Vision and Language

Lisbon Machine Learning Summer School 2025

Desmond Elliott

Department of Computer Science

University of Copenhagen

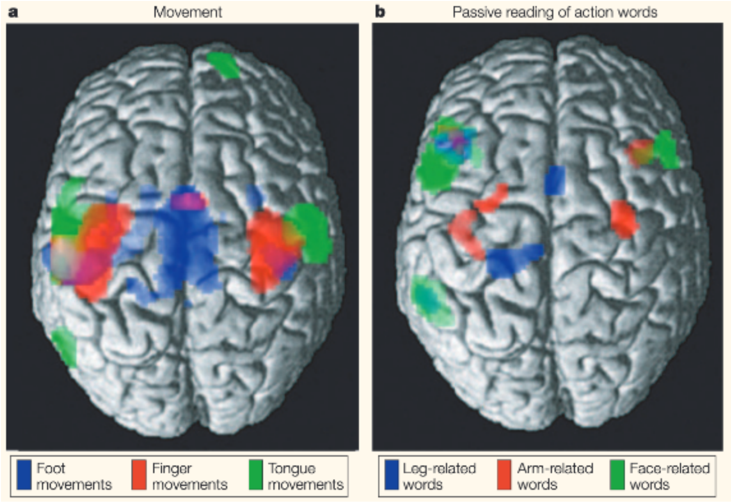
Working Definition

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

Image adapted from https://xkcd.com/1838/ (CC BY-NC 2.5)

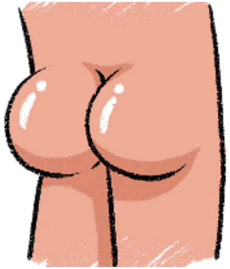
Why Multimodality?

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



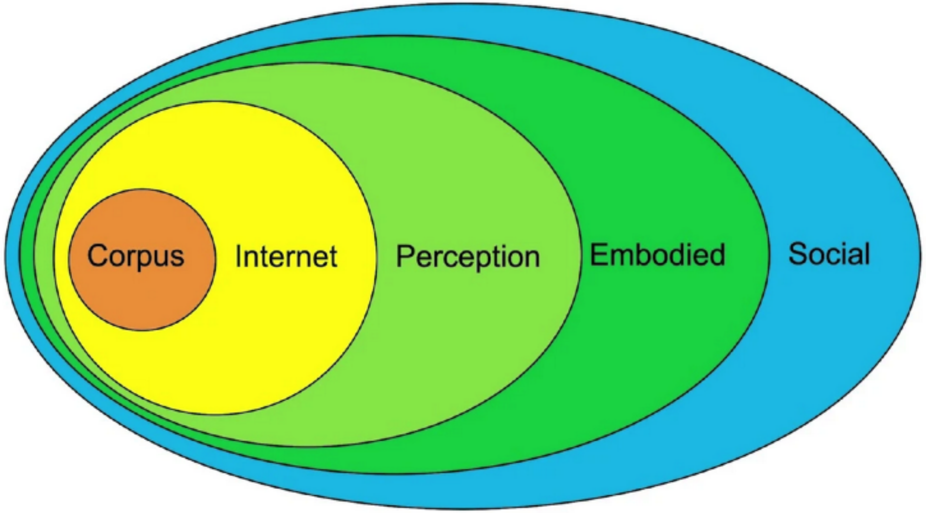
Barsalou et al. (2003). Grounding conceptual knowledge in modality-specific systems. Trends in cognitive sciences, 7(2):84–91.
Pulvermüller. (2005). Brain mechanisms linking language and action. Nature reviews neuroscience, 6(7), 576-582.

Multimodality reduces ambiguity



You Cannot Learn Language From

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



Bisk et al. EMNLP 2020. Experience Grounds Language.

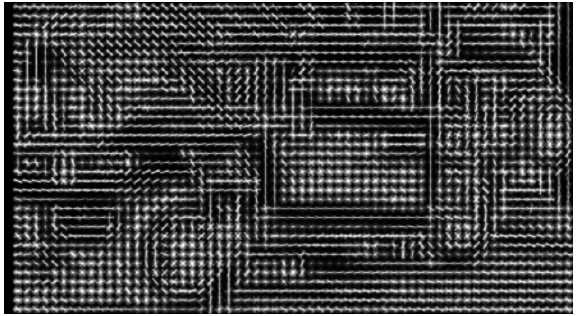
The Five Major Areas

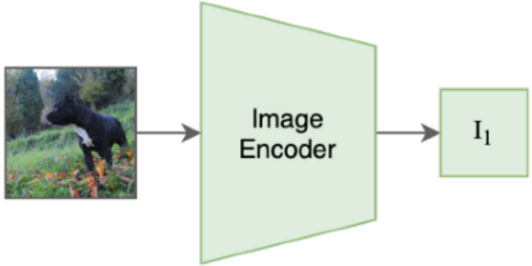
- **Representation**: how to convert raw inputs into a usable format
- **Translation**: transform from one modality to another
- **Alignment**: predict relationships between elements across modalities
- **Fusion**: join features from modalities to support prediction
- **Co-learning**: transferring knowledge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

Baltrušaitis, Ahuja, and Morency, (2018). Multimodal machine learning: A survey and taxonomy. IEEE PAMI, 41(2), 423-443.

Representation

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

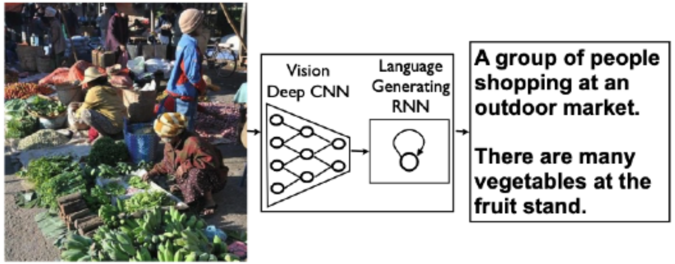


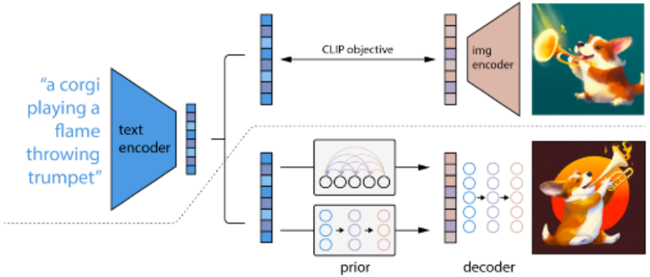


Dalal & Triggs. CVPR 2005. Histograms of oriented gradients for human detection.
Radford et al. ICML 2021. Learning transferable visual models from natural language supervision.

Translation

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

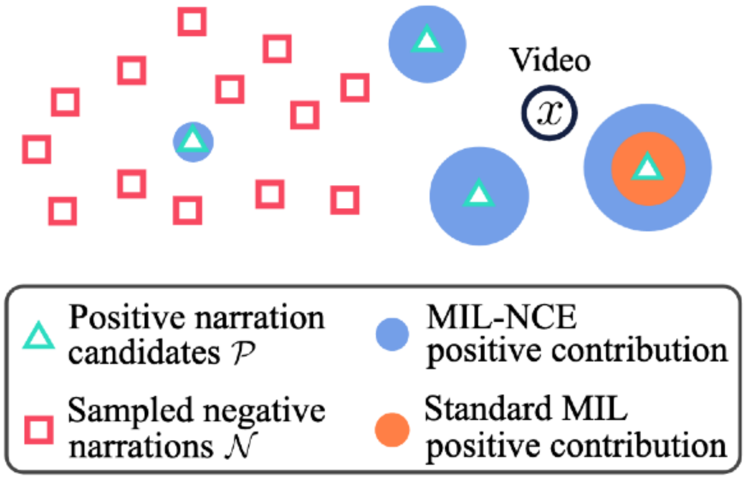




Vinyals et al. (2015). Show and tell: A neural image caption generator. CVPR.
Ramesh et al. (2022). Hierarchical Text-Conditional Image Generation with CLIP Latents. arXiv.

Alignment

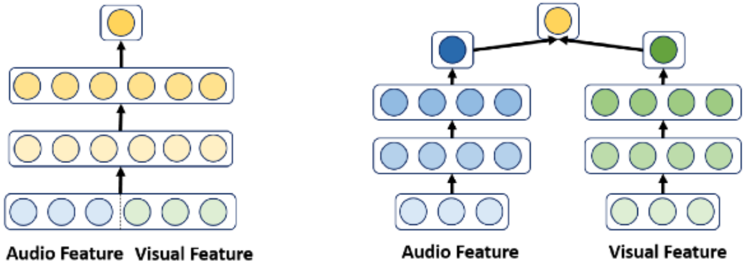
- Important for self-supervised learning and also for phrase grounding

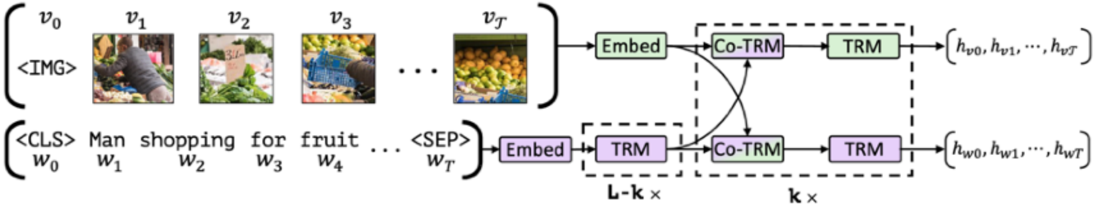


Miech et al. (2020). End-to-end learning of visual representations from uncurated instructional videos. CVPR.
Kamath et al. (2021). MDETR-modulated detection for end-to-end multi-modal understanding. ICCV.

Fusion

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

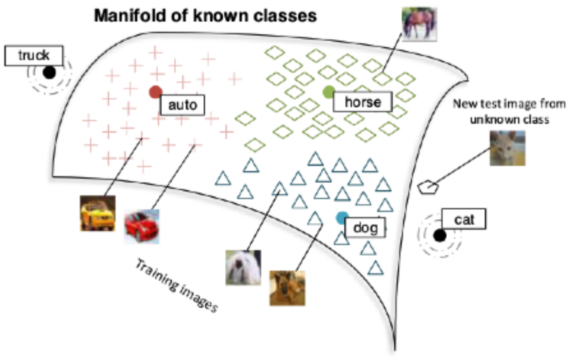


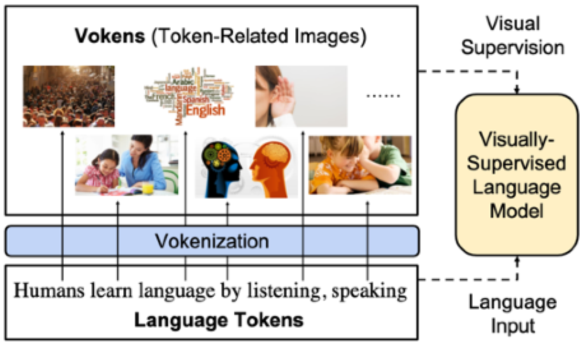


Chen and Jin (2016). Multi-modal conditional attention fusion for dimensional emotion prediction. MM.
Lu et al. (2019). ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *NeurIPS.*

Co-learning

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

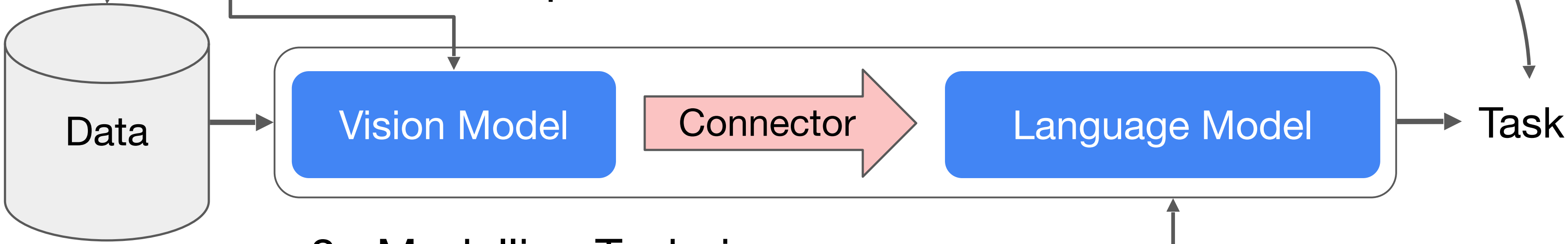




Socher et al. (2013). Zero-shot learning through cross-modal transfer. NeurIPS.
Tan & Bansal. (2020). Vokenization: Improving Language Understanding with Contextualized, Visual-Grounded Supervision. EMNLP

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
   📚 A Multimodal Spanish Dataset for Sexism Detection in Social Media Videos

1. Datasets and Tasks for Multimodal Learning

There Are No New Ideas in AI… Only New Datasets

LLMs were invented in four major developments... all of which were datasets

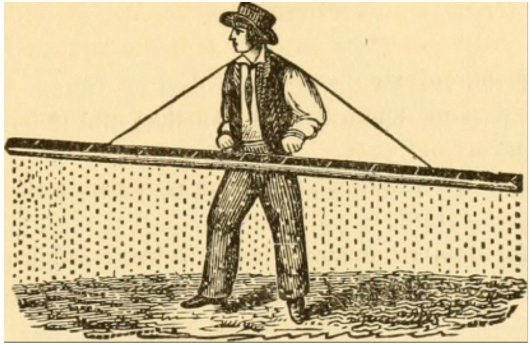
JACK MORRIS

APR 09, 2025

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)

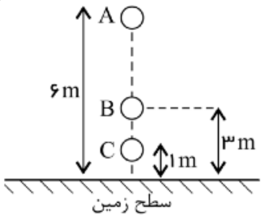
(b) *Görsellerden birinde dizlerinde kanun bulunan birden çok insan var.* ("In one of the images, there are multiple people with qanuns on their knees.", concept: *Kanun (çalgı)* (QANUN, a popular instrument in Turkey), label: TRUE)

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:
گلوله‌ای مسیری مطابق شکل را طی می کند. کار نیروی وزن از A تا B چند برابر کار نیروی وزن از B 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: $\textbf{Distance}(x, v)$
  - Image↔Text Retrieval

    Referring Expression Localization

$P(x|v)$

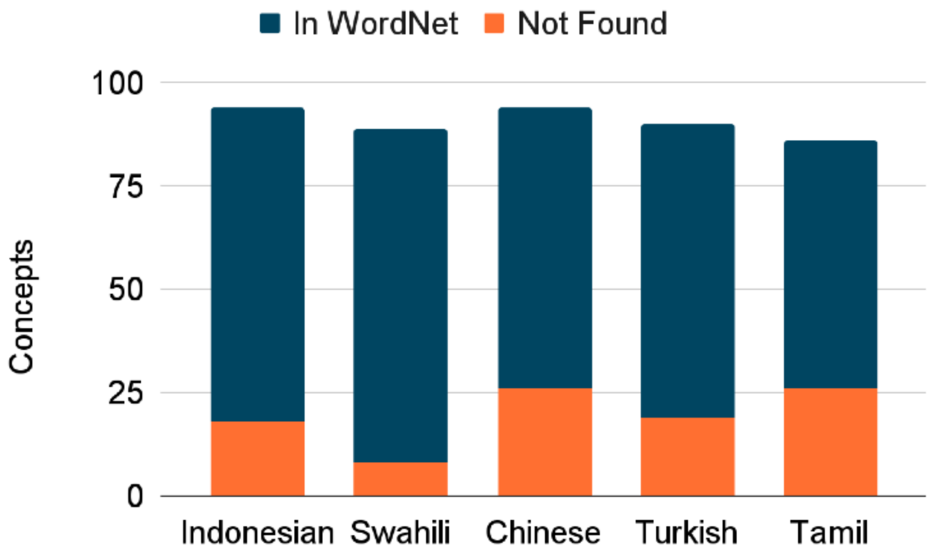
Multi30K: Replicate

A brown dog is running after the black dog.

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

$P(y|x, v)$

MaRVL: Concepts beyond ImageNet



$P(y|x,v)$

FoodieQA: Fine-grained Chinese Food

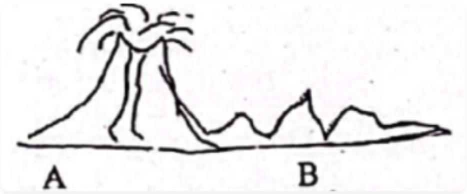
$P(y|v)$

Kaleidoscope: Exams

Geography

Question #5357

'A' চিত্রের প্রদর্শিত ঘটনাটি কী?



1) সঞ্চয়

2) অগ্ন্যুৎপাত

3) ক্ষয়

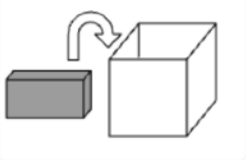
4) ভঙ্গিল

Social Sciences | Bengali | Bangladesh

Mathematics

Question #9326

Alice a beaucoup de blocs, tous de dimensions 1 cm x 2 cm x 4 cm. Elle en met le plus possible dans une boîte de dimensions 4 cm x 4 cm x 4 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?

Hired	Volunteers	Collaborators	Indirect
CulturalVQA			
MaRVL			
FoodieQA			
Multi30K		Kaleidoscope	LAION

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 ....

Source: machine translated extract from crowdworker discussion forum about our crowdsourcing task.

Multi30K: Speculations

#: 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.*

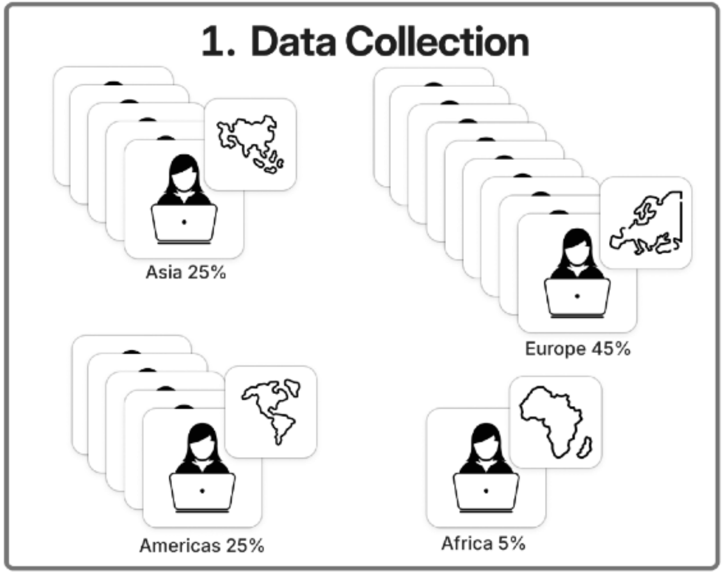
Yes good chance when I look so what he has been previously employed. I find something fascinating 😬

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

Source: machine translated extract from crowdworker discussion forum about our crowdsourcing task.

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?

