Multimodal NLP

AthNLP 2025

Raquel Fernández

Institute for Logic, Language and Computation

University of Amsterdam

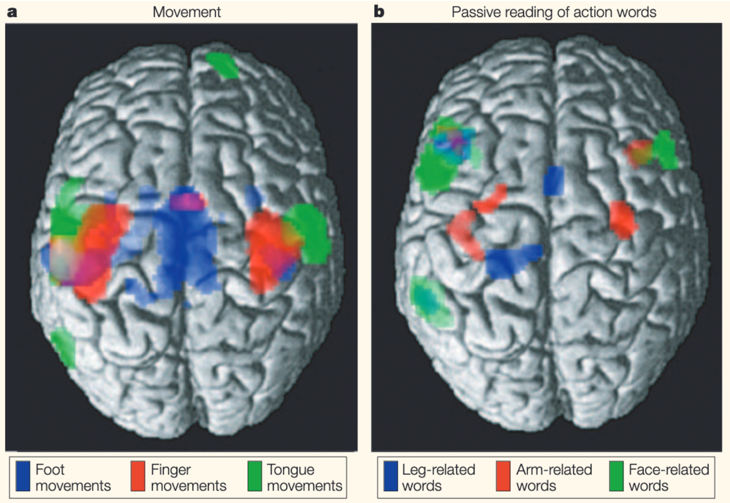
About me

- Barcelona ➤ London ➤ Potsdam ➤ Stanford ➤ Amsterdam
- Background in computational linguistic/NLP and cognitive science
- General interest in language use for communication – how language is shaped by perception and social interaction

http://www.illc.uva.nl/~raquel

Why multimodal NLP?

Pulvermüller. (2005). Brain mechanisms linking language and action. *Nature Reviews Neuroscience*, 6(7), 576-582.



According to theories of embodied cognition, conceptual knowledge encoded in language is grounded in our sensory-motor experience.

(Barsalou et al. 1998, Harvard, 1990, and many others)

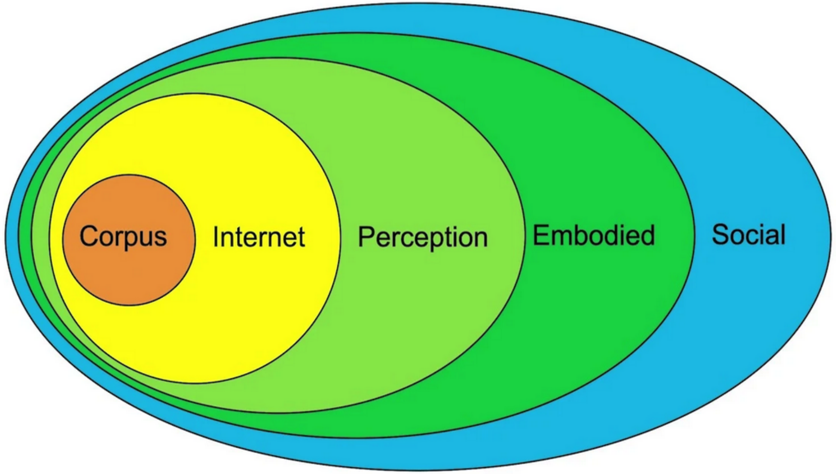
Why multimodal NLP?

… eating a **banana** after exercising…
… ripe **bananas** from Costa Rica…
… a fruit salad with **banana**, kiwi, …

According to theories of embodied cognition, conceptual knowledge encoded in language is grounded in our sensory-motor experience.

(Barsalou et al. 1998, Harvard, 1990, and many others)

Why multimodal NLP?



Bisk et al. Experience Grounds Language. EMNLP 2020.

According to theories of embodied cognition, conceptual knowledge encoded in language is grounded in our sensory-motor experience.
(Barsalou et al. 1998, Harvard, 1990, and many others)

Why multimodal NLP?

AI And The Limits Of Language

An artificial intelligence system trained on words and sentences alone will never approximate human understanding.

