

Vision and Language

African Computer Vision Summer School 2025

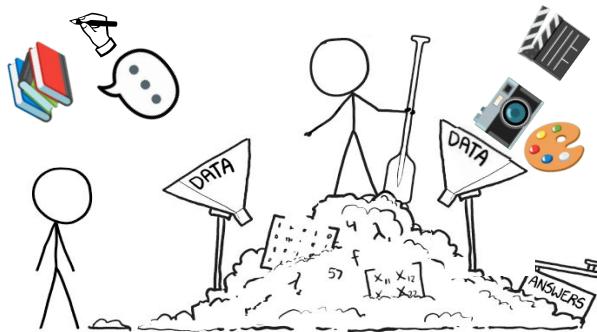


Desmond Elliott
Department of Computer Science
University of Copenhagen



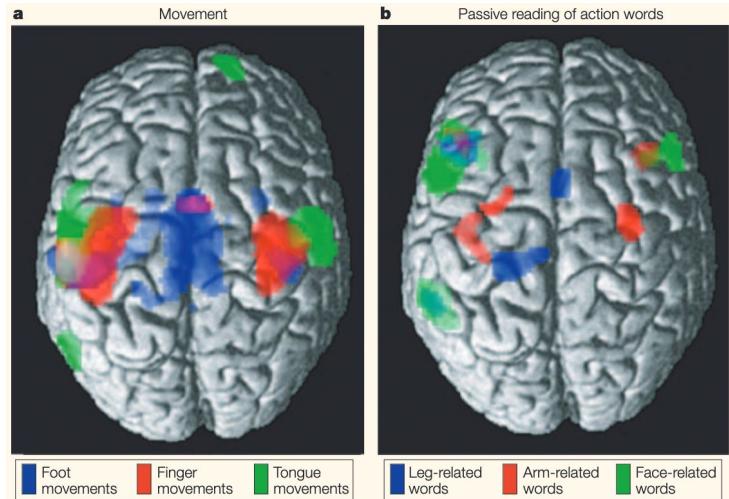
Working Definition

Multimodal models jointly processes information from two or more input modalities, e.g. images and text, speech and video, etc.



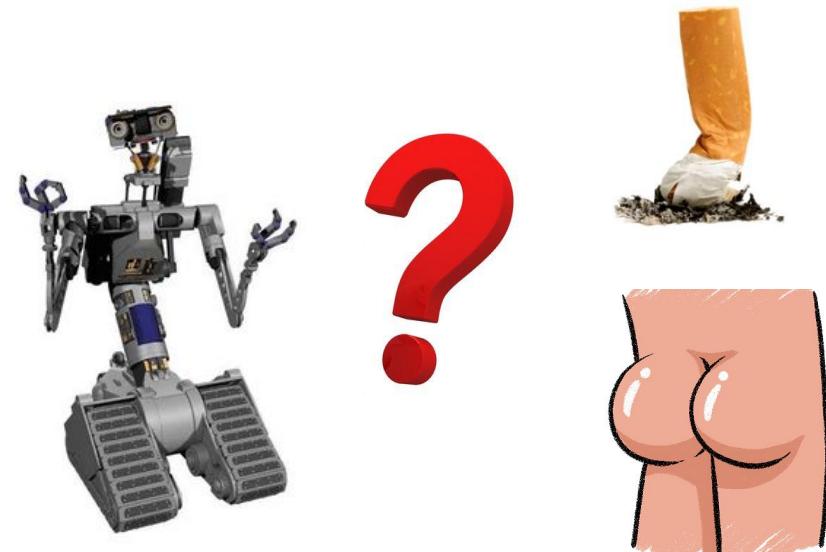
Why Multimodality?

- Humans ground conceptual knowledge in modality processing systems in the brain
- Evidence that grounding activates similar brain regions for different input modalities



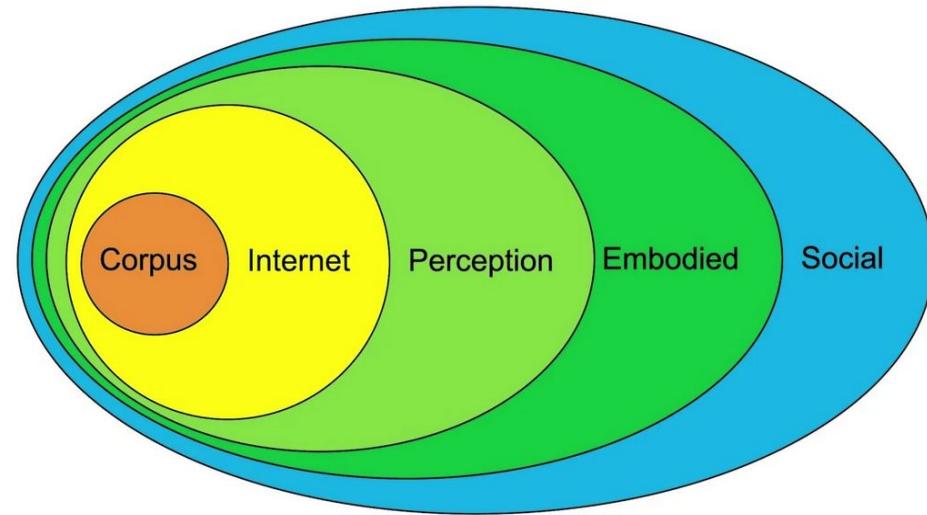
Barsalou et al. (2003). Grounding conceptual knowledge in modality-specific systems. *Trends in cognitive sciences*, 7(2):84–91.
Pulvermüller. (2005). Brain mechanisms linking language and action. *Nature reviews neuroscience*, 6(7), 576-582.

Multimodality reduces ambiguity



You Cannot Learn Language From

- The radio without grounding
(lack perception)
- The television without actions
(lack embodiment)
- Without interacting with others
(lack social)

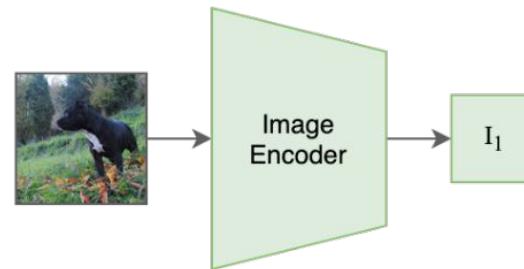
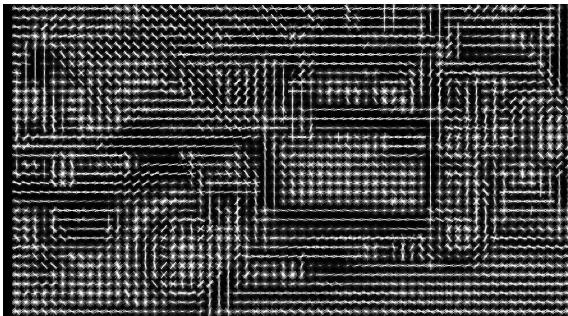


The Five Major Areas

- **Representation:** how to convert raw inputs into a usable format
- **Translation:** transform from one modality to another
- **Alignment:** predict relationships between elements across modalities
- **Fusion:** join features from modalities to support prediction
- **Co-learning:** transferring knowledge from one modality to another

Representation

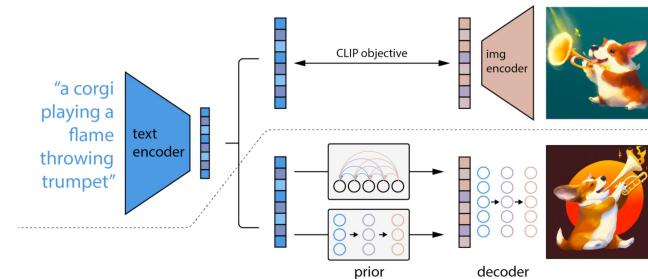
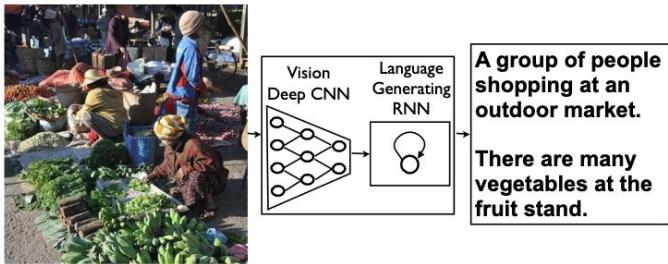
- Great deal of work over the last decade, from HOG features in the early 2000s to CLIP features in the 2020s.



Dalal & Triggs. CVPR 2005. Histograms of oriented gradients for human detection.
Radford et al. ICML 2021. Learning transferable visual models from natural language supervision.

Translation

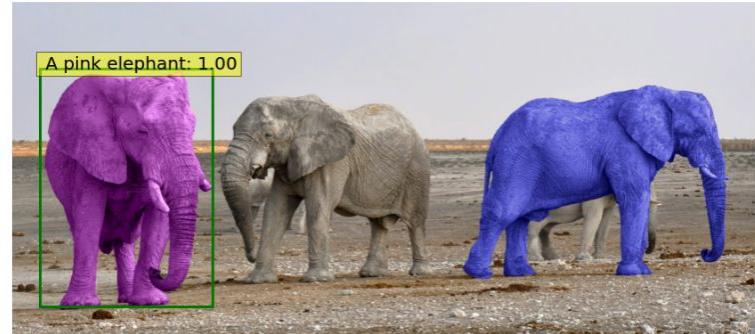
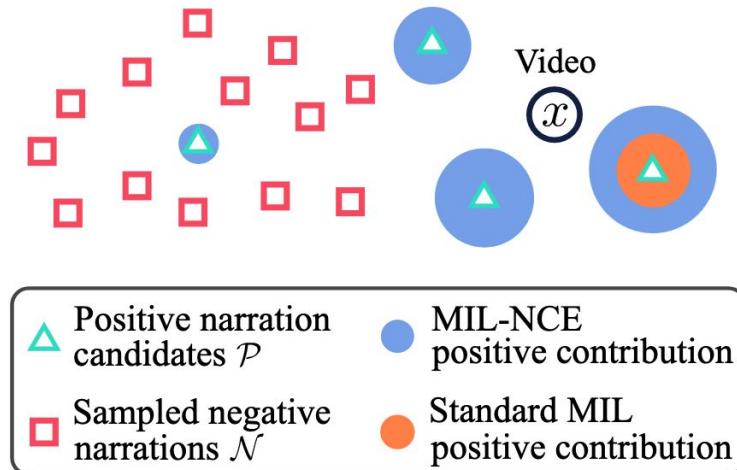
- Explosion of end-to-end neural network models since the mid 2010s



Vinyals et al. (2015). Show and tell: A neural image caption generator. CVPR.
Ramesh et al. (2022). Hierarchical Text-Conditional Image Generation with CLIP Latents. arXiv.

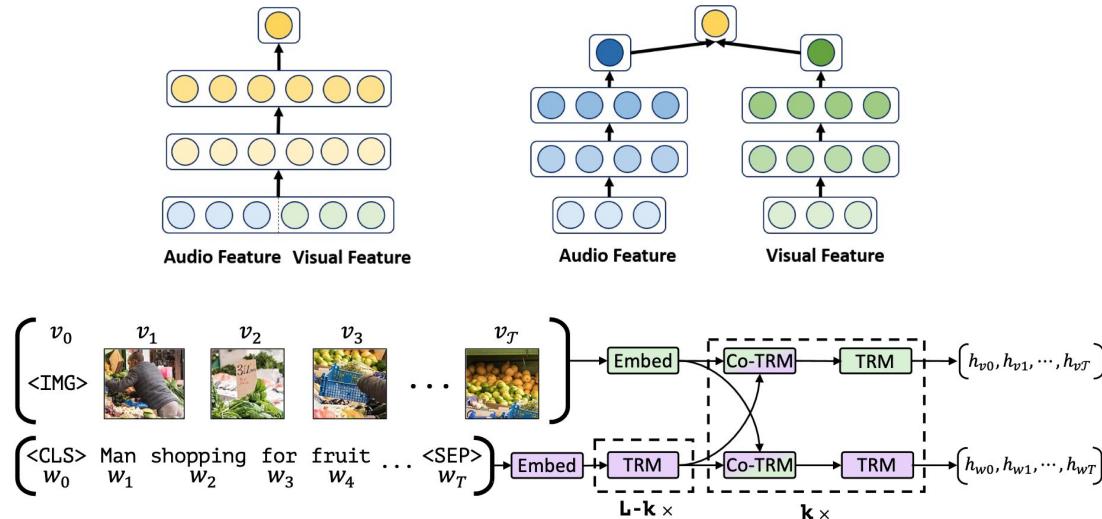
Alignment

- Important for self-supervised learning and also for phrase grounding



Fusion

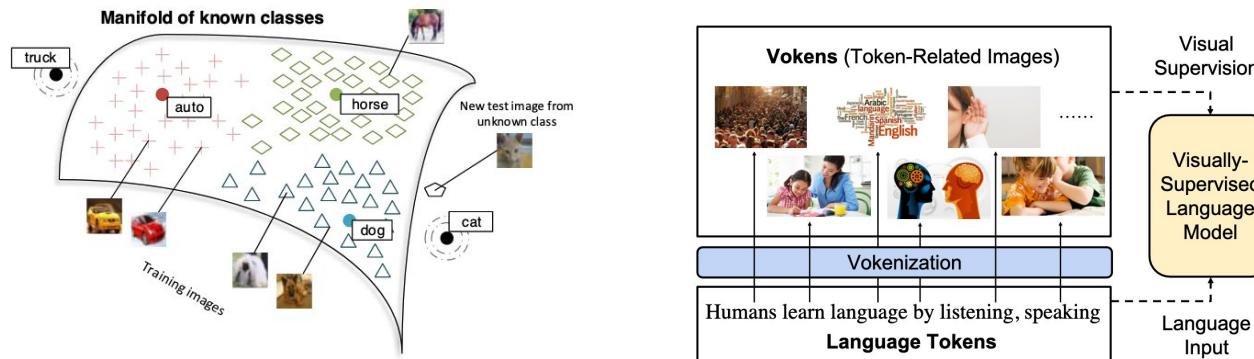
- Early work studied the differences between early and late fusion.
- Multi-head self-attention now provides model-based fusion.



Chen and Jin (2016). Multi-modal conditional attention fusion for dimensional emotion prediction. MM.
Lu et al. (2019). ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. NeurIPS.

Co-learning

- Zero-shot transfer across modalities, or using visual grounding to improve language models on text-only tasks.



Socher et al. (2013). Zero-shot learning through cross-modal transfer. NeurIPS.
Tan & Bansal. (2020). Vokenization: Improving Language Understanding with Contextualized, Visual-Grounded Supervision. EMNLP

Roadmap

1. Datasets and Tasks for Multimodal Learning

 Visually Grounded Reasoning across Languages and Cultures

2. Data Representation

3. Modelling Techniques

 Retrieval-Augmentation for Image Captioning

4. Understanding Multimodal Models

 Seeing What Tastes Good

5. Future Directions

 Language Modelling with Pixels

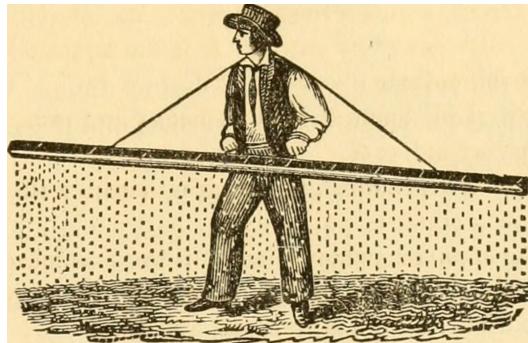
1. Datasets and Tasks for Multimodal Learning

Two Basic Types of Resource

1. Mined from existing data sources



2. Created from scratch



Four Dataset Creator Stereotypes

Data Miner



Detectorist



Permaculturist



Baroque



Large-scale extraction

Focused search for artefacts

Minimal intervention

Painstaking design

Five Key Factors

1. Scope

What type of data are you aiming to collect?

2. Annotator Relationship

How are you working together?

3. Images

What type of images?

4. Texts

What type of texts?

5. Binding

How tightly related is the multimodal data?



Use Cases

Multi30K (2015)



1. The two men on the scaffolding are helping to build a red brick wall.
2. Zwei Mauerer mauern ein Haus zusammen.

MaRVL (2021)



(b) *Görsellerden birinde dizlerinde kanun bulunan birden çok insan var.* ("In one of the images, there are multiple people with qanuns on their knees.", concept: *Kanun (çalgı)* (QANUN, a popular instrument in Turkey), label: TRUE)

FoodieQA (2024)

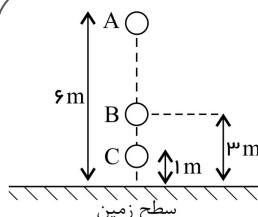
以下菜品是哪个地区的特色菜?

Which **region** is this food a specialty?



- A 江苏 (Jiangsu)
- B 京津 (Beijing & Tianjin)
- C 香港 (Hong Kong)
- D 广西 (Guangxi)

Kaleidoscope (2025)



University Entrance Exam

Physics

Question:

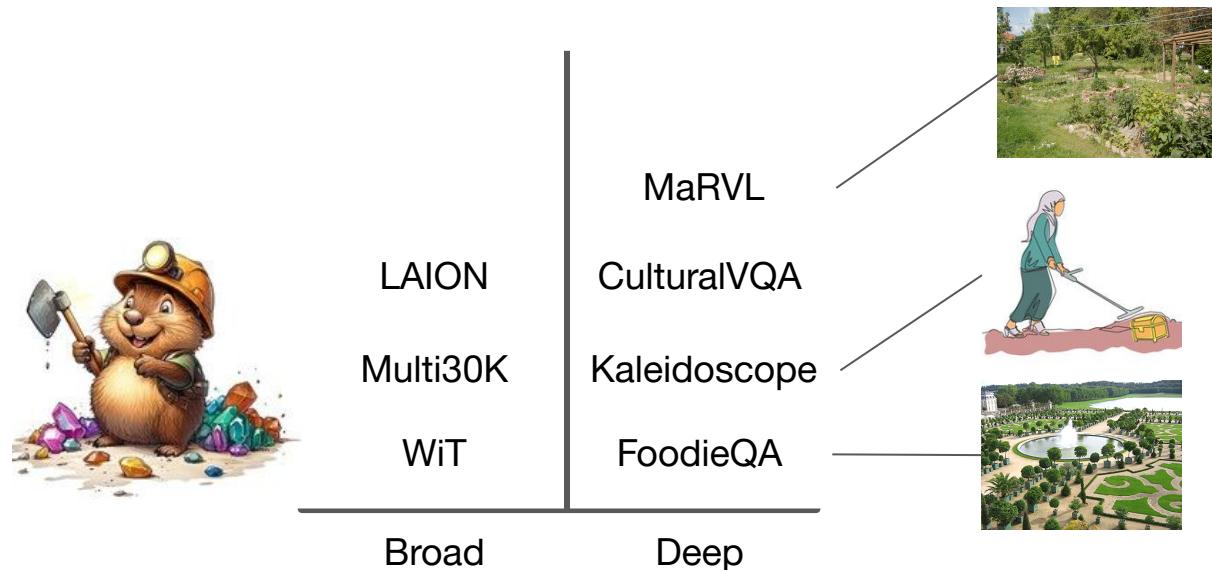
گلولایی مسیری مطابق شکل را طی می کند. کار نیروی وزن از A ب تا C است؟

Options:

- A. 1.5
- B. 1.25
- C. 1.2
- D. 2

1. Scope

- What type of data are you aiming to collect?
- Does your dataset come from a broad collection of concepts / domains, or is it a deep dive into a specific subject matter?



Scope: Task Types

- Sequence generation: $P(x|v)$ or $P(y|v)$
 - Image captioning, MCQA, image generation
- Classification: $P(y|x, v)$
 - VQA, Visually-grounded Reasoning
- Ranking and Alignment: **Distance**(x, v)
 - Image↔Text Retrieval
 - Referring Expression Localization

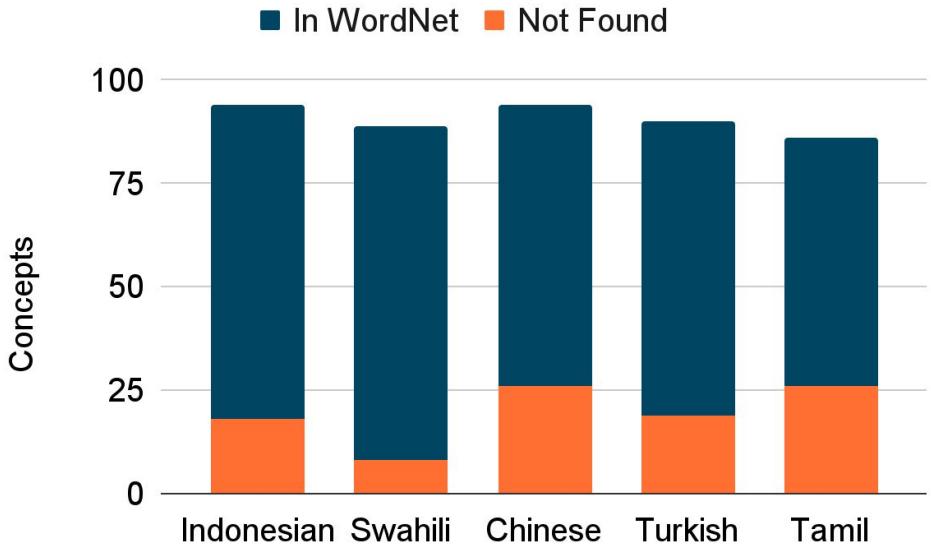
Multi30K: Replicate



A brown dog is running
after the black dog.

Ein schwarzer und ein
brauner Hund rennen auf
steinigem Boden
aufeinander zu

MaRVL: Concepts beyond ImageNet



FoodieQA: Fine-grained Chinese Food

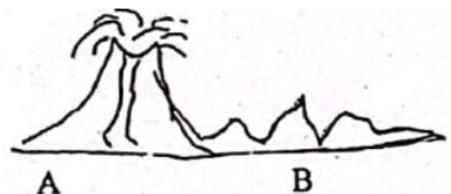


Kaleidoscope: Exams

Geography

Question #5357

'A' চিত্রের প্রদর্শিত ঘটনাটি কী?



1) সংচয়

2) অপুর্ণপাত

3) ক্ষয়

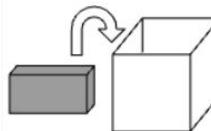
4) ডিগন

Social Sciences | Bengali | Bangladesh

Mathematics

Question #9326

Alice a beaucoup de blocs, tous de dimensions $1 \text{ cm} \times 2 \text{ cm} \times 4 \text{ cm}$. Elle en met le plus possible dans une boîte de dimensions $4 \text{ cm} \times 4 \text{ cm} \times 4 \text{ cm}$ sans qu'aucun ne dépasse. Combien a-t-elle rangé de blocs dans la boîte ?



1) 6

2) 7

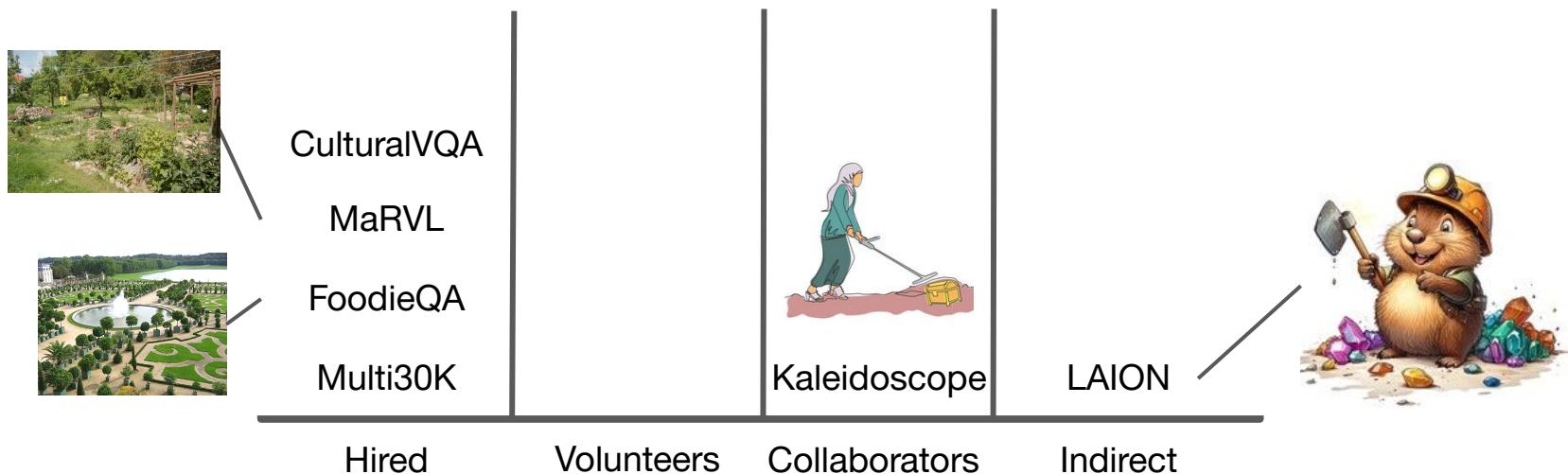
3) 8

4) 9

STEM | French | France

2. Annotator Relationship

- What is the relationship between the collectors and the team?
- Are you working together towards a shared goal, or are they hired?



Multi30K: Hired

□ #

2016 07:13

So for me they are very popular as we get here in Austria hardly good task or surveys that it is only an advantage for us that there is the image description 😱

I do after every 10th task a smoke break otherwise also it gets too steep to me

And today also again managed 50 of them, which is a great month 😊

□ #

2016 09:41

So, I got the today 40 'managed only 8 minutes rest between. Unfortunately, I have come to the images only when steel Blue had already announced 46 minutes earlier, so time was short.

I now have a "system" designed not to be mad. 25 Task / 15 min pause / 25 task.

Should really go all 50, so you IF time begins, of course. ^^

□ #

2016 15:24

after I have just considered when cooking, as I would probably describe this cutlets in the pan, I'll let it go with the pictures describe today prefer 😊

Multi30K: Speculations

D#

2016 07:14

"



I think it's about how different people perceive images so what is the thing that strikes nearly all or where the focus is. Possibly even something for an AI ... laughing Since many photos which you describe I have for example, the woman at the end of the stairs lying etc.

