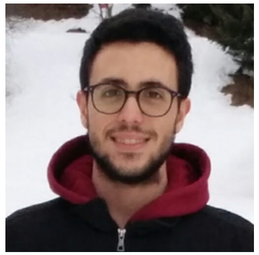
EMNLP 2021

Visually Grounded Reasoning across Languages and Cultures

Fangyu Liu*

University of Cambridge
fl399@cam.ac.uk



Emanuele Bugliarello*

University of Copenhagen
emanuele@di.ku.dk

Edoardo M. Ponti

Mila/McGill University

Siva Reddy

Mila/McGill University

Nigel Collier

University of Cambridge

Desmond Elliott

University of Copenhagen

MaRVL-ta ஏறுதழுவல் (Bull taming)

இரு படங்களில் ஒன்றில் இரண்டிற்கும் மேற்பட்ட மஞ்சள் சட்டை அணிந்த வீரர்கள் காளையை அடக்கும் பணியில் ஈடுப்பட்டிருப்பதை காணமுடி.

In one of the two photos, more than two yellow-shirted players are seen engaged in bull taming.

Label: True

MaRVL

Multicultur**a**l **R**easoning over **V**ision and **L**anguage

Evaluation data for cross-lingual V&L transfer

Task

Predict if a caption is True/False for 2 images

V&L Data

Languages

- Mostly in English
- Or in a few Indo-European Languages

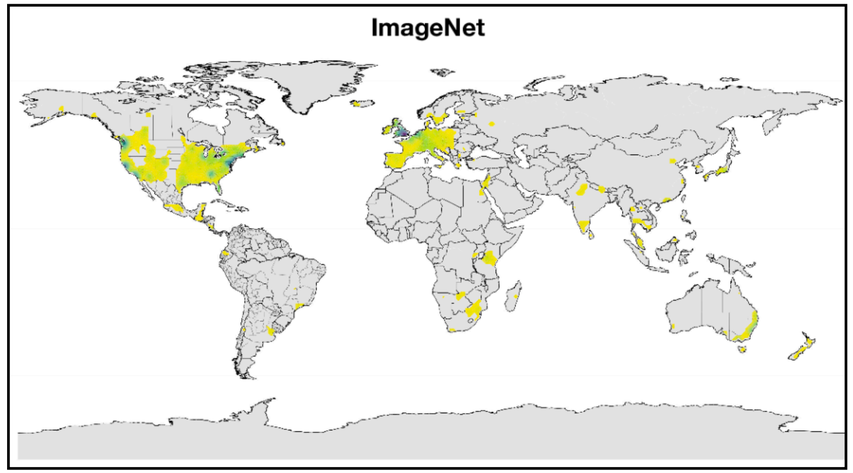
A street organ stands in front of a …

An unusual looking vehicle parked …

Example from van Miltenburg+ (INLG'17)

Image sources

- Mostly from ImageNet, MS COCO and Visual Genome
- Reflecting North American and European cultures



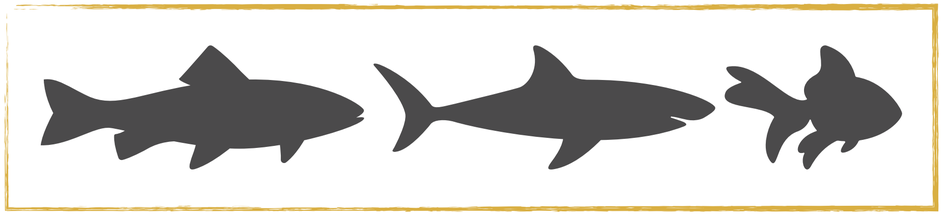
Density map of geographical distribution of images in ImageNet (DeVries+, CVPRW'19)

Implications for V&L models

- Narrow linguistic/cultural domain
- No way to assess their real-world comprehension

