Visually Grounded Reasoning across Languages and Cultures

Desmond Elliott
University of Copenhagen

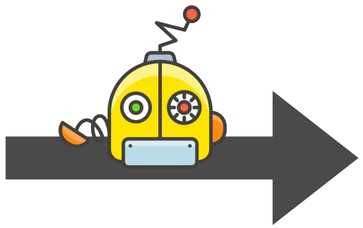
de@di.ku.dk @delliott

UNIVERSITY OF COPENHAGEN

Advances in Vision-and-Language

Advances in Vision-and-Language

IMAGE CAPTIONING

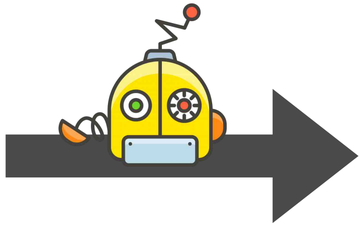


A sandwich on
a cutting board
with a knife.

Cornia+ (CVPR'20)

Advances in Vision-and-Language

IMAGE CAPTIONING

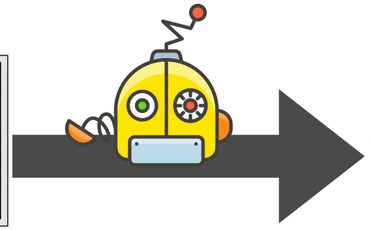


A sandwich on a cutting board with a knife.

Cornia+ (CVPR'20)

MULTIMODAL REASONING

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

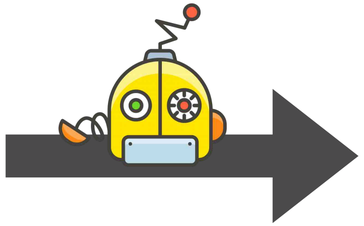


True

Suhr+ (ACL'19)

Advances in Vision-and-Language

IMAGE CAPTIONING

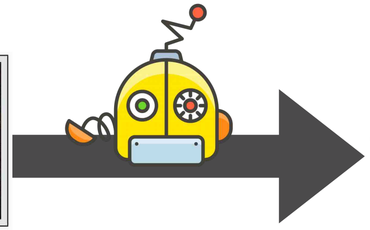


A sandwich on a cutting board with a knife.

Cornia+ (CVPR'20)

MULTIMODAL REASONING

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

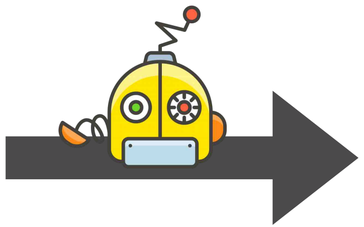


True

Suhr+ (ACL'19)

TEXT-TO-IMAGE SYNTHESIS

A baby daikon radish in a tutu walking a dog



Ramesh+ (arXiv'20)

Advances in Vision-and-Language

IMAGE CAPTIONING

A sandwich on a cutting board with a knife.

Cornia+ (CVPR'20)

MULTIMODAL REASONING

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

True

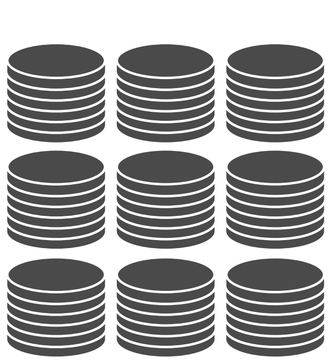
Suhr+ (ACL'19)

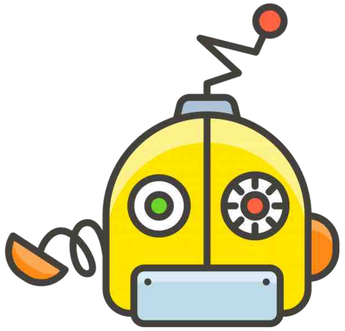
TEXT-TO-IMAGE SYNTHESIS

A baby daikon radish in a tutu walking a dog

Ramesh+ (arXiv'20)

SELF-SUPERVISED LEARNING





CAPTIONING

VQA

REASONING

GROUNDING

Advances in Vision-and-Language

IMAGE CAPTIONING

A sandwich on a cutting board with a knife.

Cornia+ (CVPR'20)

MULTIMODAL REASONING

age contains twice the number of dogs as the and at least two dogs in total are standing.

True

L'19)

DATA

TEXT-TO-IMAGE SYNTHESIS

A baby daikon radish in a tutu walking a dog

Ramesh+ (arXiv'20)

SELF-SUPERVISED LEARNING

CAPTIONING

VQA

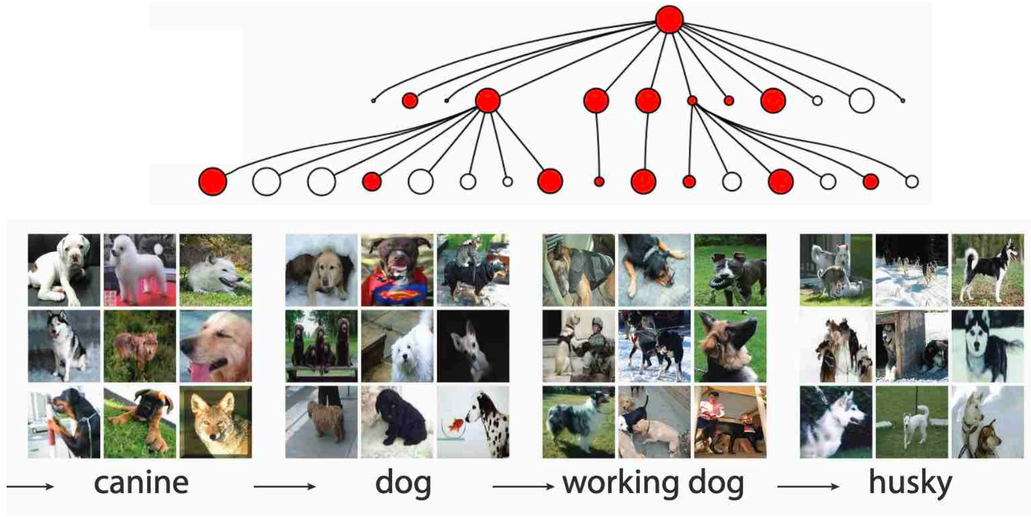
REASONING

GROUNDING

Typical Vision and Language

ImageNet (Deng et al. 2009)

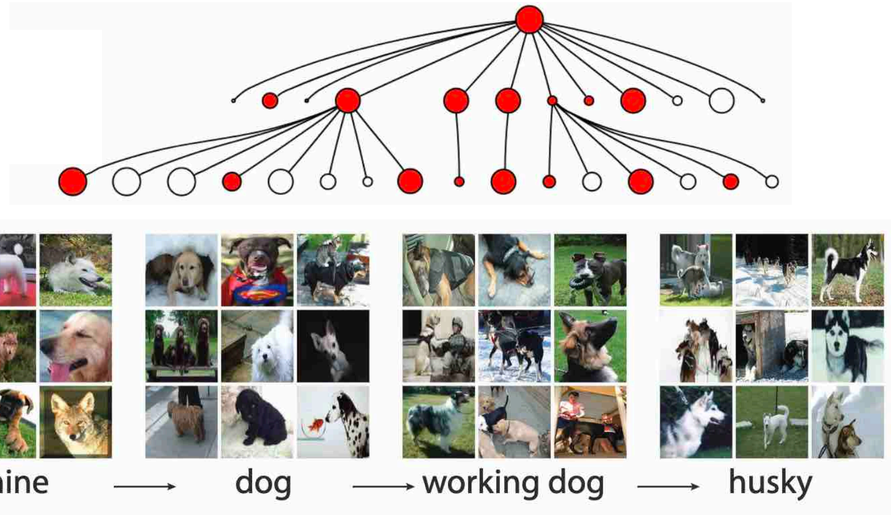
- Train visual encoders
- Millions of labelled images
- Derived from the WordNet hierarchy



Typical Vision and Language

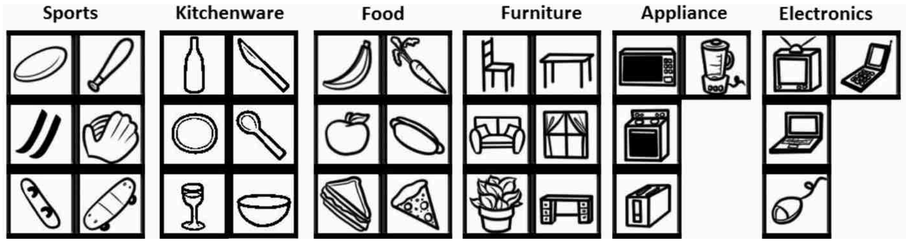
ImageNet (Deng et al. 2009)

- Train visual encoders
- Millions of labelled images
- Derived from the WordNet hierarchy



Common Objects in Context (Lin et al. 2014)

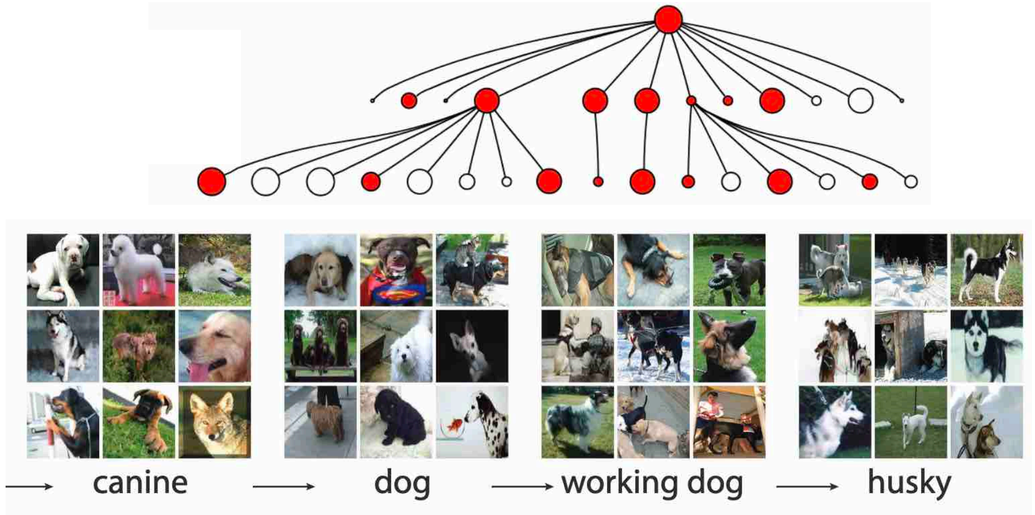
- Train and evaluate multimodal models
- 330K labelled images
  - 80 types of commonly occurring objects



Typical Vision and Language

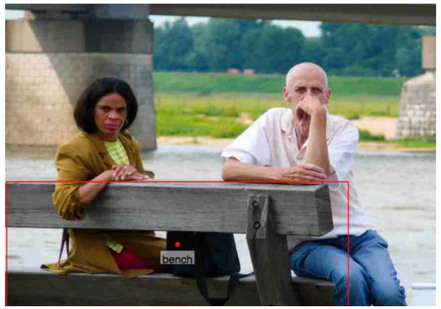
ImageNet (Deng et al. 2009)

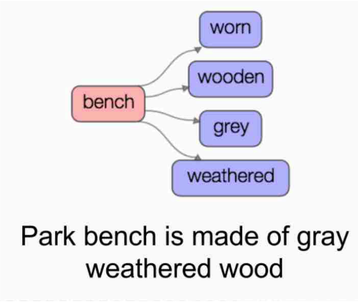
- Train visual encoders
- Millions of labelled images
- Derived from the WordNet hierarchy



Visual Genome (Krishna et al. 2017)

- Train and evaluate multimodal models
- 110K densely annotated images
- Derived from COCO and YFCC100M dataset





Rethinking Vision and Language

Languages

- Mostly in English
- Or some Indo-European Languages

Rethinking Vision and Language

Languages

- Mostly in English
- Or some Indo-European Languages

ENG: An unusual looking vehicle …

NLD: Een mobiel draaiorgel …

Example from van Miltenburg+ 2017

Rethinking Vision and Language

Languages

- Mostly in English
- Or some Indo-European Languages

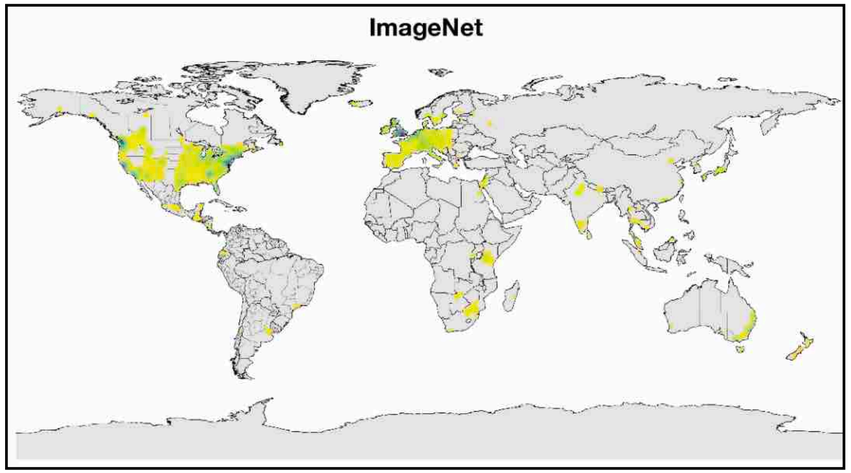
ENG: An unusual looking vehicle …

NLD: Een mobiel draaiorgel …

Example from van Miltenburg+ 2017

Image sources

- Mostly from ImageNet or COCO
- Reflecting North American and European cultures



Density map of geographical distribution of images in ImageNet (DeVries+, 2019)

Rethinking Vision and Language

Languages

- Mostly in English
- Or some Indo-European Languages

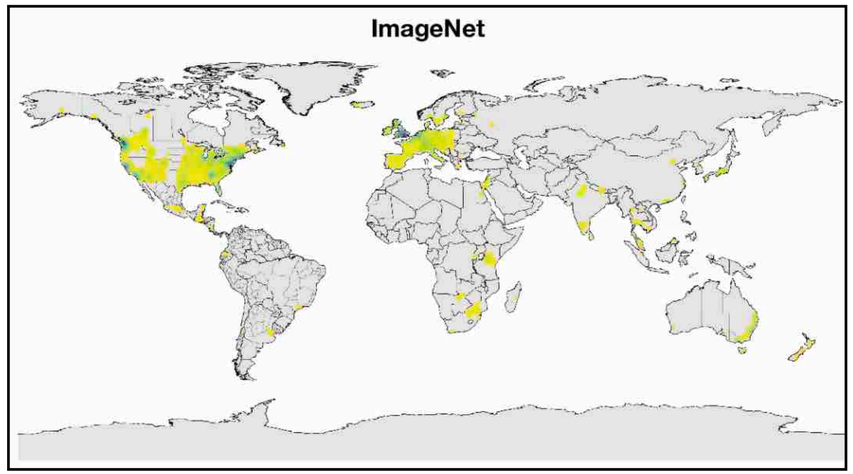
ENG: An unusual looking vehicle …

NLD: Een mobiel draaiorgel …

Example from van Miltenburg+ 2017

Image sources

- Mostly from ImageNet or COCO
- Reflecting North American and European cultures



Density map of geographical distribution of images in ImageNet (DeVries+, 2019)

Implications for V&L models

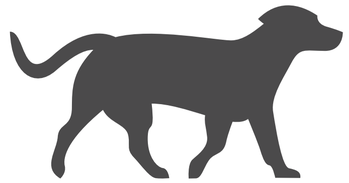
- Narrow linguistic/cultural domain
- No way to assess their real-world comprehension

Concepts in Language and Vision

Concepts and Hierarchies

Category: objects with similar properties (Aristotle 40 BCE, …)

Concept: mental representation of a category (Rosch 1973)



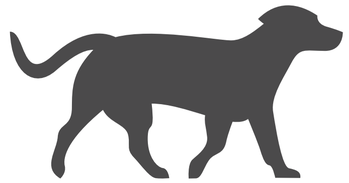
"Dog" concept

"Dog" category

Concepts and Hierarchies

Category: objects with similar properties (Aristotle 40 BCE, …)

Concept: mental representation of a category (Rosch 1973)

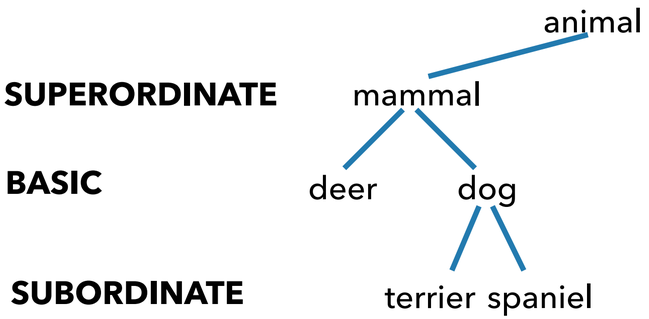


"Dog" concept

"Dog" category

Categories form a *hierarchy*

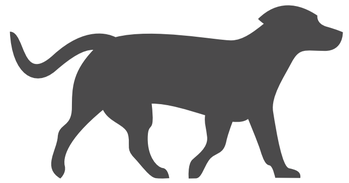
- Basic-level categories (Rosch 1976)



Concepts and Hierarchies

Category: objects with similar properties (Aristotle 40 BCE, …)

Concept: mental representation of a category (Rosch 1973)

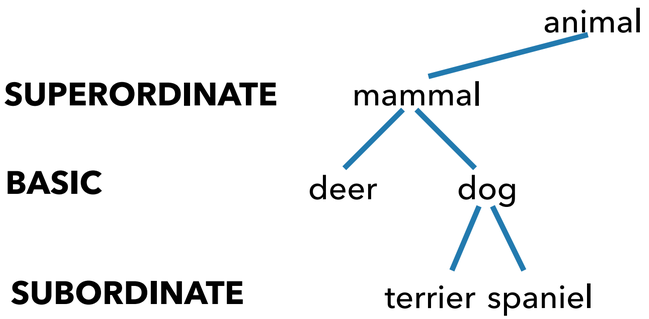


"Dog" concept

"Dog" category

Categories form a *hierarchy*

- Basic-level categories (Rosch 1976)



Somewhat universal

- Different cultures (Berlin 2014)
- Familiarity of individuals (Wisniewski and Murphy, 1989)

Concrete Concepts in Cultural Context

- Some concepts are most immediately understood within a cultural background

Culture: The way of life of a collective of people that distinguishes them from other people (Mora, 2013; Shweder et al. 2007).

See Hershcovich et al. ACL 2022 for an overview of Cross-cultural strategies in NLP

Concrete Concepts in Cultural Context

- Some concepts are most immediately understood within a cultural background

Culture: The way of life of a collective of people that distinguishes them from other people (Mora, 2013; Shweder et al. 2007).

See Hershcovich et al. ACL 2022 for an overview of Cross-cultural strategies in NLP

Concrete Concepts in Cultural Context

- Some concepts are most immediately understood within a cultural background

Culture: The way of life of a collective of people that distinguishes them from other people (Mora, 2013; Shweder et al. 2007).

Pilota / Jai-alai

See Hershcovich et al. ACL 2022 for an overview of Cross-cultural strategies in NLP

Concrete Concepts in Cultural Context

- Some concepts are most immediately understood within a cultural background

Culture: The way of life of a collective of people that distinguishes them from other people (Mora, 2013; Shweder et al. 2007).

Pilota / Jai-alai

Sanxian / Shamisen

See Hershcovich et al. ACL 2022 for an overview of Cross-cultural strategies in NLP

Concrete Concepts in Cultural Context

- Some concepts are most immediately understood within a cultural background

Culture: The way of life of a collective of people that distinguishes them from other people (Mora, 2013; Shweder et al. 2007).

Pilota / Jai-alai

Sanxian / Shamisen

Clavie

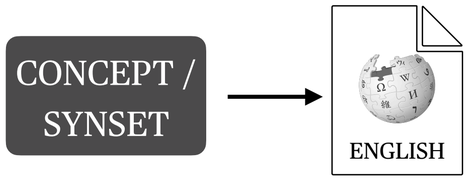
See Hershcovich et al. ACL 2022 for an overview of Cross-cultural strategies in NLP

Are ImageNet Concepts Cross-Lingual?

- The ImageNet, COCO and Visual Genome datasets use English WordNet concepts
- Idea: estimate cross-linguality using Wikipedia as a proxy

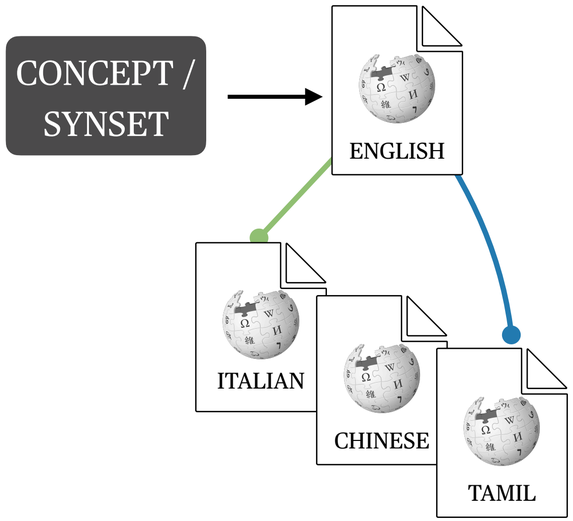
Are ImageNet Concepts Cross-Lingual?

- The ImageNet, COCO and Visual Genome datasets use English WordNet concepts
- Idea: estimate cross-linguality using Wikipedia as a proxy



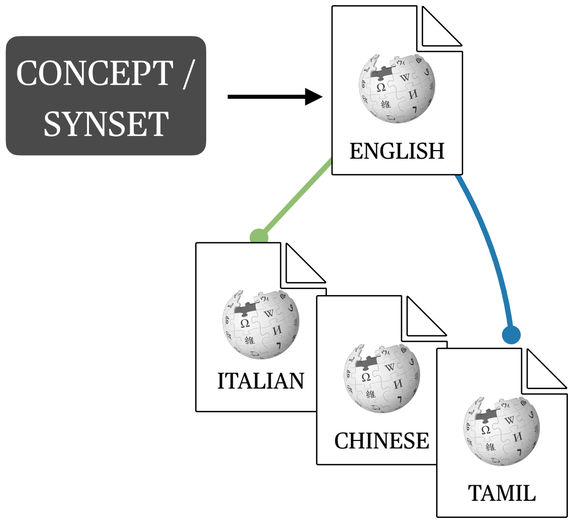
Are ImageNet Concepts Cross-Lingual?

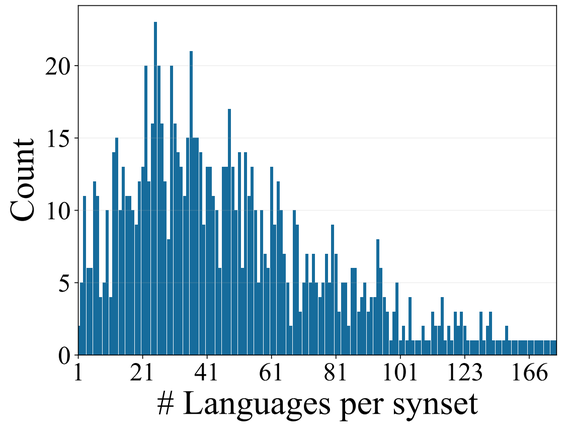
- The ImageNet, COCO and Visual Genome datasets use English WordNet concepts
- Idea: estimate cross-linguality using Wikipedia as a proxy



Are ImageNet Concepts Cross-Lingual?

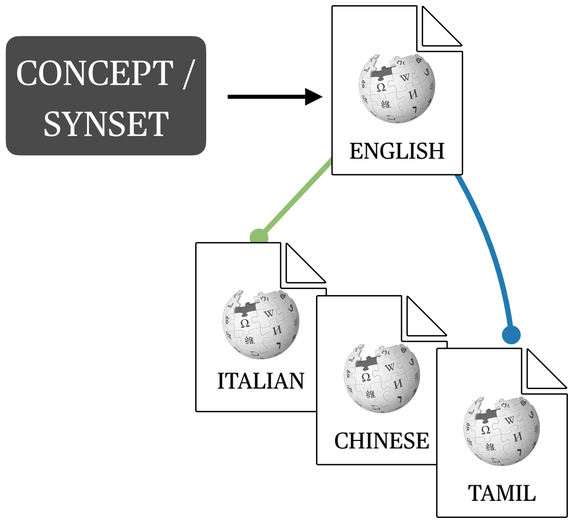
- The ImageNet, COCO and Visual Genome datasets use English WordNet concepts
- Idea: estimate cross-linguality using Wikipedia as a proxy

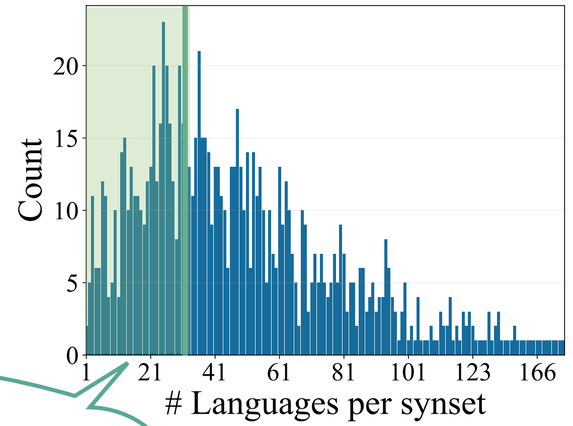




Are ImageNet Concepts Cross-Lingual?