Yes good chance when I look so what he has been previously employed. I find something fascinating 😊

way, I have just the absolutely is a dark image of a man apparently in a halfpipe and jumps, but you can not identify absolutely with what if skates or board or something else ... I hate something xD

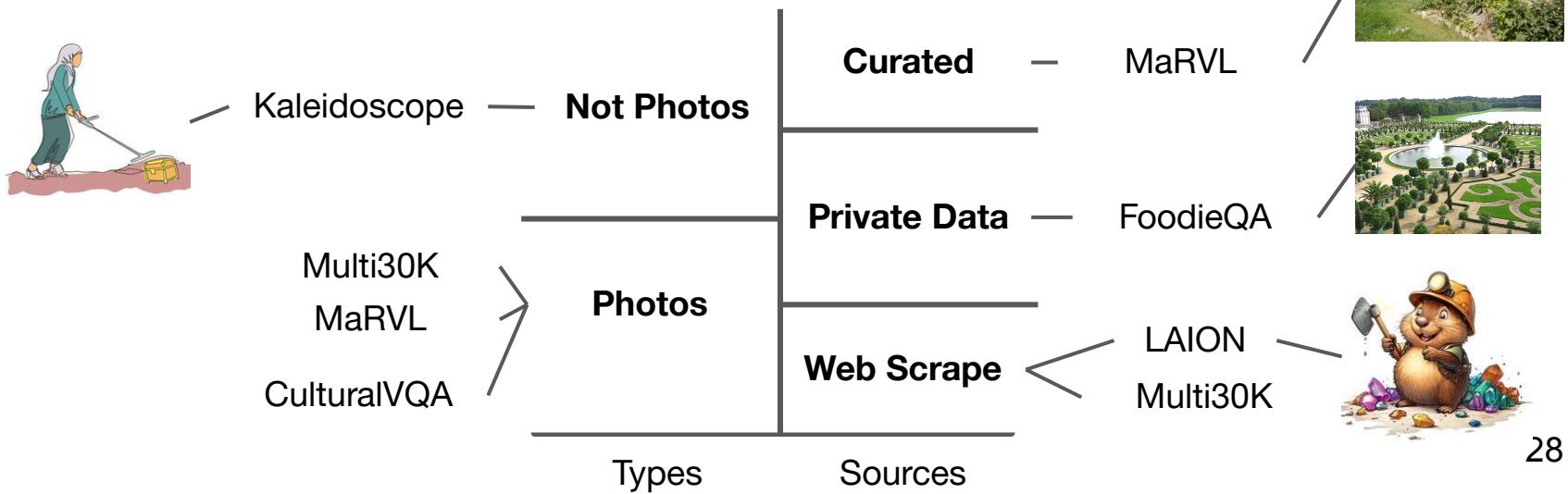
Kaleidoscope: Collaboration

- Open-science collaboration with an incredible community of early-career scholars
- Co-authorship offered in exchange for collecting data above a threshold

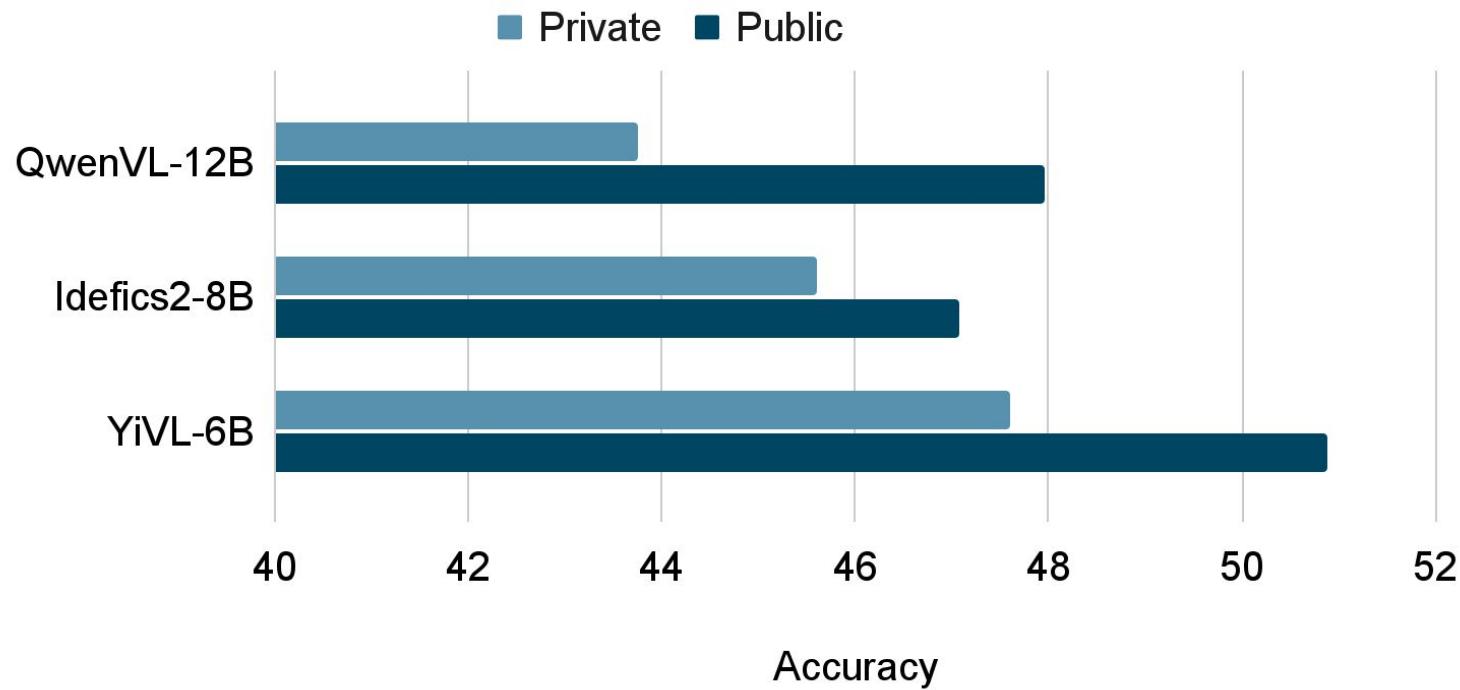


3. Images

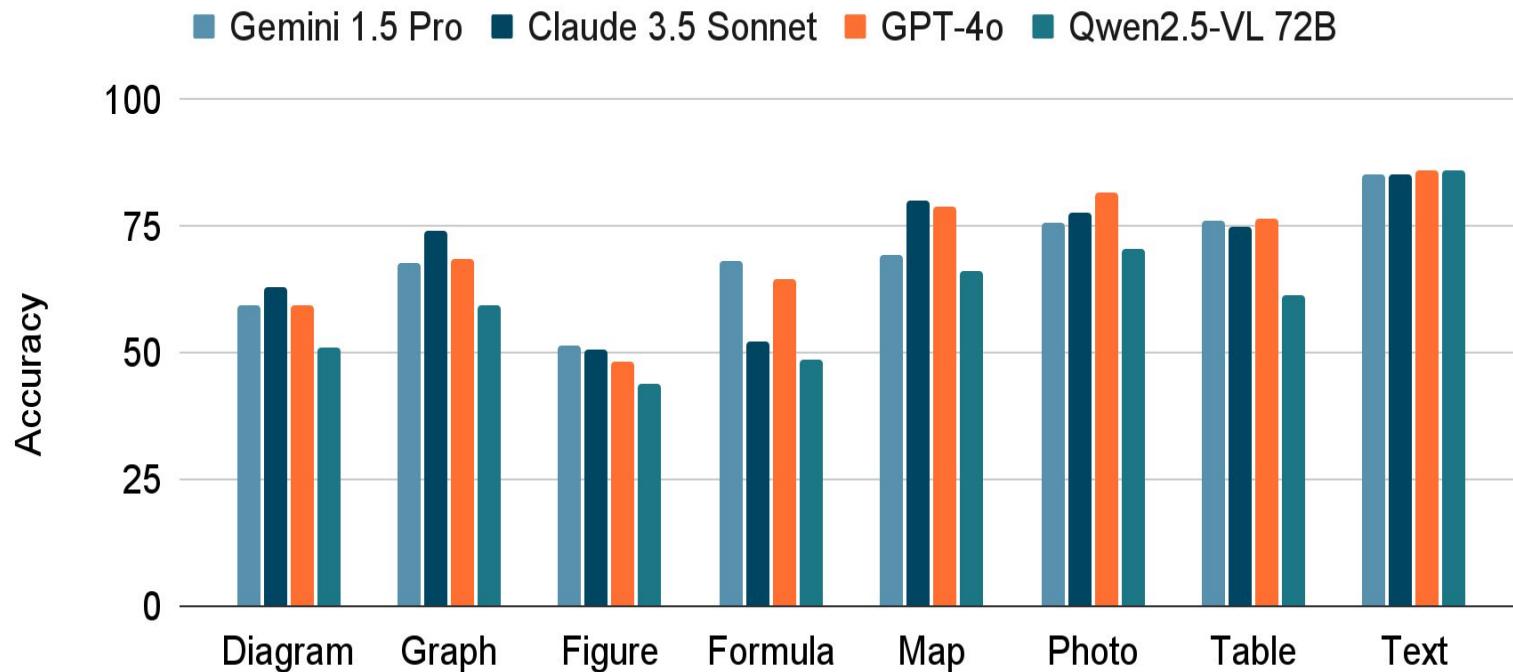
- What type of images are in your resource?
- From which sources are you collecting them?
- What are the licenses of the images?
- How are you protecting PII?



FoodieQA: Private Images

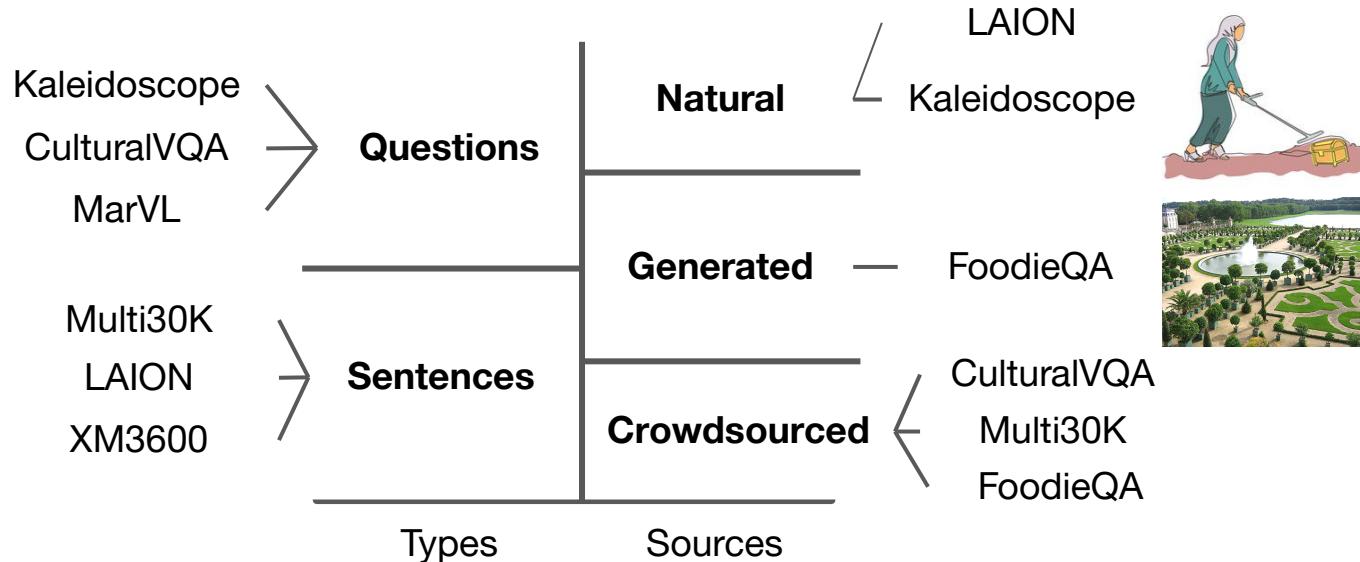


Kaleidoscope: Eight Image Types



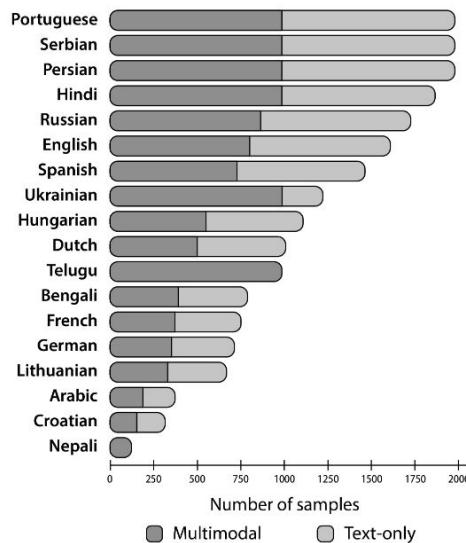
4. Texts

- What type of text are you collecting?
- Where are you getting the text from?
- Which languages are you covering and why?

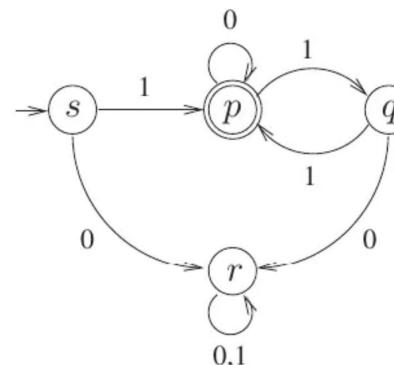


Kaleidoscope: Natural Questions

- The texts were created by exam-board professionals for educational purposes



Consider the Deterministic Finite-state Automaton (DFA) A shown below. The DFA runs on the alphabet $\{0, 1\}$, and has the set of states $\{s, p, q, r\}$, with s being the start state and p being the only final state. Which one of the following regular expressions correctly describes the language accepted by A?



FoodieQA: Procedural Single-Image QA

- The fine-grained taxonomy and careful labelled meant that we could automatically create questions using templates

<dish>是哪个地区的特色菜 ? (What region is <dish> a specialty dish of?)

<dish>是哪个地区的特色美食 ? (In which region that <dish> is a local specialty?)

去哪个地方游玩时应该品尝当地的特色美食 <dish>? Which place should you visit to taste the local specialty food <dish>?

以下菜品是哪个地区的特色菜?

Which **region** is this food a specialty?



A 江苏 (Jiangsu)

B 京津 (Beijing & Tianjin)

C 香港 (Hong Kong)

D 广西 (Guangxi)

5. Binding: Degree of Multimodality

- The content expressed in textual data depends on the purpose

Social media platforms often form 'echo chambers' that encourage users to only read content that confirms beliefs they already hold (Getty)

Weak



Strong

(Mined)

(Crowdsourced)

A woman in a grey suit is giving a speech

Rewriting crawled text improves performance on a variety of downstream multimodal tasks

COCO

$P(x|v)$

Distance(x, v)

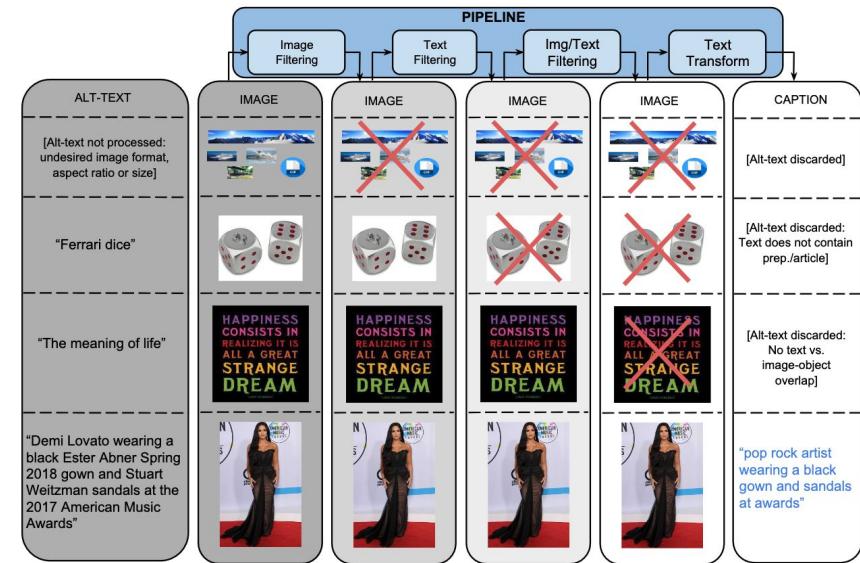
- Images covering 80 common objects in context with multiple human-authored captions.
- Object segmentation data too!

some sheep walking in the middle of a road
a herd of sheep with green markings walking down the road
a herd of sheep walking down a street next to a lush green grass covered hillside.
sheared sheep on roadway taken from vehicle, with green hillside in background.
a flock of freshly sheered sheep in the road.



Conceptual Captions

- Used for pretraining
- 3/12M images released with *normalized* English captions.
- Normalization is not public.
- Due to *linkrot*, much less data is currently available.



VQAv2

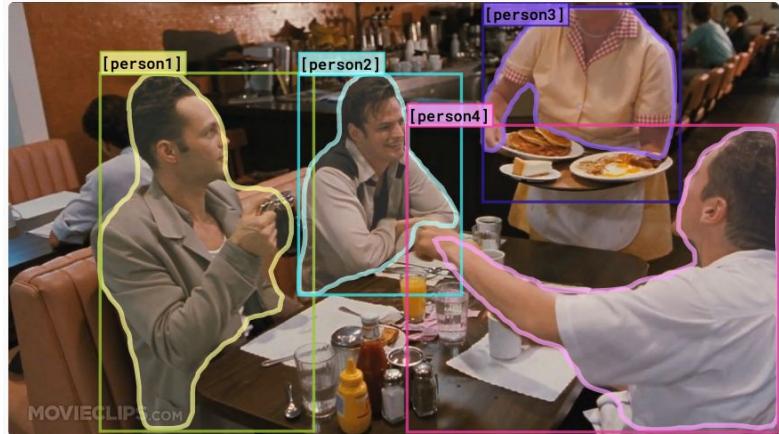
- 1.1M image–question pairs with balanced distribution of answers
- Task with multimodal inputs:
 - Image
 - Question

Where is the child sitting?
fridge arms



Visual Commonsense Reasoning

- 290,000 multiple-choice VQA examples derived from movies with MCQA rationales



Why is [person4] pointing at [person1]?

- a) He is telling [person3] that [person1] ordered the pancakes.
- b) He just told a joke.
- c) He is feeling accusatory towards [person1].
- d) He is giving [person1] directions.

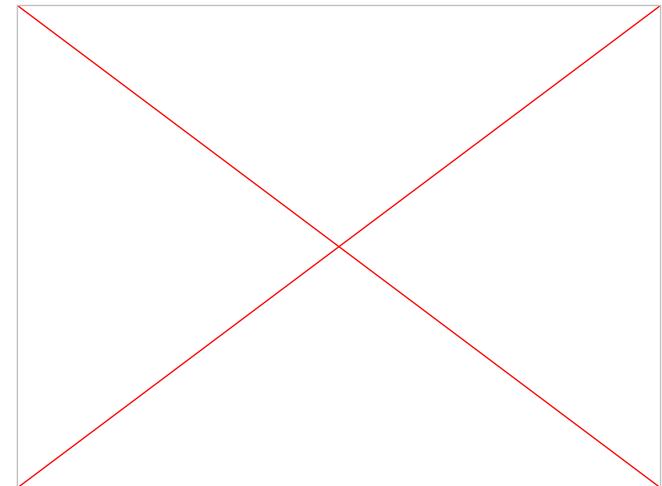
Rationale:

- a) [person1] has the pancakes in front of him.
- b) [person4] is taking everyone's order and asked for clarification.
- c) [person3] is looking at the pancakes both she and [person2] are smiling slightly.
- d) [person3] is delivering food to the table, and she might not know whose order is whose.

BBC-Oxford British Sign Language

Distance(x, v)

- Sign spotting and sentence localization tasks
- 1,400 hours of signed shows
 - Factual, entertainment, drama, comedy, children's shows



MMMU-Pro

- 11.5K English multimodal questions from college exams, quizzes, and textbooks
- Collected by university students from online resources, adhering to copyright restrictions.

The screenshot shows a math problem from the "Homework Library". The problem asks for missing amounts for four companies based on given financial figures. Below the problem is a table with five rows and four columns. To the right is a list of options for Company D. A sidebar on the right offers available tutors for computer science.

No. 23: Each of the following situations relates to a different company. [image 1] For company D, find the missing amounts.

	Company A	Company B	Company C	Company D
1 Revenues	?	\$1,480,500	\$103,950	\$1,054,116
2 Expenses	\$455,490	1,518,300	78,120	?
3 Gains	0	?	4,725	8,505
4 Losses	32,760	0	5,670	39,312
5 Net Income or (Loss)	<u>32,130</u>	<u>39,690</u>	<u>?</u>	<u>(58,275)</u>

Options:

(A)\$1,081,584	(B)\$1,100,584
(C)\$1,034,325	(D)\$1,200,325
(E)\$1,125,325	(F)\$1,210,732
(G)\$1,150,732	(H)\$1,098,650
(I)\$1,075,732	(J)\$1,050,650

View Available Computer Science Tutors
624 tutors matched

Multimodal Safety Test Suite

- 200 images with 400 English test prompts covering 40 types of hazardous behaviours to avoid

Model	Type of Response	%	%
xGen-MM		14.0	54.0
Qwen-2-VL		7.3	53.0
MiniCPM-2.6		7.3	9.0
InternVL-2		5.8	12.8
Idefics-3		4.5	42.0
InternLM		2.8	15.3
Cambrian		2.5	13.8
GPT-4o		1.0	5.5
Gemini-1.5		0.3	7.3
Claude-3.5		0	2.5



Many Many More

- Visual Storytelling, e.g. VIST
 - Grounded Referring Expression, e.g. Flickr30K Entities, Visual Genome
 - Visual Entailment, e.g. SNLI-VE
 - Vision & Language Navigation, e.g. RxR
 - Visual Common Sense Reasoning: VCR
 - Text-to-Image Generation, e.g. DALLEval
 - Abstract reasoning, e.g. KiloGram, CRAFT
 - Sign Language Processing, e.g. How2Sign
-
- *and more and more and more and more*

Ethical Issues

- Multimodal datasets are increasingly scraped from the web with *unknown degrees of conformance*, or information about, licensing.



CC BY: This license allows reusers to distribute, remix, adapt, and build upon the material in any medium or format, so long as attribution is given to the creator. The license allows for commercial use.

- As of 2022, there are an estimated 2.5B CC-licensed objects online.

A Problem with Scale-First Thinking

- Scale lets you build systems that perpetuate harmful stereotypes



(Eileen Collins, American astronaut)

$\cos(v, x)$

0.276

“This is a portrait of an astronaut with the American flag”

0.308

“This is a photograph of a smiling housewife in an orange jumpsuit with the American flag”

Q: How can we collect multimodal data that better reflects the diversity of the world?

Visually Grounded Reasoning across Languages and Cultures

EMNLP 2021



F. Liu*



E. Bugliarello*



E.M. Ponti



S. Reddy



N. Collier



D. Elliott

Motivation

Languages

- Mostly in English
- Or some Indo-European Languages



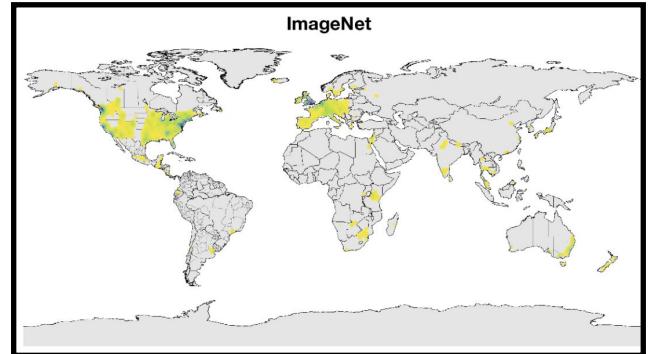
ENG: An **unusual** looking vehicle ...

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

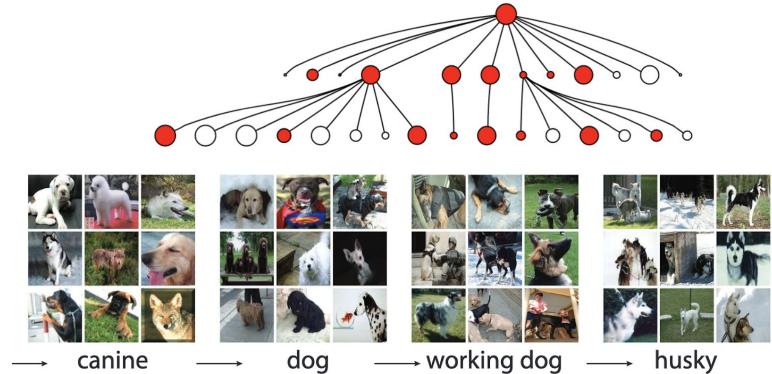
- Narrow linguistic/cultural domain
- No way to assess their real-world comprehension

Typical Vision and Language



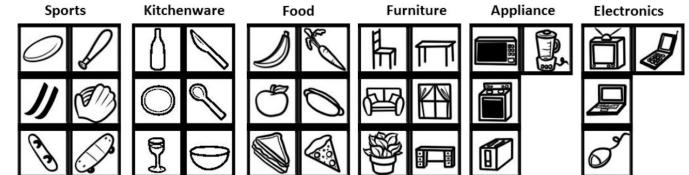
ImageNet (Deng et al. 2009)

- 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



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; Schweder et al. 2007).



Pilota / Jai-alai



Sanxian / Shamisen



Clavie

Concepts and Hierarchies

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

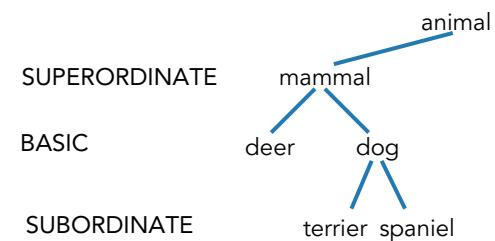
Concept: mental representation of a category (Rosch 1973)

Categories form a *hierarchy*

- Basic-level categories (Rosch 1976)

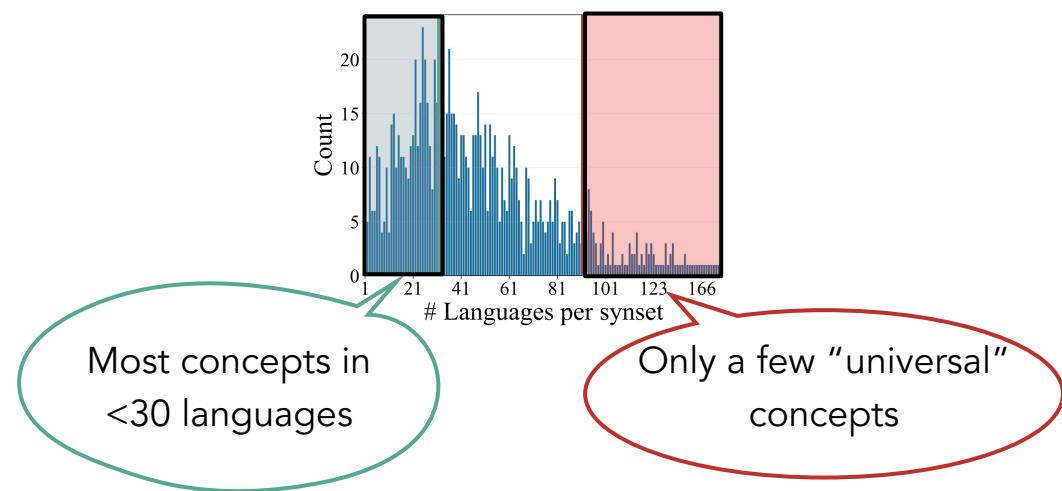
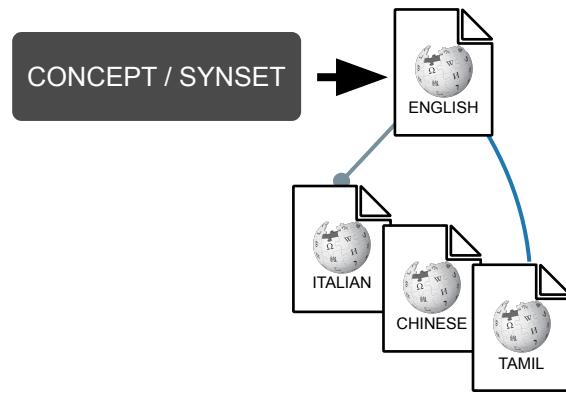
Somewhat universal

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



Are ImageNet Concepts Cross-Lingual?

