Appendix

A. Pre-Training and Fine-Tuning Details of Downstream V&L Tasks

A.1. Pre-Training Setups

GIVL is initialized with pre-trained parameters of BERT-base model [47]. It is pre-trained for at most 1M steps with a batch size of 720. The learning rate is $1e-4$ with linear decay. The maximum numbers of tokens in input texts and visual objects are 70 and 50, respectively. All the pre-training experiments for GIVL and ablated baselines are implemented with 8 A100 GPUs with 40GB GPU memory.

A.2. Fine-Tuning Setups

MaRVL and NLVR2. We fine-tune GIVL on NLVR2 for 20 epochs, with batch size 72 and learning rate $3e-5$. The maximum number of tokens in input texts and visual objects is 55. Because each sample has two input images, we include a maximum of 80 visual objects in the model input, with each image having a maximum of 40 input visual objects. As mentioned in Section 4.2, since MaRVL is a testing set following NLVR2's formulation, the fine-tuning results are based on the fine-tuning method discussed here.

GD-VCR. We fine-tune GIVL on original VCR dataset for 5 epochs, with batch size 128 and learning rate $3e-5$. For model input, we concatenate the question and four answer choices together, along with the visual embeddings of the input image. The maximum numbers of tokens and visual objects are 100 and 50, respectively.

WIT Image-Text Retrieval. We fine-tune GIVL on the WIT Image-Text retrieval training set for 20 epochs, with a batch size of 128 and a learning rate of $2e-5$. The maximum numbers of tokens in input texts and visual objects are 70 and 70, respectively. We use the translated English training and dev set provided in IGLUE [3].

COCO Captioning. We fine-tune GIVL on COCO captioning dataset for 60 epochs, with batch size 256 and learning rate $3e-5$ with Seq2Seq objective [6, 41]. The maximum numbers of tokens in input texts and visual objects are 70 and 50, respectively. After that, we further optimize GIVL with the CIDEr metric for 75 epochs with a batch size of 64 and a learning rate of $2e-6$. We use beam search with beam size 5 [1] to sample the generation results, and the maximum length of the generated captions is 20 words.

GQA. We fine-tune GIVL on GQA for 5 epochs, with batch size 128 and learning rate $5e-5$. The maximum numbers of tokens in input texts and visual objects are 165 and 45, respectively.

B. Detailed Results on Common V&L Tasks

As mentioned in Section 4, we also conduct experiments on common Vision-Language (V&L) tasks. We show detailed experimental results in Table 5, 6 and 7 for GQA, NLVR2 and COCO captioning, respectively.

In Table 5, we show that GIVL outperforms many prior Vision-Language Pre-trained Models (VLPs) on GQA. We emphasize that GIVL is trained with significantly less data than most of the prior VLPs, while GIVL also uses fewer parameters compared to these VLPs. For fair comparison, VinVL* uses the same pre-training data as GIVL.

Model	#Param	Data	Acc.
Prior VLPs			
LXMERT [42]	240M	-	60.00
Oscar [26]	-	4.1M	61.19
CLIP-ViL [36]	178M	-	61.34
MDETR [18]	-	-	62.48
VinVL* [53]	112M	3.17M	62.58
Ours			
GIVL	112M	3.17M	**63.44**

Table 5. Results on GQA test-dev set.

We also evaluate the proposed GIVL on the NLVR2 dataset. Similar to results in GQA, according to Table 6, GIVL also outperforms all the listed prior VLPs with much less pre-training data and smaller model size.

Model	#Param	Data	Acc.
Prior VLPs			
VL-T5 [5]	224M	-	74.60
LXMERT [42]	240M	-	74.90
VLMixer [45]	-	4M	75.28
ViLT [20]	87M	4M	75.70
PixelBERT [14]	114M	-	76.73
SOHO [13]	-	-	76.37
UNITER [4]	300M	4M	77.18
ViCHA [37]	-	-	77.27
ViLBERT [29]	274M	3.3M	77.40
Oscar [26]	-	4.1M	78.07
VILLA [10]	-	4M	78.39
VinVL* [53]	112M	3.17M	78.54
Ours			
GIVL	112M	3.17M	**79.03**
GIVL (900K)	112M	3.17M	**79.87**

Table 6. Results on NLVR2 test-dev set.

