Broaden the Vision: Geo-Diverse Visual Commonsense Reasoning

Da Yin Liunian Harold Li Ziniu Hu Nanyun Peng Kai-Wei Chang
Computer Science Department, University of California, Los Angeles
{da.yin,liunian.harold.li,bull,violetpeng,kwchang}@cs.ucla.edu
gd-vcr.github.io

Abstract

Commonsense is defined as the knowledge that is shared by everyone. However, certain types of commonsense knowledge are correlated with culture and geographic locations and they are only shared locally. For example, the scenarios of wedding ceremonies vary across regions due to different customs influenced by historical and religious factors. Such regional characteristics, however, are generally omitted in prior work. In this paper, we construct a **G**eo-**D**iverse **V**isual **C**ommonsense **R**easoning dataset (**GD-VCR**) to test vision-and-language models' ability to understand cultural and geo-location-specific commonsense. In particular, we study two state-of-the-art Vision-and-Language models, VisualBERT and ViLBERT trained on VCR, a standard multimodal commonsense benchmark with images primarily from Western regions. We then evaluate how well the trained models can generalize to answering the questions in **GD-VCR**. We find that the performance of both models for non-Western regions including East Asia, South Asia, and Africa is significantly lower than that for Western region. We analyze the reasons behind the performance disparity and find that the performance gap is larger on QA pairs that: 1) are concerned with culture-related scenarios, e.g., weddings, religious activities, and festivals; 2) require high-level geo-diverse commonsense reasoning rather than low-order perception and recognition. Dataset and code are released at https://github.com/WadeYin9712/GD-VCR.

1 Introduction

Commonsense reasoning endows machines with high-level reasoning ability to understand situations with implicit commonsense knowledge. Suppose that there is a scene where a woman is wearing a bridal gown at a party. An ideal AI system with commonsense knowledge should be able to infer that this woman is attending a wedding and likely to be the bride.

Recently, the field of commonsense reasoning is progressing with the development of large-scale benchmark datasets (Zellers et al., 2018; Talmor et al., 2019), intended to cover a wide range of commonsense knowledge, such as physical interactions (Bisk et al., 2020), social conventions (Sap et al., 2019), and commonsense grounded in vision (Zellers et al., 2019).

However, existing benchmarks are often composed by data from sources in certain regions (e.g., Western movies) and overlook the differences across groups in different regions[1] due to factors including cultural differences. In the aforementioned wedding example, while brides are usually in white in Western weddings, they often wear red and their faces are covered with a red cloth in traditional Chinese weddings. If a model is unaware of regional characteristics or incapable of capturing the nuanced regional characteristics, it leads to a disparity in performance across different regions (Acharya et al., 2020). This motivates us to study how well a model trained on existing commonsense annotations can generalize to tasks requiring commonsense beyond Western regions.

In this paper, we mainly focus on regional commonsense with *visual scenes*. As shown in Figure 1, the three images all describe a wedding but the dresses of the grooms and brides are different, reflecting the regional characteristics of the wedding scenario. In this paper, we introduce a new *evaluation* benchmark, **G**eo-**D**iverse **V**isual **C**ommonsense **R**easoning (**GD-VCR**), following the settings of the visual commonsense reasoning (VCR) task (Zellers et al., 2019). VCR consists of multiple-choice questions paired with images

[1] Due to resource constraints, we use regions as a proxy to evaluate commonsense among different groups. We note that groups of individuals in the same region may have different beliefs, cultures, and behaviors. Please see discussion in the section of Ethical Considerations.

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 2115–2129
November 7–11, 2021.

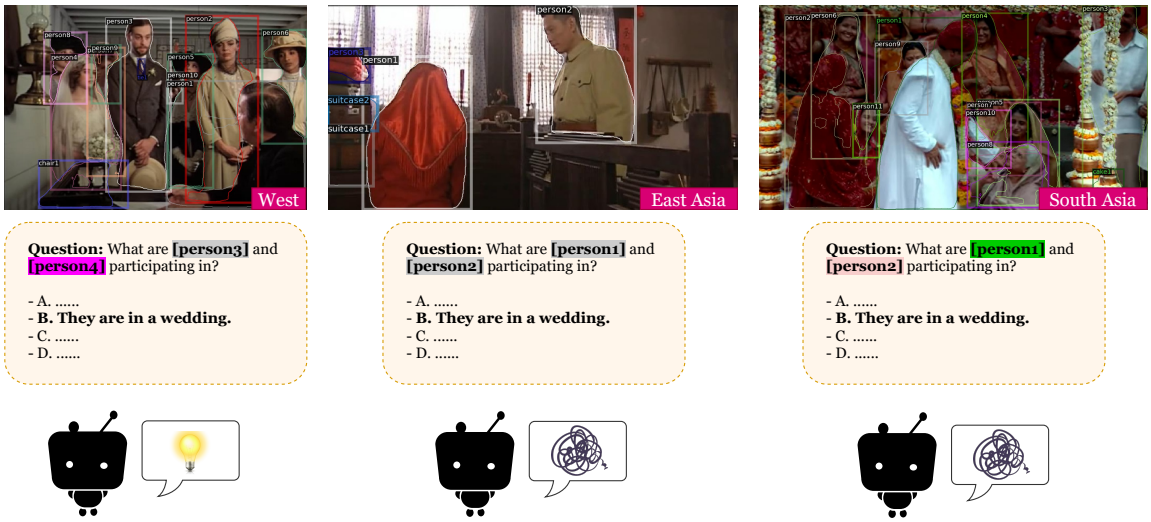


Figure 1: Examples in GD-VCR. The three images are all about weddings but from different regions (left-to-right order: Western, East Asian, South Asian). Current Vision-and-Language models perform well on answering questions about Western weddings but often make mistakes when encountering wedding scenarios in other regions.

extracted from movies or TV series primarily in *Western* regions. GD-VCR includes 328 images, which are mainly sourced from movies and TV series in East Asian, South Asian, African, and Western countries. The images are paired with 886 QA pairs, which need to be answered with geo-diverse commonsense and thorough understanding of the images. An example is given in Figure 1. GD-VCR benchmark addresses geo-diverse commonsense, such as "*What are* ***[person1]*** *and* ***[person2]*** *participating?*". With the help of these questions, it can manifest how models behave differently and reveal potential issues with geographical bias in commonsense reasoning. GD-VCR is one of the first benchmarks to evaluate model's reasoning ability on the task which requires geo-specific commonsense knowledge.

We first study *how well a model trained on VCR can generalize to questions involving geo-specific commonsense*. Experimenting with two pre-trained vision-and-language (V&L) models, VisualBERT (Li et al., 2019) and ViLBERT (Lu et al., 2019), we observe that two models achieve 64%-68% accuracy over the QA pairs on images from Western regions, while their accuracy on images from East Asian region ranges around 45%-51%. The significant performance disparity suggests that the commonsense learned in these models cannot be generalized well across different regions.

