Image-Text Modeling for Multilingual NLP

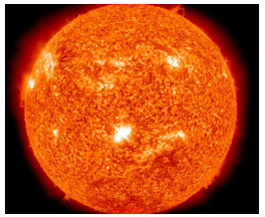
Simran Khanuja

Carnegie Mellon University

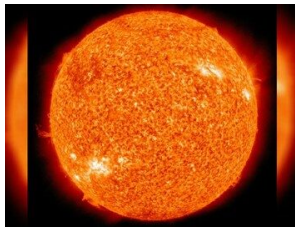
In this lecture we will talk about …

- **Why "*vision*" in multilingual NLP?**
- **Four types of multilingual image-text models**
  - Cross-encoder based (understanding)
  - Dual-encoder based (retrieval)
  - Encoder-decoder based (understanding + text generation)
  - Diffusion based (image generation)
- **For each type we will cover**
  - Downstream task capabilities
  - Model Architecture
  - Pre-training data resources
  - Deep Dive into one model
- **Discussion of biases**
- **Revisit motivation + Open questions**

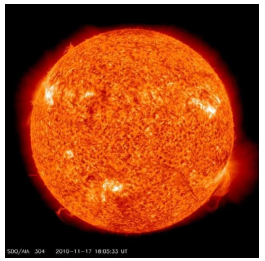
Some concepts *look* the same across languages and cultures



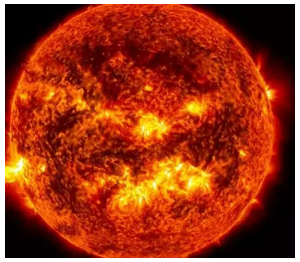
sun



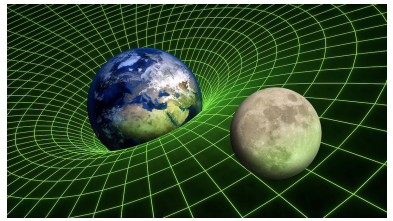
सूरज



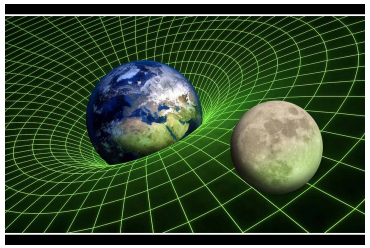
太陽



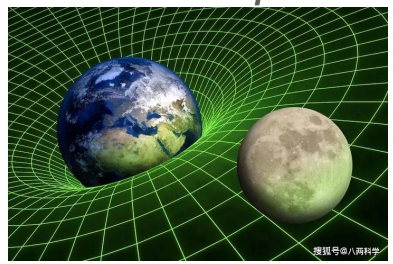
güneş



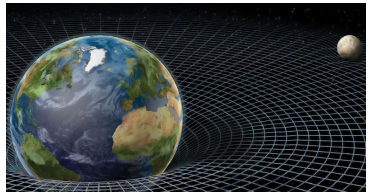
gravity



गुरुत्वाकर्षण

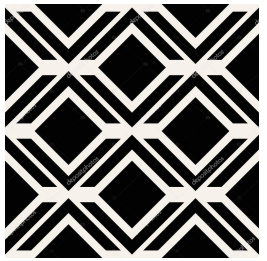


重力

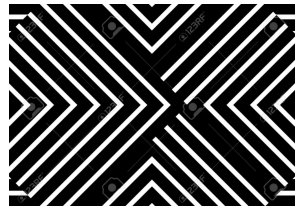


yer çekimi

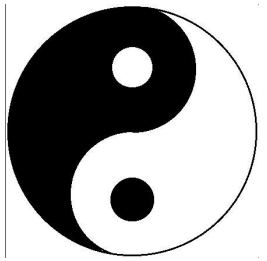
black and white



काला और सफेद



黒と白



siyah ve beyaz

[List of Human Universals (Donald Brown, 1991)](); [Psychological universals: what are they and how can we know?]()

Image Source: Google Search

Some may universally exist but yet *look* different

wedding

शादी

結婚式

düğün

music

संगीत

音楽

müzik

sports

खेल

スポーツ

Spor Dalları

[List of Human Universals (Donald Brown, 1991)](); [Psychological universals: what are they and how can we know?]()

Image Source: Google Search

Some may not exist cross-culturally at all …

Some concepts are only 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

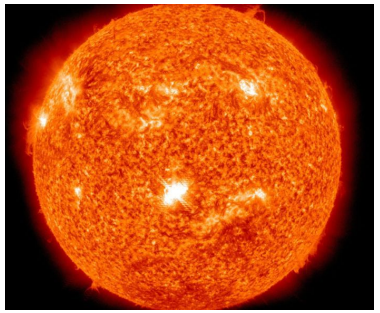
Sanxian / Shamisen

Clavie

Source: Desmond Elliott's LxMLS Slides

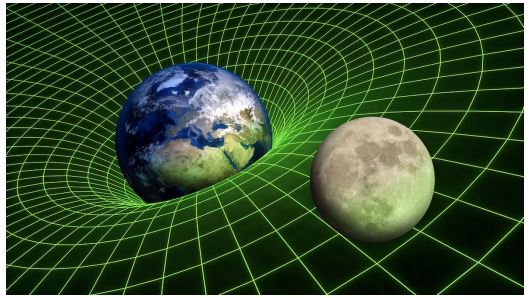
As an anchor point for multiple languages for universal concepts

sun



太陽

gravity



yer çekimi

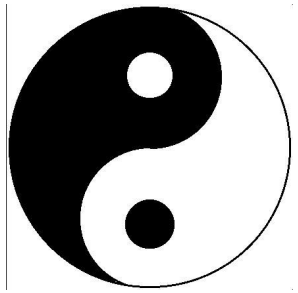
सूरज

güneş

गुरुत्वाकर्षण

重力

black and white



siyah ve beyaz

काला और सफेद

黒と白

As a mechanism to enhance cultural diversity within concepts




wedding
düğün
結婚式
शादी




音楽
müzik
music
संगीत




Spor
Dalları
sports
खेल
スポーツ

To represent culturally unique objects/events beyond text

Pilota / Jai-alai

sport
racquet
court
actions

Sanxian / Shamisen

musical instrument
strings
action
pose

Clavie

fire
people
night

Status quo

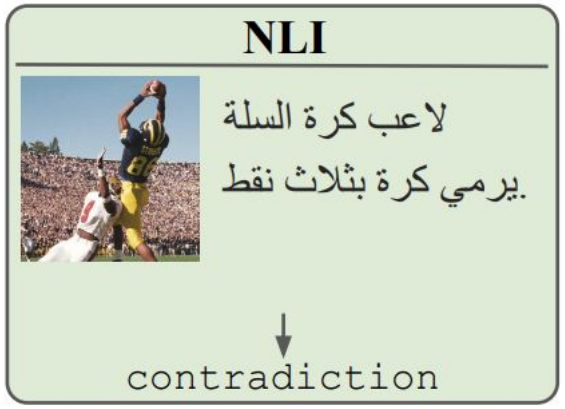
So what progress has been made thus far in the field to support research in image-text multilingual NLP?

