
# Multilingual Multimodal Pre-training for Zero-Shot Cross-Lingual Transfer of Vision-Language Models

http://github.com/berniebear/Multi-HT100M

Po-Yao (Bernie) Huang*, Mandela Patrick*, Junjie Hu,
Graham Neubig, Florian Metze, Alexander Hauptmann

NAACL 2021

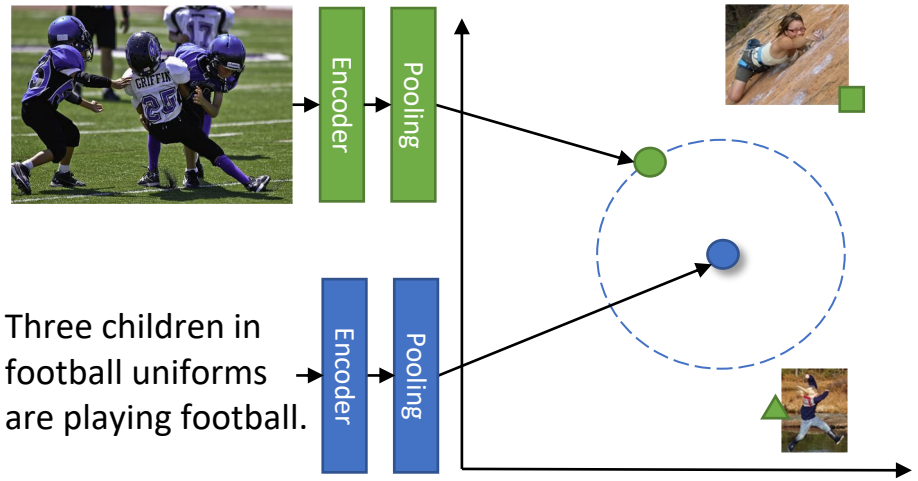
Carnegie Mellon University
Language Technologies Institute

UNIVERSITY OF OXFORD

FACEBOOK

# Introduction

- Most vision-language models/tasks are English-centered. It is challenging to generalize these V-L models to other 7000 languages.
- Possible solutions:
  - Re-collect vision-language datasets for all languages => $$$
  - Machine translation => low-resource/distant languages and Inference time memory/computation cost
  - Our solution: **Zero-shot cross-lingual transfer of V-L models** (one model to rule them all)!
    - Multilingual multimodal Transformers + Multilingual multimodal pre-training
    - English-only fine-tuning then directly inference with non-English inputs




What is the mustache made of?


A young boy in a white striped shirt holding a tennis racket .

Cross-modal search

# Cross-Lingual Transfer for NLP

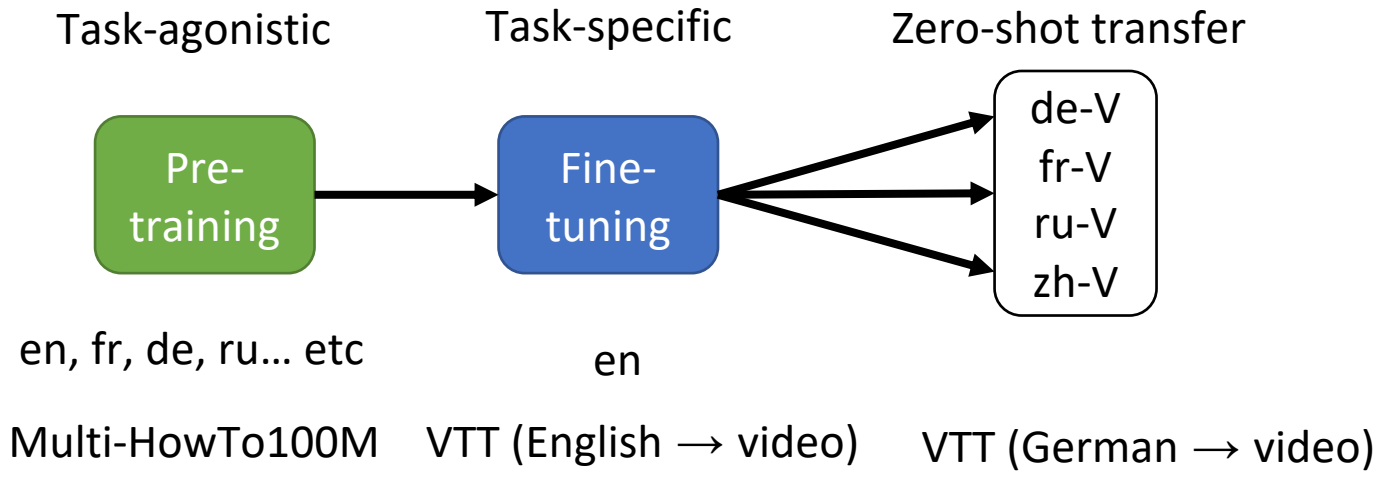
- Common practice in NLP
  - Multilingual pre-training
  - Task-specific English fine-tuning
  - Zero-shot transfer of English fine-tuned NLP model to other languages
- Related work in NLP:
  - XNLI
  - Multilingual BERT, XLM-Roberta
  - EXTREME