We further investigate *the reasons why the model exhibits such disparity* based on the results of VisualBERT. We first find that the performance gap on the images from Western and non-Western regions is large on the scenarios involving regional characteristics, such as weddings, religion and festivals. We also discover that the disparity is related to the reasoning difficulty of QA pairs. On the QA pairs only requiring basic visual recognition, e.g., "*What's* ***[person3]*** *wearing?* ***[person3]*** *is wearing a suit.*", the model achieves relatively similar performance over the four regions; however, the gap enlarges when the questions involve higher-level reasoning with commonsense and rich visual contexts.

By presenting the GD-VCR benchmark, we call upon the researchers to empower AI systems with geo-diverse commonsense such that they are capable of conducting high-level reasoning on data from different regions.

2 Related Work

Commonsense Reasoning Benchmarks. Recently, there has been an emergence of commonsense reasoning benchmarks (Zellers et al., 2019; Talmor et al., 2019; Sap et al., 2019; Zhou et al., 2019; Huang et al., 2019; Bhagavatula et al., 2020; Bisk et al., 2020), which cover a great variety of commonsense knowledge including visual, social, physical, and temporal commonsense. However, these commonsense benchmarks are mostly constructed by annotators from certain regions (e.g., the US and UK) using specific languages (e.g., English). XCOPA (Ponti et al., 2020) and X-CSR (Lin et al., 2021) are two multilingual benchmarks, but

most samples in both benchmarks are simply translated from English and cannot reflect the regional characteristics. Different from previous benchmarks, GD-VCR focuses on geo-diverse commonsense instead of viewing commonsense as a universal monolith.

Vision-and-Language Tasks. A long line of research seeks to build vision-and-language datasets that test a model's ability to understand the visual world and how it is grounded in natural language. The tasks take on various forms, such as phrase grounding (Kazemzadeh et al., 2014; Plummer et al., 2015), visual question answering (Antol et al., 2015; Goyal et al., 2017), and visual reasoning (Zellers et al., 2019; Suhr et al., 2019). To solve these tasks, a wide range of visual grounding skills are required. However, in existing tasks, little consideration is taken into reasoning on the images with regional characteristics.

Geographic Bias. Geographic bias is a serious issue that may cause harmful effects on certain groups of people. In computer vision, researchers (Shankar et al., 2017; de Vries et al., 2019) find that most images from two large-scale image datasets, ImageNet (Deng et al., 2009) and OpenImages (Krasin et al., 2017), are amerocentric and eurocentric. When a model trained on these datasets is applied to images from other regions, the performance will drop drastically. There also exists geographic bias in language technology. For example, it underlies natural language processing (Blodgett et al., 2016; Jurgens et al., 2017; Ghosh et al., 2021) and automatic speech recognition (Tatman, 2017; Koenecke et al., 2020) models. Our work seeks to reveal and test the geographic bias in the visual commonsense reasoning task and models.

3 Benchmark Construction

To build a geo-diverse visual commonsense reasoning benchmark, we design a three-stage annotation pipeline, following the original VCR dataset. 1) We first ask annotators to collect images from movies and TV series in Western, East Asian, South Asian, and African countries. 2) We request annotators to design questions and write down the right answers according to the collected images. 3) We generate answer candidates automatically and formulate multiple-choice questions. The overview of our pipeline is illustrated in Figure 2. We elaborate on the three stages in the following.

3.1 Image Collection

In the image collection stage, we request annotators to follow two principles:

Images with Regional Characteristics. In our annotation instruction, we require that the collected images should have representative scenarios containing cultural elements of the annotators' regions. We further recommend annotators choose scenarios that are ubiquitous but have specific characteristics across regions, e.g., wedding, funeral, festival, religion events, etc. For the purpose of analysis, during image collection, annotators are required to write down keywords on each of their collected images for categorization. For example, the keywords of the middle image in Figure 1 are labeled as "*wedding, soldier, bride, groom, couple, war, countryside*".

Sources of Images. Follow the settings of the original VCR dataset, we ask annotators to select diverse and informative images by taking screenshots from movies, TV series, documentaries, and movie trailers from websites including Youtube[2], BiliBili[3], IQIYI[4], etc. These videos usually include various scenarios and rich contents containing a large amount of actions and human interactions. Note that our collected images from Western regions share the same source[5] with those in the original VCR development set. We use them as a control set in the experiments. Details are in Appendix A.1.

3.2 QA Pair Annotation

We recruit another batch of annotators who are familiar with the culture in one of the four regions to annotate QA pairs upon the collected images in English. The annotation stage is divided into two parts: 1) designing questions according to the image contents; 2) annotating the correct answers of the questions.

Following the pre-processing of the original VCR dataset, we first apply the Mask R-CNN object detector[6] to mark bounding boxes of objects in each image, and the annotators can use the labels (e.g., person, car, bowl, etc.) to design QA pairs.

[2] www.youtube.com/
[3] www.bilibili.com/
[4] www.iqiyi.com/
[5] www.youtube.com/c/MOVIECLIPS/videos
[6] github.com/facebookresearch/detectron2. COCO-pretrained Mask R-CNN.

