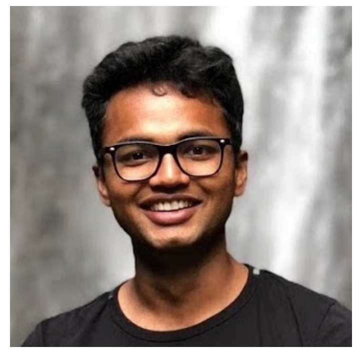
ICML 2022

IGLUE: A Benchmark for Transfer Learning across Modalities, Tasks, and Languages

Emanuele Bugliarello
emanuele@di.ku.dk
@ebugliarello

F. Liu

J. Pfeiffer


S. Reddy

D. Elliott

EM. Ponti

I. Vulić

UNIVERSITY OF CAMBRIDGE

TECHNISCHE UNIVERSITÄT DARMSTADT

McGill UNIVERSITY

Mila

UNIVERSITY OF COPENHAGEN

Vision-and-Language Data

Vision-and-Language Data

- MS COCO, Flickr30K, Visual Genome
- Conceptual Captions, Conceptual 12M, RedCaps
- VQA, GQA, Visual7W, VizWiz, RefCOCO, NLVR2, …

Vision-and-Language Data

- MS COCO, Flickr30K, Visual Genome
- Conceptual Captions, Conceptual 12M, RedCaps
- VQA, GQA, Visual7W, VizWiz, RefCOCO, NLVR2, …

Vision-and-Language Data

- MS COCO, Flickr30K, Visual Genome
- Conceptual Captions, Conceptual 12M, RedCaps
- VQA, GQA, Visual7W, VizWiz, RefCOCO, NLVR2, …

But this trend is reversing

- Flickr30K-{🇨🇳,🇯🇵} STAIR Captions (🇯🇵) COCO-{🇨🇳,🇪🇸,🇮🇹,🇮🇳,🇻🇳}
- Multi30K (🇩🇪🇫🇷🇨🇿) XTD (10 langs) GEM (20 langs) WIT (108 langs)
- MultiSubs (🇩🇪🇫🇷🇪🇸🇵🇹) MuCO-VQA (🇮🇳) xGQA (🇩🇪🇵🇹🇨🇳🇮🇩🇧🇩🇷🇺🇰🇷) MaRVL (🇨🇳🇮🇩🇮🇳🇹🇷🇰🇪)

Vision-and-Language Data

- MS COCO, Flickr30K, Visual Genome
- Conceptual Captions, Conceptual 12M, RedCaps
- VQA, GQA, Visual7W, VizWiz, RefCOCO, NLVR2, …

But this trend is reversing

- Flickr30K-{🇨🇳,🇯🇵} STAIR Captions (🇯🇵) COCO-{🇨🇳,🇪🇸,🇮🇹,🇮🇳,🇻🇳}
- Multi30K (🇩🇪🇫🇷🇨🇿) XTD (10 langs) GEM (20 langs) WIT (108 langs)
- MultiSubs (🇩🇪🇫🇷🇪🇸🇵🇹) MuCO-VQA (🇮🇳) xGQA (🇩🇪🇵🇹🇨🇳🇮🇩🇧🇩🇷🇺🇰🇷) MaRVL (🇨🇳🇮🇩🇮🇳🇹🇷🇰🇪)

Mostly Indo-European languages

Mostly translations from English

IGLUE: A Benchmark to the Rescue

Benchmarks have driven progress in machine learning

🖼️ ImageNet

🇺🇸 GLUE, SuperGLUE 🇮🇩 IndoNLU 🇰🇷 KLUE 🇷🇺 RussianSuperGLUE 🇷🇴 Liro

🌍 XGLUE, XTREME, XTREME-R

IGLUE: A Benchmark to the Rescue

Benchmarks have driven progress in machine learning

🖼️ ImageNet

🇺🇸 GLUE, SuperGLUE 🇮🇩 IndoNLU 🇰🇷 KLUE 🇷🇺 RussianSuperGLUE 🇷🇴 Liro

🌍 XGLUE, XTREME, XTREME-R

IGLUE: Image-Grounded Language Understanding Evaluation

🌍 20 languages: 11 families, 9 scripts, 3/5 WALS macro-areas

⚙️ 4 V&L tasks requiring different levels of syntactic-semantic understanding

🗂️ 5 datasets, both pre-existing and new ones

📈 Zero-shot & Few-shot learning setups

IGLUE: Tasks & Datasets

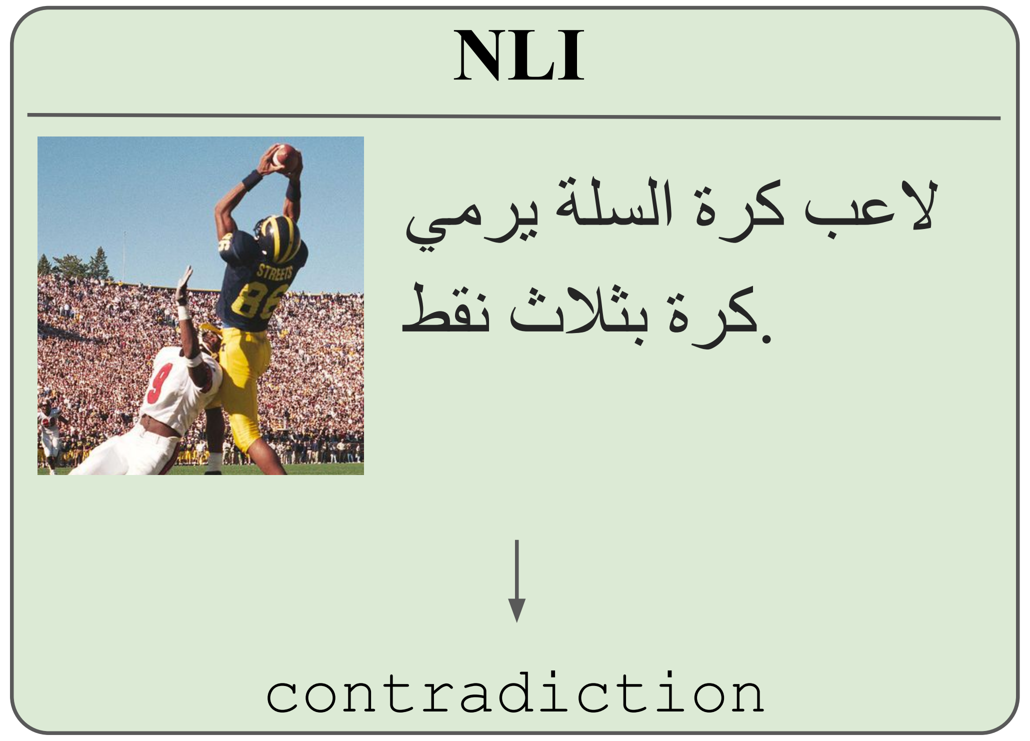
IGLUE: Tasks & Datasets

NATURAL LANGUAGE INFERENCE

Given an *image*-premise, predict if a *text*-hypothesis `entails`, `contradicts`, or is `neutral` to it