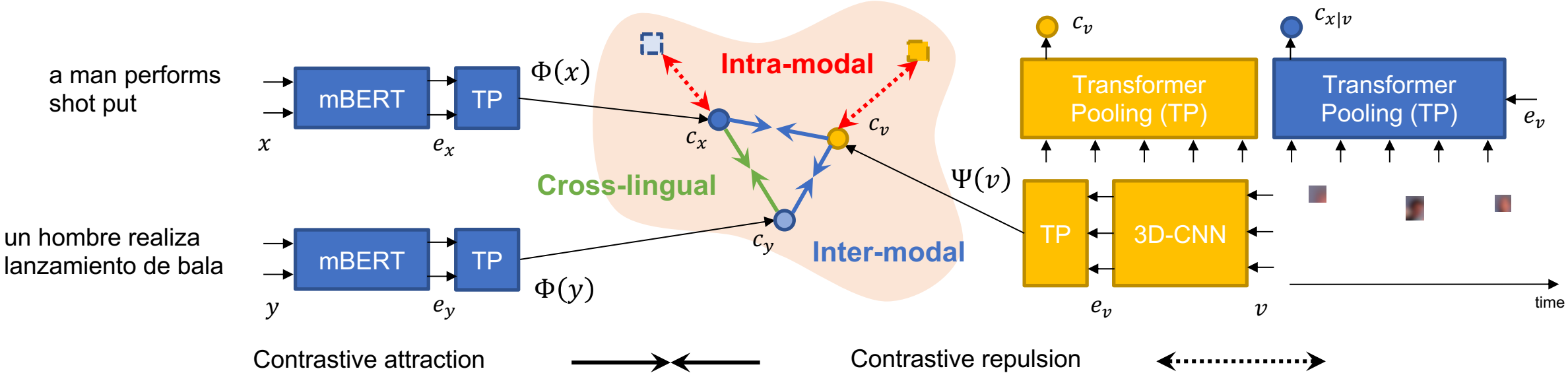

Animation from Junjie Hu (EXTREME)

# Cross-Lingual Transfer for V-L models

- Proposed framework:
  - **Multilingual multimodal Transformers**
  - **Multilingual multimodal pre-training**
  - Task-specific English-vision fine-tuning
  - Zero-shot transfer of English-vision fine-tuned model to other languages




# Multilingual Multimodal Transformer




$$\mathcal{L}^{\text{inter}} = \mathcal{L}(\mathcal{X}, \mathcal{V}) + \mathcal{L}(\mathcal{Y}, \mathcal{V})$$

$$\mathcal{L}^{\text{intra}} = \mathcal{L}(\mathcal{X}, \mathcal{X}^m) + \mathcal{L}(\mathcal{Y}, \mathcal{Y}^m) + \mathcal{L}(\mathcal{V}, \mathcal{V}^m)$$

$$\mathcal{L}^{\text{cross}} = \mathcal{L}(\mathcal{X}|\mathcal{V}, \mathcal{Y}|\mathcal{V})$$

$$\mathcal{L}(\mathcal{X}, \mathcal{V}) = -\frac{1}{B}\sum_{i=1}^{B} \log \ell^{\text{NCE}}(\Phi(x_i), \Psi(v_i))$$

$$\ell^{\text{NCE}}(c_x, c_v) = \frac{e^{s(c_x, c_v)}}{e^{s(c_x, c_v)} + \sum_{(x', v') \sim \mathcal{N}} e^{s(c_{x'}, c_{v'})}}$$

# Multilingual Multimodal Pre-training

- Multi-HowTo100M: a multilingual version of HowTo100M
  - 1.2 million instructional videos, 138 million video clips
  - Video transcriptions in 9 languages
    - English, German, French, Russian, Spanish, Czech, Swahili, Chinese, Vietnamese

# Experiment Setup

- Choice of text backbone:
  - mBERT
  - XLM-R (large)
- (Multilingual) Multimodal Pre-training
  - HowTo100M
  - Multi-HowTo100M
- Fine-tuning: MSR-VTT
  - Evaluation task:
    - English → video search
    - (Zero-shot) non-English → video search