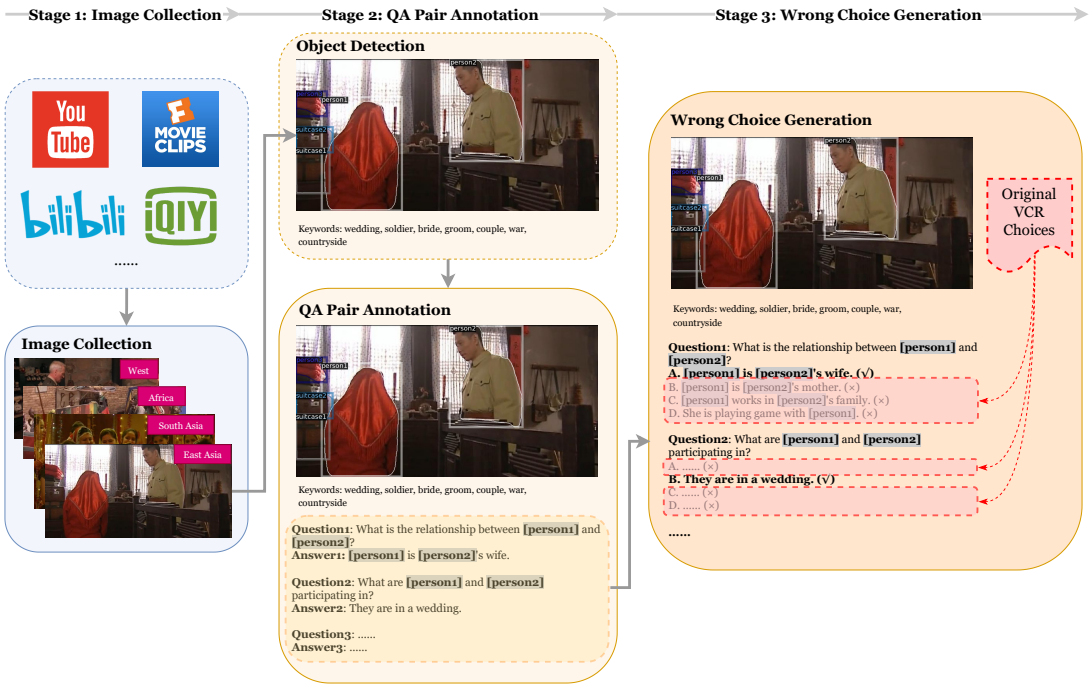


Figure 2: Overall annotation pipeline. It is divided into three stages: Stage 1 is to collect images with regional characteristics; Stage 2 is to design QA pairs based on the detected objects; Stage 3 is to generate answer candidates to complete the dataset with the help of the answer choices in the original VCR development set.

3.2.1 Designing Questions

Annotators are asked to design questions based on the following three instructions.

Usage of the Detected Objects. Annotators are requested to choose the named objects in the bounding boxes to construct questions. As shown in Figure 1, annotators can design questions such as "*What is the relationship between* ***[person1]*** *and* ***[person2]*** *?*". This requirement is aligned with the question design in the original VCR dataset.

High-Order Cognitive Questions. Following the original VCR dataset, we ask annotators to design *high-order* cognitive questions which require geo-specific commonsense knowledge and visual understanding to be answered. Take the rightmost image in Figure 1 as an example. "*Why is* ***[person11]*** *so happy?*" is a *qualified* question because people have to observe the surroundings including **[person1]** and **[person2]** 's wearing and others' facial expression, and conclude that it is a wedding. Moreover, **[person1]** is wearing a wedding dress and others are celebrating for her. Thus, people can infer that **[person11]** is happy because it is **[person1]** and **[person2]** 's wedding. Overall, to answer this question, we need to combine the image context and commonsense knowledge, and reason with multiple turns. A *disqualified* example of question is "*What is* ***[person3]*** *wearing?*" in the left image of Figure 1. It is defined as a *low-order* cognitive question because it can be directly answered by recognizing that the woman is wearing a suit. This type of question does not need commonsense reasoning based on the context information.

Question Templates. Since models trained on the original VCR dataset will be evaluated on GD-VCR dataset, we attempt to eliminate the discrepancy of questions used between the original VCR dev set and GD-VCR to mitigate the effect of different question formats. Hence, we ask annotators to design questions by referring to the templates summarized from the original VCR development set. To generate the templates, we first replace nouns, verbs, adjectives, and adverbs of the questions in VCR development set with their POS tags (e.g., NN, VB, JJ, etc.) labeled by NLTK[7], while keeping question words such as "*what*", "*why*" and "*how*", and auxiliary verbs like "*is*", "*do*" and "*have*". In this way, we remove the terms associated with specific questions, while keeping the general patterns.

[7] www.nltk.org

We then apply K-Means (MacQueen et al., 1967) algorithm[8] to the question patterns, and manually summarize 17 templates, e.g., "*What (n.) is sb. v.+ing sth.?*", "*What is sb.'s relationship with sb.?*". Details of the clustering method and the list of question templates are in Appendix A.2.

3.2.2 Annotating Correct Answers

Annotators are required to write down *only* the *right* answer for the questions they designed. This is to reduce annotation cost and avoid potential annotation artifacts (Zellers et al., 2019). We require that the right answers to the questions should be consistent with the image contents. However, we remind annotators to avoid writing answers that are too specific to the video plots because the answers should be inferred without prior knowledge about the plots. In addition, instead of writing named entities or proper names specific to one region, annotators are required to use common expressions in English. These instructions would help us maintain the difficulty of GD-VCR to a proper extent. Finally, we invite 3 annotators from each QA pair's corresponding regions to validate the correctness of each question and its answer. If 2 of them have an agreement to approve a certain QA pair, we keep it in the final dataset.

3.3 Answer Candidate Generation

In this stage, for each question, we generate three wrong answer candidates (i.e., wrong answer choices), to construct a 4-way multiple-choice QA example. We follow the answer candidate generation algorithm in VCR (Zellers et al., 2019):

Answer Candidate Pool. Instead of generating answer candidates from scratch by language models, we leverage the right choices in the original VCR development set, and treat them as an answer candidate pool. All the answer candidates of GD-VCR are derived from this pool.

Answer Candidate Selection. The principles for selecting answer candidates from the pool are two-fold: 1) answer candidates should be relevant to questions; 2) they should be dissimilar with the right choice and other selected answer candidates, so that they would not be the right answer incidentally. Details of the candidate selection algorithm are in Appendix B.2.

[8] We concatenate the question words and the POS tags of all the other words (e.g., from "*What is [person1] doing?*" to "*What VBZ NNP VBG?*"). Then we use sentence representations of the converted sentences as real-valued vectors in K-Means. The representations are obtained from Sentence-Transformers (Reimers and Gurevych, 2019) based on RoBERTa-base (Liu et al., 2019).

3.4 Dataset Statistics

Table 1 summarizes the statistics of the GD-VCR benchmark and the original VCR development set.

Texts. We observe that the average lengths of questions and answers in GD-VCR are similar to those in the original VCR development set. Aside from the text lengths, we also consider out-of-vocabulary (OOV) rate with respect to the original VCR training set. This indicates how much unseen knowledge (e.g., entities specific to certain region) are involved in GD-VCR. As shown in Table 1, we find that the OOV rate of the entire GD-VCR dataset is 6.75%, while that of the original VCR development set is 12.70%. This shows that GD-VCR has a similar distribution of the vocabulary with the original VCR dataset and the difficulty of GD-VCR does not come from the vocabulary gap.

Images. The average number of the detected objects 10.18 is similar to that of the original VCR development set. Moreover, since the objects mentioned in questions and answers are directly relevant to the reasoning process on each QA pair, we consider statistics of the average number of the relevant objects. The average number of relevant objects in image in GD-VCR is 2.38, which nearly equals that of the VCR development set.

4 Model Performance and Human Evaluation over GD-VCR

We are interested in the following questions: 1) Can a model trained on the original VCR dataset (mostly Western scenarios) generalize well to solve reasoning questions require commonsense specific to other regions? 2) Do humans show similar trend when dealing with questions require regional commonsense that they are not familiar with?