XVNLI *

🌍 5 Languages: Arabic, French, Russian and Spanish



ENG: The basketball player shoots a three pointer

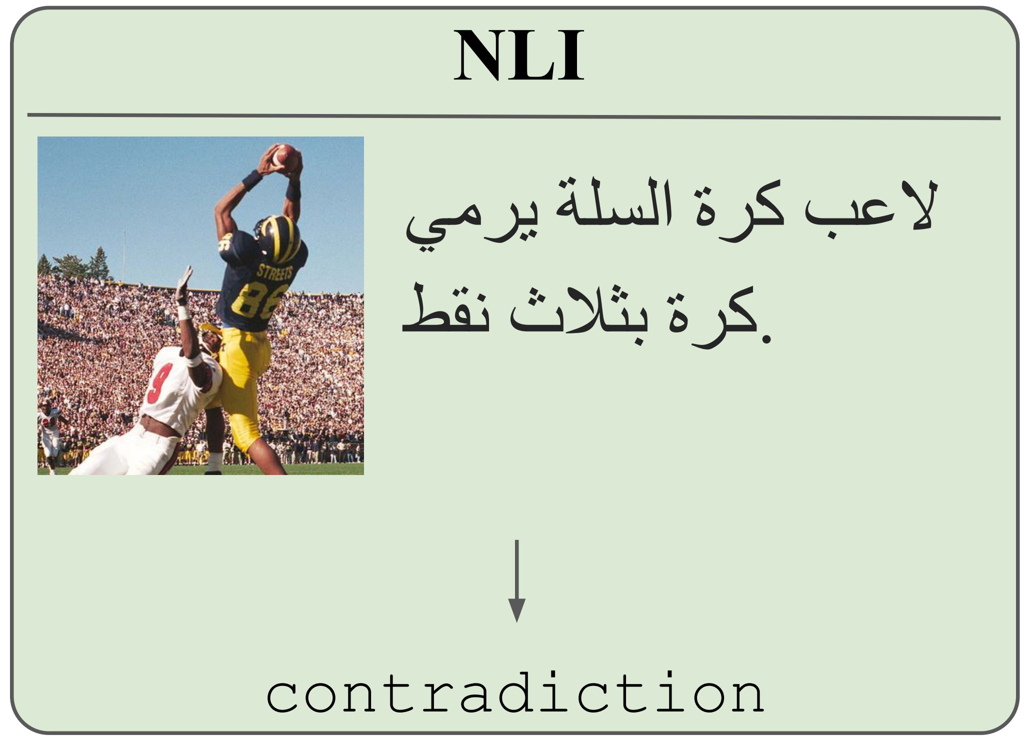
IGLUE: Tasks & Datasets

NATURAL LANGUAGE INFERENCE

Given an *image*-premise, predict if a *text*-hypothesis `entails`, `contradicts`, or is `neutral` to it

XVNLI *

🌍 5 Languages: Arabic, French, Russian and Spanish



NLI

لاعب كرة السلة يرمي
كرة بثلاث نقط.

↓

`contradiction`

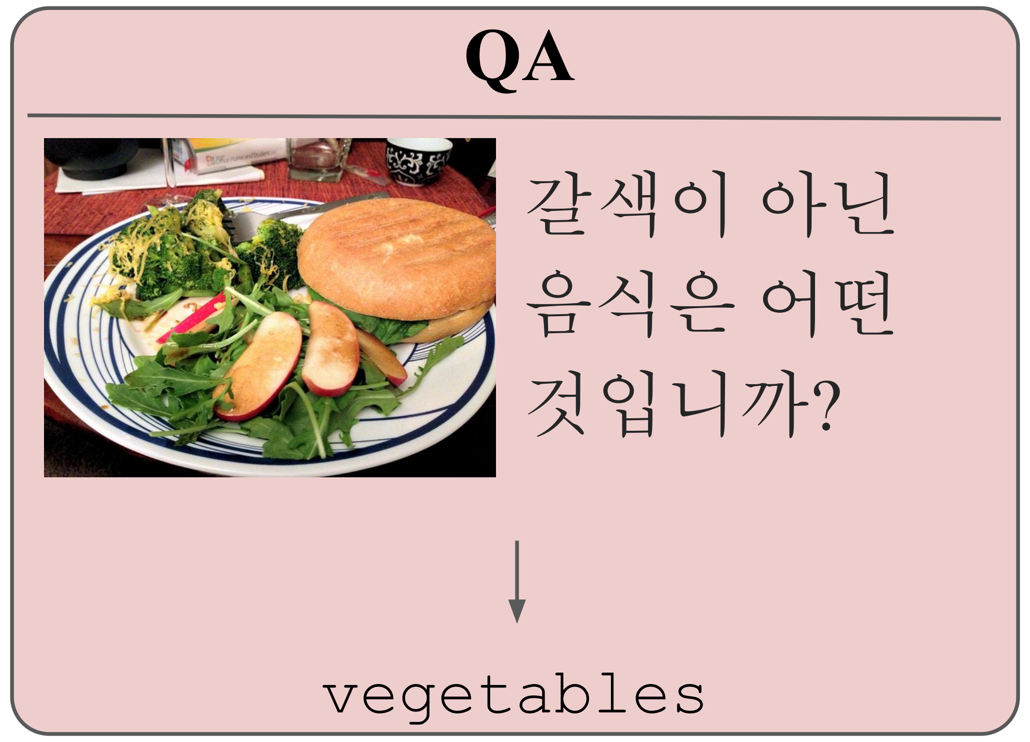
ENG: The basketball player shoots a three pointer

QUESTION ANSWERING

Given an image and question about it, predict the answer

xGQA (Pfeiffer+, 2022)

🌍 8 Languages: Bengali, German, Indonesian, Korean, Mandarin, Portuguese, Russian



QA

갈색이 아닌
음식은 어떤
것입니까?