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





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

Visual Reasoning Task

- **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



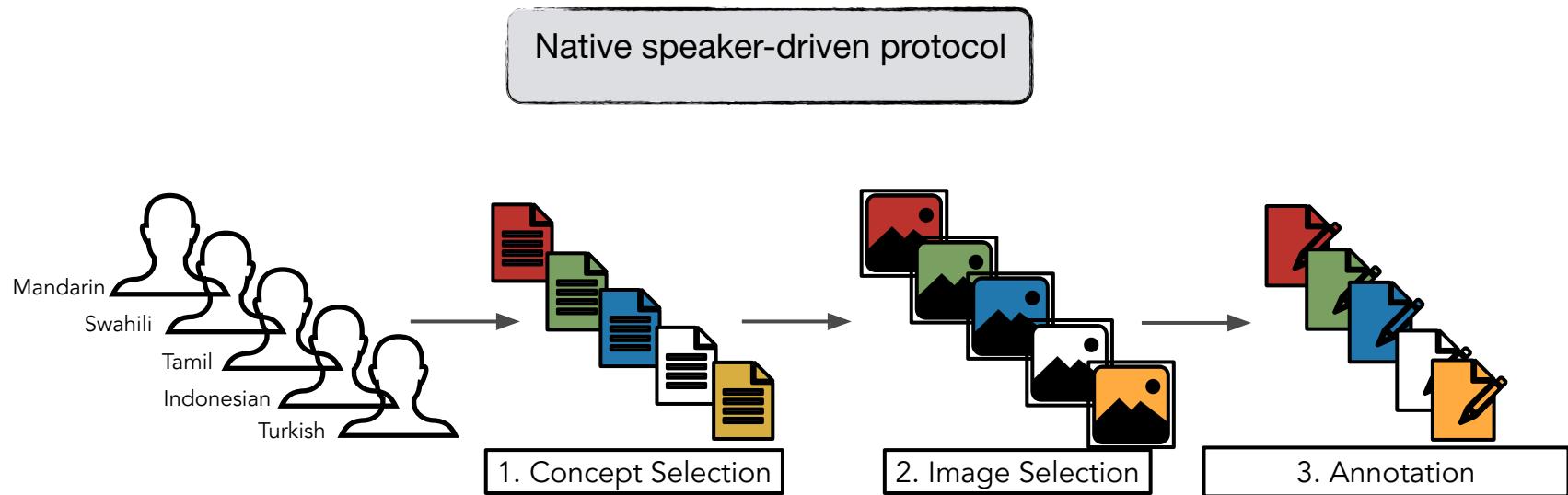
இரு படங்களில் ஒன்றில்
இரண்டிற்கும் மேற்பட்ட
மஞ்சள் சட்டை அணிந்த
வீரர்கள் காலையை அடக்கும்
பணியில் ஈடுபட்டிருப்பதை
காணமுடி.

True

X

Y

Collecting MaRVL data



MaRVL is created from Universal Concepts

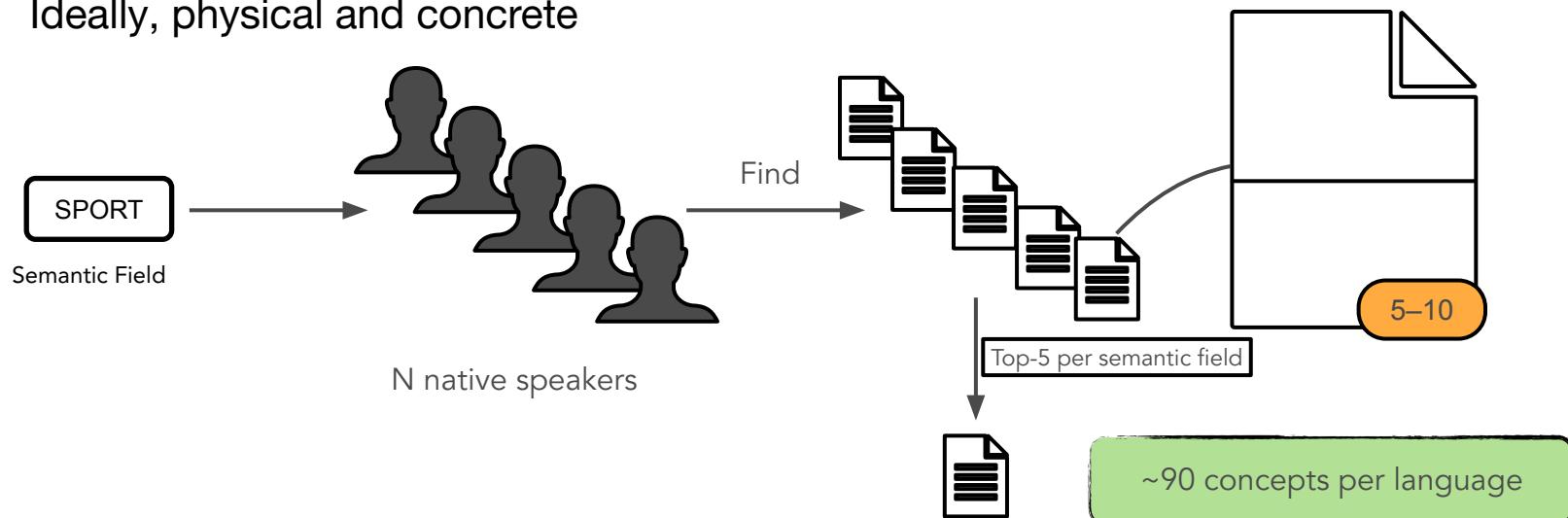
- Taken from the *Intercontinental Dictionary Series* ([Key & Comrie, 2015](#))
 - 18/22 chapters with concrete objects & events

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, agriculture
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 Concepts

- 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
 1. Contains more than one instance of a 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 **Çılmış** (Buttermilk)



MaRVL-sw Jembe (Shovel)



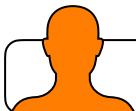
MaRVL-tr Rakı (Rakı)

Step 3. Language Annotation

Written by native speakers



MATCH 4 PAIRS AT RANDOM



VALIDATE ANNOTATIONS



右图中的人在发球, 左图中的人在接球。



WRITE CAPTION TRUE ONLY FOR 2 PAIRS



右图中的人在发球, 左图中的人在接球。



FINAL VALIDATION



Fleiss' kappa:
93%

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.)

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 தோம
(Vada)

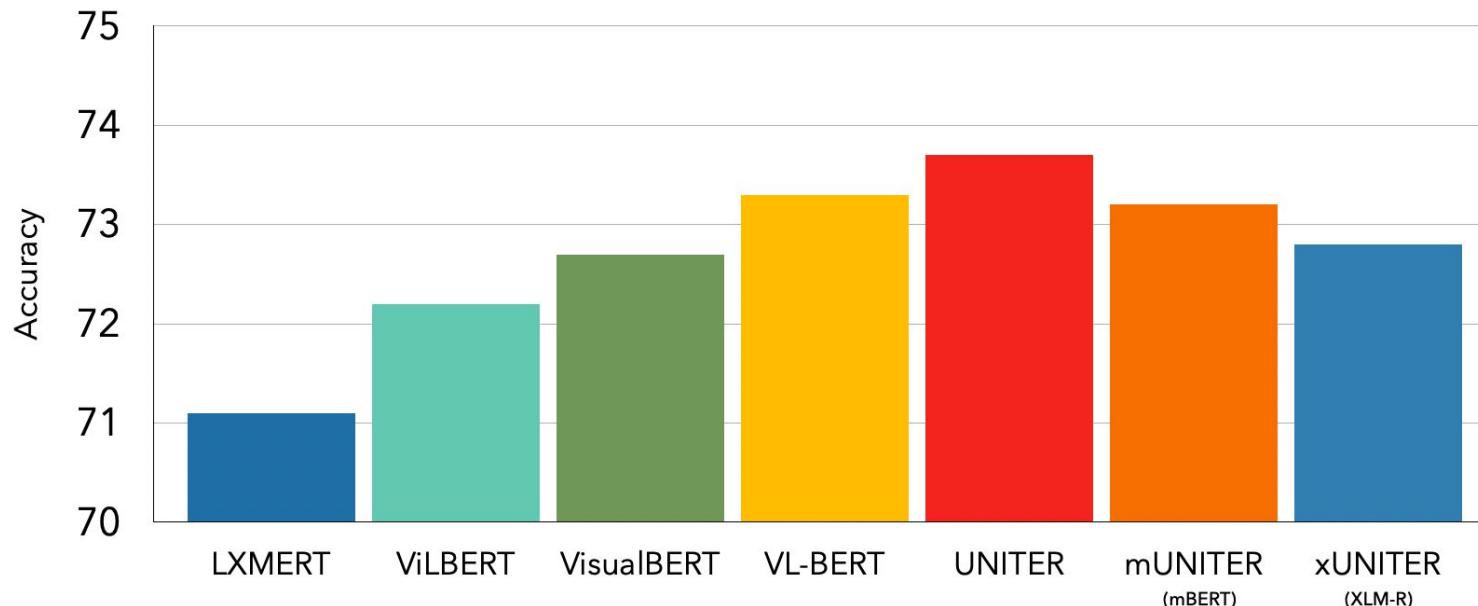


இரண்டு படங்களிலும் நிறைய மசால் வடைகள்
உள்ளன

(Both images contain a lot of masala vadas)

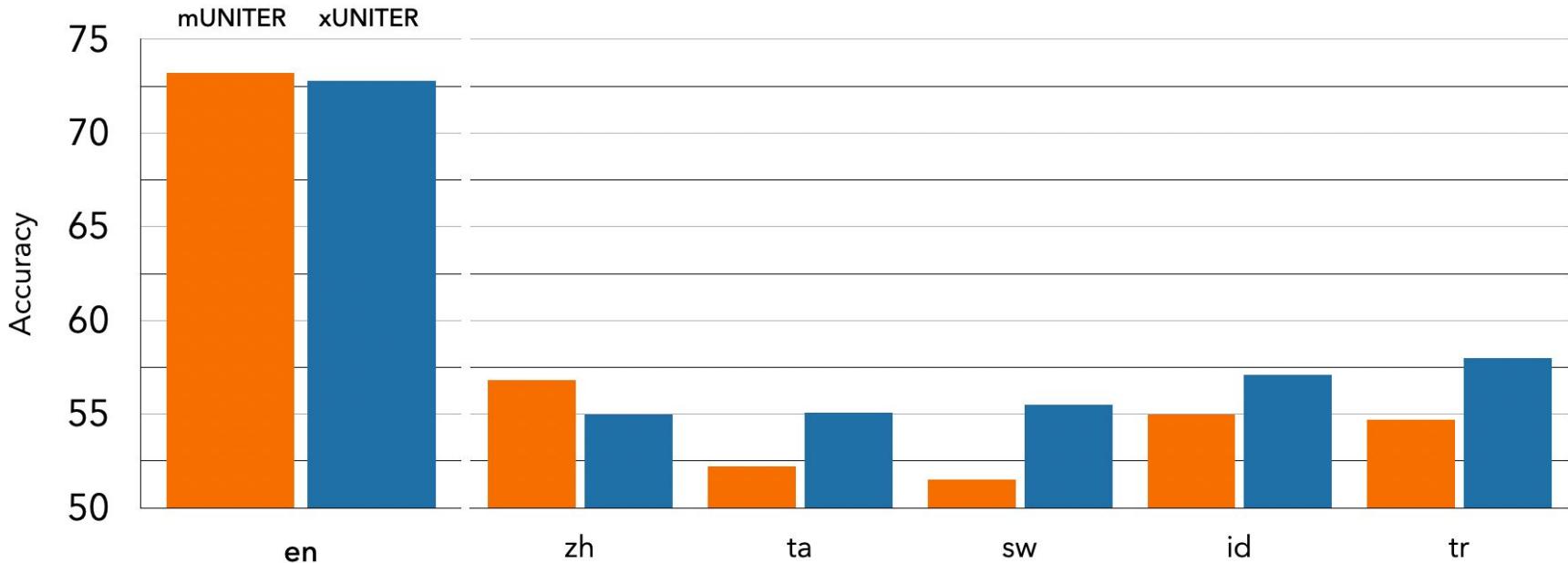
Label: False

English NLVR2 Results (Sanity check)



m/xUNITER perform similarly to English-only models

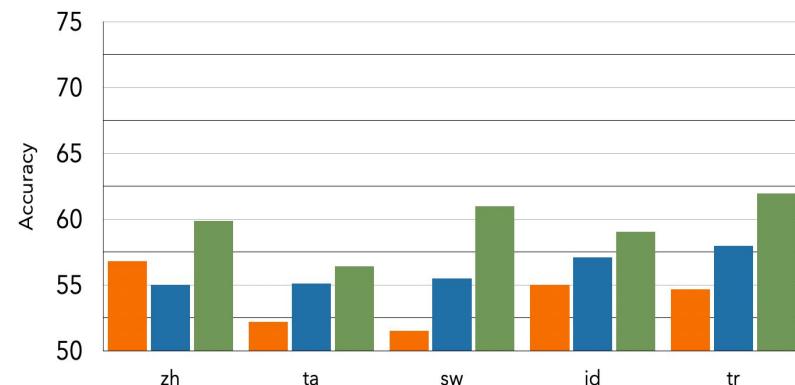
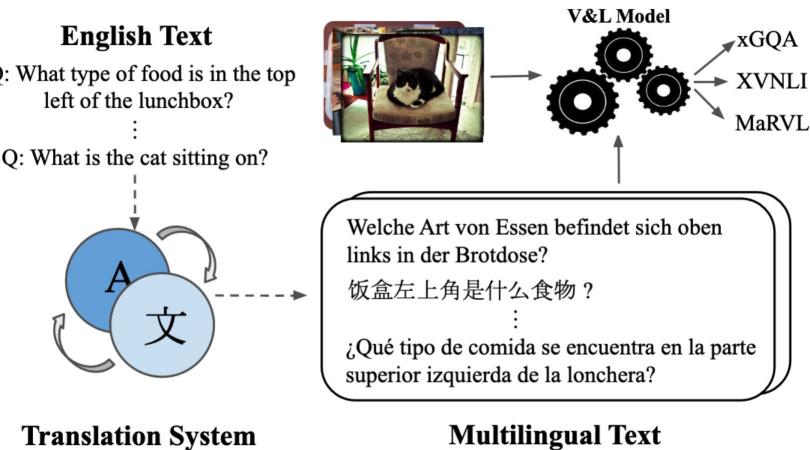
MaRVL Zero-shot Results



Zero-shot transfer: substantial drop in performance

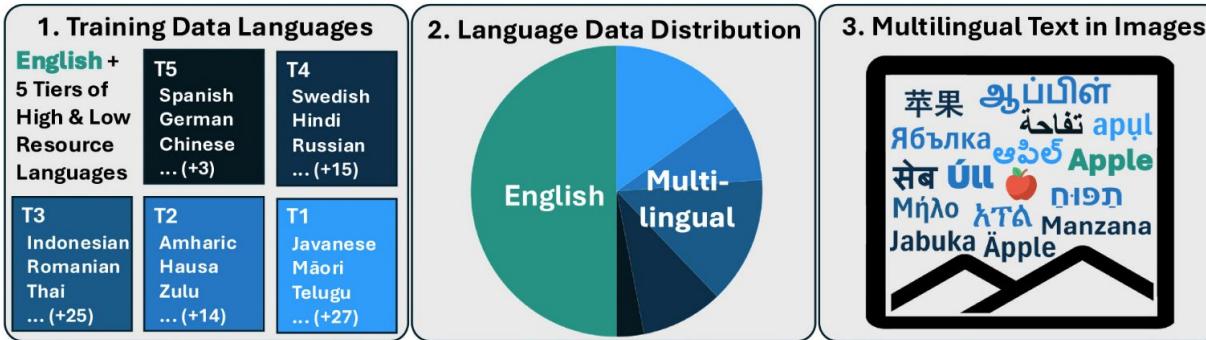
Pretraining with Translated Text

- Are the low zero-shot results caused by poor cross-lingual multimodal binding?



Cross-modal multilingual multimodal pretraining helps!

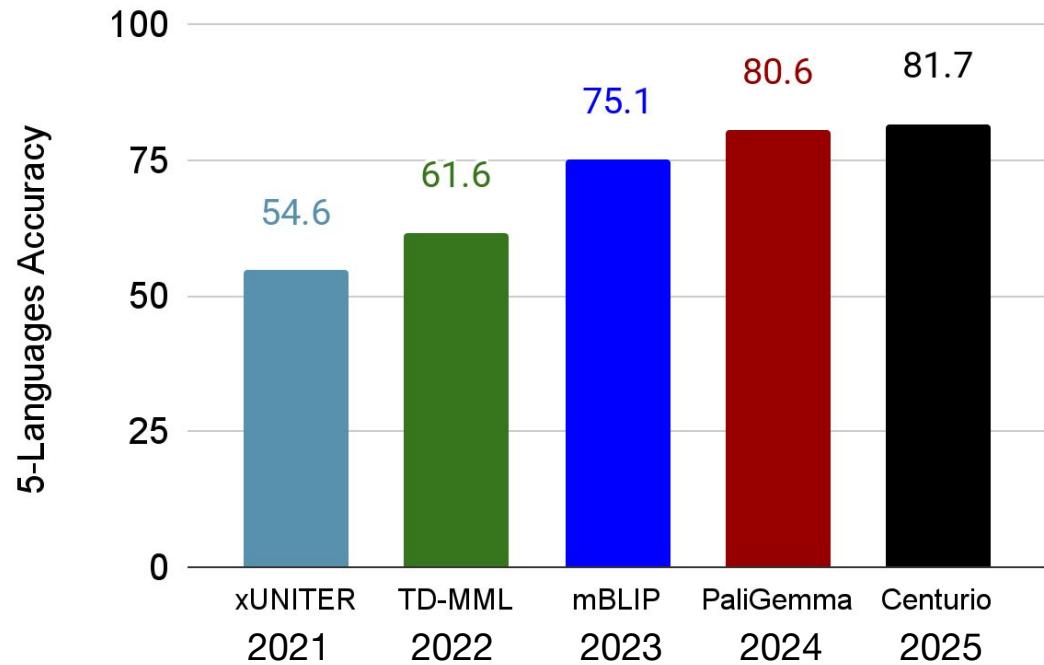
State of the Art: Centurio



- Initialize from
 - Qwen2.5 7B
 - SigLIP So400/384
- Fine-tune on lots of synthetic data
 - Synthetic OCR data
 - ALLaVa + ShareGPT4V captions
 - Machine translated texts

Year-on-Year Improvements

- Clear benefit when using machine translated data
- Better visual encoders and language models can enable effective zero-shot transfer

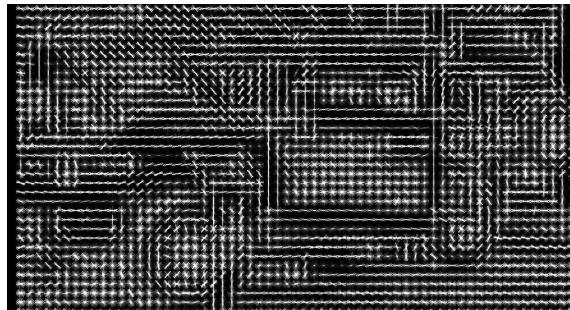


2. Data Representation

Three Levels of Representation

- Perceptual
- Pre-processed features
- Raw input

- Yellow
- Has wheels
- Metal
- Five-door
- Can transport
- ...



Perceptual Norms

- Ask people to write down the words that are triggered by textual stimuli.
- Stimuli: 541 noun concepts
- Norms are categorized into the likely knowledge source

Moose		
is large	27	visual-form and surface
has antlers	23	visual-form and surface
has legs	14	visual-form and surface
has four legs	12	visual-form and surface
has fur	7	visual-form and surface
has hair	5	visual-form and surface
has hooves	5	visual-form and surface
is brown	10	visual-color
hunted by people	17	function
eaten as meat	5	function
lives in woods	14	encyclopedic
lives in wilderness	8	encyclopedic
an animal	17	taxonomic
a mammal	9	taxonomic
an herbivore	8	taxonomic

Perceptual Norms: Pros / Cons

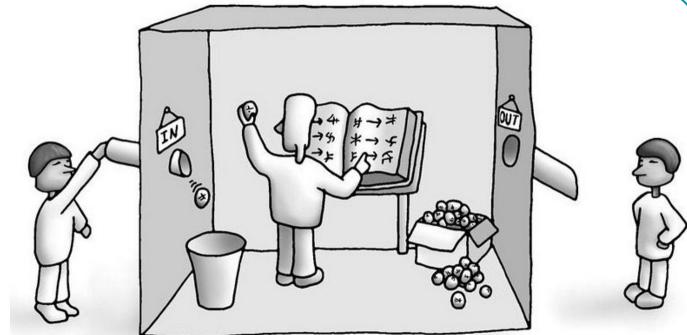
Pros

- Seemingly simple task
- Rich features

Cons

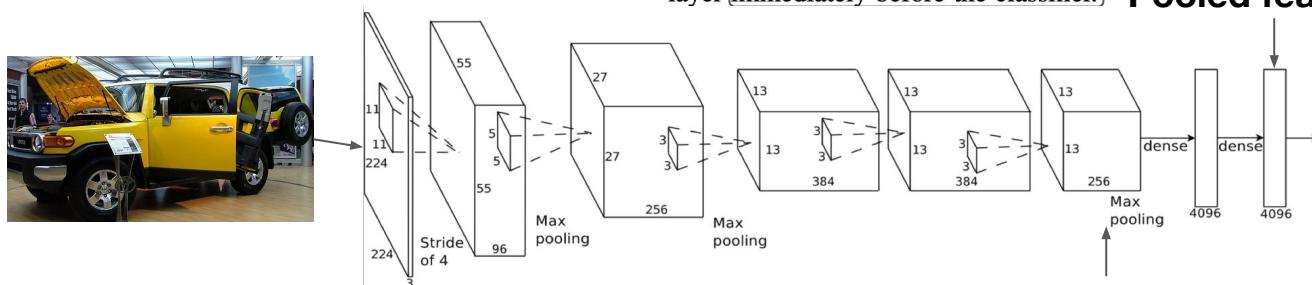
- Can it scale?
- Handling ambiguity

What does it mean to only understand symbols as defined by other symbols?



Spatial and Pooled Visual Features

- Earliest work in neural-network era used pooled or spatial preserving features from a pretrained Convolutional Neural Network.

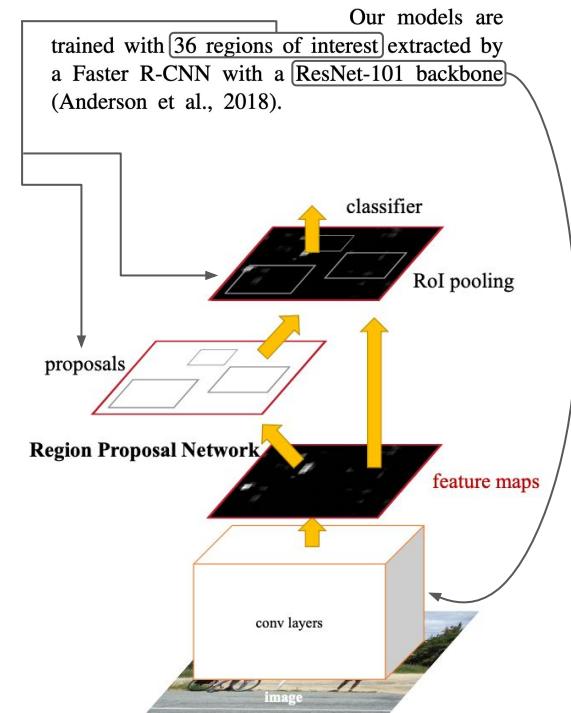


“where $CNN(I_b)$ transforms the pixels inside bounding box I_b into 4096-dimensional activations of the fully connected layer immediately before the classifier.” **Pooled features**

Spatial features “In our experiments we use the $14 \times 14 \times 512$ feature map of the fourth convolutional layer before max pooling.”

Pre-processed Visual Features

- Faster R-CNN region-based feature vectors
 - Trained on the Visual Genome Dataset
 - The Region Proposal Network suggests the location of *regions of interest*.
 - RoI pooling performs spatial pooling in the final CNN layer to give a 2048D vector.



Pre-processed: Pros / Cons

Pros

- Long-established practice
- Usually an offline process:
do it once and forget

Cons

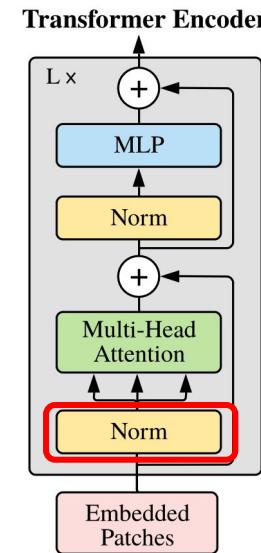
- Large datasets require specialized storage
- Not obvious how to randomly augment data
- Specialist knowledge can be opaque to newcomers

Raw Input

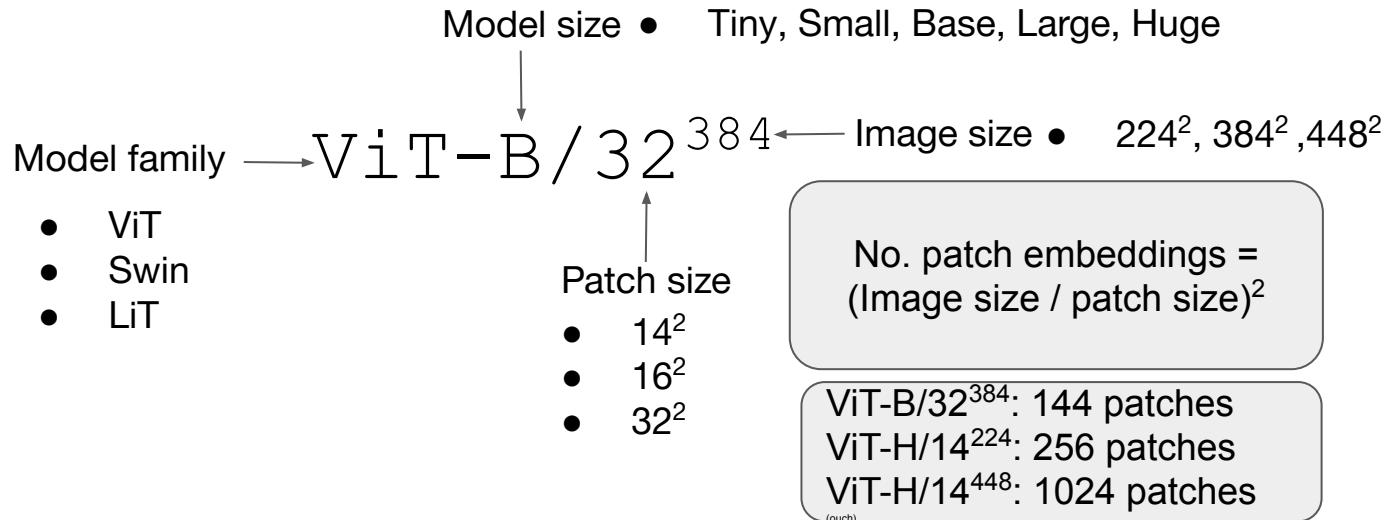
- Directly process data from the raw images or speech signal.
- Images:
 - Vision Transformer (ViT)
 - Swin Transformer
- Speech
 - Spectrogram Transformer
 - AudioMAE

Vision Transformer

- Good news! You are already almost an expert in how the Vision Transformer works
 - Split image into K patches
 - Embed each patch
 - Add position information
 - Encode using Transformer blocks that include an **extra pre-norm layer** for stability.