Kaleidoscope — Not Photos

Multi30K
MaRVL
CulturalVQA
— Photos

Types

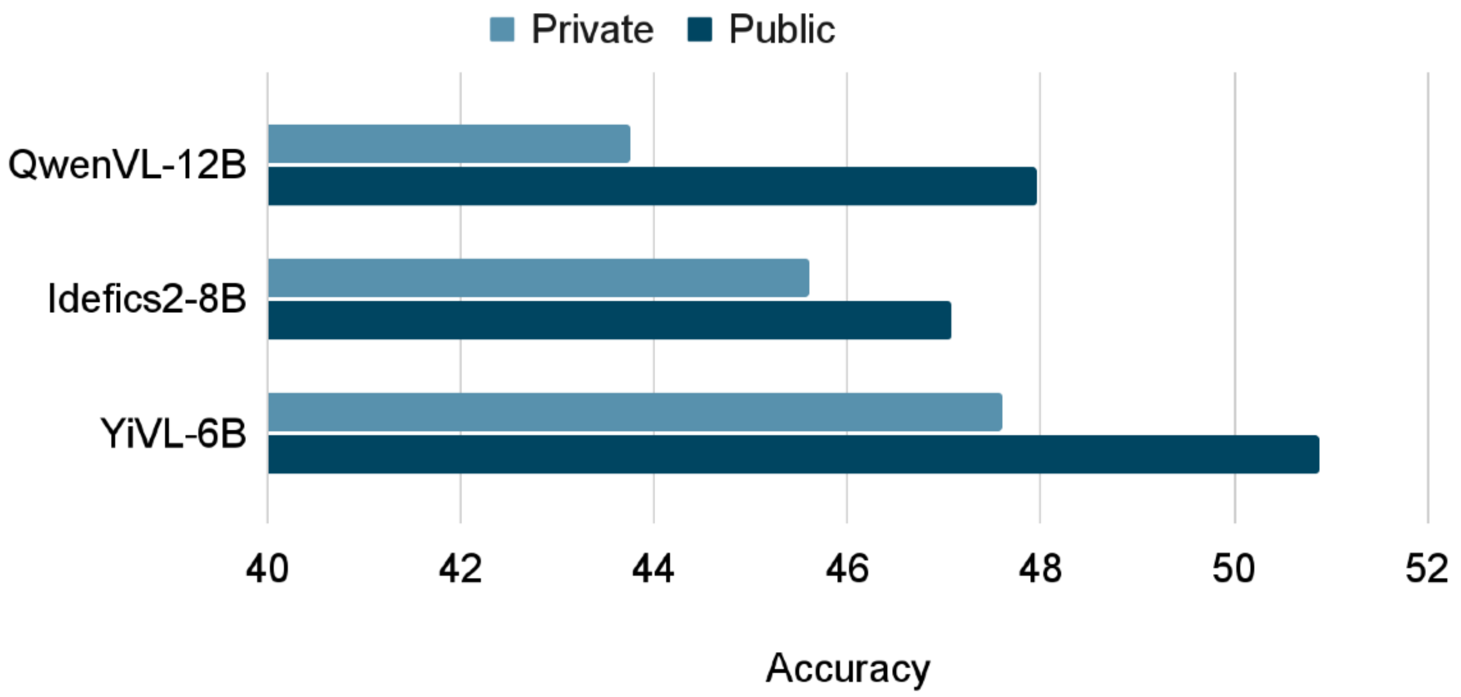
Curated — MaRVL

Private Data — FoodieQA

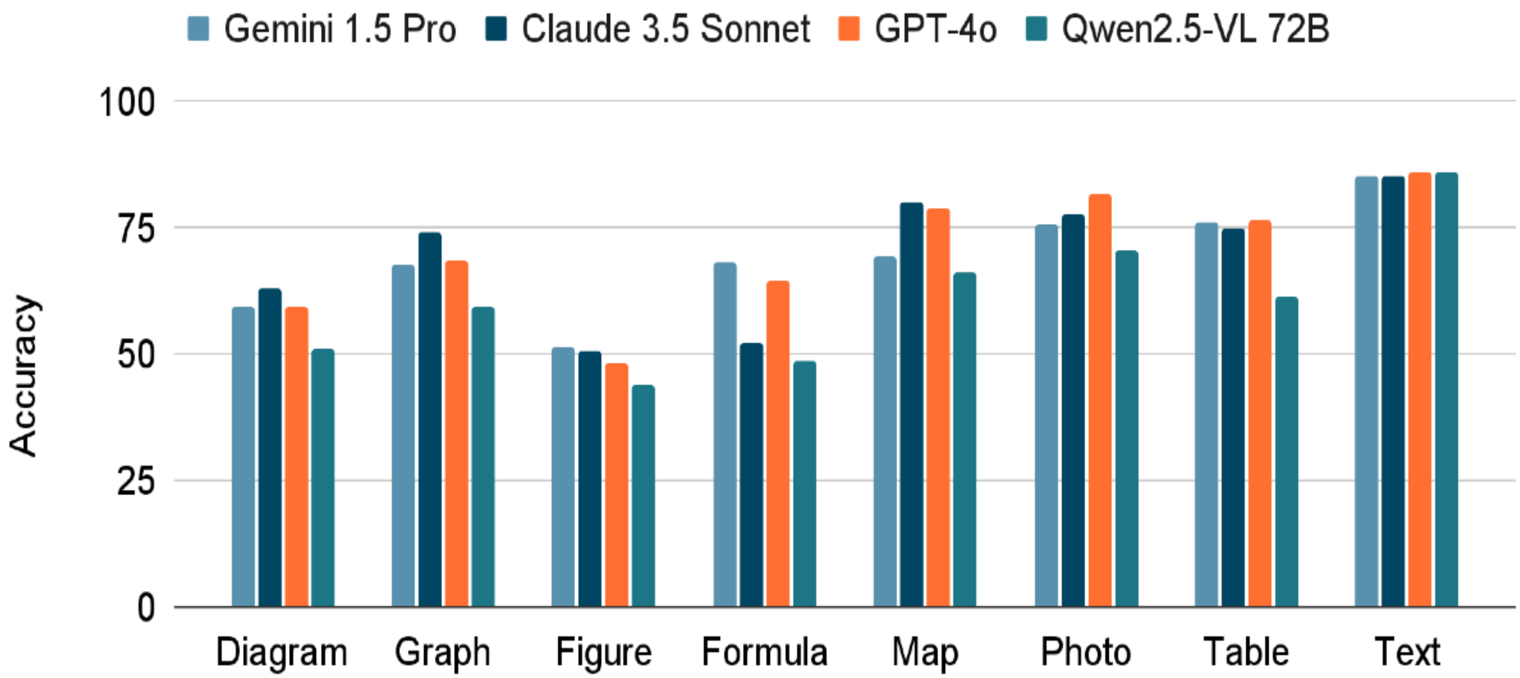
Web Scrape — LAION, Multi30K

Sources

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?

Types

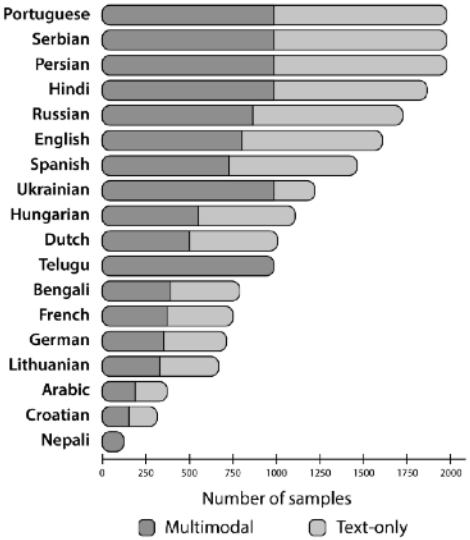
- **Questions**: Kaleidoscope, CulturalVQA, MarVL
- **Sentences**: Multi30K, LAION, XM3600

Sources

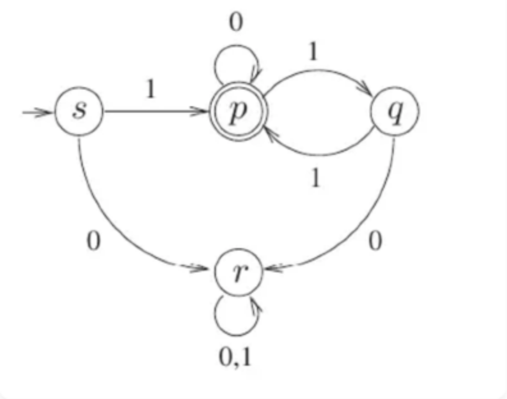
- **Natural**: LAION, Kaleidoscope
- **Generated**: FoodieQA
- **Crowdsourced**: CulturalVQA, Multi30K, FoodieQA

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>?

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)

A woman in a grey suit is giving a speech

Weak ← → Strong

(Mined) (Crowdsourced)

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

Li et al. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. ICML 2022.
Nguyen et al. Improving multimodal datasets with image captioning. NeurIPS 2023.

$P(x|v)$
$\textbf{Distance}(x, v)$

COCO

- Images covering 80 *co*mmon *o*bjects in *co*ntext with multiple human-authored captions.
- Object segmentation data too!

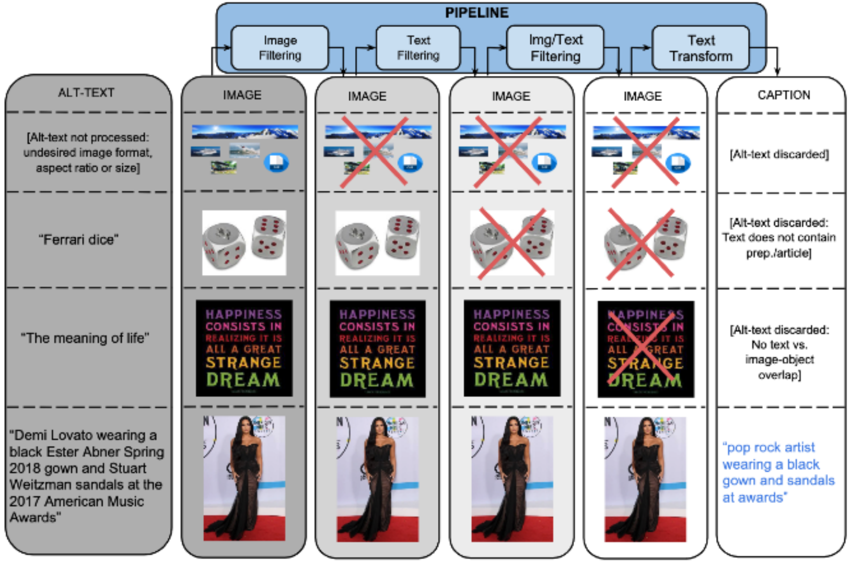
some sheep walking in the middle of a road
a herd of sheep with green markings walking down the road
a herd of sheep walking down a street next to a lush green grass covered hillside.
sheared sheep on roadway taken from vehicle, with green hillside in background.
a flock of freshly sheered sheep in the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

Chen et al. (2015). Microsoft COCO captions: Data collection and evaluation server. *arXiv.*

$P(x|v)$

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.



Sharma et al. (2018). Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. ACL.
Changpinyo et al. (2021). Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. CVPR.
Download your own: https://github.com/igorbrigadir/DownloadConceptualCaptions

$P(y|x, v)$

VQAv2

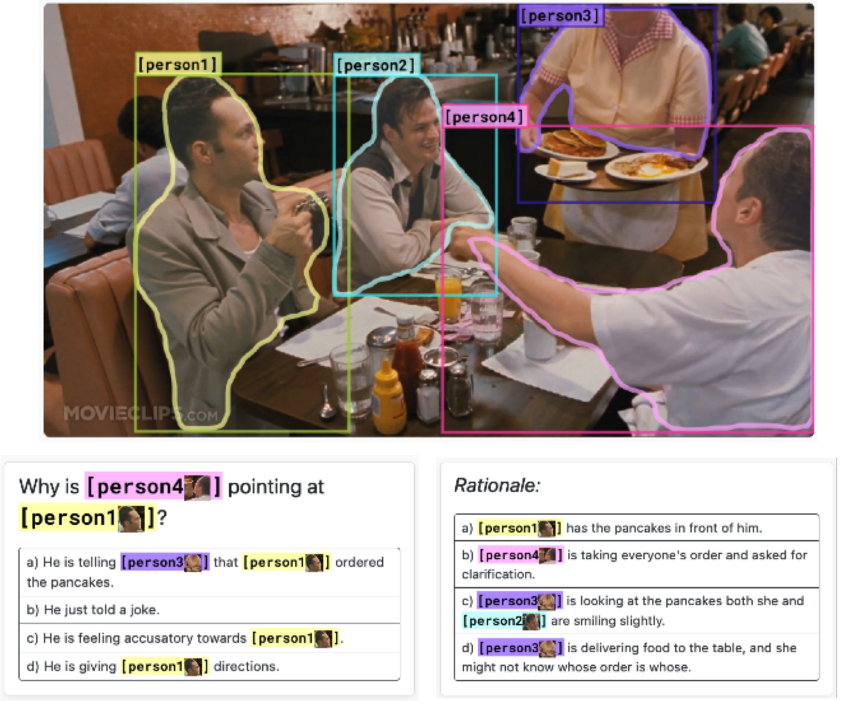
- 1.1M image–question pairs with balanced distribution of answers
- Task with multimodal inputs:
    - Image
    - Question

Goyal et al. (2017). Making the V in VQA matter: Elevating the role of image understanding in visual question answering. CVPR.

$P(y|x, v)$

Visual Commonsense Reasoning

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



Zellers et al. (2019). From Recognition to Cognition: Visual Commonsense Reasoning. CVPR.

$$P(y|x, v)$$
$$\mathbf{Distance}(x, v)$$

BBC-Oxford British Sign Language

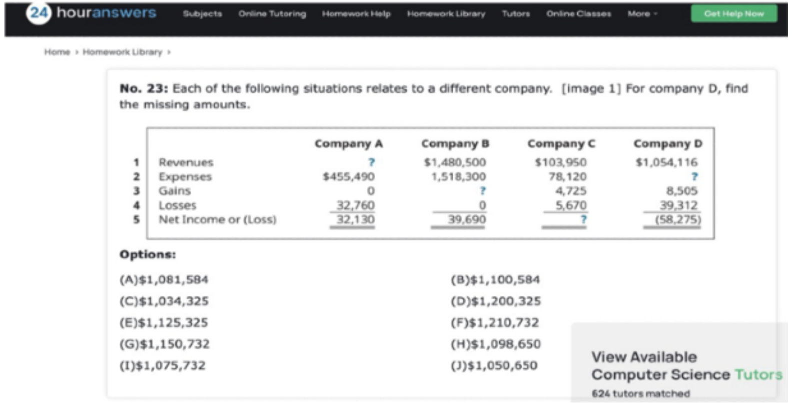
- Sign spotting and sentence localization tasks
- 1,400 hours of signed shows
  - Factual, entertainment, drama, comedy, children’s shows

Albanie et al. (2021). BBC-Oxford British Sign Language Dataset. arXiv.

$P(y|v)$

MMMU-Pro

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


24 houranswers Subjects Online Tutoring Homework Help Homework Library Tutors Online Classes More Get Help Now

Home › Homework Library ›

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)	32,130	39,690	?	(58,275)

Options:

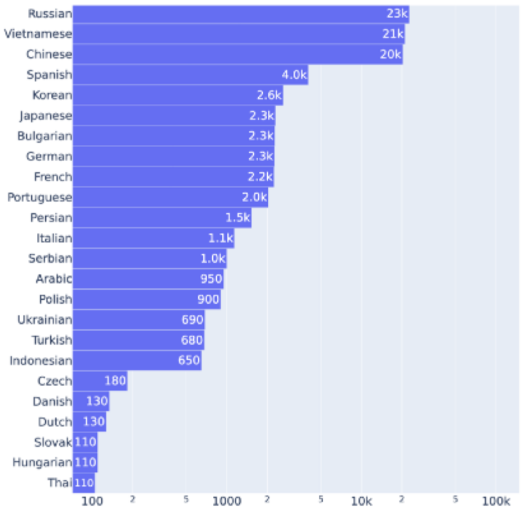
(A)$1,081,584 (B)$1,100,584
(C)$1,034,325 (D)$1,200,325
(E)$1,125,325 (F)$1,210,732
(G)$1,150,732 (H)$1,098,650
(I)$1,075,732 (J)$1,050,650

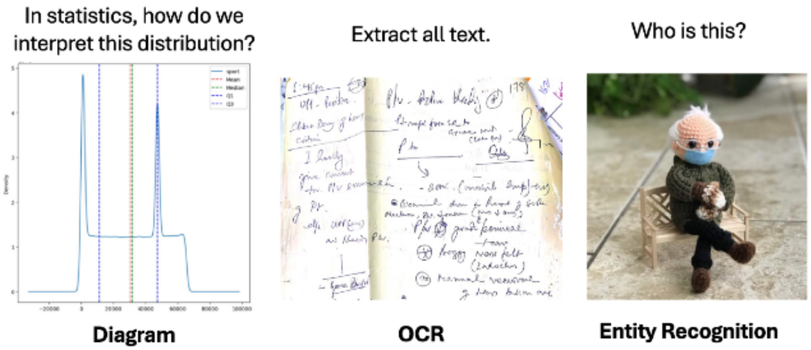
View Available Computer Science Tutors
624 tutors matched

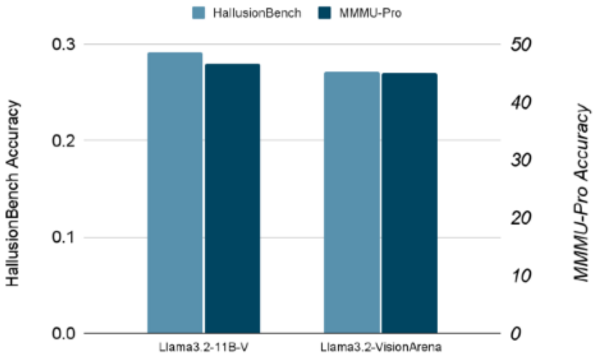
$P(y|x, v)$

VisionArena

- Diverse multilingual dataset with 238K human conversations with VLMs







Chou et al. CVPR 2025. VisionArena: 230K Real World User-VLM Conversations with Preference Labels.

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.

State of the Commons 2022. Creative Commons Foundation.

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"

Birhane et al. (2021) Multimodal datasets: misogyny, pornography, and malignant stereotypes. arXiv.

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

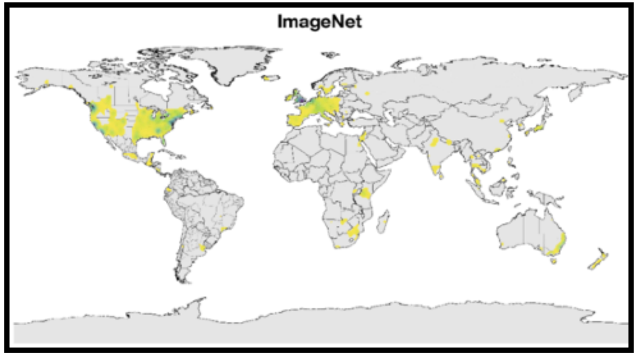
ENG: An unusual looking vehicle …

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

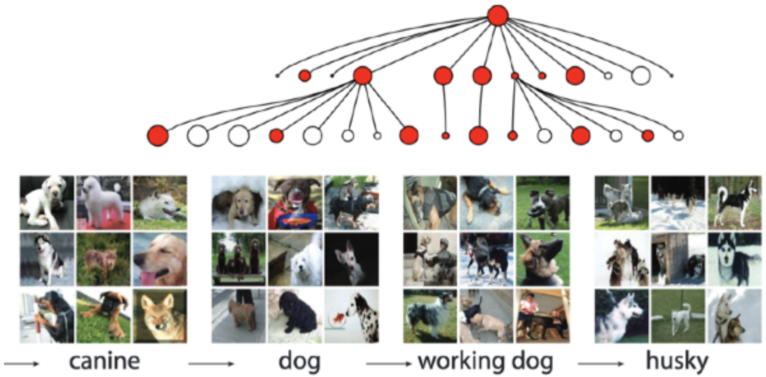
- Narrow linguistic/cultural domain
- No way to assess their real-world comprehension

F. Liu*, E. Bugliarello*, et al. Visually Grounded Reasoning across Languages and Cultures. EMNLP 2021.

Typical Vision and Language DATA

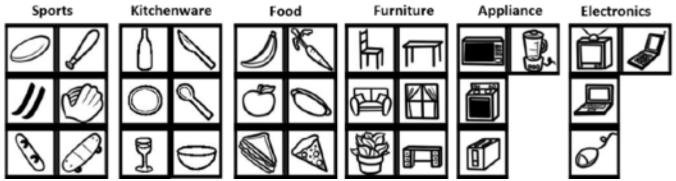
ImageNet (Deng et al. 2009)

- 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; Shweder et al. 2007).

Pilota / Jai-alai

Sanxian / Shamisen

Clavie

Concepts and Hierarchies

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

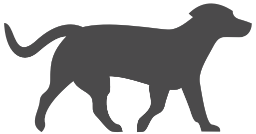
Concept: mental representation of a category (Rosch 1973)

Categories form a *hierarchy*

- Basic-level categories (Rosch 1976)

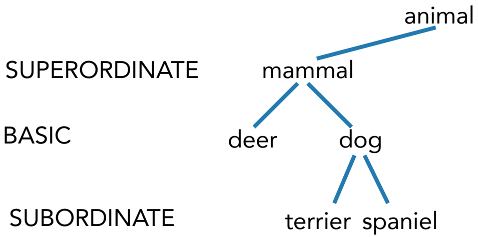
Somewhat universal

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



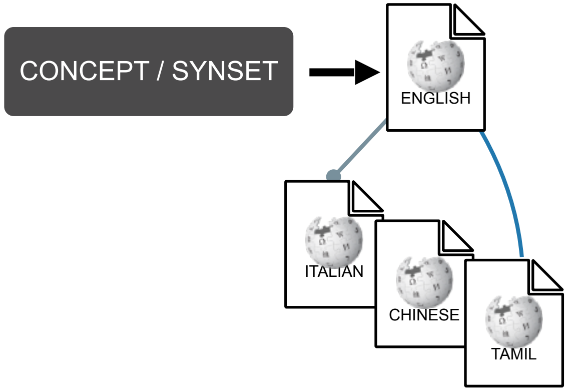
“Dog” concept

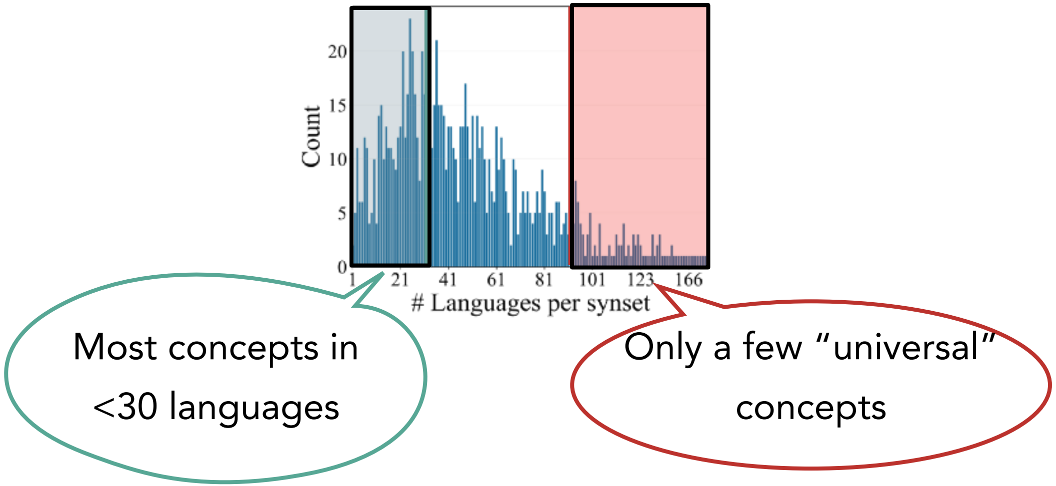
“Dog” category



Are ImageNet Concepts Cross-Lingual?

