Broaden the Vision: Geo-Diverse Visual Commonsense Reasoning

Da Yin, Liunian Harold Li, Ziniu Hu,
Nanyun Peng, Kai-Wei Chang

Background

*Commonsense is … a basic ability to perceive, understand, and judge in a manner that is shared by **nearly all people**.*

--- Wikipedia (Common Sense)

Background

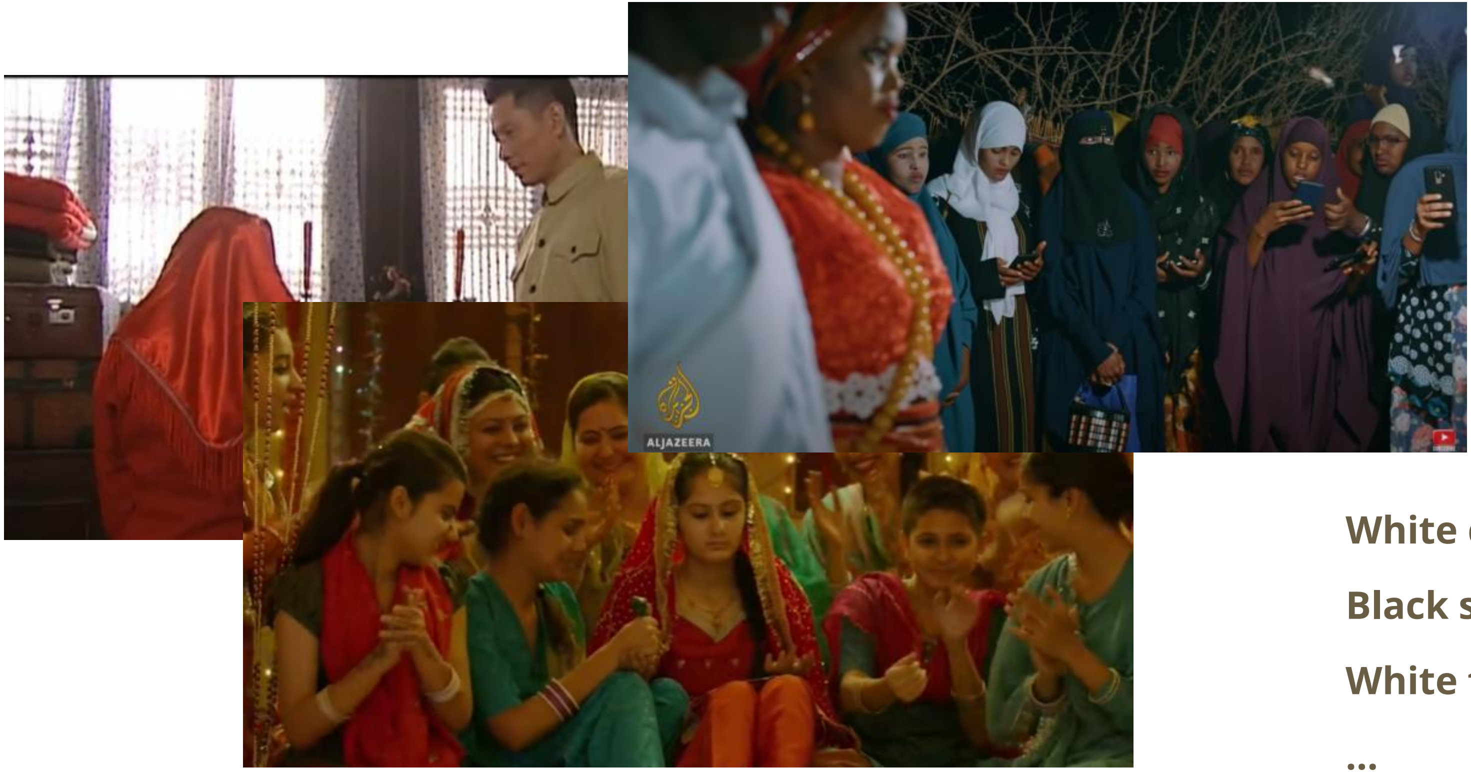
White dress ✓

Black suits ✓

White flowers ✓

...