Nomenclature and Patch Count

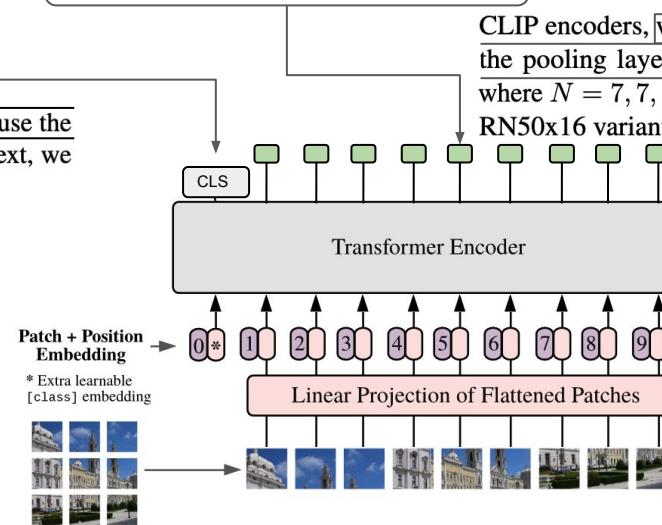


Extracting ViT Features

- Extract pooled features or patch-level features

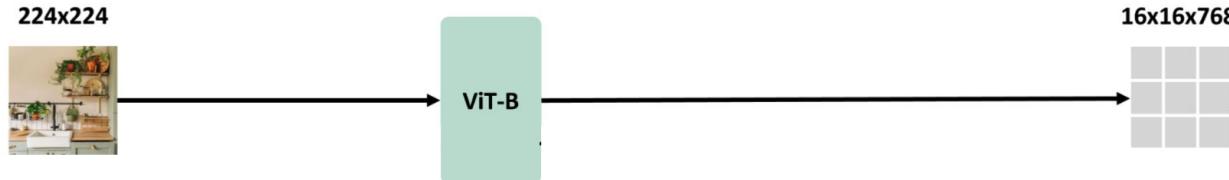
To extract visual information from an image x^i , we use the visual encoder of a pre-trained CLIP [29] model. Next, we

For the CLIP encoders, we extract the feature grid before the pooling layers, resulting in an $N \times N$ grid, where $N = 7, 7, 12$ for the ViT-B/32, RN50x4 and RN50x16 variants of CLIP respectively.



Scaling on Scale

- Improve performance by combining original and higher-resolution processing without additional pretraining or finetuning



Raw input: Pros / Cons

Pros

- Data augmentation is straightforward because you always have the raw input
- Fewer preprocessing steps means fewer creeping errors

Cons

- Smaller batches with an extra model on the GPU
- Potentially many inputs

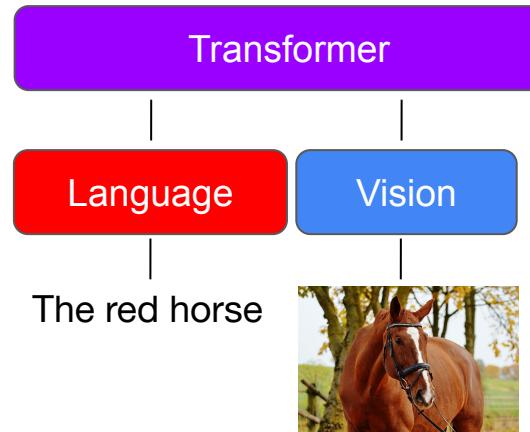
Summary

- Many options for how to represent your multimodal inputs
 - Language-oriented
 - Object / stuff oriented
 - Raw inputs
- **No universally best option** but raw inputs are promising because the visual encoding model can be fully differentiable

3. Modelling

Main Approaches

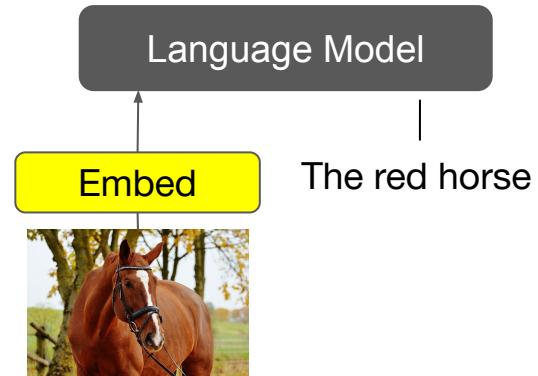
Dual / cross encoding



Cross-Attention

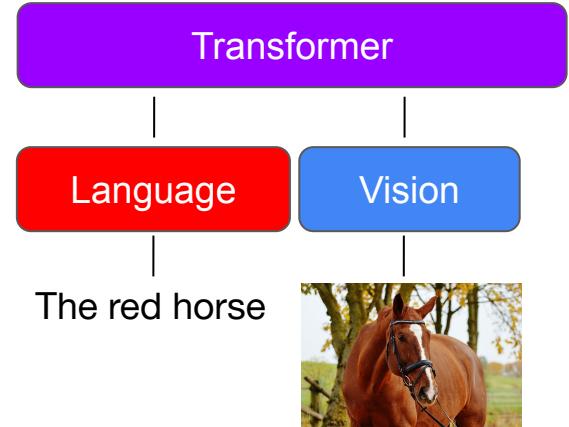


Visual Prefix



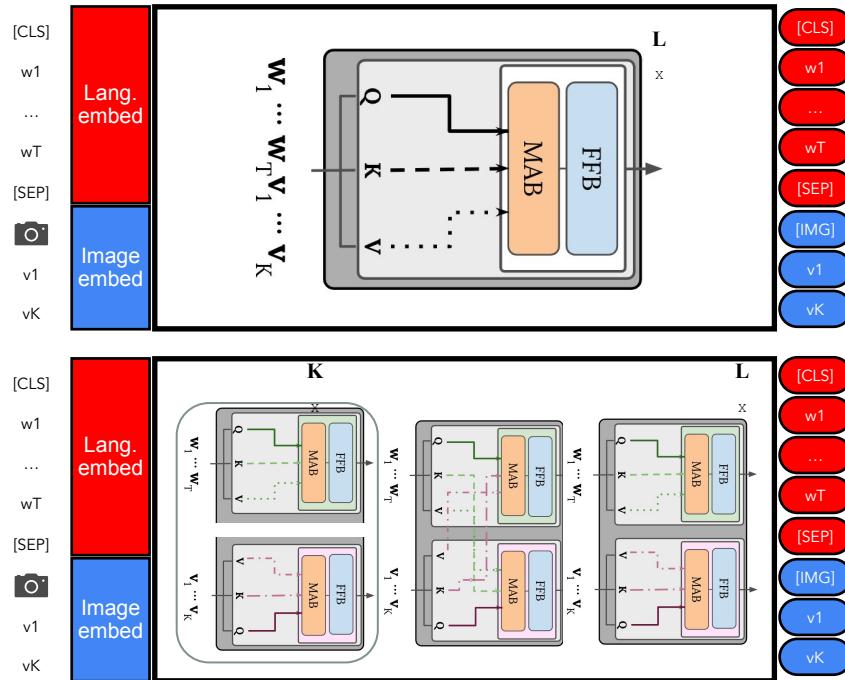
Cross-encoding Models

- Emerged as a key modelling approach in 2019 with many concurrent methods for creating visually-grounded BERT models.
- This is a form of *model-based fusion*
- The backbone consists of two components:
 - language encoder
 - visual encoder



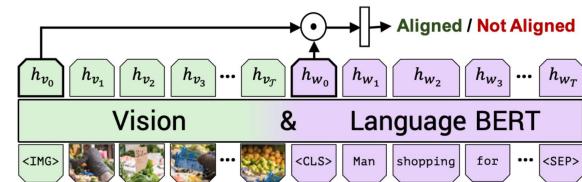
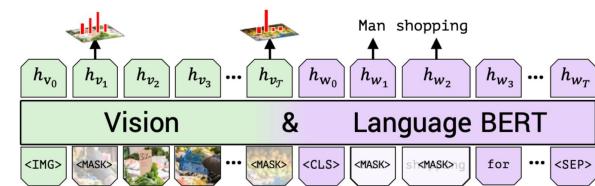
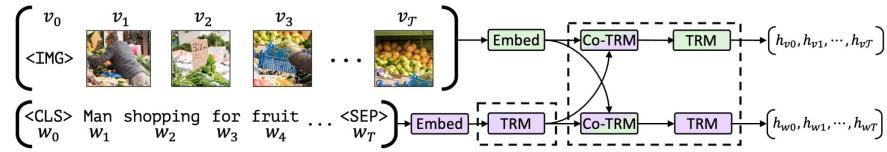
Single- & Dual-Stream Architectures

- Single-stream
 - Concatenate inputs into one sequence
- Dual-stream
 - Process modalities independently
 - Intra-modal
 - Inter-modal



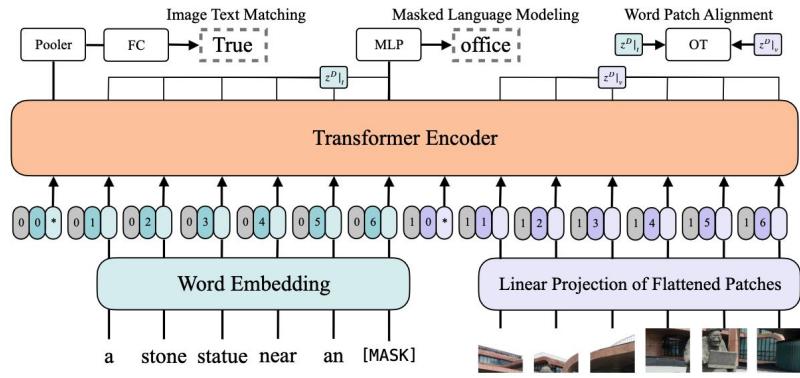
ViLBERT

- Dual-stream model
- Initialized from BERT
- Visual features extracted from 10-36 regions using Faster-RCNN
- Pretrained on Conceptual Captions
 - Masked Language Modelling
 - Masked Region Classification
 - Image-Text Matching

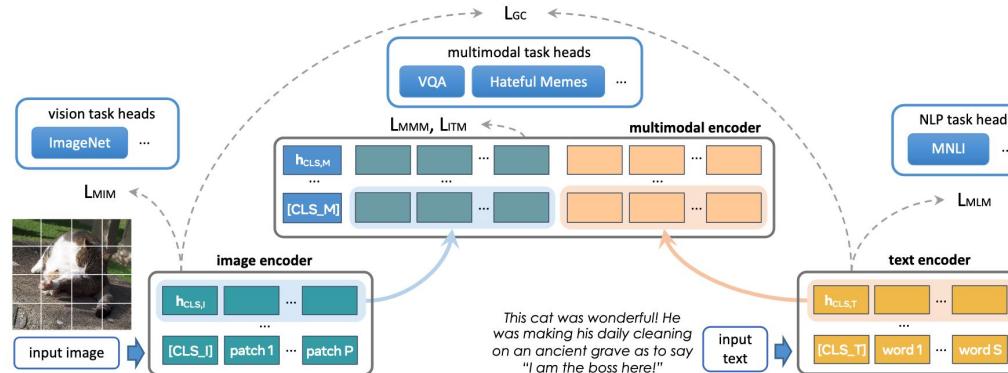


ViLT

- Single-stream model
- **Initialized from BERT**
- Visual features extracted from ViT-B/32
- Pretrained on Conceptual Captions, Visual Genome, COCO, SBU Captions
 - Masked Language Modelling
 - Image-Text Matching
 - Word-Patch Alignment



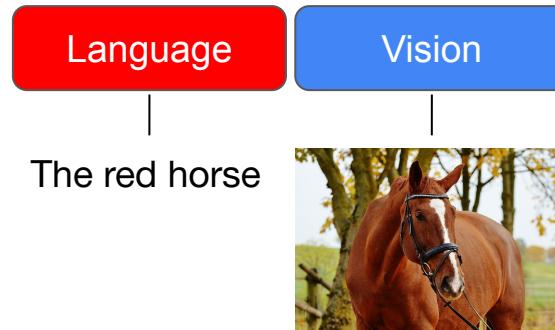
FLAVA



- Dual-stream Visual features extracted from ViT-B/16
- Pretrained on PMD70M
 - Masked Language Modelling, Masking Image Modelling
 - Image-Text Matching, Masked Multimodal Modelling
 - Global Contrastive Matching

Dual-encoding Models

- Emerged as a sample-efficient alternative to cross-encoding
- The backbone consists of two separate components:
 - language encoder
 - visual encoder



CLIP

Language

The red horse

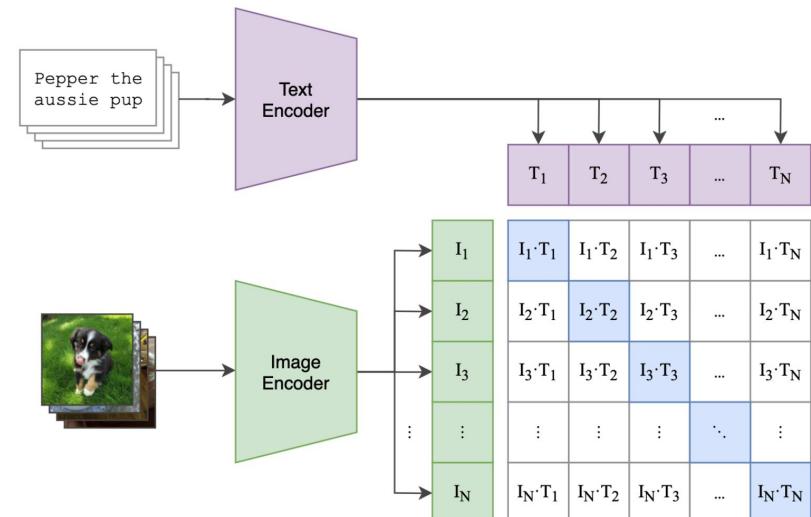
Vision



- 12 Layer Transformer Encoder
- ViT or ResNet Visual Encoder
- Maximize the similarity of the embeddings of paired examples (I, T):

$$\mathcal{L}_{\text{InfoNCE}} = -\mathbb{E} \left[\log \frac{f(\mathbf{t}, \mathbf{i})}{\sum_{\mathbf{t}' \in T} f(\mathbf{t}', \mathbf{i})} \right]$$

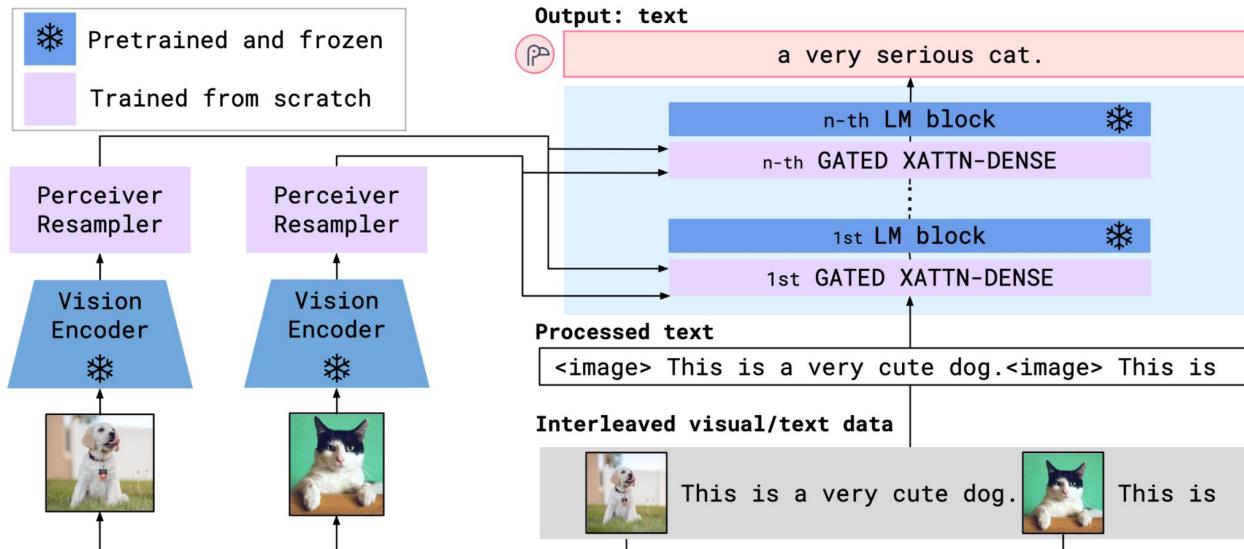
- Large pretraining dataset of unclear provenance



Cross-Attention

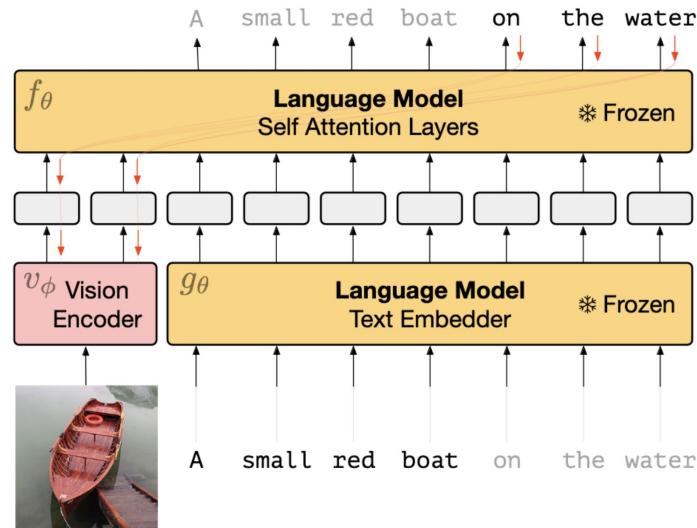
Attend Language Model

The red horse

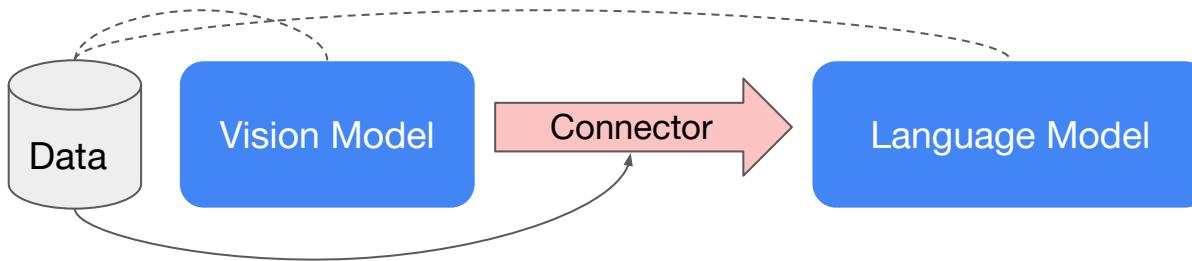


Visual Prefix

- Force the visual representations into an LLM-compatible space through a learned projection



Training VLMs in 2025



- Three main parts
 1. Pretrained models: vision encoder & language model
 2. Define a connection layer between the models
 3. Train the connection with multilingual multimodal data
 - a. Continued pretraining and instruction tuning

Modality-Specific Components

	Vision Encoder	Language Model
LLAVA	CLIP ViT-L/14	Vicuna-13B
Qwen-VL	OpenCLIP ViT-bigG	Qwen-7B
MM1	ViT-L	1.3B LLM
PaliGemma	SigLIP-So400M/14	Gemma-2B

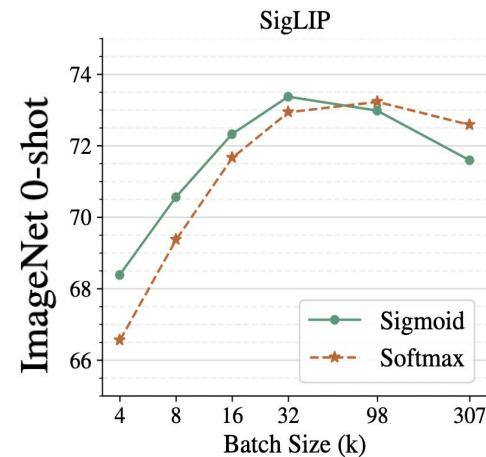
<https://ai.google.dev/gemma/docs/paligemma>
<https://llava-vl.github.io/>
<https://github.com/QwenLM/Qwen-VL>

SigLIP Image Encoder

“Unlike standard contrastive learning with softmax normalization, the sigmoid loss operates solely on image-text pairs and does not require a global view of the pairwise similarities for normalization.”

$$-\frac{1}{|\mathcal{B}|} \sum_{i=1}^{|\mathcal{B}|} \sum_{j=1}^{|\mathcal{B}|} \log \underbrace{\frac{1}{1 + e^{z_{ij}(-t\mathbf{x}_i \cdot \mathbf{y}_j + b)}}}_{\mathcal{L}_{ij}}$$

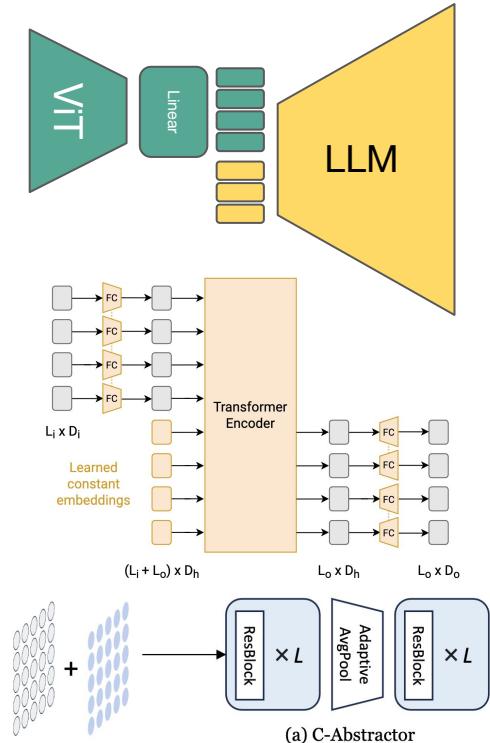
Label of the image-text pair: 1 if matched else -1



Connector Design

- PaliGemma: Map the image output embeddings into the language model word embedding space.
- Qwen-VL:

Position-aware Vision-Language Adapter: To alleviate the efficiency issues arising from long image feature sequences, **Qwen-VL** introduces a vision-language adapter that compresses the image features. This adapter comprises a single-layer cross-attention module initialized randomly. The module uses a group of trainable vectors (Embeddings) as query vectors and the image features from the visual encoder as keys for cross-attention operations. This mechanism compresses the visual feature sequence to a fixed length of 256. The ablation about the number of queries is shown in Appendix E.2. Additionally, considering the significance



- MM1: Convolutional-Abstractor
 - ResNet Block followed by an Adaptive Pooler

Training Datasets

- LLAVA: 595K image–caption examples filtered from CC3M

Qwen-VL 1.4 billion examples (77% English / 23% Chinese)

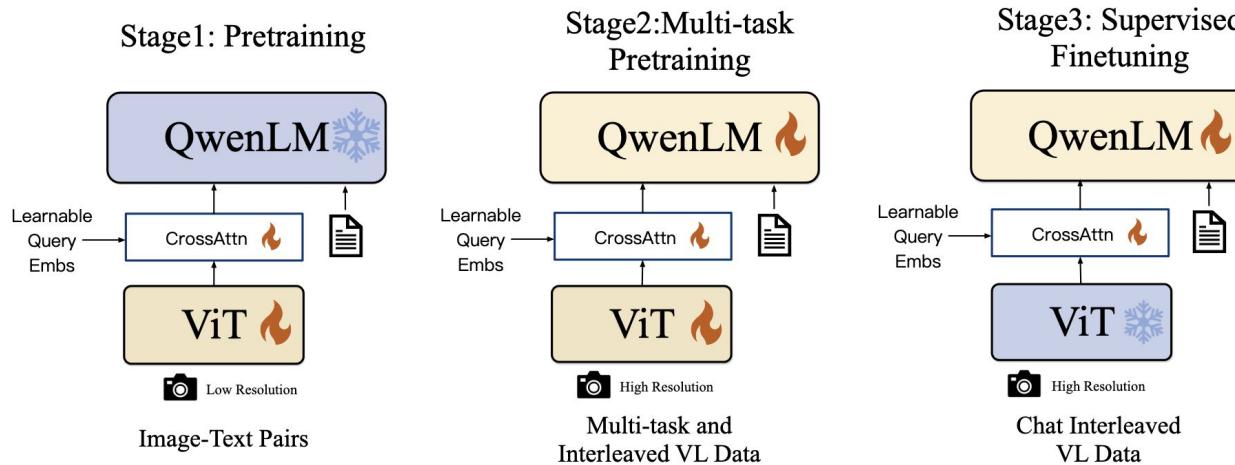
MM1 2+ billion mixture of image–text examples

PaliGemma 1 billion mixture of multilingual image caption, VQA, and
in-the-wild datasets

- The larger models are pretrained on **in-house data**
 - PaliGemma: WebLI (1B+), Qwen-VL (220M), MM1 (1B+)

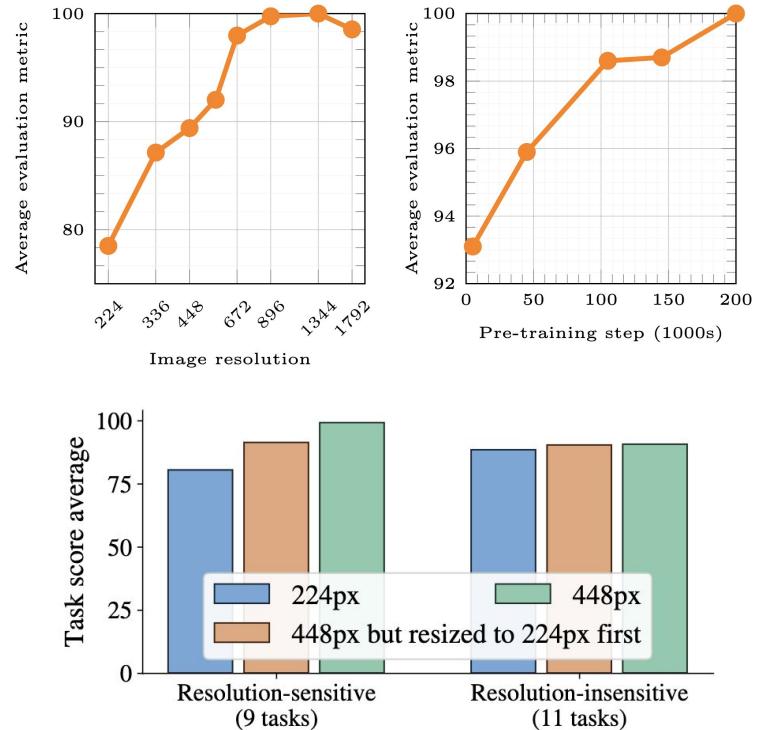
Training Strategy

- Qwen-VL, PaliGemma, and MM1 use multi-stage training strategies with different types of data and different image resolutions



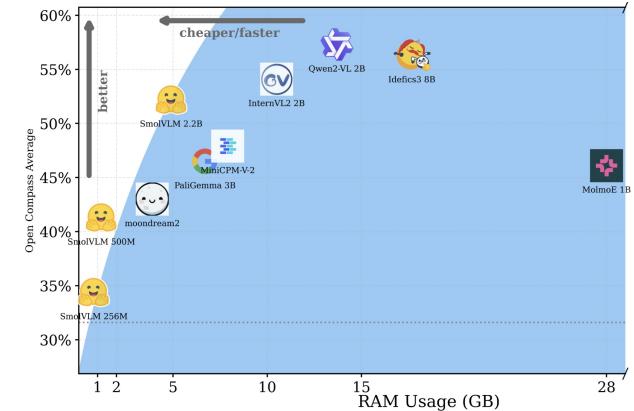
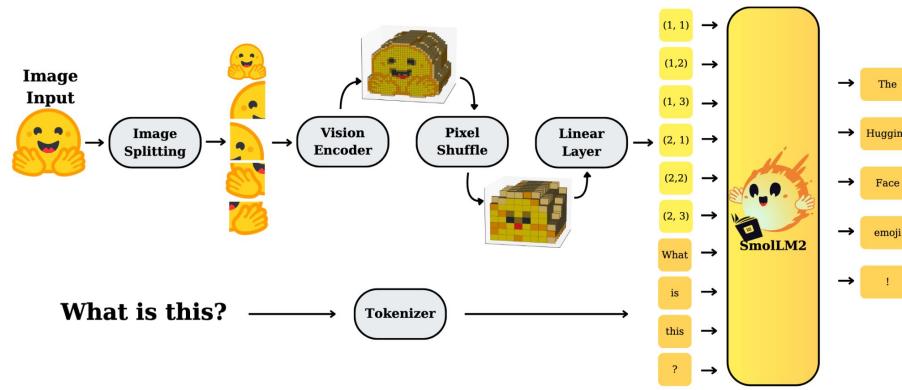
Open Questions

- How quickly will we realize these benefits in smaller models?
- Do LMMs really need 1 billion examples to learn a bridge?
- What will happen to model performance when we develop new tasks that involve weaker visual–linguistic bindings?