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





Multicultural Reasoning over Vision and Language

Representative of annotators' cultures

5 typologically diverse languages

Independent, culture-specific annotations


Bola basket


Mpira wa kikapu


Basketbol


篮球


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 காளையை அடக்கும் பணியில்
ஈடுப்பட்டிருப்பதை காணமுடி.

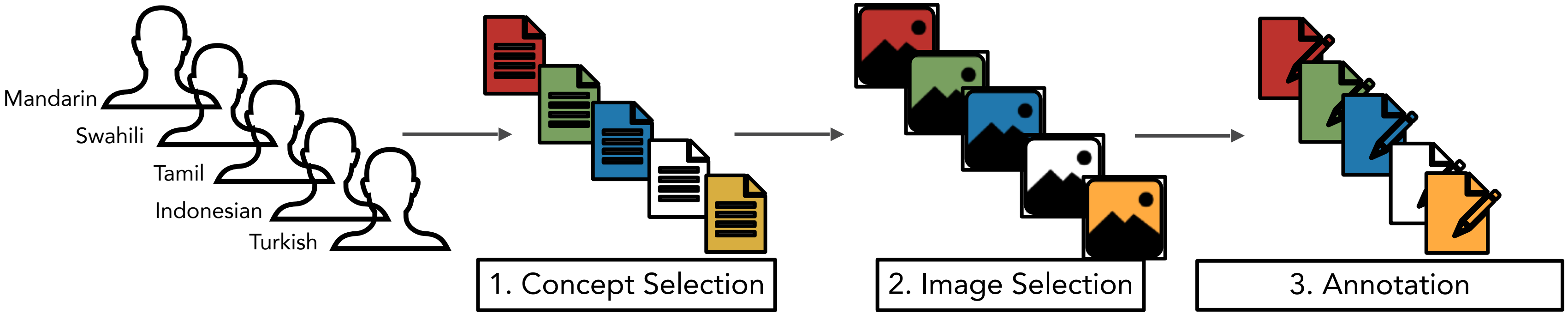
True

x y

Suhr et al. A Corpus for Reasoning about Natural Language Grounded in Photographs. ACL 2019.

Collecting MaRVL data

Native speaker-driven protocol



MaRVL is created from Universal Concepts

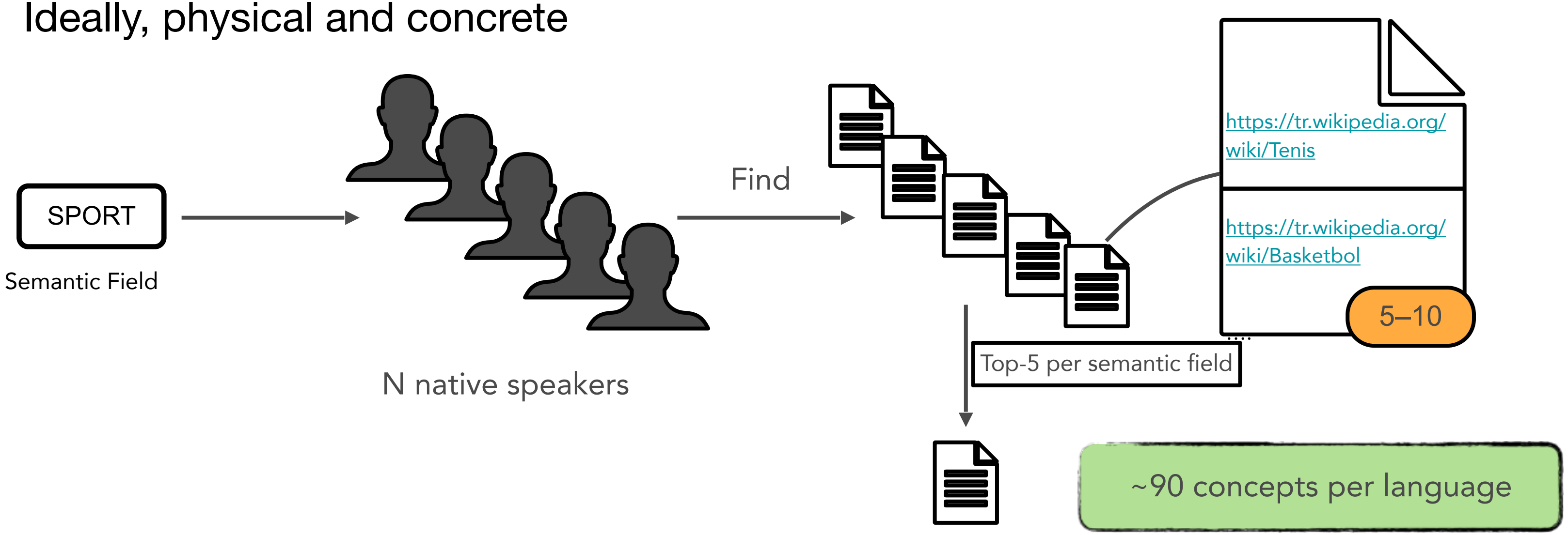
- Taken from the Intercontinental Dictionary Series (Key & Comrie, 2015)
  - 18/22 chapters with concrete objects & events

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

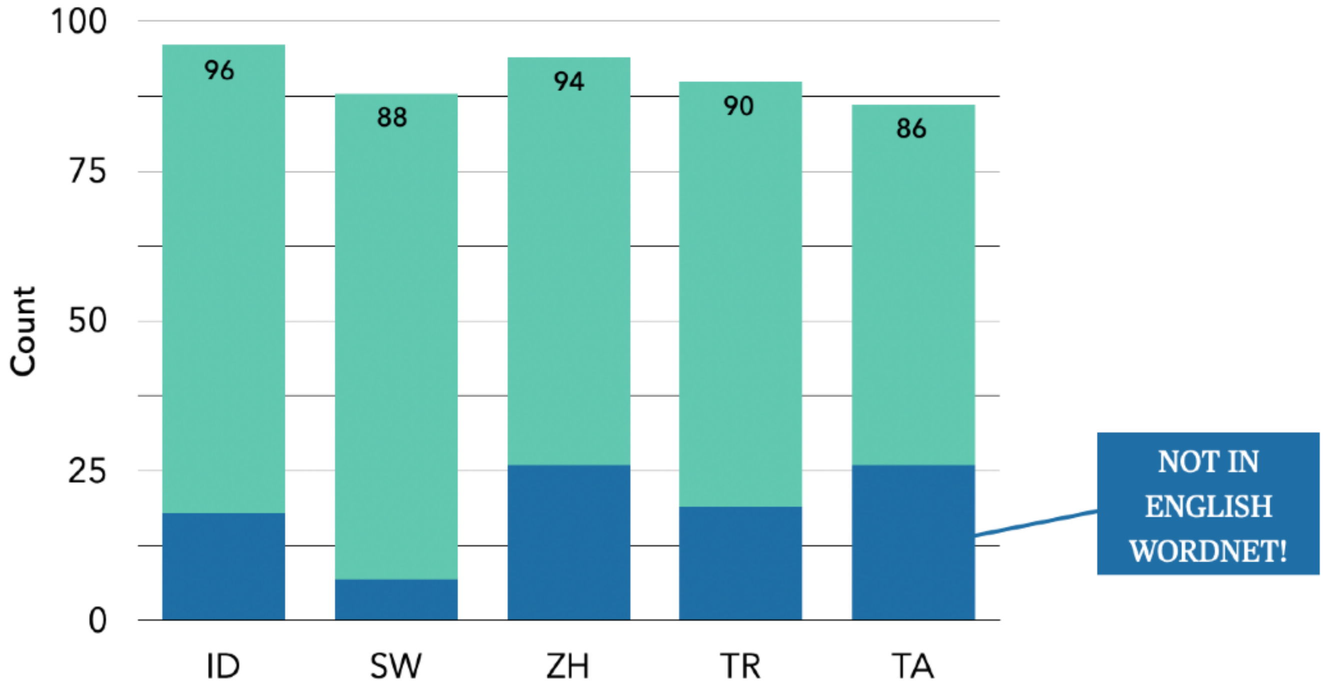
Step 1. Language-Specific Concepts

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

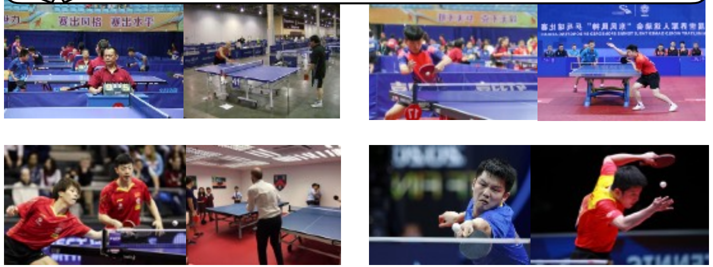
- Representative of the language population
- NLVR2 (Suhr et al. ACL 2019) requirements

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
3. Shows an instance of the concept performing an activity
4. Displays a set of diverse objects or features

Step 3. Language Annotation

Written by native speakers

MATCH 4 PAIRS AT RANDOM



WRITE CAPTION TRUE ONLY FOR 2 PAIRS

右图中的人在发球，左图中的人在接球。

VALIDATE ANNOTATIONS

右图中的人在发球，左图中的人在接球。

图中的人在发球，左图中的人在接球。

FINAL VALIDATION

右图中的人在发球，左图中的人在接球。

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

Fleiss' kappa: 93%

Dataset Examples

Görsellerden birinde dizlerinde kanun bulunan birden çok insan var

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

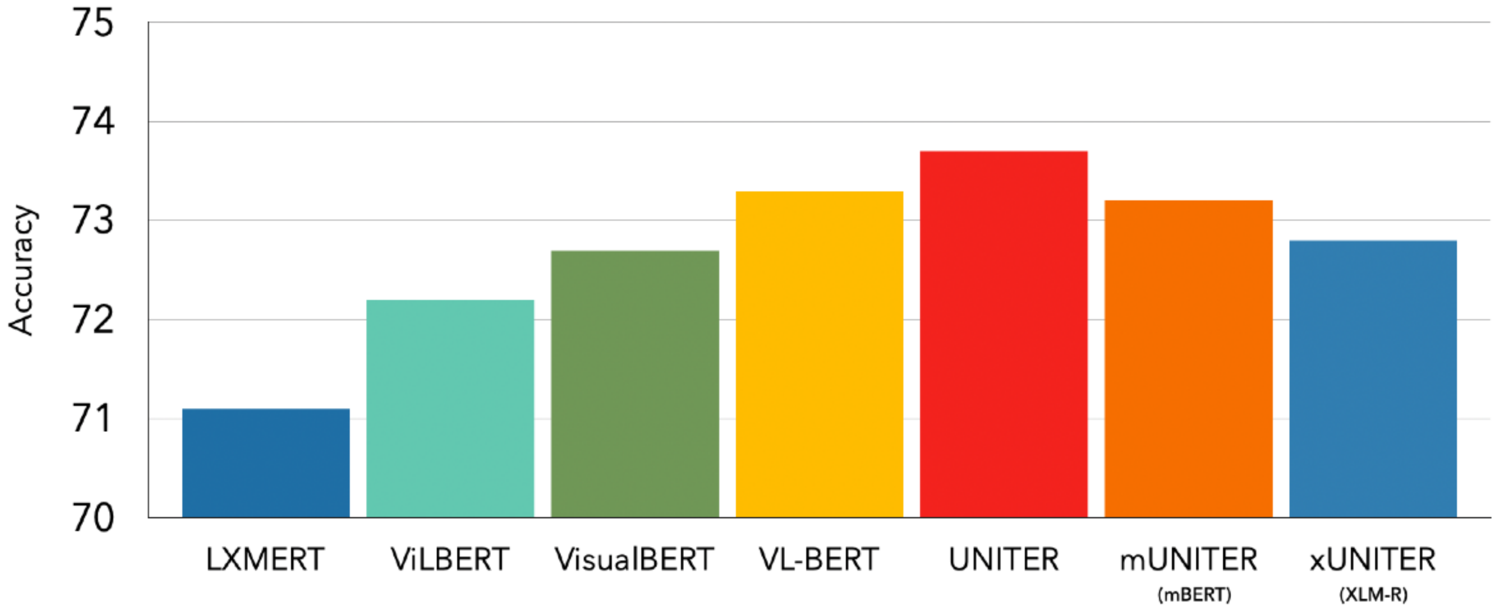
Label: True

Picha ya upande wa kushoto mtu mmoja tu anapiga zumari
na picha ya upande wa kulia watu wawili wanapiga zumari

(Picture on the left is just one person blowing the flute and in the
picture on the right two people are blowing the flute)

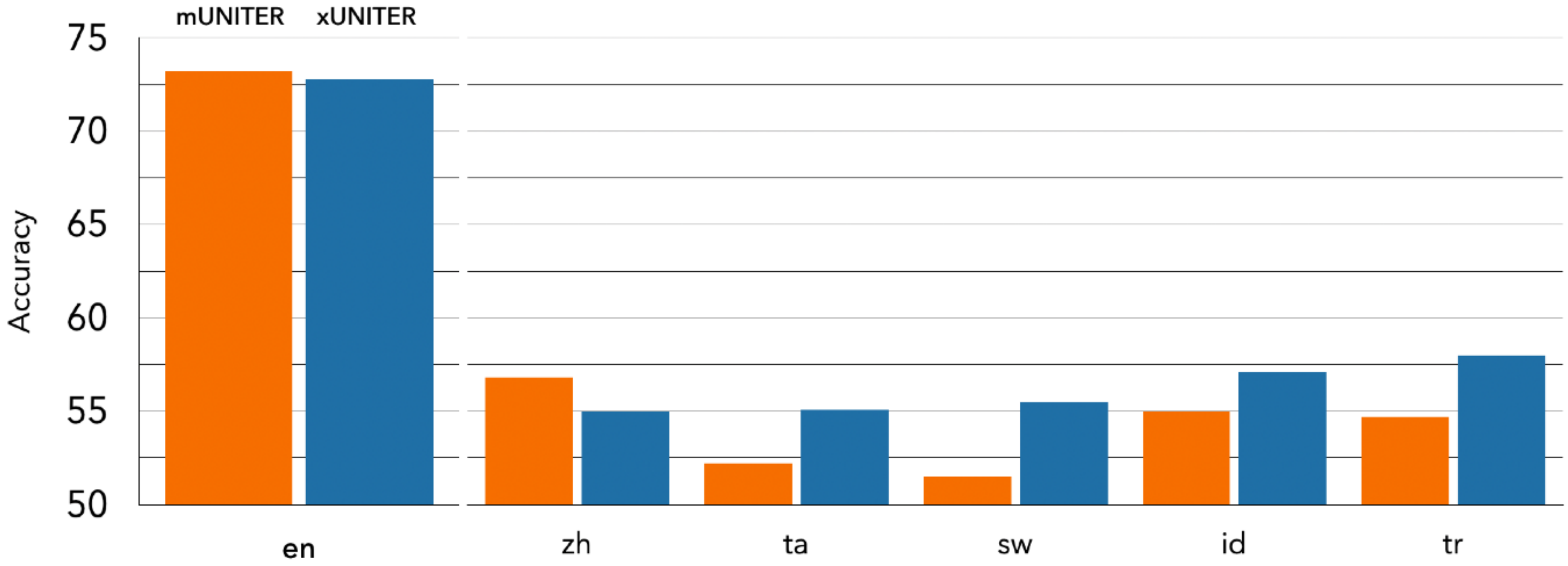
Label: True

English NLVR2 Results (Sanity check)



m/xUNITER perform similarly to English-only models

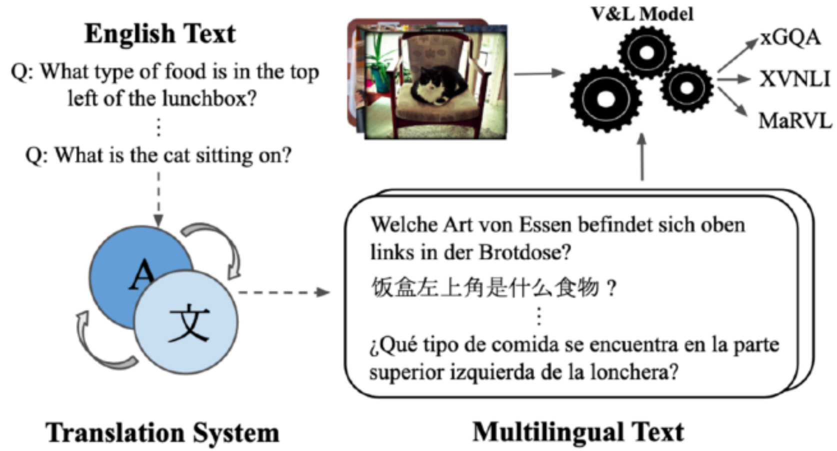
MaRVL Zero-shot Results

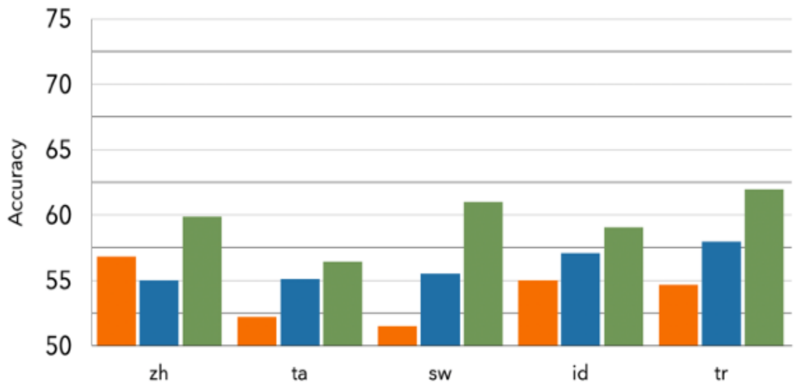


Zero-shot transfer: substantial drop in performance

Pretraining with Translated Text

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

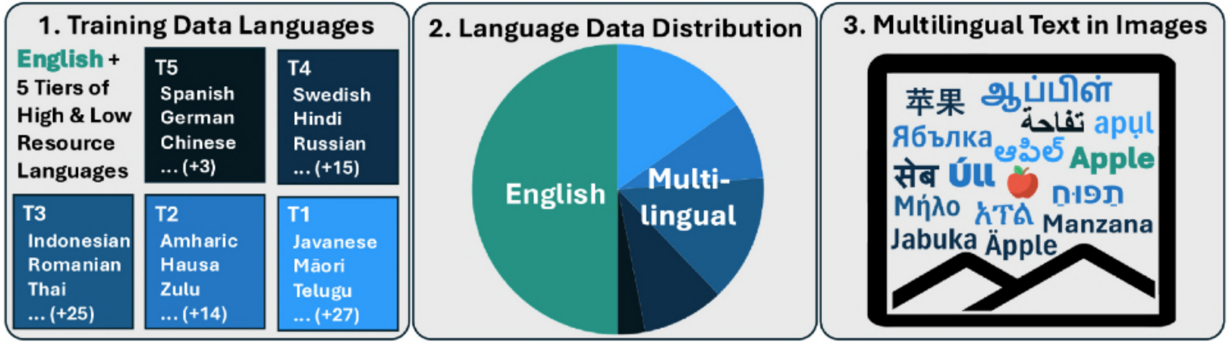




Cross-modal multilingual multimodal pretraining helps!

Qui et al. Multilingual Multimodal Learning with Machine Translated Text. Findings of EMNLP 2022.

State of the Art: Centurio



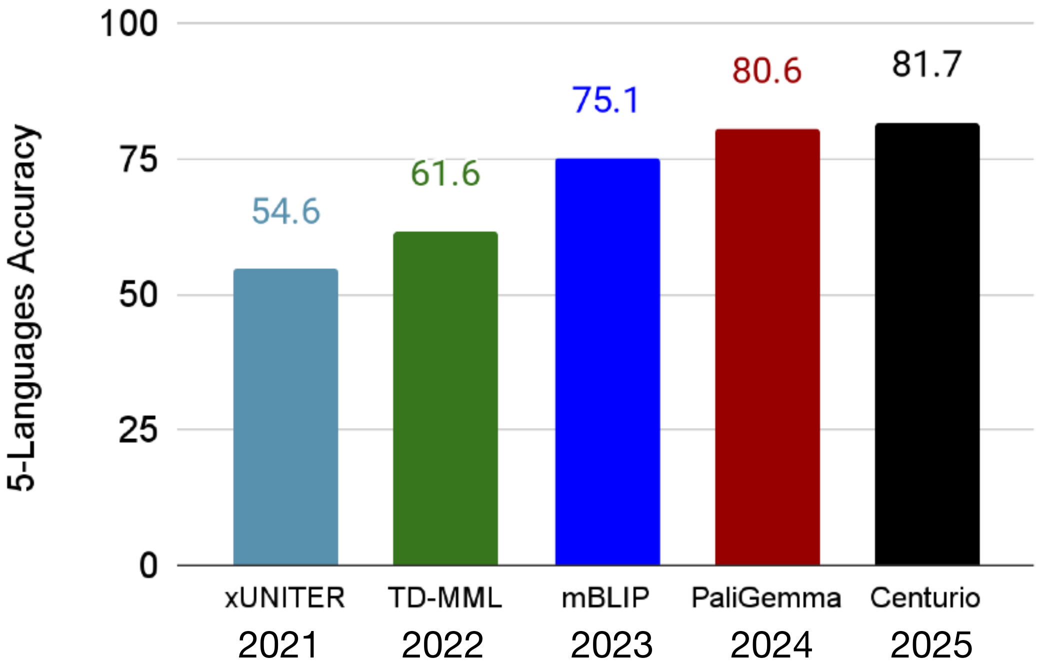
- Initialize from
  - Qwen2.5 7B
  - SigLIP So400/384
- Fine-tune on lots of synthetic data
  - Synthetic OCR data
  - ALLaVa + ShareGPT4V captions
  - Machine translated texts

Geigle et al. Centurio: On Drivers of Multilingual Ability of Large Vision-Language Model. ACL 2025.