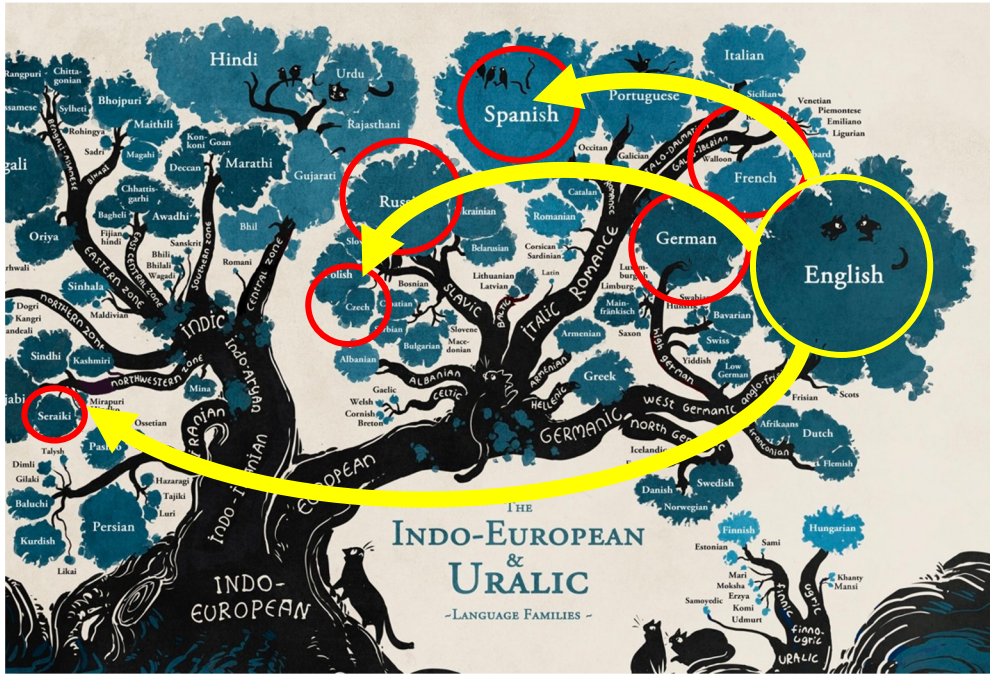

Instructional videos in HT100M


***Query 537:*** one or more people swimming in a swimming pool

Video and its caption in VTT

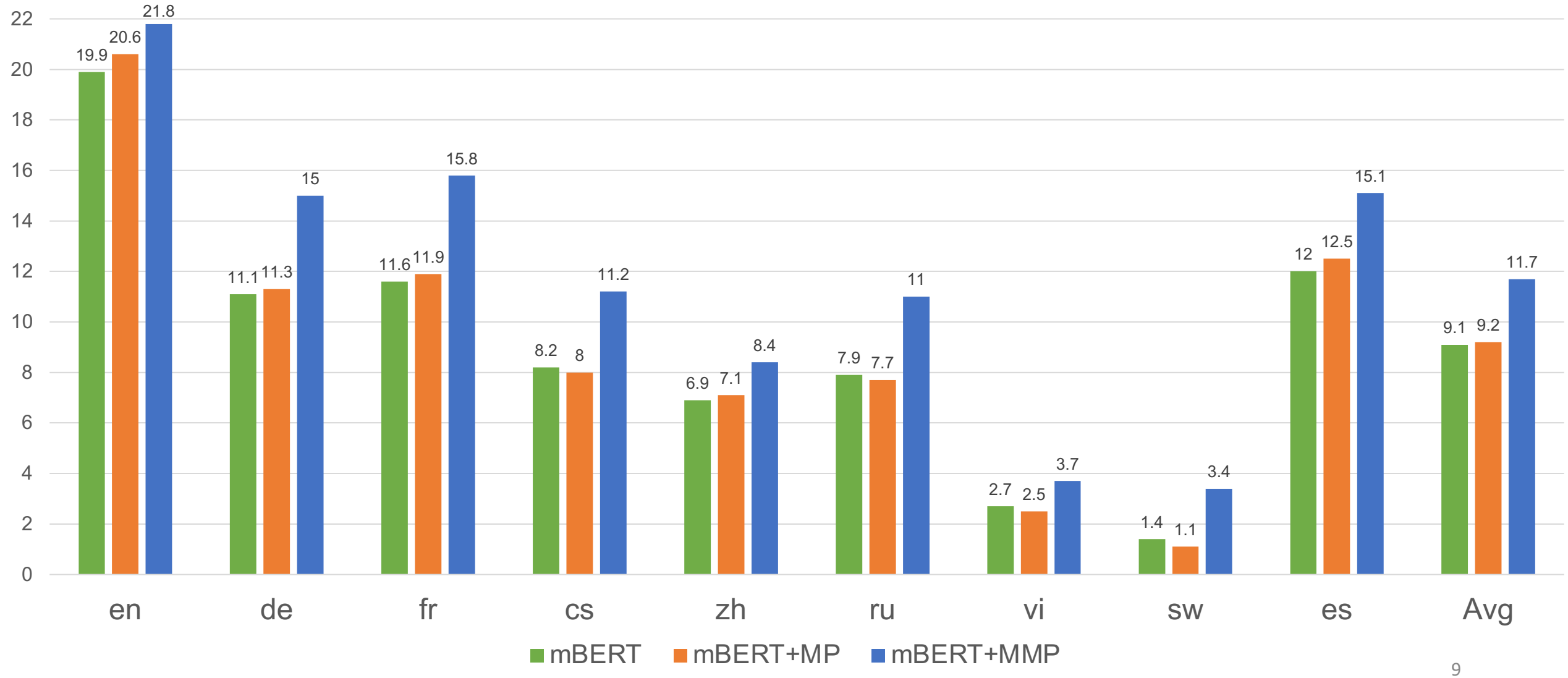
# We will answer the following research questions:

- Is multilingual BERT sufficient for zero-shot cross-lingual transfer of V-L models?
- Does English-video model benefit from multilingual multimodal pre-training (MMP)?
- Does MMP transfer well to distant languages?
- Does additional language(s) help?