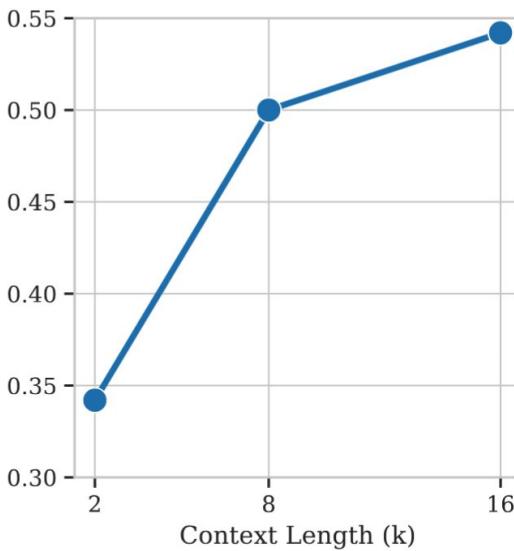
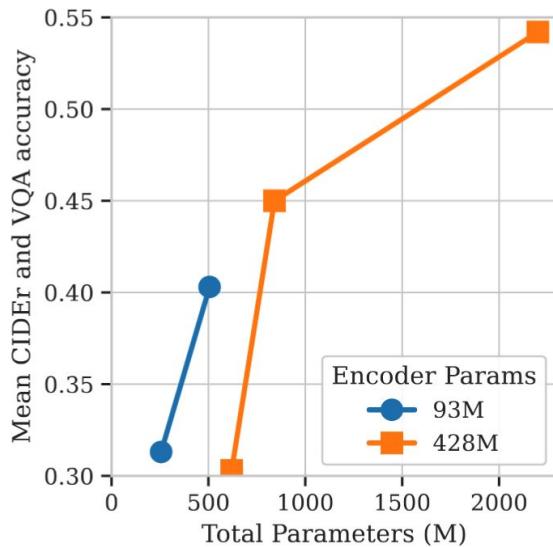


Trends in Smaller Models

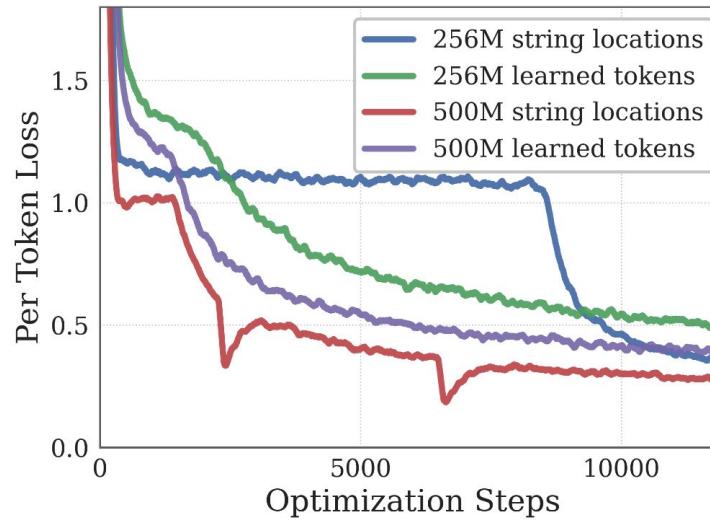
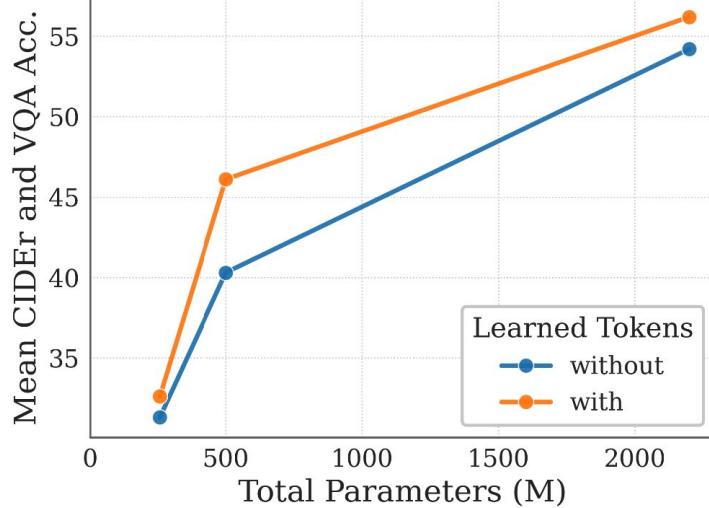
- How quickly will we realize these benefits in smaller models?
 - Less than six months



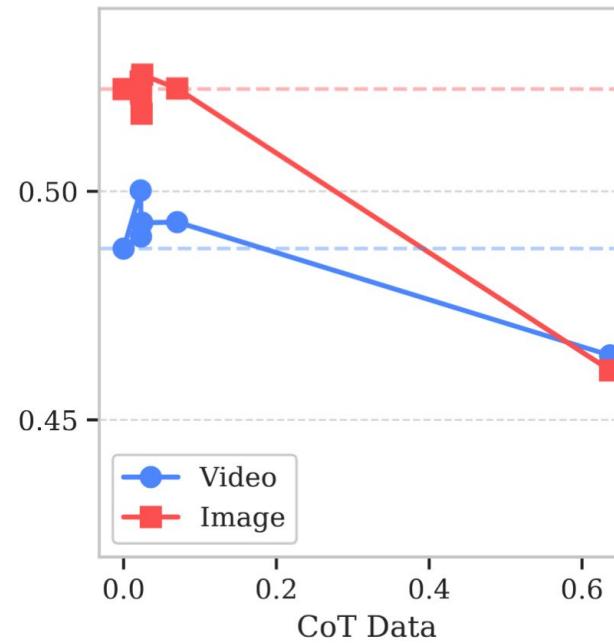
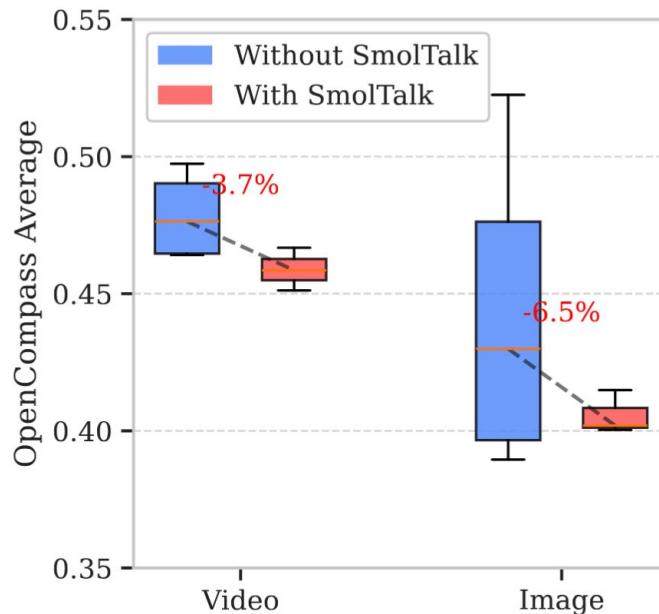
Balance Models and Increase Context



Learned Position Embeddings Help!



Chain-of-Thought Data Hurts



Summary

- Cross-encoding:
 - Many advances in which parts of the input contribute to loss
 - Shift from regions-of-interest to Vision Transformers
- Dual-encoding:
 - Excellent cross-domain transfer to a wide range of problems
- Visual Prefix Learning:
 - Exploit the benefits of single-modality pretraining

**Q: Does an image captioning model
need to learn everything in-weights?**

PAELLA: Parameter-Efficient Lightweight Language-agnostic Captioning Model

Findings of NAACL 2024



R. Ramos



E. Bugliarello



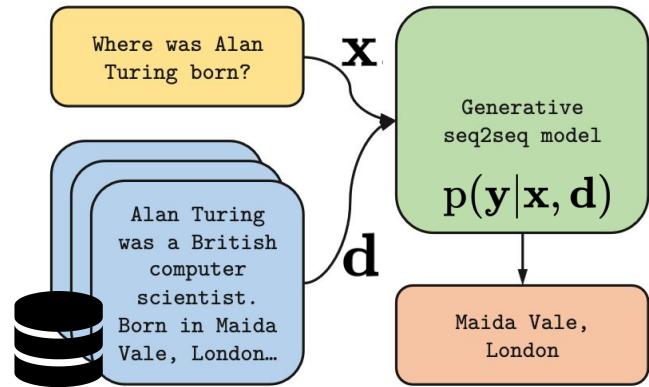
B. Martins



D. Elliott

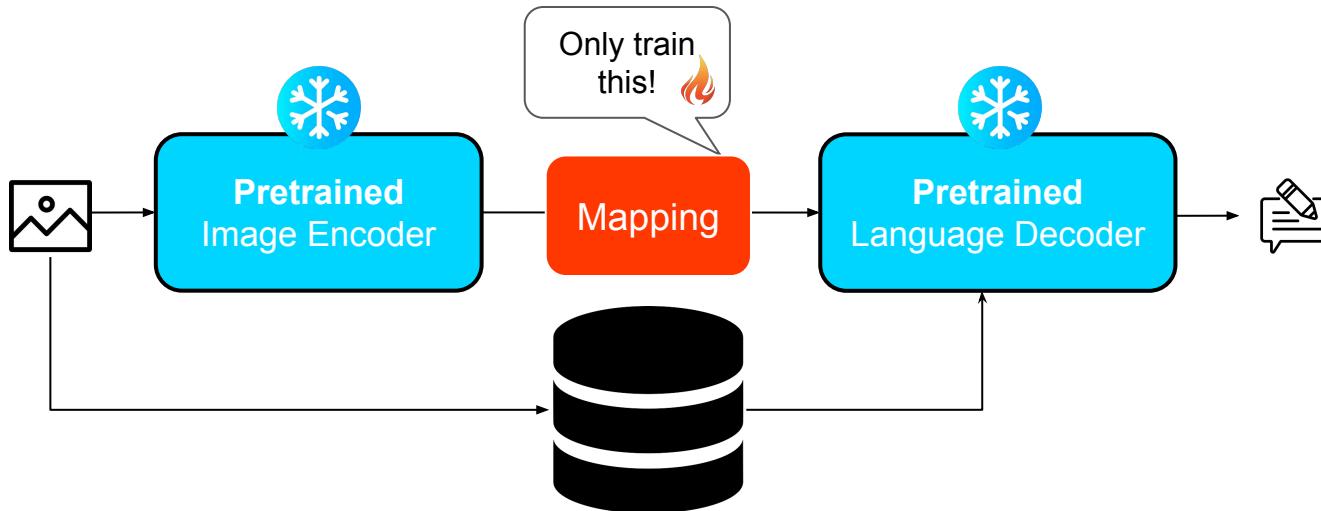
Retrieval Augmented Generation

- Combine the power of in-weights learning with in-context adaptation through retrieval augmentation
- Given a datastore of facts, knowledge, documents, etc.
 - Combine the most relevant items from the datastore (d) with the input (x) for your task



Motivation

- Main trend in V&L is training bigger models on more data
- Alternative is emerging that re-uses independent backbone models
- Can we further improve performance with retrieval augmentation?

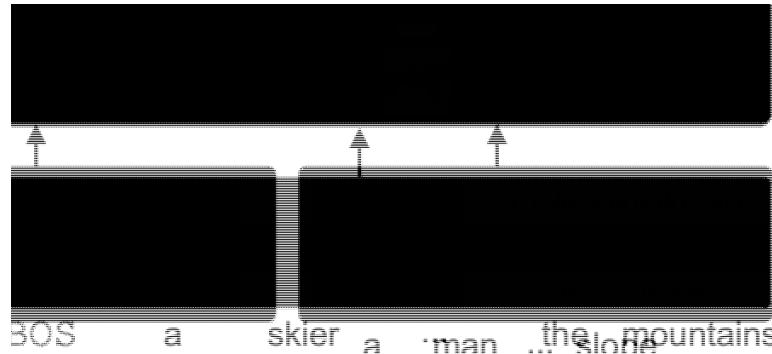


PAELLA Model



Retrieval System

- Build a FAISS datastore: store high-dimensional vectors
 - Captions of images represented with CLIP embeddings
- Retrieve k nearest-neighbours captions from datastore
 - Image embedding compared against datastore caption vectors

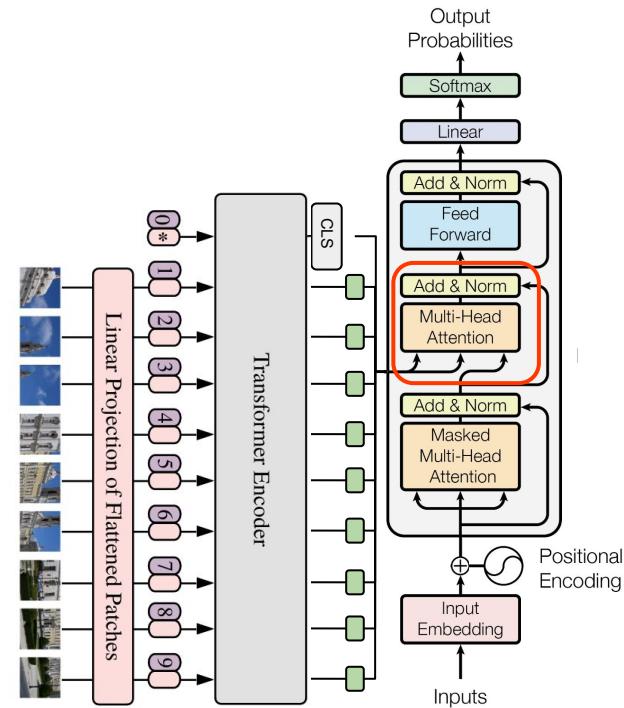


Trained Cross-Attention Layers

- We insert a randomly initialized **cross-attention mechanism** to attend to the visual encoder output embeddings

Rank	Params
d=128	553M
d=8	34M

$$\text{Att}(\mathbf{XW}^Q, \mathbf{XW}^K, \mathbf{XW}^V)$$
$$\mathbf{W}^Q, \mathbf{W}^K, \mathbf{W}^V \in \mathbb{R}^{d_{\text{enc}} \times d}$$



Experimental Protocol

- Encoder: Multilingual CLIP
- Decoder: XGLM-2.9B
- Training data:
 - 566K captions sampled uniformly from COCO-35
- Evaluation: XM-3600
 - 3600 geographically-diverse images
 - 36 languages: 100 captions per image
 - 5 low-resource languages (L5):
 - Bengali, Cusco Quechua,
Maori, Swahili, Telugu



Examples images from XM3600

Results

	Data	Trained Θ	L36	L5
PaLI	12B	17B	53.6	-
Lg _{COCO-35}	19M	2.6B	15.0	12.5
mBLIP: BLOOMZ-7B	135M	800M	23.4	6.7
BB+CC _{COCO-35 + CC-35}	135M	800M	28.5	22.4
mBLIP: mT0-XL	489M	124M	28.3	7.9
PAELLA	566K	30M	26.2	20.7

PAELLA is competitive against models with 35-863x more training data, and 4-87x more trained parameters

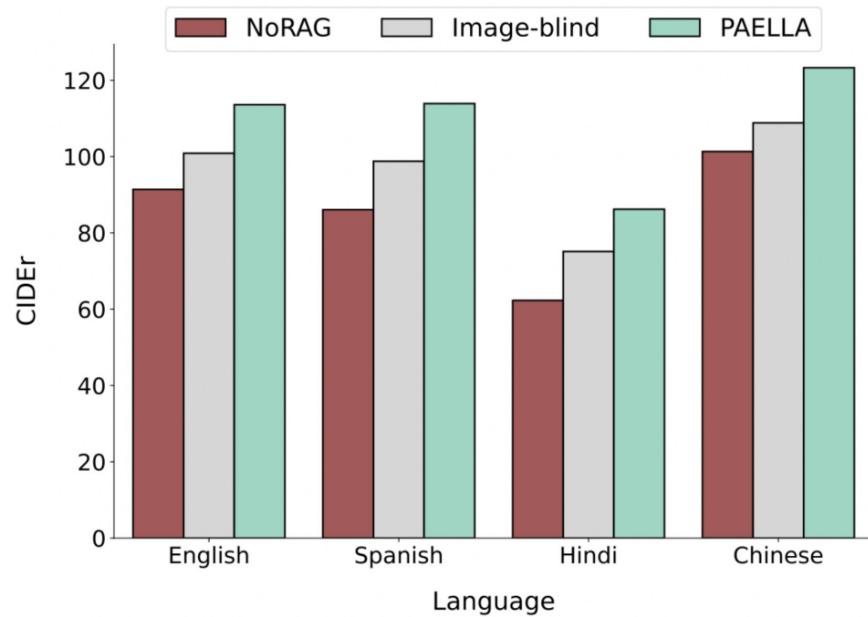
Zero-shot Multilingual Transfer

- **PAELLA_{MONO}** is a variant trained on 566K examples in English COCO
- Outperforms **Lg** trained on 19.8M examples in the machine translated COCO-35 dataset

	Data	Trained Θ	L36	L5
Lg: Thapliyal et al. <small>coco-35</small>	19M	2.6B	15.0	12.5
PAELLA_{MONO}	566K _{en}	30M	15.5	12.1

Value of Retrieval Augmentation

Consistent improvements from multilingual retrieval augmentation across the core languages in the XM3600 evaluation data



Qualitative Example



类似图片显示：

ऐसी ही तस्वीरें दिखाती हैं:

Imágenes similares muestran:

Similar images show:

the owl is perched outside in front of the people
an owl sitting a top a table during the daytime
an owl is sitting on a perch at a camp site
the fuzzy owl is sitting on a tree branch

A caption I can generate to describe this image in english is:

PAELLA

en: "an owl sitting on top of a tree"

es: "un búho sentado en una rama de un árbol"
(an owl sitting on a tree branch)

hi: "एक उल्लू एक पेड़ की टहनी पर बैठा है"
(an owl is sitting on a tree branch)

zh: "一只 猫头鹰 站在 树上"
(an owl standing in a tree)

NoRAG

en: "a large black and white picture of a bird"

es: "un pájaro posado en la parte superior de un edificio"
(a bird perched on the top of a building)

hi: "एक पेड़ के पास खड़ा एक पक्षी"
(a bird standing near a tree)

zh: "一只 长颈鹿 坐在 树枝 上"
(a giraffe sitting on a branch)

Q: Do you even need to train?

LMCap: Few-shot Multilingual Image Captioning by Retrieval Augmented Language Model Prompting

Findings of ACL 2023



R. Ramos



B. Martins

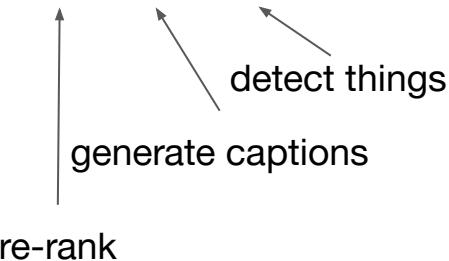


D. Elliott

Socratic Models

- Enable models to “communicate” with each other through their output labels, prompting, and ranking

$$f_{\text{VLM}}^3(f_{\text{LM}}^2(f_{\text{VLM}}^1(\text{image})))$$

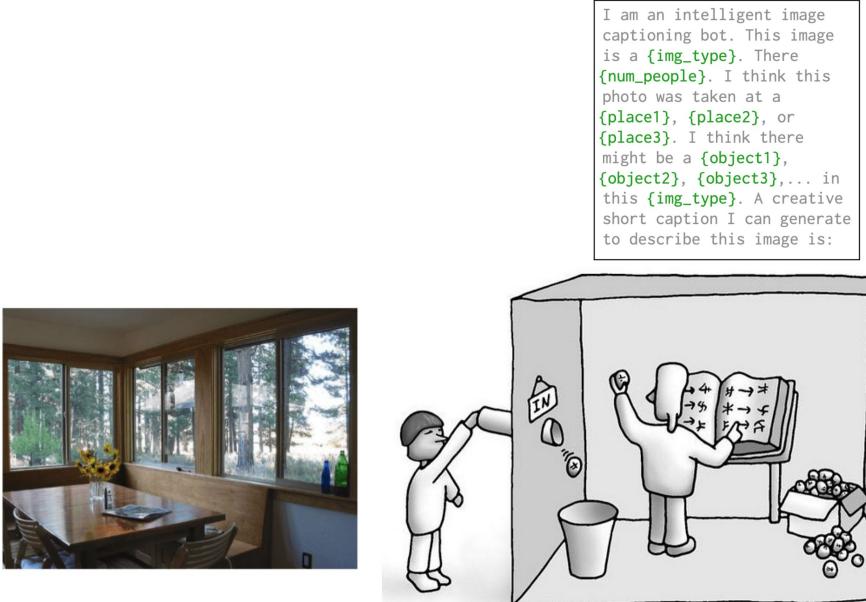


I am an intelligent image captioning bot. This image is a {img_type}. There {num_people}. I think this photo was taken at a {place1}, {place2}, or {place3}. I think there might be a {object1}, {object2}, {object3},... in this {img_type}. A creative short caption I can generate to describe this image is:



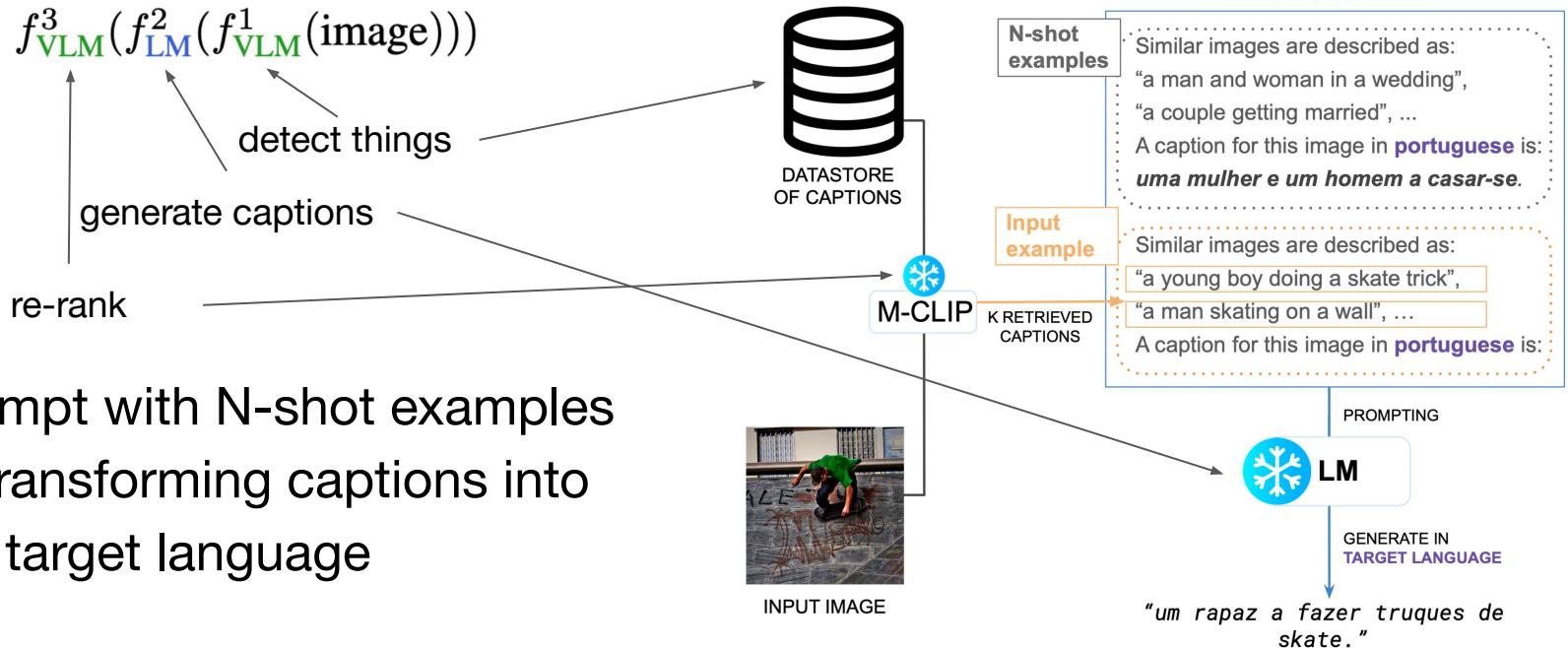
SM (ours): This image shows an inviting dining space with plenty of natural light.

ClipCap: A wooden table sitting in front of a window.



What does it mean to only understand symbols as defined by other symbols?