Background



- White dress ✗
- Black suits ✗
- White flowers ✗

…

Background

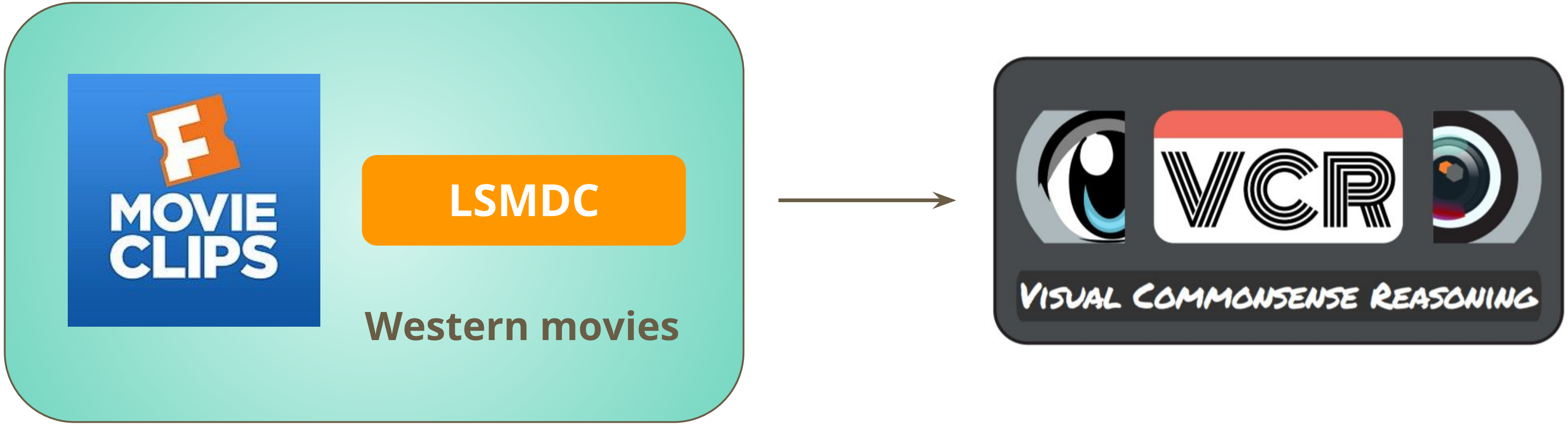
Commonsense is more diverse than we thought!

Background

Commonsense is geo-diverse!

Background

- Some datasets are composed by data from sources in certain regions
  - E.g., VCR (Visual Commonsense Reasoning)