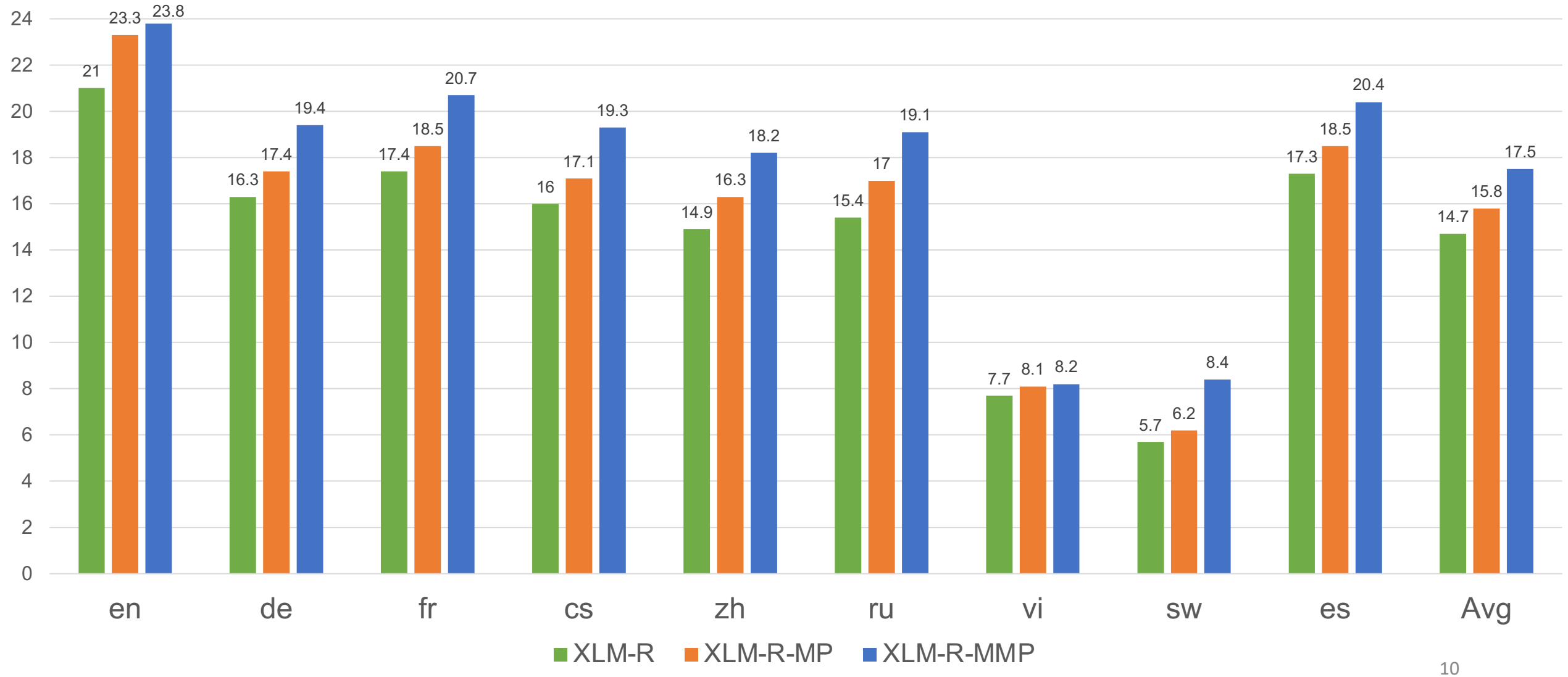
# Multilingual-Text → Video Search on VTT (mBERT)

MP: Multimodal Pre-training (HT100M)
MMP: Multilingual MP (Multi-HT100M)