Year-on-Year Improvements

- Clear benefit when using machine translated data

- Better visual encoders and language models can enable effective zero-shot transfer

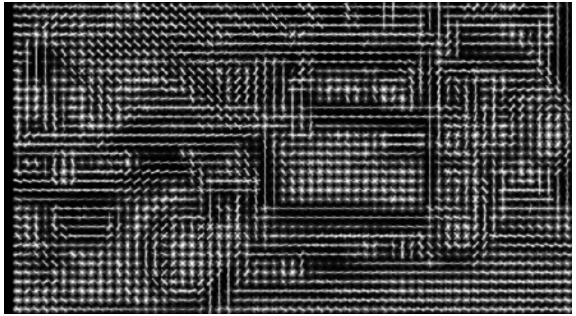


2. Data Representation

Three Levels of Representation

- Perceptual
- Pre-processed features
- Raw input

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



Images: http://www.cs.columbia.edu/~vondrick/ihog/

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

McRae et al. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior research methods*, *37*(4), 547.

Perceptual Norms: Pros / Cons

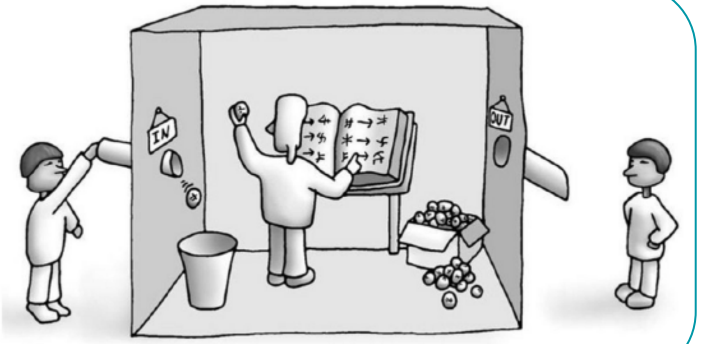
Pros

- Seemingly simple task
- Rich features

Cons

- Can it scale?
- Handling ambiguity

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

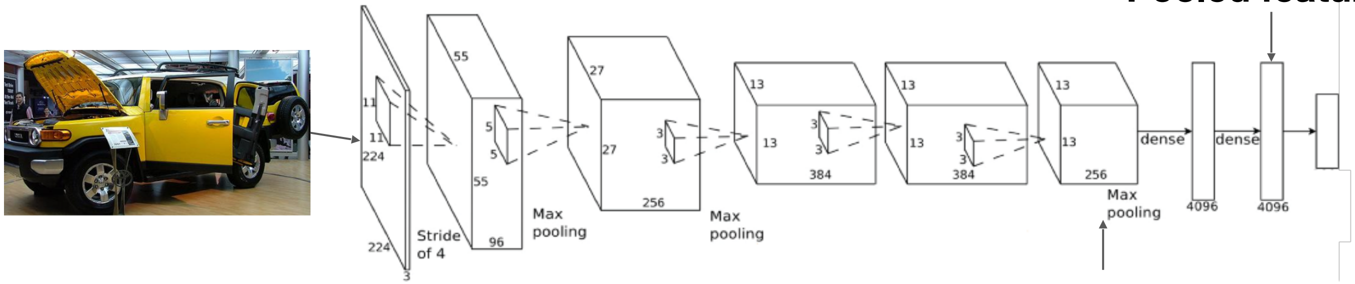


Searle. (1980). Minds, Brains and Programs. *Behavioral and Brain Sciences*, 3: 417–57

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 $\boxed{\text{4096-dimensional activations}}$ of the fully connected layer $\boxed{\text{immediately before the classifier.}}$" **Pooled features**

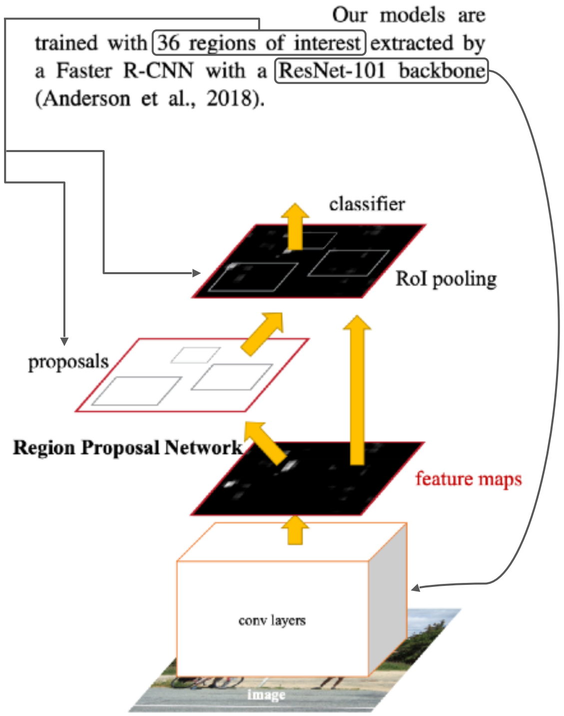


Spatial features "In our experiments we use the $\boxed{14\times14\times512 \text{ feature map}}$ of the fourth convolutional layer before max pooling."

Karpathy & Fei-Fei (2015). Deep visual-semantic alignments for generating image descriptions. CVPR.
Xu et al. (2015). Show, attend and tell: Neural image caption generation with visual attention. ICML.

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.



Ren et al. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. *NeurIPS.*
Bugliarello et al. (2021). Multimodal Pretraining Unmasked: A Meta-Analysis and a Unified Framework of Vision-and-Language BERTs. TACL.

Pre-processed: Pros / Cons

Pros

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

Cons

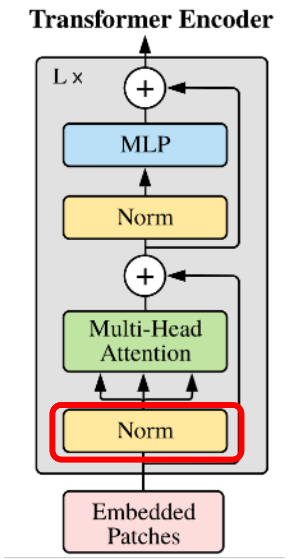
- Large datasets require specialized storage
- Not obvious how to randomly augment data
- Specialist knowledge can be opaque to newcomers

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.



Dosovitskiy et al. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. ICLR.
Baevski & Auli (2018). Adaptive input representations for neural language modeling. *ICLR.*

Nomenclature and Patch Count

Model size
- Tiny, Small, Base, Large, Huge

`ViT-B/32`^{384}

Model family
- ViT
- Swin
- LiT

Image size
- 224^2, 384^2, 448^2

Patch size
- 14^2
- 16^2
- 32^2

No. patch embeddings =
(Image size / patch size)^2

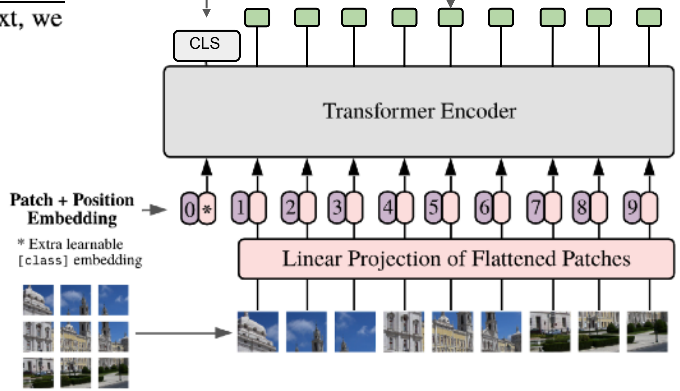
ViT-B/32^{384}: 144 patches
ViT-H/14^{224}: 256 patches
ViT-H/14^{448}: 1024 patches
(ouch)

Extracting ViT Features

- Extract pooled features or patch-level features

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.

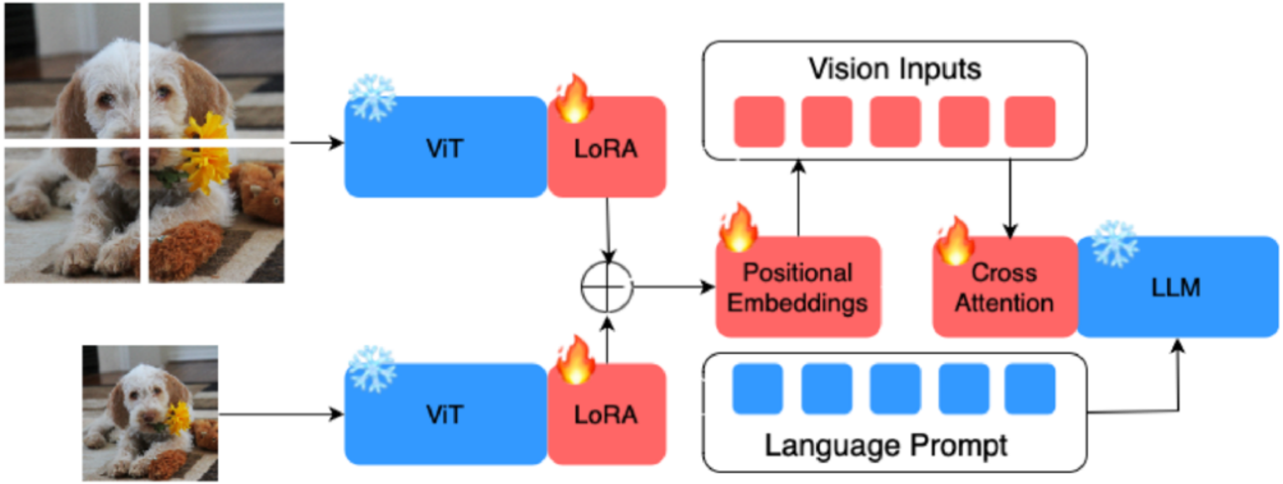
To extract visual information from an image x^i, we use the visual encoder of a pre-trained CLIP [29] model. Next, we



Eichenberg et al. (2021). MAGMA--Multimodal Augmentation of Generative Models through Adapter-based Finetuning. *EMNLP*

Learning to Process Higher-Resolutions

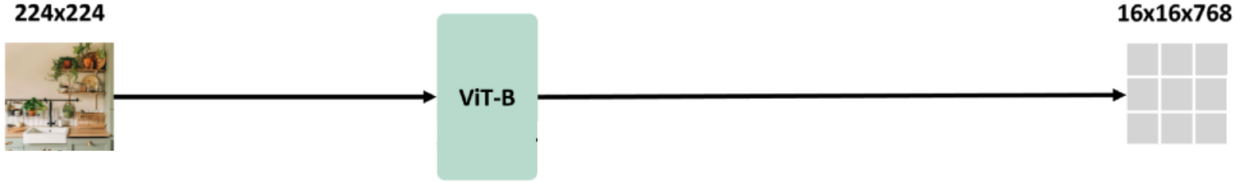
- Learn local and global LoRA adapters to combine multi-scale inputs



Carvalho and Martins. COLING 2025. Efficient Architectures for High Resolution Vision-Language Models.

Scaling on Scale

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



Shi et al. ECCV 2024. When Do We Not Need Larger Vision Models?

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

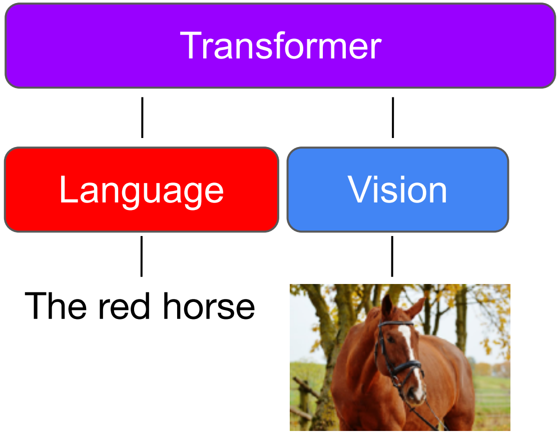
Summary

- Many options for how to represent your multimodal inputs
  - Language-oriented
  - Object / stuff oriented
  - Raw inputs
- **Raw inputs** are the current favoured approach to visual encoding because you can update the weights in the model

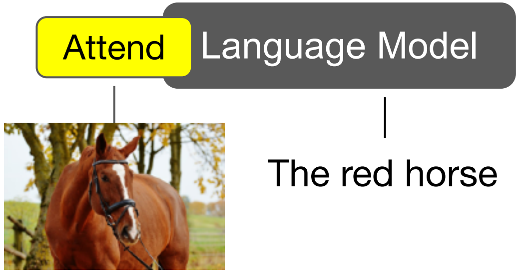
3. Modelling

Main Approaches

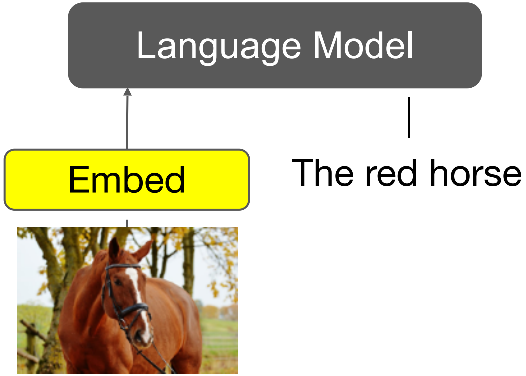
Cross encoding



Cross-Attention

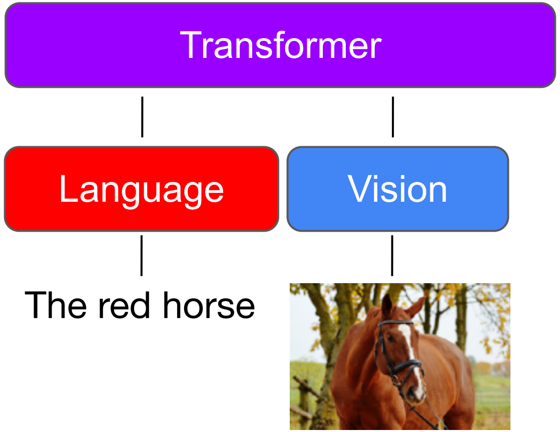


Visual Prefix



Cross-encoding Models

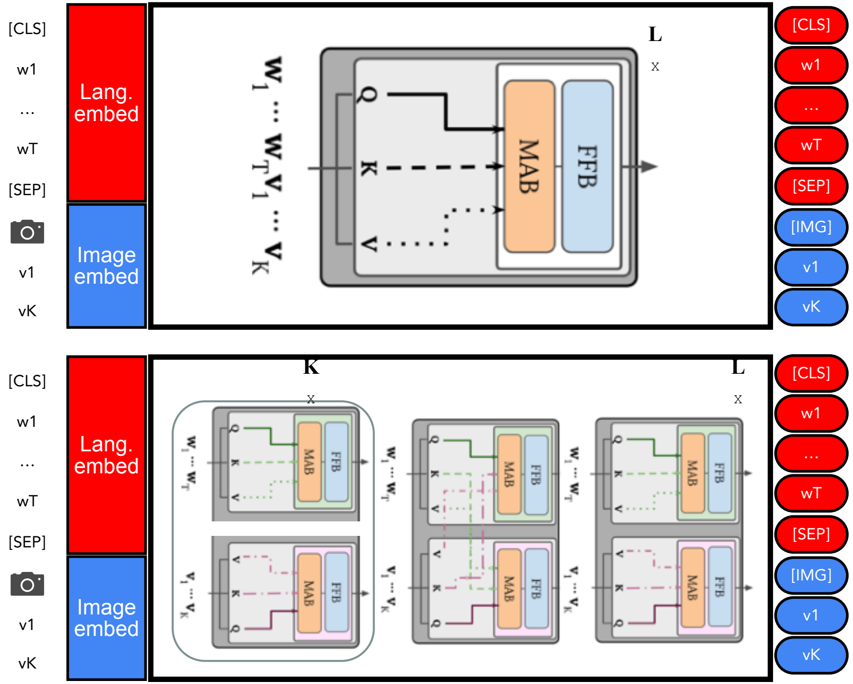
- Emerged as a key modelling approach in 2019 with many concurrent methods for creating visually-grounded BERT models.
- This is a form of *model-based fusion*
- The backbone consists of two components:
    - language encoder
    - visual encoder



Tan & Bansal (2019). LXMERT: Learning Cross-Modality Encoder Representations from Transformers. EMNLP-IJCNLP.
Lu et al. (2019). ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. NeurIPS.

Single- & Dual-Stream Architectures

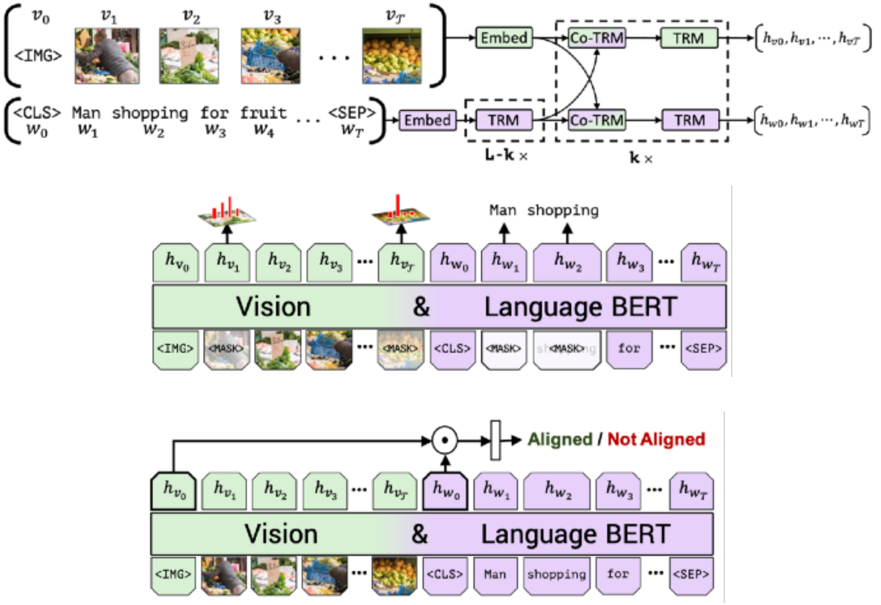
- Single-stream
    - Concatenate inputs into one sequence
- Dual-stream
    - Process modalities independently
        - Intra-modal
        - Inter-modal



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

ViLBERT

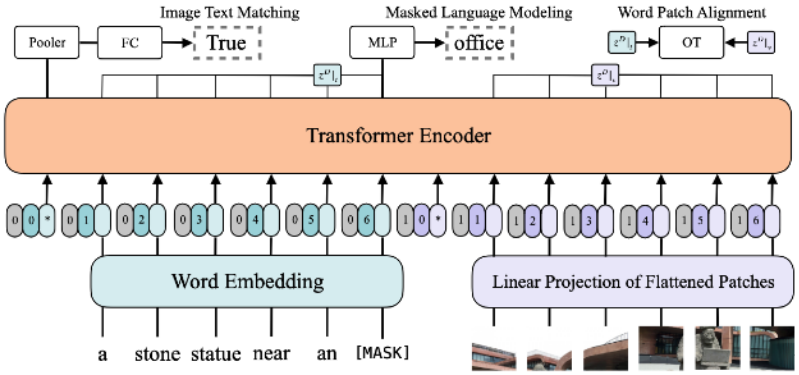
- Dual-stream model
- Initialized from BERT
- Visual features extracted from 10-36 regions using Faster-RCNN
- Pretrained on Conceptual Captions
  - Masked Language Modelling
  - Masked Region Classification
  - Image-Text Matching



Lu et al. (2019). ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. NeurIPS.