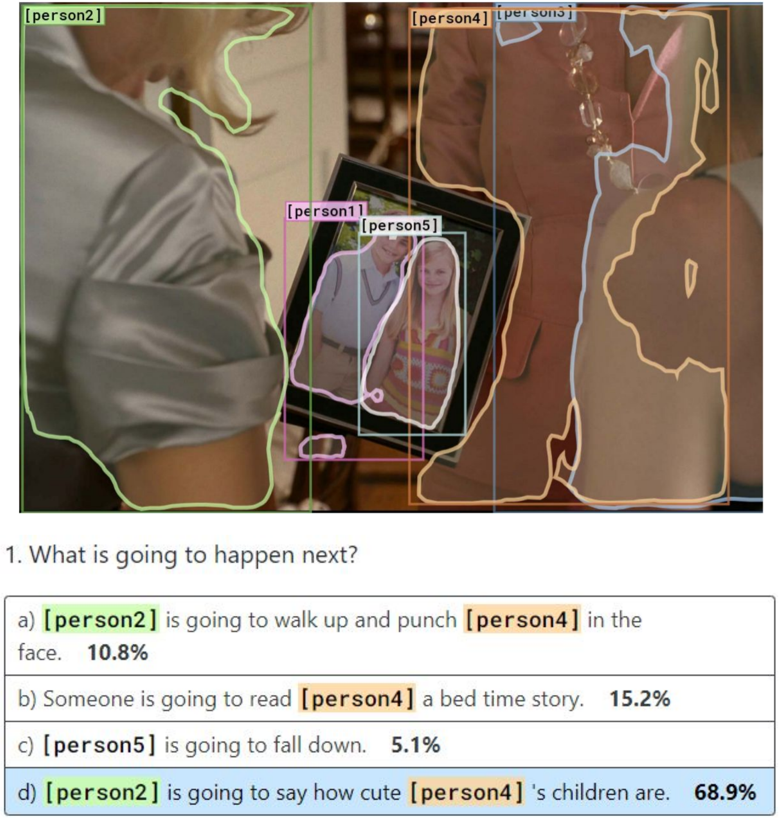
LSMDC

Western movies

From Recognition to Cognition: Visual Commonsense Reasoning (CVPR 2019)

GD-VCR Dataset

- **G**eo-**D**iverse **V**isual **C**ommonsense **R**easoning (GD-VCR) dataset
    - Follow the settings of original VCR dataset.



1. What is going to happen next?

a) [person2] is going to walk up and punch [person4] in the face. **10.8%**

b) Someone is going to read [person4] a bed time story. **15.2%**

c) [person5] is going to fall down. **5.1%**

d) [person2] is going to say how cute [person4] 's children are. **68.9%**

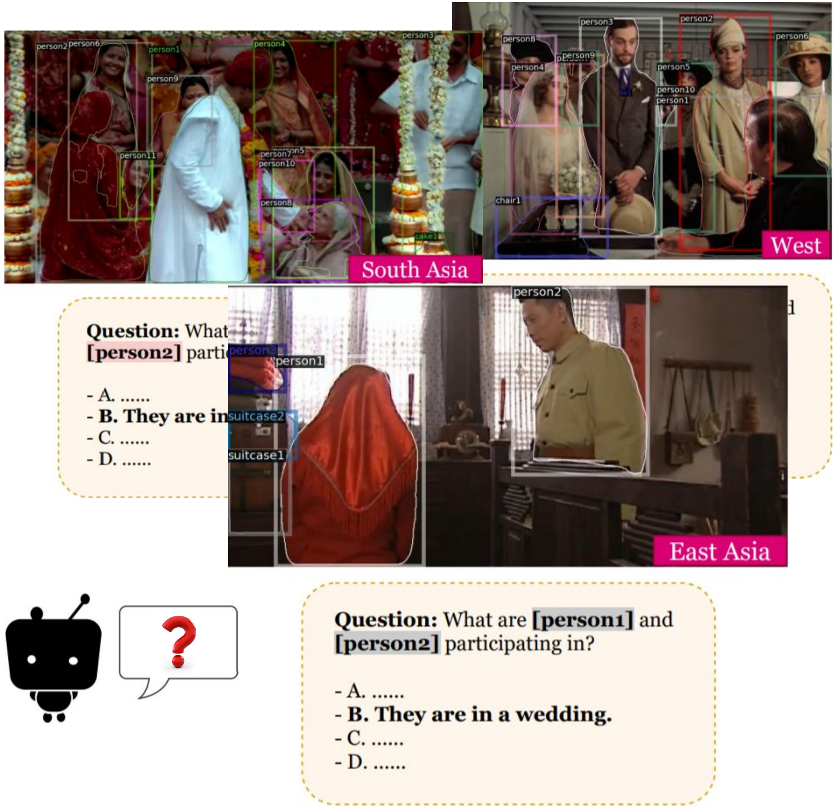
From Recognition to Cognition: Visual Commonsense Reasoning (CVPR 2019)

GD-VCR Dataset

- **G**eo-**D**iverse **V**isual **C**ommonsense **R**easoning (GD-VCR) dataset
  - Follow the settings of original VCR dataset.
  - Collect images from **East Asian, South Asian, African and Western** countries.

GD-VCR Dataset

- **G**eo-**D**iverse **V**isual **C**ommonsense **R**easoning (GD-VCR) dataset
  - Follow the settings of original VCR dataset.
  - Collect images from **East Asian, South Asian, African and Western** countries.
  - <u>**Goal:** </u>Evaluate model's reasoning ability on the task which requires geo-diverse commonsense knowledge.