Biases in Image Collections: An ImageNet Study

Concept selection



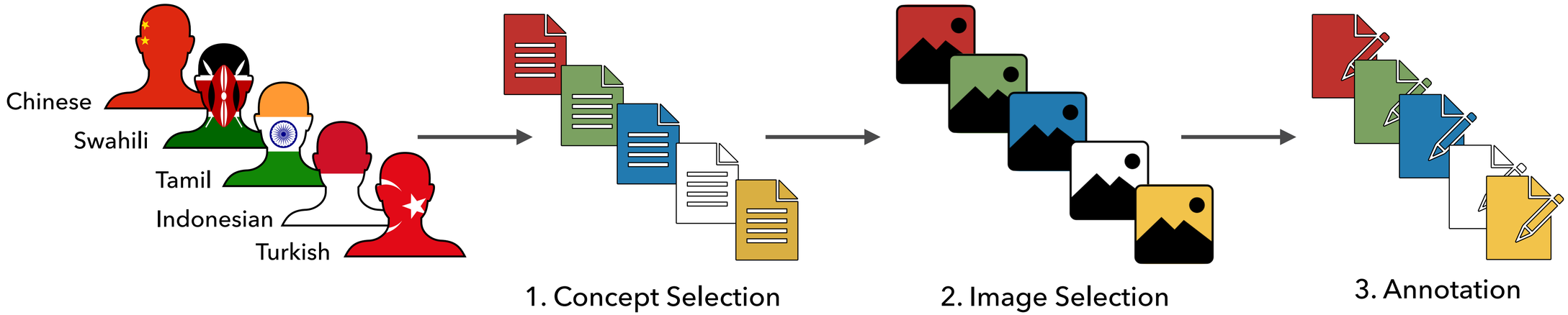
Candidate image retrieval

Manual cleanup

amazon mechanical turk

Overview of MaRVL Collection

Native speaker-driven protocol



Universal Concepts

Concepts that are shared across cultures

From the *Intercontinental Dictionary Series* (Key & Comrie, 2015)

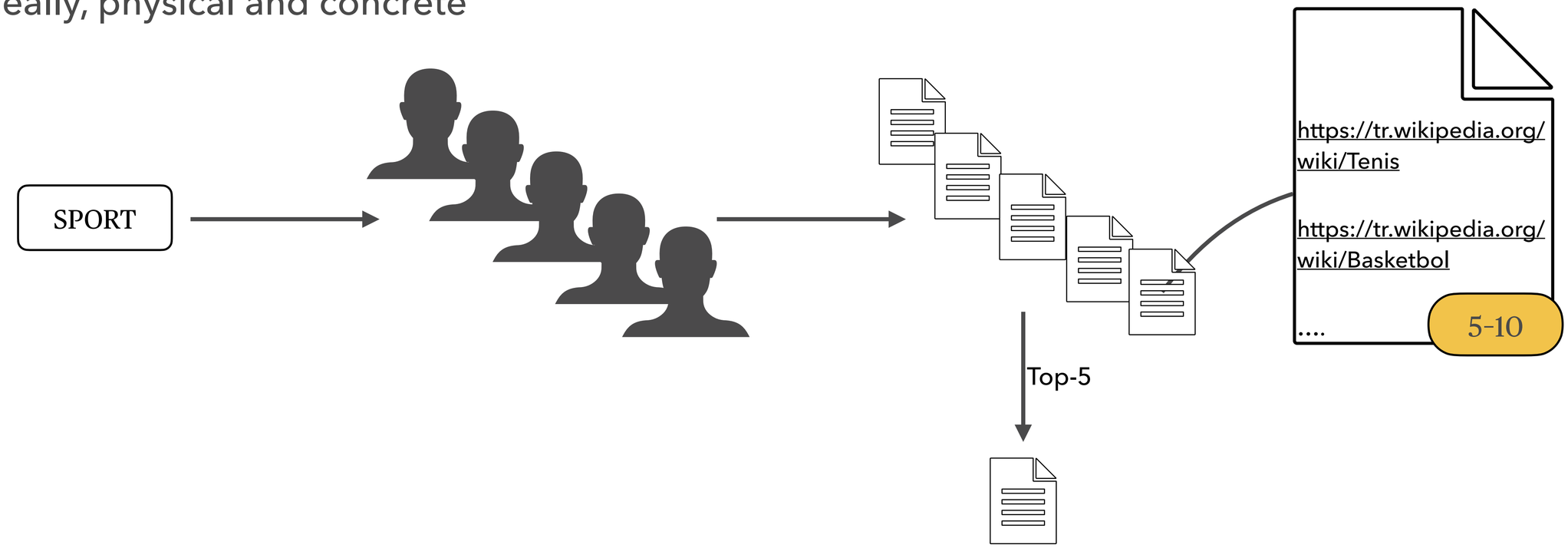
18 chapters with concrete objects & events