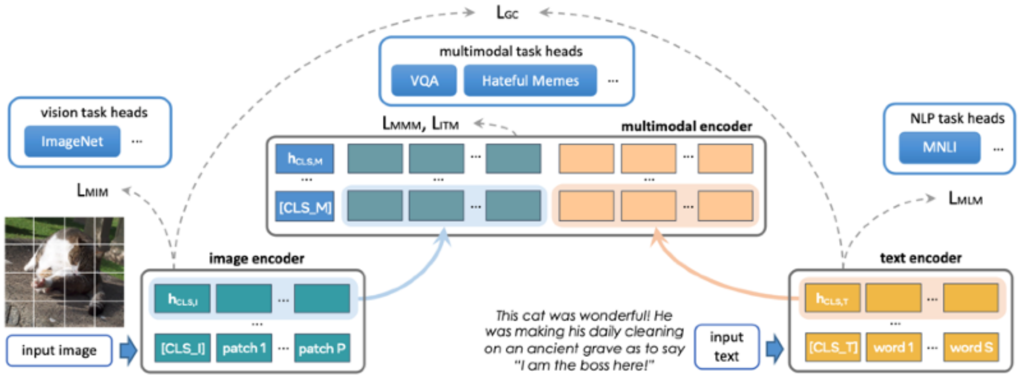
ViLT

- 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



Kim et al. (2021). ViLT: Vision-and-language transformer without convolution or region supervision. ICML.

FLAVA

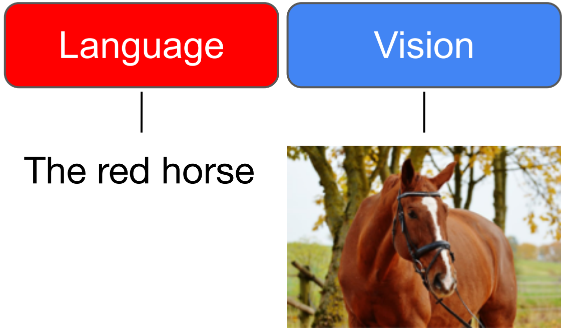


- Dual-stream Visual features extracted from ViT-B/16
- Pretrained on PMD70M
  - Masked Language Modelling, Masking Image Modelling
  - Image-Text Matching, Masked Multimodal Modelling
  - Global Contrastive Matching

Singh et al. (2022). FLAVA: A foundational language and vision alignment model. CVPR.

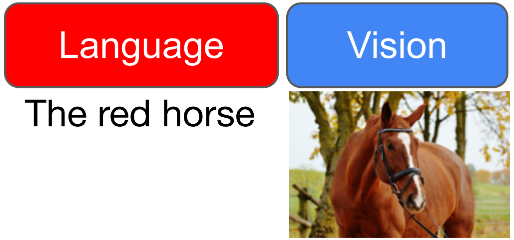
Dual-encoding Models

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

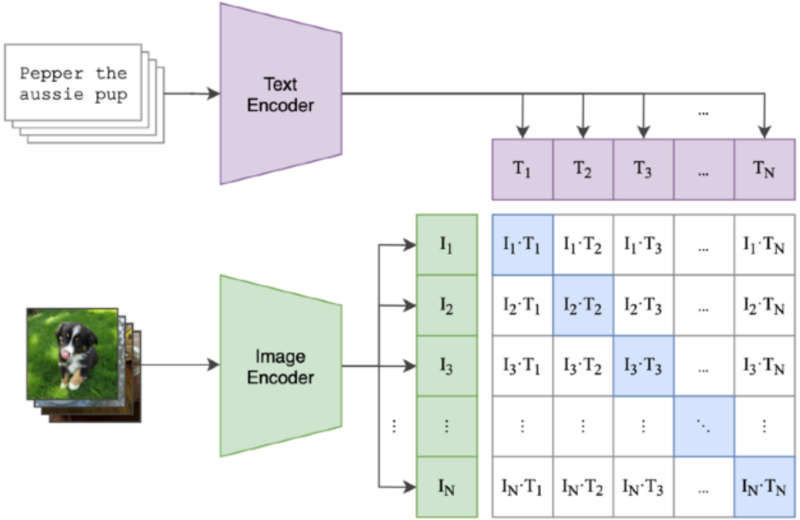


Radford et al. (2021). Learning transferable visual models from natural language supervision. ICML.

CLIP



- 12 Layer Transformer Encoder
- ViT or ResNet Visual Encoder



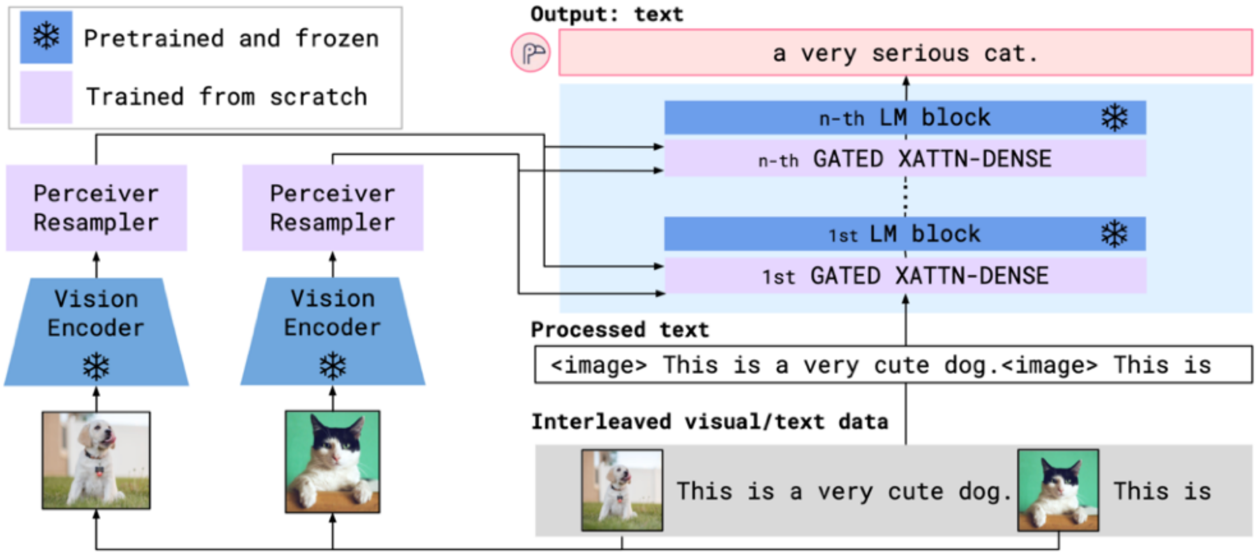
- Maximize the similarity of the embeddings of paired examples (I, T):

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

- Large pretraining dataset of unclear provenance

Radford et al. (2021). Learning transferable visual models from natural language supervision. ICML.

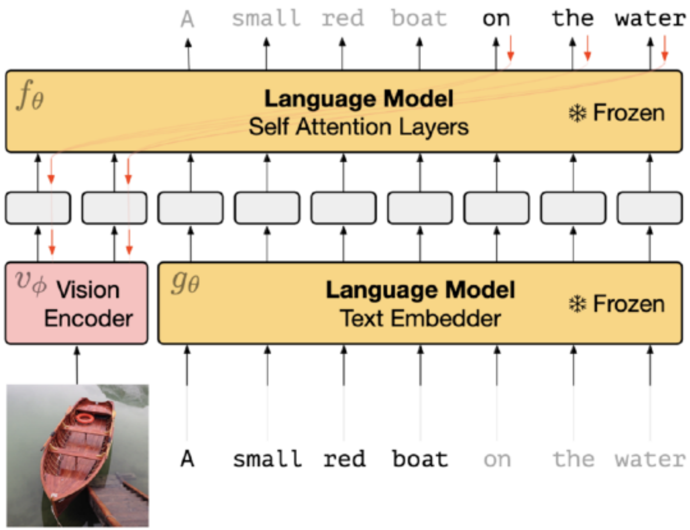
Cross-Attention



Alayrac et al. NeurIPS 2022. Flamingo: a Visual Language Model for Few-Shot Learning

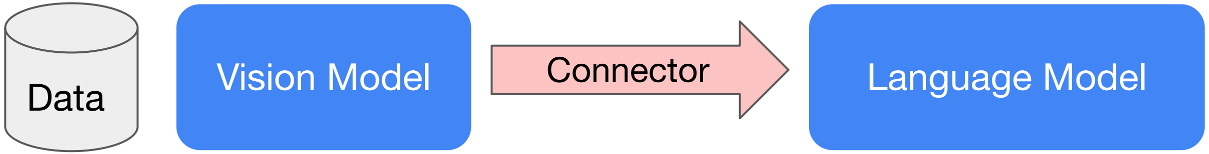
Visual Prefix

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



Tsimpoukelli et al. NeurIPS 2021. Multimodal Few-Shot Learning with Frozen Language Models.

Training VLMs in 2025



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

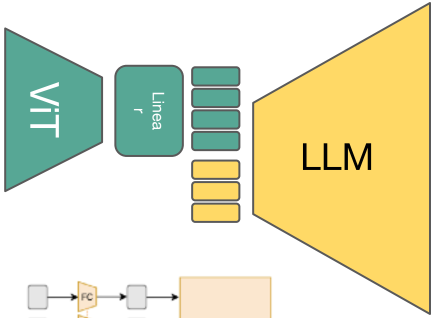
Modality-Specific Components

	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

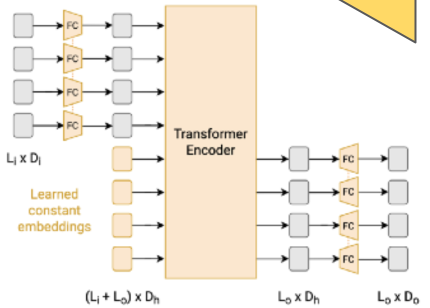
Connector Design

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

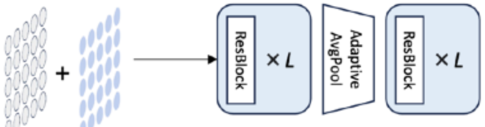


- Qwen-VL:

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



(a) C-Abstractor

Manas et al. MAPL: Parameter-efficient adaptation of unimodal pre-trained models for vision-language few-shot prompting. EACL 2024.
Cha et al. Honeybee: Locality-enhanced Projector for Multimodal LLM. CVPR 2024

Training Datasets

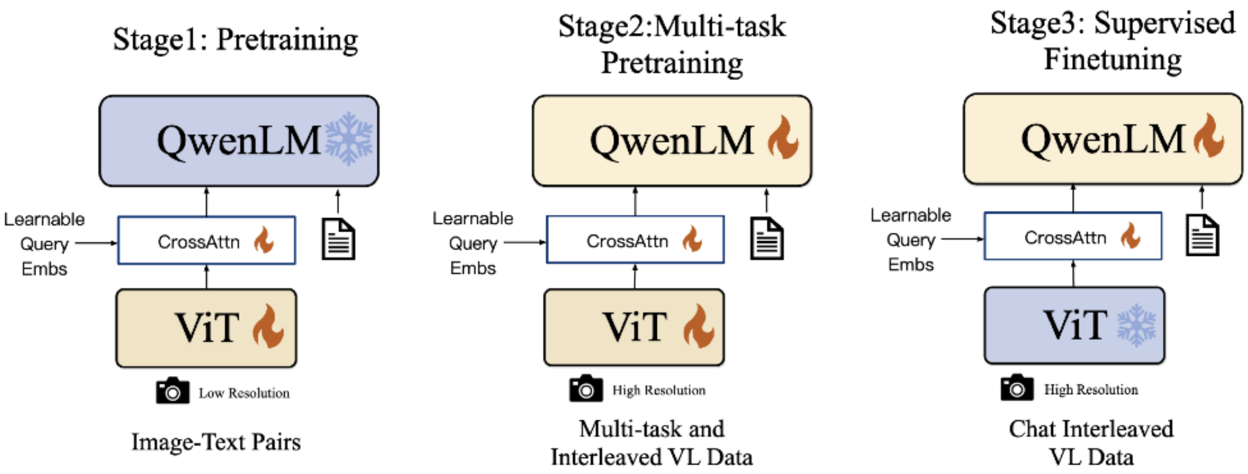
- LLAVA: 595K image–caption examples filtered from CC3M

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+)

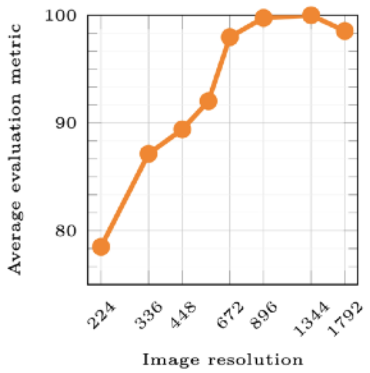
Training Strategy

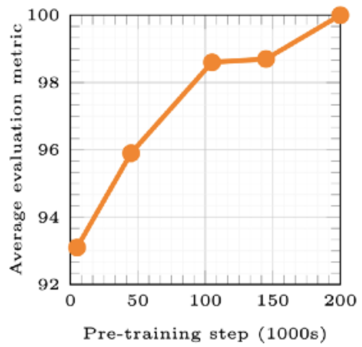
- Qwen-VL, PaliGemma, and MM1 use multi-stage training strategies with different types of data and different image resolutions

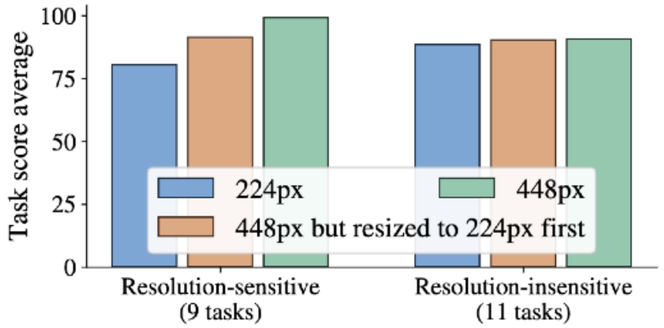


Open Questions

- 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?

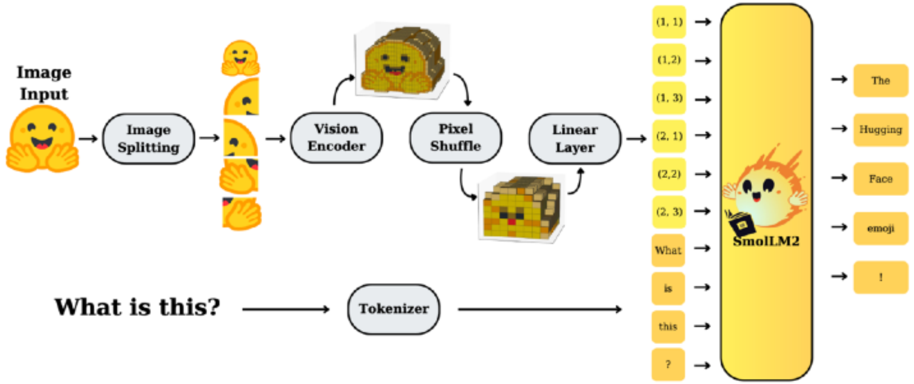


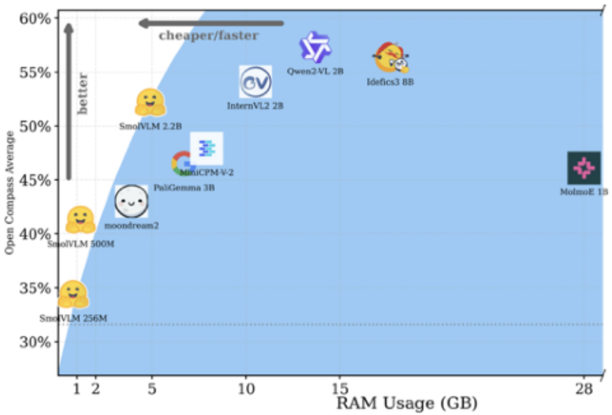




Trends in Smaller Models

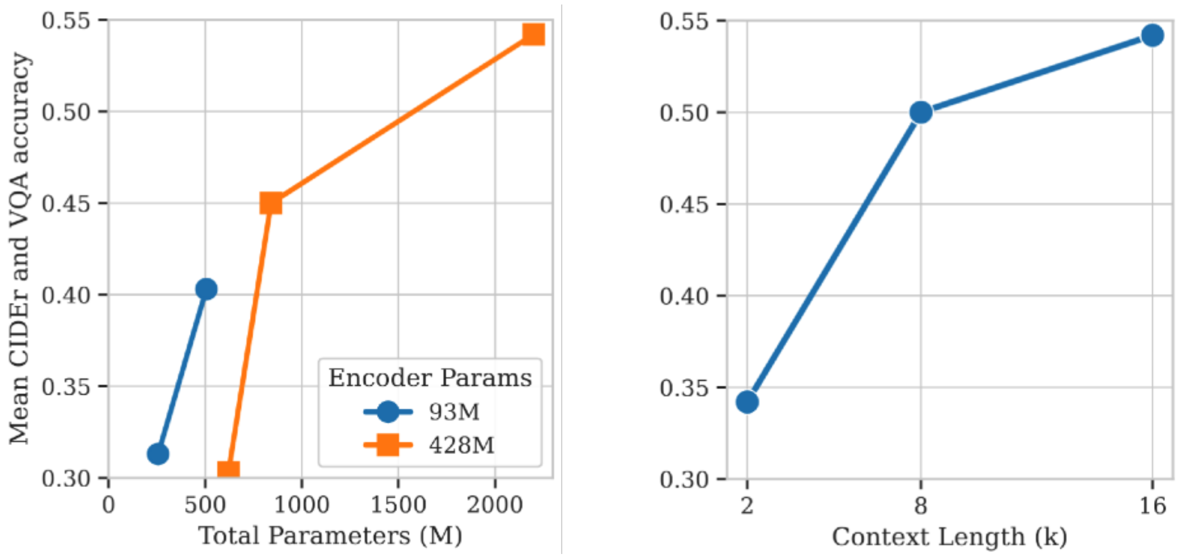
- How quickly will we realize these benefits in smaller models?
  - Less than six months



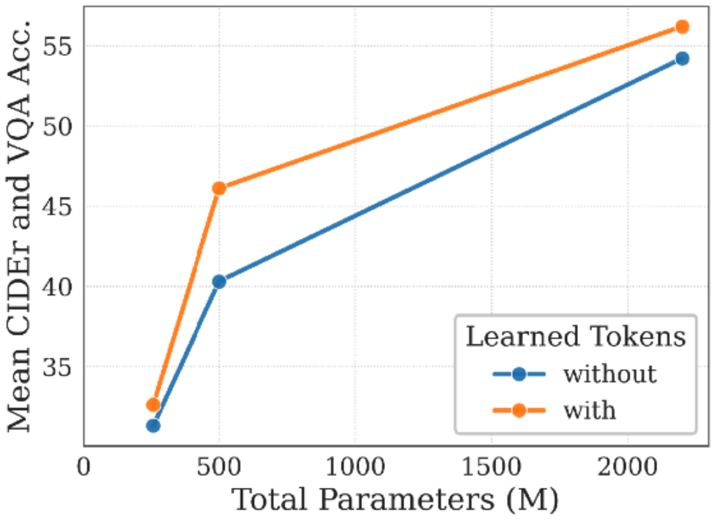


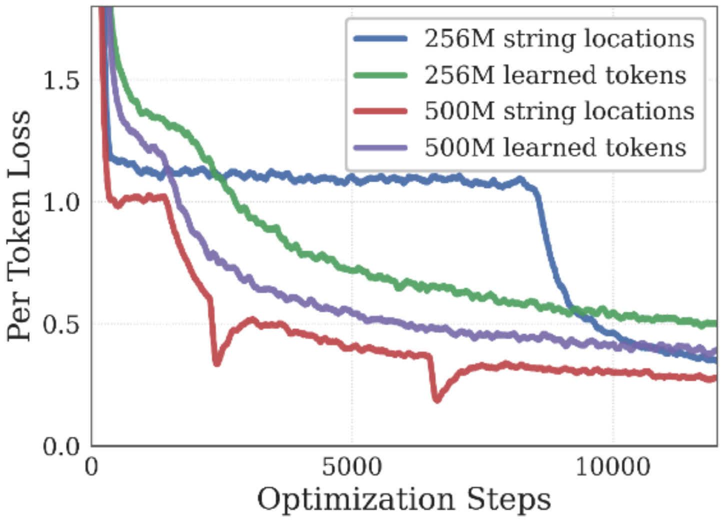
Marafioti et al. COLM 2025. SmolVLM: Redefining small and efficient multimodal models.