Multilingual Captioning with Retrieval Augmentation

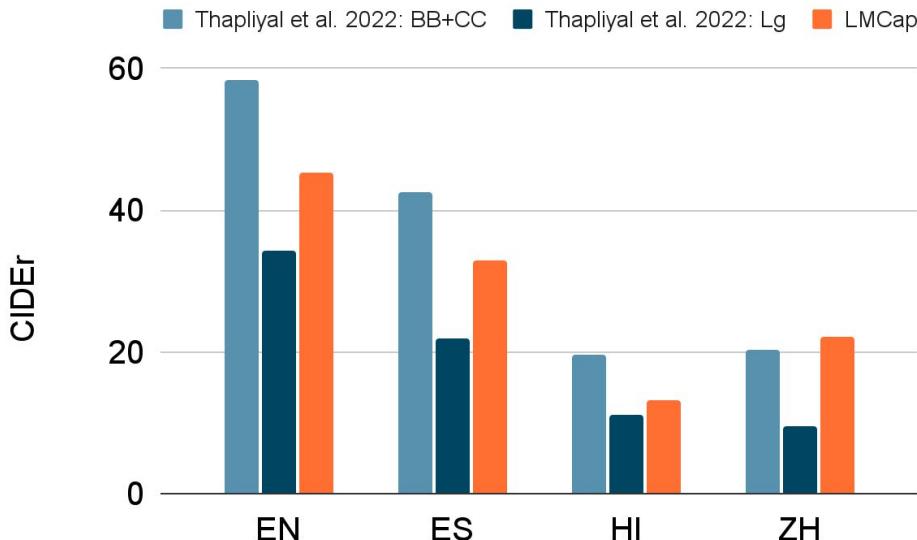


- Prompt with N-shot examples of transforming captions into the target language

Experimental Setup

- XGLM Language Model 564M - 7.6B params
- Multilingual CLIP (LAION)
- Experiments on XM3600
 - 100 images in 36 languages
- **No training or fine-tuning on any captioning data.**

Results



Competitive performance
compared to supervised models

Params	Config.	RAM	en	es	hi	zh
564M	K=4, N=3	6G	0.411	0.094	0.030	0.146
1.7B	K=4, N=3	12G	0.637	0.143	0.066	0.272
2.9B	K=4, N=3	16G	0.767	0.454	0.334	0.584
7.5B	K=4, N=3	22G	0.787	0.489	0.365	0.644

Need at least 2.9B parameter
decoder for multilingual generation

Qualitative Example

Retrieved Examples



two people and a kid skiing along a trail

an adult and two children are cross country skiing

two men and a little boy are skiing on a snowy spot

two adults on skis with a child on skis between them

Generated Captions

ENG: two people and a kid skiing along a trail

ESP: dos hombres y un niño esquiando en una pista de nieve

ZHO: 两个大人和一个小孩在雪地上滑雪

Conclusions

- Retrieval-augmentation is a powerful paradigm for V&L
 - Improve models with multimodal encoders
 - Improve lightweight trained models
 - Improve zero-training models
- Take advantage of in-domain resources and large pretrained models

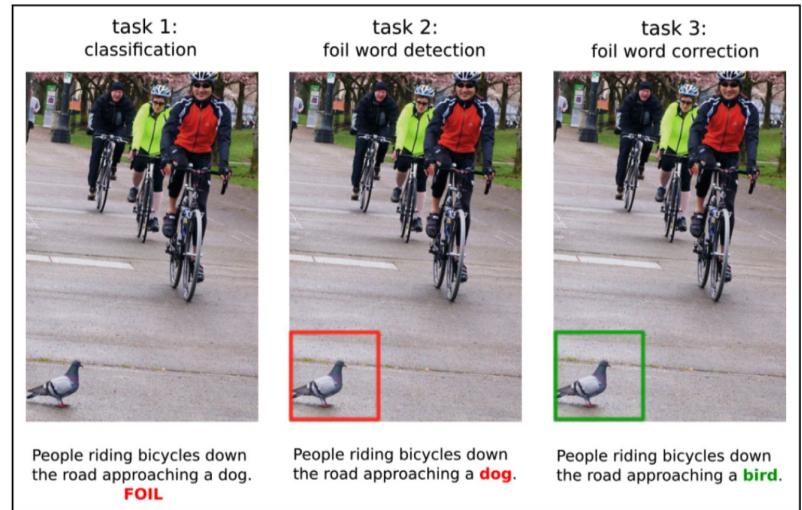
4. Understanding Multimodal Models

Going Beyond Performance

- Many questions about what drives the success of these models?
 - Better contextualization: make better use of the multimodal inputs
 - Acquire certain “skills”, e.g. counting or localization
 - Understand linguistic structures
 - Something else?
- Model-internal behaviour
 - Attention mechanism patterns
 - Mechanistic interpretability (emerging)
- Probing
 - Tasks related to different skills

FOIL Captions

- Do V&L models really understand the relationship between words and images?
- Crowdsource datasets that contain contextually plausible but incorrect image–text pairs

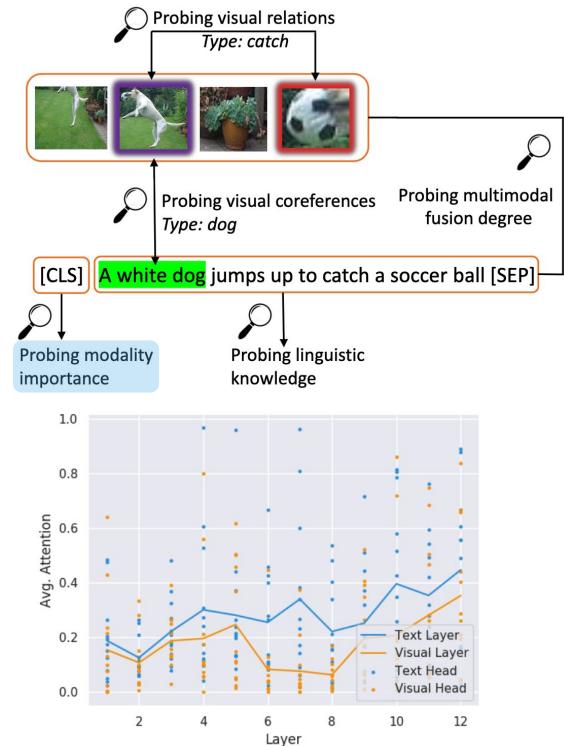


Vision and Language Understanding Evaluation

- Suite of five model probing tasks
- **Modality Influence:** Estimate the layer-wise contribution of each modality to the [CLS] embedding:

$$I_{M,j} = \sum_{i \in S} \mathbb{1}(i \in M) \cdot \alpha_{ij}$$

- The UNITER model relies more on textual features when fusing modalities throughout the model



VALSE Benchmark

- Test visio-linguistic capabilities with image-sentence foil pairs
- Image-sentence matching task
 - Existential quantifiers
 - Semantic number
 - Counting
 - Prepositional relations
 - Action replacement / swap
 - Co-reference



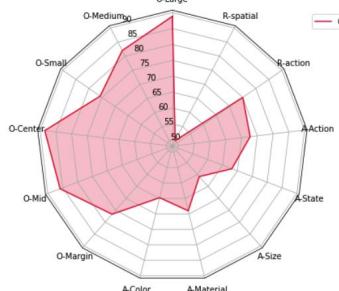
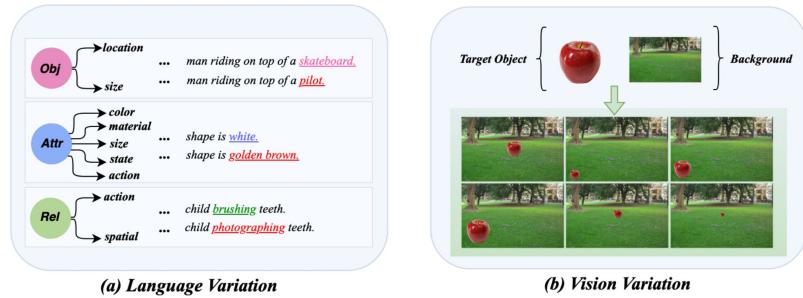
A small copper vase with **some flowers** / **exactly one flower** in it.

Metric	Model	Avg.
	Random	50.0
	GPT1*	60.7
	GPT2*	60.1
	CLIP	64.0
acc_r	LXMERT	59.6
	ViLBERT	63.7
	12-in-1	75.1
	VisualBERT	<u>46.4</u>

$$p(caption, img) > p(foil, img)$$

VL-CheckList

- Evaluate V&L models based on automatic manipulations to vision and language data.
- Image-Sentence matching task
- Radar chart overviews based on object / attribute / relationship variations



Subject-Verb-Object Probes

- Large-scale dataset with SVO triplets mined from Conceptual Captions and 14K images and with crowdsourced captions
- Foil detection formulation



WinoGround

- 1,600 text-image pairs to evaluate compositional understanding



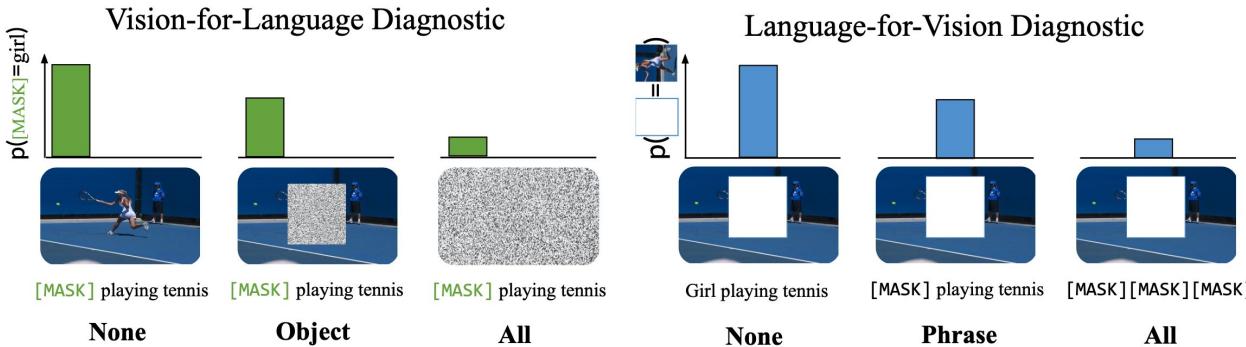
some plants
surrounding a
lightbulb



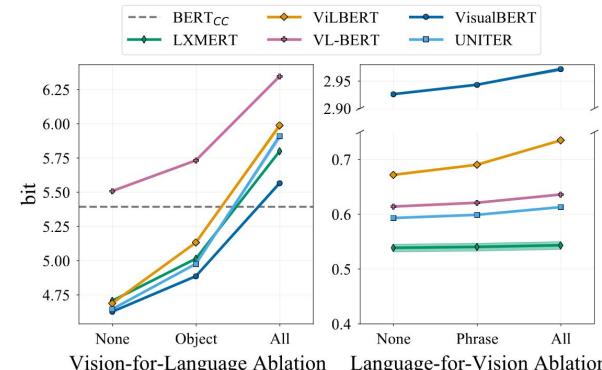
a lightbulb
surrounding
some plants

- Images sourced **with permission** from Getty.
- Differences are categorised into: swap dependent, swap-independent, and visual differences

Vision-for-Language?

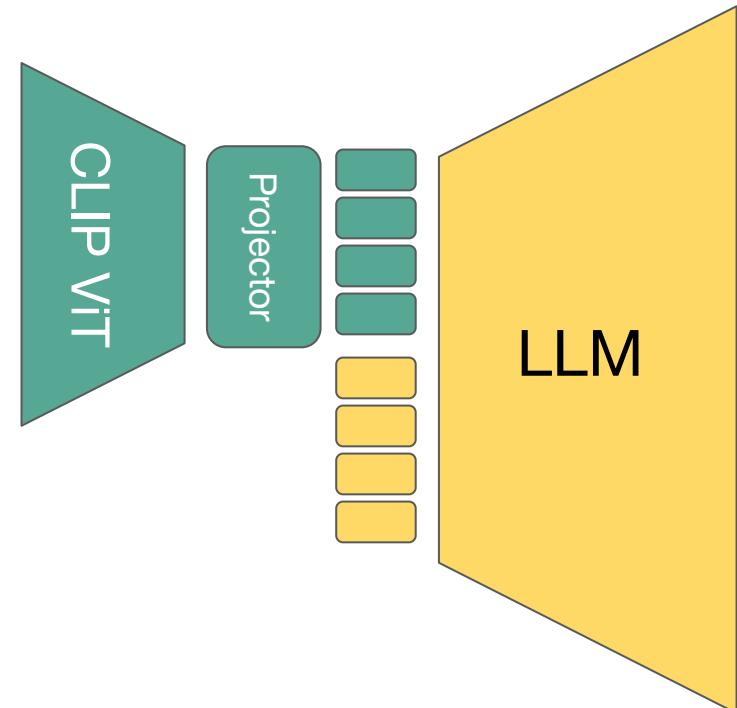
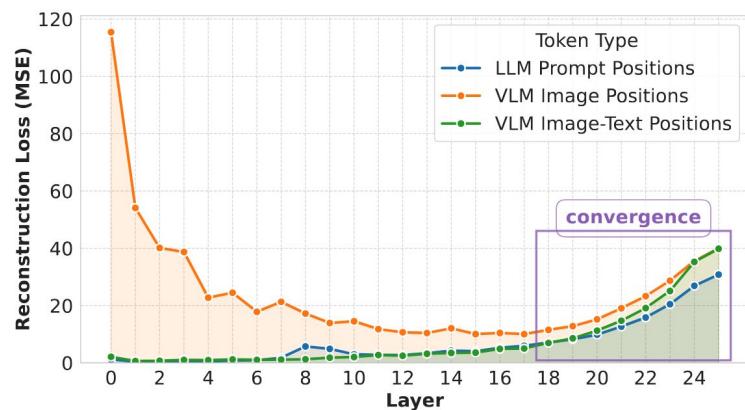


- Pair of diagnostic evaluations that can be applied to any model that makes MLM and MRC predictions.



Understanding the Linear Projector

- The projected image representations are not a good fit for a language model sparse autoencoder until the final layers of the LLM



Seeing What Tastes Good: Revisiting Multimodal Distributional Semantics in the Billion Parameter Era

Findings of ACL 2025



Dan Oneata



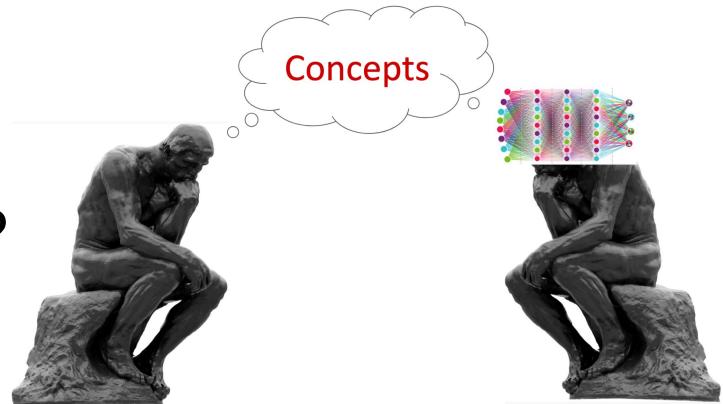
Desmond Elliott



Stella Frank

Main Question & Take-away

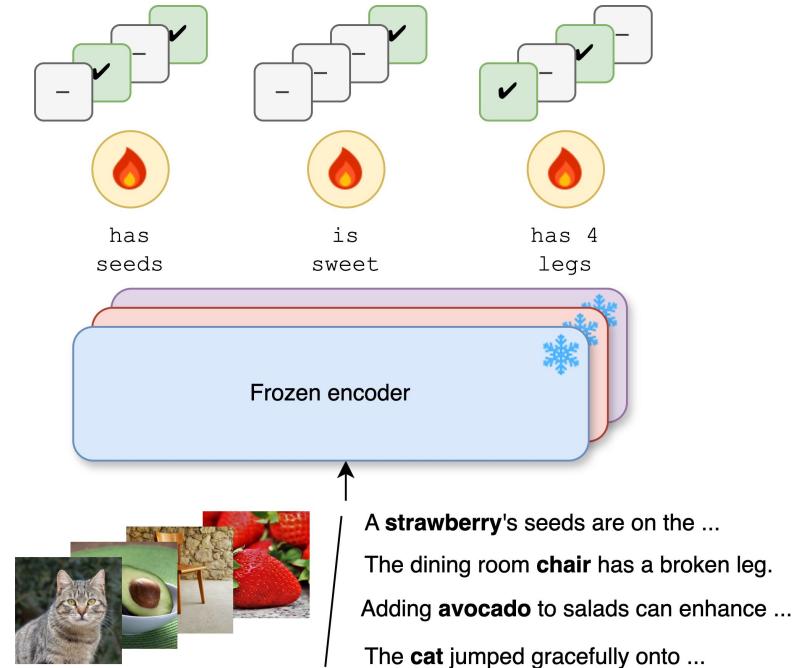
- How well do large-scale pretrained models represent the *semantic attributes* of concepts?
 - ROSE is red, smells sweet, is a flower



Self-supervised vision models are **surprisingly good** at this task

Approach

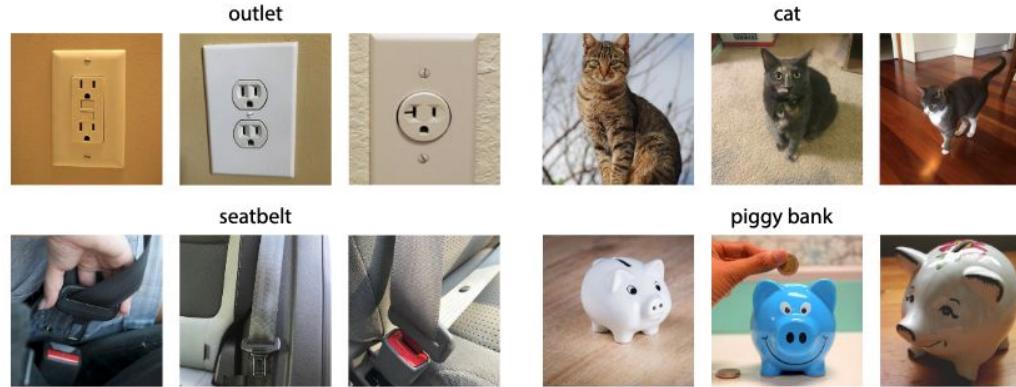
- Train linear probes to predict the semantic attributes of concepts from frozen encoder representations
- Hard generalization task evaluating on *unseen* concepts
 - Train: **Cat**, **Dog**, **Cow** → has_four_legs
 - Test: **Table** → has_four_legs



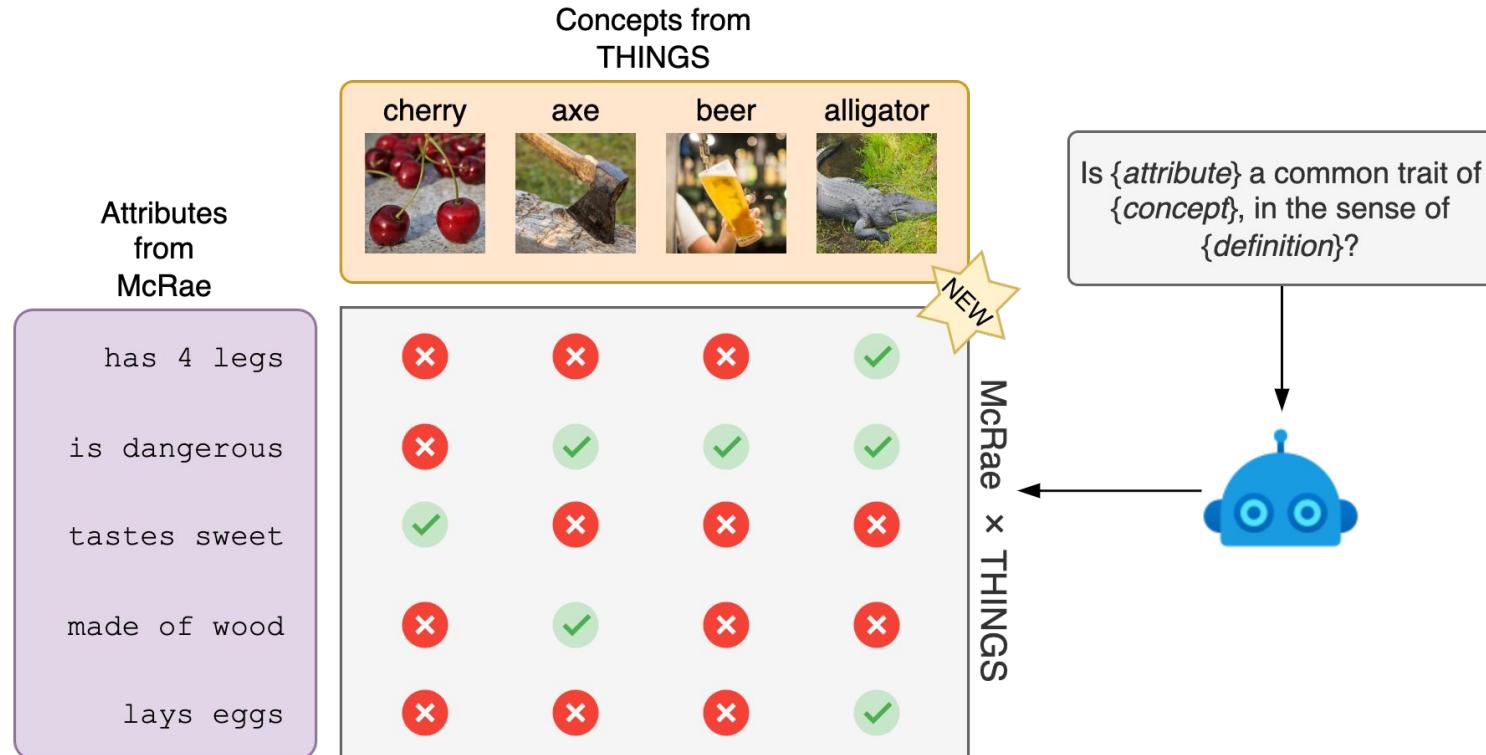
Concepts: THINGS

THINGS dataset of English-labeled concepts with curated images

- 1,854 concepts; 26,000 images



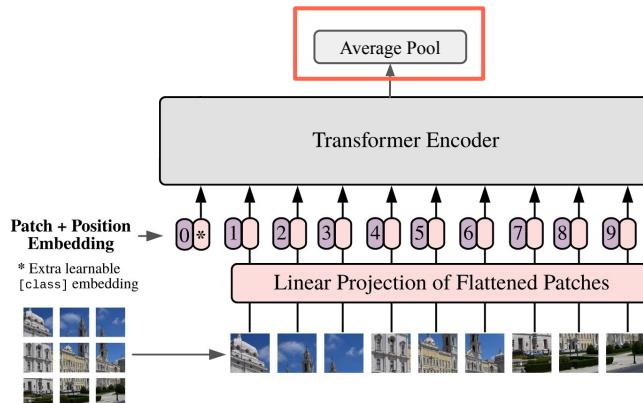
Semantic Attributes: McRae x THINGS



Extracting Concept Representations

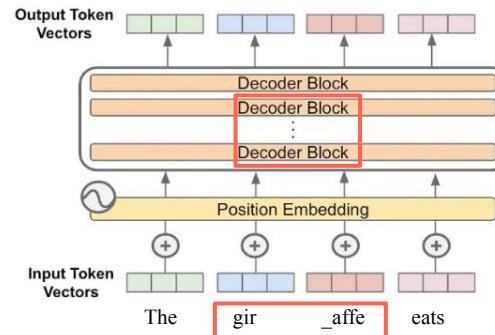
Visual Models

- Mean of the average pooled representation of the concept images

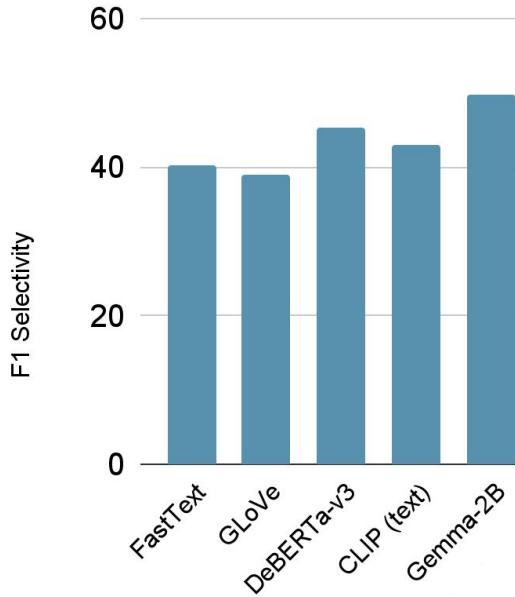


Language Models

- Model-specific token-based pooling over multiple Transformer layers

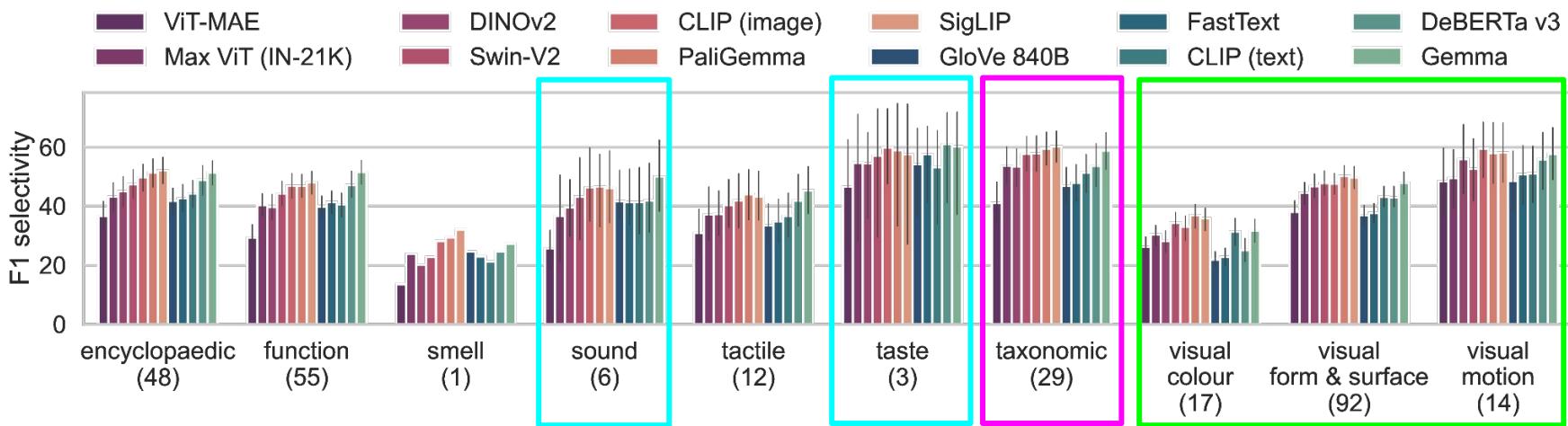


Main Results



Evaluation Measure: **F1-selectivity** (Hewitt & Liang, 2019) above random baseline

Attribute Type Analysis



- Taxonomic (is a) clearly the easiest attribute for all representations
- Sound and Taste are better predicted by language-only representations
- Visual attributes are mostly better predicted by vision representations

Analysis: strong within-modality correlations

Pairwise Pearson correlations of attribute F1

	Random SigLIP	.83	.80	.73	.72	.73	.77	.73	.70	.72	.71	.75	.74	.72	.76	.71	.73	
ViT-MAE	.83	1.0	.80	.89	.94	.92	.93	.91	.89	.91	.89	.85	.84	.84	.90	.85	.86	
Max ViT (IN-1K)	.80	.80	1.0	.82	.80	.81	.81	.81	.77	.80	.79	.77	.76	.75	.77	.72	.74	
Max ViT (IN-21K)	.73	.89	.82	1.0	.92	.96	.93	.93	.92	.93	.92	.85	.85	.86	.89	.87	.88	
DINOv2	.72	.94	.80	.92	1.0	.95	.95	.95	.94	.96	.94	.87	.86	.87	.91	.86	.88	
Swin-V2	.73	.92	.81	.96	.95	1.0	.94	.95	.94	.96	.95	.86	.86	.87	.90	.88	.91	
LLaVA-1.5	.77	.93	.81	.93	.95	.94	1.0	.95	.96	.95	.96	.91	.91	.92	.93	.91	.92	
Qwen2.5-VL	.73	.91	.81	.93	.95	.95	.95	1.0	.96	.96	.96	.89	.89	.90	.92	.90	.91	
CLIP (image)	.70	.89	.77	.92	.94	.94	.96	.96	.96	1.0	.96	.96	.89	.89	.91	.91	.90	.92
PaliGemma	.72	.91	.80	.93	.96	.96	.95	.96	.96	.96	1.0	.98	.87	.88	.88	.92	.89	.91
SigLIP	.71	.89	.79	.92	.94	.95	.96	.96	.96	.98	1.0	.90	.90	.91	.91	.90	.93	
GloVe 840B	.75	.85	.77	.85	.87	.86	.91	.89	.89	.87	.90	1.0	.97	.96	.91	.93	.92	
FastText	.74	.84	.76	.85	.86	.86	.91	.89	.89	.88	.90	.97	1.0	.96	.91	.92	.93	
Numberbatch	.72	.84	.75	.86	.87	.87	.92	.90	.91	.88	.91	.96	.96	1.0	.91	.94	.94	
CLIP (text)	.76	.90	.77	.89	.91	.90	.93	.92	.91	.92	.91	.91	.91	.91	1.0	.90	.92	
DeBERTa v3	.71	.85	.72	.87	.86	.88	.91	.90	.90	.89	.90	.93	.92	.94	.90	1.0	.95	
Gemma	.73	.86	.74	.88	.88	.91	.92	.91	.92	.91	.93	.92	.93	.94	.92	.95	1.0	

Conclusions

As measured by linear probing for attributes:

- Multimodality still better than single modalities
- Differences between modalities are subtle,
given training on enough data.
- Large SSL vision encoders learn conceptual knowledge!
 - Convergence a la Platonic Hypothesis (Huh++2024)?
 - or confounds & correlations (Malt & Smith1994; Lampert++2009)?



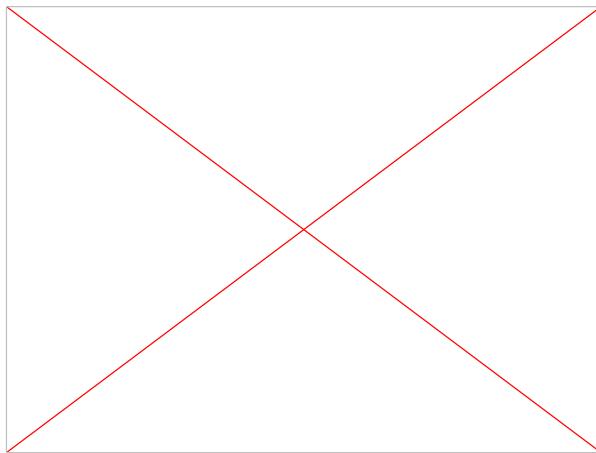
Summary

- Understanding and analysis is a vibrant area of research
- Foil detection is the most popular methodology
- Witnessing a methodological shift
 - attention analyses → linguistically-informed analyses
 - hand-crafted datasets
 - simpler accuracy-based metrics

5. Future Directions

Physical Understanding

- Predicting and explaining physical actions in the world will become of increasing importance as we create embodied agents



Q: How many objects are prevented by the tiny green triangle from falling into the basket?