- The ImageNet, COCO and Visual Genome datasets use English WordNet concepts
- Idea: estimate cross-linguality using Wikipedia as a proxy

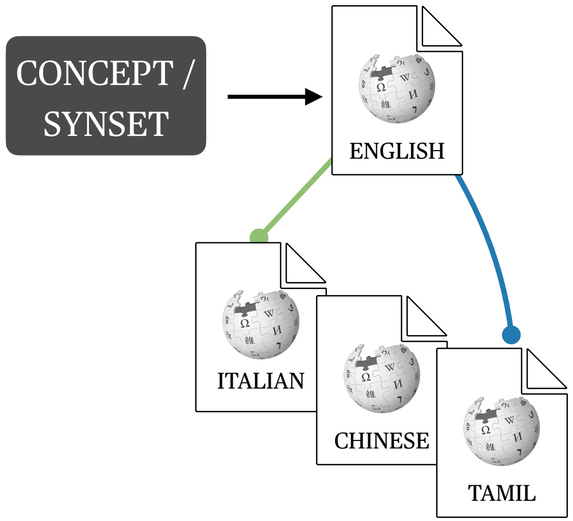


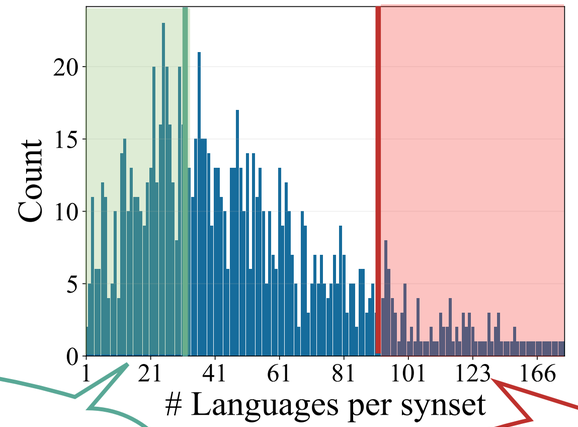


Most concepts in <30 languages

Are ImageNet Concepts Cross-Lingual?

- The ImageNet, COCO and Visual Genome datasets use English WordNet concepts
- Idea: estimate cross-linguality using Wikipedia as a proxy





Most concepts in <30 languages

Only a few "universal" concepts

Are ImageNet Concepts Over-Specific?

American robin

Weimaraner

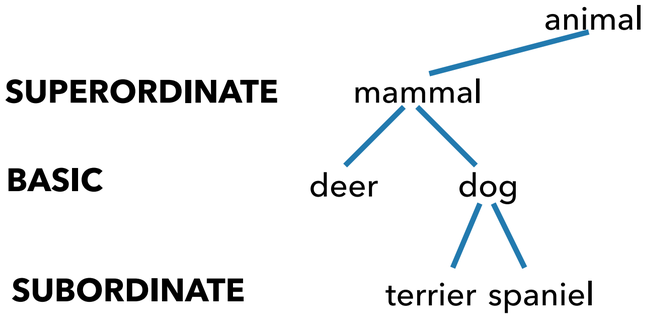
Idea: map 447 human concepts from Ordonez+ (IJCV'13) onto WordNet and measure depth of ImageNet and Human labels

Are ImageNet Concepts Over-Specific?

American robin

Weimaraner

Idea: map 447 human concepts from Ordonez+ (IJCV'13) onto WordNet and measure depth of ImageNet and Human labels



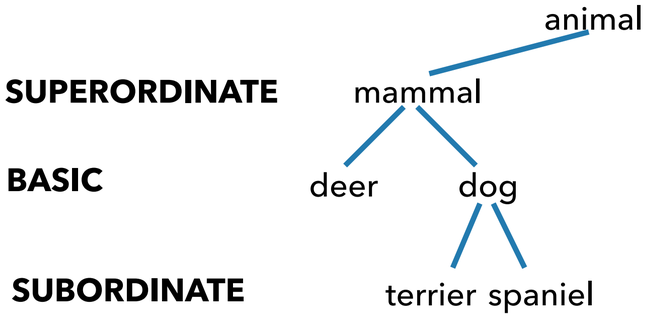
	Depth
Humans	8.92 ± 3.94
ImageNet	10.61 ± 6.13

Are ImageNet Concepts Over-Specific?

American robin

Weimaraner

Idea: map 447 human concepts from Ordonez+ (IJCV'13) onto WordNet and measure depth of ImageNet and Human labels



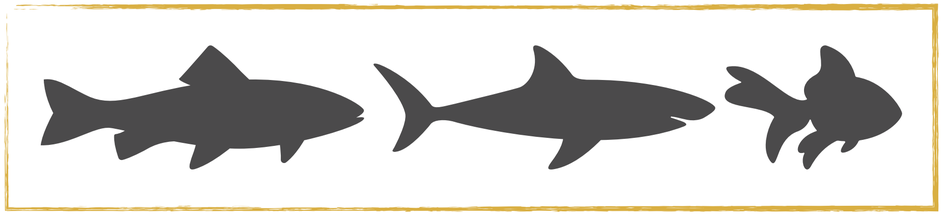
	Depth
Humans	8.92 ± 3.94
ImageNet	10.61 ± 6.13

Human annotations are *less specific* than ImageNet labels

Biases in Image Collections: An ImageNet Study

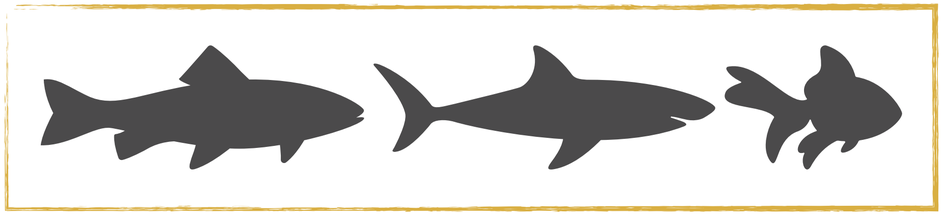
Biases in Image Collections: An ImageNet Study

Concept selection



Biases in Image Collections: An ImageNet Study

Concept selection

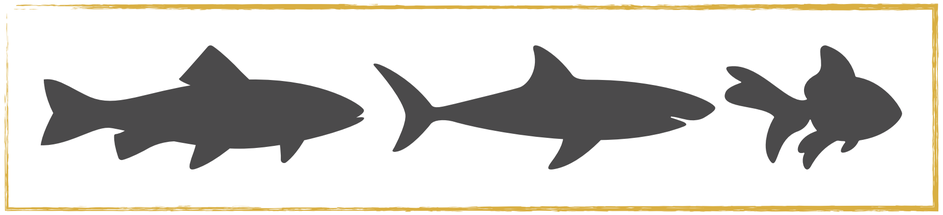


Candidate image retrieval

American English

Biases in Image Collections: An ImageNet Study

Concept selection



Candidate image retrieval

American English

Manual cleanup

amazon mechanical turk

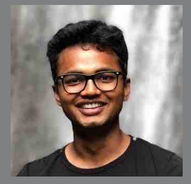
Visually Grounded Reasoning across Languages and Cultures

EMNLP 2021


F. Liu*


E. Bugliarello*


E.M. Ponti


S. Reddy


N. Collier


D. Elliott

MaRVL Multicultural Reasoning over Vision and Language

Representative of annotators' cultures

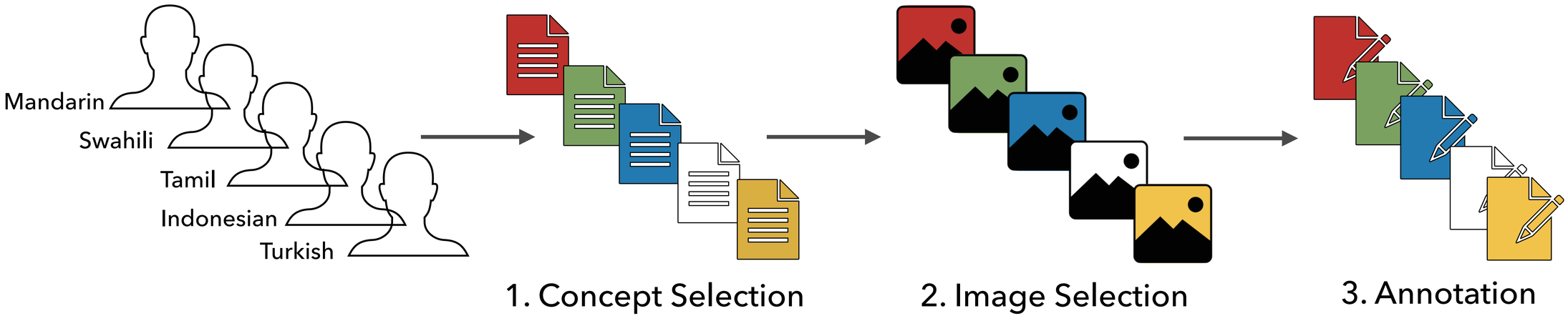
5 typologically diverse languages

Independent, culture-specific annotations

MaRVL-id Bola basket | MaRVL-sw Mpira wa kikapu | MaRVL-tr Basketbol | MaRVL-zh 篮球 | MaRVL-ta கூடைப்பந்தாட்டம்

Collecting MaRVL data

Native speaker-driven protocol



Visual Reasoning Task (Suhr et al. ACL 2019)

- **Datapoint**: two images (v_1, v_2) paired with a sentence (x)

v_1

v_2

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.)

x

Visual Reasoning Task (Suhr et al. ACL 2019)

- **Datapoint**: two images (v_1, v_2) paired with a sentence (x)
- **Task**: Predict whether x is a true description of the pair of images v_1 v_2

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.)

True

v_1 v_2 x y

MaRVL is created from Universal Concepts

- Taken from the *Intercontinental Dictionary Series* (Key & Comrie, 2015)
  - 18/22 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

Step 1. Language-Specific Concept Selection

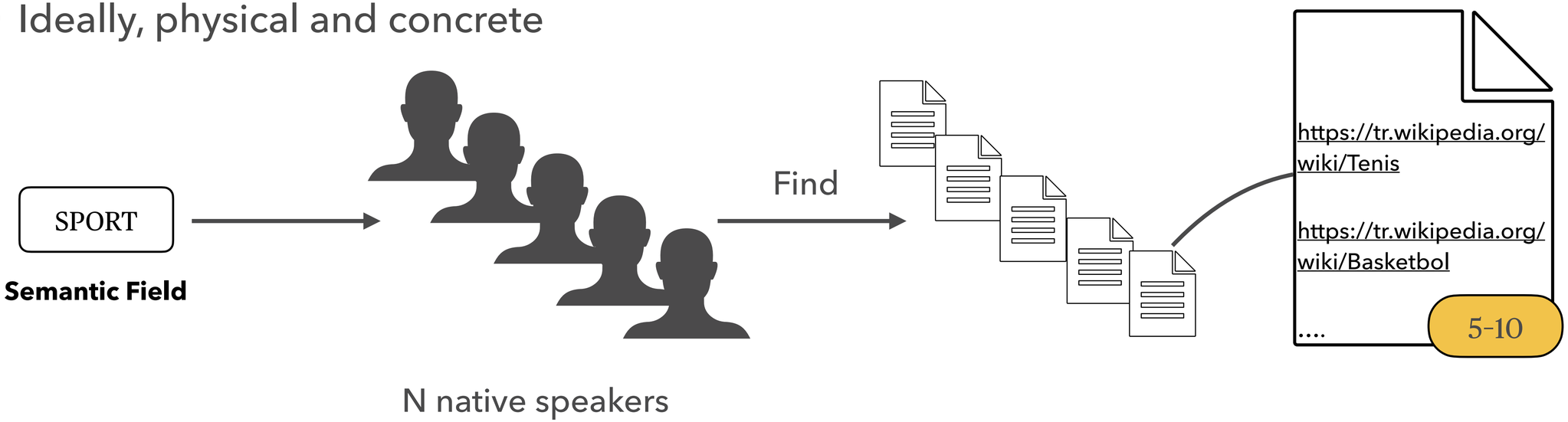
Defined by native speakers

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

Step 1. Language-Specific Concept Selection

Defined by native speakers

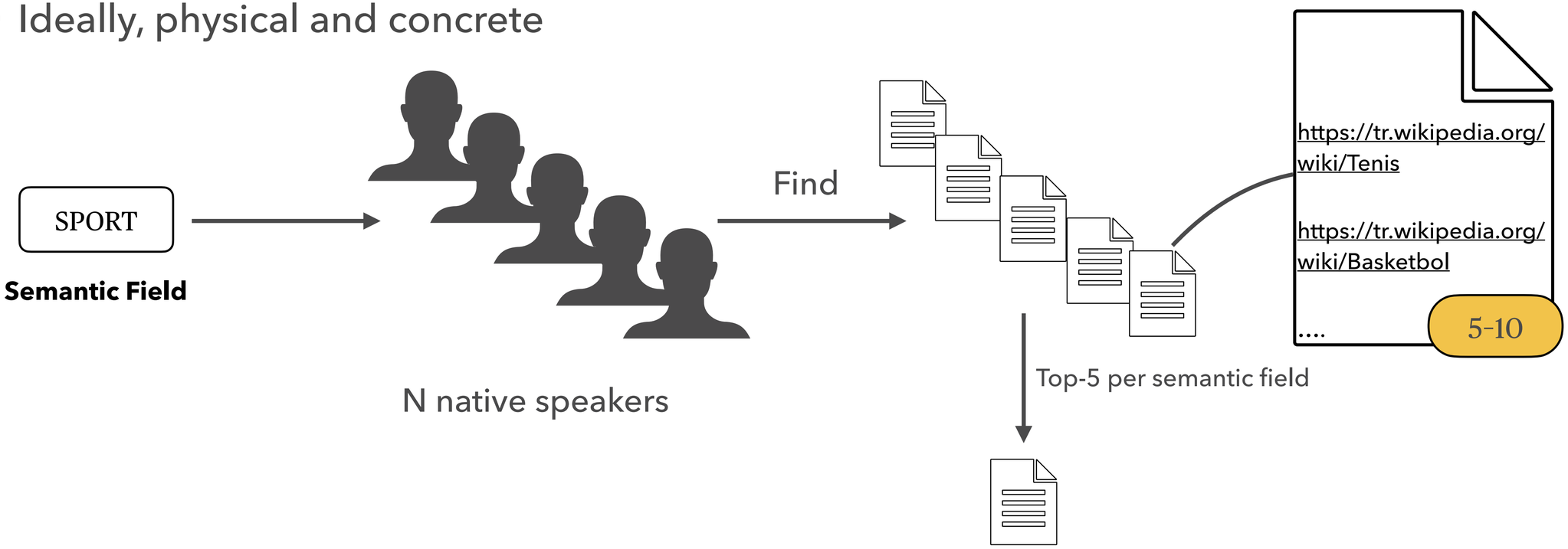
- Commonly seen or representative in their culture
- Ideally, physical and concrete



Step 1. Language-Specific Concept Selection

Defined by native speakers

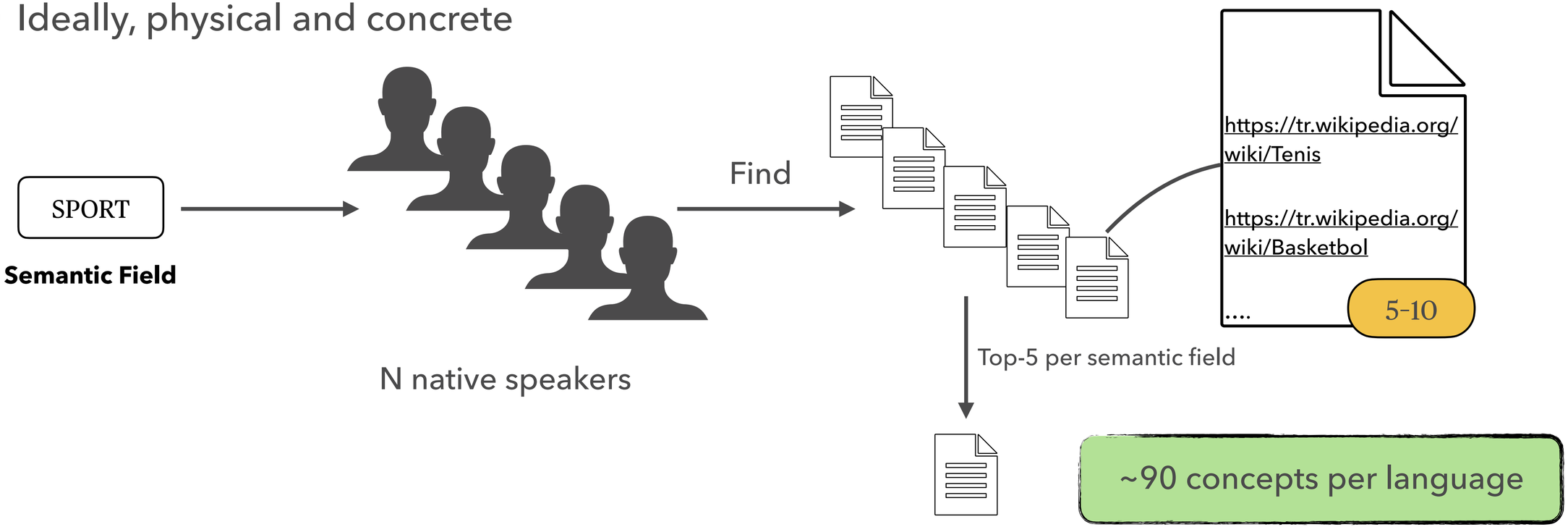
- Commonly seen or representative in their culture
- Ideally, physical and concrete



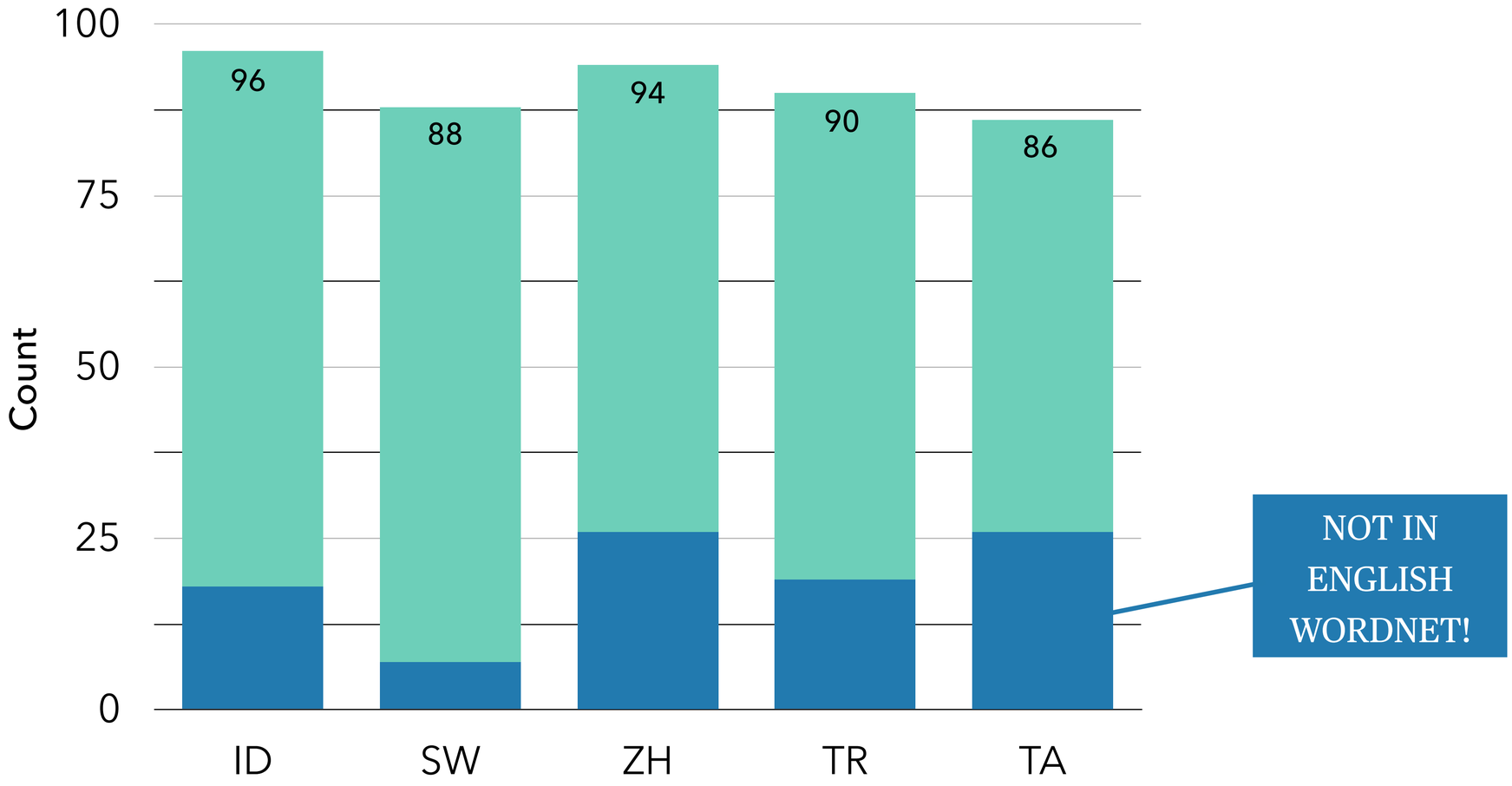
Step 1. Language-Specific Concept Selection

Defined by native speakers

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



Overview of Resulting Concepts



Step 2. Image Collection

Collected by native speakers

- Representative of the language population
- NLVR2 (Suhr et al. ACL 2019) requirements

Step 2. Image Collection

Collected by native speakers

- Representative of the language population
- NLVR2 (Suhr et al. ACL 2019) requirements

  1. Contains more than one instance of the concept

MaRVL-zh 花椰菜 (Cauliflower)

Step 2. Image Collection

Collected by native speakers

- Representative of the language population
- NLVR2 (Suhr et al. ACL 2019) requirements

  1. Contains more than one instance of the concept

  2. Shows an instance of the concept interacting with other objects

MaRVL-zh 花椰菜 (Cauliflower)

MaRVL-ta மோர் (Buttermilk)

Step 2. Image Collection

Collected by native speakers

- Representative of the language population
- NLVR2 (Suhr et al. ACL 2019) requirements

1. Contains more than one instance of the concept

2. Shows an instance of the concept interacting with other objects

3. Shows an instance of the concept performing an activity

MaRVL-zh 花椰菜 (Cauliflower)

MaRVL-ta மோர் (Buttermilk)

MaRVL-sw Jembe (Shovel)

Step 2. Image Collection

Collected by native speakers

- Representative of the language population
- NLVR2 (Suhr et al. ACL 2019) requirements

1. Contains more than one instance of the concept

2. Shows an instance of the concept interacting with other objects

3. Shows an instance of the concept performing an activity

4. Displays a set of diverse objects or features

MaRVL-zh 花椰菜 (Cauliflower)

MaRVL-ta மோர் (Buttermilk)

MaRVL-sw Jembe (Shovel)

MaRVL-tr Rakı (Raki)