Balance Models and Increase Context

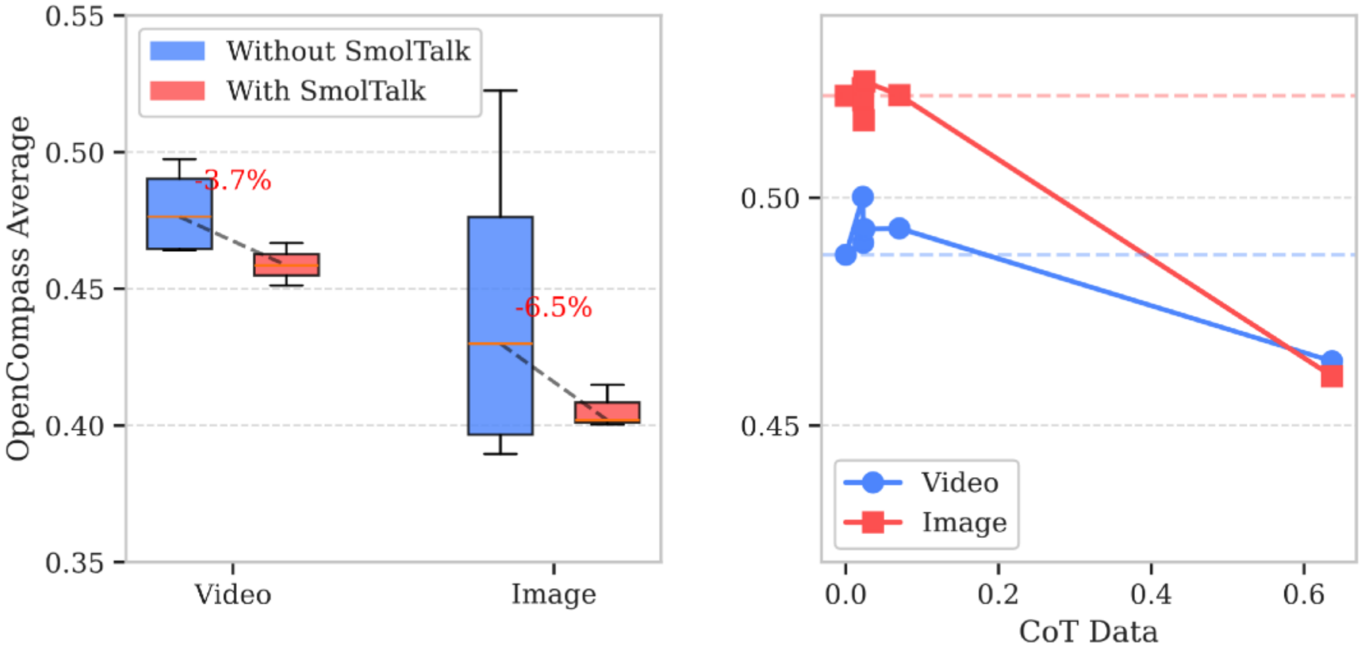


Learned Position Embeddings Help!





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

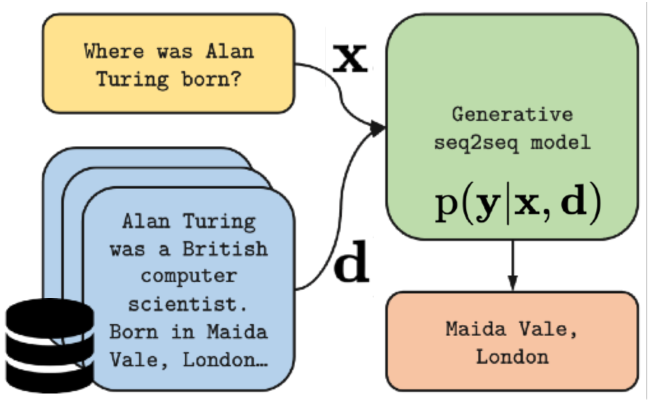
Findings of NAACL 2024




R. Ramos   E. Bugliarello   B. Martins   D. Elliott

Retrieval Augmented Generation

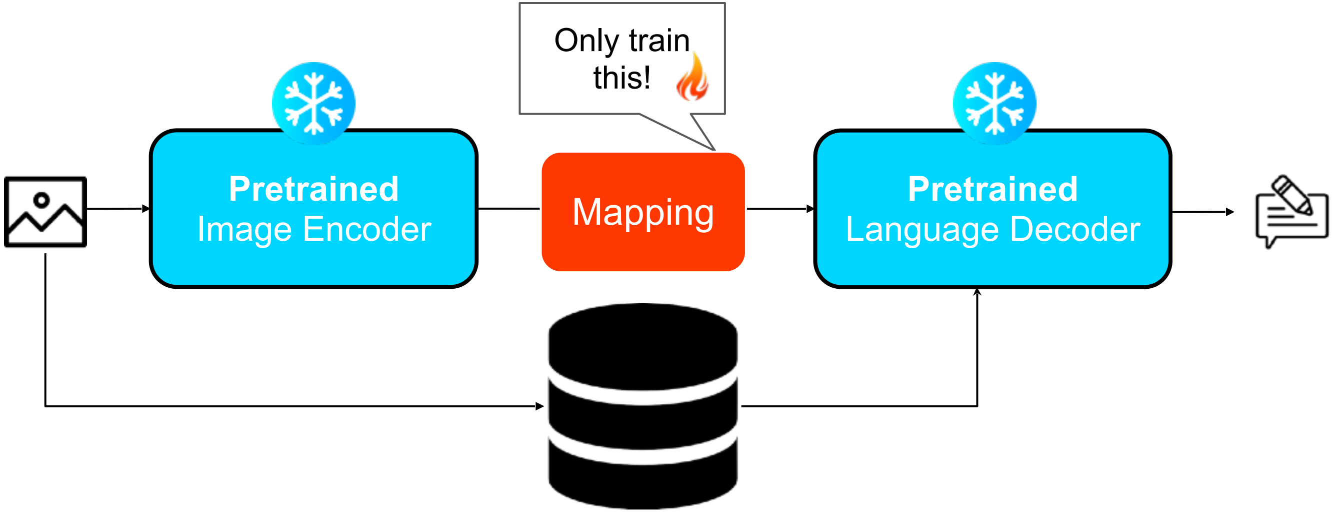
- Combine the power of in-weights learning with in-context adaptation through retrieval augmentation
- Given a datastore of facts, knowledge, documents, etc.
    - Combine the most relevant items from the datastore (d) with the input (x) for your task



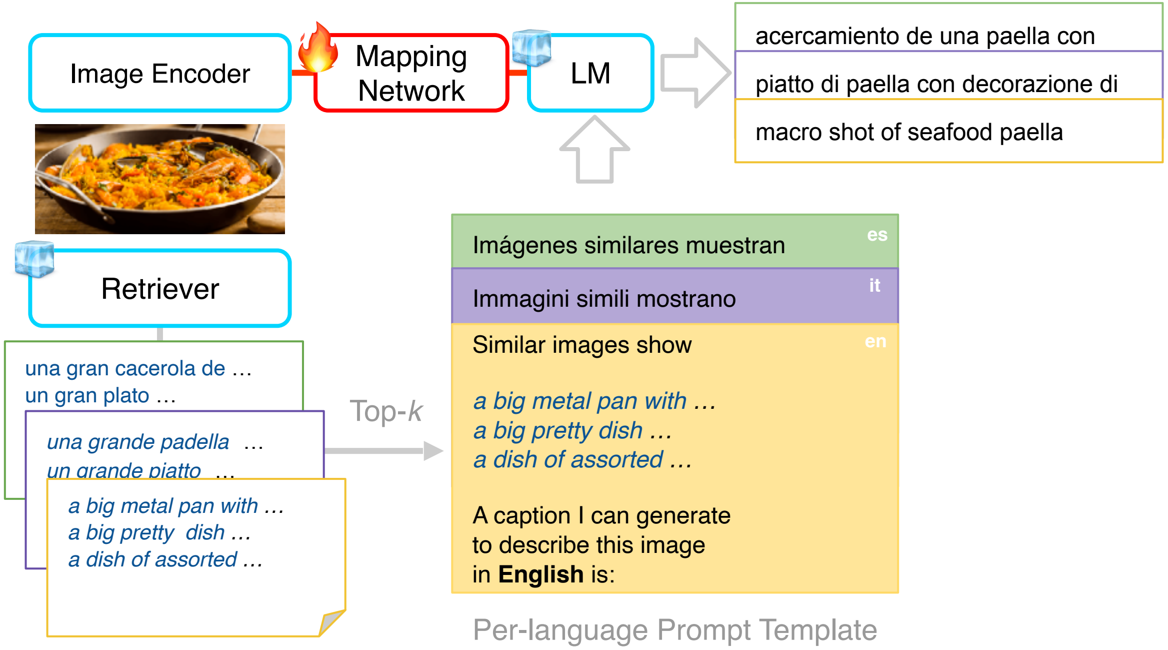
Izacard and Grave. EACL 2021. Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering.

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?

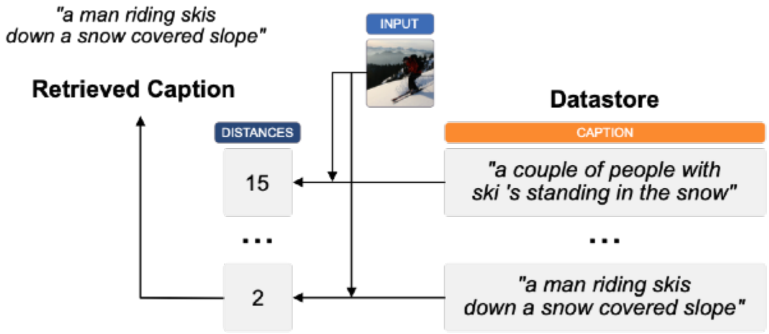


PAELLA Model



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



Johnson et al. (2019). Billion-scale similarity search with GPUs. IEEE Big Data

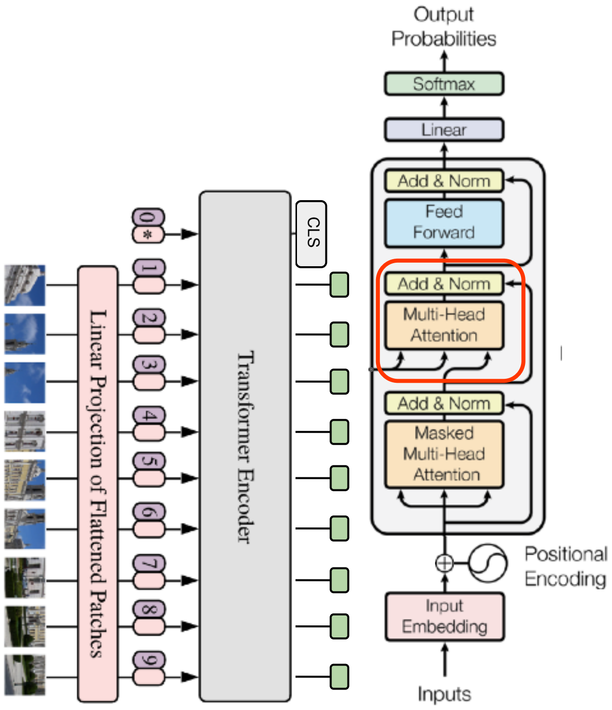
Trained Cross-Attention Layers

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

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

$$\mathrm{Att}(\mathbf{X}\mathbf{W}^Q, \mathbf{X}\mathbf{W}^K, \mathbf{X}\mathbf{W}^V)$$

$$\mathbf{W}^Q, \mathbf{W}^K, \mathbf{W}^V \in \mathbb{R}^{\mathrm{d\_enc} \times d}$$



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 $_{\text{COCO-35}}$	19M	2.6B	15.0	12.5
mBLIP: BLOOMZ-7B	135M	800M	23.4	6.7
BB+CC$_{\textit{COCO-35 + CC-35}}$	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

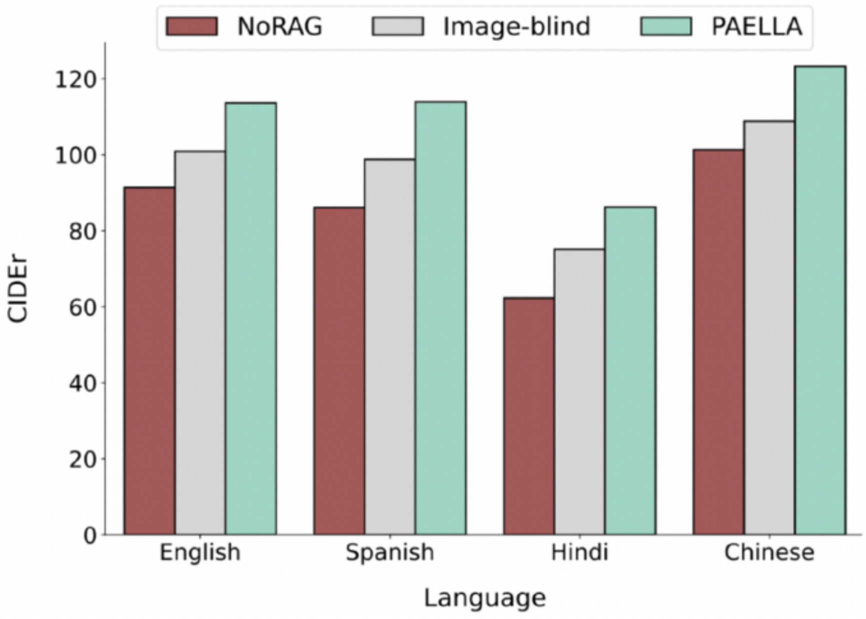
Zero-shot Multilingual Transfer

- **PAELLA$_{\text{mono}}$** 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. COCO-35	19M	2.6B	15.0	12.5
PAELLA$_{\text{mono}}$	566K$_{\text{en}}$	30M	15.5	12.1

Value of Retrieval Augmentation

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



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

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

detect things

generate captions

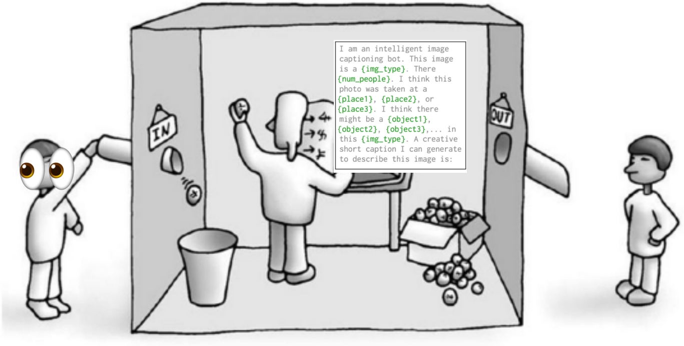
re-rank

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.

Zeng et al. (2023). Socratic Models: Composing Zero-Shot Multimodal Reasoning with Language. ICLR.



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

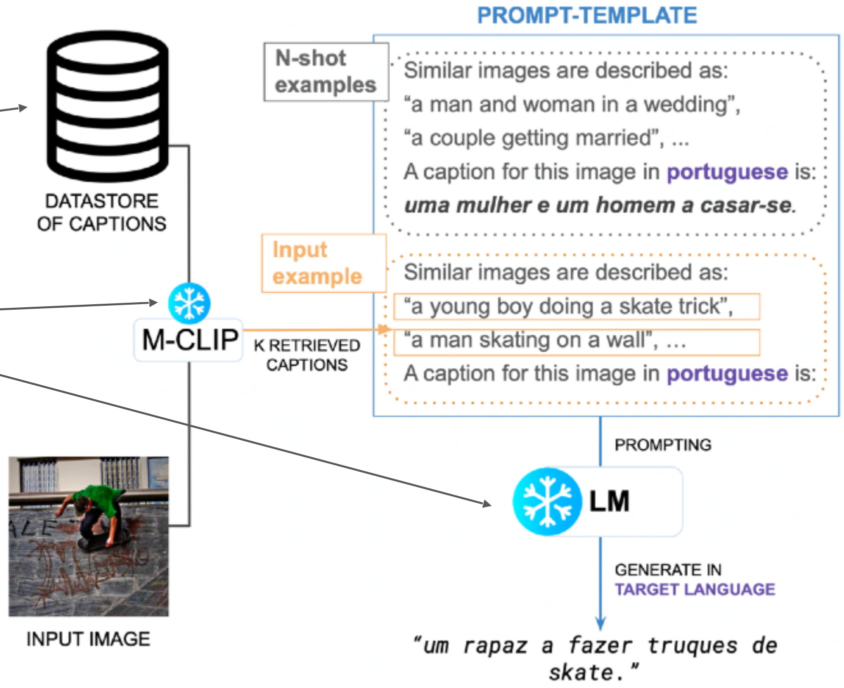
Multilingual Captioning with Retrieval Augmentation

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

detect things

generate captions

re-rank



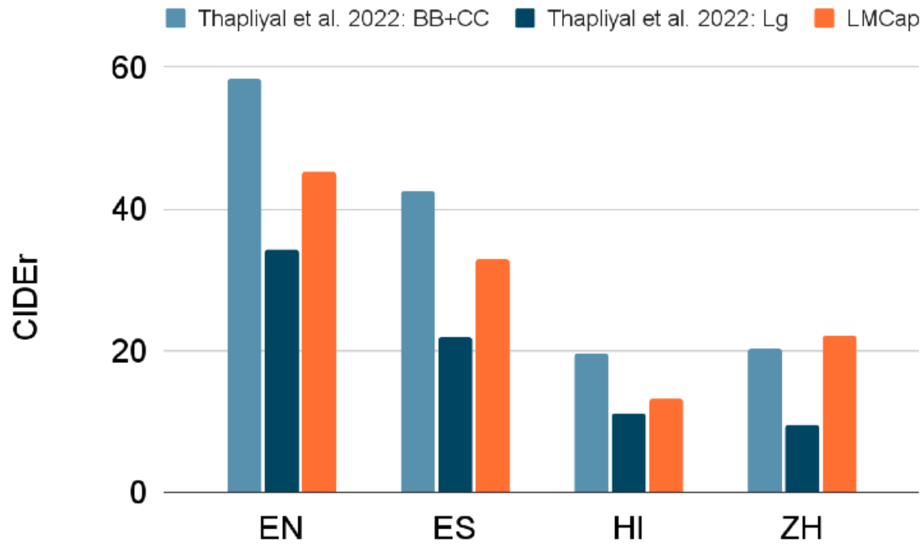
- Prompt with N-shot examples of transforming captions into the target language

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.**

Thapliyal et al. (2022). Crossmodal-3600: A massively multilingual multimodal evaluation dataset. EMNLP.

Results



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**

Competitive performance compared to supervised models

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: 两个大人和一个小男孩在雪地上滑雪

More examples in the paper

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

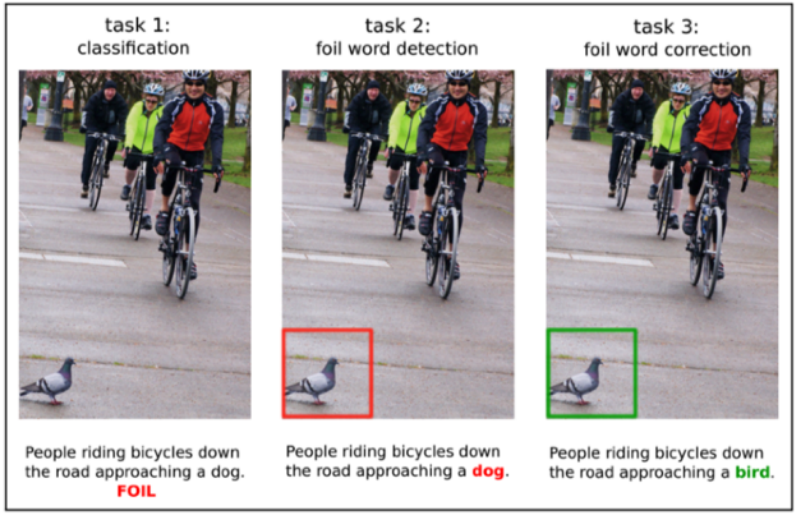
4. Understanding Multimodal Models

Beyond Task 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



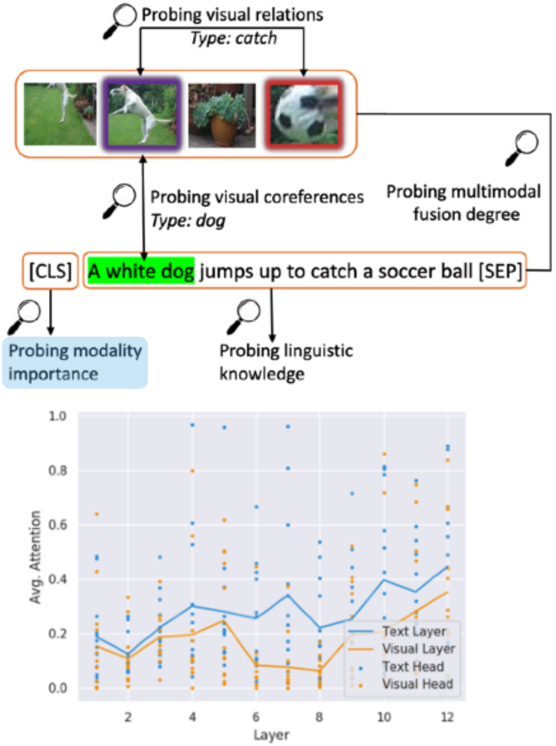
Shekhar et al. (2017). FOIL it! Find One mismatch between Image and Language caption. ACL.

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



Cao et al. (2020). Behind the scene: Revealing the secrets of pre-trained vision-and-language models. ECCV.

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.

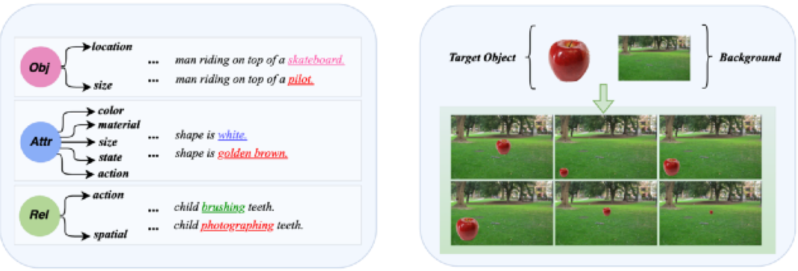
Metric	Model	Avg.
	Random	50.0
acc_r	GPT1*	60.7
	GPT2*	60.1
	CLIP	64.0
	LXMERT	59.6
	ViLBERT	63.7
	12-in-1	**75.1**
	VisualBERT	46.4

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

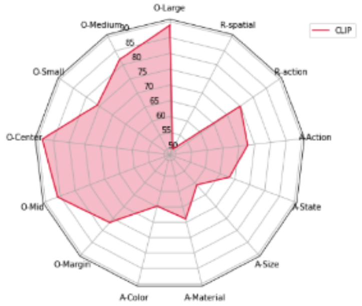
Parcalabescu et al. (2022). VALSE: A Task-Independent Benchmark for Vision and Language Models Centered on Linguistic Phenomena. ACL.