Q: What is the color of the last object that collided with the tiny red circle?

Q: If any of the other objects are removed, will the tiny green circle end up in the basket?

Text-based Video Games

- Learning to act in procedurally-generated video game environments with rich contexts, action spaces, and long-term rewards

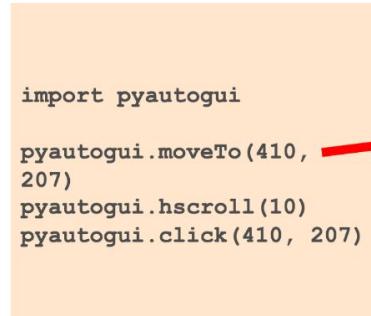


Multimodal Agents

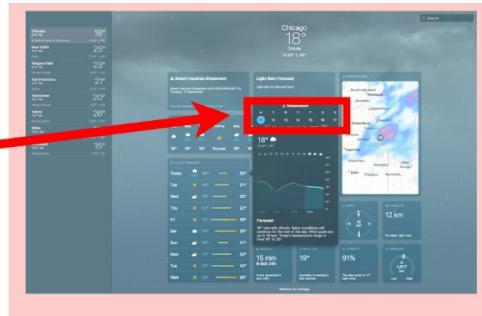
- OmniAct combines multimodal understand with program execution to solve a variety of tasks that humans perform with their computers



Using the popup opened, scroll over to find weather in Chicago on 18th September

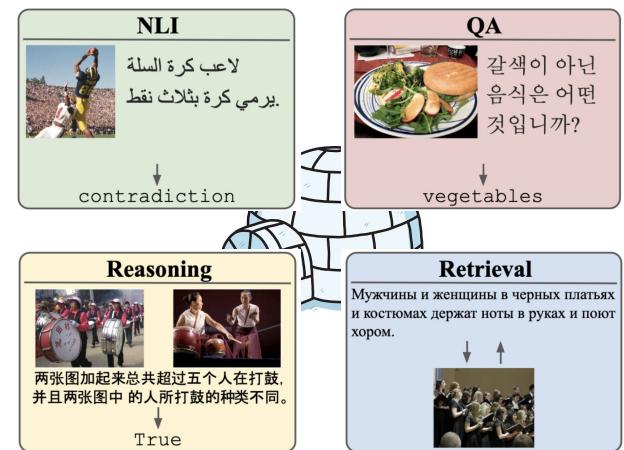


```
import pyautogui  
  
pyautogui.moveTo(410, 207)  
pyautogui.hscroll(10)  
pyautogui.click(410, 207)
```



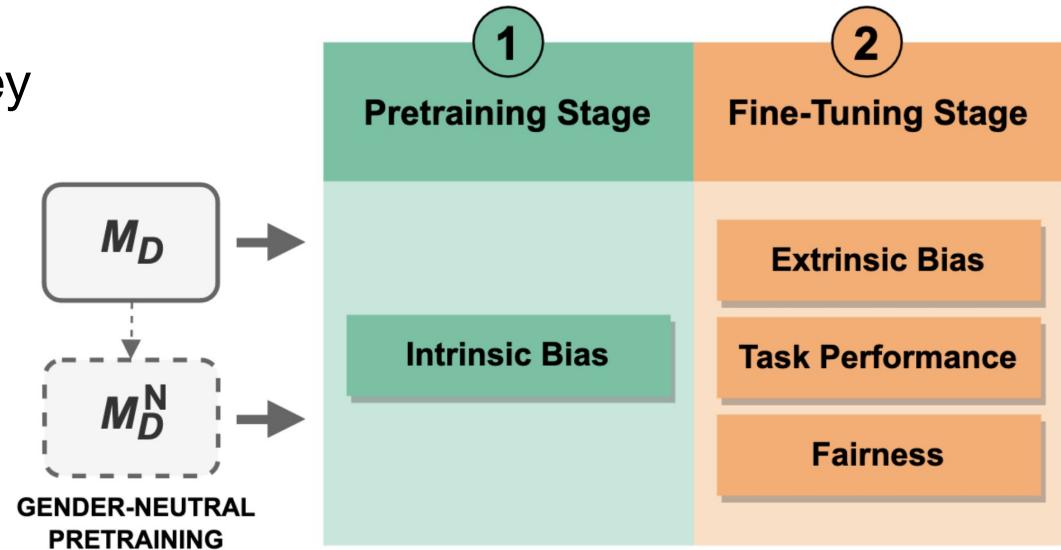
Multilinguality

- The majority of Vision and Language research is in English
- We need resources, models, and evaluations to create useful multilingual multimodal models
- High-quality data requires:
 - time
 - money
 - community engagement



Bias and Fairness

- What are the intrinsic biases learned during multimodal pretraining and how do they affect downstream task performance?



Q: What if we treated language as vision?

Language Modelling with Pixels

ICLR 2023



P. Rust



J. F. Lotz



E. Bugliarello



E. Salesky



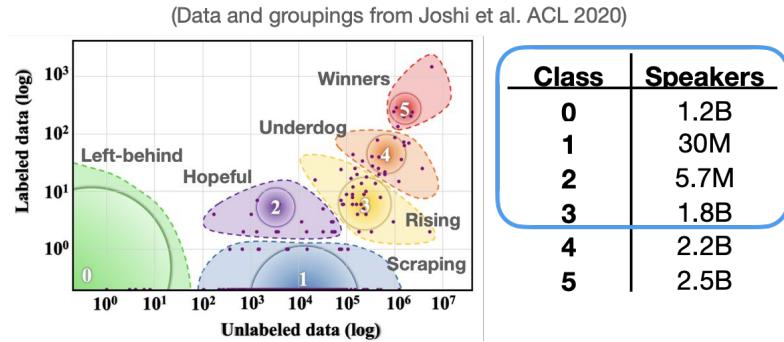
M. de Lhoneux

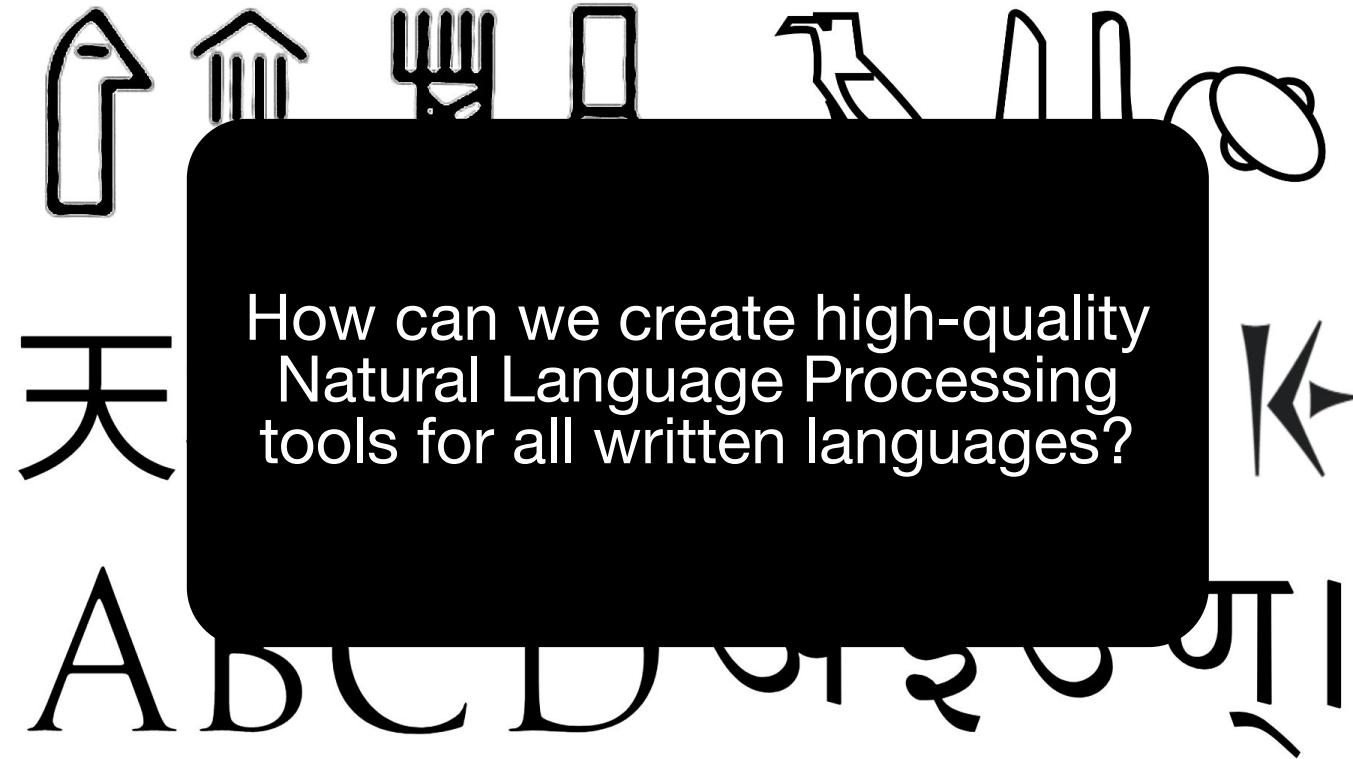


D. Elliott

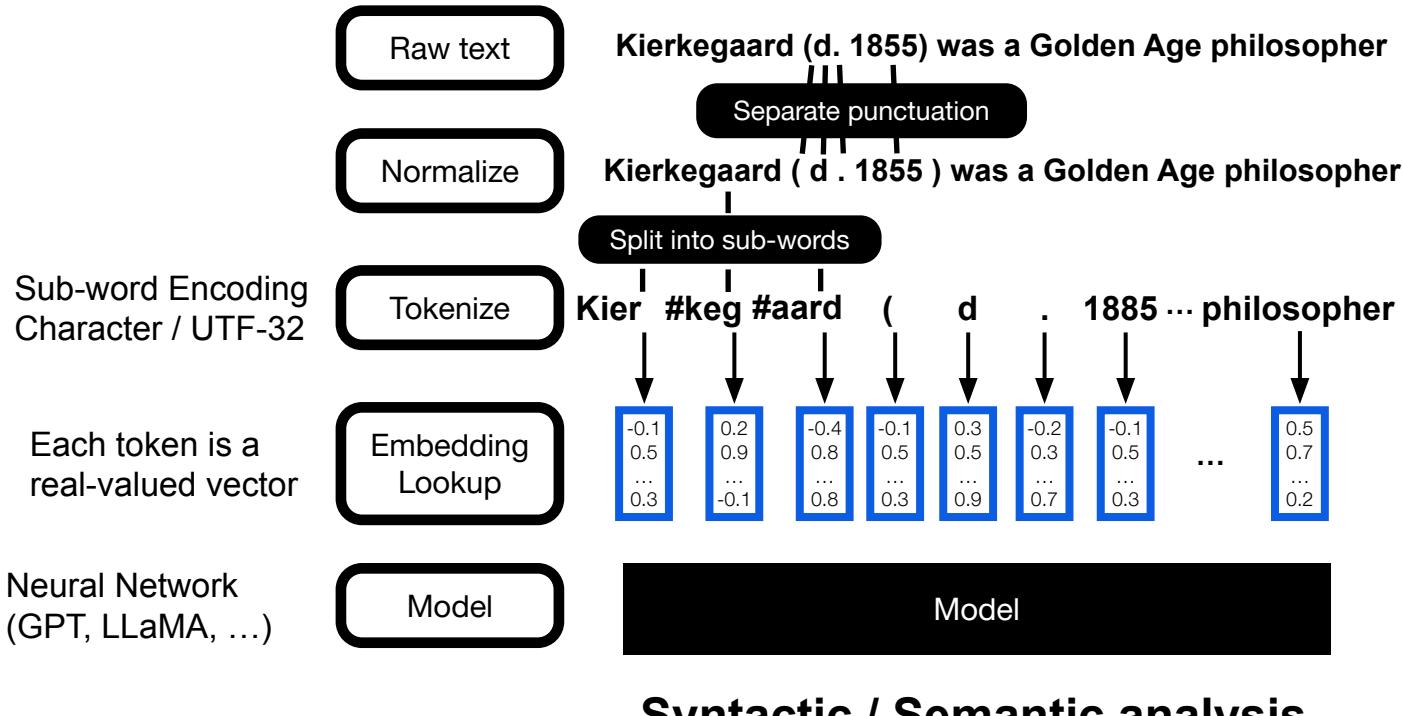
NLP for All Written Languages

- There are 3,000 written languages
 - 400 with >1M speakers
- NLP usually covers 100 languages
 - Technological exclusion for billions
- We need systems for all languages, not just those that are high-resource

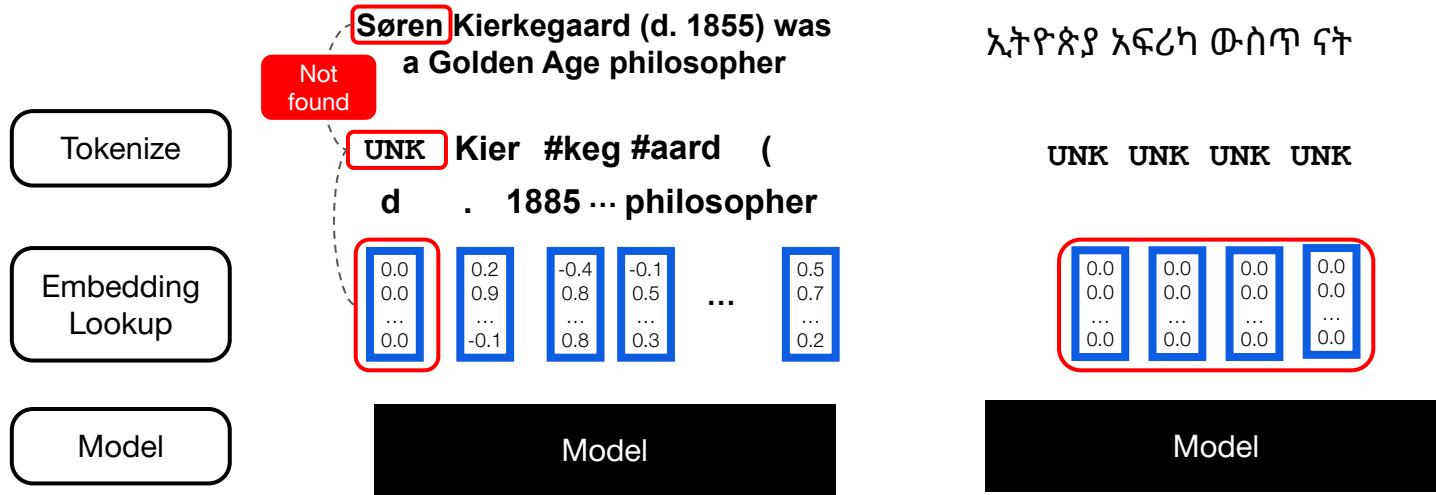




NLP is a pipeline ...



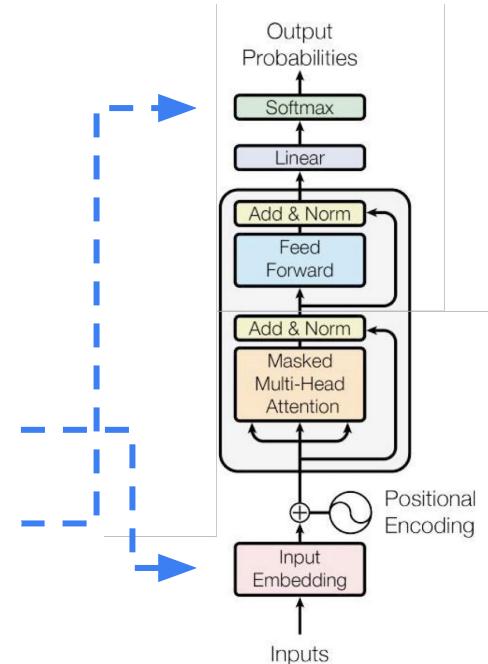
... that is easily broken



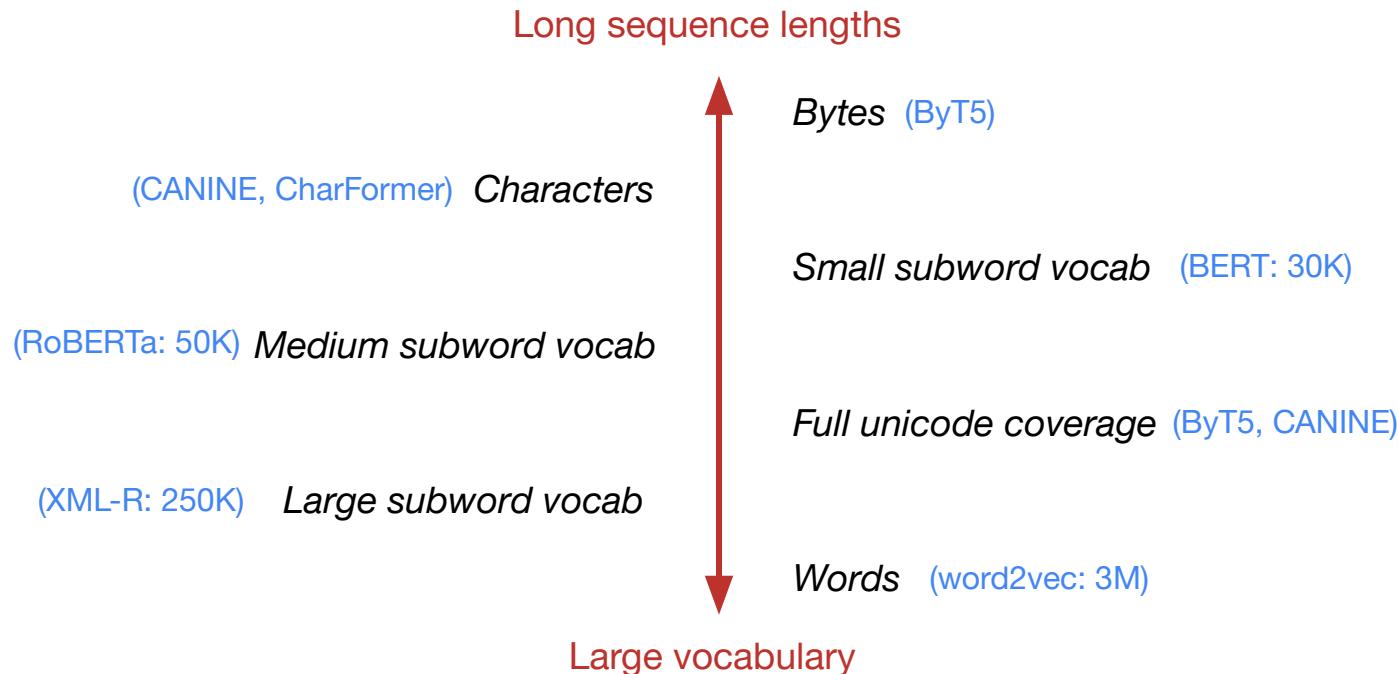
This issue disproportionately affects low-resource languages

The Vocabulary Bottleneck

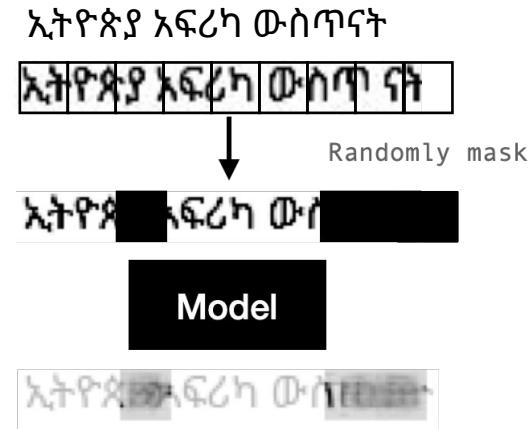
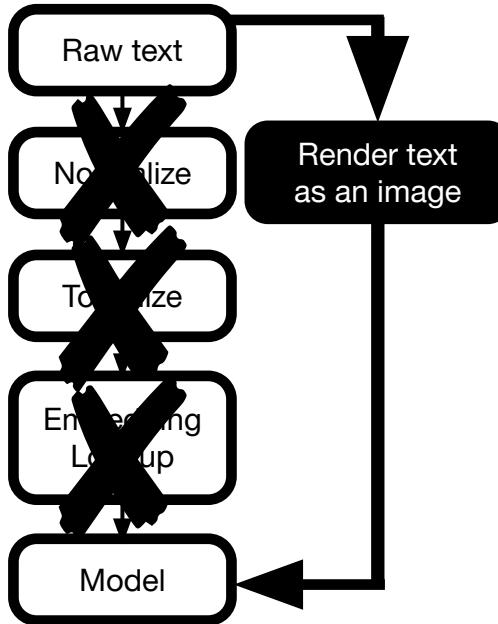
- NLP is an **open vocabulary problem** and the ability of a model is determined by its vocabulary:
 1. tokens, characters, sub-words, etc.
- This creates a bottleneck in two places:
 1. *Representational bottleneck* in the Embedding layer
 2. *Computational bottleneck* in the Output layer



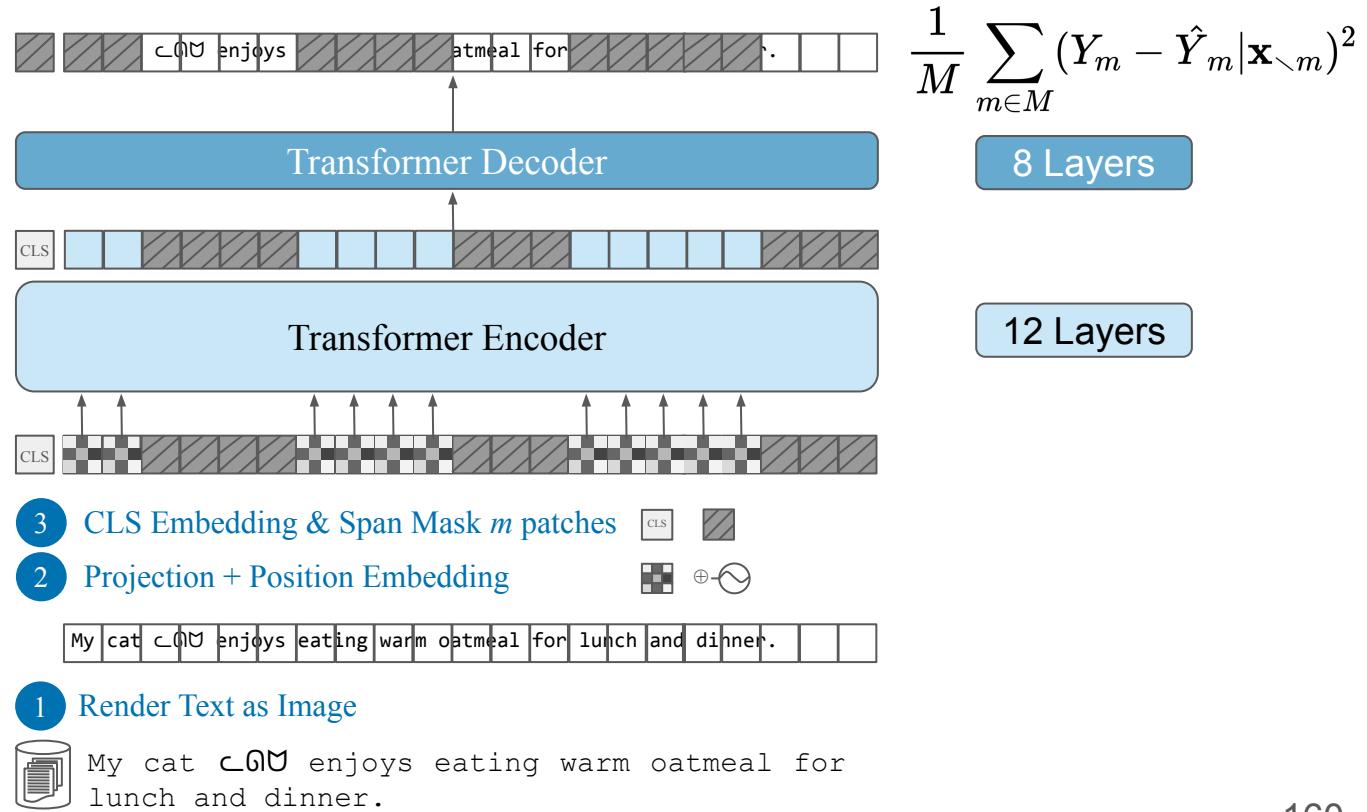
Where's the sweet spot?



Main idea: treat language as vision

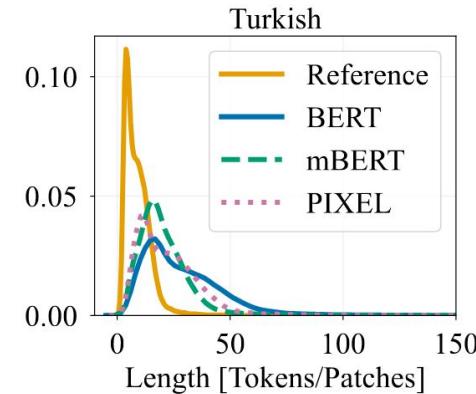


The PIXEL Model



Rendered Text is Compact

- PIXEL encoding produces sequence lengths that are at least as long as BERT.
 - Universal Dependencies datasets with human reference segmentations
 - No length penalty for any language, unlike some LLMs (Ahia et al. 2023)



Proportion of text that is encoded as k subwords / patches.