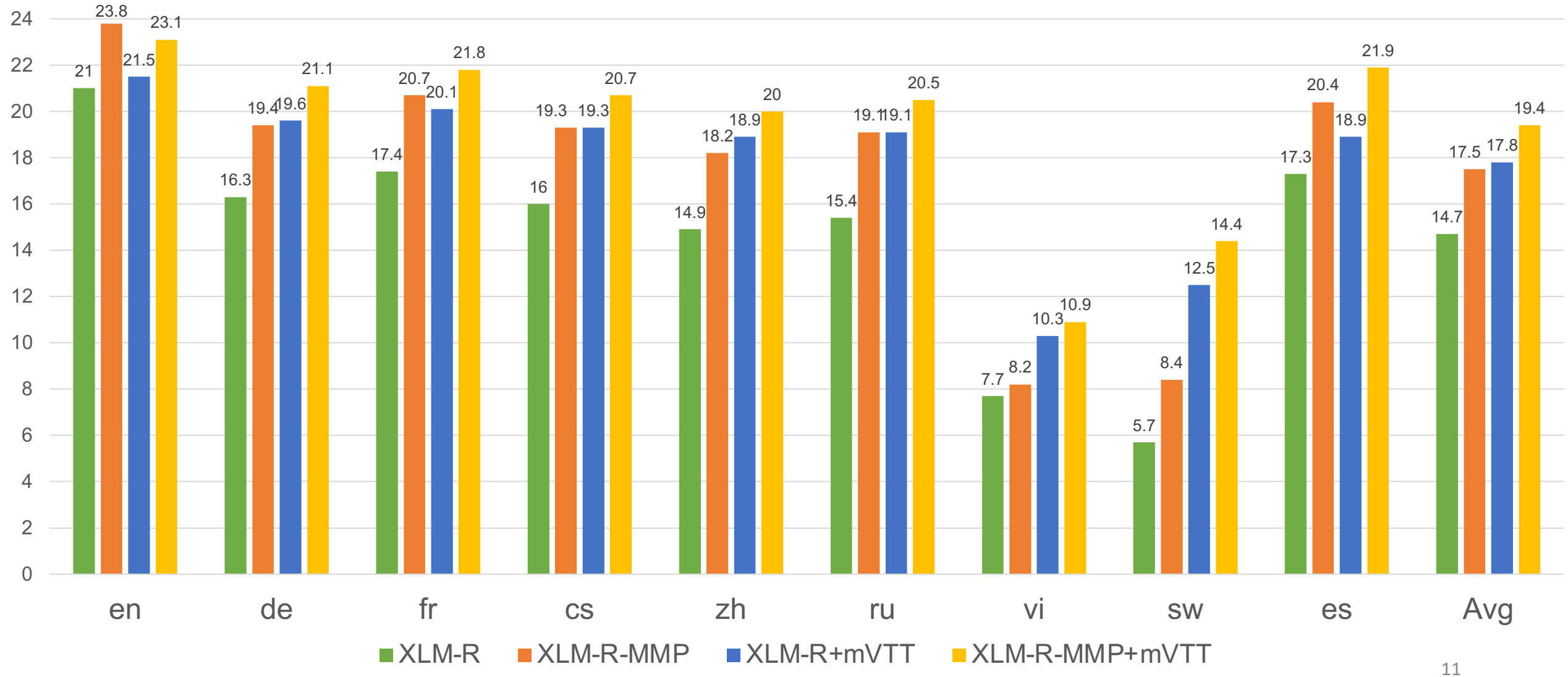


# Multilingual-Text → Video Search on VTT (XLM-R)

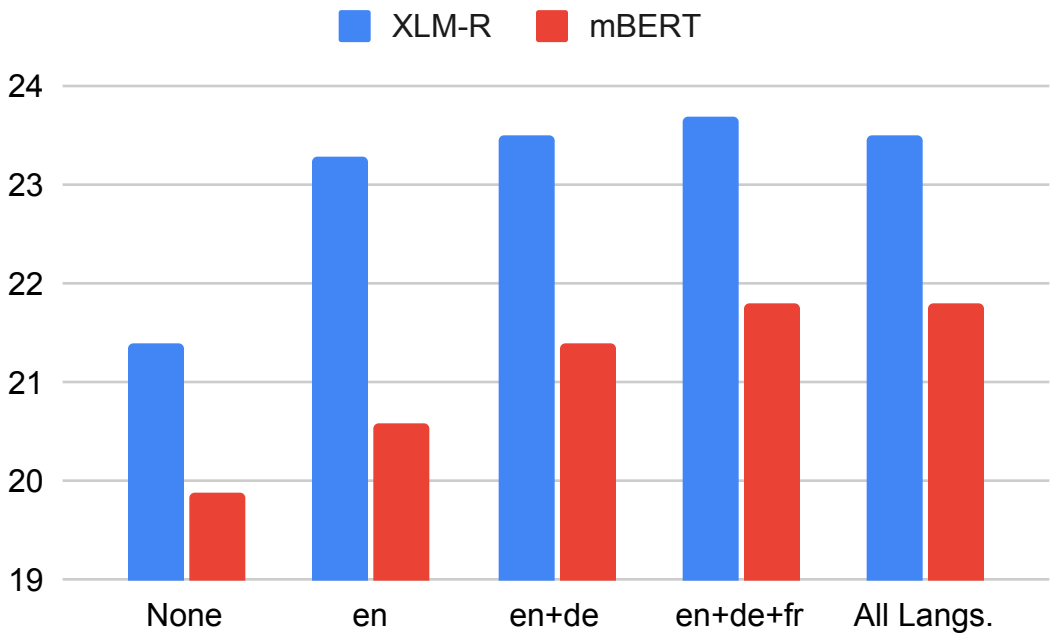
MP: Multimodal Pre-training (HT100M)
MMP: Multilingual MP (Multi-HT100M)



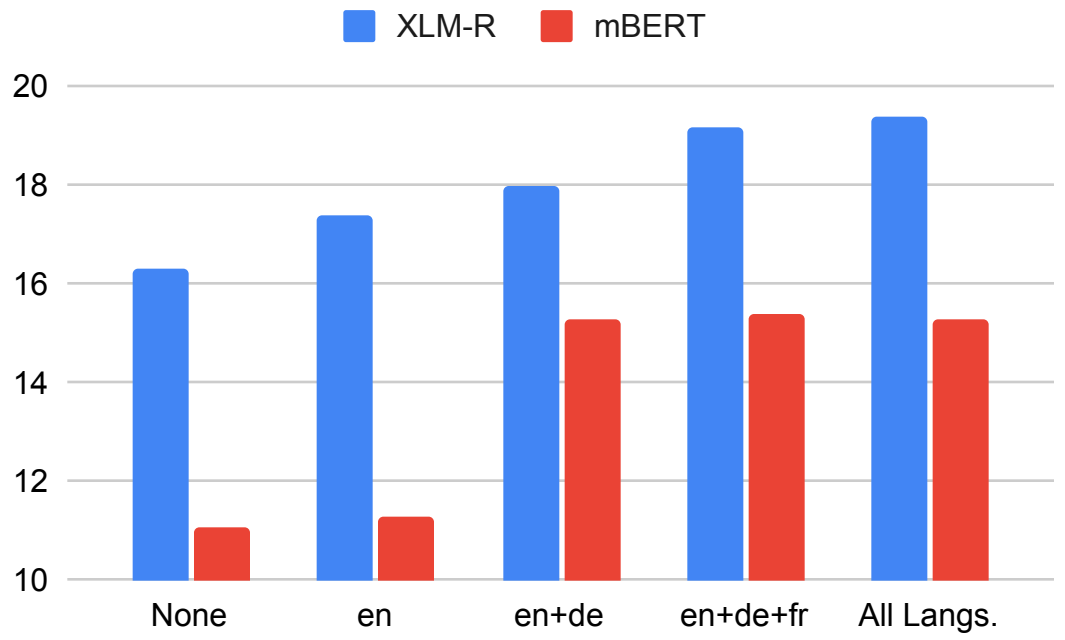
# Multilingual-Text → Video Search on VTT (w/ machine translated VTT)