Our experiments are conducted with two Vision-and-Language models VisualBERT (Li et al., 2019) and ViLBERT (Lu et al., 2019). We fine-tuned the two pre-trained models on the original VCR training set and evaluate them on GD-VCR. All the experimental results are the average of 3 runs with different seeds. Implementation details are listed in Appendix C.

Datasets	# Images	# QA Pairs	Avg. Len. of Ques.	Avg. Len. of Ans.	Avg. # Obj.	Avg. # Relevant Obj.	OOV Rate
Original VCR	9929	26534	6.77	7.67	10.34	2.39	12.70%
GD-VCR	328	886	7.38	7.68	10.18	2.38	6.75%
◦ **West**	100	275	7.36	7.19	11.10	2.28	3.44%
◦ **East Asia**	101	282	7.59	7.59	9.57	2.42	4.50%
◦ **South Asia**	87	221	6.85	8.00	10.29	2.12	5.49%
◦ **Africa**	40	108	7.98	8.54	9.29	3.03	7.34%

Table 1: Statistics of the GD-VCR benchmark. The top half of the table is the overall statistics of GD-VCR and the original VCR development set. The bottom half includes the subsets of each region in GD-VCR.

Datasets	Human	Text-only BERT	VisualBERT		ViLBERT	
			Acc.	Gap (West)	Acc.	Gap (West)
Original VCR	-	53.8*	70.10	+5.73	69.84	+2.57
GD-VCR	88.84	35.33	53.95	-10.42	59.99	-7.28
◦ **West**	91.23	37.09	64.37	0.00	67.27	0.00
◦ **South Asia**	92.98	33.48	54.90	-9.43	63.57	-3.70
◦ **Africa**	87.93	34.26	47.53	-16.84	59.73	-7.54
◦ **East Asia**	83.05	35.46	45.51	-18.86	50.18	-17.09

Table 2: Accuracy (%) on the subset of each region in GD-VCR and the original VCR development set. With regard to Western regions, two models' performance gap of the original VCR development set and other regions is shown. We also report human's accuracy (%) over each region subset. Annotators are from United Kingdom and United States according to MTurk. * denotes the reported result in Zellers et al. (2019).

4.1 Model Performance

We apply the two models on GD-VCR benchmark to study how well the two models can generalize. Results are shown in Table 2. Key observations are summarized as follows:

Western vs. Non-Western Regions. We find that the models perform significantly worse on the images from non-Western regions. According to VisualBERT's results, we observe that the gap between Western and South Asian regions is 9.43%, while it greatly amplifies to 16.84% and 18.86% when it comes to the comparison with African and East Asian regions, respectively. These results reflect significant differences of models' reasoning ability on the examples from different regions.

VisualBERT vs. ViLBERT. We find that ViLBERT outperforms VisualBERT by 6.04% on GD-VCR. We conjecture that the higher performance of ViLBERT partly results from the pre-training data: VisualBERT is pre-trained on COCO Captions which includes 80K images (Chen et al., 2015), while ViLBERT's pre-training data are from Conceptual Captions containing 3.3M images (Sharma et al., 2018). The larger coverage of image contents may help ViLBERT generalize to the images with regional characteristics. It is also shown that the performance gap over the images from Western and non-Western regions shrinks when applying ViLBERT. However, the gap is still significant, ranging from 3.70% to 17.09%.

Western v.s. Original VCR Dataset. We observe a performance gap around 2%-6% between images from Western and the original VCR dataset. We speculate that the gap is caused by one main aspect: the requirements in the image collection stage are slightly different. We expect to collect images containing regional characteristics, including cultural elements like customs. It may add to the complexity of the reasoning process as cultural commonsense is needed. However, the gap is much smaller compared with the gap between Western and other regions.

4.2 Human Evaluation

Apart from the model performance, we investigate how well human beings perform on GD-VCR. We randomly select 40 QA pairs of each region, and there are 160 QA pairs in total for evaluation. We recruit qualified annotators living in United Kingdom and United States from MTurk[9] to accomplish the evaluation. Assuming them to be familiar with

[9]The annotators should complete at least 1000 HITs, with an approval rate above 95%.

Western culture, we are interested to see their performance on the examples from other regions.

Human evaluation results are shown in Table 2. We notice that human performance is much better than models. More importantly, we observe that the performance gap among regions is much smaller than that of two V&L models. For example, annotators from Western can achieve 87.93% accuracy on East Asian images, and the gap reduces to 8.18% from 18.86% and 17.09%. It implies that human beings are more capable of applying their commonsense knowledge and transferring it to the comprehension in geo-diverse settings, while models are still far away from this level.

5 Analyses of Performance Disparity

As we observe large performance gaps between Western and non-Western data in Section 4.1, in this section, we inspect the pre-trained V&L model to analyze the reasons behind such performance disparity on two aspects, 1) *regional differences of scenarios* and 2) *reasoning level of QA pairs*. We analyze the *VisualBERT* model, since its performance gap is more evident.

5.1 Regional Differences of Scenarios

As shown in Figure 1, even the same scenarios such as wedding can take different visual forms across regions. Motivated by this, we investigate how large the performance gap is when we apply VisualBERT to the images of the same scenarios happening in different regions.

We select the scenarios that frequently appear in the annotated keyword set of GD-VCR. Specifically, we choose the scenarios which appear at least 10 times in *not only* Western images, *but also* the images from any of the other regions. We visualize each scenario's performance gap between the images from Western and non-Western regions in Figure 3. The scenarios whose gap is above 8% are colored in **red**; otherwise, they are labeled by **blue**.

As shown in Figure 3, we find that on the scenarios which often contain regional characteristics (e.g., wedding, religion, festival), the performance gap is much larger, from 8.28% to 23.69%. One interesting finding is that, aside from festival, wedding and religion, which are generally considered to be different across regions, the gap is considerably large over the scenarios involving customers. We speculate that it is also due to regional characteristics. As shown in Figure 5 of Appendix F, in East

Figure 3: Visualization of the performance gap on images of the same scenarios in Western and non-Western regions. The larger the characters, the larger the performance gap over the scenarios. The **red** and **blue** words are the scenarios whose performance gap is *above* and *below* 8%, respectively. Detailed accuracy on these scenarios is shown in Appendix D.

Asia and South Asia, many customers would buy things from the local merchants along the streets, while in Western regions, customers typically shop in supermarkets and restaurants. The visual discrepancy may result in errors on the "customer" scenarios in GD-VCR.