↓

`vegetables`

ENG: Which kind of food is not brown?

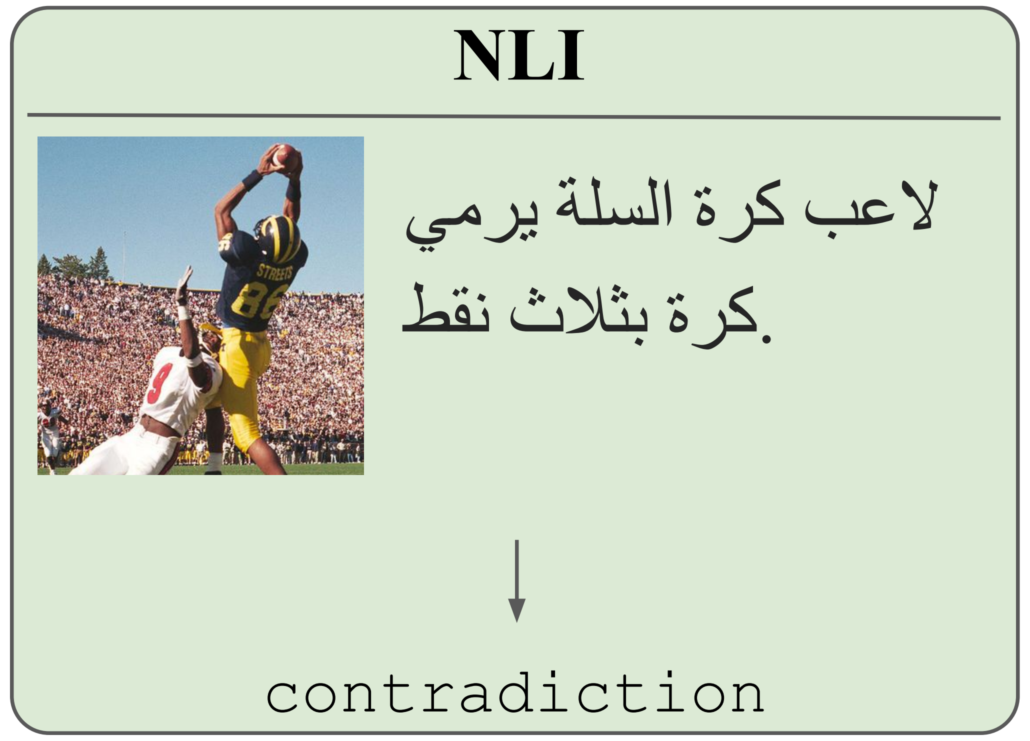
IGLUE: Tasks & Datasets

NATURAL LANGUAGE INFERENCE

Given an *image*-premise, predict if a *text*-hypothesis `entails`, `contradicts`, or is `neutral` to it

XVNLI *

🌍 5 Languages: Arabic, French, Russian and Spanish



NLI

لاعب كرة السلة يرمي كرة بثلاث نقط.

↓

`contradiction`

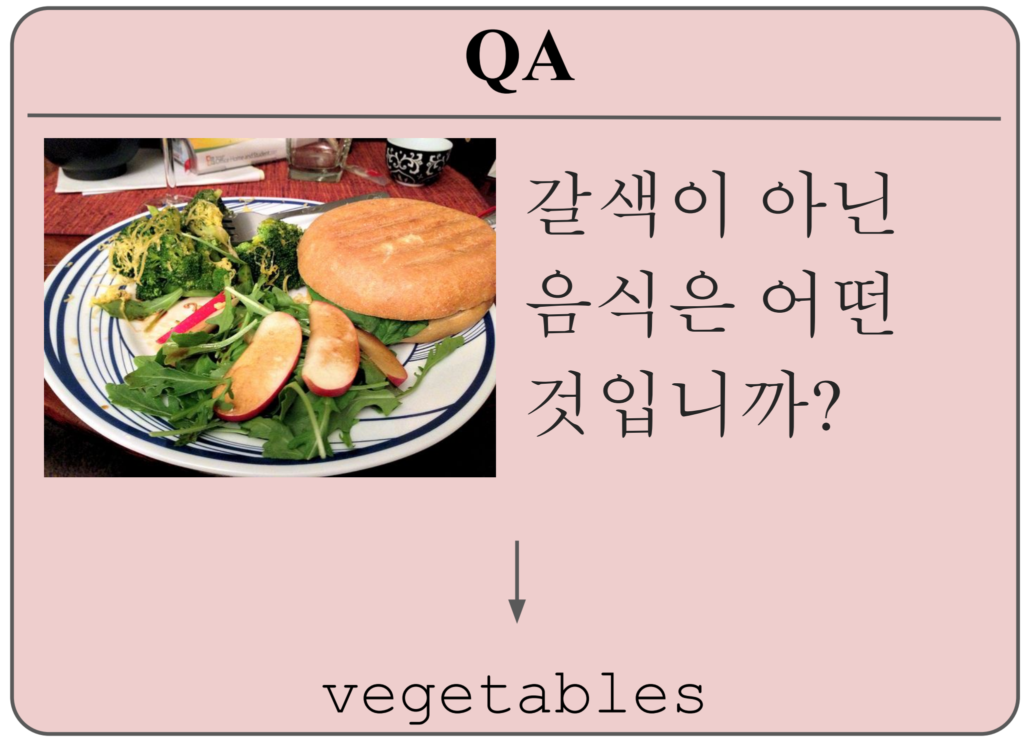
ENG: The basketball player shoots a three pointer

QUESTION ANSWERING

Given an image and question about it, predict the answer

xGQA (Pfeiffer+, 2022)

🌍 8 Languages: Bengali, German, Indonesian, Korean, Mandarin, Portuguese, Russian



QA

갈색이 아닌 음식은 어떤 것입니까?