# Does additional language(s) help?

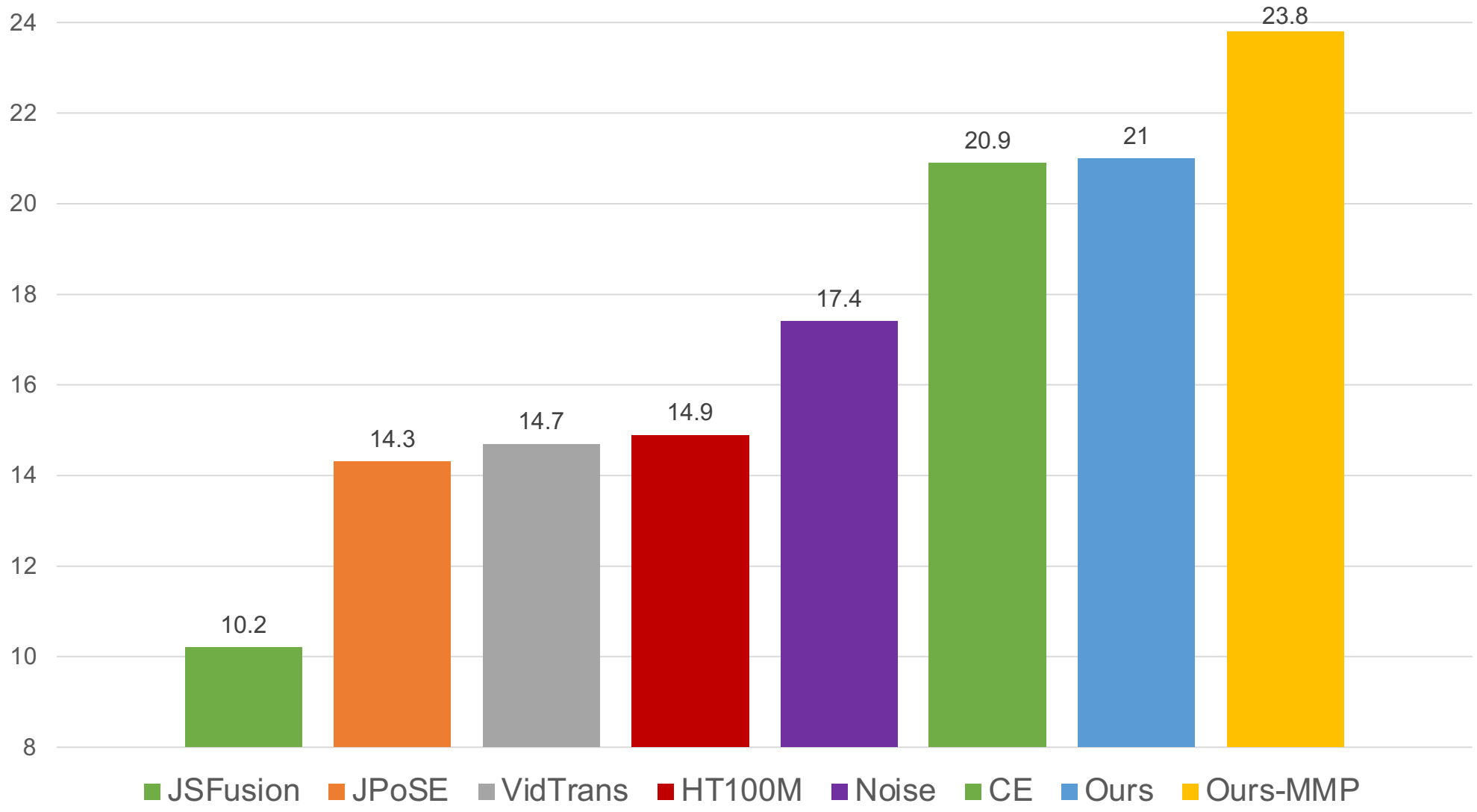



English → Video

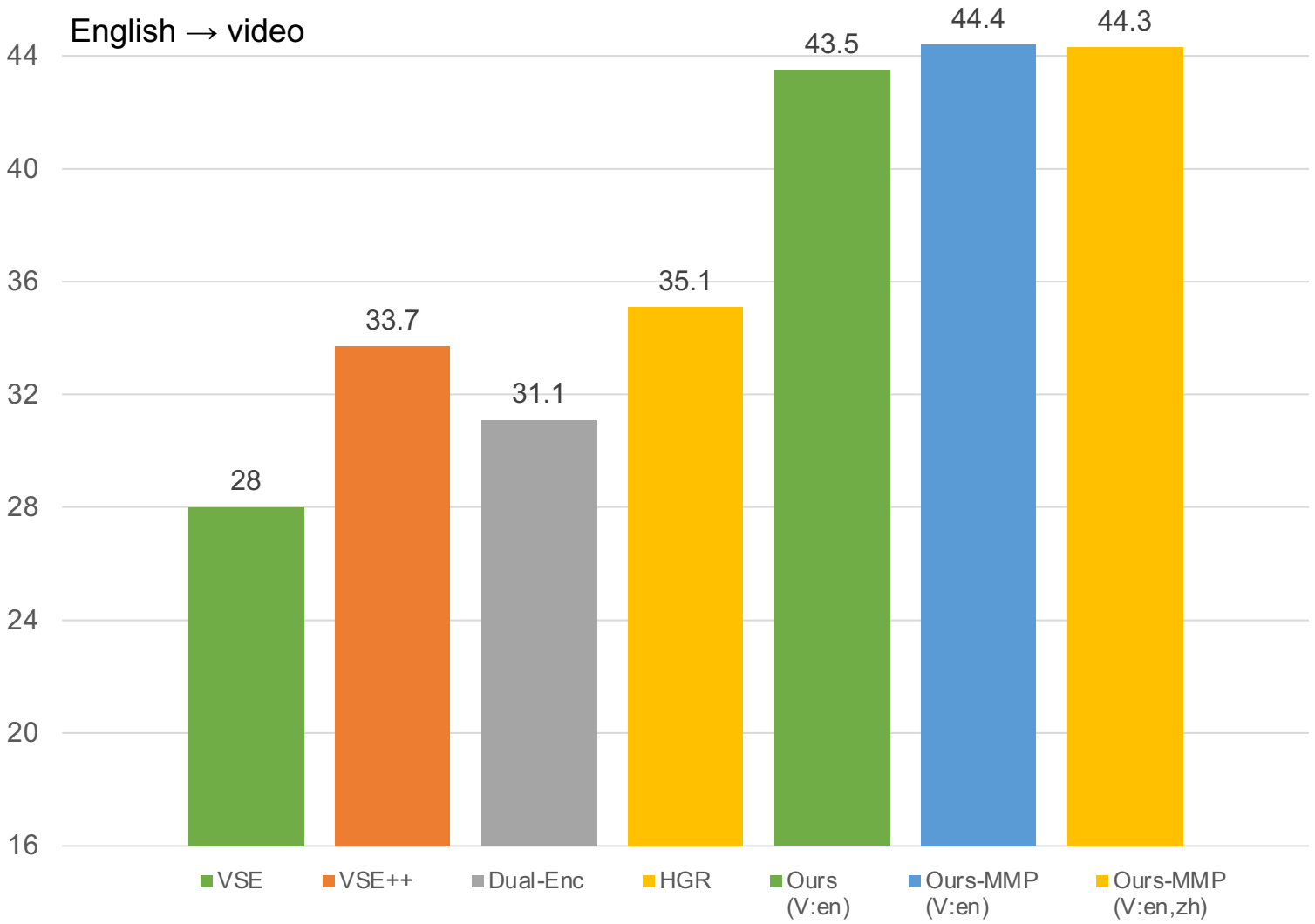


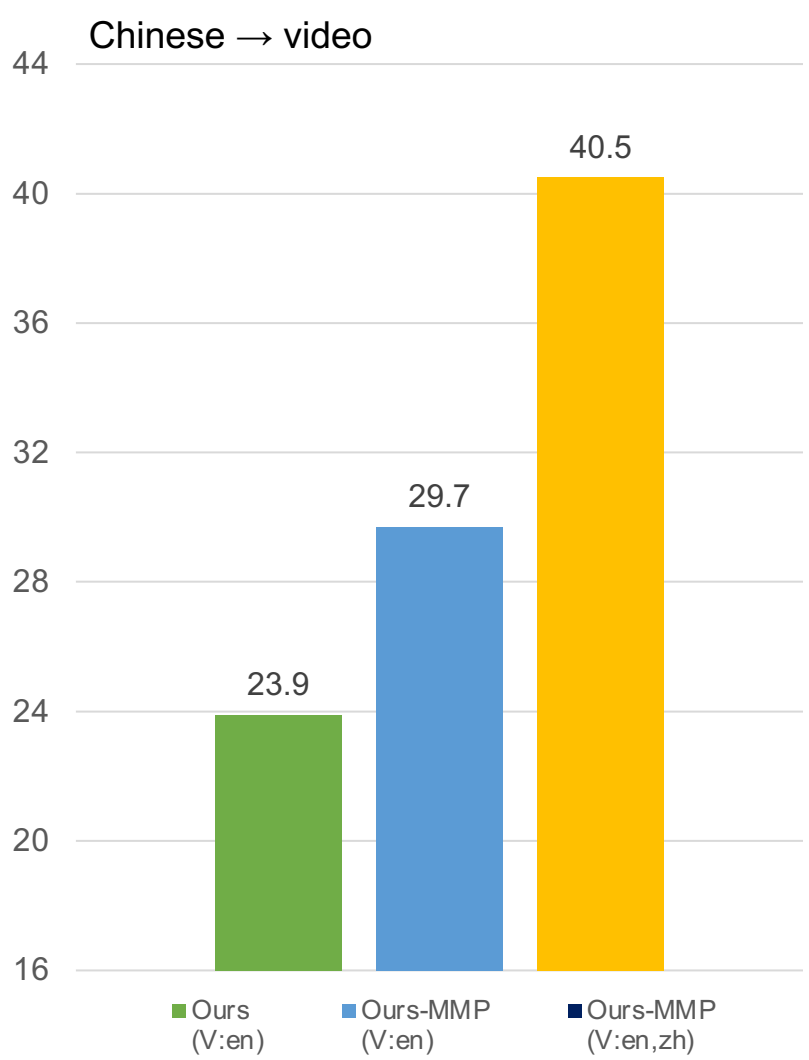

Zero-Shot German → Video

# Comparison to SoTA English-video models on VTT



# Multilingual-text → video search on VATEX (English/Chinese)

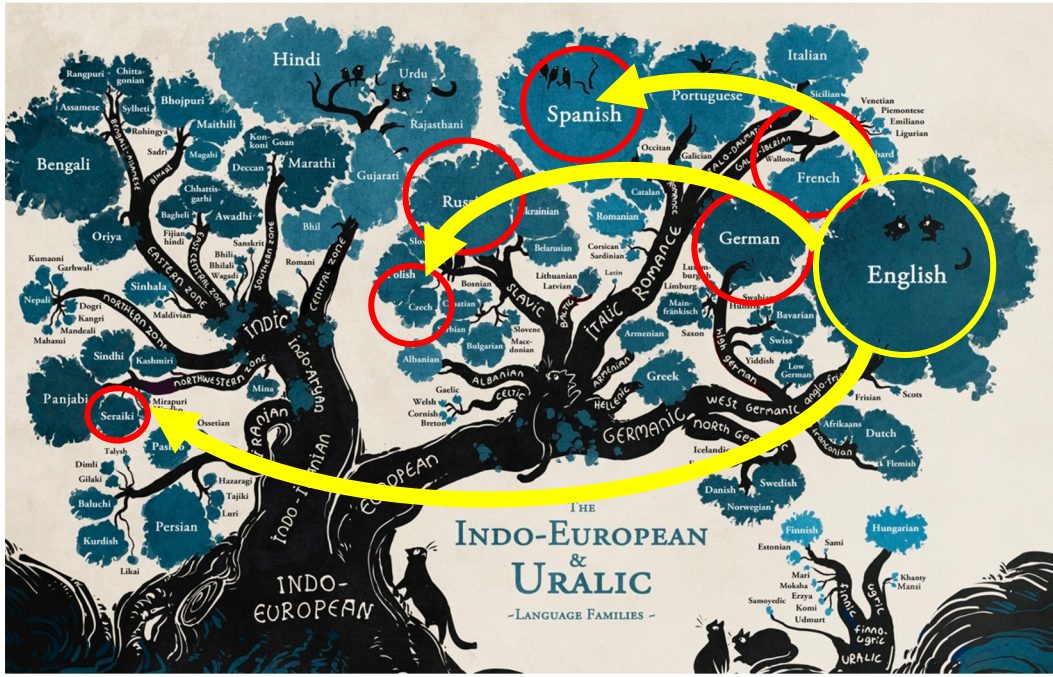


# We answered the following research questions:

- Multilingual BERT is insufficient for zero-shot cross-lingual transfer of V-L models.
  - De: P(en+de) + F(en) > P(en) + F(en) > F(en)