Image captioning is a classic task to evaluate the performance of VLPs. As illustrated in Table 7, GIVL shows comparable performance to prior VLPs in different evalua-

Model	#Param	Data	BLEU@4	CIDEr	METEOR	SPICE
Prior VLPs						
VL-T5 [5]	224M	-	-	116.5	-	-
BUTD [1]	-	-	36.3	120.1	27.7	21.4
VLP [54]	-	-	39.5	129.8	29.3	22.4
Unimo-Large [25]	300M	-	39.6	127.7	-	-
Oscar [26]	-	4.1M	40.5	137.6	29.7	22.8
CLIP-ViL [36]	178M	-	40.2	134.2	29.7	23.8
SimVLM-base [46]	-	1.8B	39	134.8	32.9	24
VinVL* [53]	112M	3.17M	39.6	136.5	30.4	24.4
Ours						
GIVL	112M	3.17M	39.6	135.1	30.3	24.3

Table 7. Results on COCO captioning.

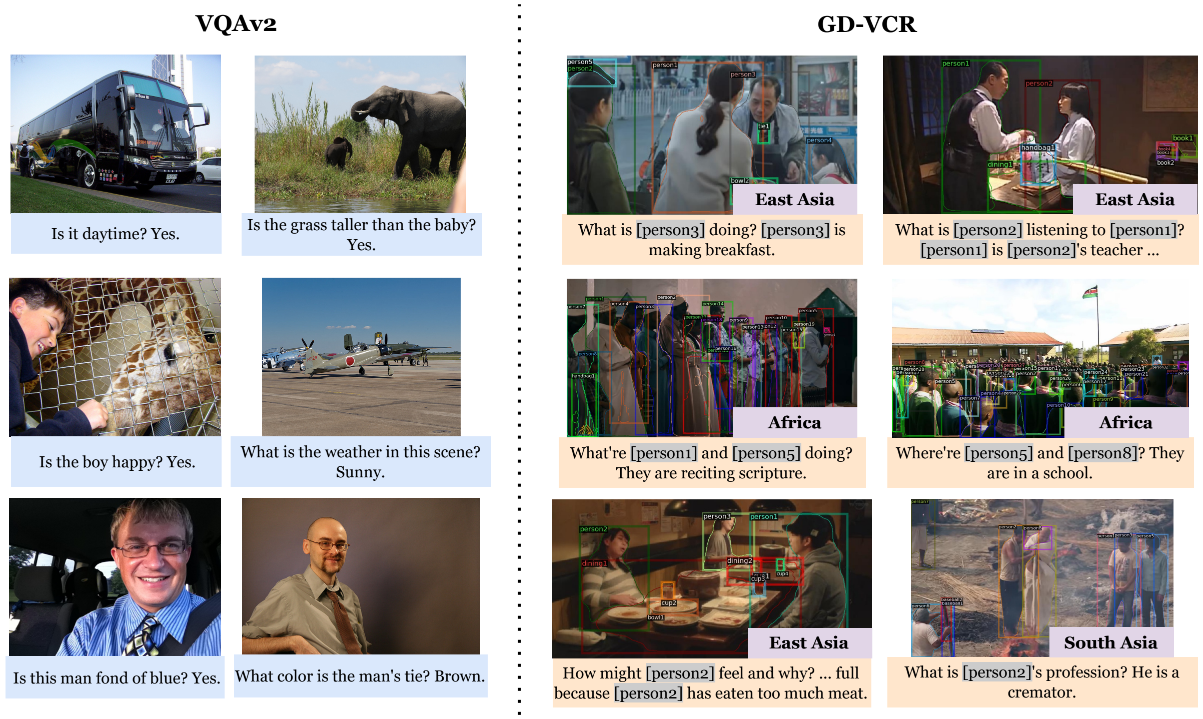

VQAv2
GD-VCR
Is it daytime? Yes.
Is the grass taller than the baby? Yes.
East Asia
What is [person3] doing? [person3] is making breakfast.
East Asia
What is [person2] listening to [person1]? [person1] is [person2]'s teacher ...
Is the boy happy? Yes.
What is the weather in this scene? Sunny.
Africa
What're [person1] and [person5] doing? They are reciting scripture.
Africa
Where're [person5] and [person8]? They are in a school.
Is this man fond of blue? Yes.
What color is the man's tie? Brown.
East Asia
How might [person2] feel and why? ... full because [person2] has eaten too much meat.
South Asia
What is [person2]'s profession? He is a cremator.