ESSAY TECHNOLOGY & THE HUMAN

BY JACOB BROWNING AND YANN LECUN

AUGUST 23, 2022

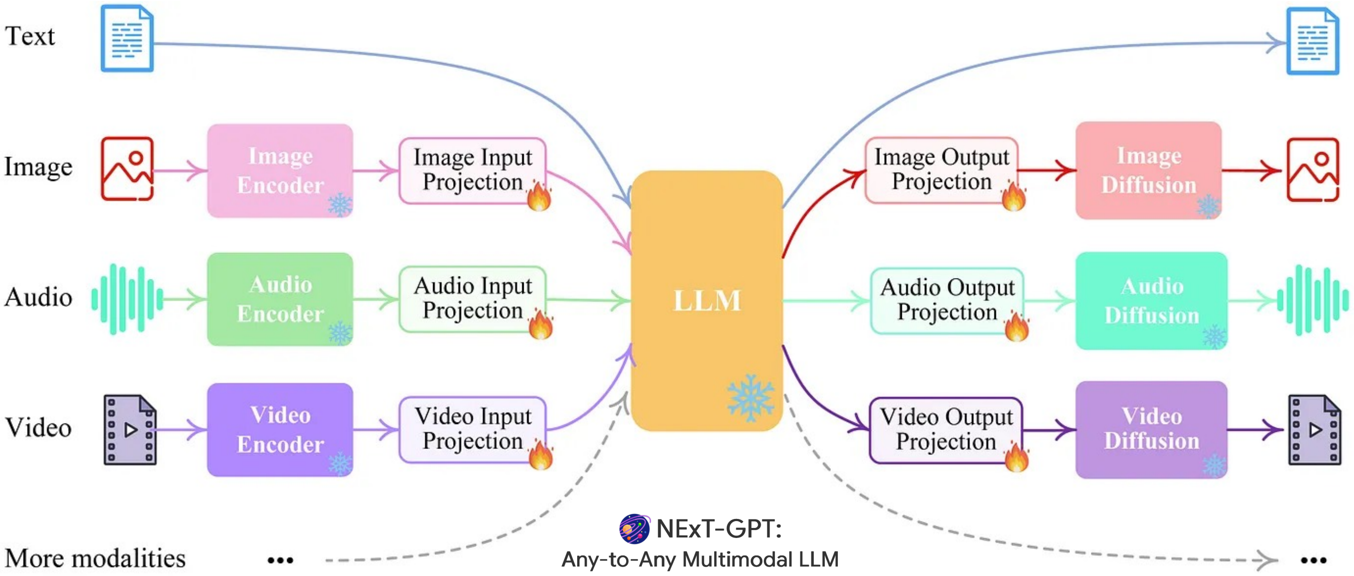
According to theories of embodied cognition, conceptual knowledge encoded in language is grounded in our sensory-motor experience.

(Barsalou et al. 1998, Harvard, 1990, and many others)

Current multimodal models

More than cognitive plausibility, the goal is to handle useful multimodal applications

Multimodal models process information from two or more *modalities* (i.e., means to convey information): text, speech, images, video, smells, sounds, actions, code, ….



NExT-GPT:
Any-to-Any Multimodal LLM

Current multimodal models

Multimodal models process information from two or more *modalities* (i.e., means to convey information): text, speech, images, video, smells, sounds, actions, code, ….

I will focus on the interplay between language & vision, from an NLP perspective

The Plan

Part 1

- Task-specific approaches: datasets and modelling techniques
- General purpose, pre-trained vision-language models (VLMs)

Part 2

- Evaluation of VLMs
- New directions in multimodal NLP

Task-specific approaches (historical notes)

- Datasets and architectures designed to tackle specific tasks

Key tasks in the early deep learning era (2014-2017)

Image Captioning

A group of people eating noodles

Visual Question Answering

What are the people eating?

- Noodles

Representing visual information

You have already seen how to learn text representations.

How do we represent information from other modalities, in particular vision?

- In very early approaches, **symbolic features** were used to represent objects or scenes, *without any vision*
- As computer vision methods started to be further developed, the focus shifted towards **automatically learning to represent visual information**