We will be looking at **four types** of image-text multilingual models that exist today. For each type of models, we will cover :

- *Evaluation Tasks*
- *Model Architecture*
- *Data*: Pre-training datasets and languages
- *Pre-training tasks*
  - Deep-dive into a state-of-the-art model
- *[Optional] Demo*
- *[Optional] Bias Discussion*

Overview of Downstream Tasks (Understanding)

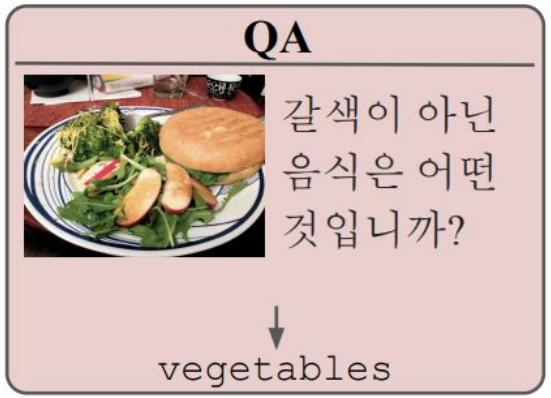
Cross-lingual Visual NLI (XVNLI)



ENG: The basketball player shoots a three pointer

- Infer whether a text-hypothesis *entails*, *contradicts*, or is *neutral* to an image-premise.
- *Comprised of*: SNLI (Bowman et al., 2015), with its multimodal (Xie et al., 2019) and cross-lingual (Agic & Schluter., 2018) counterparts.
- English, Arabic, French, Russian, Spanish

Cross-lingual Grounded QA (xGQA)

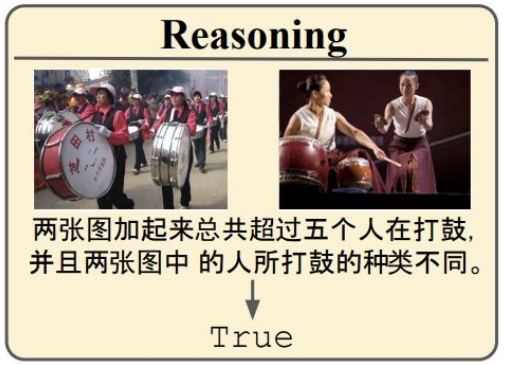


ENG: Which kind of food is not brown?

- Answer several types of structured questions about an image
- *xGQA* (Pfeiffer et al. 2022) which is translated from GQA (Hudson & Manning, 2019), into 7 languages
- *MaXM* (Changpinyo et al. 2022), automatically generate QA pairs from XM3600 for 7 langs

Overview of Downstream Tasks (Understanding+Retrieval)

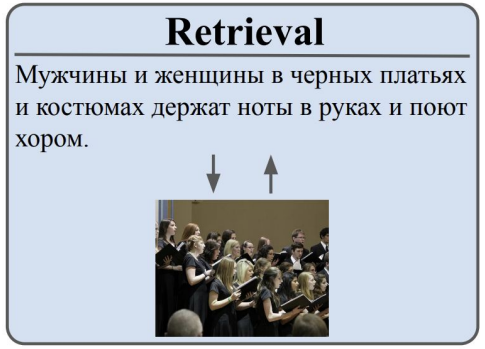
Multicultural V-L Reasoning



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.

- Infer whether statement is true or false about a pair of images.
- MaRVL (Liu et al., 2021): Dataset constructed from scratch by native speakers **[EMNLP Best Paper]**
- Indonesian, Mandarin, Tamil, Turkish, Swahili

Multi-X Retrieval



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

- *xFlickr&CO.*: 1000 from Flickr30K (Young et al., 2014) + COCO (Lin et al., 2014); captions crowdsourced for 7 langs
- *WIT* (Srinivasan et al., 2021): Wikipedia in 108 languages (IGLUE covers 10)
- *XM3600* (Thapliyal et al., 2022): Captions for 3600 geographically diverse images in 36 languages

Overview of Downstream Tasks (Generation)

Image Captioning

Source: Porsche Museum, Stuttgart by Brian Solis.

Dutch	nl	klassieke raceauto's op een rij een museum
English	en	The branded classic cars in a row at display.
Persian	fa	تعدادی اتومبیل مسابقه ای اسپرت در یک نمایشگاه سرپوشیده
Filipino	fil	mga klasikong sasakyan na nakadisplay sa tindahan ng kotse
Finnish	fi	Antiikkisia urheiluautoja näyttelyssä
French	fr	Une série de voitures de course vintage exposé dans un musée
German	de	Verschiedenfarbige Rennwagen mit Nummern auf einer Autoausstellung
Greek	el	έκθεση ρετρό αγωνιστικών αυτοκινήτων
Hebrew	he	רכב פורשה וינטג' עם גג נפתח בצבע כסוף עומד בתצוגה ליד רכבים נוספים.
Hindi	hi	सफ़ेद फर्श पर ठैरी कोरे रंग की गाडी और उसके पीछे और भी गाड़ियां

- The task is to generate a caption for an image in a language of choice.
- Are image descriptions consistent across cultures? Research shows otherwise (Ye et al. 2023, Vossen et al., 2017).
- [Open Question] Beyond BLEU, CIDEr, how do we capture the cultural relevance of a caption?