Figure 8. Comparison between VQAv2 and GD-VCR's images and corresponding question-answer pairs.

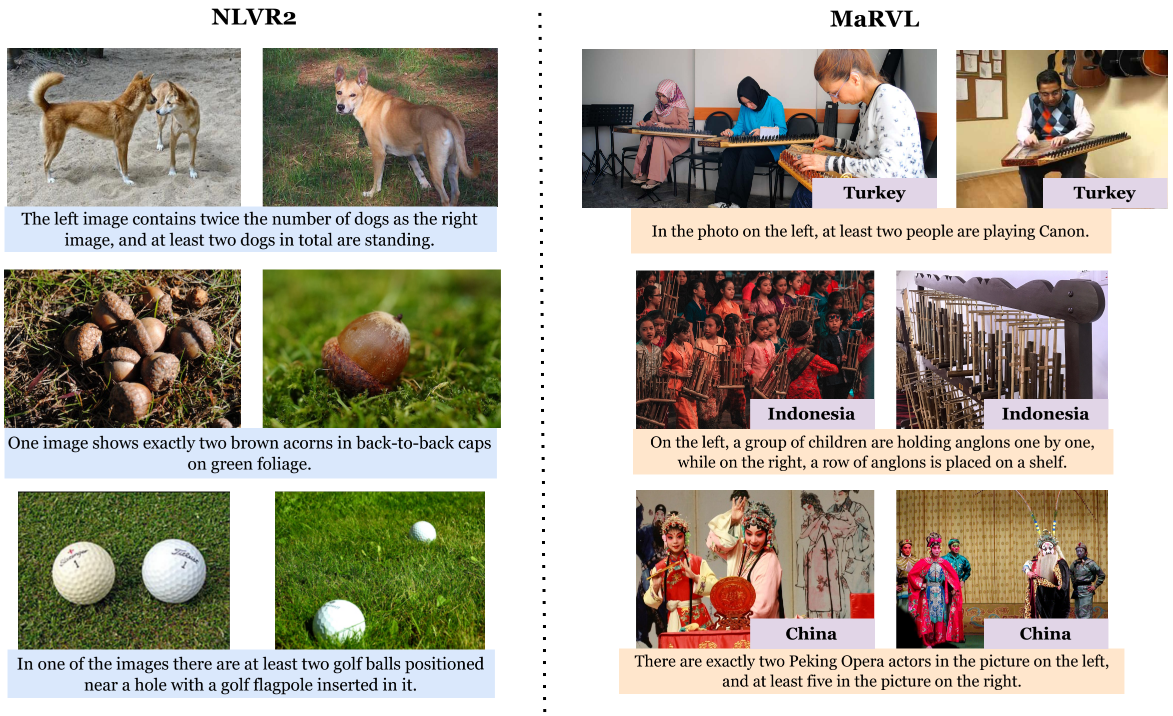
tion metrics. Most of the prior image captioning VLPs use much more data than GIVL, for example, SimVLM-base. All three experiments above demonstrate the effectiveness and data efficiency of GIVL.

C. Qualitative Examples

C.1. Common v.s. Geo-Diverse V&L Tasks

Since geo-diverse Vision-Language (V&L) tasks are not widely studied in Computer Vision (CV) community, it may not be intuitive enough for the audience to understand the differences between common V&L tasks and geo-diverse V&L tasks. In this section, we use some examples to illustrate it.