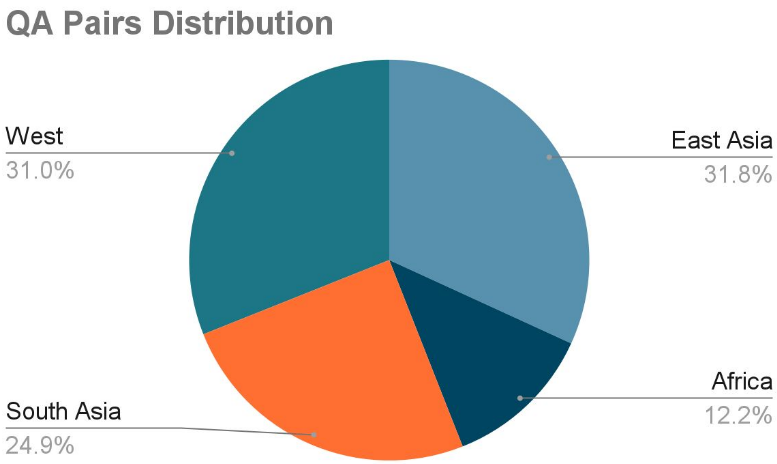
Question: What are **[person1]** and **[person2]** participating in?

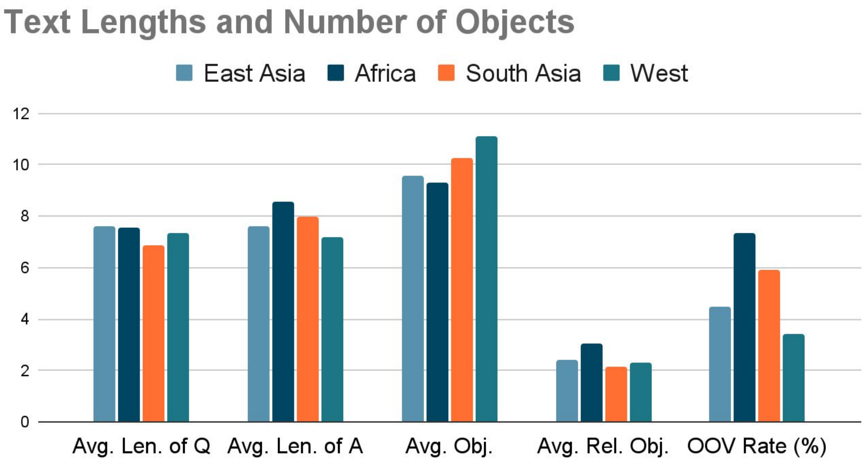
- A. ......
- **B. They are in a wedding.**
- C. ......
- D. ......

GD-VCR Dataset

OOV rate: the ratio of words that appear in GD-VCR but **not** in original VCR training set.

- Statistics about GD-VCR
  - Total: 328 images, 886 QA pairs
  - Text lengths, numbers of image bounding boxes, and OOV rate are similar across regions.