Step 3. Language Annotation

Written by native speakers

Step 3. Language Annotation

Written by native speakers

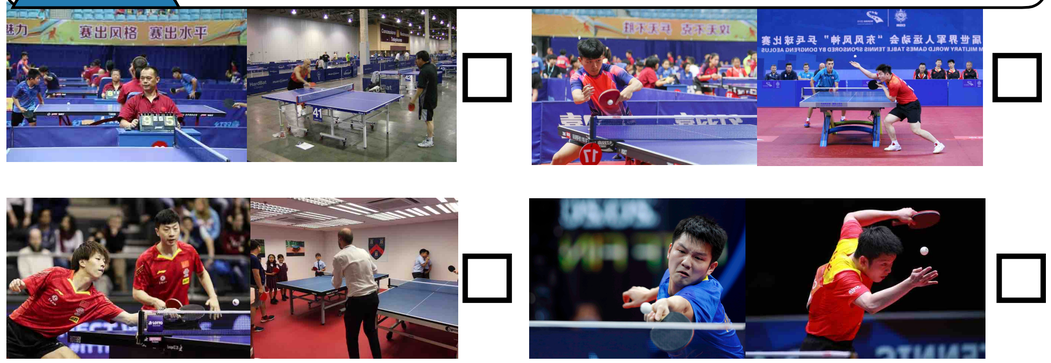
MATCH 4 PAIRS AT RANDOM

Step 3. Language Annotation

Written by native speakers

MATCH 4 PAIRS AT RANDOM

WRITE CAPTION TRUE ONLY FOR 2 PAIRS

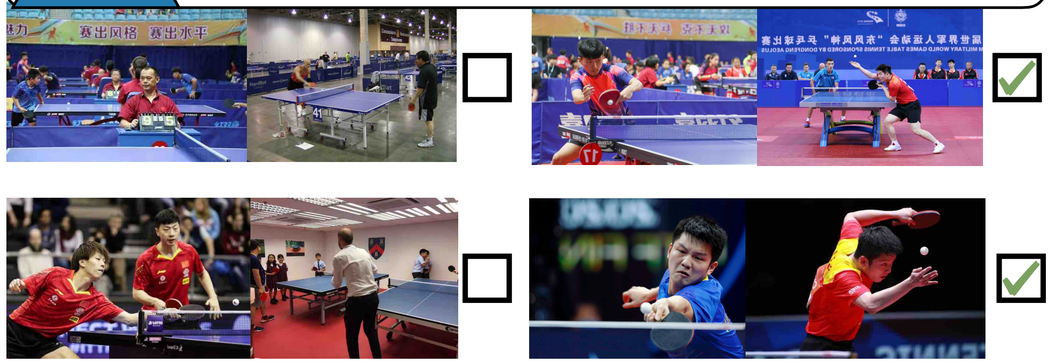


Step 3. Language Annotation

Written by native speakers

MATCH 4 PAIRS AT RANDOM

WRITE CAPTION TRUE ONLY FOR 2 PAIRS



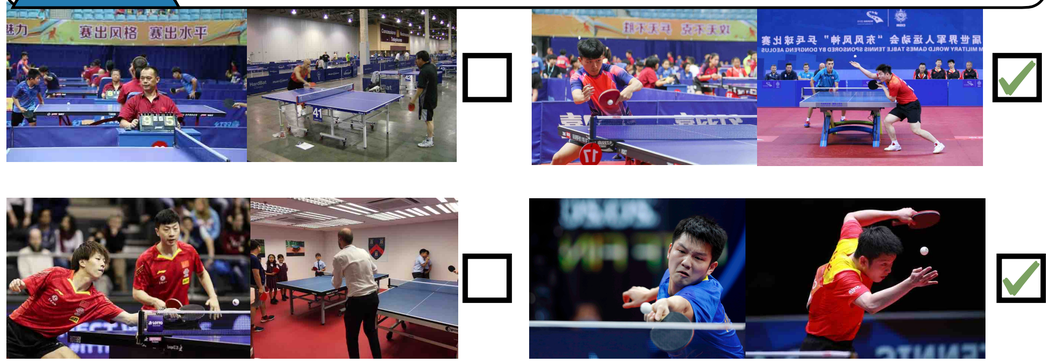
右图中的人在发球，左图中的人在接球。

Step 3. Language Annotation

Written by native speakers

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。

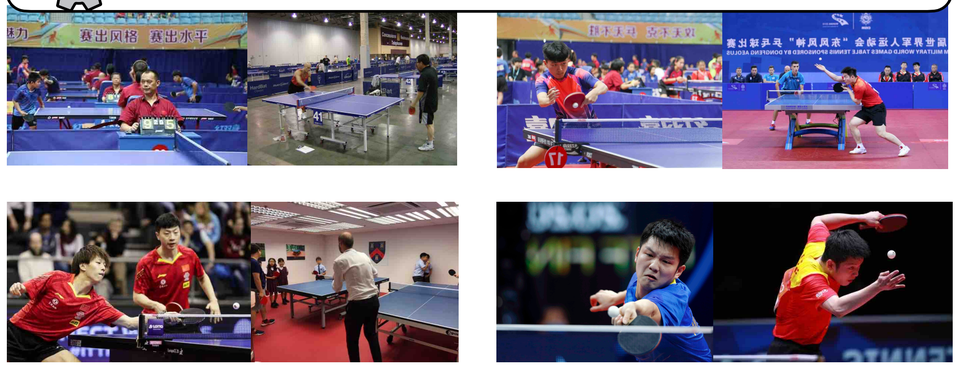
VALIDATE ANNOTATIONS

右图中的人在发球，左图中的人在接球。

Step 3. Language Annotation

Written by native speakers

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。

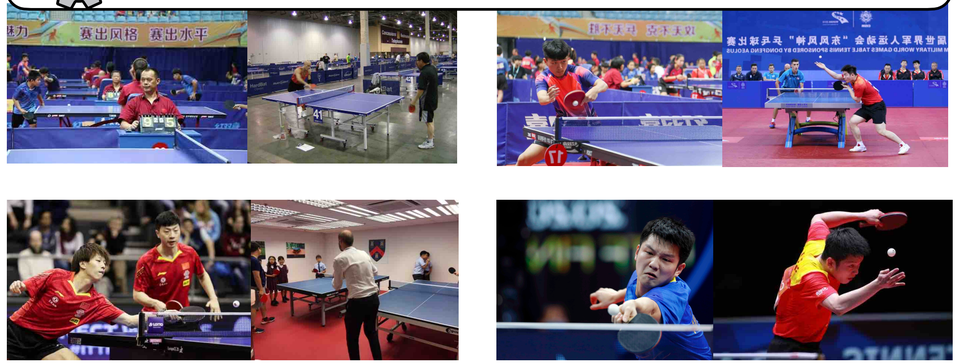
VALIDATE ANNOTATIONS

右图中的人在发球，左图中的人在接球。

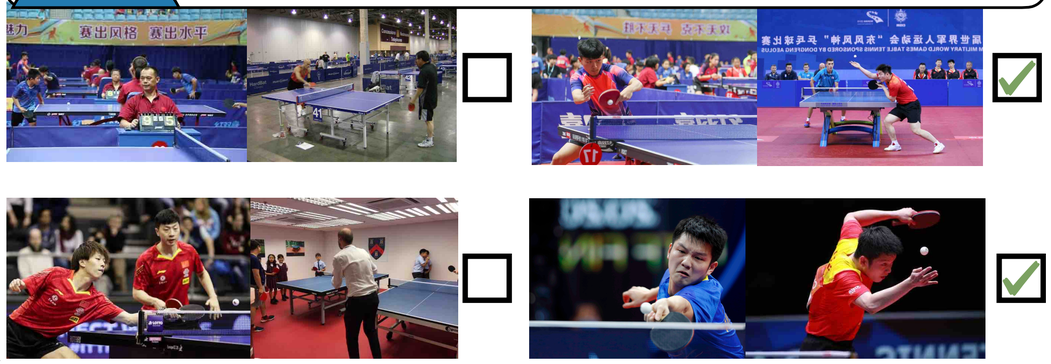
Step 3. Language Annotation

Written by native speakers

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。

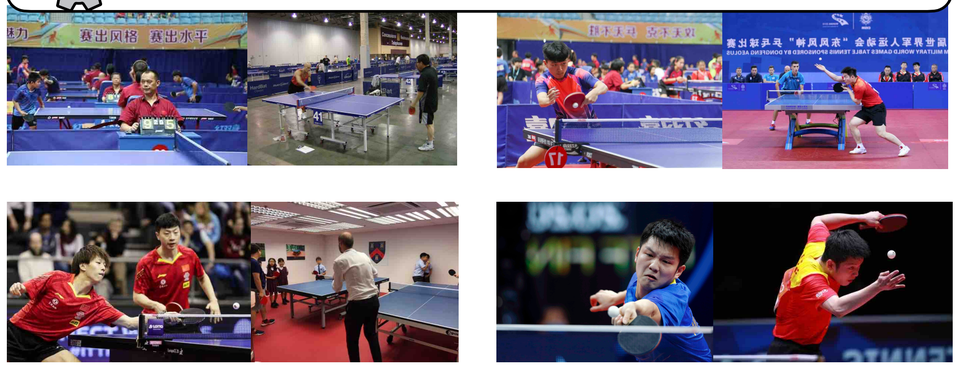
VALIDATE ANNOTATIONS

右图中的人在发球，左图中的人在接球。

Step 3. Language Annotation

Written by native speakers

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。

VALIDATE ANNOTATIONS

右图中的人在发球，左图中的人在接球。

FINAL VALIDATION

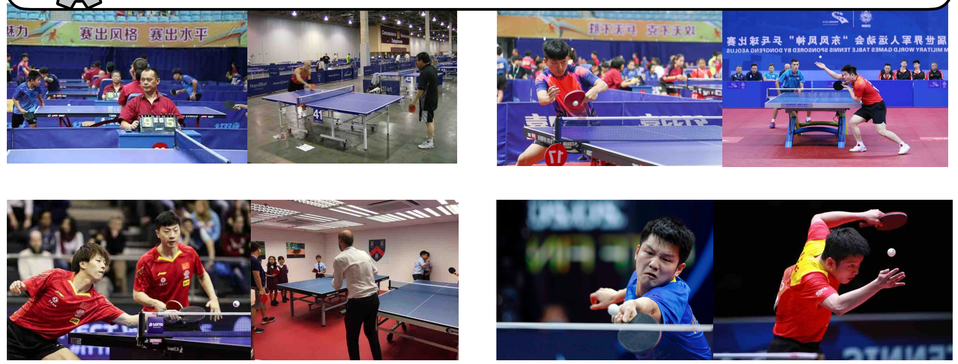
右图中的人在发球，左图中的人在接球。

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

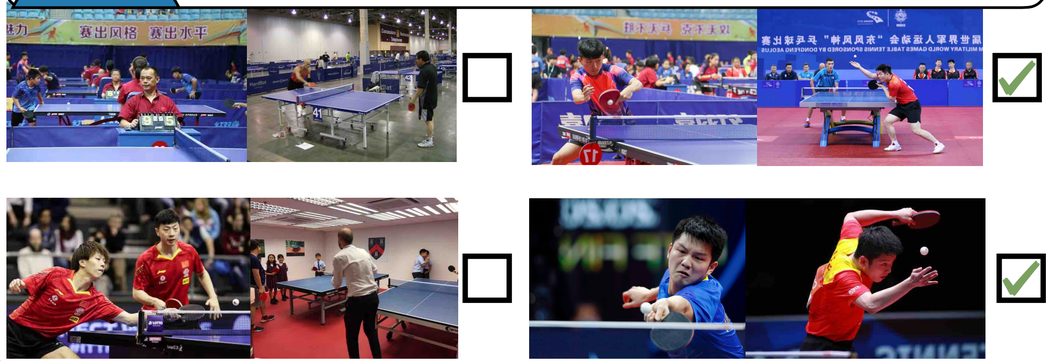
Step 3. Language Annotation

Written by native speakers

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。

VALIDATE ANNOTATIONS

右图中的人在发球，左图中的人在接球。

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.)

Fleiss' kappa: 93%

Dataset Examples

MaRVL-tr **Kanun (çalgı)**

Görsellerden birinde dizlerinde kanun bulunan birden çok insan var

(In one of the images, there are multiple people with qanuns on their knees)

Dataset Examples

MaRVL-tr **Kanun (çalgı)**

Görsellerden birinde dizlerinde kanun bulunan birden çok insan var

(In one of the images, there are multiple people with qanuns on their knees)

Label: `True`

Dataset Examples

MaRVL-tr **Kanun (çalgı)**

Görsellerden birinde dizlerinde kanun bulunan birden çok insan var

(In one of the images, there are multiple people with qanuns on their knees)

Label: True

MaRVL-ta **மை (Vada)**

இரண்டு படங்களிலும் நிறைய மசால் வடைகள் உள்

(Both images contain a lot of masala vadas)

Dataset Examples

MaRVL-tr **Kanun (çalgı)**

Görsellerden birinde dizlerinde kanun bulunan birden çok insan var

(In one of the images, there are multiple people with qanuns on their knees)

Label: True

MaRVL-ta **மை (Vada)**

இரண்டு படங்களிலும் நிறைய மசால் வடைகள் உள்

(Both images contain a lot of masala vadas)

Label: False

Summary Statistics

- 5560 data points across 5 languages
- 423 concepts (96 not in WordNet)
- 1390 unique captions

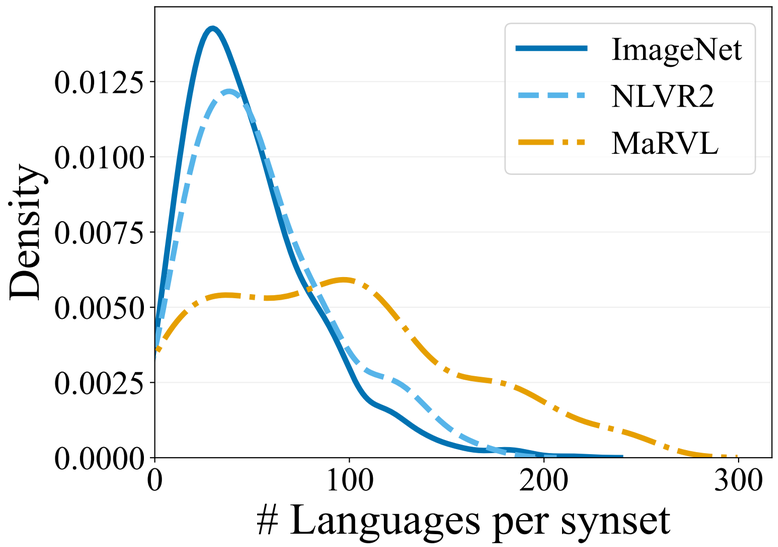
Summary Statistics

- 5560 data points across 5 languages
- 423 concepts (96 not in WordNet)
- 1390 unique captions

MaRVL covers more languages

MaRVL covers more language families

MaRVL covers more macroareas



Modelling: Vision-and-Language BERTs

Vision & Language BERT Models

Vision & Language BERT Models

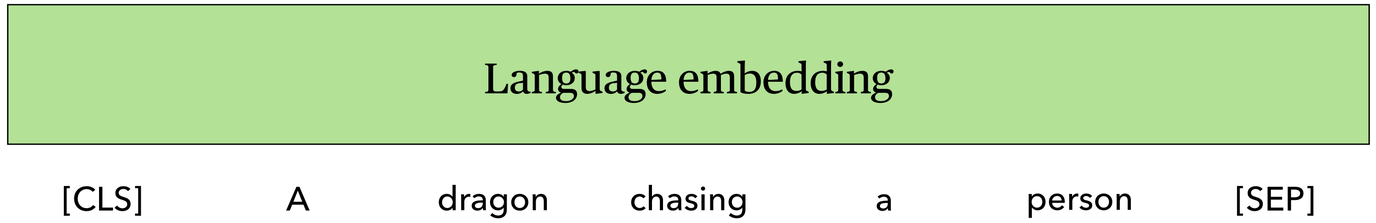
A dragon chasing a person

Vision & Language BERT Models

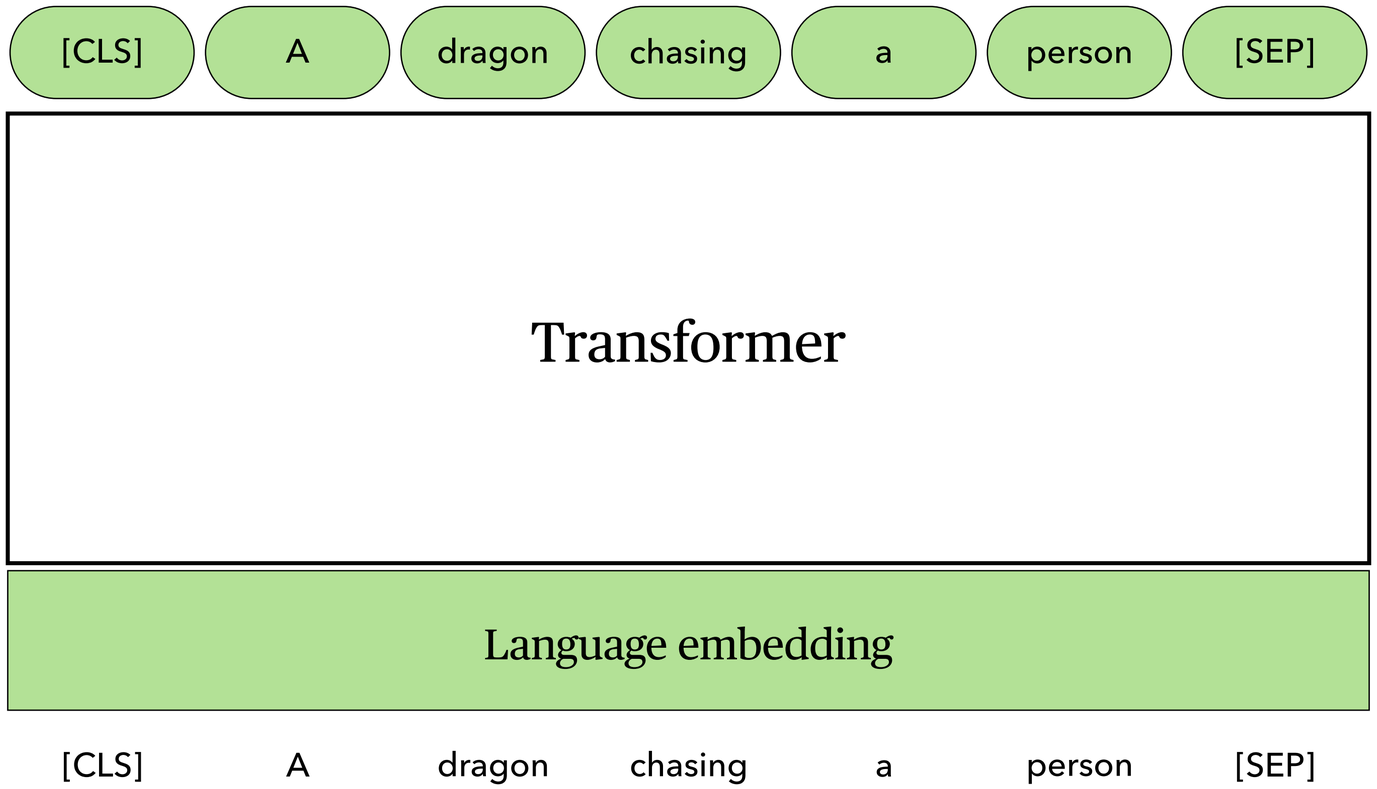
[CLS] A dragon chasing a person [SEP]