Text → Image Generation

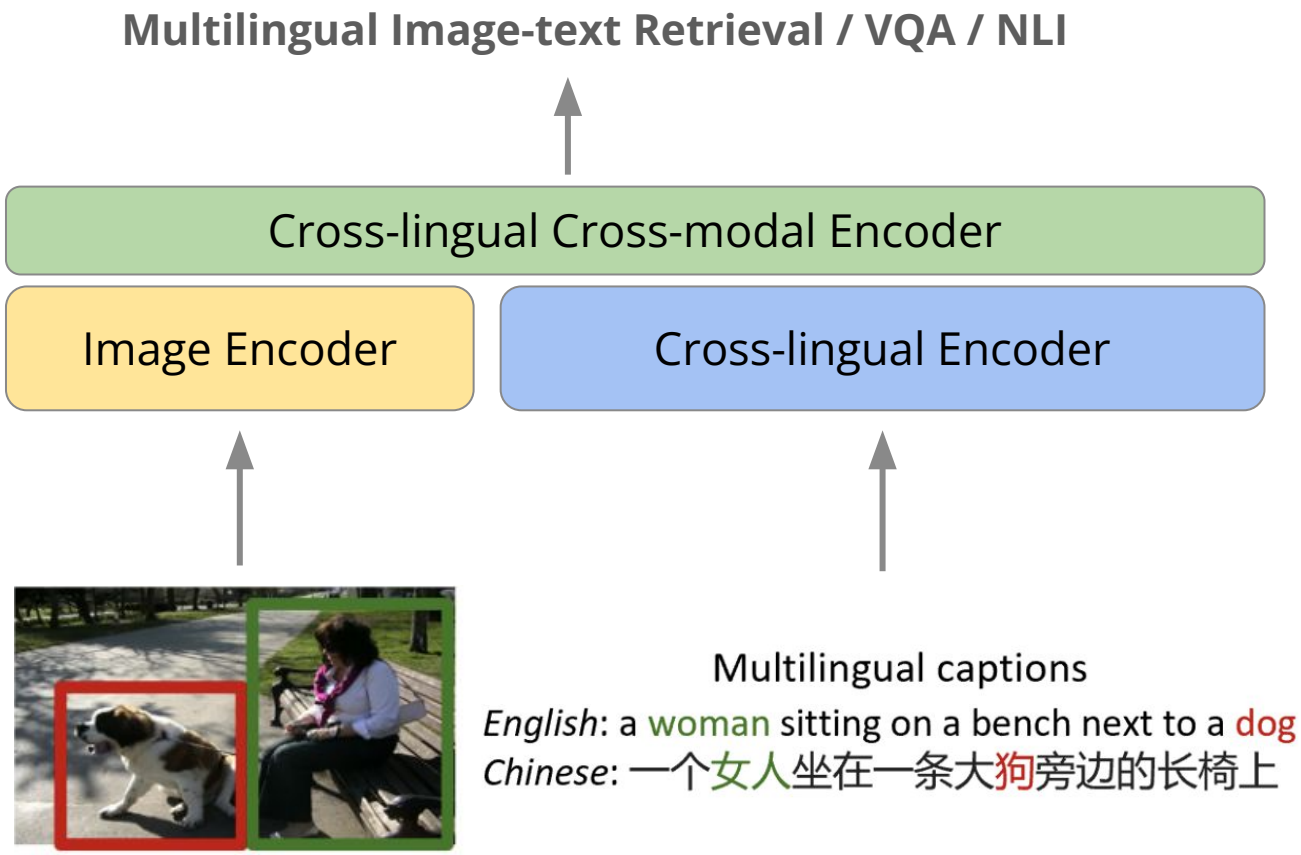
Chinese:一只柯基犬在丛林中奔跑

French:Un chien corgi court dans la jungle

English: A corgi dog runs in the jungle

- The task is to generate an image given a multilingual caption
- Evaluation metrics include FID (can be biased due to training on ImageNet), and image-caption similarity using multilingual CLIP
- [Open Question] How do we capture cultural representativeness of output, while making sure the model is not stereotyping/biased?

Category 1: Multi-X Understanding (Cross-Encoding Based Models)

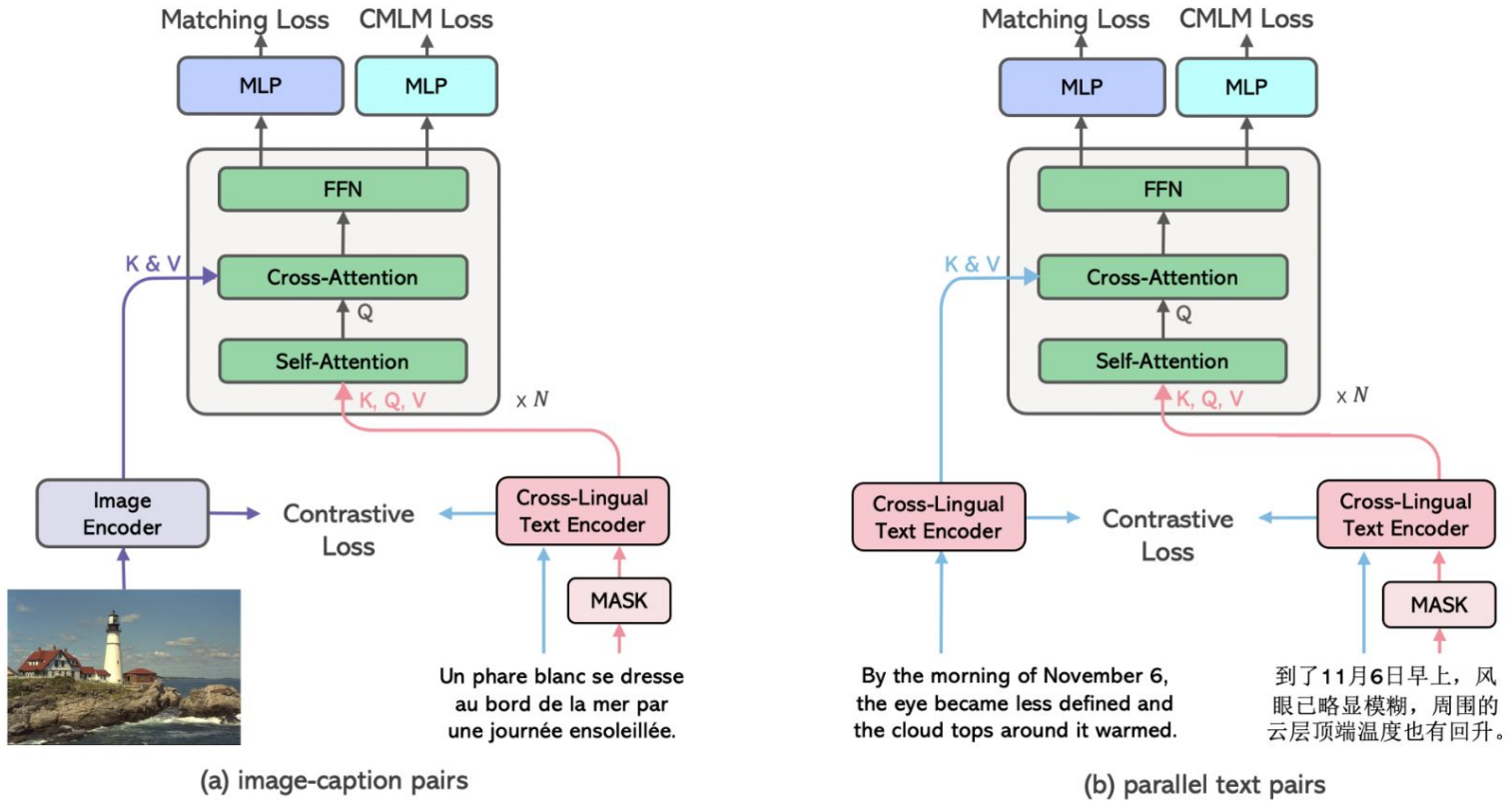


M3P: Learning Universal Representations via Multitask Multilingual Multimodal Pre-training (Ni et. al, CVPR 2021)

UC2 : Universal Cross-lingual Cross-modal Vision-and-Language Pre-training (Zhou et. al, CVPR 2021)