On the other hand, for the scenarios such as party, restaurant and student, the gap is only 0.42%, 1.29% and 1.12%, respectively. We notice that these scenarios are more similar across regions. For example, parties are related to actions like drinking, dancing, and celebration, which are common and take on similar visual forms across regions. Such similarity may contribute to model's high transfer performance on "party".

5.2 Reasoning Level of QA Pairs

The QA pairs in GD-VCR are *high-order* cognitive QA pairs, which require several reasoning steps to be solved. For example, to infer that "***[person1]*** *and* ***[person2]*** *are in a wedding*" in the middle image of Figure 1, human beings must first recognize basic facts such as **[person1]** is wearing in red and her face is covered by a red cloth. Only by combining the recognized facts and regional commonsense can human make correct predictions. Therefore, the model's failure on these high-order cognitive QA pairs from non-Western regions may be attributed to two reasons: 1) the model fails to recognize the basic facts from the image, 2) or the model succeeds on the basic facts but fails eventually due to lack of geo-specific commonsense.

To determine at which stage the model fails to generalize, we aim to answer the following two questions: **Q1.** *Can the model perform similarly on recognizing basic visual information in the images from different regions?* **Q2.** *Is the performance*

Regions	Low-order		High-order		$\lvert Low - High \rvert$
	Acc. (Low)	Gap (West)	Acc. ($High$)	Gap (West)	
West	65.15	0.00	66.60	0.00	1.45
South Asia	54.37	-10.78	52.37	-14.23	2.00
East Asia	58.74	-6.41	50.47	-16.13	8.27
Africa	56.06	-9.09	40.35	-26.25	15.71

Table 3: VisualBERT's accuracy (%) on low-order and high-order cognitive QA pairs. "Gap (West)" denotes performance gap over the QA pairs of images from Western and non-Western regions. "$\lvert Low - High \rvert$" denotes the performance gap between low-order and high-order cognitive QA pairs from the same regions.

disparity attributed to the failure of understanding more advanced or basic visual information?

According to the standard of reasoning level discrimination mentioned in Section 3.2.1, we categorize QA pairs into two types: *low-order* and *high-order* cognitive QA pairs. *Low-order* cognitive QA pairs correspond to the inquiry on basic visual information, while *high-order* QA pairs involve more advanced information. Our analysis is mainly concerned with the two types of QA pairs.

Q1. Can model perform similarly on understanding basic visual information across regions? We evaluate model's performance on *low-order* cognitive QA pairs to analyze this aspect.

As mentioned in Section 3.2.1, GD-VCR is composed of high-order cognitive QA pairs but without low-order pairs. Therefore, we additionally annotate low-order cognitive QA pairs on the images of GD-VCR. Specifically, we randomly select 30 images per region and design low-order QA pairs based on these selected images. Finally, we collect 22 QA pairs on Western images, 26 on East Asian images, 16 on South Asian images, and 22 on African images.

Results are shown in Table 3. We observe that the performance over the low-order cognitive QA pairs is all around 60% for the four regions. Performance over Western images is still the highest among the four regions. But note that the performance gap between the images from Western and non-Western regions is not so large as the overall gap shown in Table 2. For example, the overall performance gap between East Asia and Western is around 19%, but it decreases to 6.41% when the model deals with simpler situations. It demonstrates that, when encountering the QA pairs focusing on simple recognition, VisualBERT can narrow down the gap on the images from different regions. In other words, VisualBERT shows more similar ability to process basic visual information, no matter where the images are from.

Q2. Is the performance disparity attributed to understanding on more advanced or basic visual information? We analyze the performance over *low-order* and *high-order* cognitive QA pairs. For a fair comparison, both types of QA pairs share *the same images*.

Results are shown in Table 3. We observe that VisualBERT's performance over low-order cognitive QA pairs is higher than that over high-order QA pairs on images from East Asia, South Asia, and Africa. Especially, on the images from African regions, the performance gap between these two types of QA pairs is 15.71%.

Furthermore, from Table 3, we notice that the performance gap between Western and non-Western regions on high-order cognitive QA pairs is much larger than that on low-order QA pairs. For the images from East Asian regions, the performance gap with regard to Western regions on low-order pairs is 6.41%. The gap amplifies to 16.13% when VisualBERT is applied to high-order QA pairs. For African images, the gap changes rapidly from 9.09% to 26.25%. These results show that VisualBERT trained on VCR lacks the ability to perform complex reasoning on the scenarios in non-Western regions. We hope our findings could inspire future work to model high-level reasoning process better with geo-diverse commonsense knowledge in commonsense reasoning tasks.

6 Conclusion

We propose a new benchmark, GD-VCR, for evaluating V&L models' reasoning ability on the QA pairs involving geo-diverse commonsense knowledge. Experimental results show that the V&L models cannot generalize well to the images regarding the regional characteristics of non-Western regions. Based on VisualBERT's results, we find that 1) the scenarios such as wedding, religion and

festival, which require geo-diverse commonsense knowledge to be understood, and 2) the reasoning difficulty of QA pairs are highly associated with the performance disparity. For broader impact, we hope that the GD-VCR benchmark could broaden researchers' vision on the scope of commonsense reasoning field and motivate researchers to build better commonsense reasoning systems with more inclusive consideration.

Acknowledgement

We thank Xiao Liu, Ming Zhong, Te-Lin Wu, Masoud Monajatipoor, Nuan Wen, and other members of UCLANLP and UCLA PlusLab groups for their helpful comments. We also greatly appreciate the help of anonymous annotators for their effort into constructing the benchmark. This work was partially supported by DARPA MCS program under Cooperative Agreement N66001-19-2-4032. The views expressed are those of the authors and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

Ethical Considerations

In this work, we propose a geo-diverse visual commonsense reasoning dataset GD-VCR. Since the paper introduces new dataset, we discuss the potential ethical issues about data collection.

Intellectual Property and Privacy Rights. We ensure that intellectual property and privacy rights of the original authors of videos and recruited annotators are respected during the dataset construction process with permission of licence[10][11]. We also claim that the collected data would not be used commercially.

Compensation for Annotators. We recruit annotators from Amazon Mechanical Turk platform[12] and college departments of foreign languages and culture. In image collection stage, we paid annotators \$0.5-0.7 per collected image. In QA pair annotation stage, the payment is \$0.2 per QA pair. For validation and human evaluation, we pay them \$0.02-0.03 per QA pair. The pay rate is determined by a preliminary annotation trial to ensure the average hourly pay rate is around \$12 per hour. The annotations on the images from East Asian regions are partly done by the authors of this work.