↓

`vegetables`

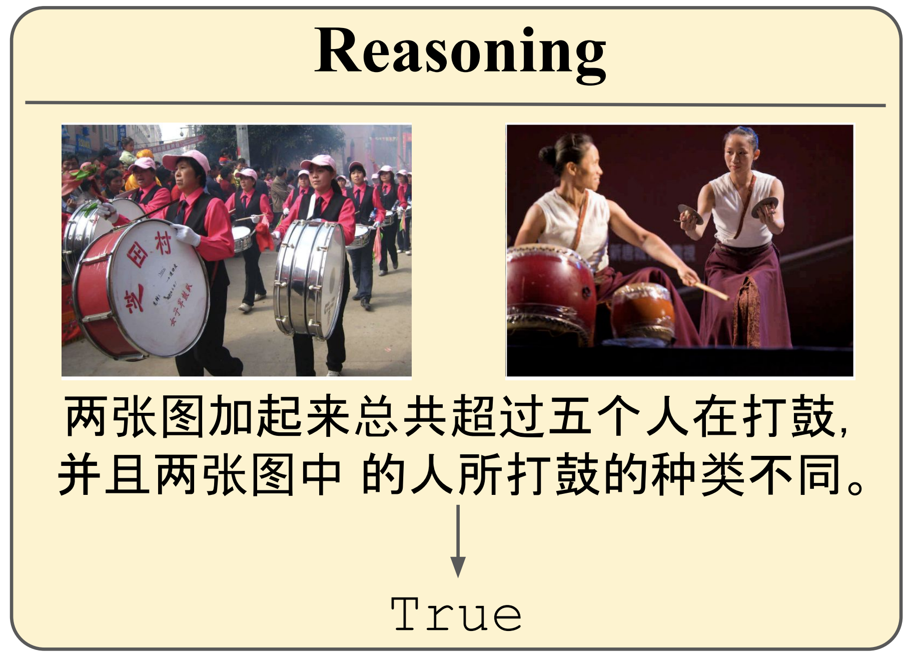
ENG: Which kind of food is not brown?

VISUAL REASONING

Given two images and a textual description, predict if the description applies to both images (`true`/`false`)

MaRVL (Liu&Bugliarello+, 2021)

🌍 6 Languages: Indonesian, Mandarin, Swahili, Tamil, Turkish



Reasoning

两张图加起来总共超过五个人在打鼓，并且两张图中 的人所打鼓的种类不同。

↓

`True`

ENG: In total, there are more than five people playing drums in the two images combined and people in the two images are playing different kinds of drums.

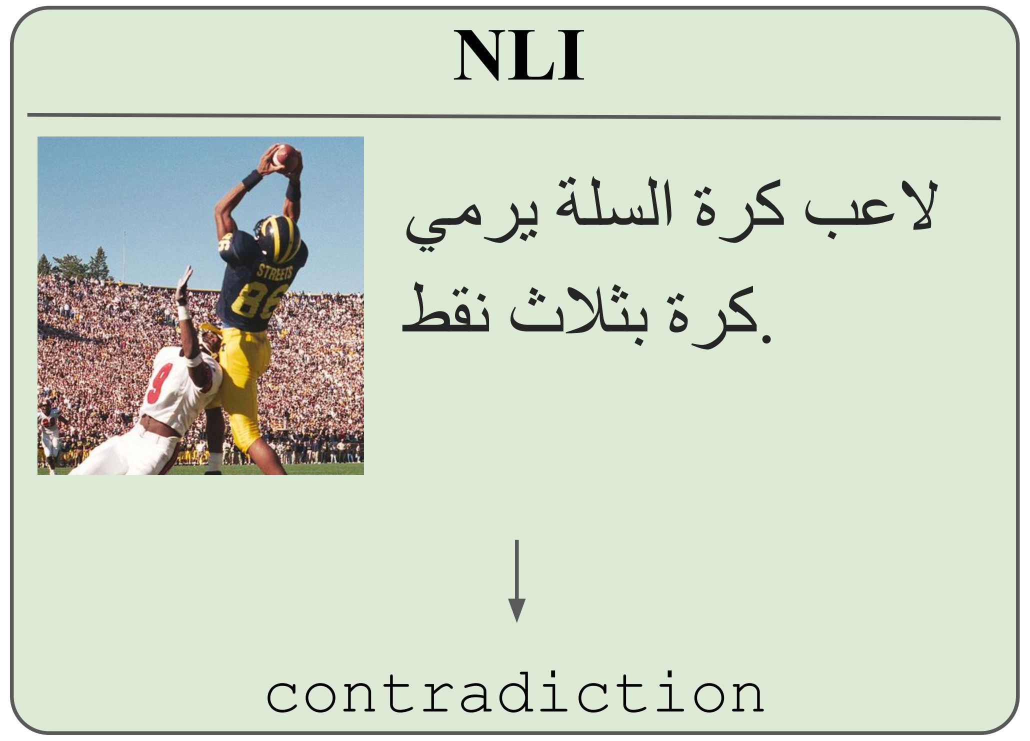
IGLUE: Tasks & Datasets

NATURAL LANGUAGE INFERENCE

Given an *image*-premise, predict if a *text*-hypothesis `entails`, `contradicts`, or is `neutral` to it

XVNLI *

🌍 5 Languages: Arabic, French, Russian and Spanish



NLI

لاعب كرة السلة يرمي كرة بثلاث نقط.