[CCLM] Cross-View Language Modeling: Towards Unified Cross-Lingual Cross-Modal Pre-training (Zeng et. al, ACL 2023)

Deep Dive: CCLM (SoTA Encoder-based model)

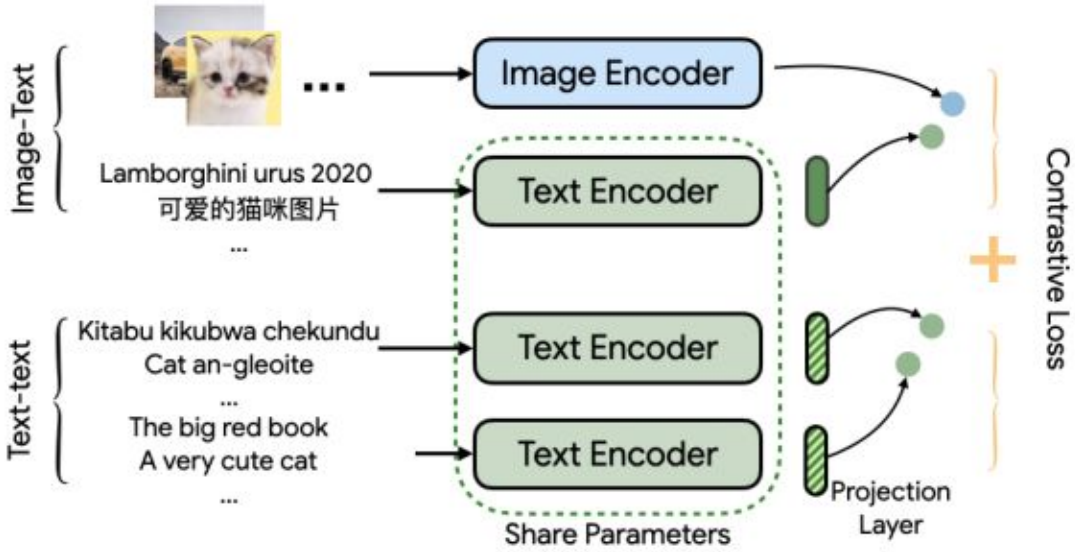


[CCLM] Cross-View Language Modeling: Towards Unified Cross-Lingual Cross-Modal Pre-training (Zeng et. al, ACL 2023)

Pre-training data for Multi-X understanding models

Data Resource Type →	Multilingual (text)	Multimodal (English-only)	Multi-X (translated)	Multi-X
Datasets	Wikipedia (100L); Wikimatrix (para; 20L);	CC3M (en; 3.3M)	CC3M (5L; 3.3M);	
Models	M3P (Wikipedia)	M3P (CC3M)		
			UC2 (CC3M-translated)	
	CCLM (Wikimatrix)		CCLM (CC3M-trans)	

Category 2: Multi-X Understanding (Dual-encoder Based Models)



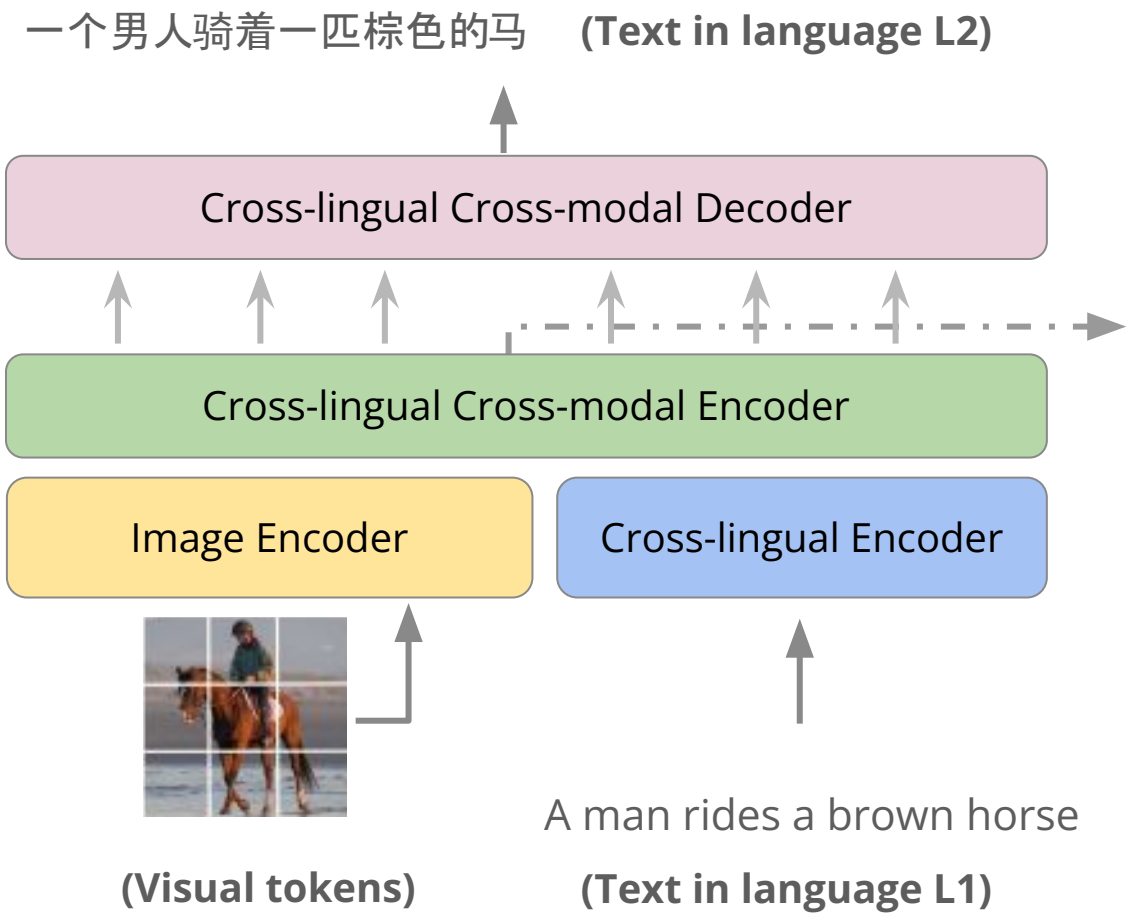
[MURAL: Multimodal, Multitask Retrieval Across Languages](#) (Jain et. al, 2021)

[Cross-Lingual and Multilingual CLIP](#) (Carlsson et al., LREC 2022)

Pre-training data for Multi-X understanding models

Data Resource Type →	Multilingual (text)	Multimodal (English-only)	Multi-X (translated)	Multi-X
Datasets	Wikipedia (100L); Wikimatrix (para; 20L); EOBT (wikimatrix+paracrawl+ europarl);	CC3M (en; 3.3M); CC12M (en)	CC3M (5L; 3.3M); MSCOCO; GCC; VizWiz	WIT (101L), Alt-Text
Models	M3P (Wikipedia)	M3P (CC3M)		
			UC2 (CC3M-translated)	
	CCLM (Wikimatrix)		CCLM (CC3M-trans)	
	MURAL (EOBT)	MURAL (CC12M)		MURAL (Alt-Text)
			mCLIP (COCO+GCC+ VizWiz)	