Potential Problems. Although we have considered the potential geographic bias in the benchmark construction process, GD-VCR may still contain unwanted bias. First, due to the resource constraints, GD-VCR dataset is unable to cover diverse regional characteristics at once. For instance, we do not take Southeast Asian, Arabic and Latin American regions into account. Moreover, even groups in the same region may have different beliefs. For the regions like Africa, the regional differences between West Africa, East Africa, and North Africa are evident. However, in GD-VCR, the images from Africa are mainly sourced from East Africa. It inevitably introduces geographic bias into our benchmark. More fine-grained analysis should be conducted to scale up this study, especially before the visual commonsense reasoning model is used in the commercial product.

[10]Fair use on YouTube. support.google.com/youtube/answer/9783148?hl=en

[11]Copyright Law of the People's Republic of China (Article 22). http://www.gov.cn/flfg/2010-02/26/content_1544458.htm.

[12]www.mturk.com

References

A. Acharya, Kartik Talamadupula, and Mark A. Finlayson. 2020. An Atlas of Cultural Commonsense for Machine Reasoning. *ArXiv*, abs/2009.05664.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Wen tau Yih, and Yejin Choi. 2020. Abductive Commonsense Reasoning. In *International Conference on Learning Representations*.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2020. PIQA: Reasoning about Physical Commonsense in Natural Language. In *Thirty-Fourth AAAI Conference on Artificial Intelligence*.

Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 2016. Demographic Dialectal Variation in Social Media: A Case Study of African-American English. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1119–1130, Austin, Texas. Association for Computational Linguistics.

Xinlei Chen, H. Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. L. Zitnick. 2015. Microsoft COCO Captions: Data Collection and Evaluation Server. *ArXiv*, abs/1504.00325.

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

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*.

Sayan Ghosh, Dylan Baker, David Jurgens, and Vinodkumar Prabhakaran. 2021. Cross-geographic Bias Detection in Toxicity Modeling. *NAACL 2021*.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the Role of Image Understanding in Visual Question Answering. In *CVPR*.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos QA: Machine Reading Comprehension with Contextual Commonsense Reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2391–2401, Hong Kong, China. Association for Computational Linguistics.

David Jurgens, Yulia Tsvetkov, and Dan Jurafsky. 2017. Incorporating Dialectal Variability for Socially Equitable Language Identification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 51–57, Vancouver, Canada. Association for Computational Linguistics.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. ReferItGame: Referring to Objects in Photographs of Natural Scenes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 787–798, Doha, Qatar. Association for Computational Linguistics.

Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R Rickford, Dan Jurafsky, and Sharad Goel. 2020. Racial Disparities in Automated Speech Recognition. *Proceedings of the National Academy of Sciences*, 117(14):7684–7689.

Ivan Krasin, Tom Duerig, Neil Alldrin, Vittorio Ferrari, Sami Abu-El-Haija, Alina Kuznetsova, Hassan Rom, Jasper Uijlings, Stefan Popov, Andreas Veit, Serge Belongie, Victor Gomes, Abhinav Gupta, Chen Sun, Gal Chechik, David Cai, Zheyun Feng, Dhyanesh Narayanan, and Kevin Murphy. 2017. OpenImages: A Public Dataset for Large-scale Multi-label and Multi-class Image Classification. *Dataset available from https://github.com/openimages*.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. VisualBERT: A Simple and Performant Baseline for Vision and Language. *ArXiv*.

Bill Yuchen Lin, Seyeon Lee, Xiaoyang Qiao, and Xiang Ren. 2021. Common Sense Beyond English: Evaluating and Improving Multilingual Language Models for Commonsense Reasoning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1274–1287, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR*, abs/1907.11692.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. In *NeurIPS*.

James MacQueen et al. 1967. Some Methods for Classification and Analysis of Multivariate Observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297. Oakland, CA, USA.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k Entities: Collecting Region-to-phrase Correspondences for Richer Image-to-sentence Models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A Multilingual Dataset for Causal Commonsense Reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376, Online. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social IQa: Commonsense Reasoning about Social Interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4463–4473, Hong Kong, China. Association for Computational Linguistics.

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

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia. Association for Computational Linguistics.

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. 2019. 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. Association for Computational Linguistics.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge. 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 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Rachael Tatman. 2017. Gender and Dialect Bias in YouTube's Automatic Captions. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 53–59, Valencia, Spain. Association for Computational Linguistics.

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. 2020. 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. Association for Computational Linguistics.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From Recognition to Cognition: Visual Commonsense Reasoning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. SWAG: A Large-Scale Adversarial Dataset for Grounded Commonsense Inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 93–104, Brussels, Belgium. Association for Computational Linguistics.

Ben Zhou, Daniel Khashabi, Qiang Ning, and Dan Roth. 2019. "Going on a vacation" takes longer than "Going for a walk": A Study of Temporal Commonsense Understanding. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3363–3369, Hong Kong, China. Association for Computational Linguistics.

A Additional Details of Annotation Pipeline

A.1 Image Collection

In addition to the requirements mentioned in Section 3.1, we have additional requirements on the contents, quality, and sources of images. The image should have at least two people, and should not be grainy and blurry. We require annotators to choose movies, TV series, documentaries and trailers which are free to access and do not have copyright issues. Together with the images and their keywords, we also collect video names, screenshot timestamps, and the links of videos. It is to help the annotators in later stages better understand the image contents with video contexts.

A.2 QA Pair Annotation

Question Template List. As mentioned in Section 3.2, we recommend annotators to design questions based on question templates. The template list is shown in Table 5. For clustering methods to summarize templates, we use K-Means algorithm to cluster similar question patterns. Specifically, the maximum number of clusters is at most 20 clusters. The algorithm will automatically stop until the 200-th iteration.

Other Annotation Details. To pursue diversity of question types, we require annotators to design questions via different question templates. Besides, we ask annotators to avoid annotating too simple answers, such as "*yes*" and "*no*".

B Details of Answer Candidate Generation Algorithm

B.1 Relevance Model

Relevance model is to evaluate the relevance score between questions and answers. Higher relevance scores indicate that the answers are more relevant with the questions. We train the relevance model based on pre-trained BERT-base parameters (Wolf et al., 2020). Specifically, the training data are all from the *original VCR training set* and composed by relevant and irrelevant QA pairs. The relevant QA pairs are the ones consisting of questions and their corresponding right answers; the irrelevant pairs are the ones consisting of questions and random answers sampled from the whole set of answer choices. We build a binary classifier upon these training data to classify whether an answer is relevant with a question or not. The relevance score is the probability of being relevant pairs.

B.2 Pseudo Code of Answer Candidate Generation Algorithm

Algorithm 1 Answer Candidate Generation Alg.

Input: Question $Q = \{q_1, q_2, ..., q_n\}$, the question's right answer $Corr = \{c_1, c_2, ..., c_n\}$, answer candidate pool $\mathbf{A} = \{A_1, A_2, ..., A_m\}$, relevance model Rel, similarity model Sim. q_i and c_i indicate tokens.
Output: The whole set of four answer choices **ansList** of the given question Q, including one right choice $Corr$ and three answer candidates W_1, W_2, and W_3.