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



(a) Language Variation (b) Vision Variation



Zhao et al. (2022). An Explainable Toolbox for Evaluating Pre-trained Vision-Language Models. EMNLP.

Subject-Verb-Object Probes

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

Hendricks and Nematzadeh. (2021). Probing Image–Language Transformers for Verb Understanding. ACL.

WinoGround

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

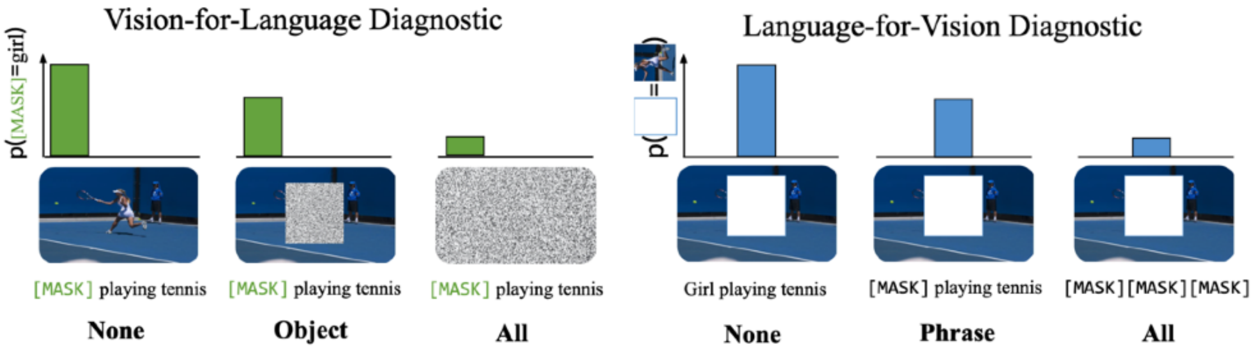
some plants surrounding a lightbulb

a lightbulb surrounding some plants

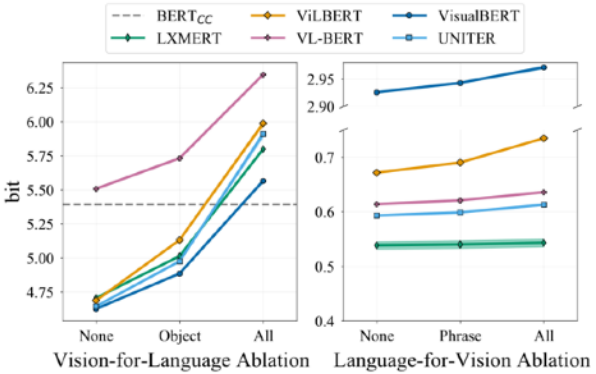
- Images sourced **with permission** from Getty.
- Differences are categorised into: swap dependent, swap-independent, and visual differences

Thrush et al. (2022). Winoground: Probing vision and language models for visio-linguistic compositionality. CVPR.

Vision-for-Language?



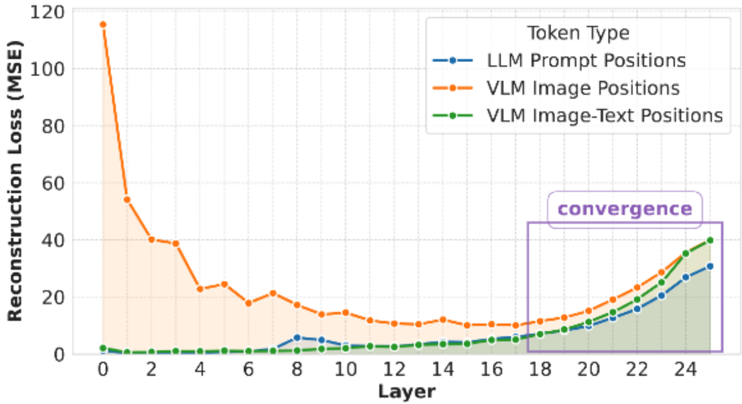
- Pair of diagnostic evaluations that can be applied to any model that makes MLM and MRC predictions.

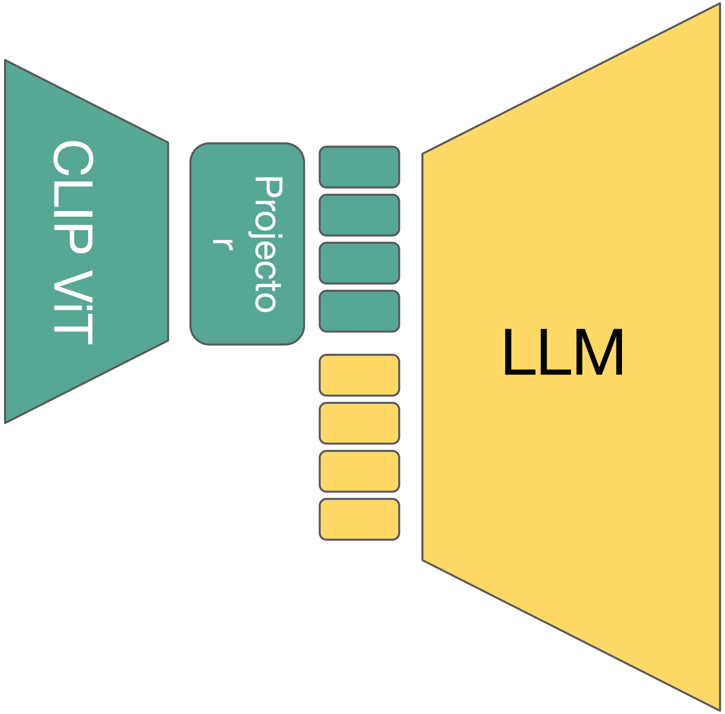


Frank et al. (2021). Vision-and-Language or Vision-for-Language? On Cross-Modal Influence in Multimodal Transformers. EMNLP.

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

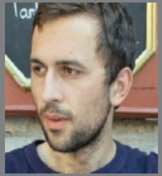




Venhoff et al. 2025. How Visual Representations Map to Language Feature Space in Multimodal LLMs

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

Findings of ACL 2025

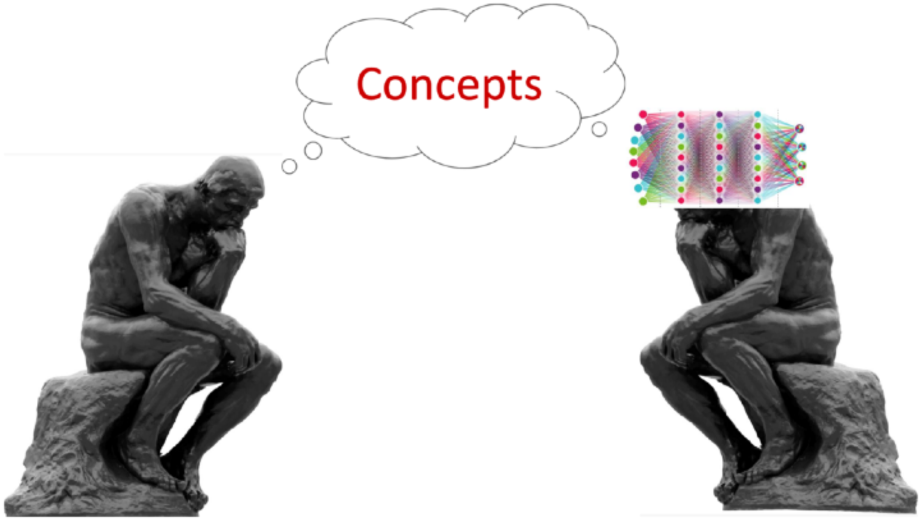


Dan Oneata

Desmond Elliott

Stella Frank

Main Question

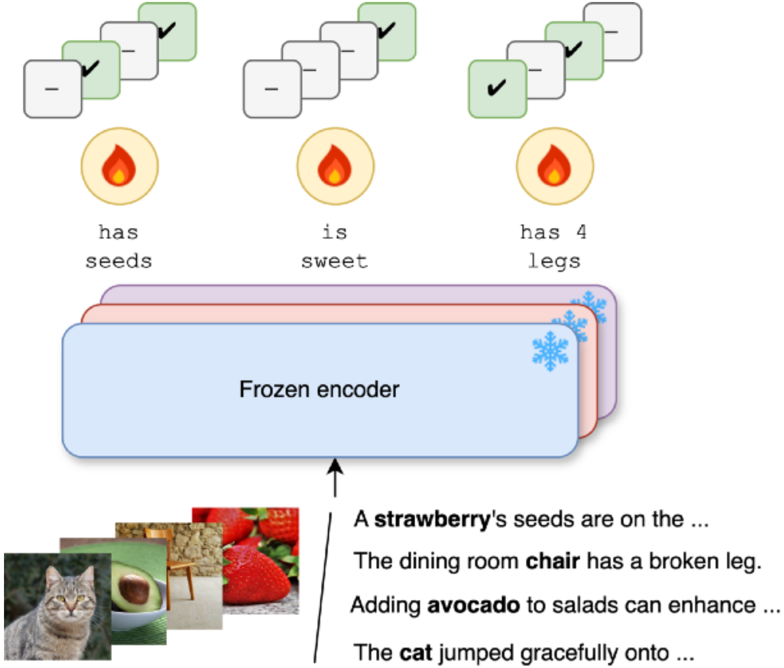


- How well do large-scale pretrained models represent the *semantic attributes* of concepts?
  - `ROSE` is *red*, *smells sweet*, *is a flower*

Approach

- Train linear probes to predict the semantic attributes of concepts from frozen encoder representations

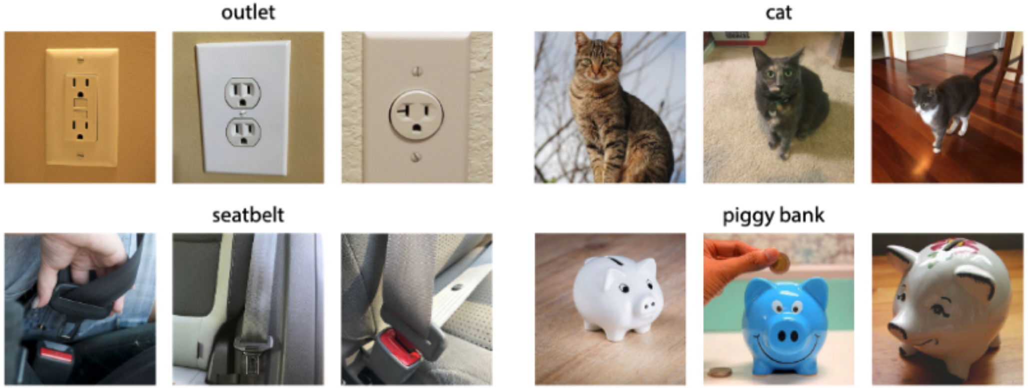
- Hard generalization task evaluating on *unseen* concepts
  - Train: **Cat**, **Dog**, **Cow**→ `has_four_legs`
  - Test: **Table**→ `has_four_legs`



Concepts: THINGS

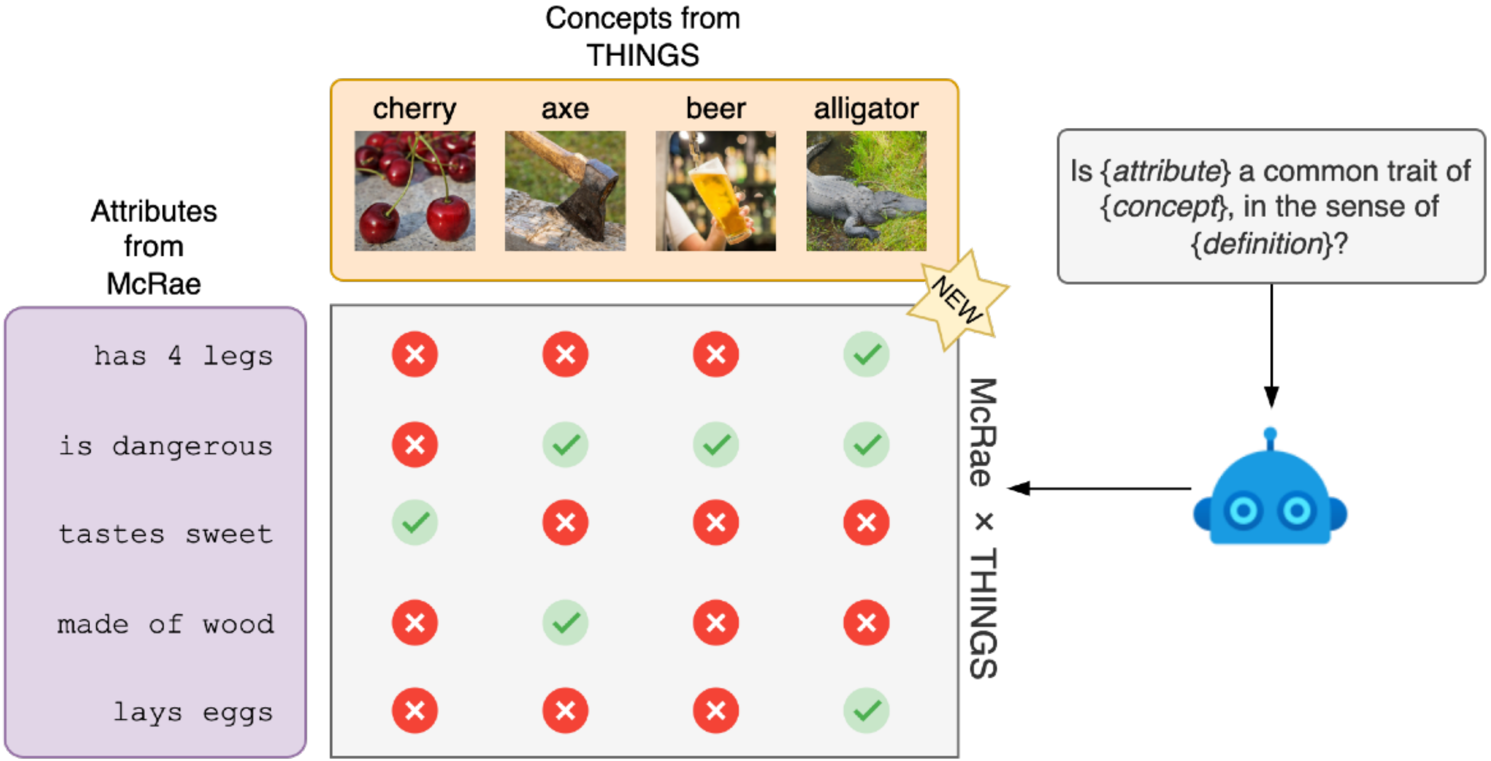
THINGS dataset of English-labeled concepts with curated images

- 1,854 concepts; 26,000 images



Hebart et al. 2019. PLOS One. THINGS: A database of 1,854 object concepts and more than 26,000 naturalistic object images

Semantic Attributes: McRae x THINGS

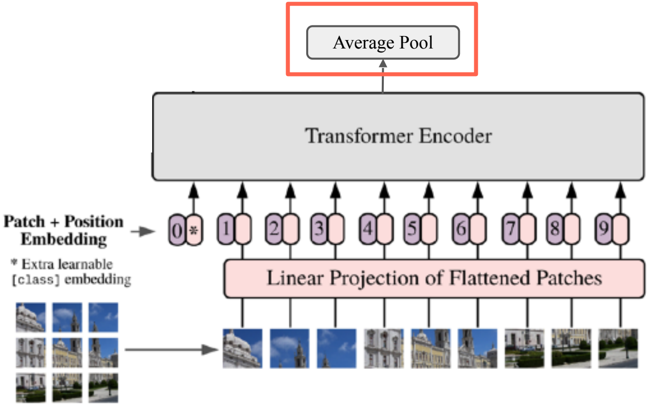


McRae et al. 2005. Behaviour Research Methods. Semantic feature production norms for a large set of living and nonliving things.

Extracting Concept Representations

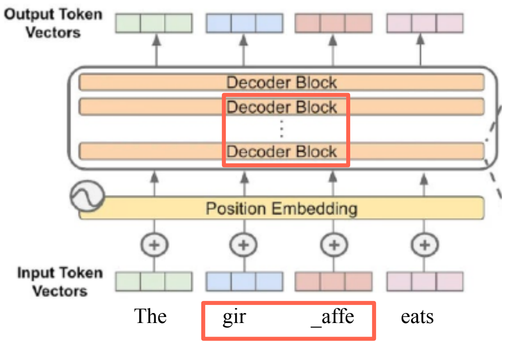
Visual Models

- Mean of the average pooled representation of the concept images

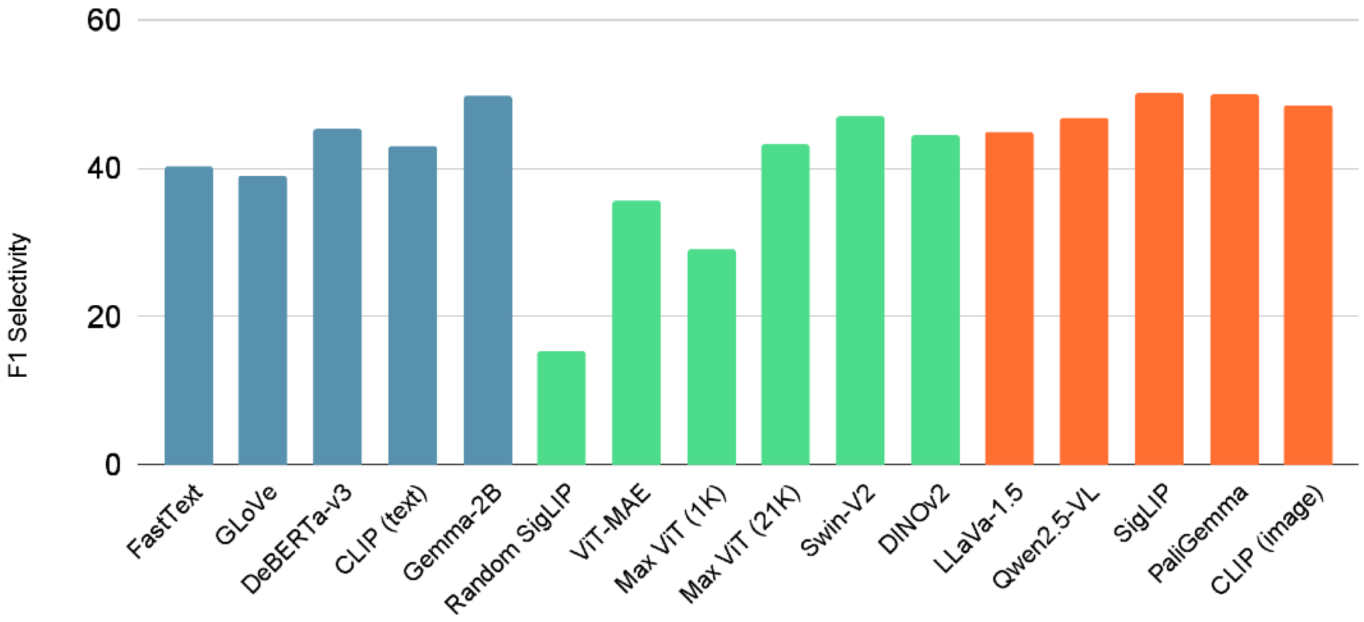


Language Models

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

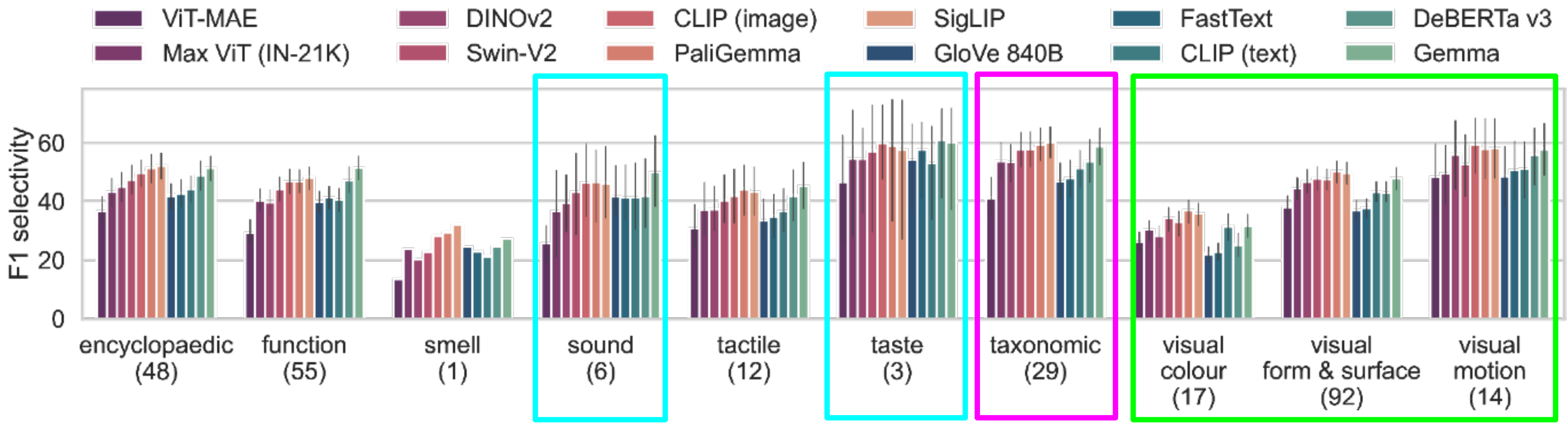


Main Results



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	ViT-MAE	Max ViT (IN-1K)	Max ViT (IN-21K)	DINOv2	Swin-V2	LLaVA-1.5	Qwen2.5-VL	CLIP (image)	PaliGemma	SigLIP	GloVe 840B	FastText	Numberbatch	CLIP (text)	DeBERTa v3	Gemma
Random SigLIP	1.0	.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	1.0	.96	.96	.89	.89	.91	.91	.90	.92
PaliGemma	.72	.91	.80	.93	.96	.96	.95	.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?