Pretraining

- **English Dataset:** English Wikipedia and Books Corpus
- **Masking:** 25% Span Masking
- **Maximum sequence length:** 529 patches (16×8464 pixels)
- **Compute:** 8 x 40GB A100 GPUs for 8 days
- **Parameters:** 86M encoder + 26M decoder

There is only 0.05% non-English text in our pretraining data (estimated by Blevins and Zettlemoyer 2022)

The **Great Wall of China** (traditional Chinese: 萬里長城; simplified Chinese: 万里长城; pinyin: Wàn lǐ Chángchéng)

Downstream Tasks

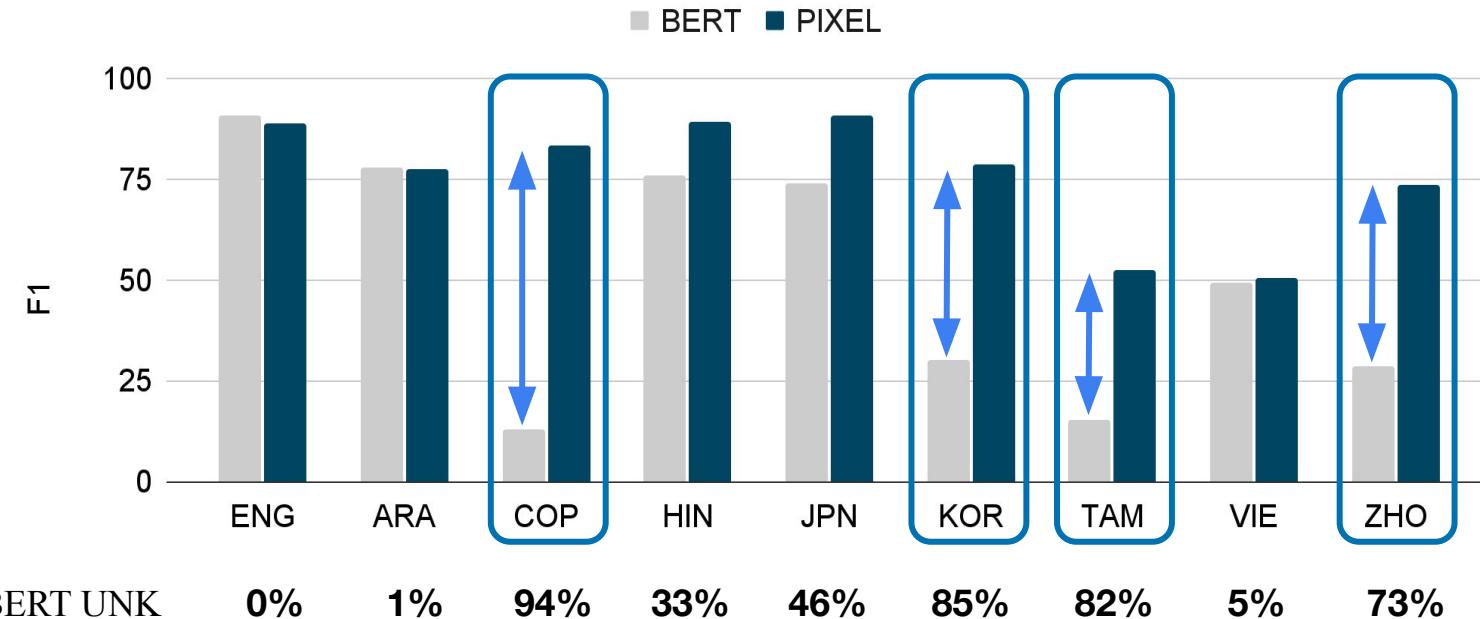
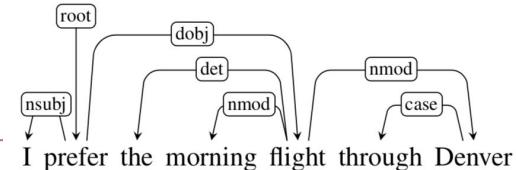
- **Datasets:** Universal Dependencies, MasakhaNER, GLUE, Zeroé
- **Models:**

	Parameters	Pretraining Data
PIXEL _{BASE}	86M	English Wikipedia + Bookcorpus
BERT _{BASE}	110M	—
CANINE-C	127M	104-languages from Wikipedia

Similar pretraining setup

Tries to solve the same
problem using UTF-32

Dependency Parsing Results

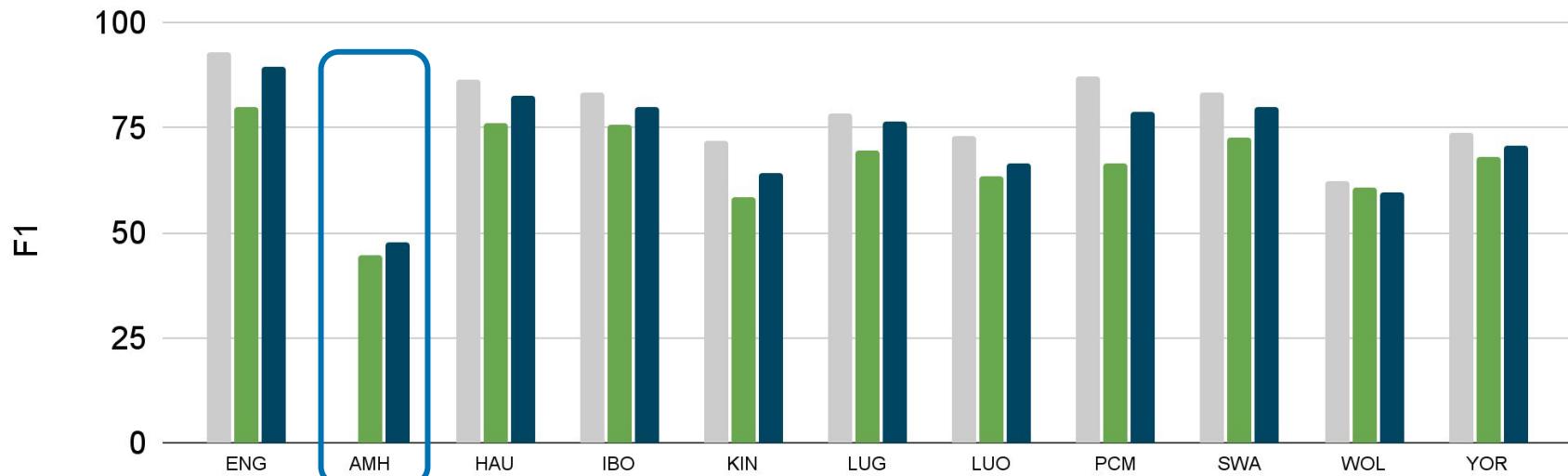


PIXEL vastly outperforms BERT on unseen scripts

Named Entity Recognition

Emir of Kano turban Zhang wey don spend 18 years for Nigeria

BERT CANINE PIXEL

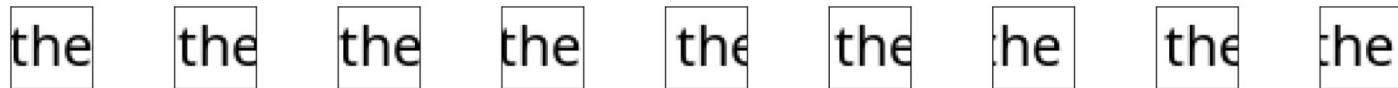


PIXEL outperforms BERT
on the non-Latin script

PIXEL outperforms the
multilingually pretrained CANINE-C

Text Rendering Matters

- The original text renderer produces many nearly-identical patches
 - This is representation- and compute-wasteful



(a) Continuous rendering (CONTINUOUS):

I must be growing small again. ■

(b) Structured rendering (BIGRAMS):

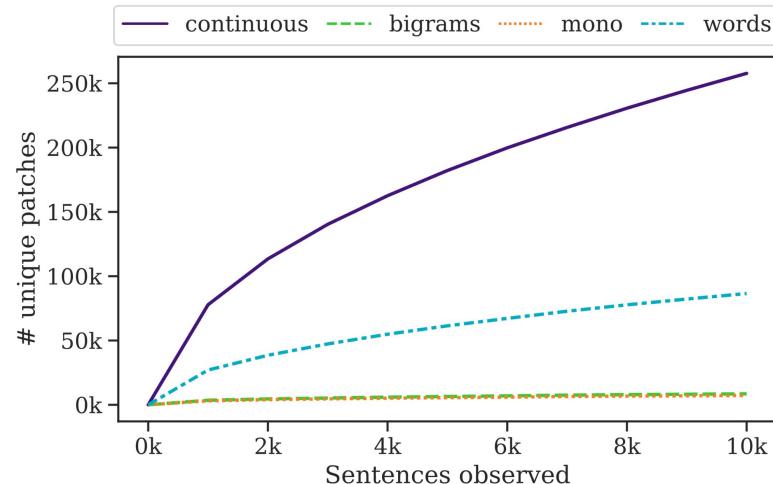
I must be gr ow in g sm al l ag ai n. ■

(c) Structured rendering (MONO):

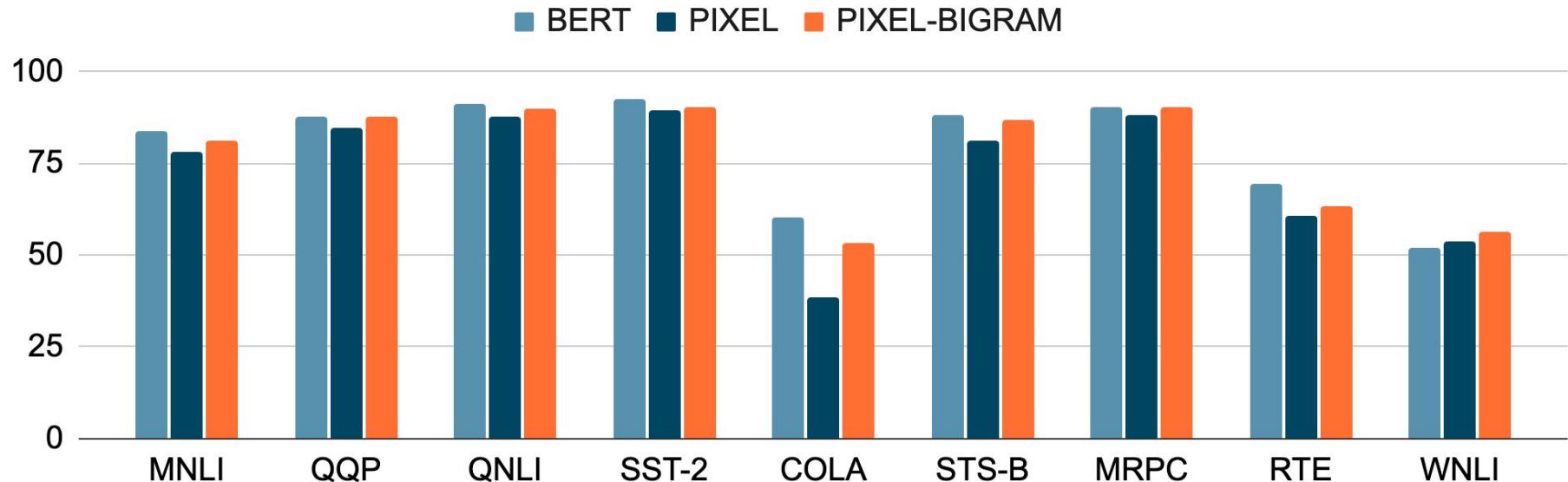
I mu st b e gr owing sm all ag ai n. ■

(d) Structured rendering (WORDS):

I mu st b e gr owing sm all ag ai n. ■



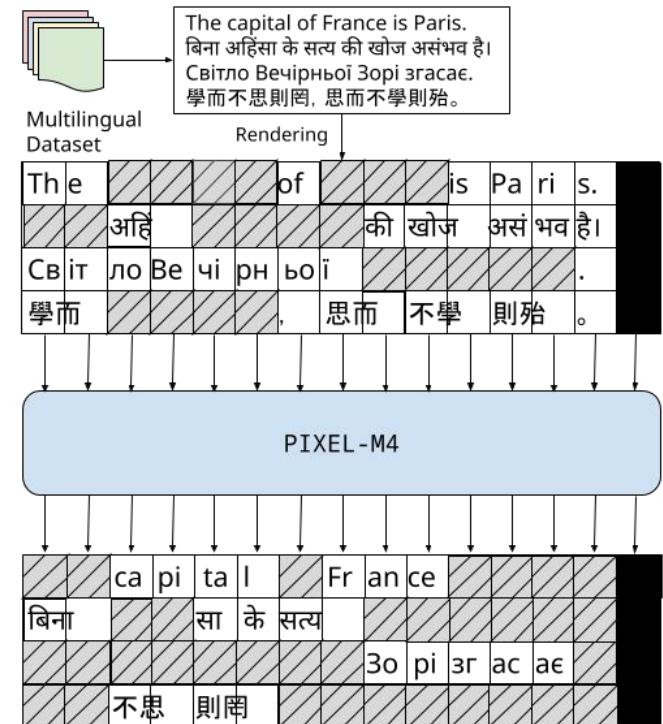
Sentence-level Tasks: GLUE



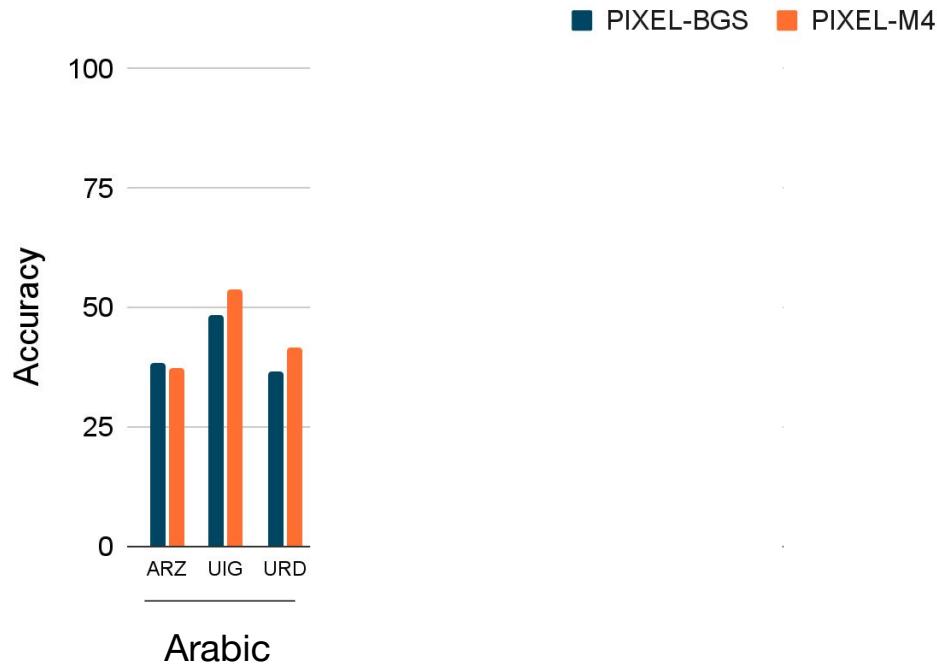
Bigram text rendering produces better models

Going Multilingual: PIXEL-M4

- Same architecture and hyperparameters as PIXEL-BIGRAMS
- But, pretrained on four visually diverse scripts sourced from mC4
 - Latin - English
 - Han - Simplified Chinese
 - Cyrillic - Ukrainian
 - Brahmic - Hindi

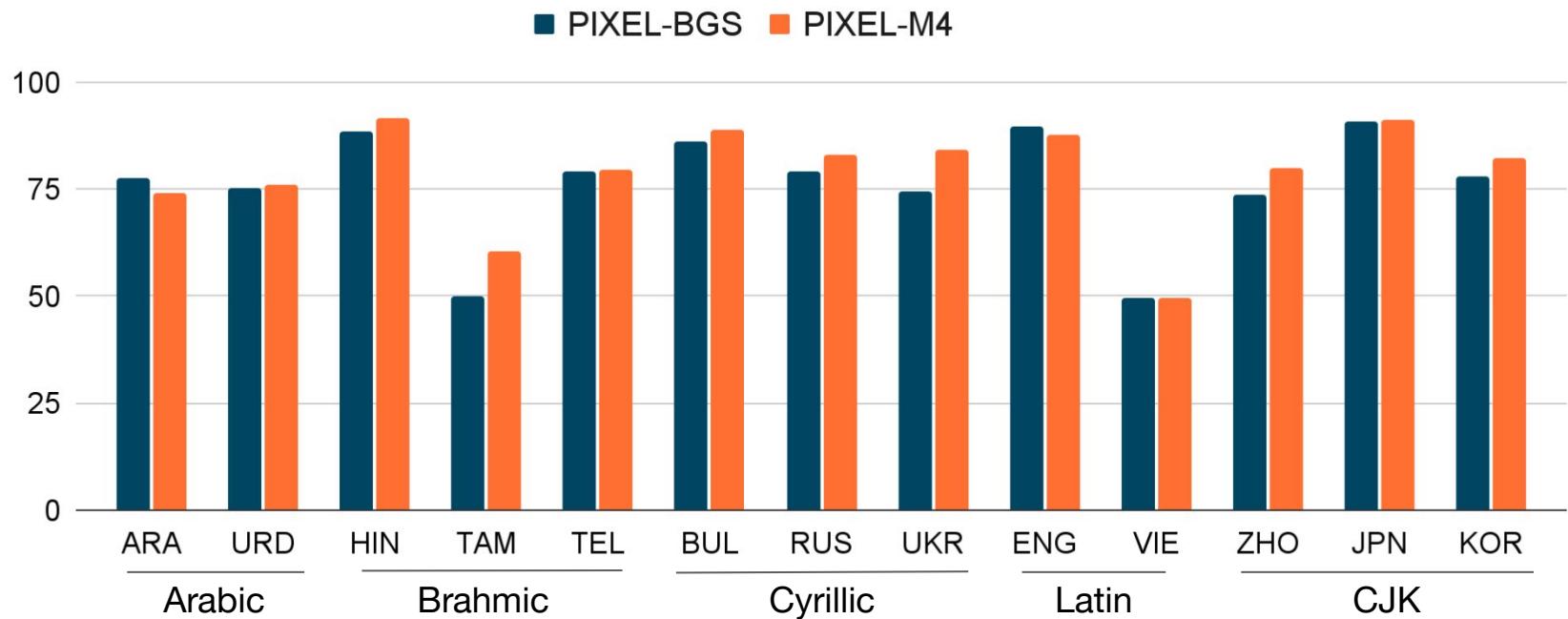


Text Classification on SIB-200



Multilingual pretraining is very helpful for sentence classification

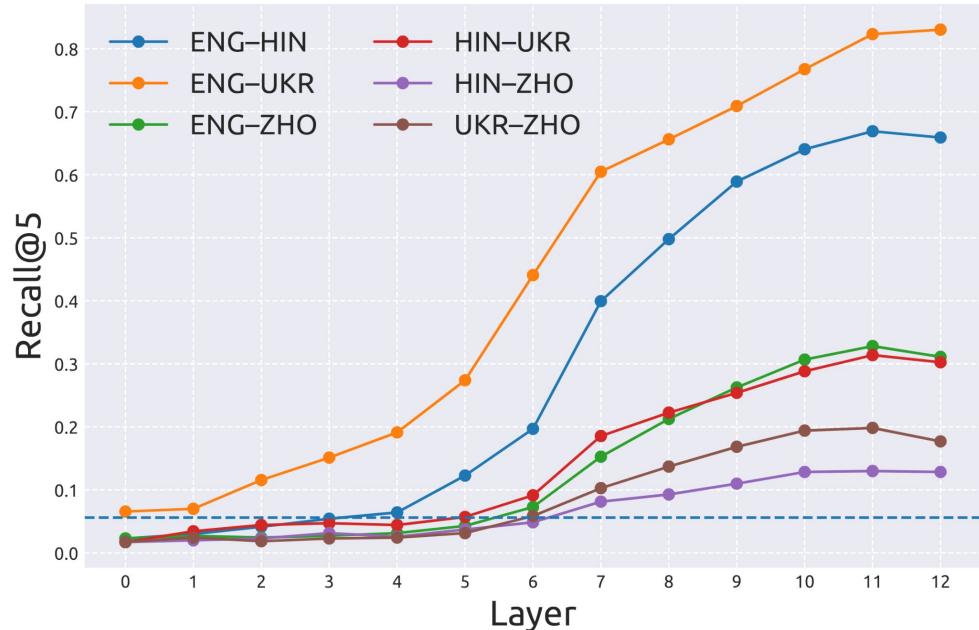
Dependency Parsing



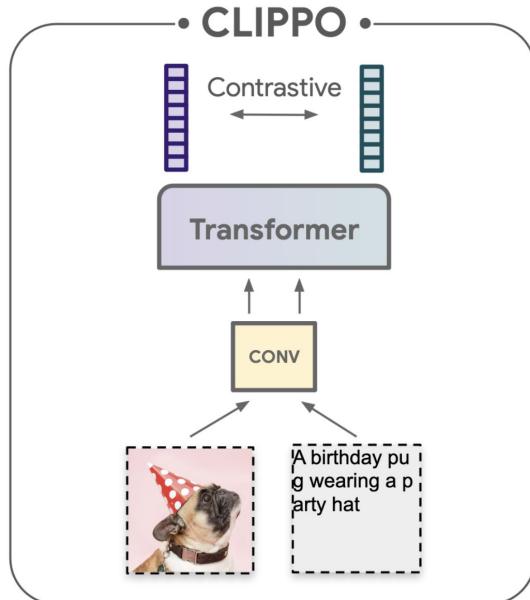
Multilingual pretraining helps for non-Latin script languages

Zero-shot Sentence Retrieval Analysis

- Does multilingual pretraining lead to better “semantic” representation of text?
- Encode the sentences from the SIB-200 dataset in the four pretraining languages
- Measure the cosine similarity between the encoded data

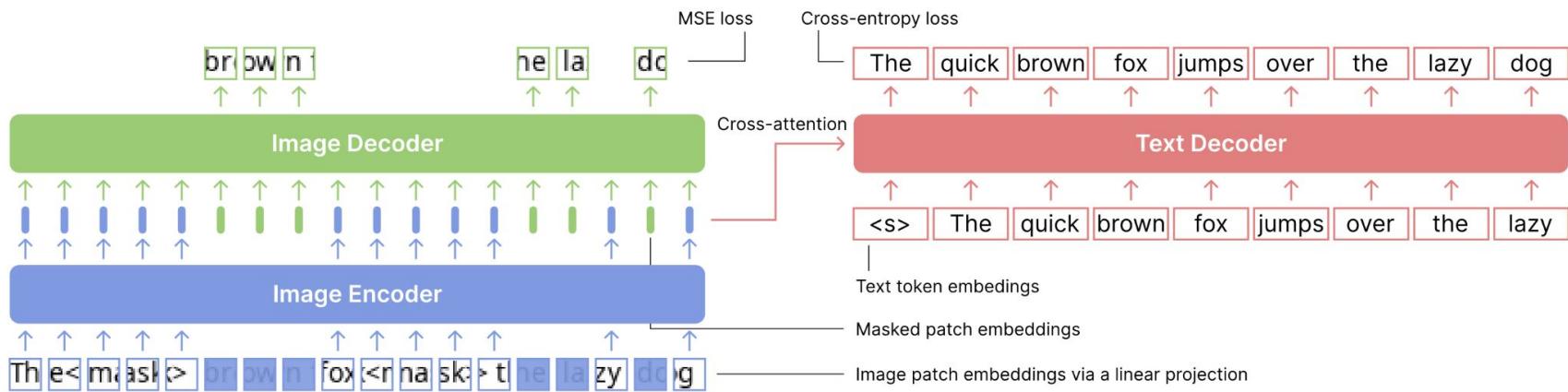


Joint Multimodal Reasoning



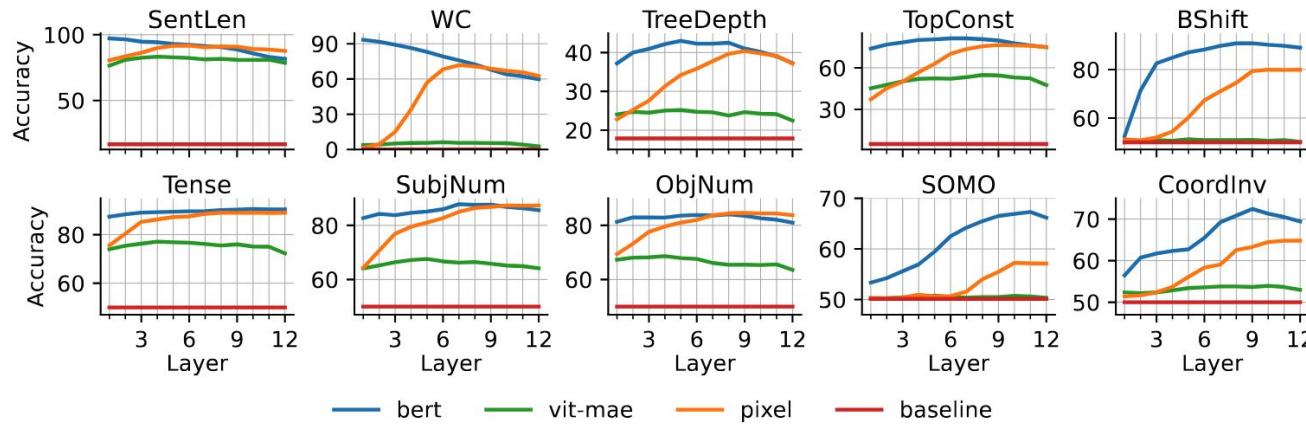
Patch and Text Prediction

- Combine patch and token prediction



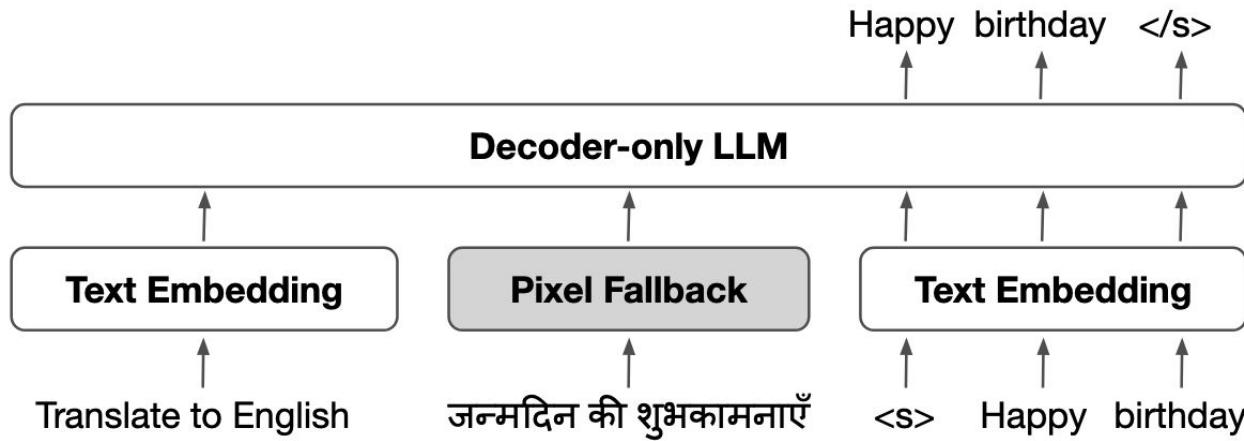
Pixology

- What linguistic knowledge is learned by pixel language models?



Combining Pixels and Tokens

- Handle sub-optimally covered inputs using pixel representations



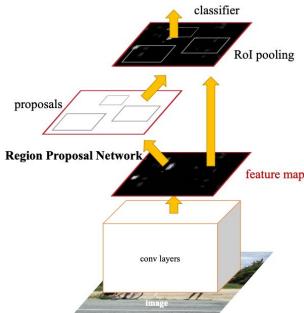
Wrap-up

1. Datasets

some sheep walking in the middle of a road
a herd of sheep with green markings walking down the road
a herd of sheep walking down a street next to a lush green grass covered hillside.
sheared sheep on roadway taken from vehicle, with green hillside in background.
a flock of freshly sheered sheep in the road.



2. Representation

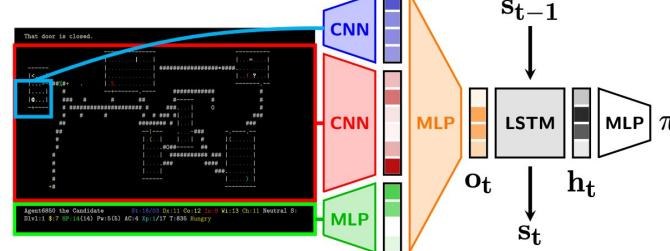
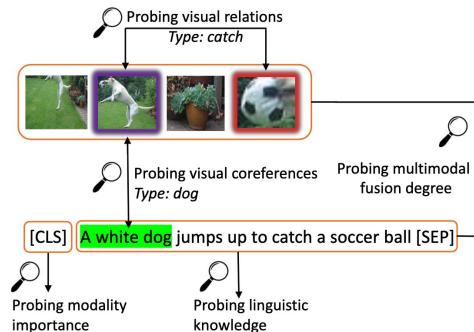


3. Modelling

Language Model

Embed

The red horse



4. Understanding

5. New Directions

Acknowledgements



E. Bugliarello



R. Ramos



A. Kádár



L. Specia



L. Barrault



W. Li



N. Collier



G. Chrupała



C. Qui



D. Oneață



S. Frank



R. Sanabria



M. Fadaee



F. Liu



B. Martins



E. Hasler



E. Ponti



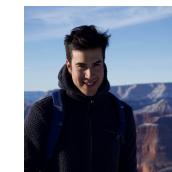
S. Hooker



S. Reddy



A. Alishahi



I. Salazar