Initialization: **ansList** $\leftarrow \{Corr\}$.
for $t = 1, 2, 3$ **do**
  for each A_i in $\mathbf{A}_{\lfloor (t-1)\times \frac{m}{3} \rfloor + 1 : \lfloor t \times \frac{m}{3} \rfloor}$ **do**
    Initialization: $score \leftarrow 0$, $minscore \leftarrow +\infty$.
    if $\mathrm{Rel}(Q, A_i) \geq 0.9$ **then**
      for each $ansList_i$ in **ansList** **do**
        $similarity \leftarrow \mathrm{Sim}(ansList_i, A_i)$
        if $similarity \leq 0.2$ **then**
          $score \leftarrow score + similarity$
        else
          $score \leftarrow score + 10$
        end if
      end for
      if $score < minscore$ **then**
        $minscore \leftarrow score$
        $W_t \leftarrow A_i$
      end if
    end if
  end for
  ansList $\leftarrow$ **ansList** $\cup \{W_t\}$
end for
return ansList

The pseudo code of the algorithm is shown in Algorithm 1. The two principles for selecting answer candidates are as follows: for each QA pair, 1) they should be relevant with the questions; 2) they should not be similar with the right choices and the selected answer candidates. The model that computes similarity is the "stsb-roberta-base" model (Reimers and Gurevych, 2019) from `github.com/UKPLab/sentence-transformers`.

C Implementation Details of Fine-tuning VisualBERT and ViLBERT

Following VisualBERT (135.07M parameters) configuration on VCR[13], we directly use the model pre-trained on COCO (Chen et al., 2015) and original VCR training set. The experiments of ViLBERT[14] (252.15M parameters) is based on the model pre-trained on Conceptual Captions (Sharma et al.,

[13] `github.com/uclanlp/visualbert`
[14] `github.com/jiasenlu/vilbert_beta`

Regions	Wedding	Festival	Religion	Bride	Groom	Restaurant	Family	Student	Customer	Party
West	62.22	68.89	58.33	66.67	69.14	61.90	59.26	61.54	66.67	55.83
Other Regions	50.00	60.61	46.21	52.78	45.45	60.61	47.27	60.42	44.44	56.25

Table 4: Accuracy (%) on the images involving the same scenarios from different regions.

Question Templates
1. What did sb. (just) v.?
2. What did sb do when/before/as CLAUSE?
3. What (n.) is sb. v.+ing prep. PHRASE?
4. What (n.) is sb. v.+ing?
5. What is sb.'s job/occupation?
6. What is sb.'s relationship with sb.?
7. Why is sb. v.+ing sth. CLAUSE?
8. Why is sb. adj.?
9. Why is sb. here?
10. Why is sb. acting adv.?
11. How does sb. feel/look?
12. Where are sb. (v.+ing)?
13. What will sb v. next/is about to do?
14. What will sb v. when/after CLAUSE?
15. What will sb v. (if) CLAUSE?
16. Where will sb. go?
17. Where was sb. previously?

Table 5: Question template list summarized from the original VCR development set.

2018) and original VCR training set. Both models are then fine-tuned for 8 epochs on 4 NVIDIA GeForce GTX 1080 Ti GPUs, with learning rate $2e-5$. The batch size of VisualBERT and ViLBERT is 32 and 16, and fine-tuning one epoch with VisualBERT and ViLBERT costs 5.28 and 6.75 hours, respectively. For both models, we choose the epoch which performs the best on the original VCR development set among 8 epochs.

D Accuracy on the QA Pairs Involving Specific Scenarios

Table 4 shows VisualBERT's accuracy of the QA pairs involving specific scenarios depicted in Figure 3. Besides the study on GD-VCR, we also make comparison between model performance on GD-VCR and the original VCR development set. We select the scenarios frequently appearing in both GD-VCR and the original VCR development set.

Results are shown in Table 6. We observe that on the images involving scenarios such as funeral, VisualBERT's performance gap is nearly 25%, which is considerably large. The results further demonstrate that the model is still incapable of tackling the QA pairs which are involving cultural differences behind scenarios well.

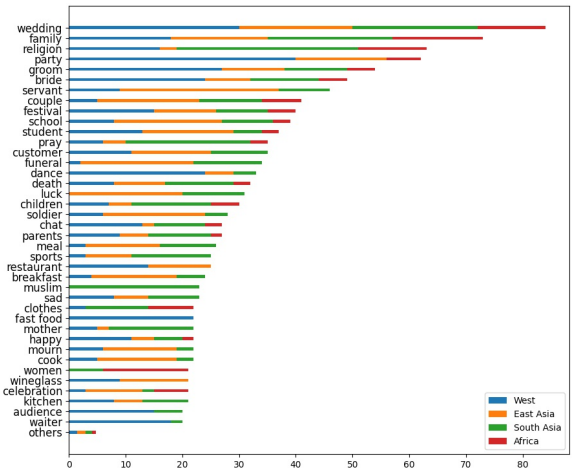


Figure 4: Statistics of keyword occurrences. "Others" denotes the average occurrences of the keywords appearing less than 20 times.

Regions/Datasets	Wedding	Funeral	Servant
Original VCR	71.78	55.00	59.72
Other Regions	50.00	30.25	48.81

Table 6: Accuracy (%) on the images involving the same scenarios from the original VCR dataset and non-Western regions from GD-VCR dataset, respectively.

E Keywords in GD-VCR Dataset

Figure 4 shows the overall statistics of keyword occurrences in GD-VCR benchmark. There are 693 keywords in total, showing the diversity of the scenarios covered by GD-VCR dataset. Besides, we observe that the keywords whose corresponding scenarios have evident regional differences, such as "*wedding*", "*religion*", "*groom*", "*bride*", appear frequently in GD-VCR.

F More Examples in GD-VCR Dataset

In this section, we showcase several examples of GD-VCR in detail. Aside from the images about "*wedding*" in Figure 1, we manifest the images regarding to "*customer*" and "*funeral*" from the four regions we study. In Figure 5 and Figure 6, we can observe the regional characteristics from the selected images. Furthermore, we visualize VisualBERT's prediction on each QA pair.