↓

`contradiction`

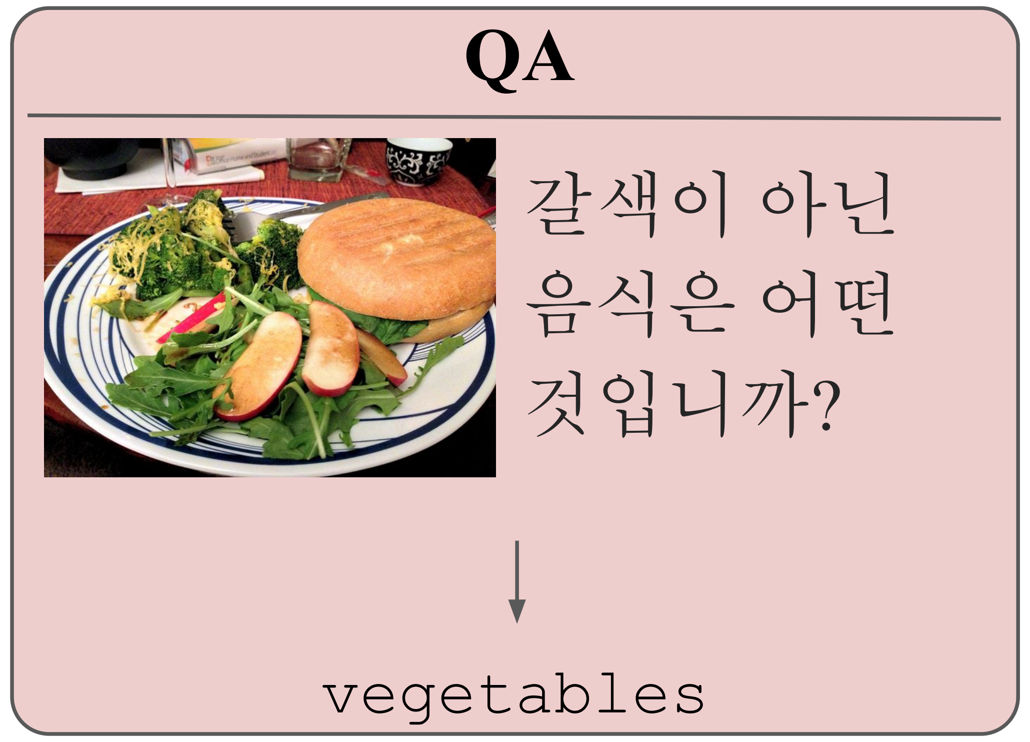
ENG: The basketball player shoots a three pointer

QUESTION ANSWERING

Given an image and question about it, predict the answer

xGQA (Pfeiffer+, 2022)

🌍 8 Languages: Bengali, German, Indonesian, Korean, Mandarin, Portuguese, Russian



QA

갈색이 아닌 음식은 어떤 것입니까?

↓

`vegetables`

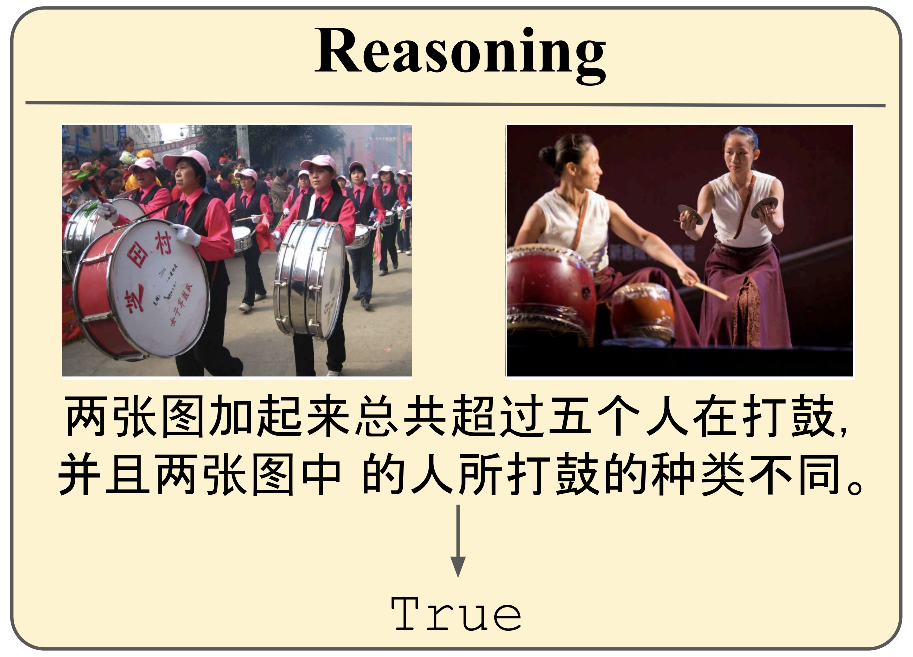
ENG: Which kind of food is not brown?

VISUAL REASONING

Given two images and a textual description, predict if the description applies to both images (`true`/`false`)

MaRVL (Liu&Bugliarello+, 2021)

🌍 6 Languages: Indonesian, Mandarin, Swahili, Tamil, Turkish



Reasoning

两张图加起来总共超过五个人在打鼓，并且两张图中 的人所打鼓的种类不同。

↓

`True`

ENG: In total, there are more than five people playing drums in the two images combined and people in the two images are playing different kinds of drums.

IMAGE—TEXT RETRIEVAL

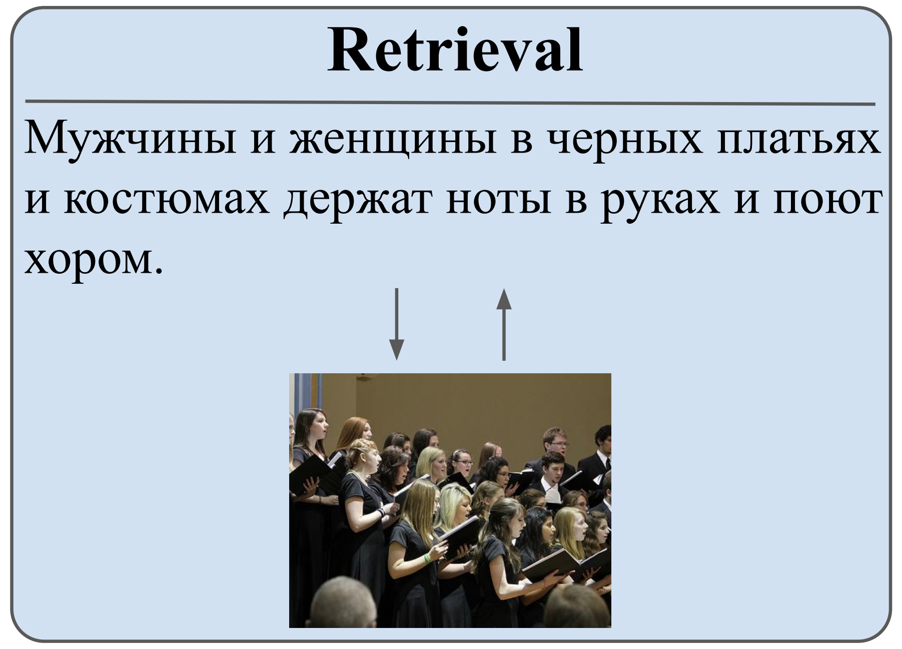
IR: Given a caption, retrieve its image
TR: Given an image, retrieve its caption

xFlickr&CO *

🌍 8 high-resource languages

WIT (Srinivasan+, 2021)

🌍 11 diverse languages



Retrieval

Мужчины и женщины в черных платьях и костюмах держат ноты в руках и поют хором.