Chapter	Semantic Field
Animal	Bird, mammal
Food and Beverages	Food, Beverages
Clothing and grooming	Clothing
The house	Interior, exterior
Agriculture and vegetation	Flower, fruit, vegetable,
Basic actions and technology	Utensil/tool
Motion	Sport
Time	Celebrations
Cognition	Education
Speech and language	Music (instruments), visual arts
Religion and belief	Religion

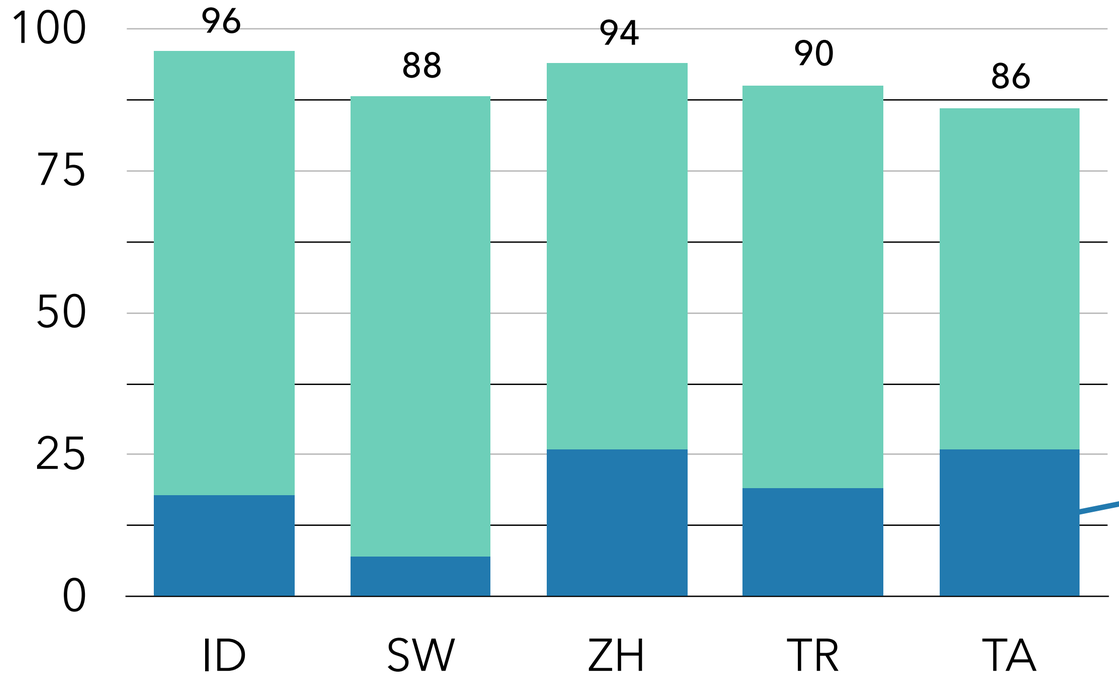
Language-Specific Concepts

Defined by native speakers

- Commonly seen or representative in their culture
- Ideally, physical and concrete