- Multilingual multimodal pre-training also benefits English-video models.
  - En: P(en+de) + F(en) > P(en) + F(en) > F(en)
- MMP transfer well to distant languages.
  - P(en+de+…+zh) + F(en) -> zh
- Additional language(s) helps
  - P(en+de+…+zh) + F(en) -> de++

P: Multimodal Pre-training (HT100M)
P: Multilingual Multimodal Pre-training (Multi-HT100M)
F: English-vision fine-tuning



# Take home message:

- Cross-lingual transfer of V-L model is feasible but challenging!
- Essential ingredients for its success:
  - Multilingual multimodal Transformers
  - Multilingual multimodal pre-training on Multi-HowTo100M


thank you!
хвала
dankon
감사합니다
dank je
dank u
gracias
ありがとう
kiitos
danke
takk
grazas
ขอขอบคุณคุณ
grazie
falemnderit
paldies
hvala
buíochas a ghabháil leat
gratias agimus tibi
di ou mèsi
شکرا
дзякуй
sukriya
teşekkür ederim
ačiū
धन्यवाद
köszönöm
ви благодариме
diolch i chi
terima kasih
d'akujem
σας ευχαριστώ
asante
aitäh
tack
谢谢
dziękuję
þakka þér
mulţumesc
dankie
gràcies
ありがとう
eskerrik asko
děkuji
merci
спасибо
obrigado