Category 3: Multi-X Understanding + Generation (Text)



Multilingual Image-text Retrieval / VQA / NLI

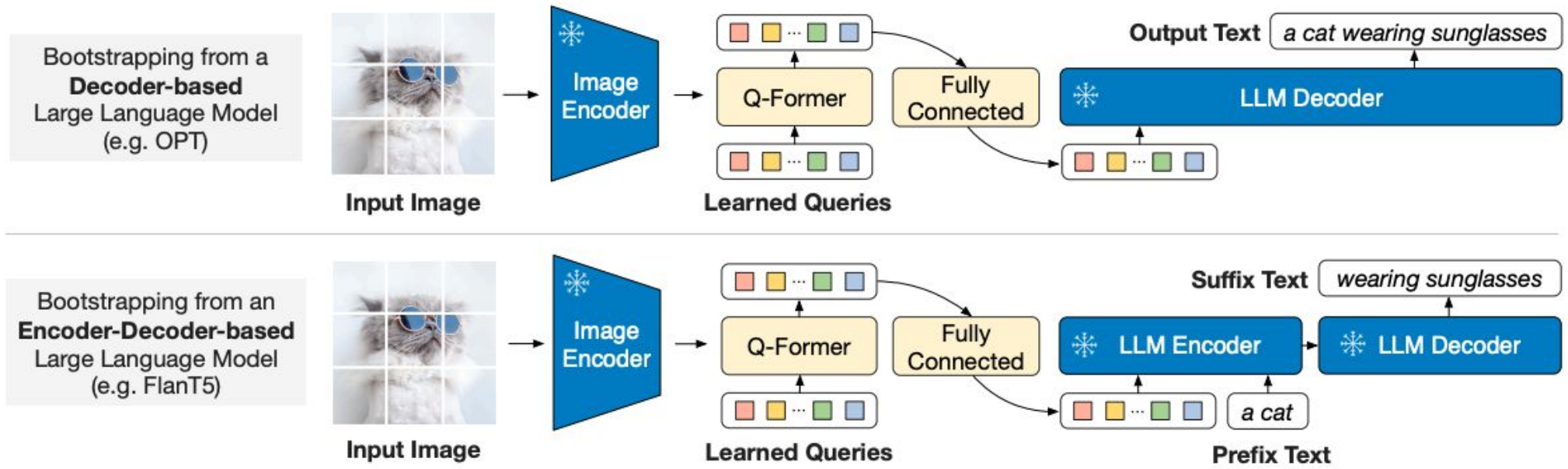
PALI-X: On Scaling up a Multilingual Vision and Language Model (Chen et al., 2023)

ERNIE-UniX2: A Unified Cross-lingual Cross-modal Framework for Understanding and Generation (Yin et al, 2022)

mBLIP: Efficient Bootstrapping of Multilingual Vision-LLMs (Geigle et al., 2023)

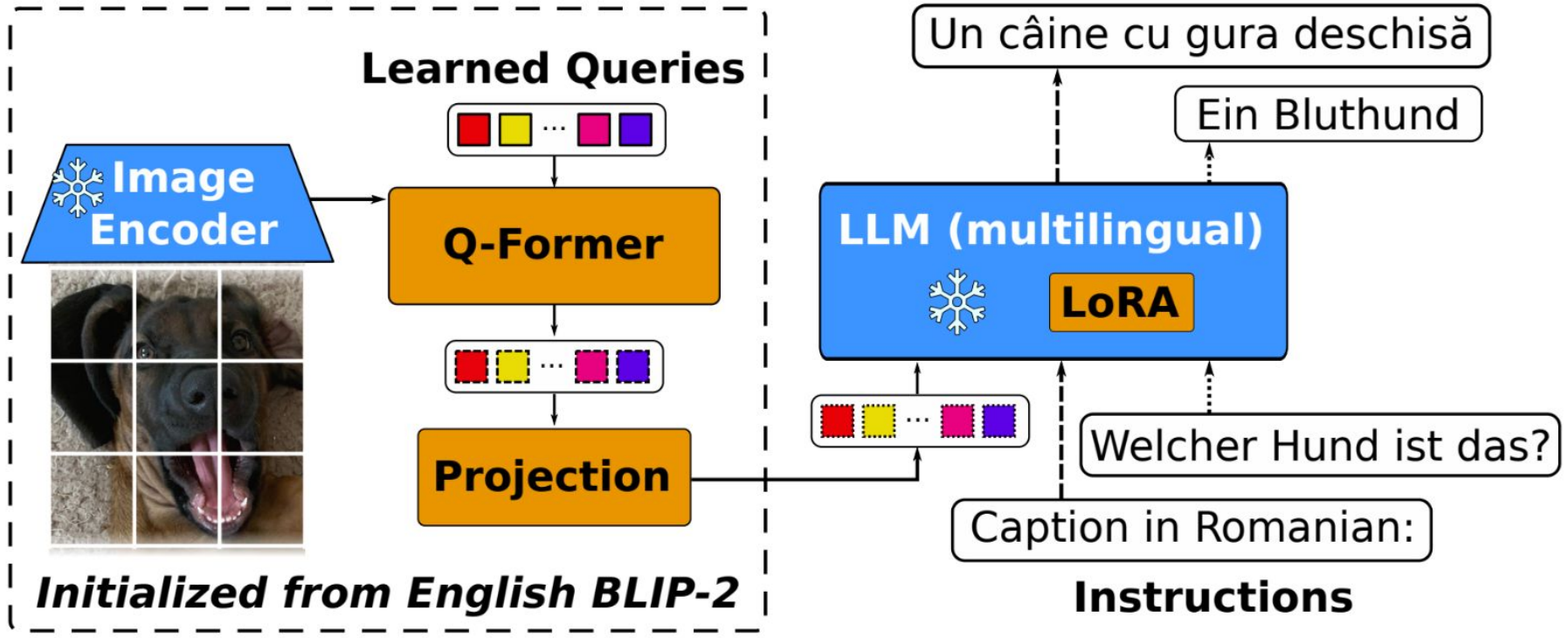
Deep Dive: mBLIP (Encoder-Decoder)

Background: BLIP-2 Architecture



[BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models](#) (Li et al., CVPR 2023)

Deep Dive: mBLIP (Encoder-Decoder)



mBLIP: Efficient Bootstrapping of Multilingual Vision-LLMs (Geigle et al., 2023)