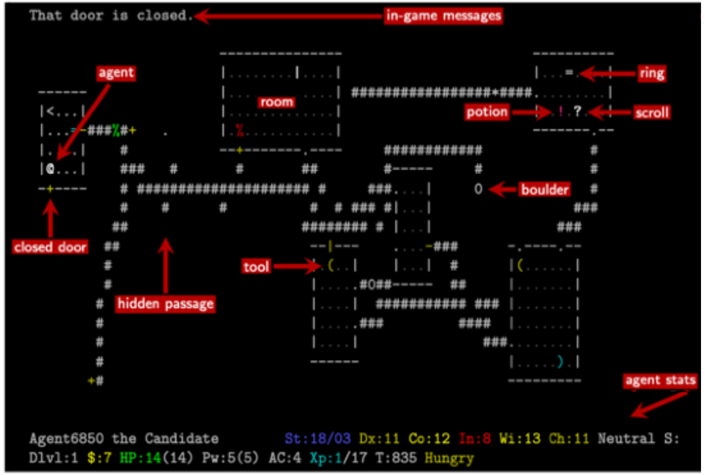
Q: What is the color of the last object that collided with the tiny red circle?

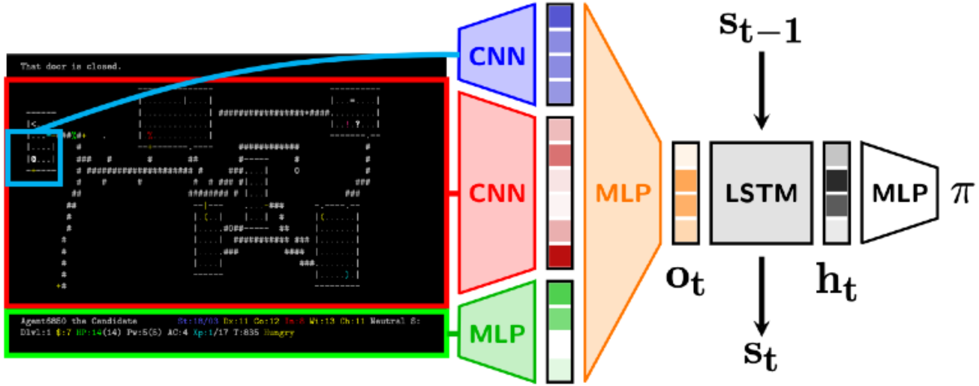
Q: If any of the other objects are removed, will the tiny green circle end up in the basket?

Ates et al. (2022). CRAFT: A Benchmark for Causal Reasoning About Forces and inTeractions. ACL.

Text-based Video Games

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

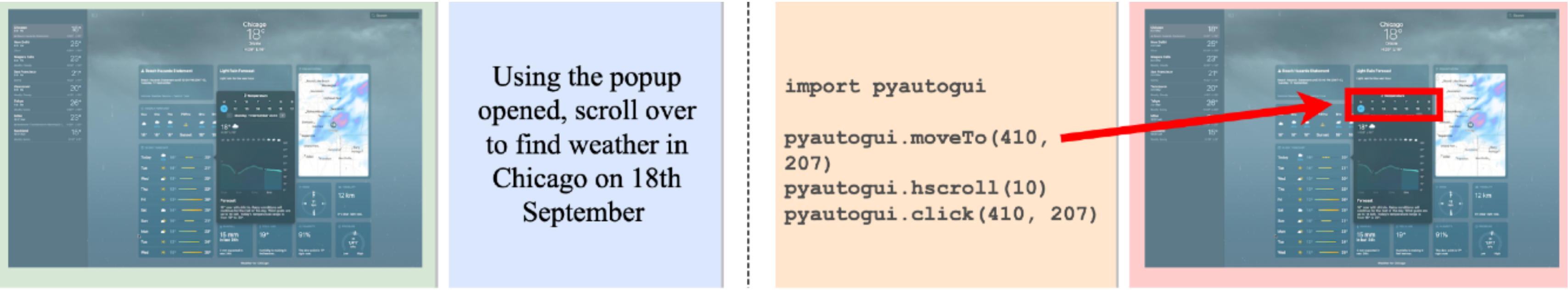




Küttler et al. (2020). The NetHack Learning Environment. NeurIPS.

Multimodal Agents

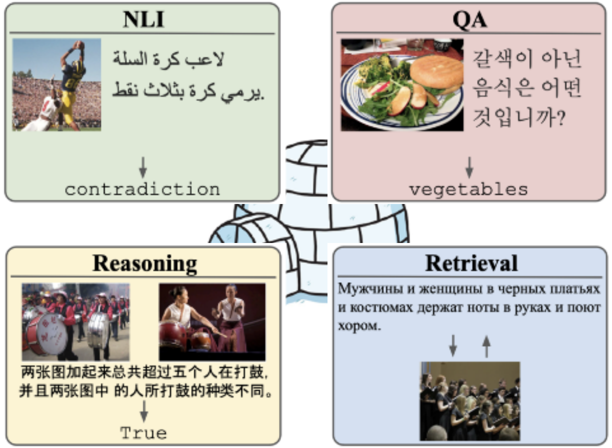
- OmniAct combines multimodal understand with program execution to solve a variety of tasks that humans perform with their computers



Kapoor et al. ECCV 2024. A Dataset and Benchmark for Enabling Multimodal Generalist Autonomous Agents for Desktop and Web.

Multilinguality

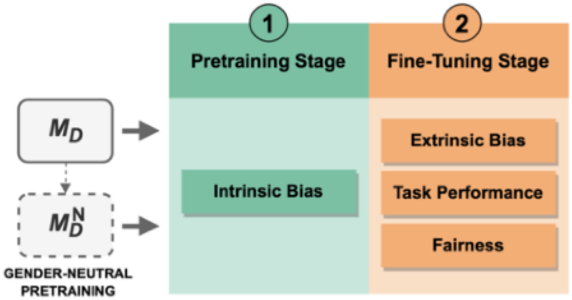
- The majority of Vision and Language research is in English
- We need resources, models, and evaluations to create useful multilingual multimodal models
- High-quality data requires:
  - time
  - money
  - community engagement

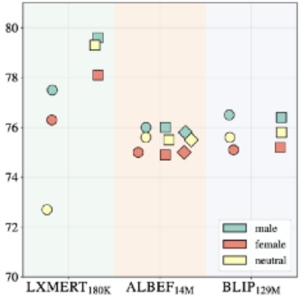


Bugliarello et al. ICML 2022. IGLUE: Image-Grounded Language Understanding Evaluation.

Multimodality for Social Good

- How can we limit the harm that models will unleash on society?





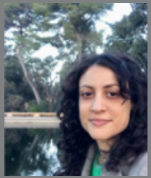
Model	Type of Response	%	%
xGen-MM		14.0	54.0
Qwen-2-VL		7.3	53.0
MiniCPM-2.6		7.3	9.0
InternVL-2		5.8	12.8
Idefics-3		4.5	42.0
InternLM		2.8	15.3
Cambrian		2.5	13.8
GPT-4o		1.0	5.5
Gemini-1.5		0.3	7.3
Claude-3.5		0	2.5

Cabello et al. EMNLP 2023. Evaluating Bias and Fairness in Gender-Neutral Pretrained Vision-and-Language Models
Röttger et al. 2025. MSTS: A Multimodal Safety Test Suite for Vision-Language Models

Q: How well do multimodal models understand implicit and explicit sexism?

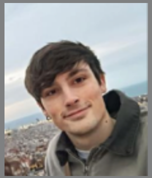
MuSeD: A Multimodal Spanish Dataset for Sexism Detection in Social Media Videos

COLM 2025



L. de Grazia

P. Pastells



M. V. Chas

D. Elliott

D. Sánchez

M. Farrús

M.Taulé

Warning: This part of the lecture contains sexist material. I use English translations of Spanish original text.

Motivation

- Online content is known to radicalize people by presenting them with an alternative and narrowly scoped world view
- Social media platforms like Facebook and Twitter have removed their content moderation teams in favour of “Community Notes”, which are not as effective as moderation (Borenstein et al. 2025)
- We need effective tools to detect and moderate harmful content to support the work of human moderators (Zeinert et al. 2021), which is harmful to those humans (Newton 2020)

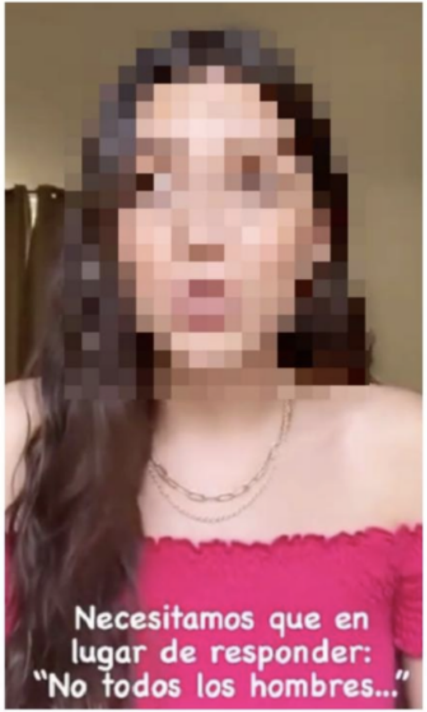
Borenstein and Warren et al. ACL 2025. Can Community Notes Replace Professional Fact-Checkers?
Zeinert et al. IJCNLP 2021. Annotating Online Misogyny.
Newton. The Verge, 2020. Facebook will pay $52 million in settlement with moderators who developed PTSD on the job.

Focus on Four Types of Sexism

Stereotyping

"My people, my people, look, enjoy these toys, enjoy these games that come with an iron, a washing machine, that come like this, because the feminists will soon protest that girls shouldn't play with this."

Non-sexist Content



- Reports sexism or denunciation of sexism

  “We need to respond not to all men. To defend that they are the exception, they better listen to us. We will be part of the solution when we understand what male privilege is. And let's question what sexist violence and toxic masculinity are.”

Prior Work

- Most work has focused on misogny and sexism in memes (Fersini et al. 2022, Plaza et al. 2024)
- No work on discrimination towards non-cisgender or heterosexual identities

Fersini et al. SemEval 2022. SemEval-2022 task 5: Multimedia automatic misogyny identification
Plaza et al. ECIR 2024. EXIST 2024: sEXism Identification in Social neTworks and Memes

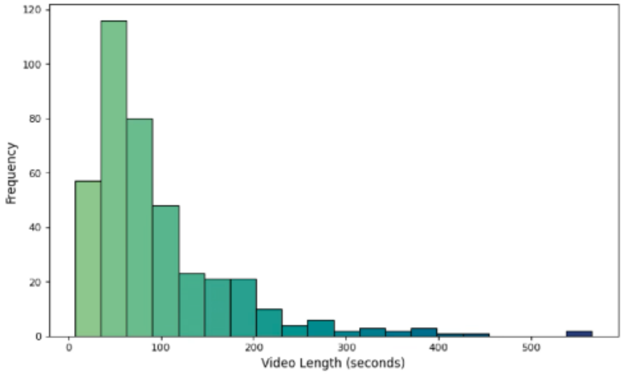
Challenges

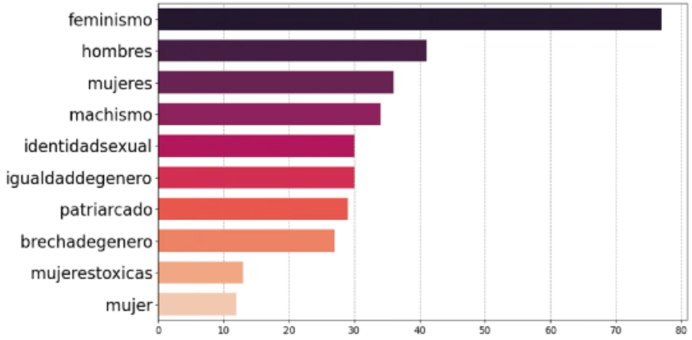
- Sexism in social media videos is complicated:
  - Indirect language
    - Irony, sarcasm, etc.
  - Many social media videos are comments on clipped content
  - Contradictions between verbal and non-verbal content

"I am a trans, non-binary person. More specifically, I am a demigirl. My sexual orientation is polysexual."

Dataset

- 400 videos scraped from TikTok and BitChute in 2024
  - 48.5% sexist / 51.5% non-sexist
  - 187 hashtags (#gender stereotypes, #feminism, #gender ideology, etc.)
  - Videos retrieved using Apify and BitChute DL
  - Transcript: Whisper-CTranslate2. Audio: `ffmpeg`. OCR: EasyOCR





Annotation

Stage 1	Stage 2	Stage 3

Annotator Teams

Two annotation Teams of three people (FFM)

- Expert annotator
- Each team annotates half of the dataset

Modality annotation

Three-step process to understand the importance of the transcript, audio, and video modalities

Sexism annotation

- Binary classification
- Fine-grained annotation
- Segment marking for video annotation

Evaluation

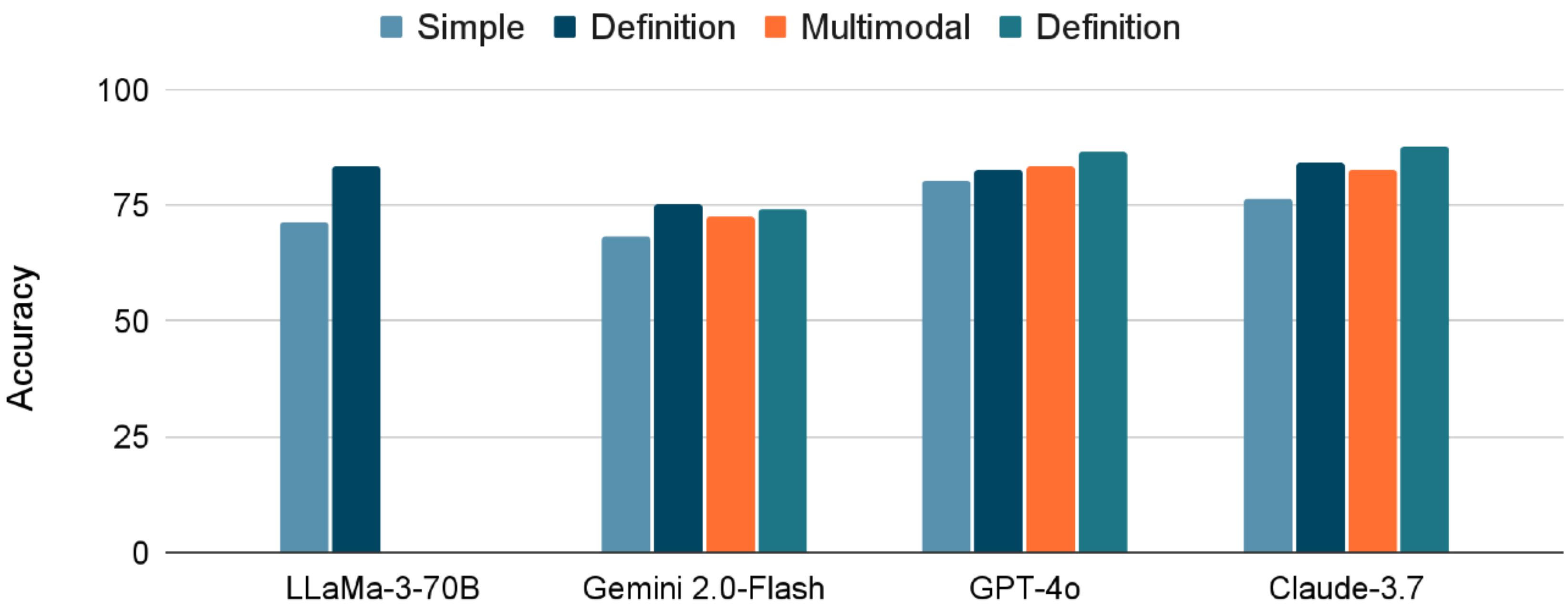
- We conducted an experiment on zero-shot recognition of sexism
  - Two prompts: (i) simple and (ii) rich definition

System prompt
Eres un asistente de clasificación de videos. Responde solo con 'Sí' o 'No'.

User prompt
Instrucción:
Determina si el texto es sexista. A continuación se presentan criterios que definen el sexismo:

- **Estereotipos:**
  - (1) Formulación de propiedades descriptivas que supuestamente distinguen a hombres y mujeres basadas en estereotipos de género.
  - (2) Formulación de propiedades prescriptivas que hombres y mujeres deben cumplir para encajar en los roles de género establecidos por la sociedad.
- **Desigualdad:**
  - (3) Contenido que niega la existencia de desigualdades (pasadas o presentes) entre hombres y mujeres.
  - (4) Contenido que se opone al feminismo, argumentando que este movimiento margina a los hombres.
- **Discriminación:**

Results



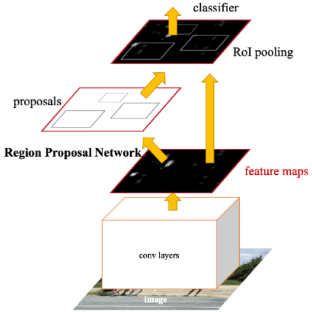
Definitions and multimodal inputs improve performance

Wrap-up

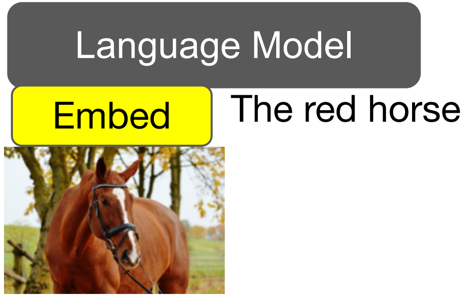
1. Datasets

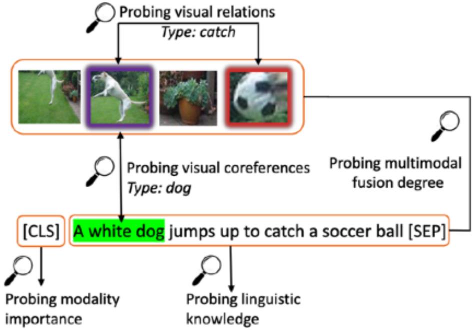
some sheep walking in the middle of a road
a herd of sheep with green markings walking down the road
a herd of sheep walking down a street next to a lush green grass covered hillside.
sheared sheep on roadway taken from vehicle, with green hillside in background.
a flock of freshly sheered sheep in the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

2. Representation

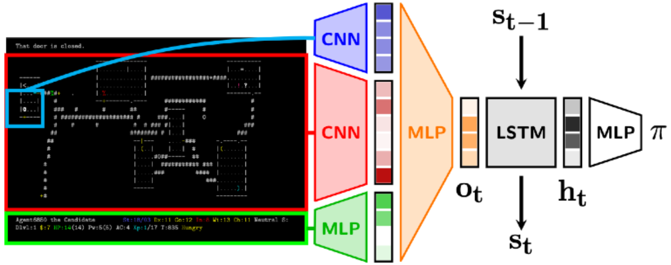


3. Modelling





4. Understanding



5. New Directions

Opportunities in Copenhagen

Danish
Data Science
Academy

Application deadlines: June

PHD FELLOWSHIP

POSTDOC FELLOWSHIP

DARA

Application deadline: August

PHD FELLOWSHIP

Acknowledgements


E. Bugliarello

R. Ramos

A. Kádár

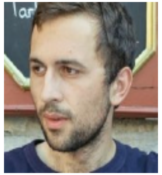
L. Specia

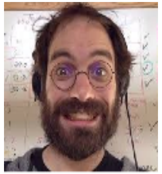
L. Barrault

D. Sánchez

N. Collier


G. Chrupała

C. Qui

D. Oneață

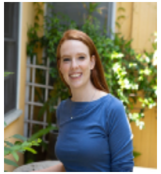
S. Frank

R. Sanabria

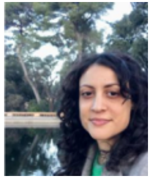
M. Fadaee

F. Liu


B. Martins

E. Hasler

E. Ponti

S. Hooker

S. Reddy

L. de Grazia

I. Salazar