IMAGENET

14,197,122 images, 21841 synsets indexed

Home Download Challenges About

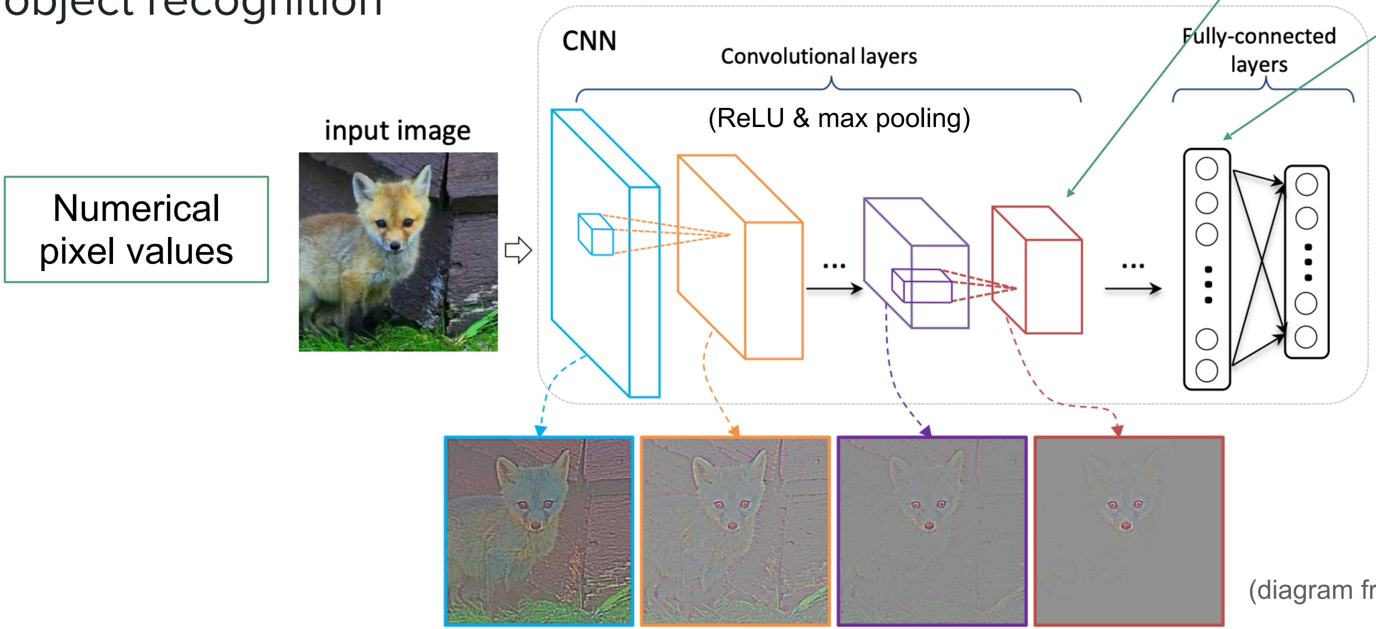
Not logged in. Login I Signup

ImageNet is an image database organized according to the WordNet hierarchy (currently only the nouns), in which each node of the hierarchy is depicted by hundreds and thousands of images. The project has been instrumental in advancing computer vision and deep learning research. The data is available for free to researchers for non-commercial use.

https://www.image-net.org

CNNs

Earliest work in the neural-network era used features from CNNs pre-trained on object recognition



class name	prob
kit fox	0.5956
red fox	0.3576
grey fox	0.0439
coyote	0.0013
Arctic fox	0.0003

(diagram from Yu et al, ICML 2016)

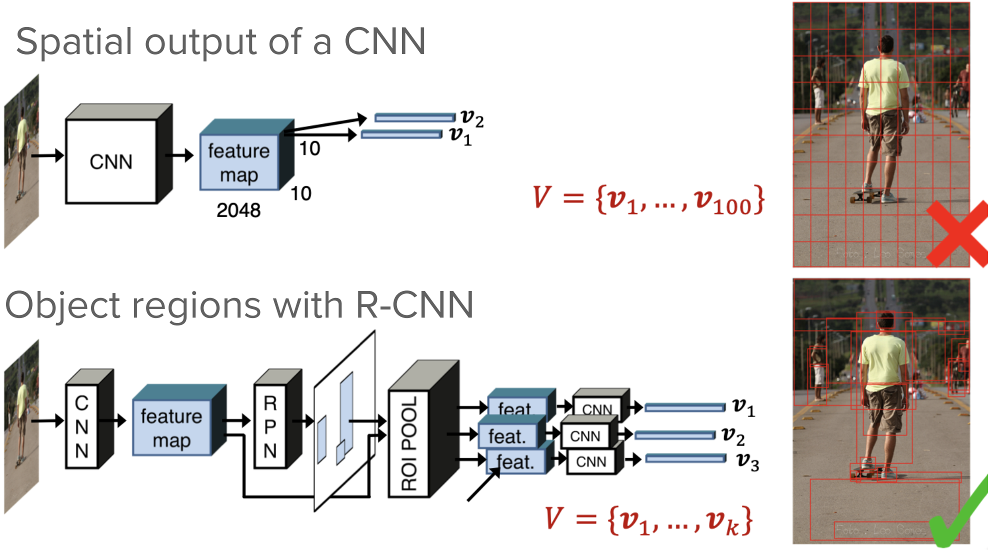
Zisserman & Simonyan, 2015, Very Deep Convolutional Networks for Large-Scale Image Recognition, ICLR.