Pretrained Language Model Tokenizer

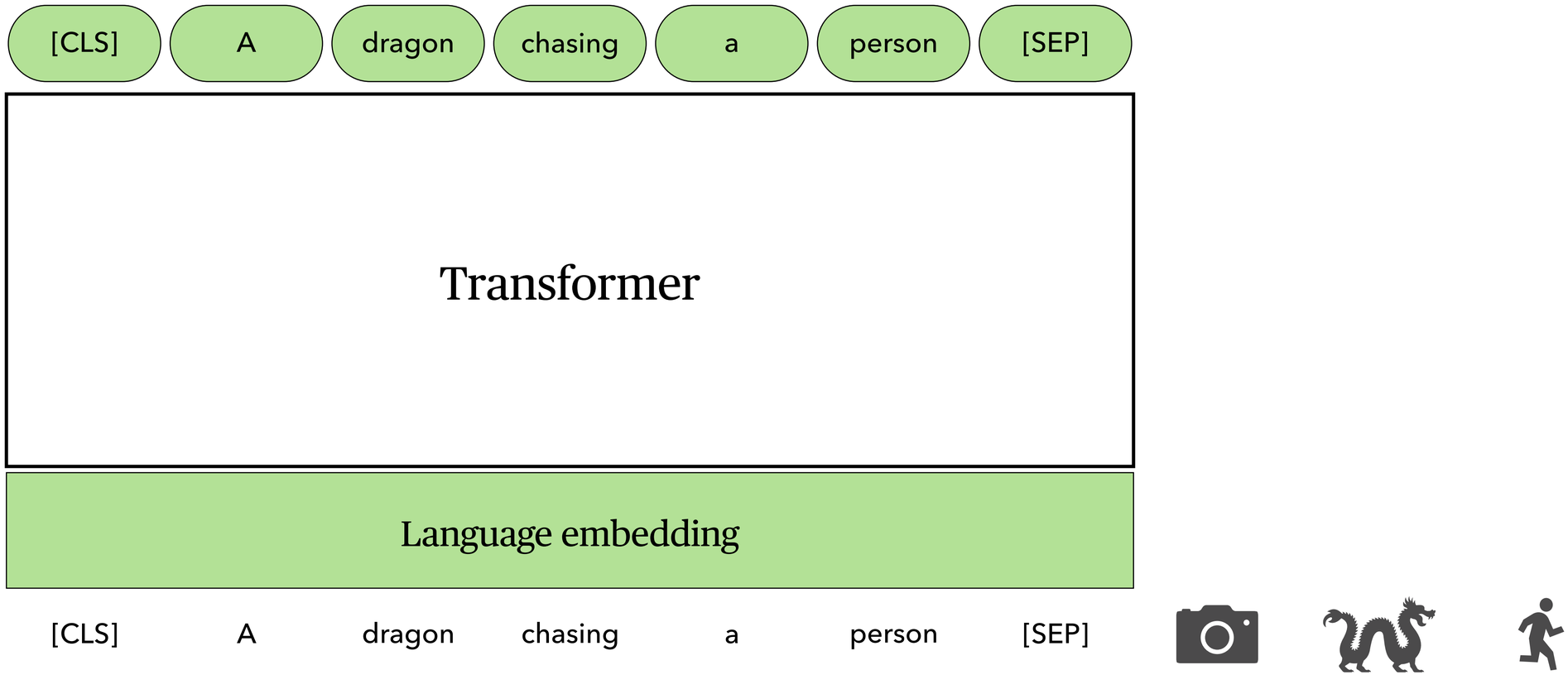
Vision & Language BERT Models



Vision & Language BERT Models



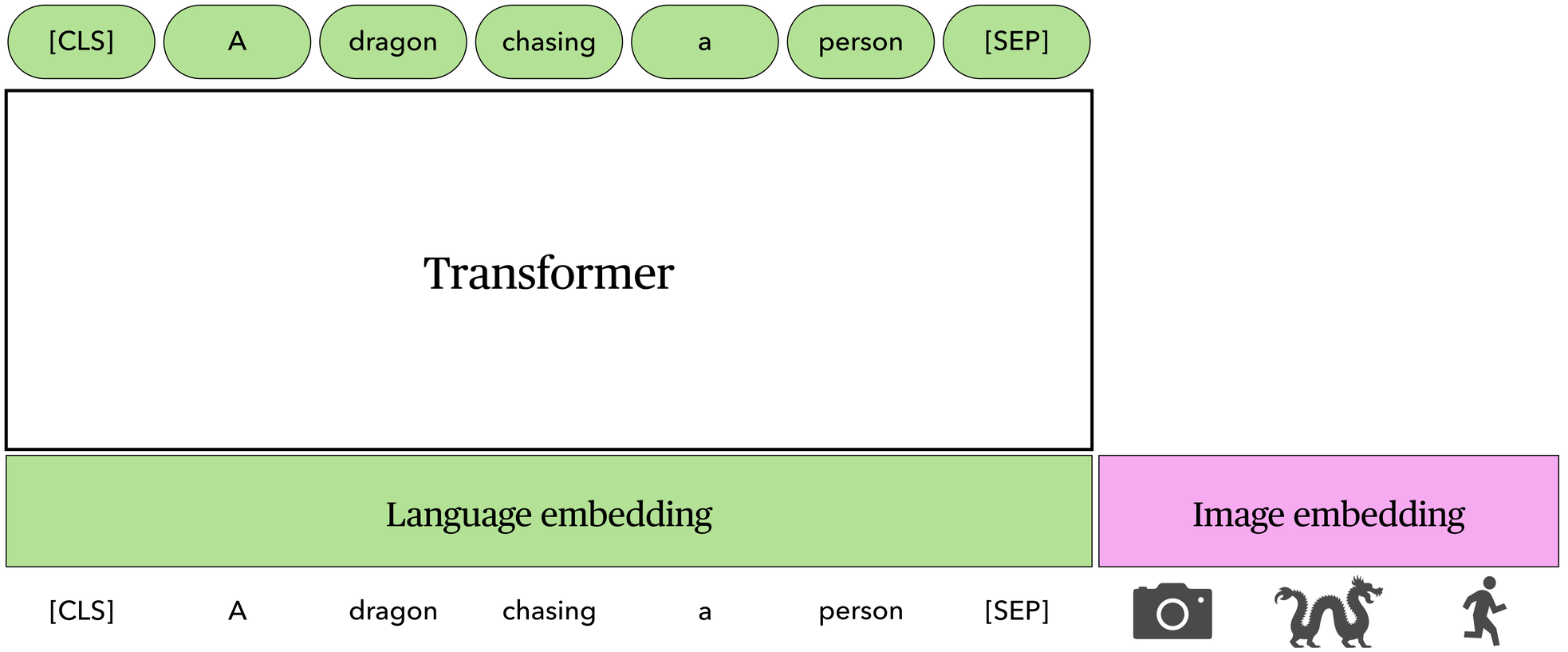
Vision & Language BERT Models



Pretrained Language Model Tokenizer

36 Regions of Interest (Faster R-CNN)

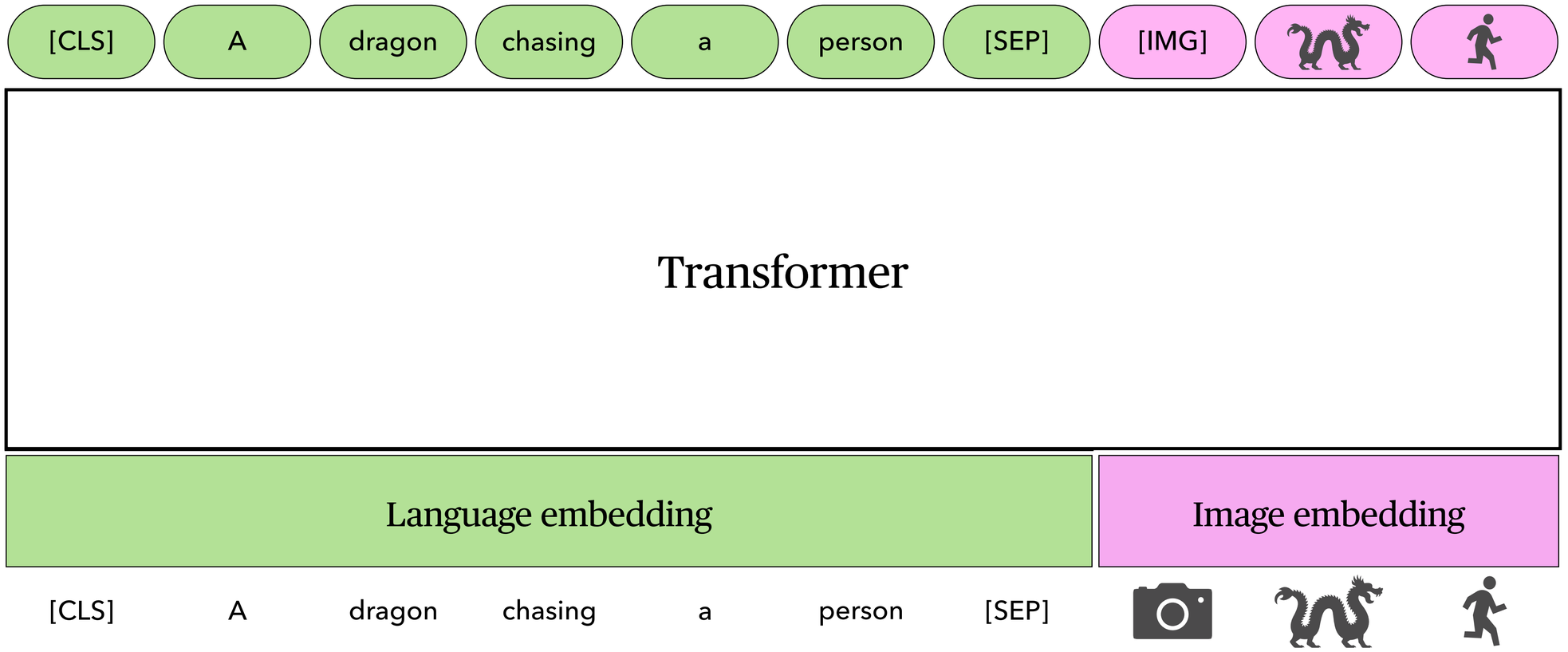
Vision & Language BERT Models



Pretrained Language Model Tokenizer

36 Regions of Interest (Faster R-CNN)

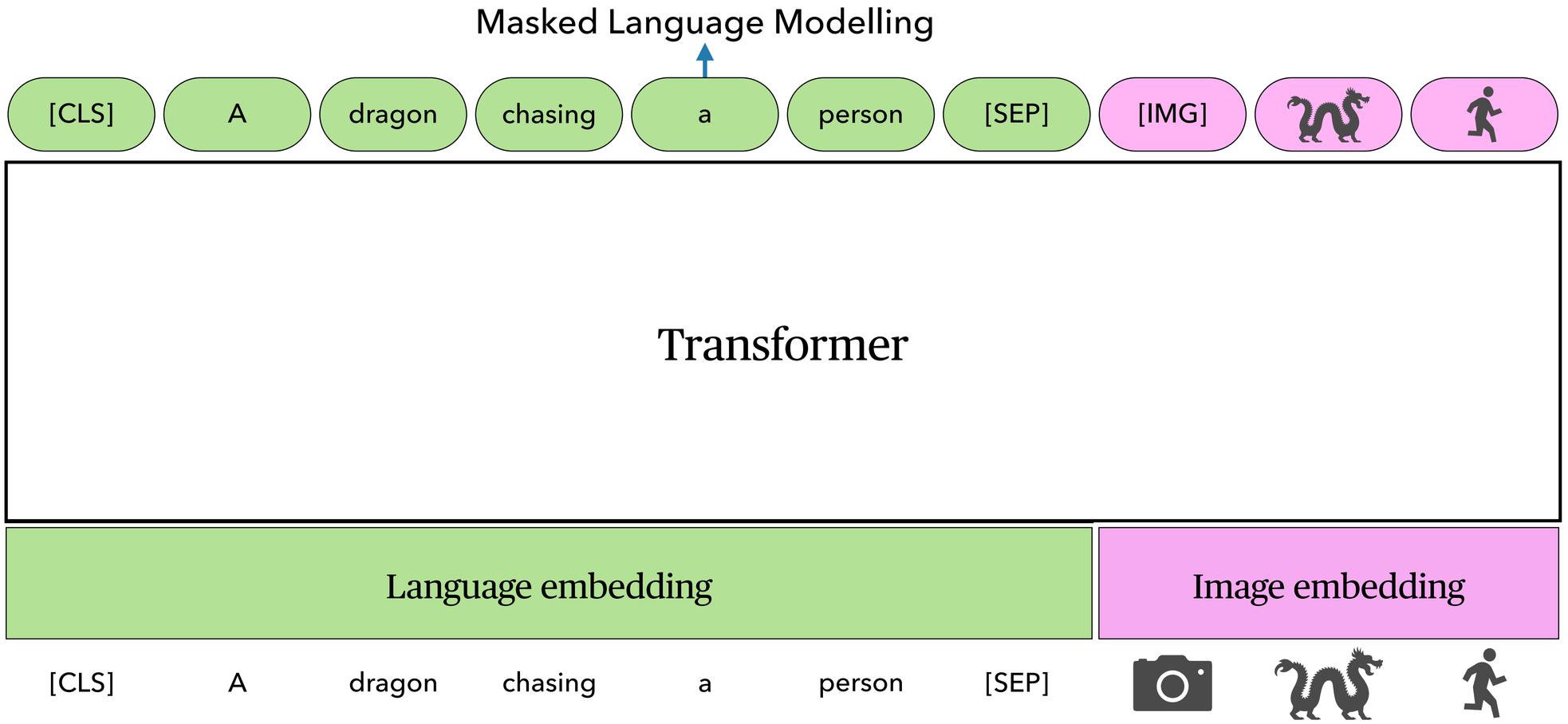
Vision & Language BERT Models



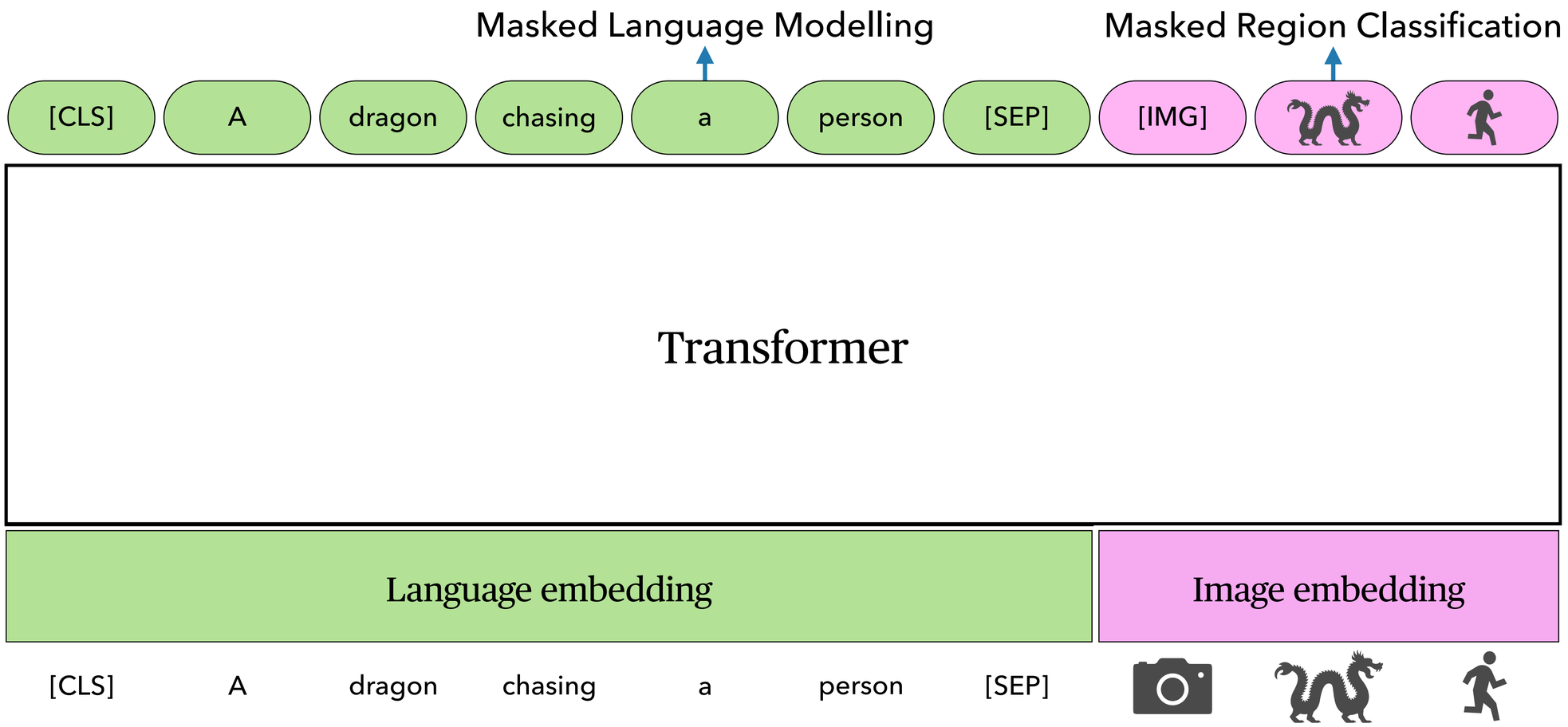
Pretrained Language Model Tokenizer

36 Regions of Interest (Faster R-CNN)

Vision & Language BERT Models



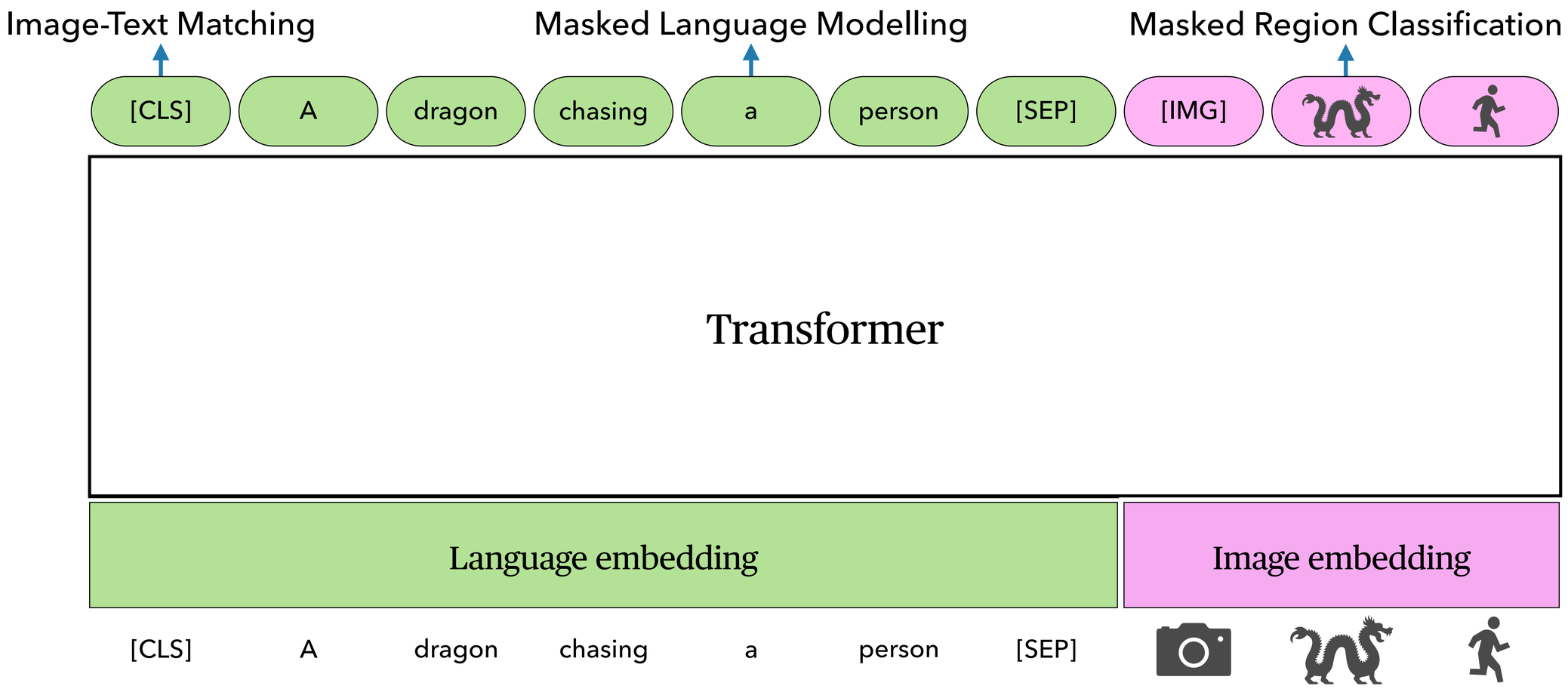
Vision & Language BERT Models



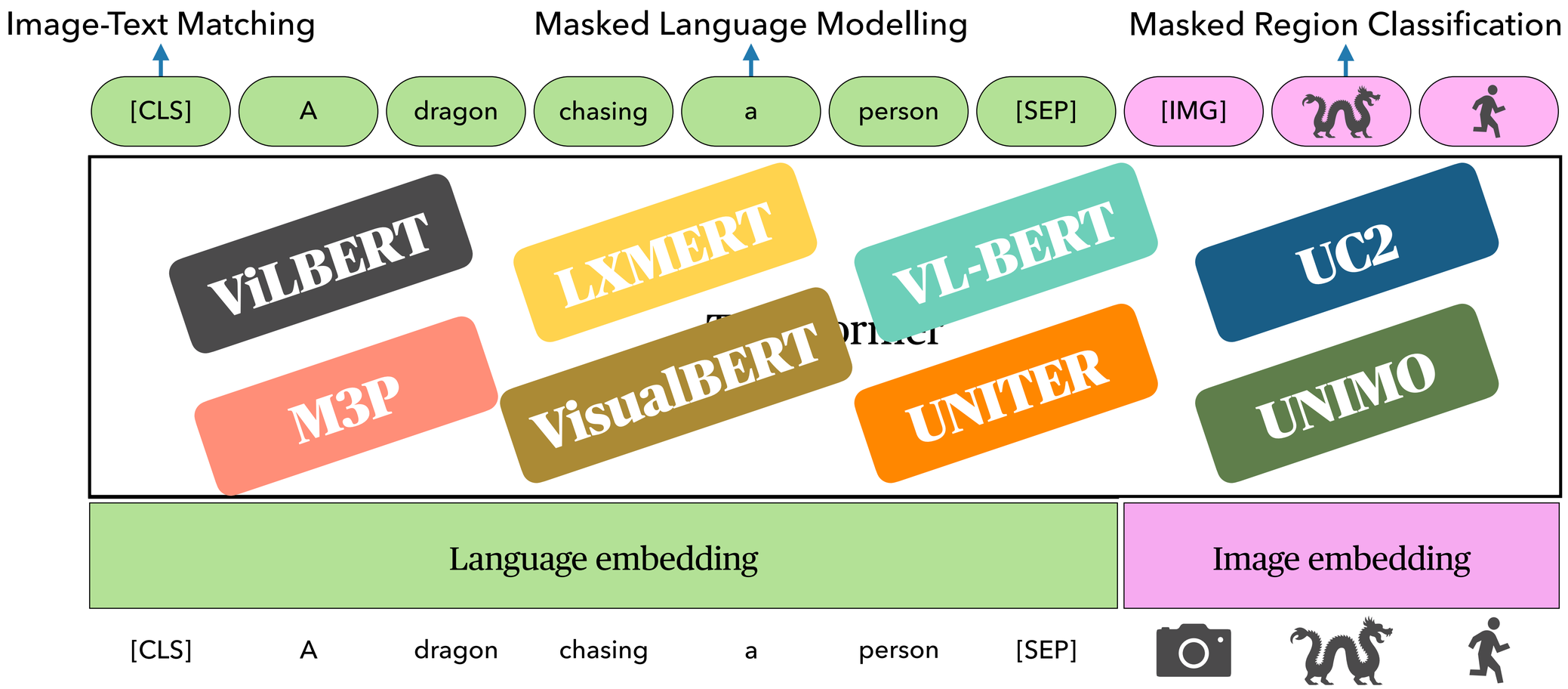
Pretrained Language Model Tokenizer

36 Regions of Interest (Faster R-CNN)

Vision & Language BERT Models



Vision & Language BERT Models



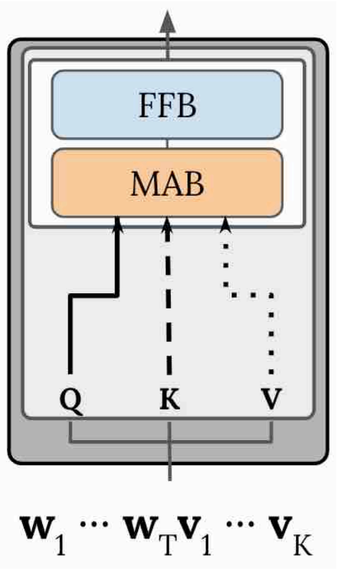
Single- & Dual-Stream Architectures

Meta-analysis: Bugliarello et al. TACL 2021. Multimodal Pretraining Unmasked: A Meta-Analysis and a Unified Framework of Vision-and-Language BERTs.

Single- & Dual-Stream Architectures

Single-Stream

- Concat image–text inputs

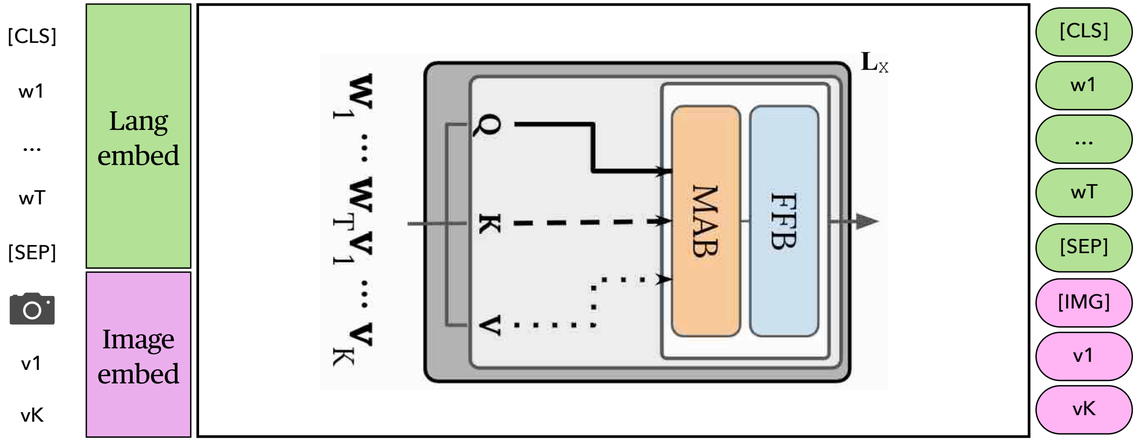


Meta-analysis: Bugliarello et al. TACL 2021. Multimodal Pretraining Unmasked: A Meta-Analysis and a Unified Framework of Vision-and-Language BERTs.

Single- & Dual-Stream Architectures

Single-Stream

- Concat image–text inputs



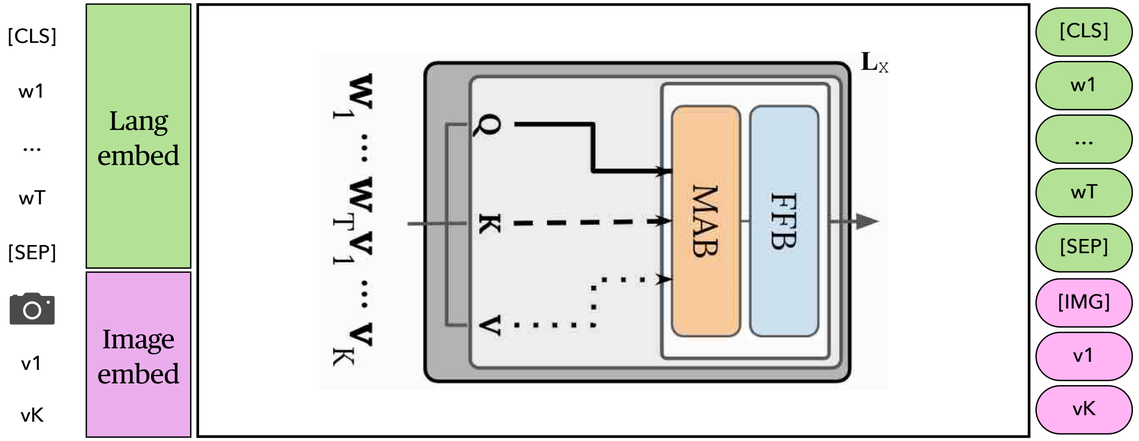
Single- & Dual-Stream Architectures

Single-Stream

- Concat image–text inputs

Dual-Stream

1. Image and text independently



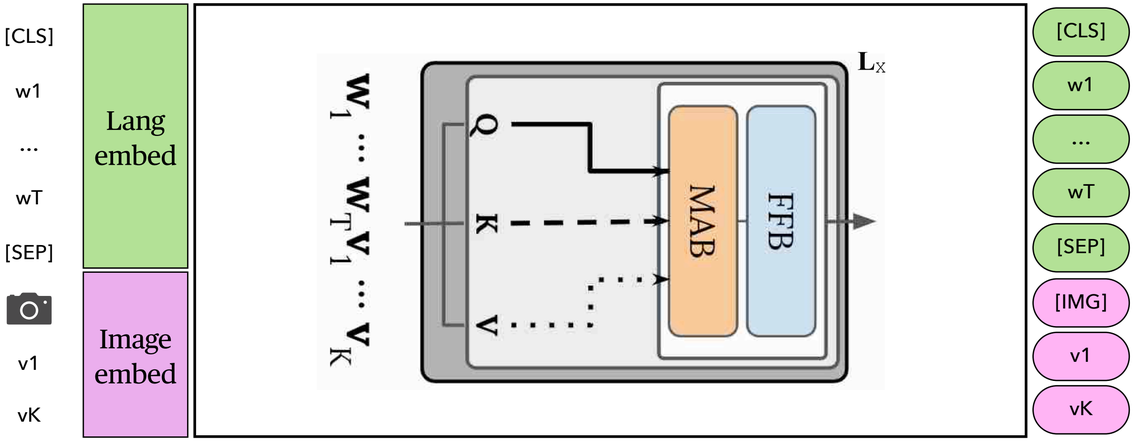
Single- & Dual-Stream Architectures

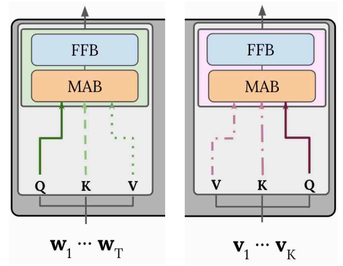
Single-Stream

- Concat image–text inputs

Dual-Stream

1. Image and text independently





Meta-analysis: Bugliarello et al. TACL 2021. Multimodal Pretraining Unmasked: A Meta-Analysis and a Unified Framework of Vision-and-Language BERTs.