Resulting Concepts

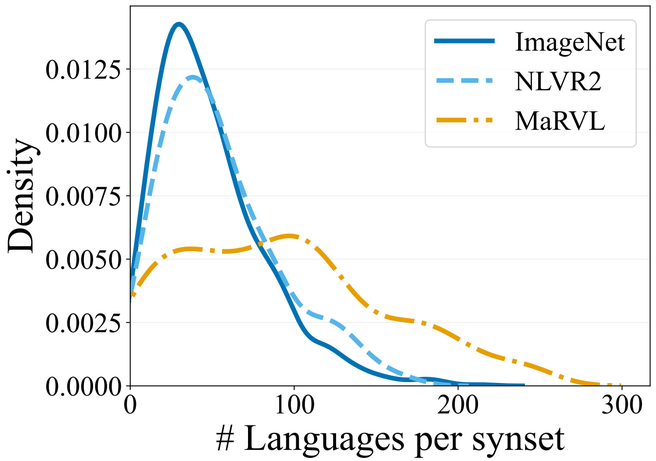


86-96 concepts per language

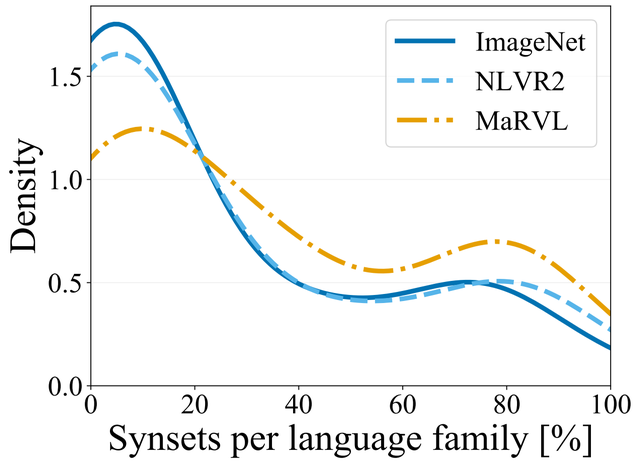
72.4% by at least 3 annotators

NOT IN ENGLISH WORDNET!

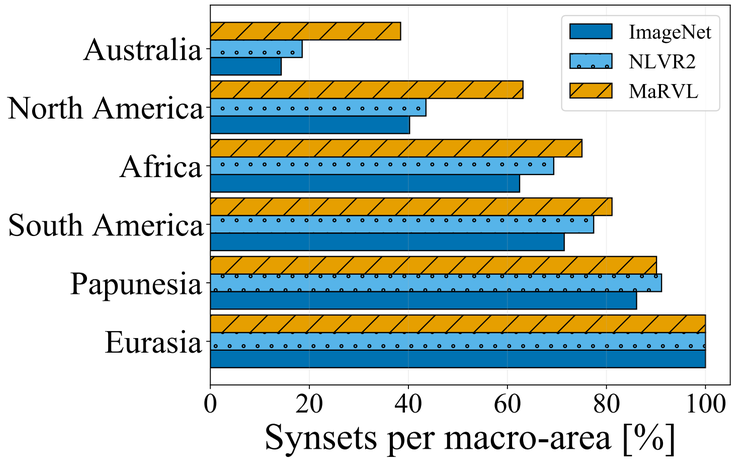
Concept Distributions



MaRVL concepts are in more ***languages***



MaRVL concepts are in more ***families***



MaRVL concepts are in more ***macroareas***

Images

Collected by native speakers

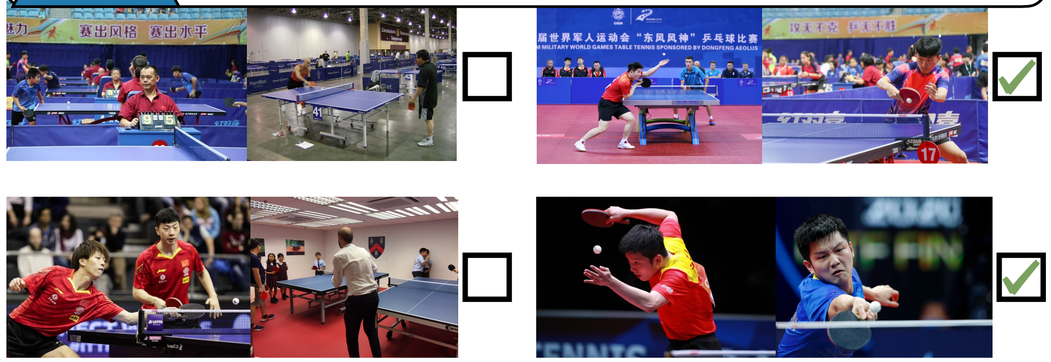
- Representative of the language population
- .....