CNNs: Features for regions of interest

From general spatial information to regions corresponding to objects/entities

R-CNN region-based feature vectors:

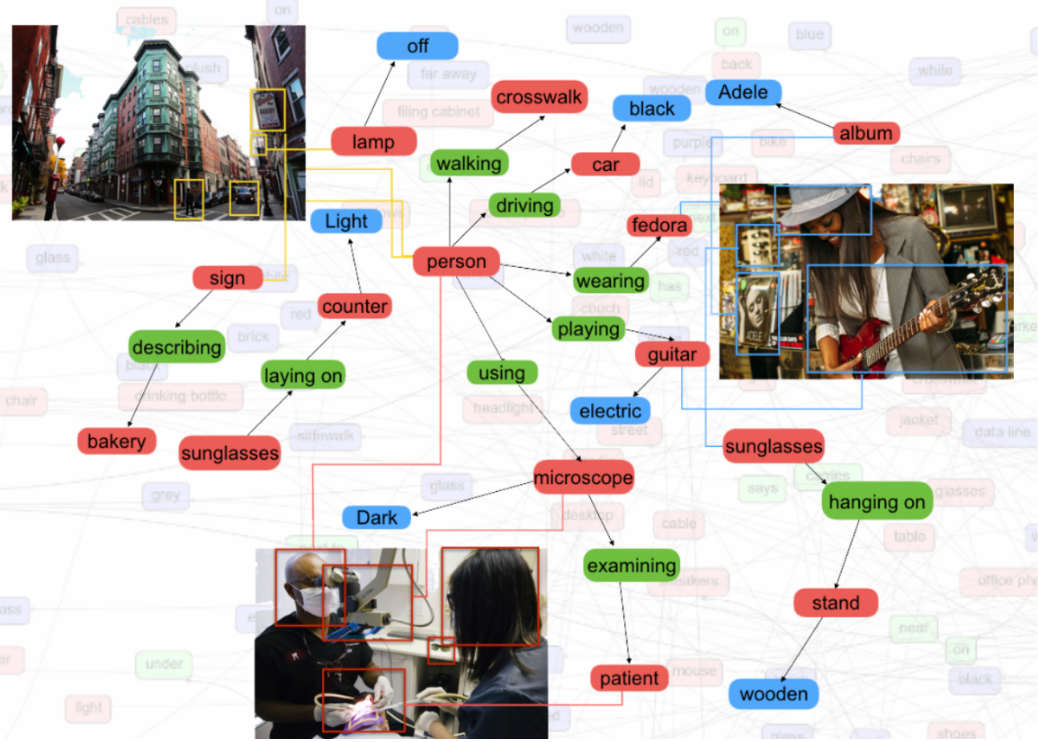
- Trained on the Visual Genome Dataset for object recognition.
- The Region Proposal Network suggests the location of *regions of interest*.



Ren et al. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. *NeurIPS*.
Anderson et al. 2018. Bottom-Up and Top-Down Attention for Image Captioning and VQA. CVPR.
Diagram from poster: `https://panderson.me/images/cvpr18_UpDown_poster.pdf`

Visual Genome

https://homes.cs.washington.edu/~ranjay/visualgenome/



Visual Genome is a dataset, a knowledge base, an ongoing effort to connect structured image concepts to language.

108,077 Images
5.4 Million Region Descriptions
1.7 Million Visual Question Answers
3.8 Million Object Instances
2.8 Million Attributes
2.3 Million Relationships
Everything Mapped to Wordnet Synsets

Krishna et al. (2017) Connecting Language and Vision using Crowdsoruced Dense Image Annotations, IJCV.

COCO: Common Objects in Context

Highly influential dataset

Multiple human-authored captions, with object segmentation.

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.

What is COCO?

COCO is a large-scale object detection, segmentation, and captioning dataset. COCO has several features:

- Object segmentation
- Recognition in context
- Superpixel stuff segmentation
- 330K images (>200K labeled)
- 1.5 million object instances
- 80 object categories
- 91 stuff categories
- 5 captions per image
- 250,000 people with keypoints

cocodataset.org

Lin et al. (2014), COCO: Common Objects in Context.
Chen et al. (2015), Microsoft COCO captions: Data collection and evaluation server.