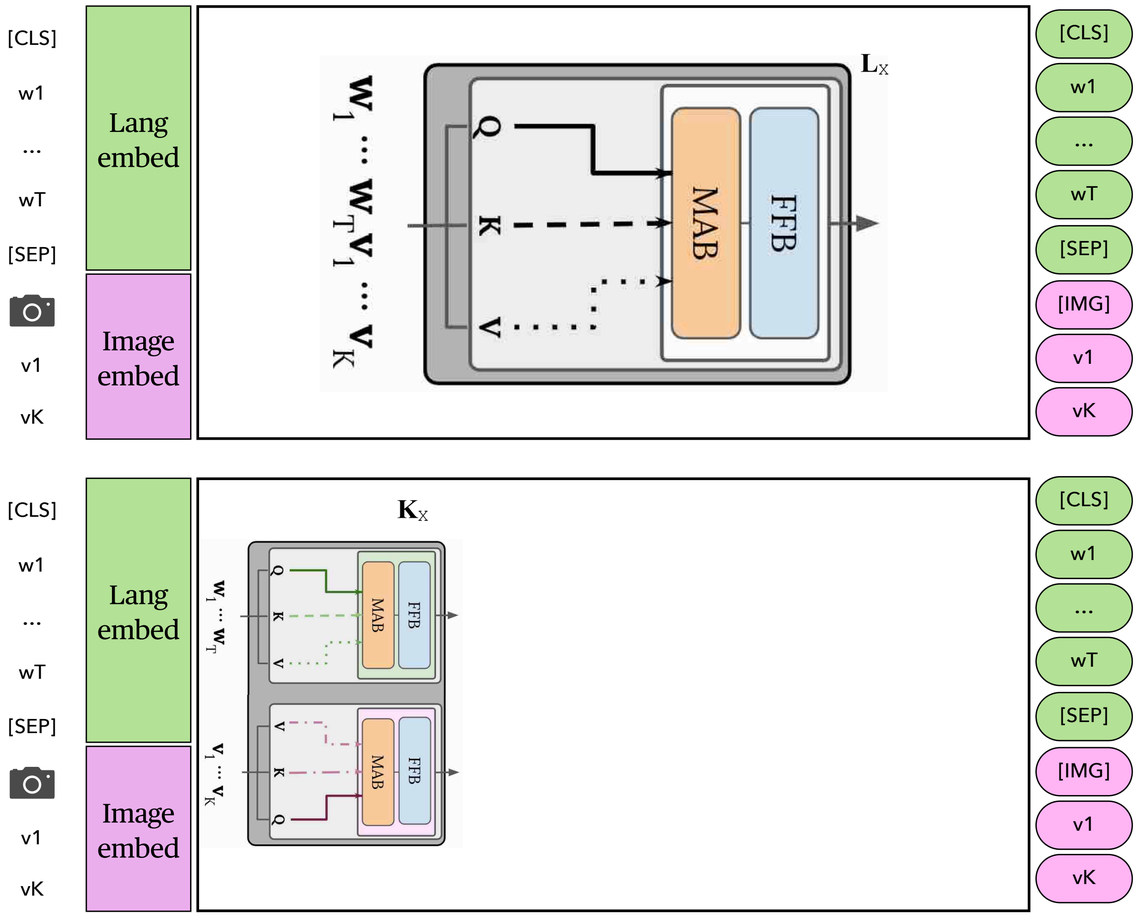
Single- & Dual-Stream Architectures

Single-Stream

- Concat image–text inputs

Dual-Stream

1. Image and text independently



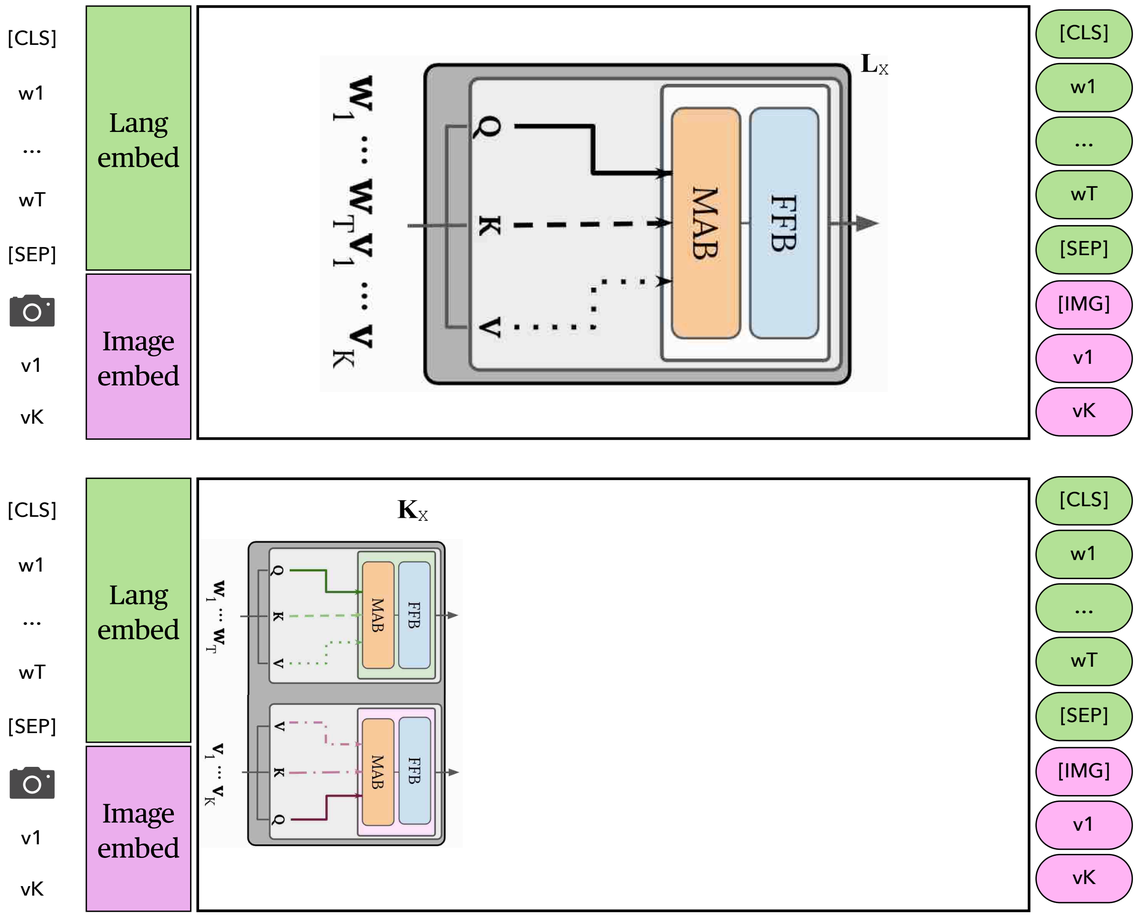
Single- & Dual-Stream Architectures

Single-Stream

- Concat image–text inputs

Dual-Stream

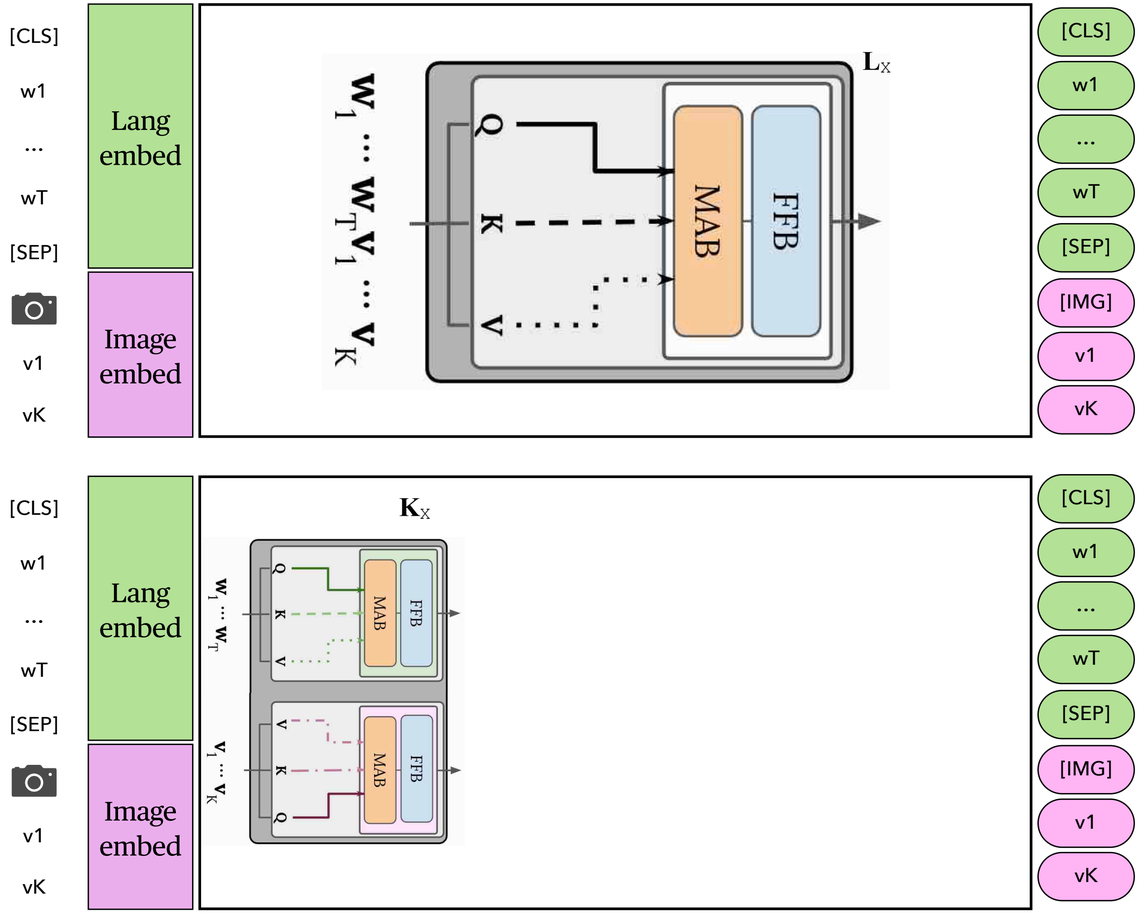
1. Image and text independently
2. Cross-modal layers



Single- & Dual-Stream Architectures

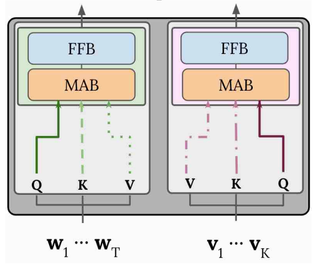
Single-Stream

- Concat image–text inputs



Dual-Stream

1. Image and text independently
2. Cross-modal layers
   - Intra-modal

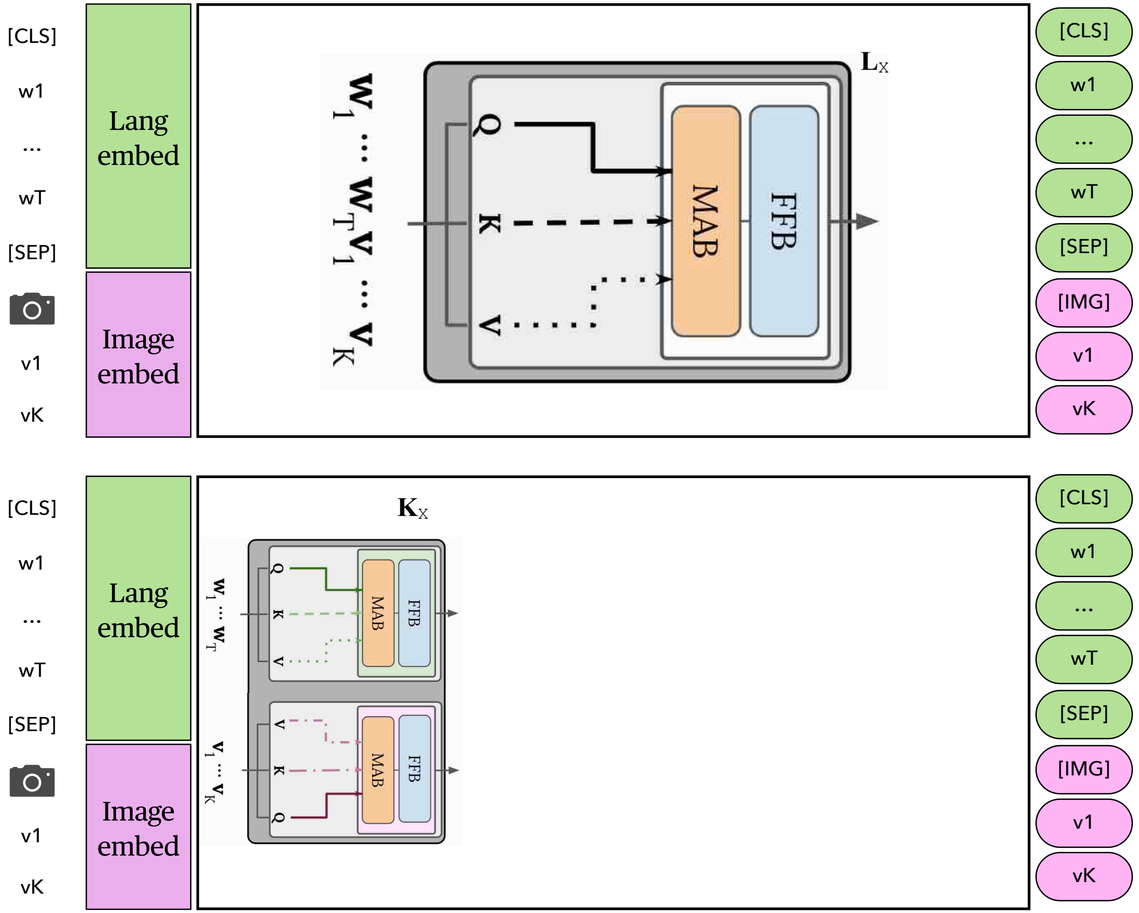


Meta-analysis: Bugliarello et al. TACL 2021. Multimodal Pretraining Unmasked: A Meta-Analysis and a Unified Framework of Vision-and-Language BERTs.

Single- & Dual-Stream Architectures

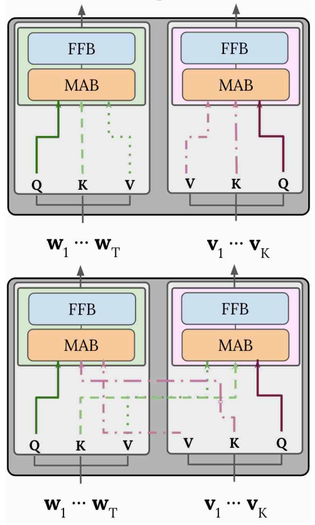
Single-Stream

- Concat image–text inputs



Dual-Stream

1. Image and text independently
2. Cross-modal layers
   - Intra-modal
   - Inter-modal

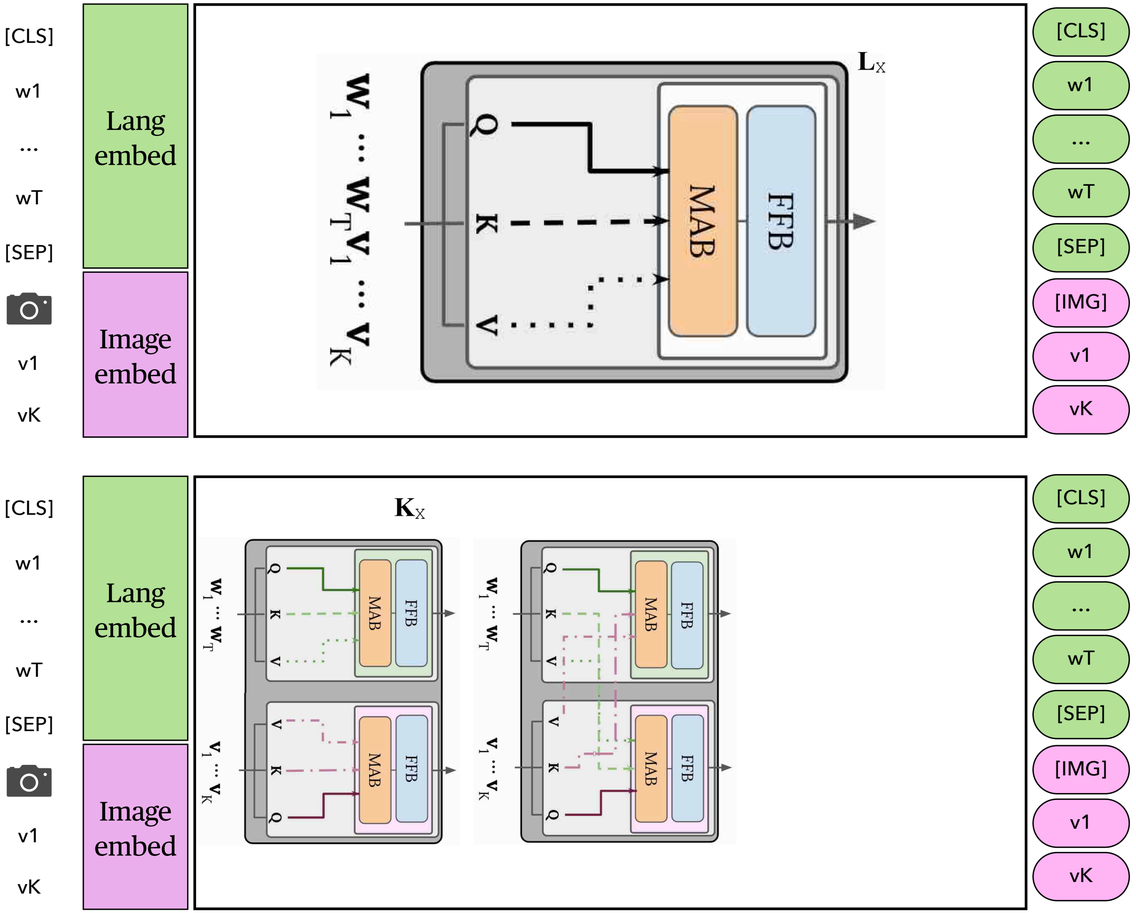


Meta-analysis: Bugliarello et al. TACL 2021. Multimodal Pretraining Unmasked: A Meta-Analysis and a Unified Framework of Vision-and-Language BERTs.

Single- & Dual-Stream Architectures

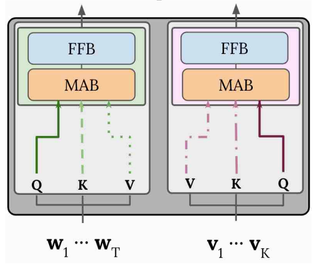
Single-Stream

- Concat image–text inputs



Dual-Stream

1. Image and text independently
2. Cross-modal layers
   - Intra-modal
   - Inter-modal



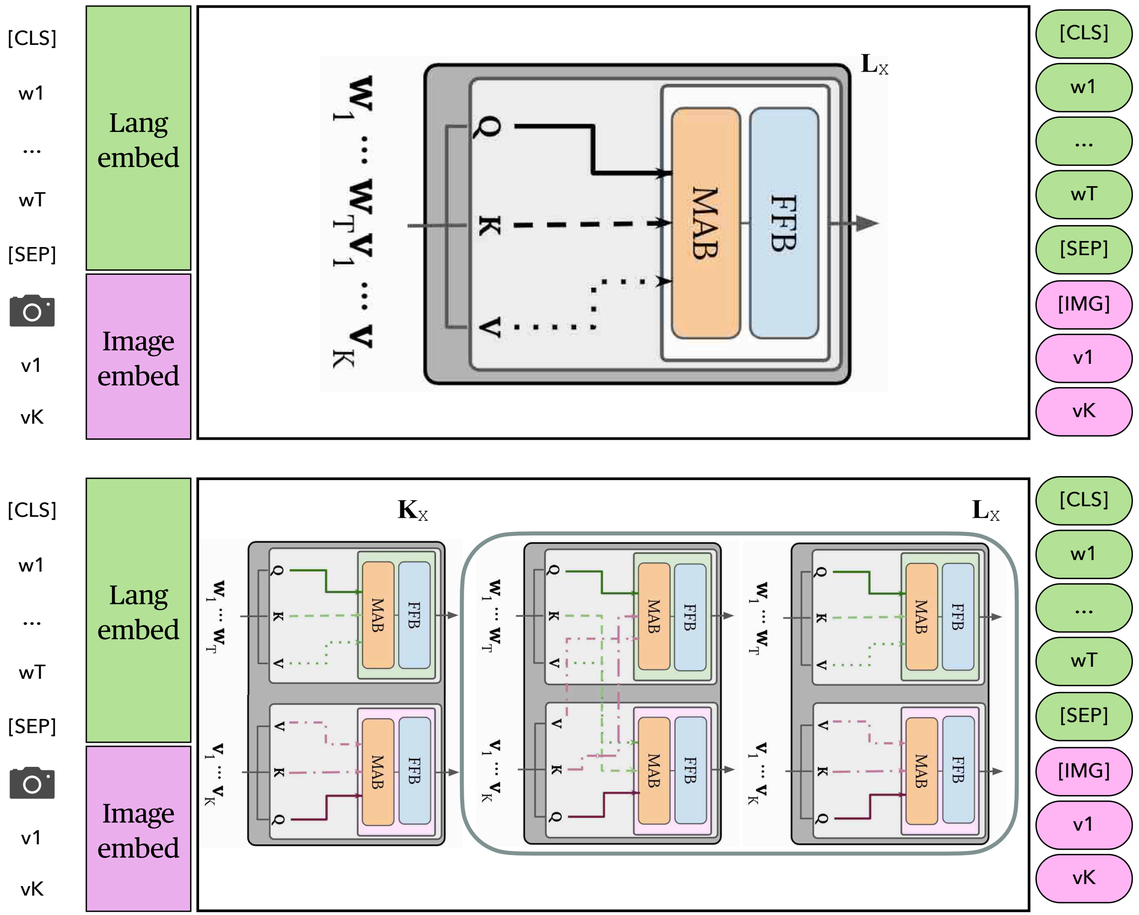
Single- & Dual-Stream Architectures

Single-Stream

- Concat image–text inputs

Dual-Stream

1. Image and text independently
2. Cross-modal layers
   - Intra-modal
   - Inter-modal



FAQ: Are these models actually cross-modal?

- Cross-modal Ablation tests how a missing modality affects model behaviour

FAQ: Are these models actually cross-modal?

- Cross-modal Ablation tests how a missing modality affects model behaviour

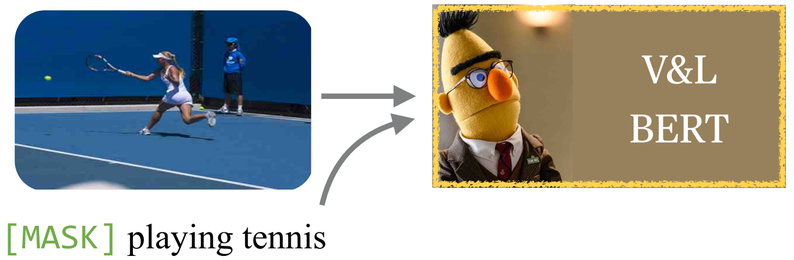
No ablation (**None**)

[MASK] playing tennis

FAQ: Are these models actually cross-modal?

- Cross-modal Ablation tests how a missing modality affects model behaviour

No ablation (**None**)



FAQ: Are these models actually cross-modal?

- Cross-modal Ablation tests how a missing modality affects model behaviour

No ablation (**None**)

FAQ: Are these models actually cross-modal?

- Cross-modal Ablation tests how a missing modality affects model behaviour

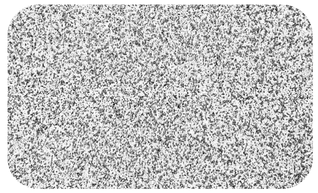
No ablation (**None**)

FAQ: Are these models actually cross-modal?

- Cross-modal Ablation tests how a missing modality affects model behaviour

No ablation (**None**)

Full ablation (**All**)



FAQ: Are these models actually cross-modal?

- Cross-modal Ablation tests how a missing modality affects model behaviour

No ablation (**None**)

Full ablation (**All**)

FAQ: Are these models actually cross-modal?

- Cross-modal Ablation tests how a missing modality affects model behaviour

No ablation (**None**)

Full ablation (**All**)

FAQ: Are these models actually cross-modal?

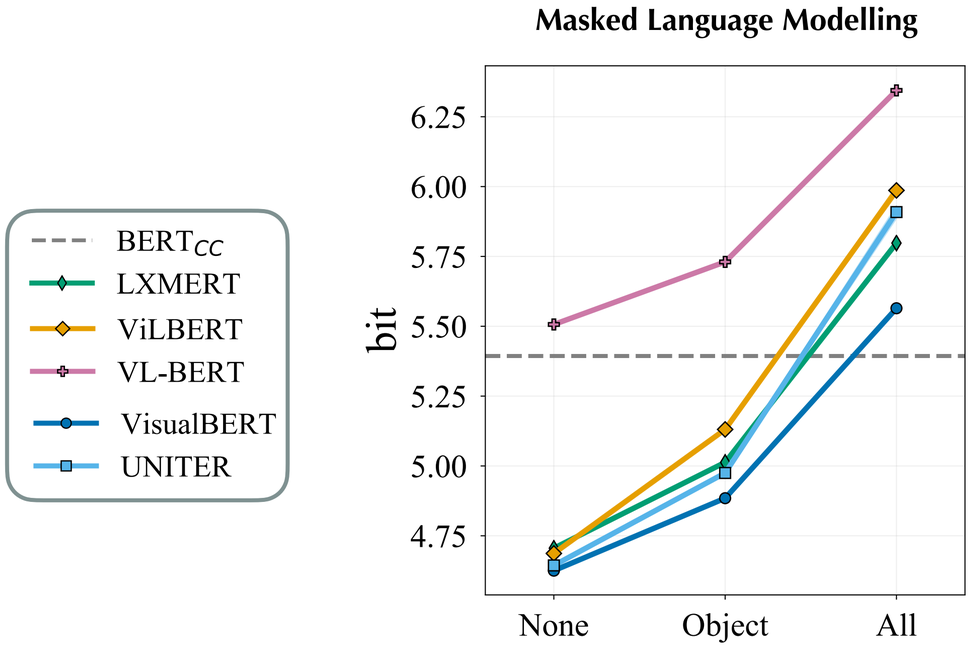
- Cross-modal Ablation tests how a missing modality affects model behaviour

No ablation (**None**)

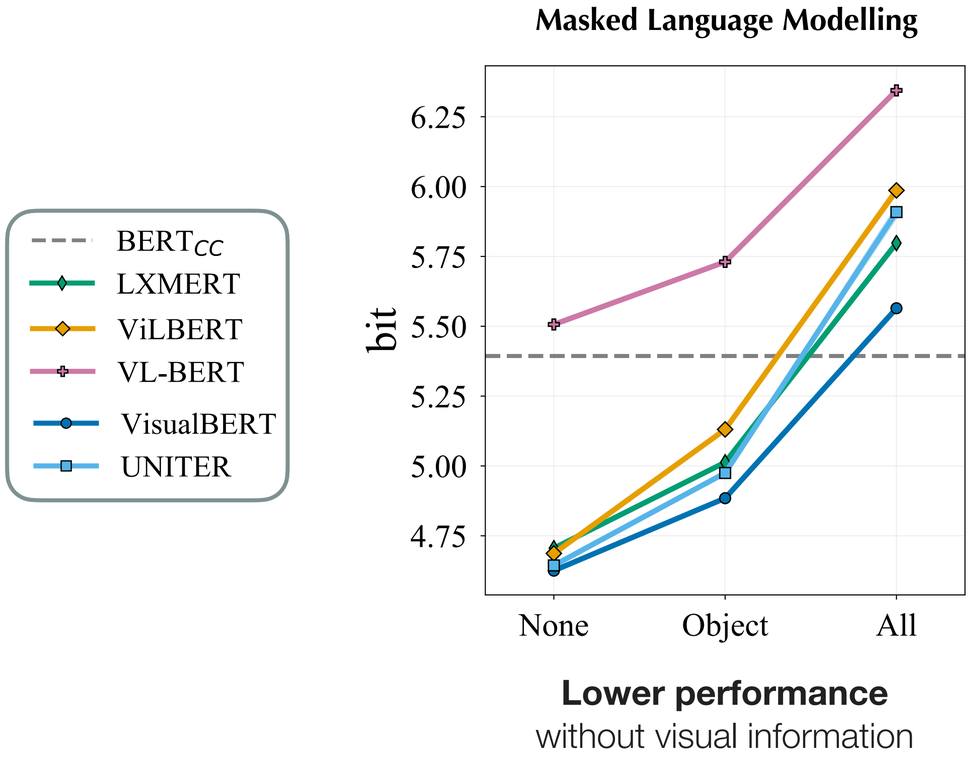
Object ablation (**Object**)

Full ablation (**All**)