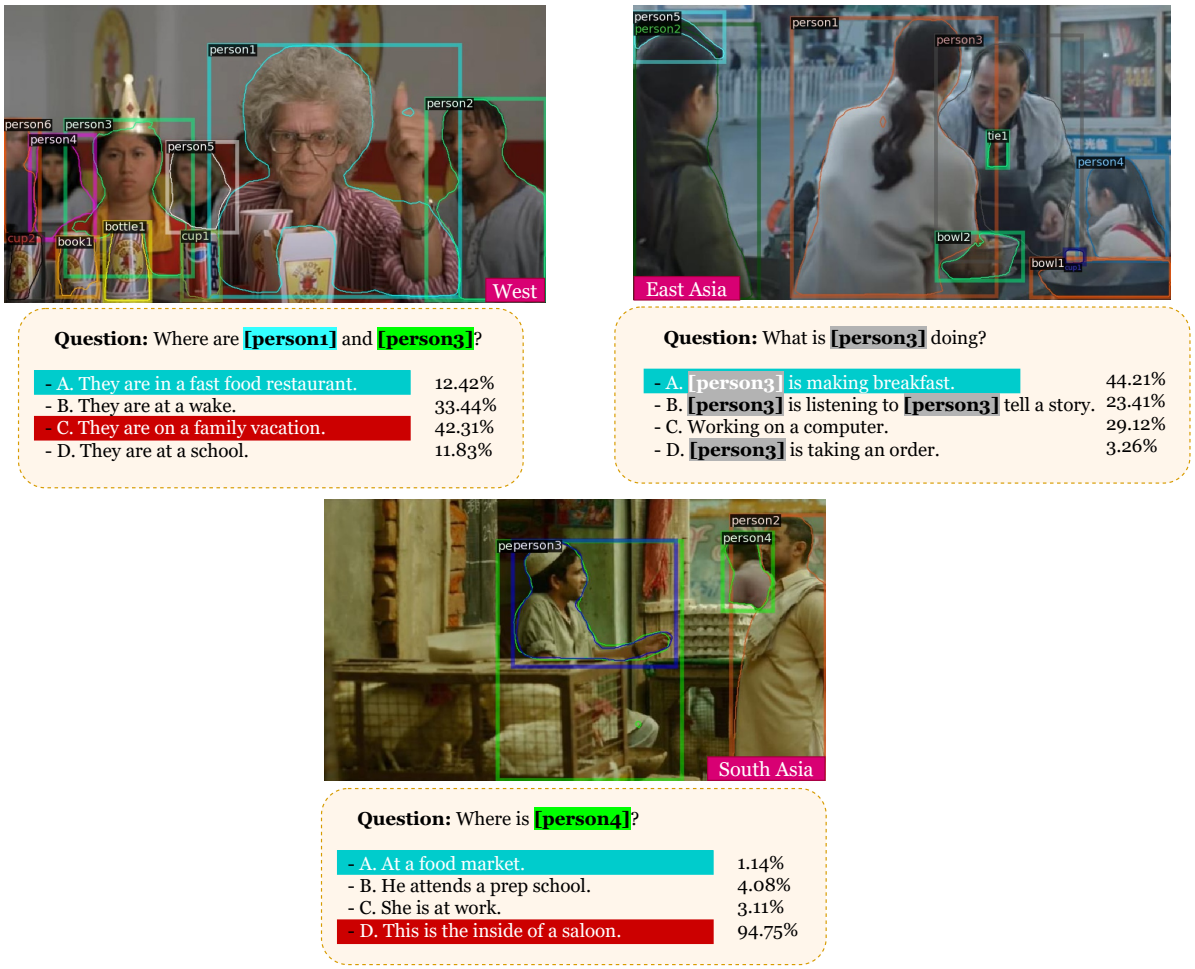
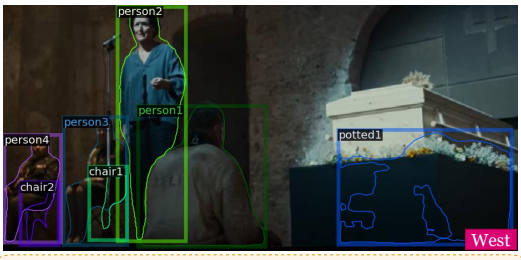
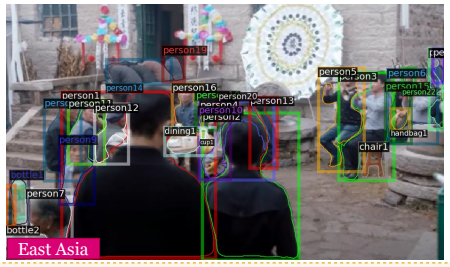


Figure 5: Examples of the images regarding "*customer*". Left-to-right order: Western, South Asia, East Asia. We visualize the prediction of the VisualBERT model fine-tuned on the original VCR training set. The blue blocks denote the right answer choices. If red block appears, it means that VisualBERT wrongly predict the answer. The rightmost value indicates the probability of the corresponding choices being selected by VisualBERT.



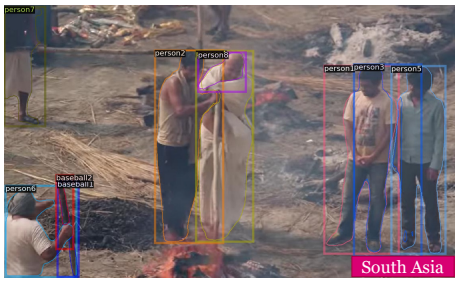
Question: Why does **[person1]** come here?

- A. **[person1]** can't believe his friend's death and wants to apologize.	72.69%
- B. **[person1]** is **[person1]**'s job to serve food to the guests.	10.28%
- C. **[person1]** wants to get a picture of **[person1]**.	0.34%
- D. **[person1]** is on vacation.	16.70%



Question: Why is **[person19]** here?

- A. **[person19]**'s father is dead and **[person19]** is mourning him.	32.73%
- B. Trying to listen to **[person19]** talk.	48.38%
- C. **[person19]** need to investigate graveyards at night.	0.99%
- D. **[person19]** might be a zoo employee.	17.90%



Question: Why are **[person3]** and **[person5]** so sad?

- A. Because the cremated body is their friend.	18.54%
- B. They are avoiding the other people at the party.	24.23%
- C. They are scared and worried for their lives thinking of what they leave behind.	58.98%
- D. They just got their food and are happy that they are finally getting a break from classes and eat something.	0.95%

Question: Why is **[person2]** crying?

- A. Because **[person2]** wants to show sympathy to **[person1]** and **[person3]**.	84.49%
- B. **[person2]** hurt **[person1]** and **[person3]**'s hand.	0.00%
- C. **[person2]** is doing some kind of medical procedure to **[person1]** and **[person3]**.	13.41%
- D. A meeting of business people has just found out some bad news that affect **[person2]**.	2.10%

Figure 6: Examples of the images regarding "*funeral*" or "*death*". Left-to-right order in the first row: Western, East Asia; Left-to-right order in the second row: South Asia, Africa.