greatly vary across the world. GIVL even outperforms VinVL 50% on those categories.

5. Conclusion

We propose GIVL, a geographically inclusive V&L pre-trained model. GIVL achieves strong and more balanced results on multiple geo-diverse V&L tasks. It can also produce competitive performance on common V&L tasks. By proposing GIVL, we call upon researchers to devise methods that can further improve geographical inclusivity of VLPs and popularize their applications for all.

---

[13]We use the translated English WIT retrieval data in [3].

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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018. 12, 13

[2] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017. 4

[3] Emanuele Bugliarello, Fangyu Liu, Jonas Pfeiffer, Siva Reddy, Desmond Elliott, Edoardo Maria Ponti, and Ivan Vulić. IGLUE: A benchmark for transfer learning across modalities, tasks, and languages. *ICML*, 2022. 3, 8, 12

[4] 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 *European conference on computer vision*, pages 104–120. Springer, 2020. 2, 8, 12

[5] Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. Unifying vision-and-language tasks via text generation. In *International Conference on Machine Learning*, pages 1931–1942. PMLR, 2021. 12, 13

[6] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar, Oct. 2014. Association for Computational Linguistics. 12

[7] Terrance de Vries, Ishan Misra, Changhan Wang, and Laurens van der Maaten. Does Object Recognition Work for Everyone? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 52–59, 2019. 3

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. 2, 3

[9] Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Pengchuan Zhang, Lu Yuan, Nanyun Peng, et al. 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, 2022. 1, 2, 7

[10] Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. Large-scale adversarial training for vision-and-language representation learning. *Advances in Neural Information Processing Systems*, 33:6616–6628, 2020. 12

[11] Feng Gao, Qing Ping, Govind Thattai, Aishwarya Reganti, Ying Nian Wu, and Prem Natarajan. Transform-retrieve-generate: Natural language-centric outside-knowledge visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5067–5077, 2022. 1

[12] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 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)*, 2017. 1, 3

[13] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12976–12985, 2021. 12

[14] 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. 12

[15] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 3

[16] Samuel P Huntington. The clash of civilizations? In *Culture and politics*, pages 99–118. Springer, 2000. 6

[17] Panagiotis G Ipeirotis. Demographics of mechanical turk. 2010. 2

[18] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1780–1790, 2021. 12

[19] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015. 1

[20] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR, 2021. 12

[21] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4. *International Journal of Computer Vision*, 128(7):1956–1981, 2020. 1, 2

[22] 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 *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11336–11344, 2020. 1, 2

[23] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021. 1, 2, 7

[24] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and perfor-

mant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019. 1, 2, 8

[25] 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. 13

[26] 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 *European Conference on Computer Vision*, pages 121–137. Springer, 2020. 2, 12, 13

[27] 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 *European conference on computer vision*, pages 740–755. Springer, 2014. 1, 2, 3

[28] Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. Visually Grounded Reasoning across Languages and Cultures. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10467–10485, Online and Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics. 2, 3, 7

[29] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019. 1, 2, 7, 8, 12

[30] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. 2

[31] Viraj Prabhu, Ramprasaath R. Selvaraju, Judy Hoffman, and Nikhil Naik. Can domain adaptation make object recognition work for everyone? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 3981–3988, June 2022. 3

[32] Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online, July 2020. Association for Computational Linguistics. 5

[33] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language Models are Unsupervised Multitask Learners. *OpenAI blog*, 1(8):9, 2019. 2, 6

[34] Shreya Shankar, Yoni Halpern, Eric Breck, James Atwood, Jimbo Wilson, and D Sculley. No Classification without Representation: Assessing Geodiversity Issues in Open Data Sets for the Developing World. *arXiv preprint arXiv:1711.08536*, 2017. 3

[35] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 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, Melbourne, Australia, July 2018. Association for Computational Linguistics. 2

[36] Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, and Kurt Keutzer. How much can clip benefit vision-and-language tasks? *arXiv preprint arXiv:2107.06383*, 2021. 2, 12, 13

[37] Mustafa Shukor, Guillaume Couairon, and Matthieu Cord. Efficient vision-language pretraining with visual concepts and hierarchical alignment. *arXiv preprint arXiv:2208.13628*, 2022. 12

[38] Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2443–2449, 2021. 2, 3

[39] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. In *International Conference on Learning Representations*, 2019. 8

[40] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A Corpus for Reasoning about Natural Language Grounded in Photographs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6418–6428, Florence, Italy, July 2019. Association for Computational Linguistics. 3, 7

[41] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27, 2014. 12

[42] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019. 2, 8, 12

[43] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All You Need. *Advances in Neural Information Processing Systems*, 30, 2017. 2

[44] Angelina Wang, Alexander Liu, Ryan Zhang, Anat Kleiman, Leslie Kim, Dora Zhao, Iroha Shirai, Arvind Narayanan, and Olga Russakovsky. Revise: A tool for measuring and mitigating bias in visual datasets. *International Journal of Computer Vision*, pages 1–21, 2022. 3

[45] Teng Wang, Wenhao Jiang, Zhichao Lu, Feng Zheng, Ran Cheng, Chengguo Yin, and Ping Luo. Vlmixer: Unpaired vision-language pre-training via cross-modal cutmix. In *International Conference on Machine Learning*, pages 22680–22690. PMLR, 2022. 12

[46] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. In *International Conference on Learning Representations*, 2021. 13

[47] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim

Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, Oct. 2020. Association for Computational Linguistics. 12

[48] Da Yin, Hritik Bansal, Masoud Monajatipoor, Liunian Harold Li, and Kai-Wei Chang. Geomlama: Geodiverse commonsense probing on multilingual pre-trained language models. *arXiv preprint arXiv:2205.12247*, 2022. 3

[49] Da Yin, Liunian Harold Li, Ziniu Hu, Nanyun Peng, and Kai-Wei Chang. Broaden the Vision: Geo-Diverse Visual Commonsense Reasoning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2115–2129, Online and Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics. 1, 2, 3

[50] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. 2, 3

[51] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6720–6731, 2019. 7

[52] Yan Zeng, Xinsong Zhang, and Hang Li. Multi-grained vision language pre-training: Aligning texts with visual concepts. *arXiv preprint arXiv:2111.08276*, 2021. 7

[53] 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*, pages 5579–5588, 2021. 1, 2, 3, 4, 6, 7, 8, 12, 13

[54] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and vqa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13041–13049, 2020. 13

[55] Wangchunshu Zhou, Yan Zeng, Shizhe Diao, and Xinsong Zhang. Vlue: A multi-task benchmark for evaluating vision-language models. *arXiv preprint arXiv:2205.15237*, 2022. 7