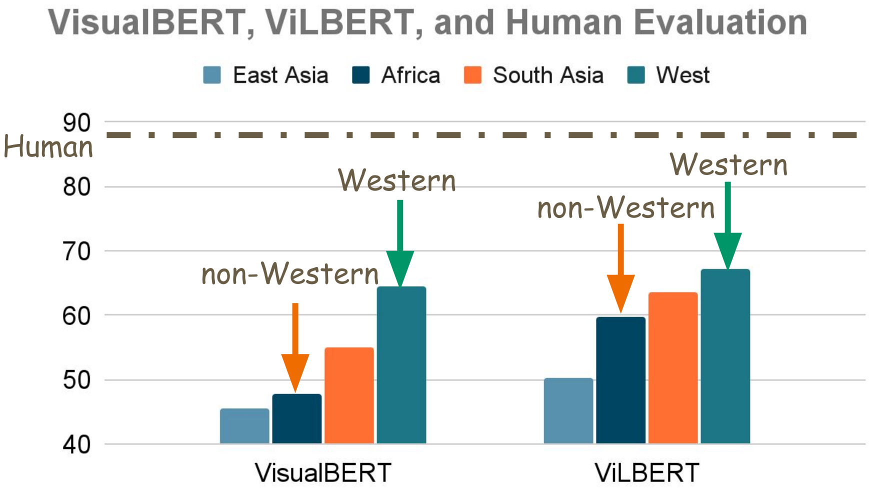
Experiments

- Evaluated Models
  - VisualBERT (Li et al., 2019)
  - ViLBERT (Lu et al., 2019)
- Evaluation Steps
  - Fine-tune models on **original VCR training dataset.**
  - Select the epoch with the highest performance on **original VCR development set.**
  - Test models on **GD-VCR**.

Results

- Model and Human Performance on Different Regions
  - Western regions vs. Non-western regions

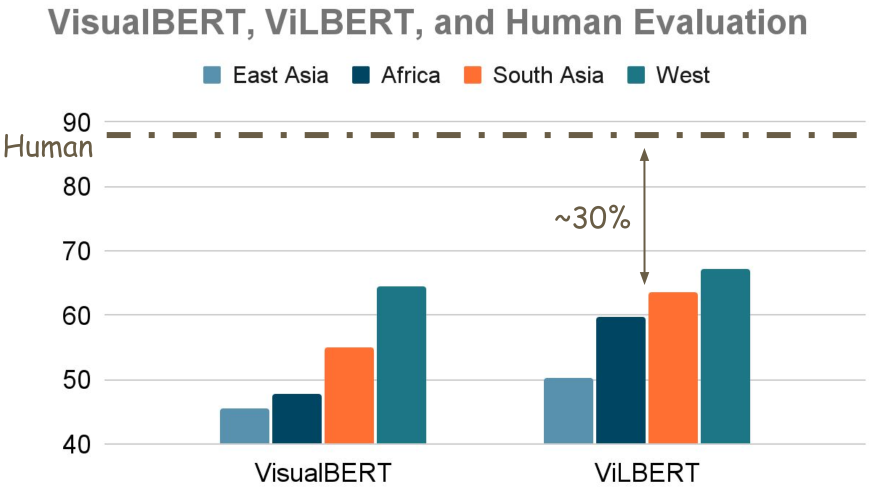
Observation 1: Models perform significantly worse on the images from non-Western regions.



Results

- Model and Human Performance on Different Regions
  - Models vs. Human

Observation 2: Though human may not be familiar with the culture, they still outperform models around 30%.



Analysis

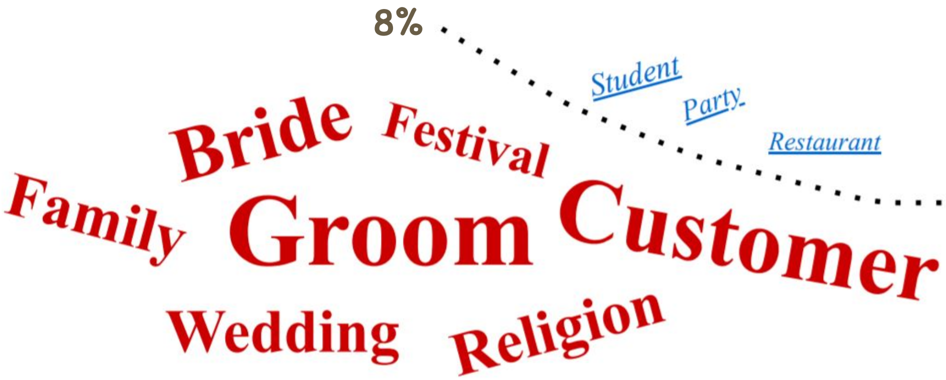
- Why such performance disparity exists?
  - Regional differences of scenarios
    - Wedding, funeral, religion, etc.

Analysis

- Why such performance disparity exists?
  - Regional differences of scenarios
    - Wedding, funeral, religion, etc.

Analysis

- Why such performance disparity exists?
  - Regional differences of scenarios
    - Compare model performance on images about the same scenarios across regions



Observation 1: For the scenarios which often involve regional characteristics, the performance gap is much larger.