MaRVL-sw **Jembe (Shovel)**

Captions

MATCH 4 PAIRS AT RANDOM

WRITE CAPTION TRUE ONLY FOR 2 PAIRS



右图中的人在发球，左图中的人在接球。

(The man in the right image is serving a ball while the man in the left image is returning a ball.)

VALIDATE ANNOTATIONS

右图中的人在发球，左图中的人在接球。

(The man in the right image is serving a ball while the man in the left image is returning a ball.)

FINAL VALIDATION

右图中的人在发球，左图中的人在接球。

(The man in the right image is serving a ball while the man in the left image is returning a ball.)

Examples

MaRVL-zh 火锅 (Hot pot)

两图中至少有一张图里面是一口鴛鴦鍋

At least one of the two pictures shows a mandarin duck pot

Label: True

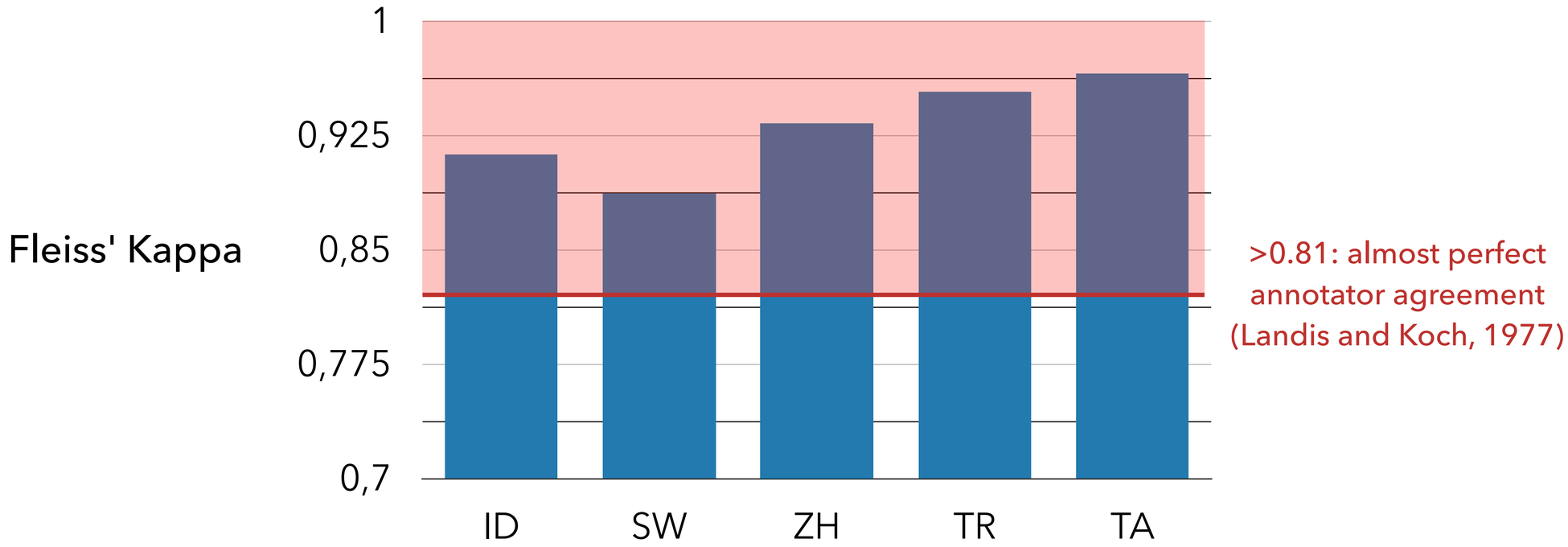
MaRVL-ta மை (Vada)

இரண்டு படங்களிலும் நிறைய மசால் வடைகள் உள்

Both images contain a lot of masala vadas

Label: False

Human Quality Assessment



Key Statistics & Limitations

Key statistics

- 423 concepts
  - 96 not in WordNet
- 5464 images
- 1390 unique captions
- 5560 data points