↓ ↑

ENG: A group of men and women dressed in formal black dresses and suits holding their music books and singing.

Experimental Setup

Baselines

🧑‍💻 Implement multilingual V&L Transformers in a single code (VOLTA; Bugliarello+ 2021)

🤖 mUNITER & xUNITER (Liu&Bugliarello+, 2021) M^{3}P (Ni+, 2021) UC^2 (Zhou+, 2021)

Experimental Setup

Baselines

👩‍💻 Implement multilingual V&L Transformers in a single code (VOLTA; Bugliarello+ 2021)

🤖 mUNITER & xUNITER (Liu&Bugliarello+, 2021) M^{3}P (Ni+, 2021) UC^2 (Zhou+, 2021)

Fine-Tuning

🇺🇸 Train on the English split

💻 On a V100 (16 GB) GPU for less than 12h

Zero-Shot Transfer

🌍 Evaluate on multilingual data

Translate-Test Transfer

🌍➡️🇺🇸 Evaluate on machine translated data

Experimental Setup

Baselines

👩‍💻 Implement multilingual V&L Transformers in a single code (VOLTA; Bugliarello+ 2021)

🤖 mUNITER & xUNITER (Liu&Bugliarello+, 2021) M^{3}P (Ni+, 2021) UC^2 (Zhou+, 2021)

Fine-Tuning

🇺🇸 Train on the English split

💻 On a V100 (16 GB) GPU for less than 12h

Zero-Shot Transfer

🌍 Evaluate on multilingual data

Translate-Test Transfer

🌍➡️🇺🇸 Evaluate on machine translated data

Few-Shot Learning

🌍 After English fine-tuning, train on few samples in each target language

📈 Performance as a function of number of shots

💻 *Max-shot* setup: evaluate with all the few-shot samples (1 run per dataset–language pair)

Experimental Setup

Baselines

🧑‍💻 Implement multilingual V&L Transformers in a single code (VOLTA; Bugliarello+ 2021)
🤖 mUNITER & xUNITER (Liu&Bugliarello+, 2021) M^{3}P (Ni+, 2021) UC^2 (Zhou+, 2021)

Fine-Tuning

🇺🇸 Train on the English split
💻 On a V100 (16 GB) GPU for less than 12h

Zero-Shot Transfer

🌍 Evaluate on multilingual data

Translate-Test Transfer

🌍➡️🇺🇸 Evaluate on machine translated data

Few-Shot Learning

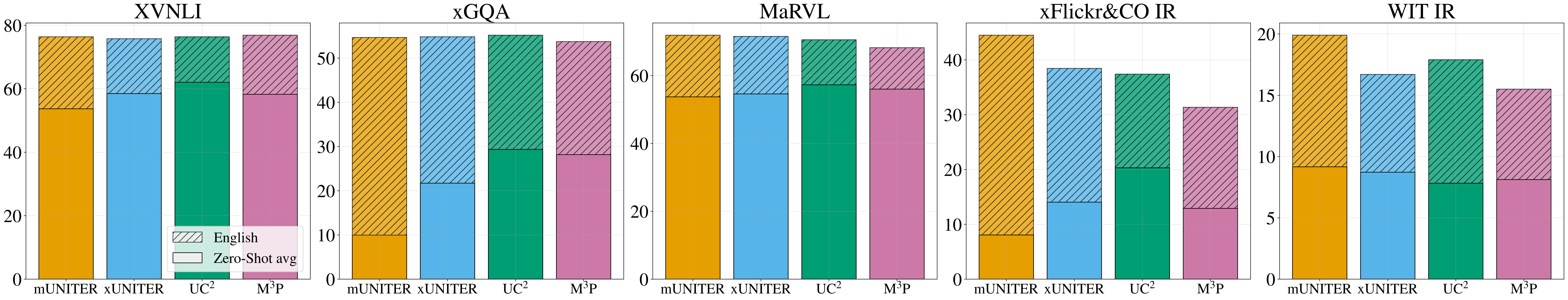
🌍 After English fine-tuning, train on few samples in each target language
📈 Performance as a function of number of shots
💻 *Max-shot* setup: evaluate with all the few-shot samples (1 run per dataset–language pair)

Metric

⚖️ Accuracy (XVNLI, xGQA, MaRVL) and Recall@1 (xFlickr&CO, WIT) – *equivalent* in our setup