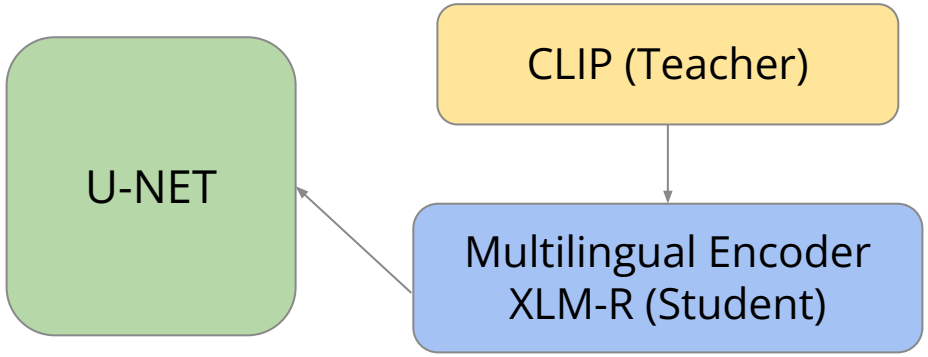
Deep Dive: mBLIP (Demo)

https://4c2e12edc5350ddb9f.gradio.live

Pre-training data for Multi-X understanding models

Data Resource Type →	Multilingual (text)	Multimodal (English-only)	Multi-X (translated)	Multi-X
Datasets	Wikipedia (100L); Wikimatrix (para; 20L); EOBT (wikimatrix+paracrawl+ europarl);	CC3M (en; 3.3M); CC12M (en)	CC3M (5L; 3.3M); MSCOCO; GCC; VizWiz	WIT (101L), Alt-Text
Models	M3P (Wikipedia)	M3P (CC3M)		
			UC2 (CC3M-translated)	
	CCLM (Wikimatrix)		CCLM (CC3M-trans)	
	MURAL (EOBT)	MURAL (CC12M)		MURAL (Alt-Text)
			mCLIP (COCO+GCC+ VizWiz)	
			mBLIP (MSCOCO+WebCapFilt+ LLaVa (IFT))	

Category 4: Multi-X Generation (Image): diffusion-based models



ERNIE-ViLG 2.0: Improving Text-to-Image Diffusion Model with Knowledge-Enhanced Mixture-of-Denoising-Experts (Feng et al, CVPR 2023)

AltDiffusion: A Multilingual Text-to-Image Diffusion Model (Ye et al, 2023)

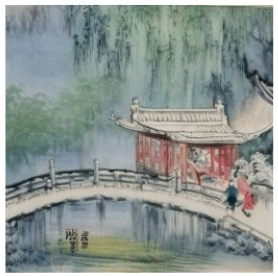
Hindi
पहाड़ी महिलाएँ, अमृता शेरगिल
(Pahari Women, Amrita Sher-Gil)

Spanish
Retrato de mujer, estilo Picasso
(Portrait of a woman, Picasso style)

Italian
Pittura ad acquerello di spaghetti
(Watercolor painting of spaghetti)

Thai
ชีวิตชนบทไทย อรรคพล เมฆพจน์
(Thai Countryside Life, Akaphol Mekpoth)

English
The Hay Wain, John Constable



Chinsse
小桥流水人家
(Small bridge,flowing water and household)

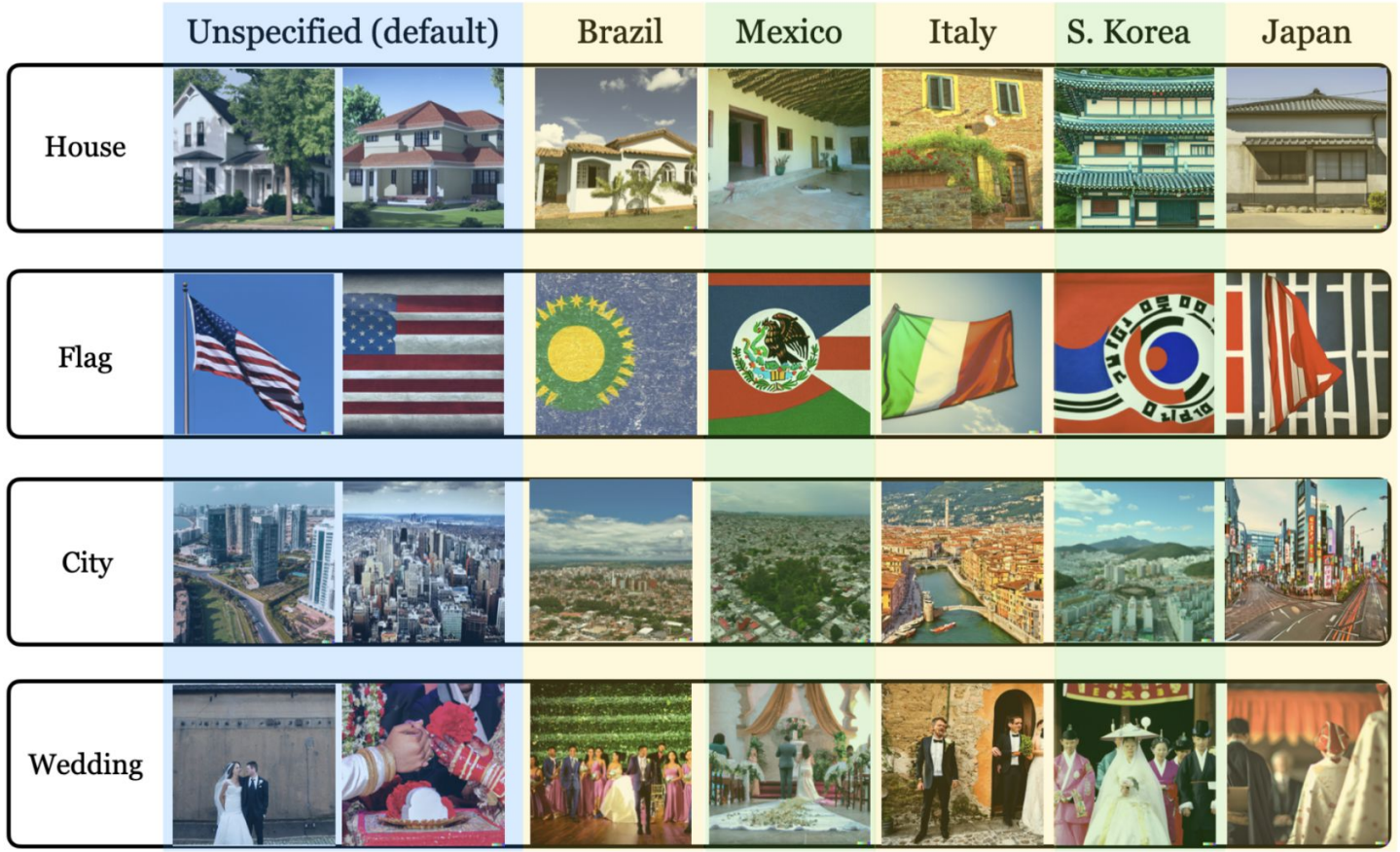
Russian
Від на Байкал, карціна алеем
(View of Baikal, oil painting)

AltDiffusion (Outputs)

Figure 14: Images generated by AltDiffusion with prompt "boy portrait with sunglasses" in various languages and a fixed seed. Note that the model demonstrates proficiency in capturing distinct facial features of young males from various cultural backgrounds, including a European-American style for English and an Asian style for Chinese.