Limitations

- Low-resource language annotators
- Wikipedia as a proxy for concepts
- ……

Experimental Setup

Models

- 5 V&L BERTs from VOLTA (Bugliarello+, 2021)
- 2 new multilingual UNITER models
  - mUNITER: Initialised from mBERT
  - xUNITER: Initialised from XLM-R

Multilingual Multimodal Pre-training

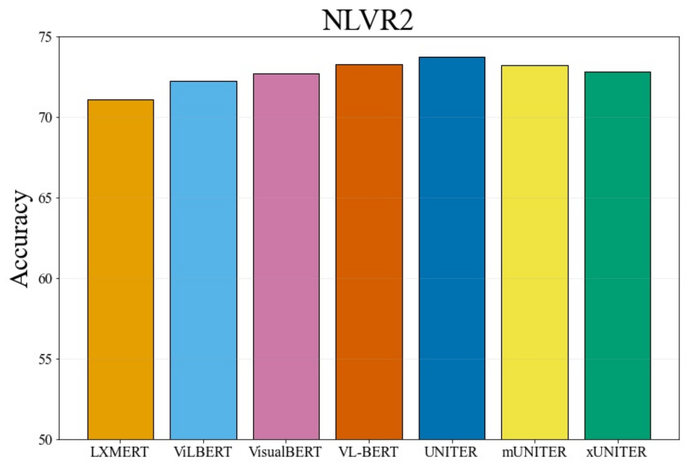
Following M^{3}P (Ni+, 2021)

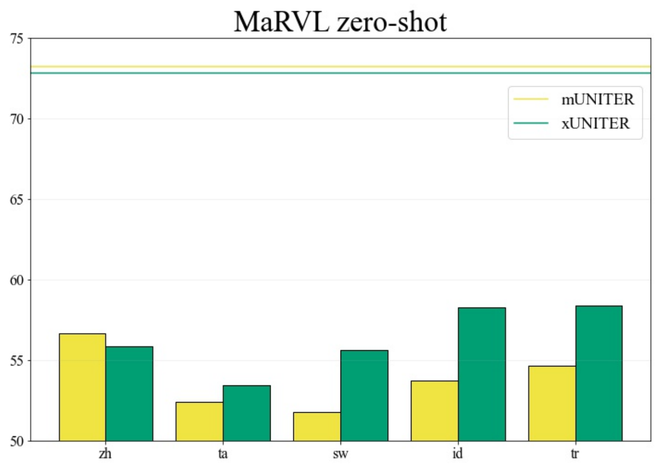
- 104 Wikipedia (mBERT): MLM
- Conceptual Captions: MLM + MRM + ITM

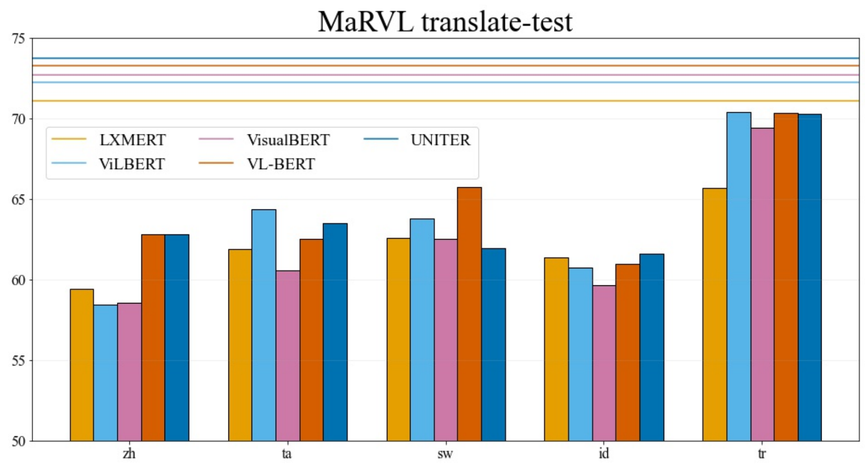
Fine-tuning

- Train on English NLVR2 (Suhr+, 2019)
- Test on MaRVL
  - Multilingual models in a "*zero-shot*", cross-lingual fashion
  - English models in a "*translate-test*" approach