Results

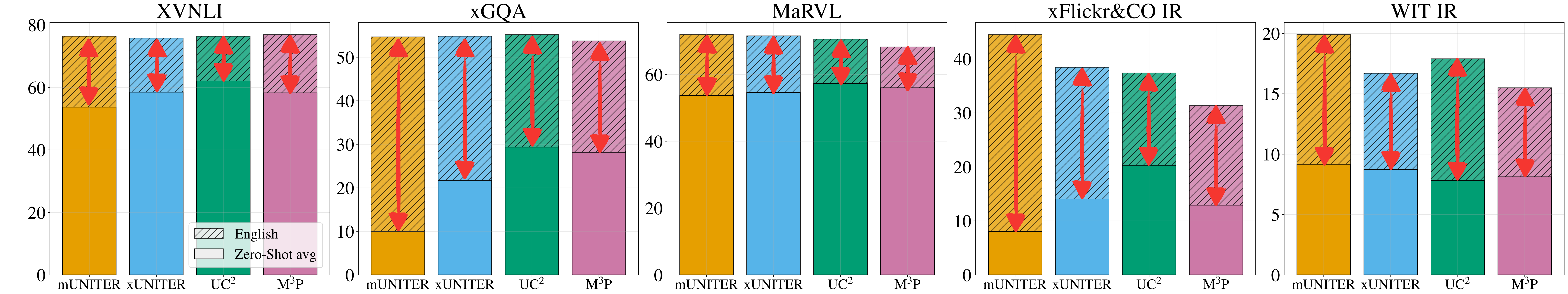
Zero-Shot Learning



Results

Zero-Shot Learning

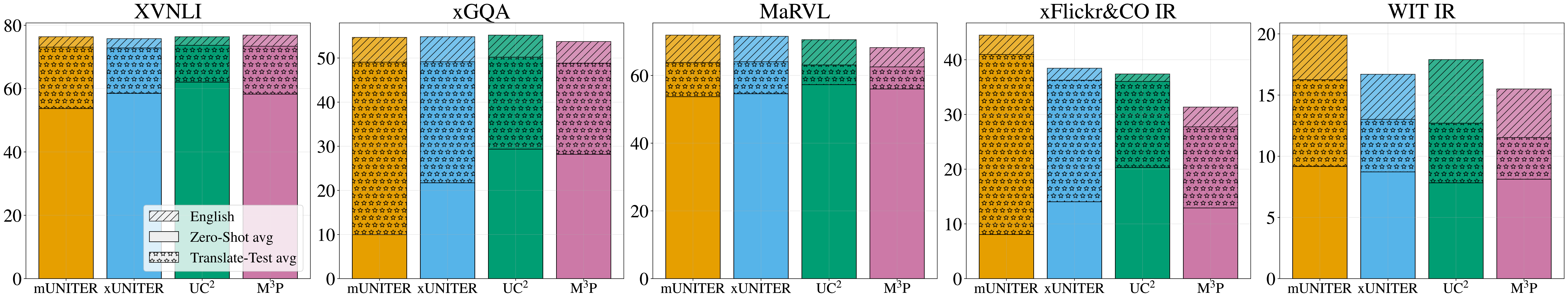
Large zero-shot transfer gap



Results

Zero-Shot Learning

Large zero-shot transfer gap

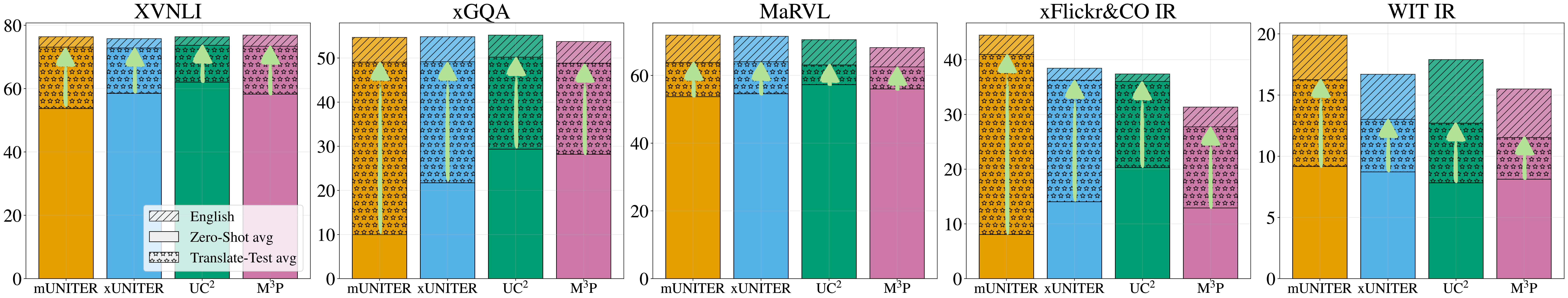


Results

Zero-Shot Learning

Large zero-shot transfer gap

Translate-test transfer » zero-shot transfer

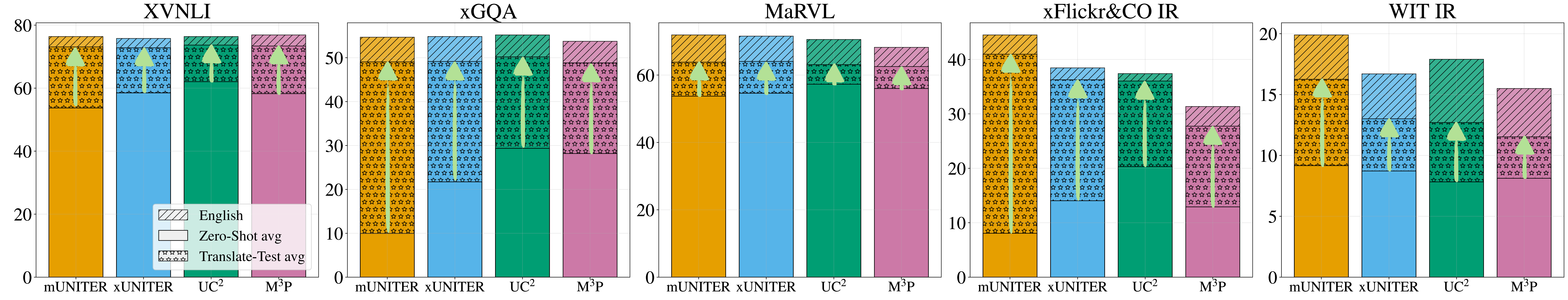


Results

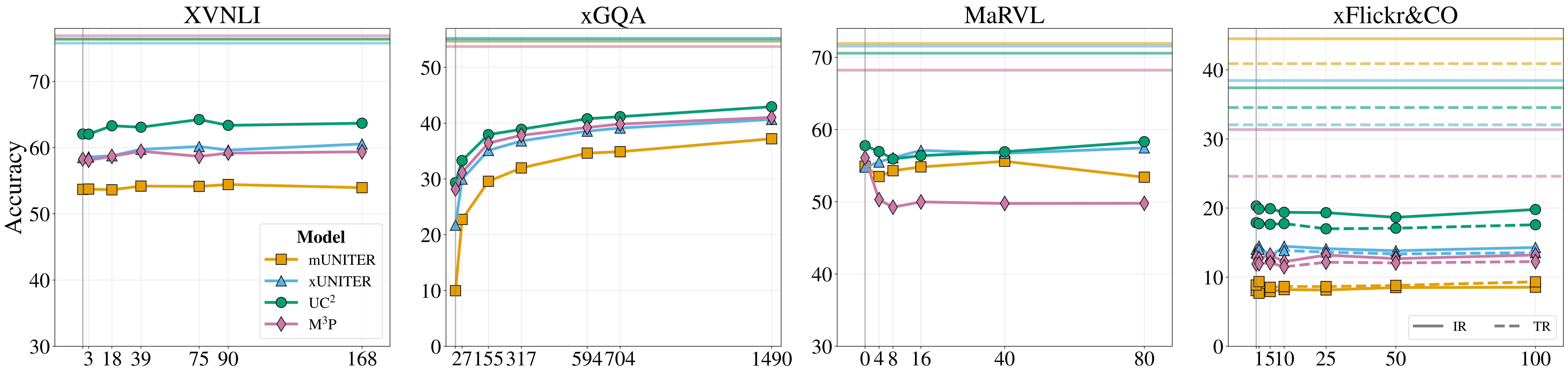
Zero-Shot Learning

Large zero-shot transfer gap

Translate-test transfer ≫ zero-shot transfer



Few-Shot Learning

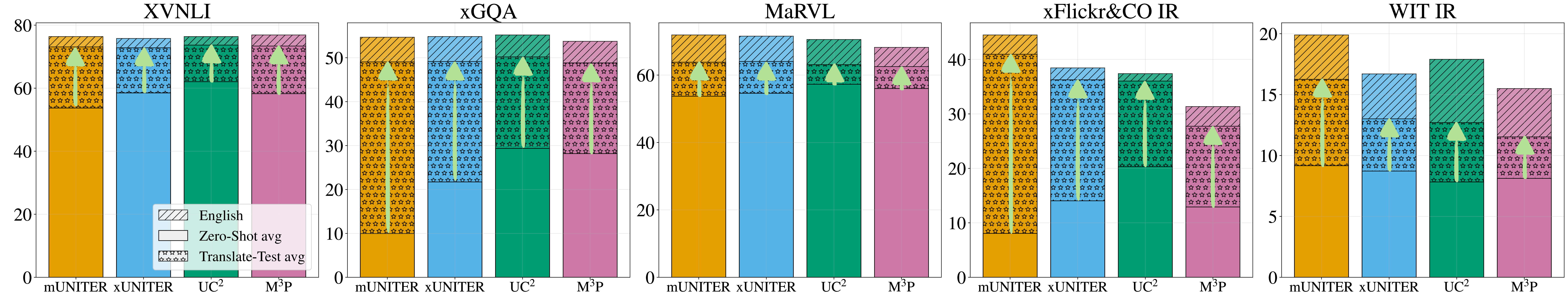


Results

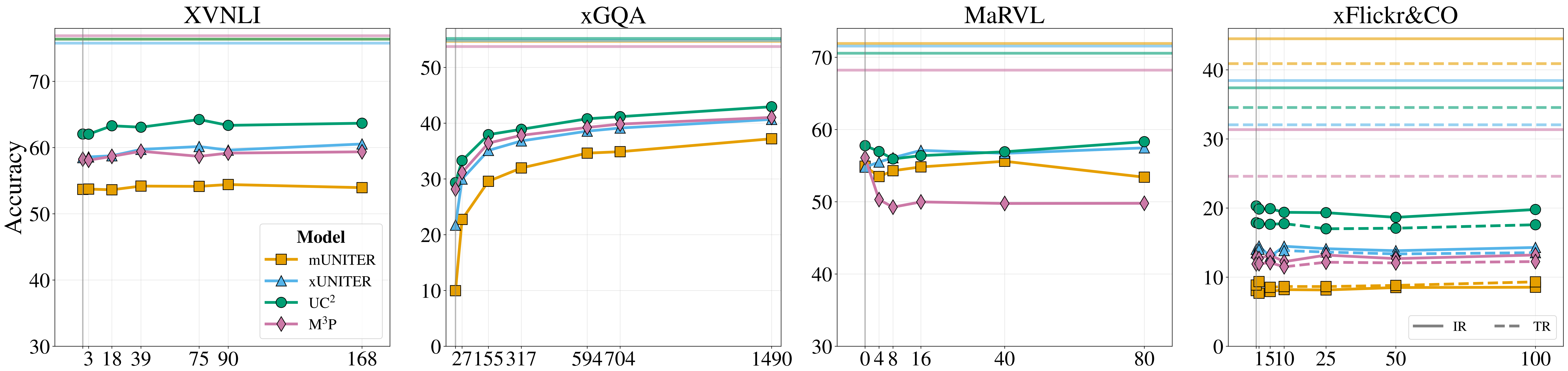
Zero-Shot Learning

Large zero-shot transfer gap

Translate-test transfer ≫ zero-shot transfer



Few-Shot Learning



Consistent but moderate gains

Conclusion & Outlook

IGLUE: The Image-Grounded Language Understanding Evaluation benchmark

- 5 datasets across 4 tasks in 20 languages
- Zero-shot & few-shot transfer setups show large drops in performance wrt English
- Code, data and pretrained models available online
  - iglue-benchmark.github.io
  - github.com/e-bug/iglue

Conclusion & Outlook

IGLUE: The Image-Grounded Language Understanding Evaluation benchmark

- 5 datasets across 4 tasks in 20 languages
- Zero-shot & few-shot transfer setups show large drops in performance wrt English
- Code, data and pretrained models available online
  - iglue-benchmark.github.io
  - github.com/e-bug/iglue

Next Steps

- Transfer learning across modalities, tasks, and languages
- Single- vs. multi-source transfer
- Beyond image-only tasks (e.g. videos and speech)

Conclusion & Outlook

Thank you

IGLUE: The Image-Grounded Language Understanding Evaluation benchmark

- 5 datasets across 4 tasks in 20 languages
- Zero-shot & few-shot transfer setups show large drops in performance wrt English
- Code, data and pretrained models available online
  - iglue-benchmark.github.io
  - github.com/e-bug/iglue

Next Steps

- Transfer learning across modalities, tasks, and languages
- Single- vs. multi-source transfer
- Beyond image-only tasks (e.g. videos and speech)