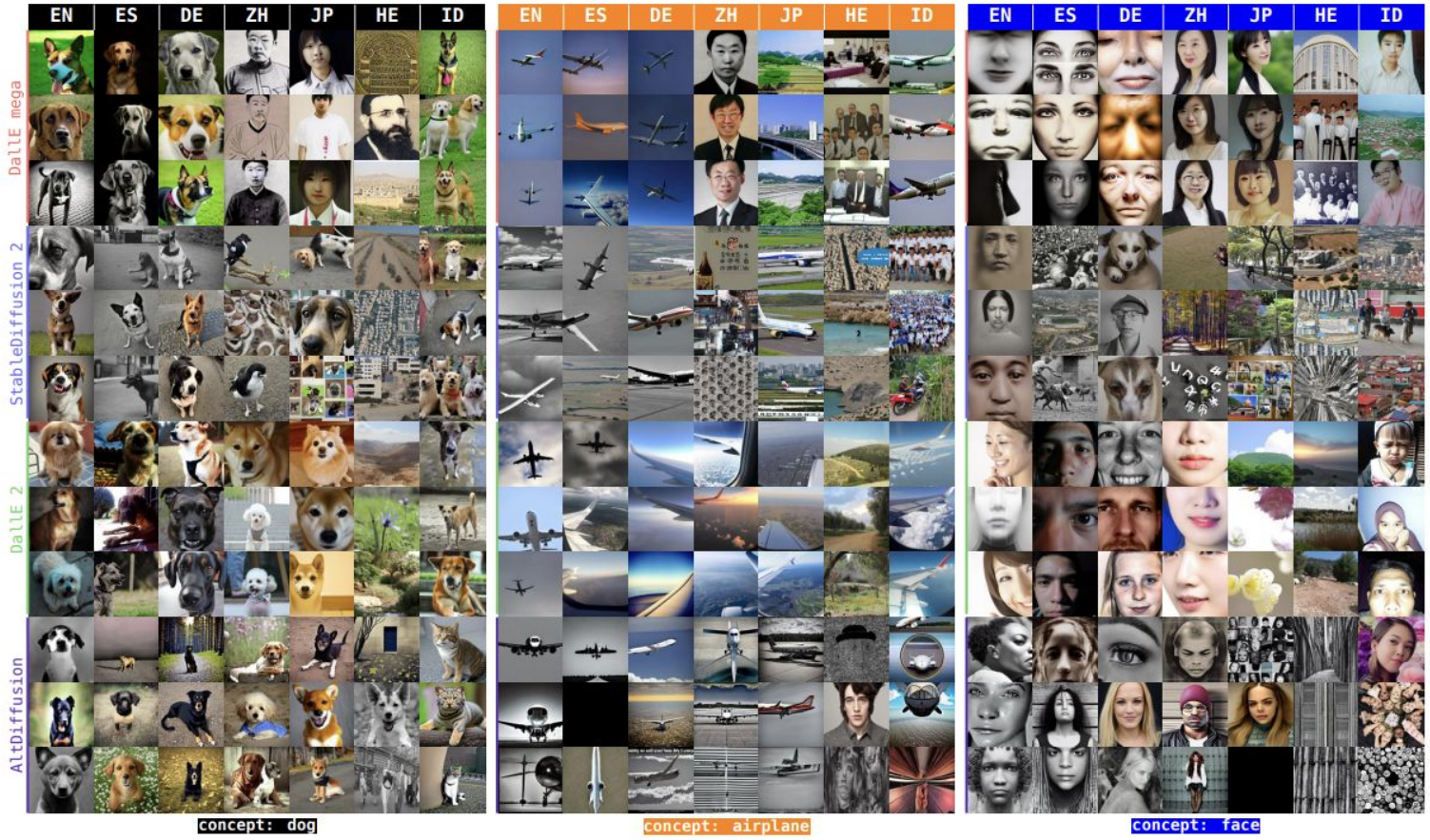
AltDiffusion: A Multilingual Text-to-Image Diffusion Model (Ye et al, 2023)

Biases in image generation models



[Inspecting the Geographical Representativeness of Images from Text-to-Image Models](#) (Basu et. al, ACL 2023)

Biases in image generation models



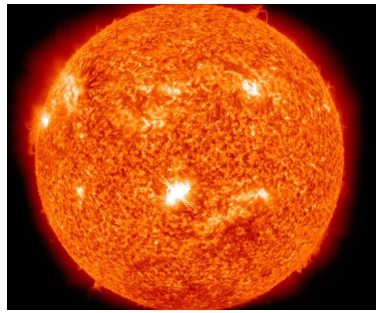
[Multilingual Conceptual Coverage in Text-to-Image Models](#) (Saxon et. al, ACL 2023)

Let's revisit our motivation in context of the prior work

Do models leverage the universality of concepts to learn better representations for multiple languages in text?

a. [Open] Hard to define *universality*
   i. [Proposal] Study similarities and differences in naturally occurring data from different countries/cultures?
b. [Open] How do we measure cross-lingual alignment in an embedding space?
c. [Open] How can we piece out contributions of the textual and visual modality in learning a cross-lingual representation?
d. [Open] Hard to evaluate with high confidence

sun *太陽*



सूरज *güneş*

Let's revisit our motivation in context of the prior work

Do models incorporate diversity in representation?

a. Initial explorations suggest otherwise
b. [Open] How do you evaluate diversity?
    i. Can you account for individual preferences?
    ii. What is the tradeoff between diversity v/s stereotyping/bias?
c. [Open] Is it right to discern culture based on language input?
    i. English is ubiquitous
    BUT also
    ii. Language has evolved within a culture and holds key information about it

Let's revisit our motivation in context of the prior work

Do models have a world view of concepts specific to every culture?

a. Probably not and may never will
   i. Not everything is present digitally
   ii. Cultures and concepts are constantly changing

b. [Open] How can we make models adept at keeping up with evolving concepts and cultures?

c. [Open] How can we incorporate cultures of communities that are not present digitally, into our models?