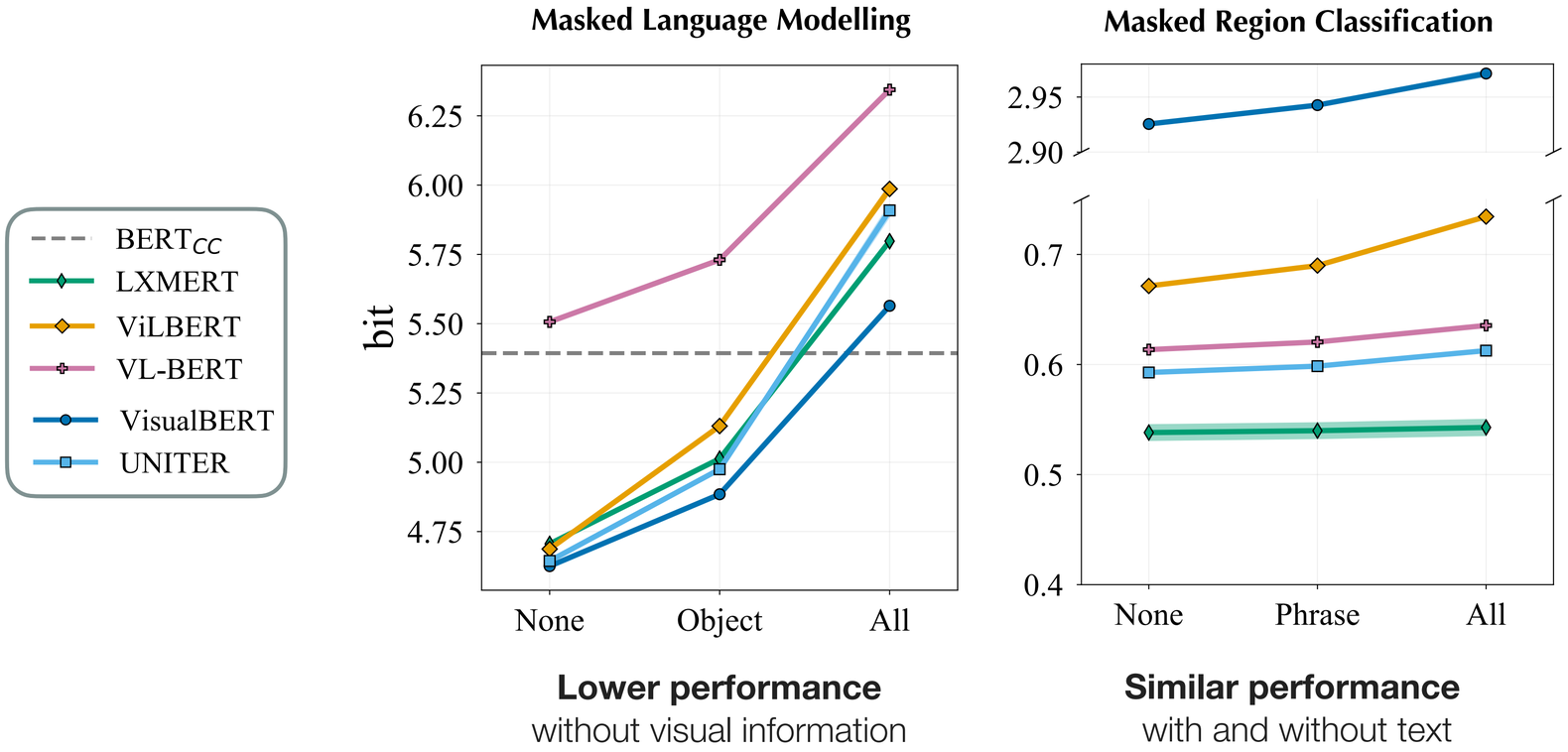
A: It's complicated



A: It's complicated



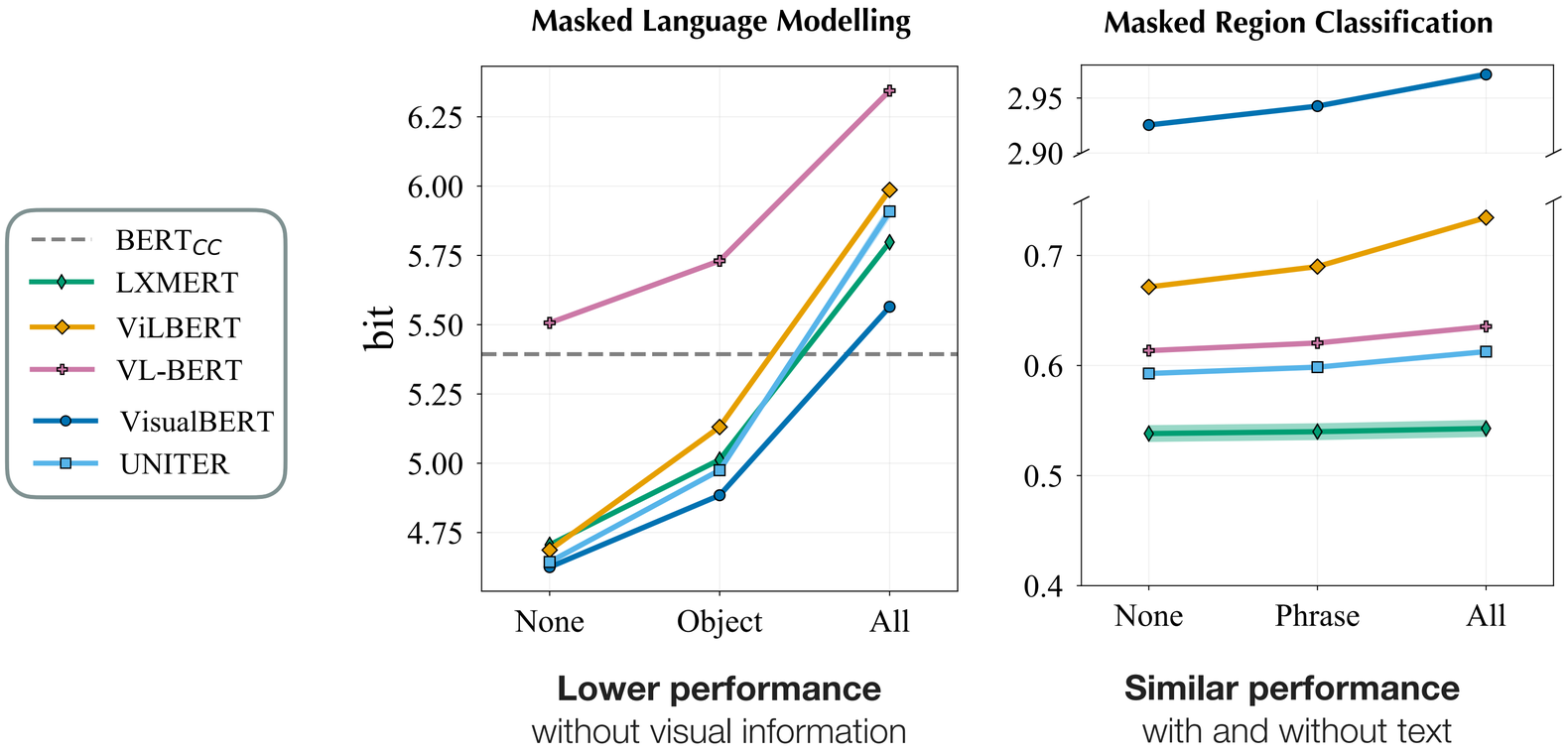
A: It's complicated



Lower performance
without visual information

Similar performance
with and without text

A: It's complicated

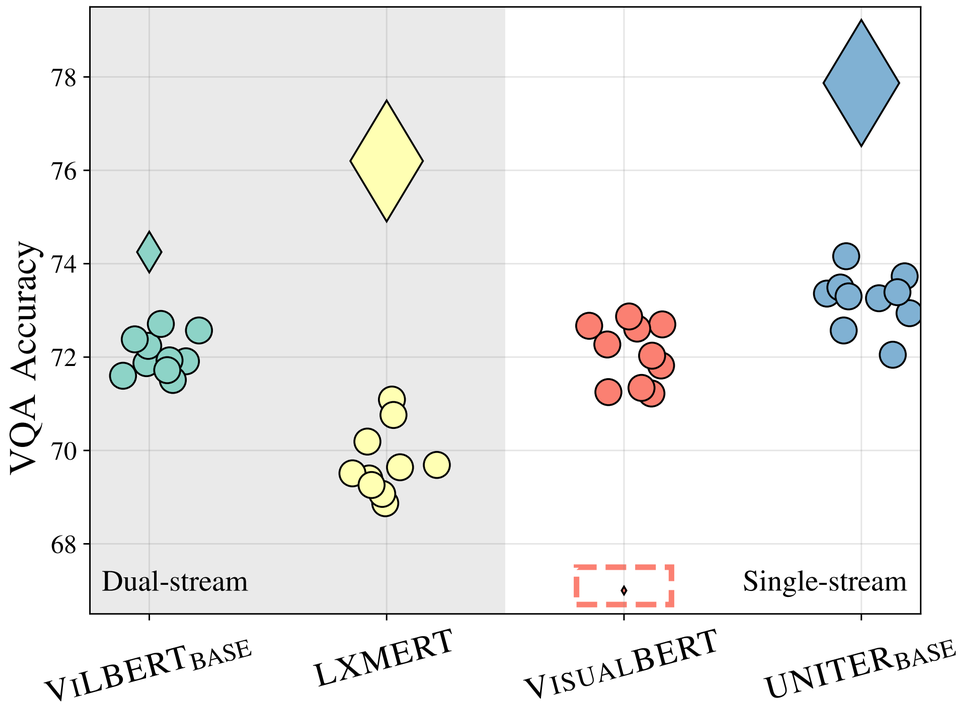


Lower performance
without visual information

Similar performance
with and without text

These models mainly learn to use vision-for-language

FAQ: Which V&L BERT Should I use?

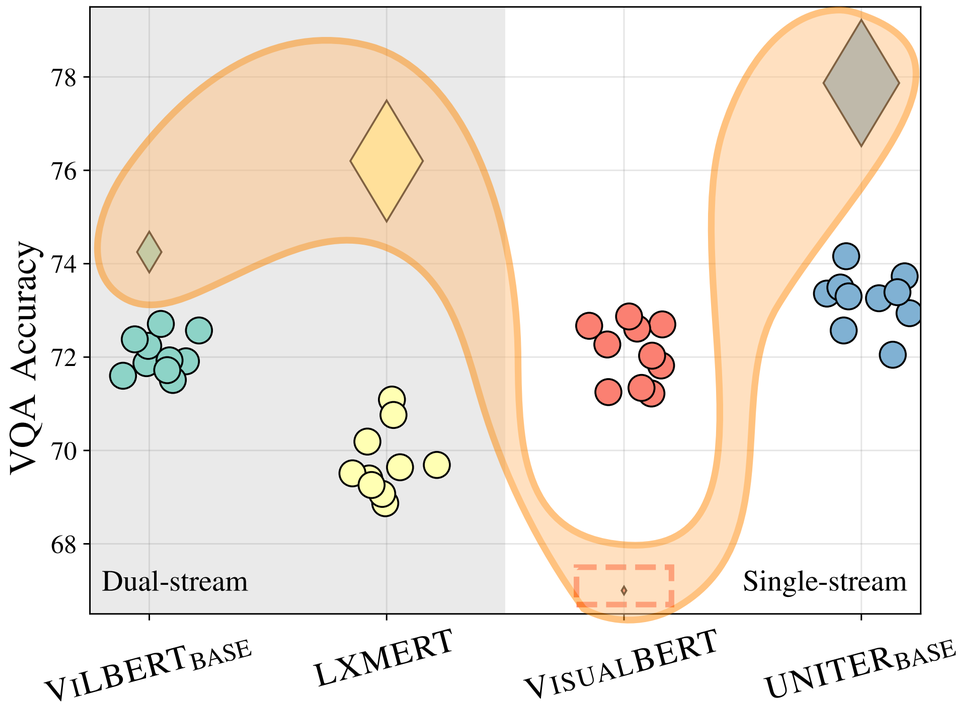


Size of **original** pretraining dataset

Size of **controlled** pretraining dataset

Meta-analysis: Bugliarello et al. TACL 2021. Multimodal Pretraining Unmasked: A Meta-Analysis and a Unified Framework of Vision-and-Language BERTs.

FAQ: Which V&L BERT Should I use?

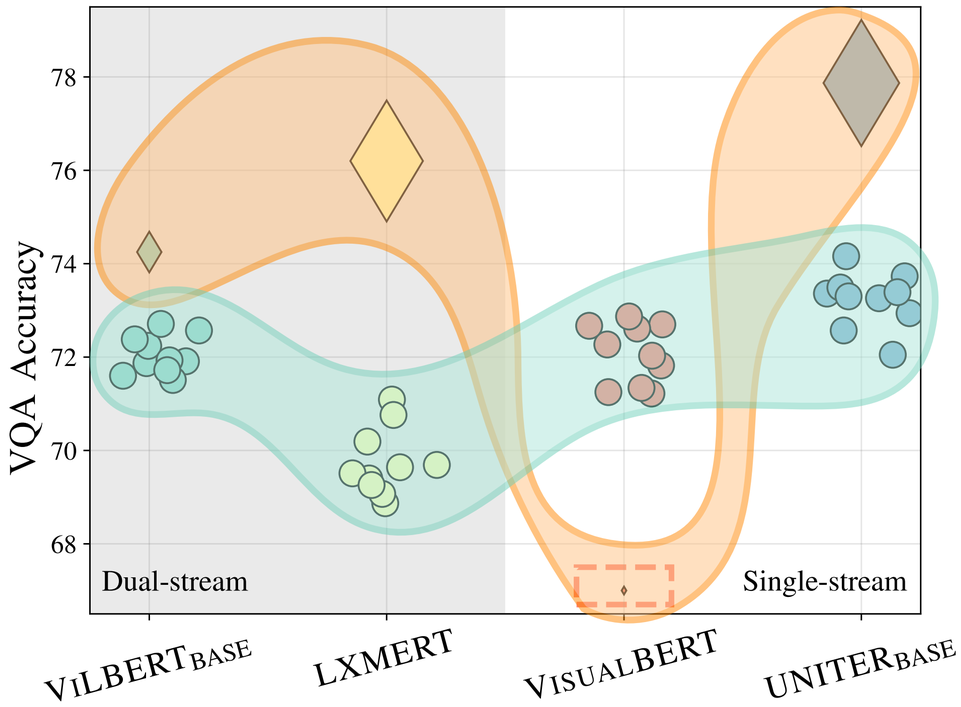


Size of **original** pretraining dataset

Size of **controlled** pretraining dataset

Meta-analysis: Bugliarello et al. TACL 2021. Multimodal Pretraining Unmasked: A Meta-Analysis and a Unified Framework of Vision-and-Language BERTs.

FAQ: Which V&L BERT Should I use?

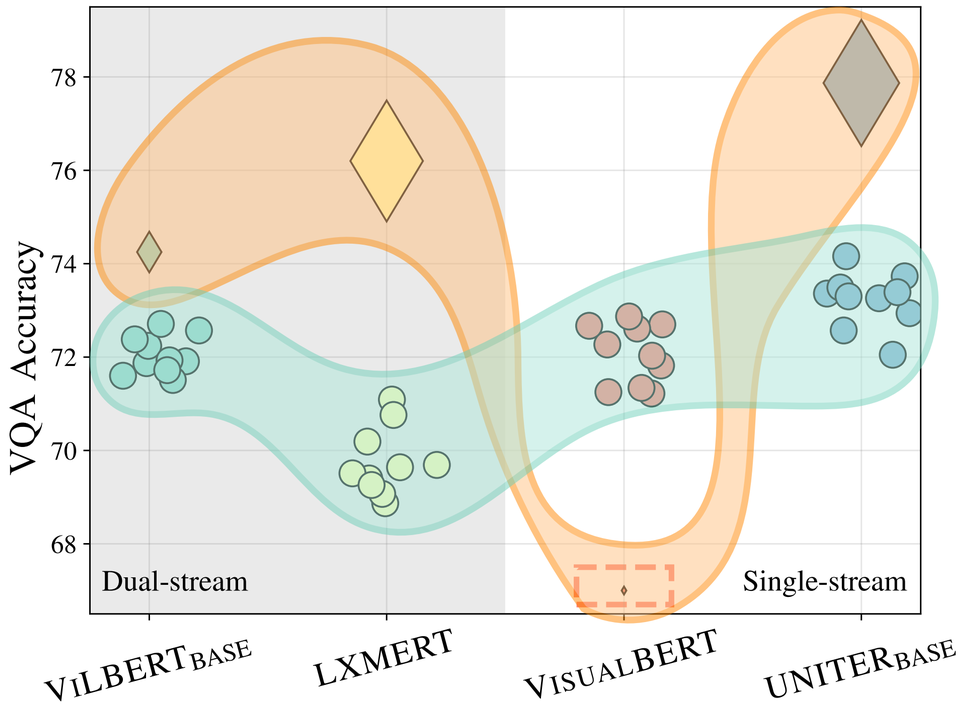


Size of **original** pretraining dataset

Size of **controlled** pretraining dataset

Meta-analysis: Bugliarello et al. TACL 2021. Multimodal Pretraining Unmasked: A Meta-Analysis and a Unified Framework of Vision-and-Language BERTs.

FAQ: Which V&L BERT Should I use?



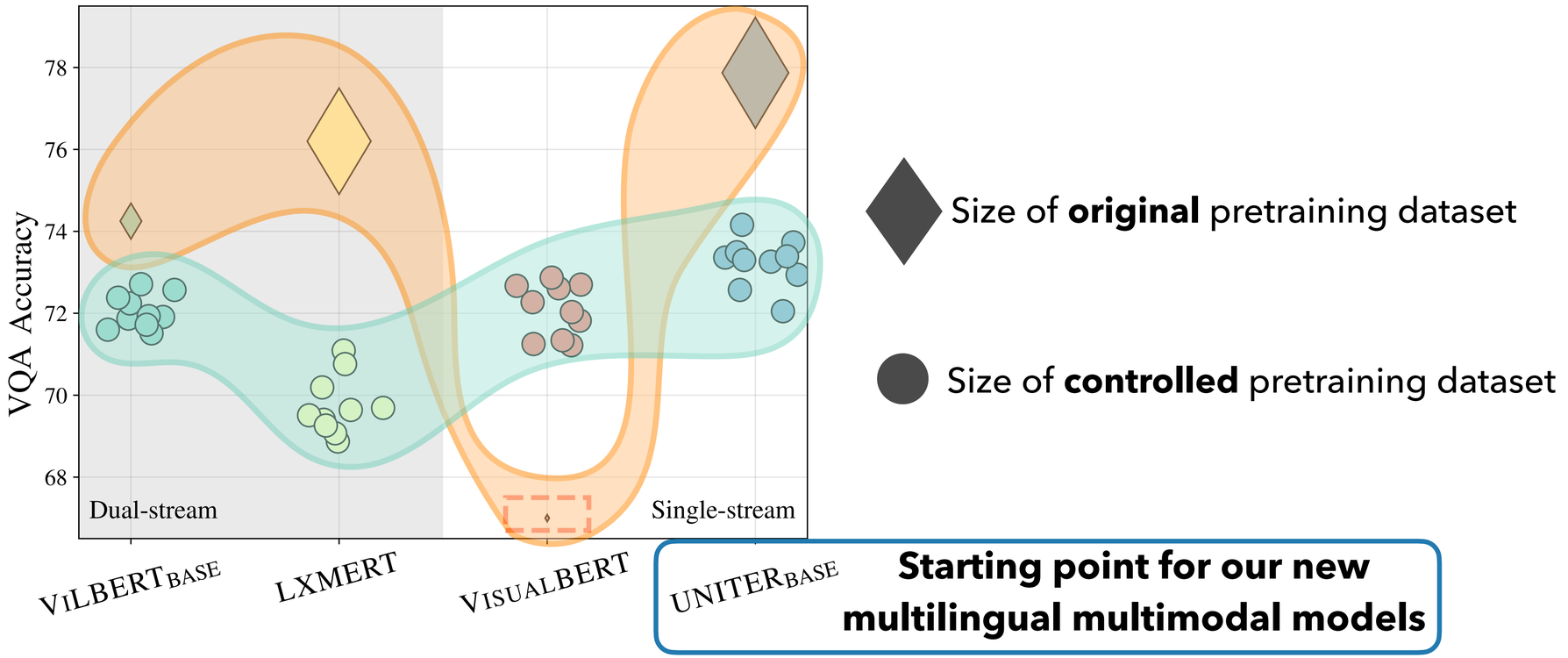
Size of **original** pretraining dataset

Size of **controlled** pretraining dataset

These models perform similarly when trained in similar conditions

Meta-analysis: Bugliarello et al. TACL 2021. Multimodal Pretraining Unmasked: A Meta-Analysis and a Unified Framework of Vision-and-Language BERTs.

FAQ: Which V&L BERT Should I use?



These models perform similarly when trained in similar conditions

Meta-analysis: Bugliarello et al. TACL 2021. Multimodal Pretraining Unmasked: A Meta-Analysis and a Unified Framework of Vision-and-Language BERTs.

Experiments and Results

MaRVL Pretraining

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

MaRVL Pretraining

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

Data

- English Conceptual Captions (Sharma+, ACL'19):

  MLM +
  Masked Region Modelling +
  Image–Text Matching

  } All Models

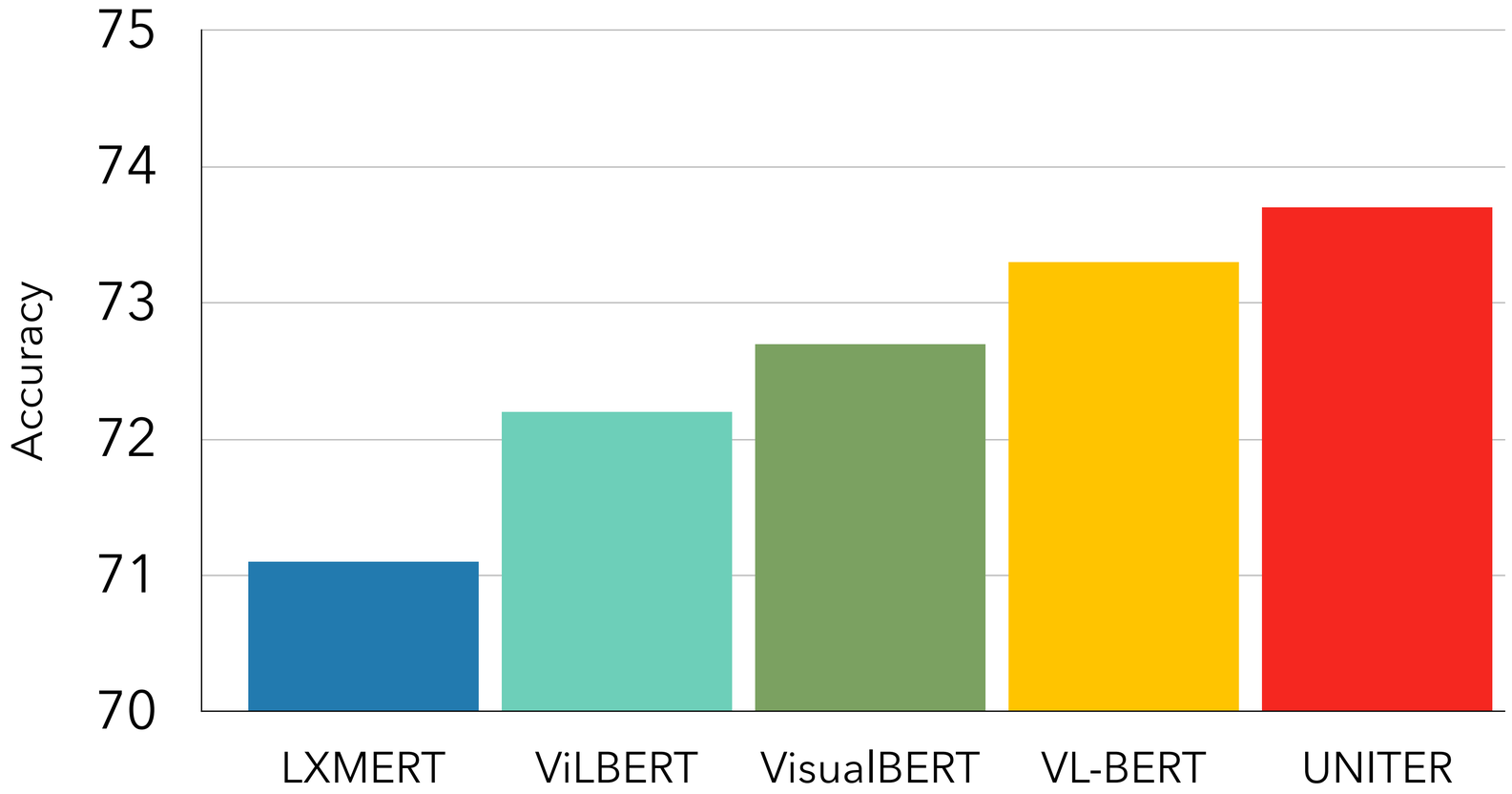
- Language-only MLM: 104 Wikipedia datasets — m/xUNITER