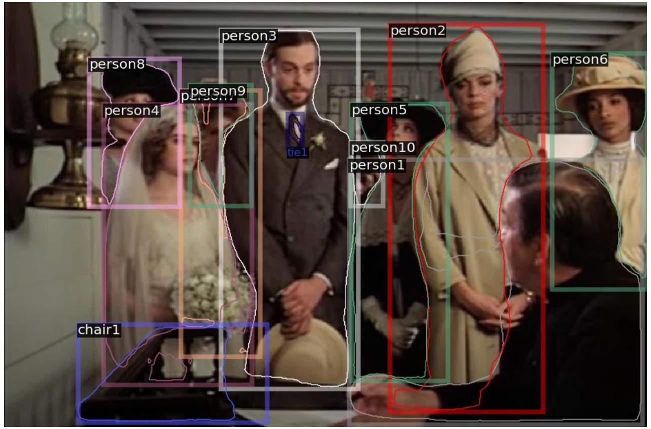
Analysis

- Why such performance disparity exists?
  - Reasoning level of QA pairs
    - Situation 1: Model even fails to recognize the basic facts from non-Western images.
    - Situation 2: Model performs similarly on the basic facts but fails eventually due to lack of geo-diverse commonsense.

Analysis

- Why such performance disparity exists?
    - Reasoning level of QA pairs
        - Design **low-order** (low reasoning level) QA pairs.
        - Assume QA pairs in GD-VCR are **high-order** (high reasoning level) QA pairs.

Example of low-order cognitive QA pair

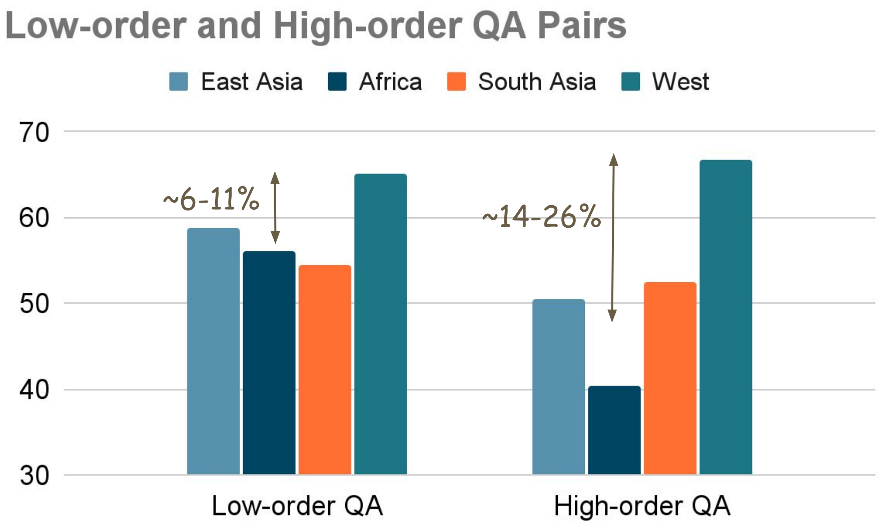


Question: What's [person3] wearing?

Answer: [person3] is wearing a suit.

Analysis

- Why such performance disparity exists?
  - Reasoning level of QA pairs



Observation 2: The disparity across regions on low-order QA pairs is much smaller than on high-order QA pairs.

Analysis

- Why such performance disparity exists?
  - Reasoning level of QA pairs
    - Situation 1: Model even fails to recognize the basic facts from non-Western images.
    - Situation 2: Model performs similarly on the basic facts but fails eventually due to lack of geo-diverse commonsense.

Conclusions and Broader Impact

- Conclusions
  - Build a new geo-diverse dataset GD-VCR
  - Evaluate model performance on GD-VCR
  - Analyze the sources of performance disparity
- Future Directions
  - Broaden researchers' vision on the scope of commonsense reasoning field
  - Motivate researchers to build better commonsense reasoning systems with more inclusive consideration

Thanks for listening!

Code & Data: https://github.com/WadeYin9712/GD-VCR

Project Page: https://gd-vcr.github.io