Before discussing the examples, we would like to introduce the setting of geo-diverse V&L tasks. First, geo-diverse V&L tasks, such as GD-VCR, only use images that are collected from different regions and cultures. It ensures that the visual concepts behind the images are highly relevant to the background regions and cultures. Second, these geo-diverse datasets require annotators from different regions and cultures to label the data, which further imposes the geo-diversity on them. Third and most importantly, questions or text descriptions in geo-diverse datasets



The left image contains twice the number of dogs as the right image, and at least two dogs in total are standing.

One image shows exactly two brown acorns in back-to-back caps on green foliage.

In one of the images there are at least two golf balls positioned near a hole with a golf flagpole inserted in it.

In the photo on the left, at least two people are playing Canon.

On the left, a group of children are holding anglons one by one, while on the right, a row of anglons is placed on a shelf.

There are exactly two Peking Opera actors in the picture on the left, and at least five in the picture on the right.

Figure 9. Comparison between NLVR2 and MaRVL's images and claims.

focus more on the visual concepts from different regions and cultures and their corresponding knowledge.

Figure 8 shows some image-question pairs from both the VQA and GD-VCR datasets. The VQA dataset contains questions that ask for generic visual concepts, such as colors, weather, size, *etc*. The visual information within the images of VQA dataset is sufficient to answer the questions. On the other hand, GD-VCR asks questions that require background knowledge from regions and cultures around the world. For example, the first example on the right-hand side describes a scenario where a person is making breakfast on a busy street. This is not a common occurrence in most Western countries, but it is very common in most East and South Asian regions.

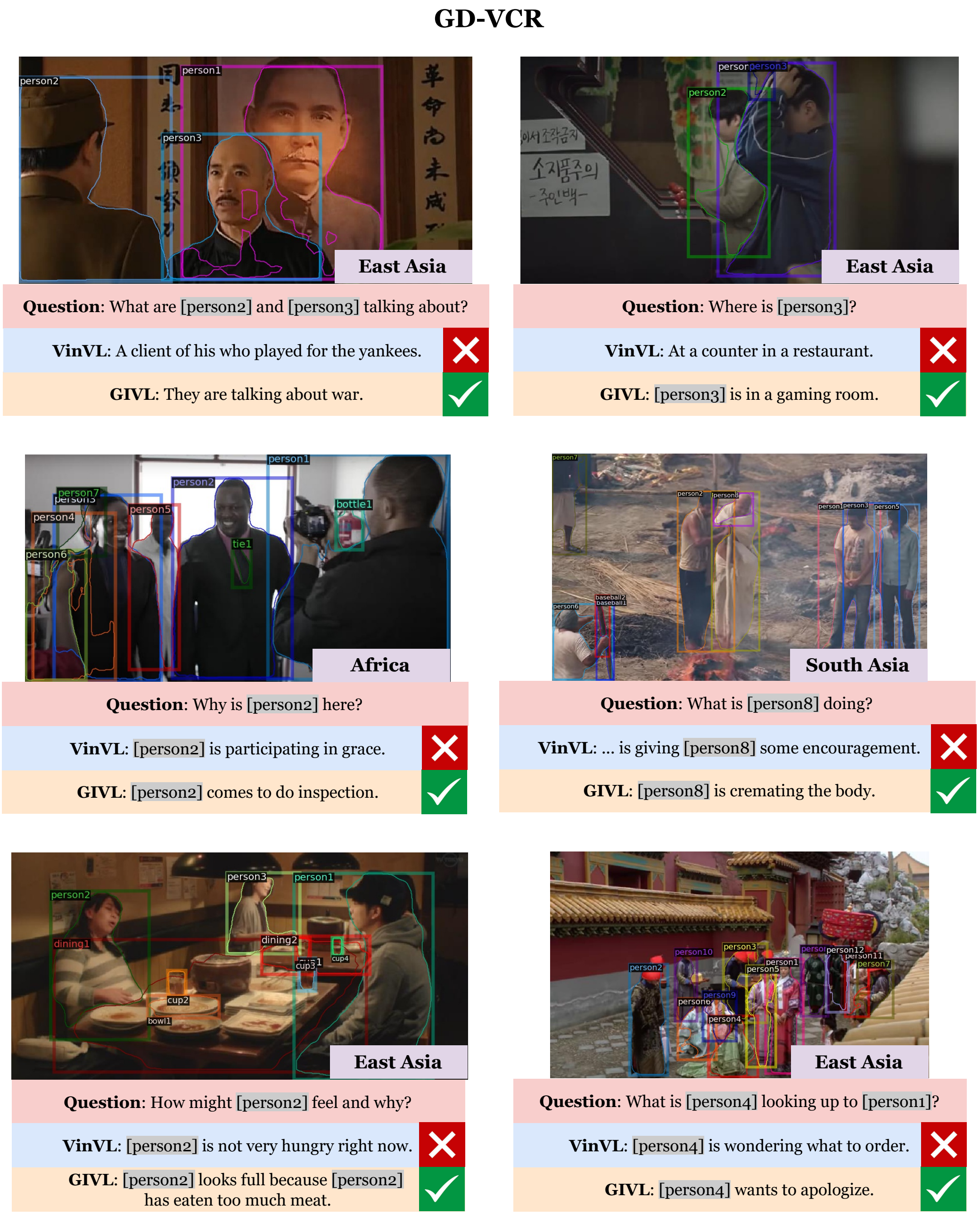
C.2. Empirical Analysis of GIVL's Performances