Fine-tuning

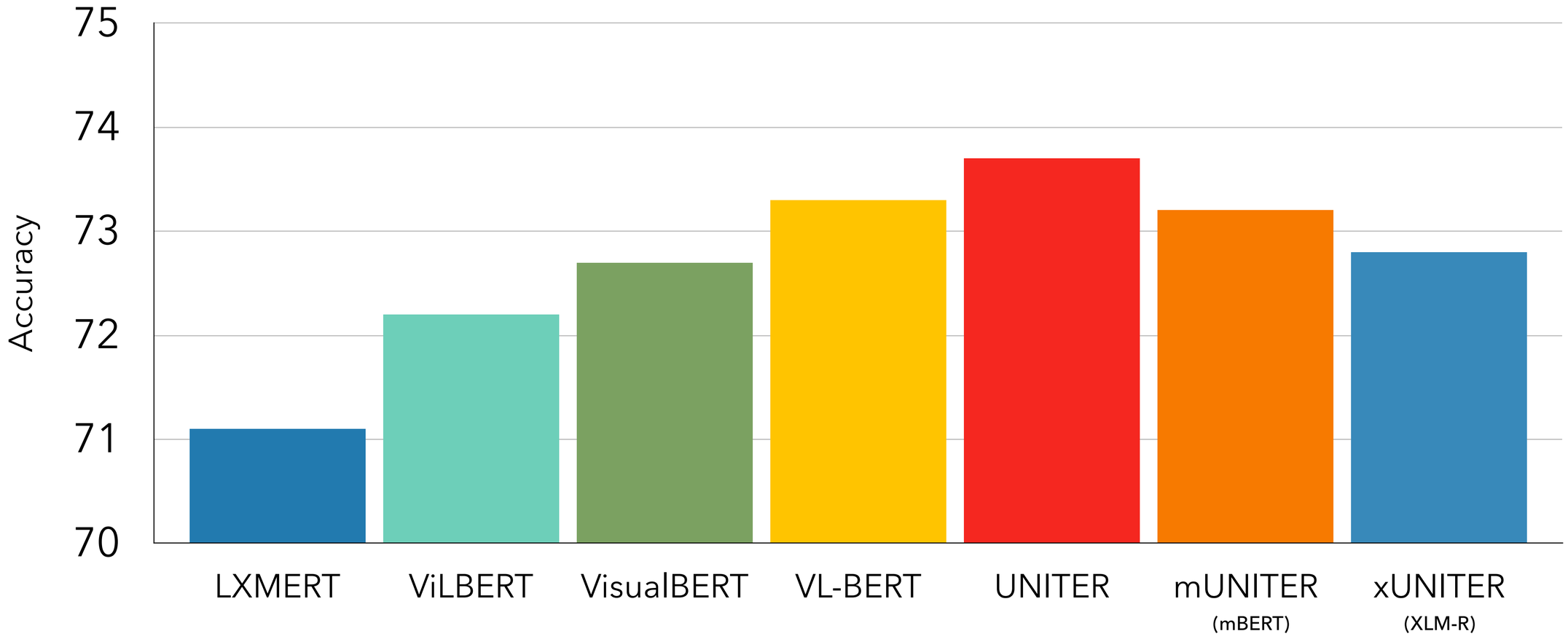
- Train on 86,373 data points in **English** NLVR2 (Suhr+, 2019)

- Test on 5,560 datapoints in MaRVL (5,560 datapoints)
  - **Zero-shot**: Multilingual inputs directly in a cross-lingual approach
  - **Translate-test**: English models by machine translating language data

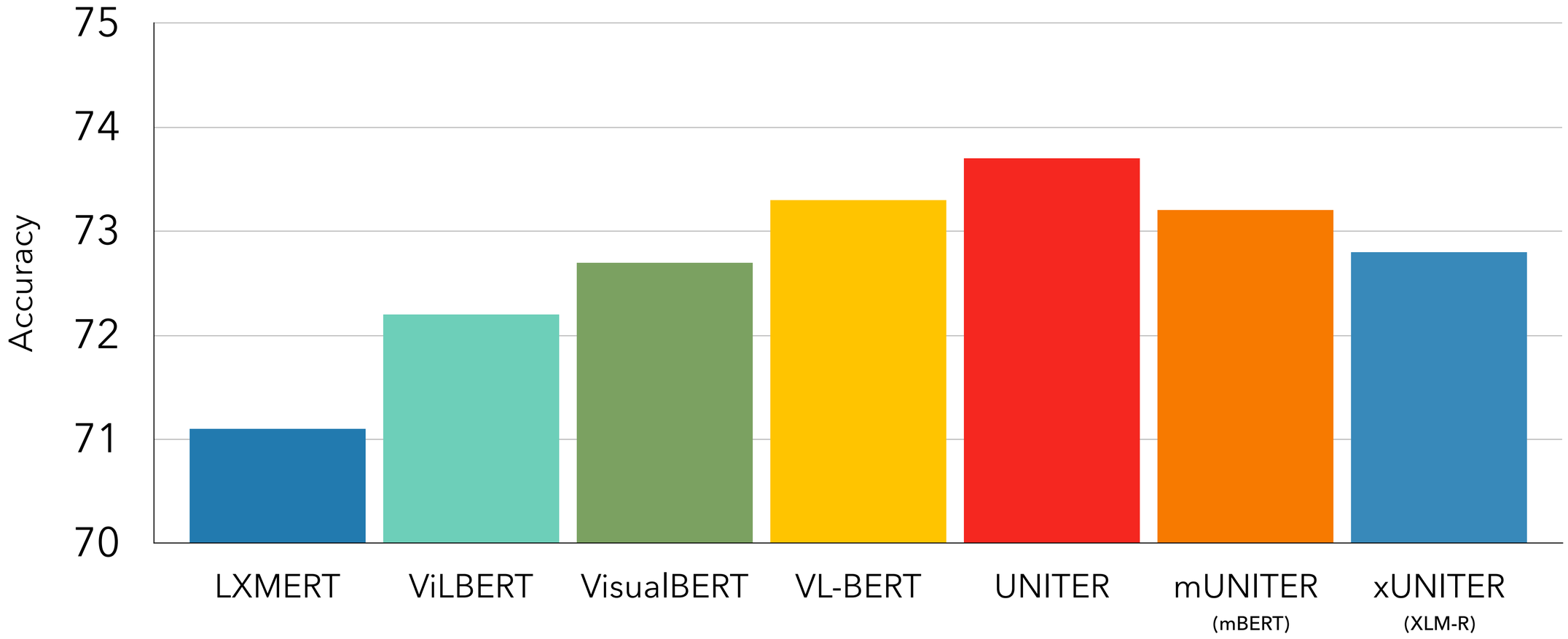
English NLVR2 Results (Sanity check)



English NLVR2 Results (Sanity check)

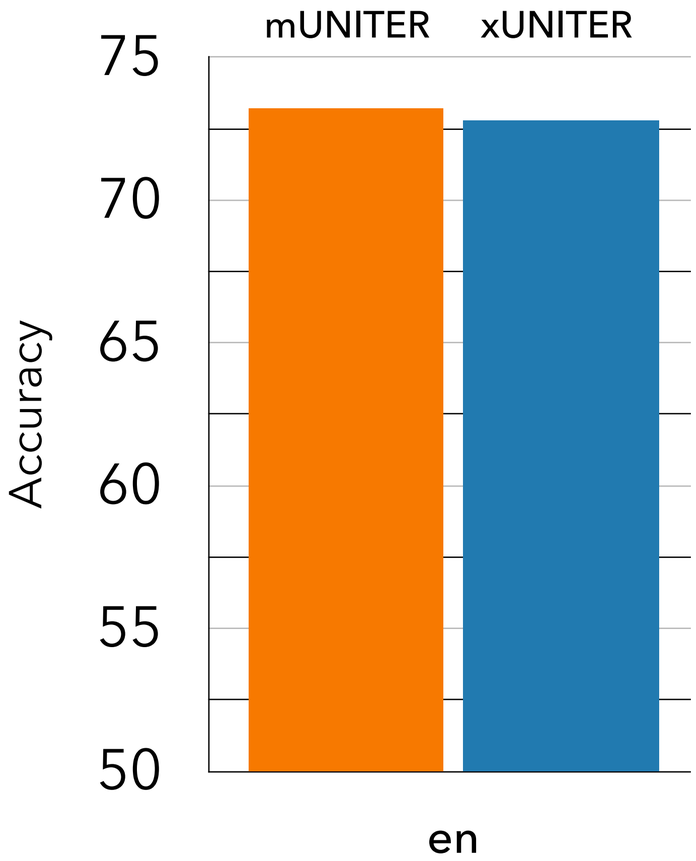


English NLVR2 Results (Sanity check)

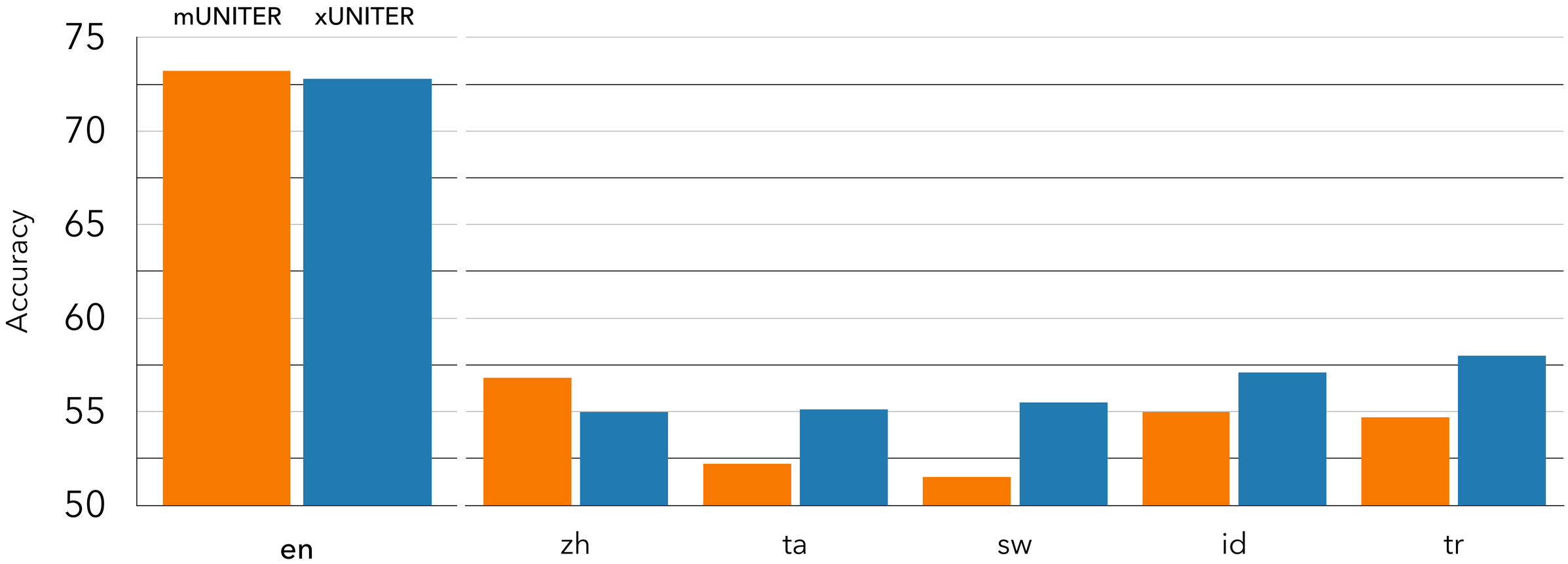


m/xUNITER perform similarly to English-only models

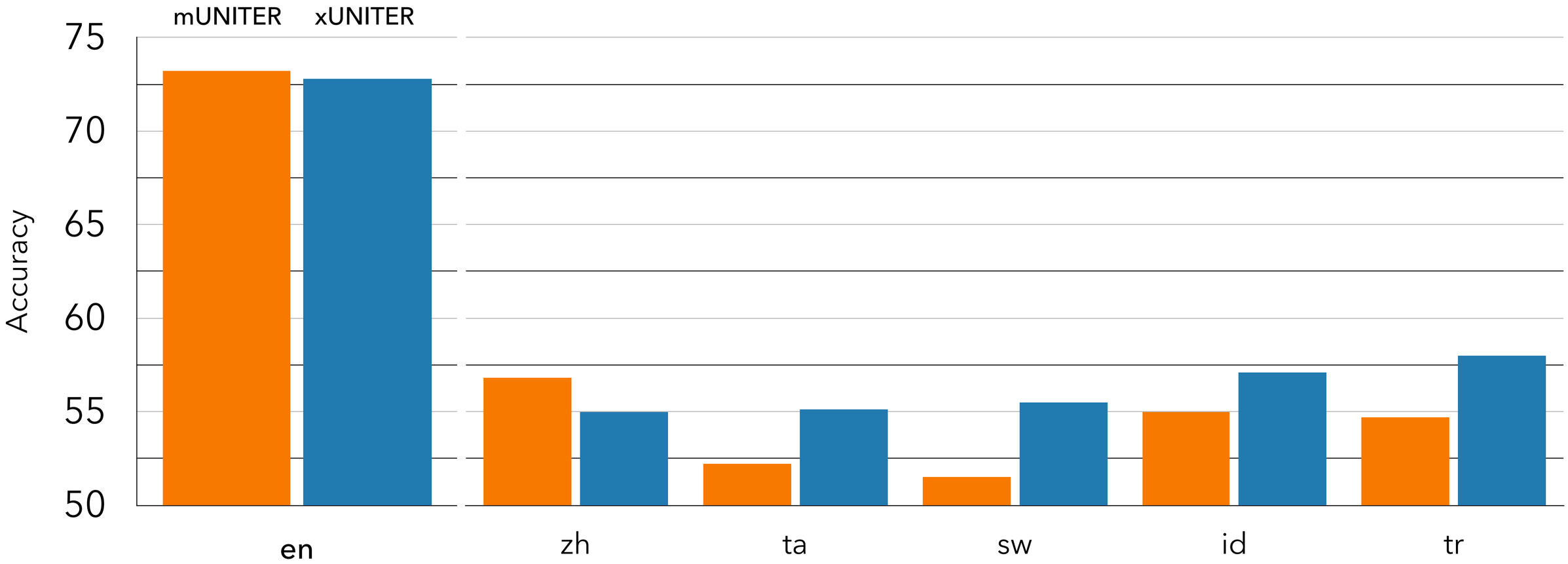
MaRVL Zero-shot Results



MaRVL Zero-shot Results

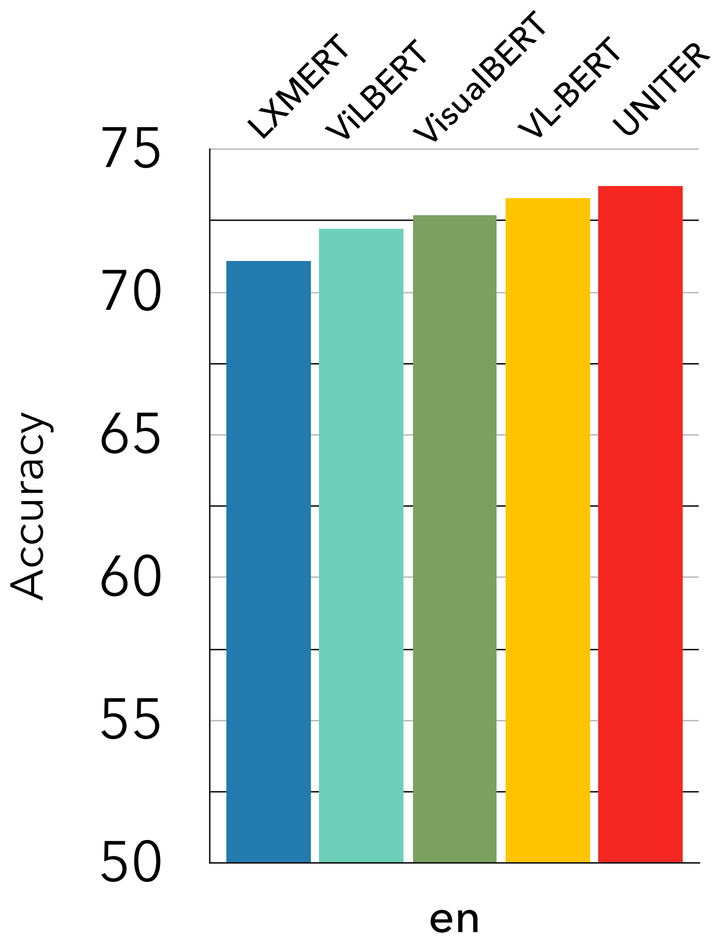


MaRVL Zero-shot Results

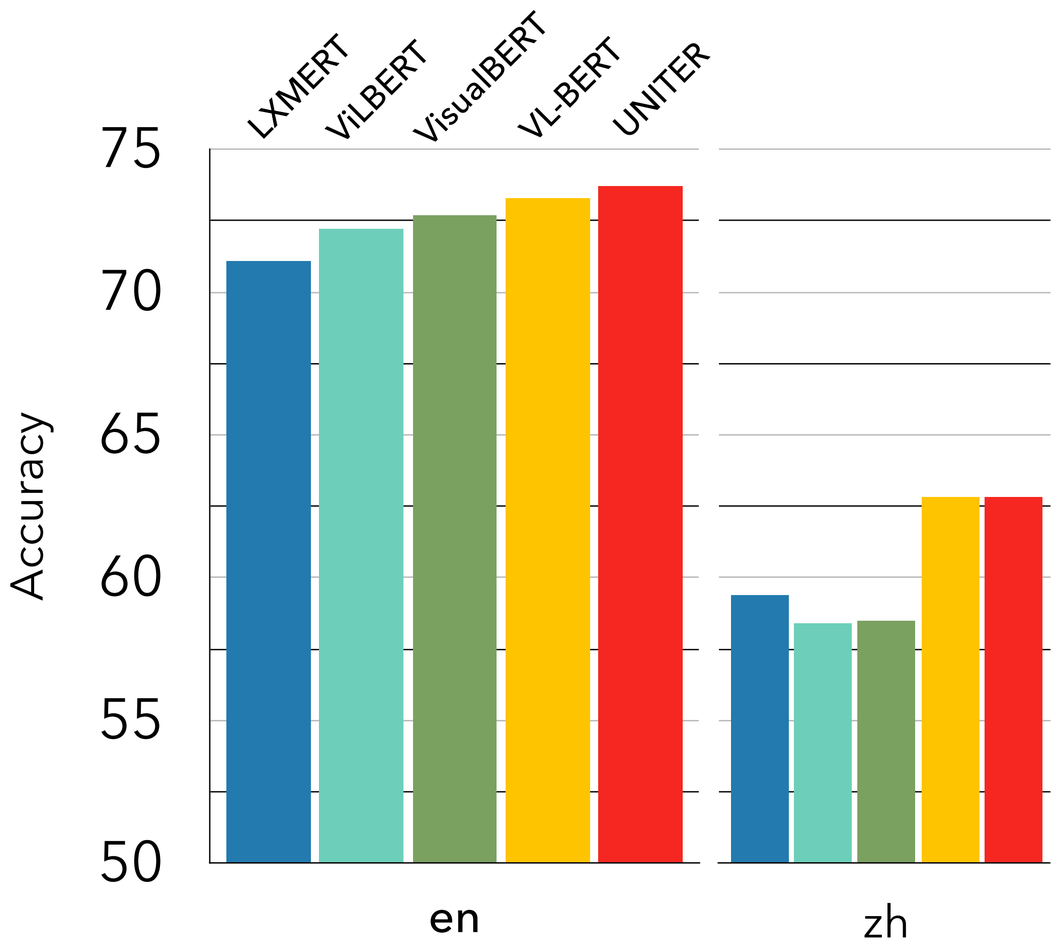


Zero-shot transfer: substantial drop in performance

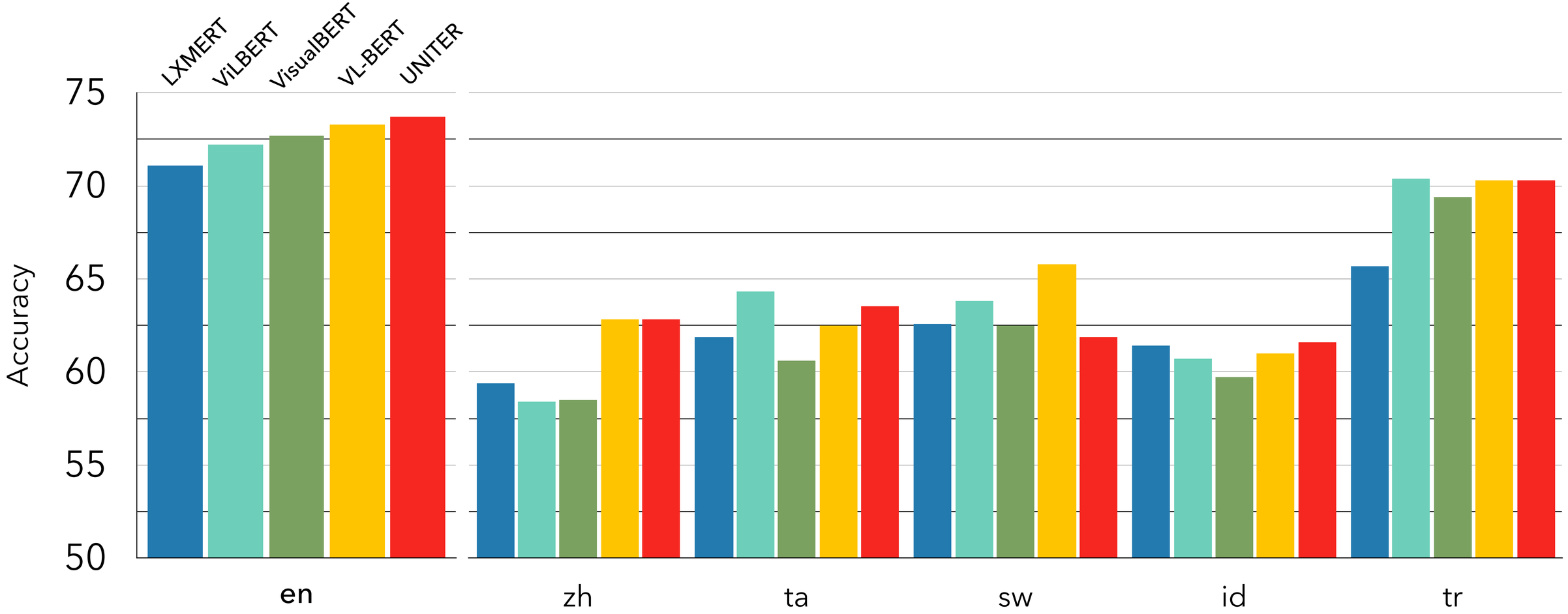
MaRVL Translate-Test



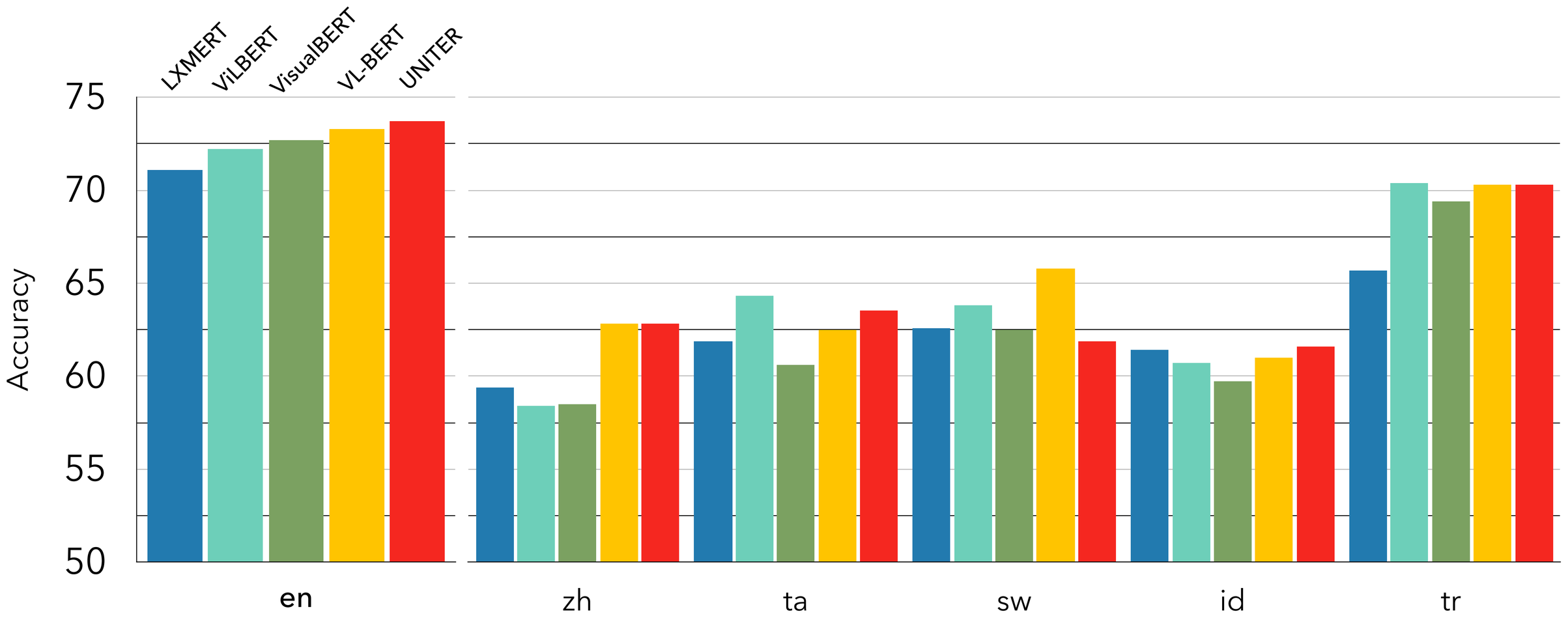
MaRVL Translate-Test



MaRVL Translate-Test



MaRVL Translate-Test



Translate-test: much better than zero-shot transfer

Why is MaRVL so difficult?