Main Results







mUNITER and xUNITER are on par in NLVR2

Zero-shot transfer: -10–20%

Translate-test: -10%

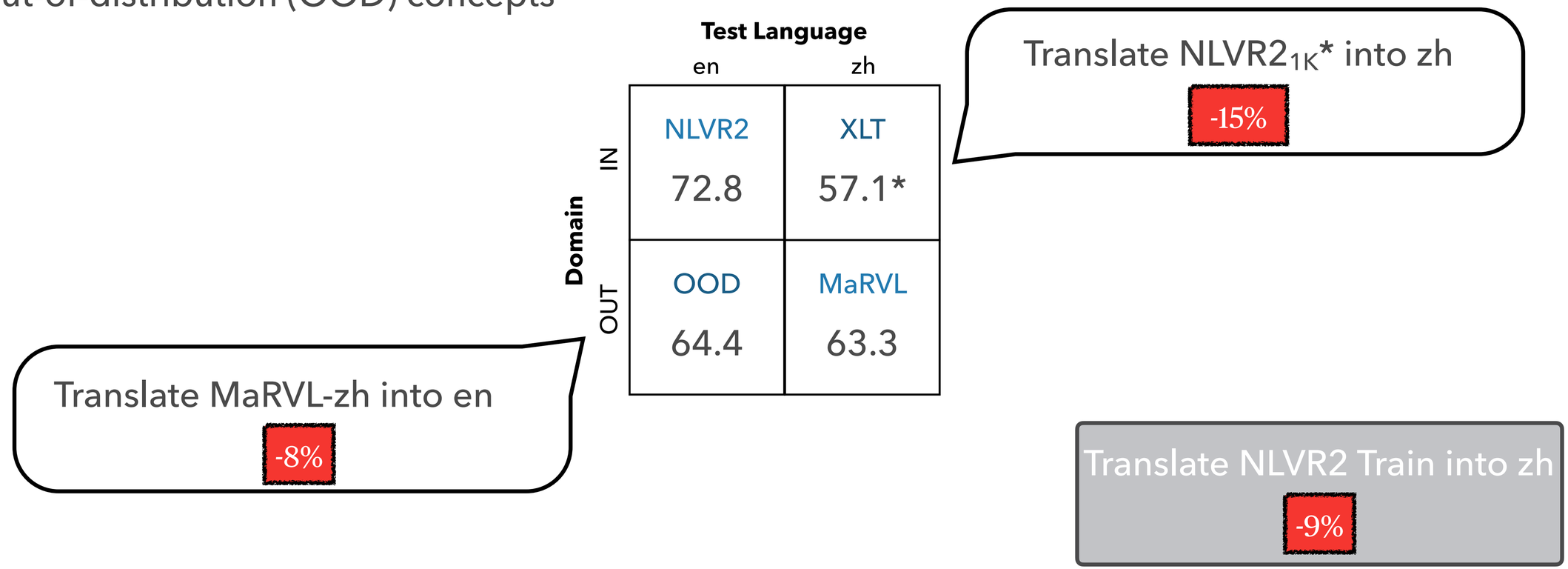
Turkish is surprisingly easy

Disentangling Distribution Shifts: A zh Study

2 distribution shifts in MaRVL

- Cross-lingual transfer (XLT)
- Out-of-distribution (OOD) concepts

Controlled study on **MaRVL-zh** with xUNITER



Conclusions & Next Steps

Thank you

Show that concepts and images in existing V&L datasets are not cross-lingual/cultural

New protocol for concept and image selection entirely driven by native speakers

Introduce MaRVL: A V&L reasoning dataset in 5 typologically diverse languages

Develop and benchmark multilingual V&L BERTs

- Performance can be at chance level

Adapting V&L models to culture-specific concepts

Multicultural object detectors

Code, data & annotation guidelines are online

- marvl-challenge.github.io
- Add your own language!