The comparison between VQA and GD-VCR also can indicate the reasons why GIVL has similar performances with other SOTAs on common V&L tasks but beats all baselines on geo-diverse tasks by a large margin. For common V&L tasks, although some images are collected around the world, they are not geo-diverse. Regardless of the geo-diverse factors in the image, the tasks only involve common visual concepts and their basic visual information. For instance, as shown in Figure 8, the second image-question pair in the VQA examples only asks for the size information of elephants in the image. But the question doesn't ask for the implicit corresponding knowledge of tropical visual concepts. To this end, on common V&L tasks, GIVL may not be able to outperform VLPs that are pre-trained with much greater V&L pre-training corpus mainly covering common visual concepts.

On the other hand, geo-diverse V&L tasks such as GD-VCR and MaRVL, require models to complete the tasks with knowledge that is related to the background regions and cultures of the images. As shown in the right hand side of Figure 9, the model needs to recognize geo-diverse visual concepts and leverage cultural knowledge beyond the image contents to make predictions. Since prior VLPs are not pre-trained to understand the underlying knowledge of geo-diverse visual concepts, GIVL can outperform the majority of SOTA VLPs on geo-diverse V&L tasks.

C.3. Case Study of GD-VCR and MaRVL

We show some cases of GD-VCR and the predictions made by GIVL and VinVL in Figure 10. VinVL is not able to solve some cases in GD-VCR while GIVL can reach the correct answers. In most shown cases, VinVL predictions do not make sense. These cases, such as the bottom-right example, are highly culturally related. People in that image wear ancient Chinese royal dress. The posture seems like they are lining up and half-squatting. In ancient China, it is a royal code for apology. More cases of MaRVL and the predictions of GIVL and VinVL are shown in Figure 11.


GD-VCR

Question: What are [person2] and [person3] talking about?

VinVL: A client of his who played for the yankees. ✗

GIVL: They are talking about war. ✓

Question: Where is [person3]?

VinVL: At a counter in a restaurant. ✗

GIVL: [person3] is in a gaming room. ✓

Question: Why is [person2] here?

VinVL: [person2] is participating in grace. ✗

GIVL: [person2] comes to do inspection. ✓

Question: What is [person8] doing?

VinVL: ... is giving [person8] some encouragement. ✗

GIVL: [person8] is cremating the body. ✓

Question: How might [person2] feel and why?