There are two distribution shifts in MaRVL:

XLT: Cross-lingual transfer

OOD: Out-of-distribution concepts

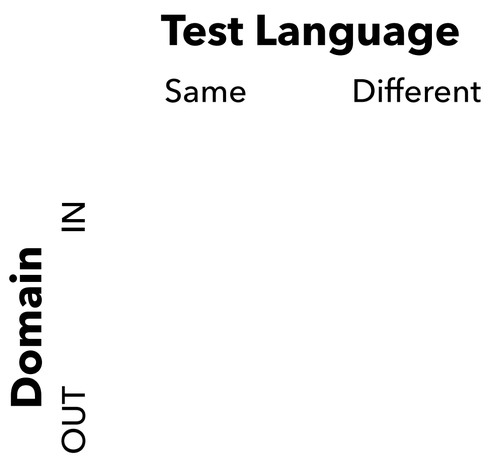
Both analyses were completed by a native Mandarin Chinese speaker (Fangyu)

Why is MaRVL so difficult?

There are two distribution shifts in MaRVL:

XLT: Cross-lingual transfer

OOD: Out-of-distribution concepts



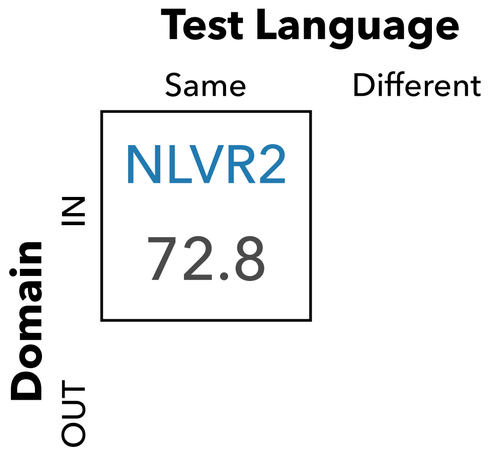
Both analyses were completed by a native Mandarin Chinese speaker (Fangyu)

Why is MaRVL so difficult?

There are two distribution shifts in MaRVL:

XLT: Cross-lingual transfer

OOD: Out-of-distribution concepts



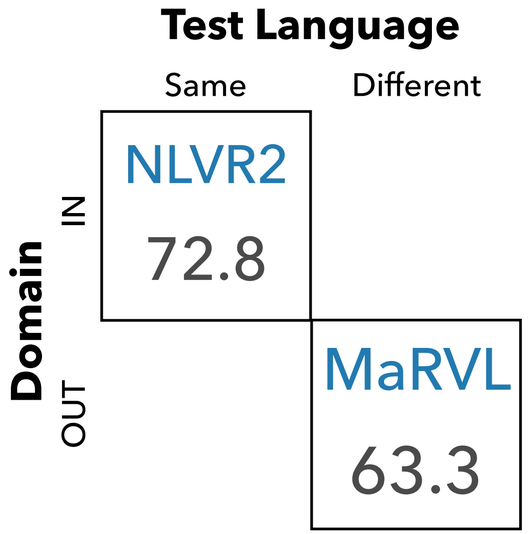
Both analyses were completed by a native Mandarin Chinese speaker (Fangyu)

Why is MaRVL so difficult?

There are two distribution shifts in MaRVL:

XLT: Cross-lingual transfer

OOD: Out-of-distribution concepts

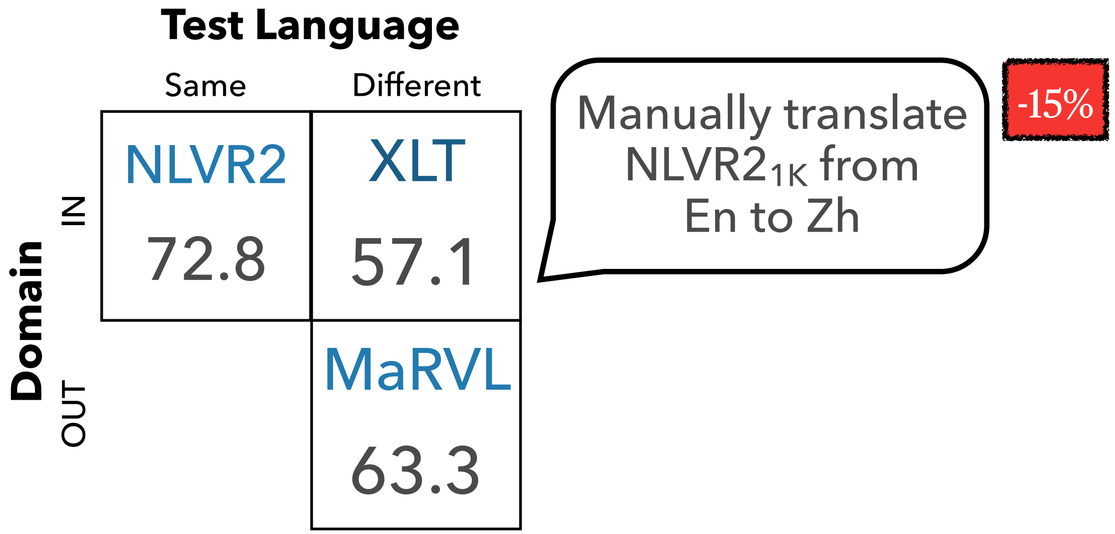


Both analyses were completed by a native Mandarin Chinese speaker (Fangyu)

Why is MaRVL so difficult?

There are two distribution shifts in MaRVL:

- XLT: Cross-lingual transfer
- OOD: Out-of-distribution concepts



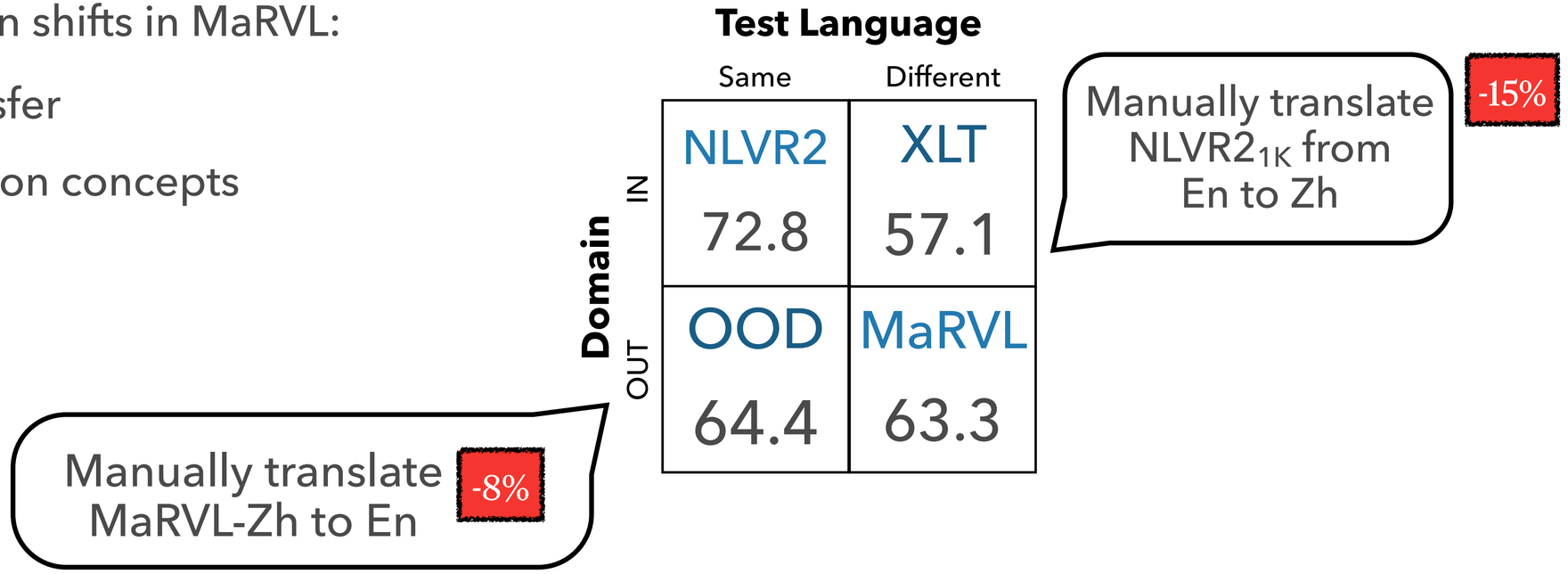
Both analyses were completed by a native Mandarin Chinese speaker (Fangyu)

Why is MaRVL so difficult?

There are two distribution shifts in MaRVL:

XLT: Cross-lingual transfer

OOD: Out-of-distribution concepts



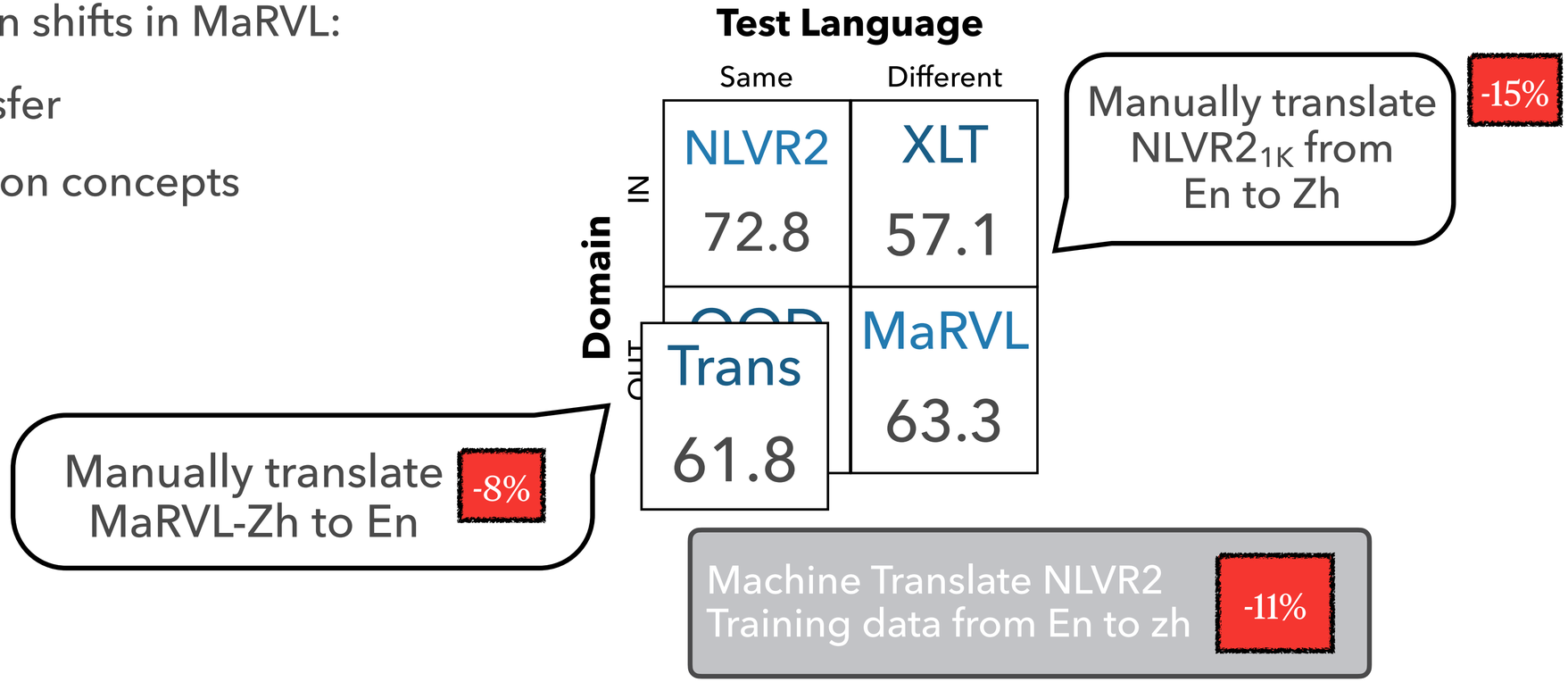
Both analyses were completed by a native Mandarin Chinese speaker (Fangyu)

Why is MaRVL so difficult?

There are two distribution shifts in MaRVL:

XLT: Cross-lingual transfer

OOD: Out-of-distribution concepts



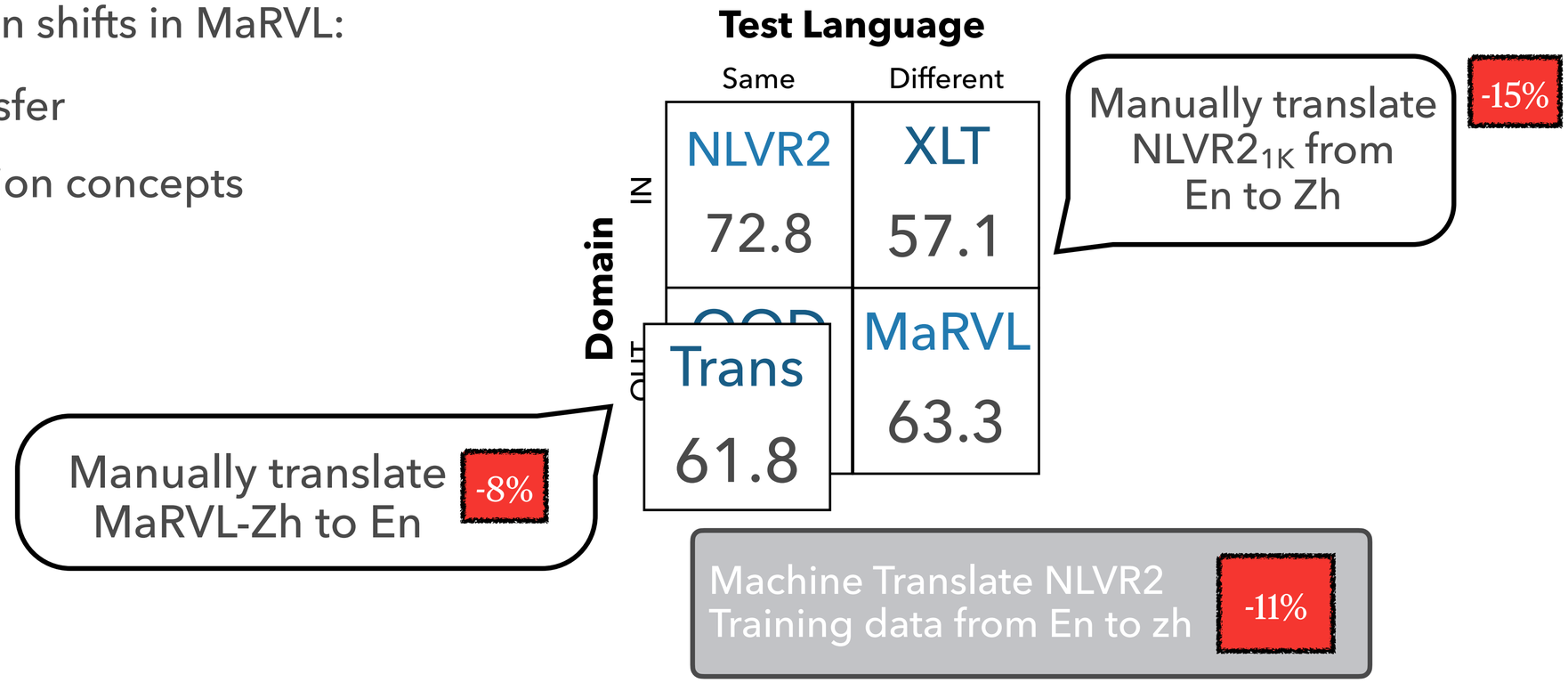
Both analyses were completed by a native Mandarin Chinese speaker (Fangyu)

Why is MaRVL so difficult?

There are two distribution shifts in MaRVL:

XLT: Cross-lingual transfer

OOD: Out-of-distribution concepts



Both of these distributional shifts are challenging

Both analyses were completed by a native Mandarin Chinese speaker (Fangyu)

Final Words

Connections to related work

- Machine Translation
  - Translation direction (Kurokawa+ 2009), “translationese” (Gellerstam, 1986)
  - Multimodal image text (Specia et al. 2016; Sulubacak et al. 2020)

- Multimodal Learning
  - GD-VCR: Geographically-diverse common sense reasoning(Yin+ 2021)
  - xGQA: Visually-grounded question answering (Pfeiffer+ 2021)
  - MURAL: Multilingual image-text retrieval (Jain+ 2021)

- Computer Vision
  - Geographically diverse replacements for ImageNet (Asano+ 2021)

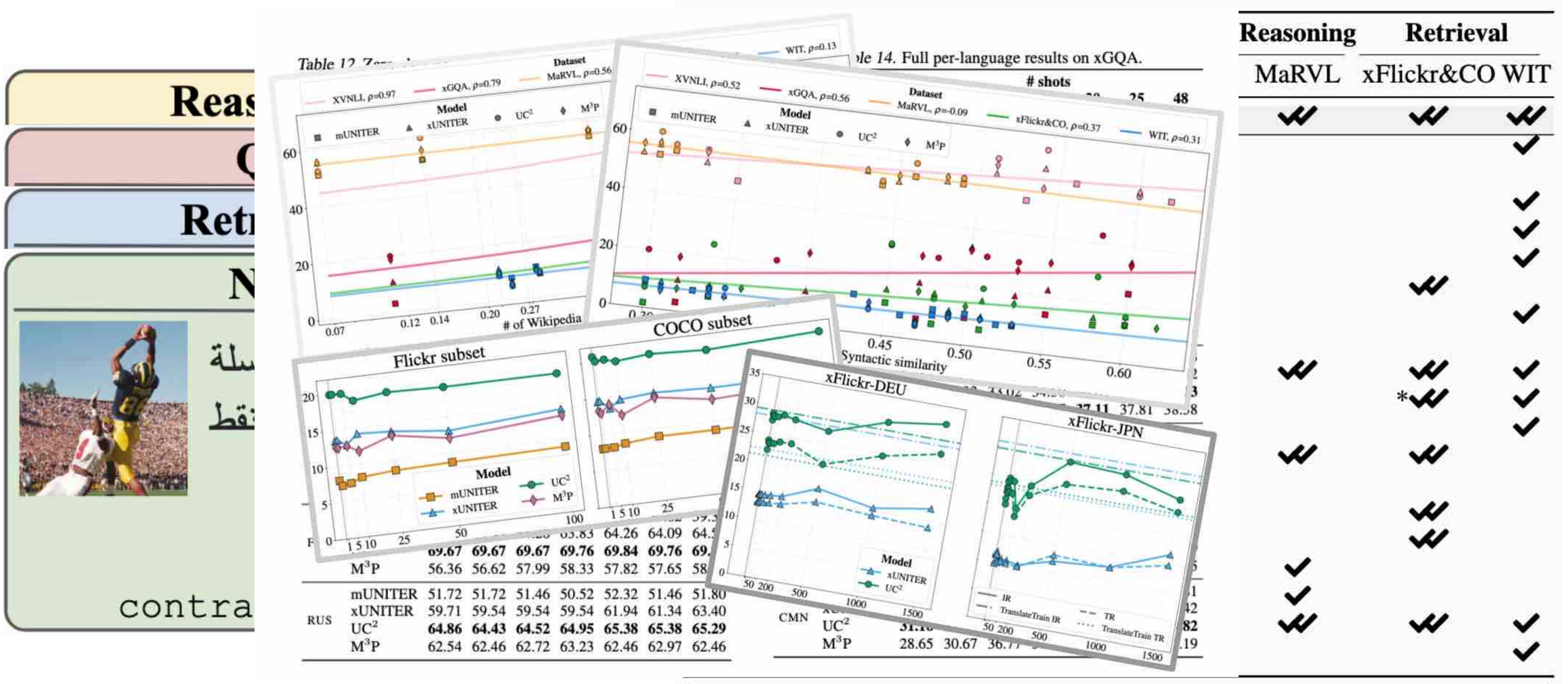
Future Work

- Data
  - Refine the data collection process
  - Collect MaRVL data on a larger scale and for more languages
- Models
  - Pretrained on data beyond ImageNet or Visual Genome
  - Deal with unseen concepts
- Tasks
  - More variety in multilingual multimodal tasks

IGLUE: More Multilingual Vision & Language Tasks

	Language			NLI	QA	Reasoning	Retrieval	
Name	Code	Family	Script	XVNLI	xGQA	MaRVL	xFlickr&CO	WIT
English	ENG	Indo-E	Latin	✓✓	✓✓	✓✓	✓✓	✓✓
Arabic	ARB	Afro-A	Arabic	✓✓				✓
Bengali	BEN	Indo-E	Bengali		✓✓			
Bulgarian	BUL	Indo-E	Cyrillic					✓
Danish	DAN	Indo-E	Latin					✓
Estonian	EST	Uralic	Latin					✓
German	DEU	Indo-E	Latin		✓✓		✓✓	
Greek	ELL	Indo-E	Greek					✓
French	FRA	Indo-E	Latin	✓✓				
Indonesian	IND	Austron	Latin		✓✓	✓✓	✓✓	✓
Japanese	JPN	Japonic	Kanji				*✓✓	✓
Korean	KOR	Koreanic	Hangul		✓✓			✓
Mandarin	CMN	Sino-T	Hanzi		✓✓	✓✓	✓✓	
Portuguese	POR	Indo-E	Latin		✓✓			
Russian	RUS	Indo-E	Cyrillic	✓✓	✓✓		✓✓	
Spanish	SPA	Indo-E	Latin	✓✓			✓✓	
Swahili	SWA	Niger-C	Latin			✓		
Tamil	TAM	Dravidian	Tamil			✓		
Turkish	TUR	Turkic	Latin			✓✓	✓✓	✓
Vietnamese	VIE	Austro-A	Latin					✓

IGLUE: More Multilingual Vision & Language Tasks



Conclusions

- Concepts and images in existing V&L datasets have an NA/EU bias
- Devise a new protocol for data creation driven by native speakers
- **MaRVL** : V&L reasoning dataset in 5 typologically diverse languages
- Develop two multilingual V&L BERTs: performance is at chance level

- Implications beyond vision and language research
  - Multilingual datasets should not just be translations of English data

Everything: https://marvl-challenge.github.io/

Examples of MaRVL Images

MaRVL-zh 京剧 (Beijing opera)

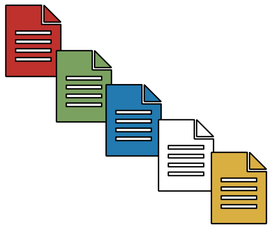
MaRVL-ta நாதசுவரம் (Nadaswaram)

MaRVL-id Cangkul (Hoe)

MaRVL-sw Chapati (Bread)

MaRVL-tr Lahmacun

Data Collection Guidelines



1. Concept Selection

The task requires you to select culture-specific concepts for 18 broad categories. You have to be a native Turkish speaker to participate in the study.

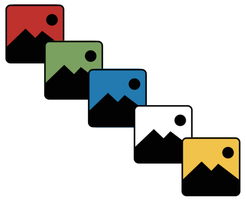
We ask you to work with the following categories.

Category list:

- Bird
- Mammal (e.g., Dog)
- Flower
- Vegetable
- Sports (e.g., Basketl...
- Celebration (e.g., Christmas)
- Utensil/tool (e.g., Bowl)
- Clothes
- Music instrument
- Food
- Beverage
- Fruits (e.g. apple, pear)
- House (interior) (e.g., types of rooms or areas in a house: kitchen, bedroom, ...)
- House (exterior) (e.g., types of buildings: villa, mansion, cottage, ...)
- Agriculture (e.g., pitchfork, harvester)
- Education (e.g., objects seen in schools: pencil, book, blackboard, ...)
- Visual arts (e.g., paintings, statues)
- Religions (e.g. Christian, Buddhism)

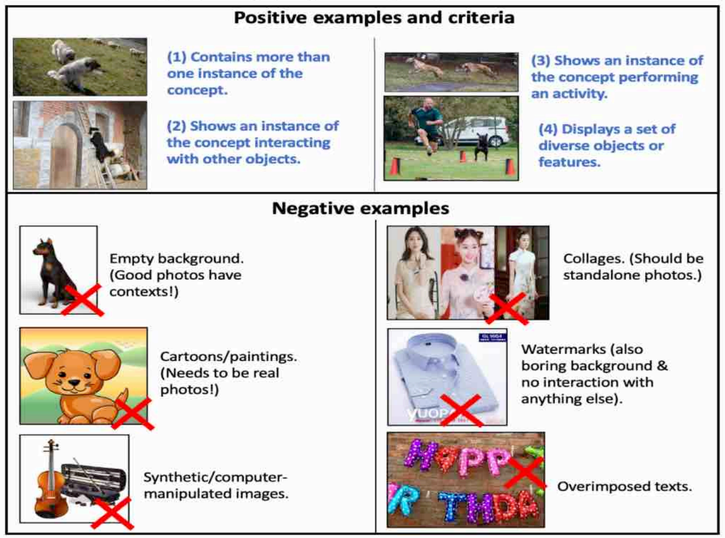
For each category, **you need to**

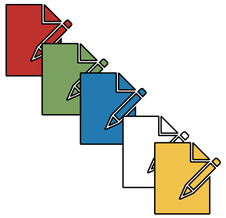
1. **come up with 5-10 specific concepts that are representative or common in the culture of Inonesian speaking population;**



2. Image Selection

The selected photos should satisfy at least **one of the four criteria** listed in the table below (the more the better).
To diversify the photos, **feel free to modify the query** you used (e.g., for "dog", you could search for "a group of dogs", "dogs running", etc.)





3. Annotation

In this task, you are required to write captions, and a caption is used to describe an image pair*. For each annotation instance, you will see 4 pairs of images, and you are supposed to write a caption **(in Turkish)** that is True for two pairs but False for each of the other two pairs.

* one image pair has two images in parallel: 

Below, we show you 5 concrete examples along with some tips. **Please spend sufficient time understanding the task.** And **it's important to follow the tips as you are paid based on the number of qualified captions!** You will receive 0.6 GBP for each valid caption.

Example 1:

./zh_images/2-crow/pair-1-b-2-8-2-9.jpg ./zh_images/2-crow/pair-1-a-2-6-2-5.jpg

https://github.com/marvl-challenge/marvl-guidelines