Sanxian / Shamisen

musical instrument
strings
action
pose

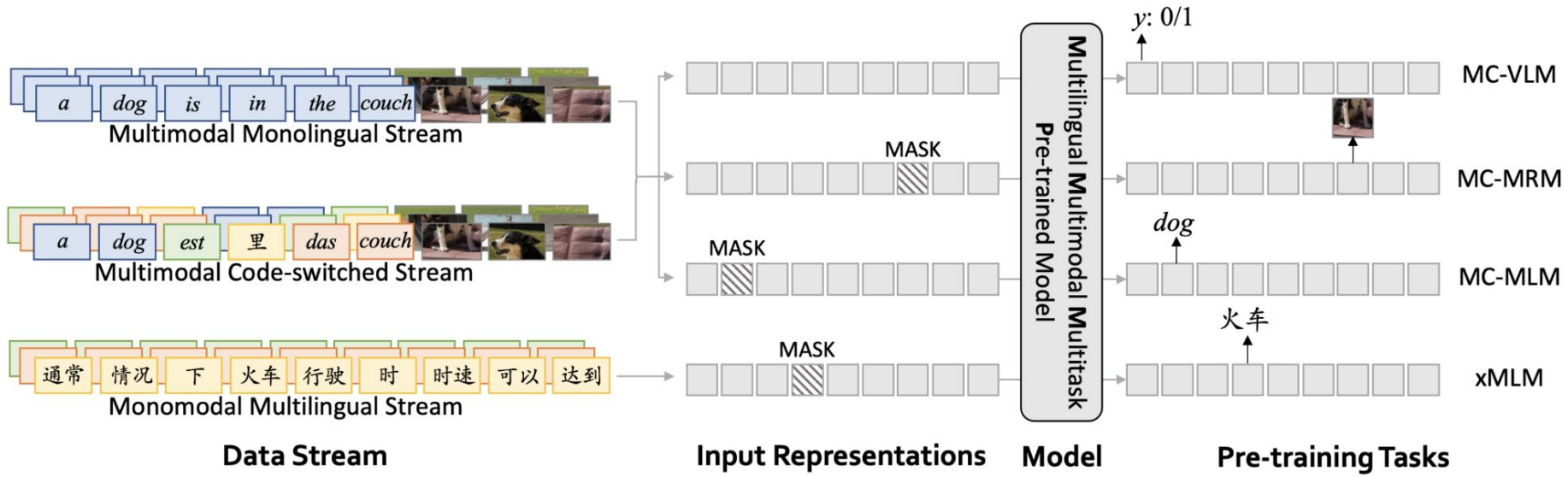
Thanks!

Questions?

References

- mBLIP: Efficient Bootstrapping of Multilingual Vision-LLMs (Geigle et. al)
- Altdiffusion
- michael saxon's work
- mclip
- santy's cultural translation paper
- M3P
- UC2
- CCLM
- PALI
- ERNIE-UniX2
- IGLUE
- XM3600
- LMCap
- xFlickrCo
- WIT

Encoder-based



Pre-training Data:

Evaluation Data:

M3P: Learning Universal Representations via Multitask Multilingual Multimodal Pre-training (Ni et. al, CVPR 2021)

The potential of visual modeling in multilingual NLP

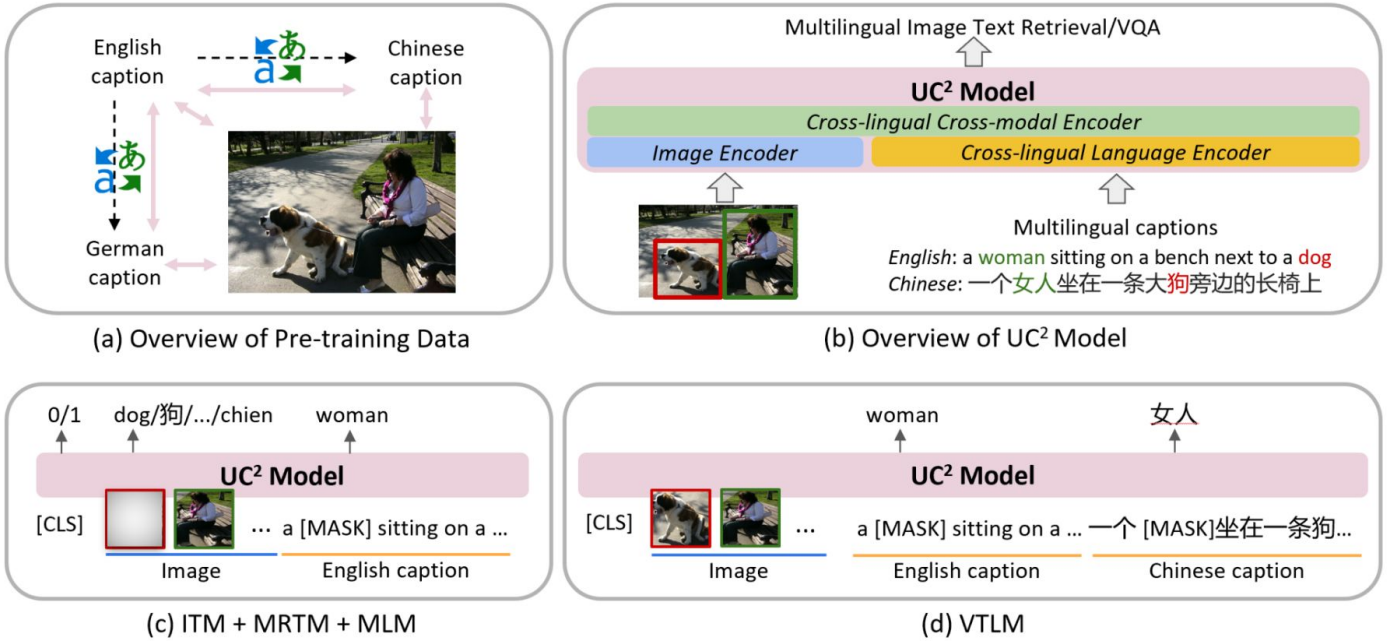
- Grounding (for universal concepts)
- Grounding + diversity
- diversity (clavie example, clavie might appear in the context of festivals but a picture might help associate it with fire, night, people etc...)

Evolution of data (with models)

Multilingual (text-only)	Multimodal (English-only)	Multi-X (translated)	Multi-X
Wikipedia (mono; 100L; 101G)	CC3M (en; 3.3M)	CC3M (en,de,fr,cs,jp,zh; 3.3M)	XM3600
Wikimatrix (parallel; 20L; 19M)	CC12M	Web CapFilt	WIT
	MSCOCO	LLaVa-IFT Data	LAION (5B

- M3P: Wikipedia (mono) + CC3M (English)
- UC2: CC3M (translated)
- CCLM: CC3M (translated) + Wikimatrix
- MURAL:
- mCLIP:
- AltDiffusion: CC3M (translated captions) + LAION
- mBLIP: Web CapFilt + LLaVa-IFT Data

Encoder-based



Pre-training Data:

Evaluation Data:

UC2 : Universal Cross-lingual Cross-modal Vision-and-Language Pre-training (Zhou et. al, CVPR 2021)

Multi-X Understanding (Dual-encoder Based Models)

Image Encoder

Cross-lingual Encoder

Multi-X Retrieval (image ⟷ text)

MURAL
mCLIP