VinVL: [person2] is not very hungry right now. ✗

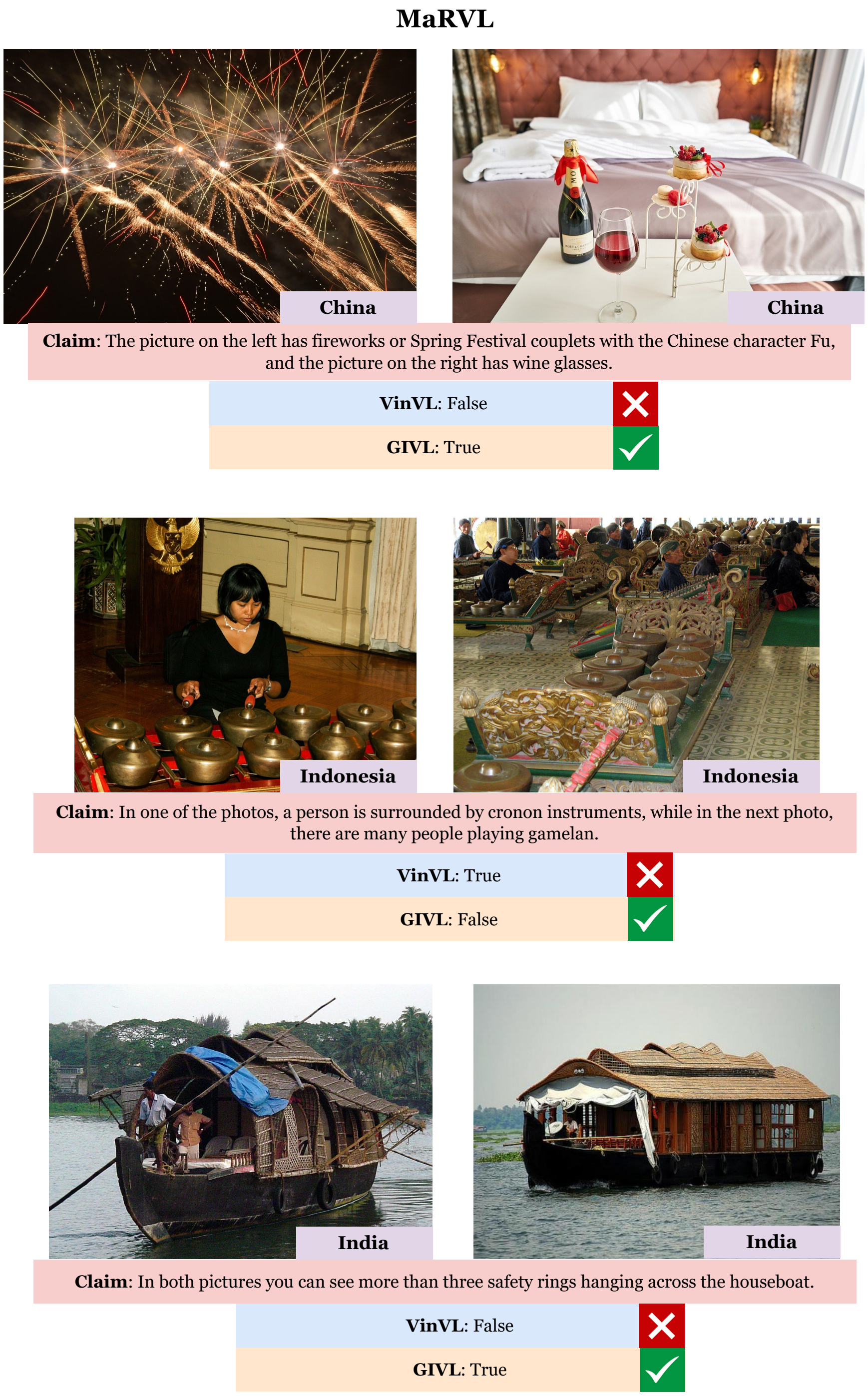
GIVL: [person2] looks full because [person2] has eaten too much meat. ✓

Question: What is [person4] looking up to [person1]?

VinVL: [person4] is wondering what to order. ✗

GIVL: [person4] wants to apologize. ✓

Figure 10. Case study of GD-VCR.



Claim: The picture on the left has fireworks or Spring Festival couplets with the Chinese character Fu, and the picture on the right has wine glasses.

VinVL: False ✗

GIVL: True ✓

Claim: In one of the photos, a person is surrounded by cronon instruments, while in the next photo, there are many people playing gamelan.

VinVL: True ✗

GIVL: False ✓

Claim: In both pictures you can see more than three safety rings hanging across the houseboat.

VinVL: False ✗

GIVL: True ✓

Figure 11. Case study of MaRVL.