Multi30K: Multilingual aligned image-sentence dataset

- English, German, French, Czech, Arabic, Japanese, Turkish, Ukrainian

A group of people are eating noodles.

Eine Gruppe von Leuten isst Nudeln.

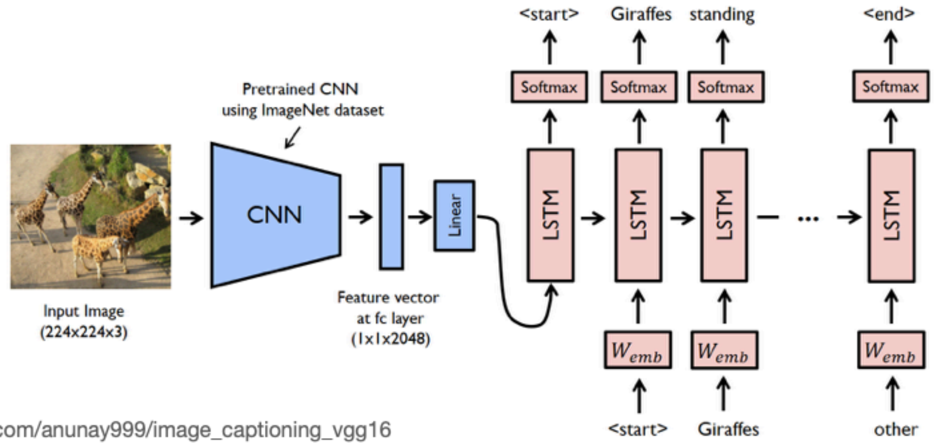
Un groupe de gens mangent des nouilles.

Skupina lidí jedí nudle.

Elliott et al. (2016). Multi30K: Multilingual English-German Image Descriptions. ACL Workshop.

Task-specific models: Image captioning

Encoder-decoder architecture: a language model conditioned on visual information



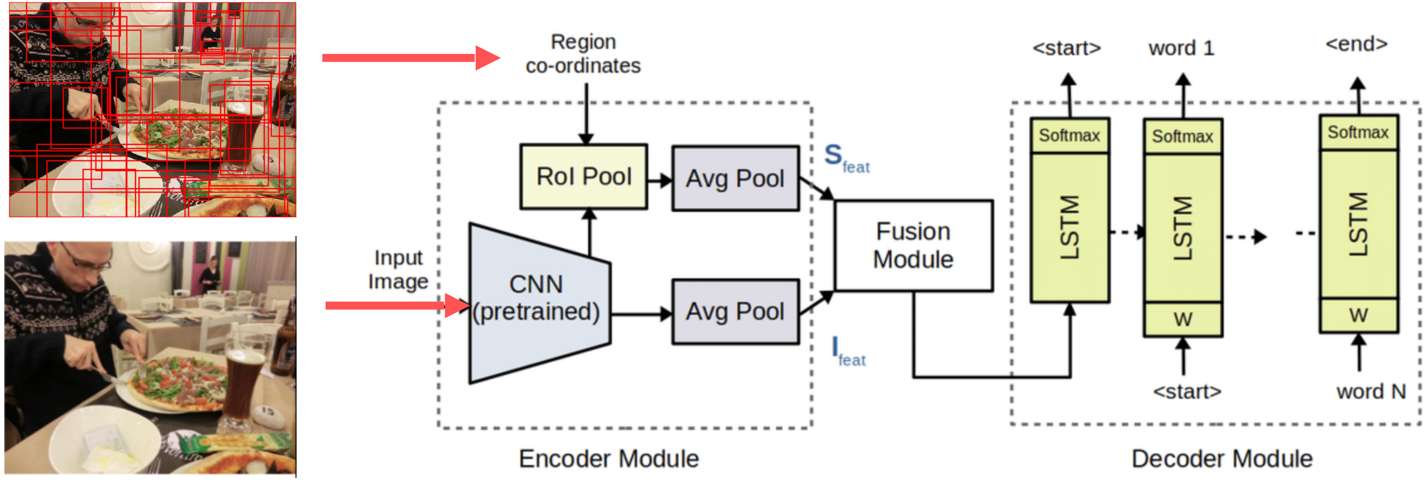
https://github.com/anunay999/image_captioning_vgg16

This basic architecture can be enriched in different ways…

Task-specific models: Image captioning

Encoder-decoder architecture: a language model conditioned on visual information

- Enriching by additionally using visual features for **regions of interest**, **attention** over these features, etc.



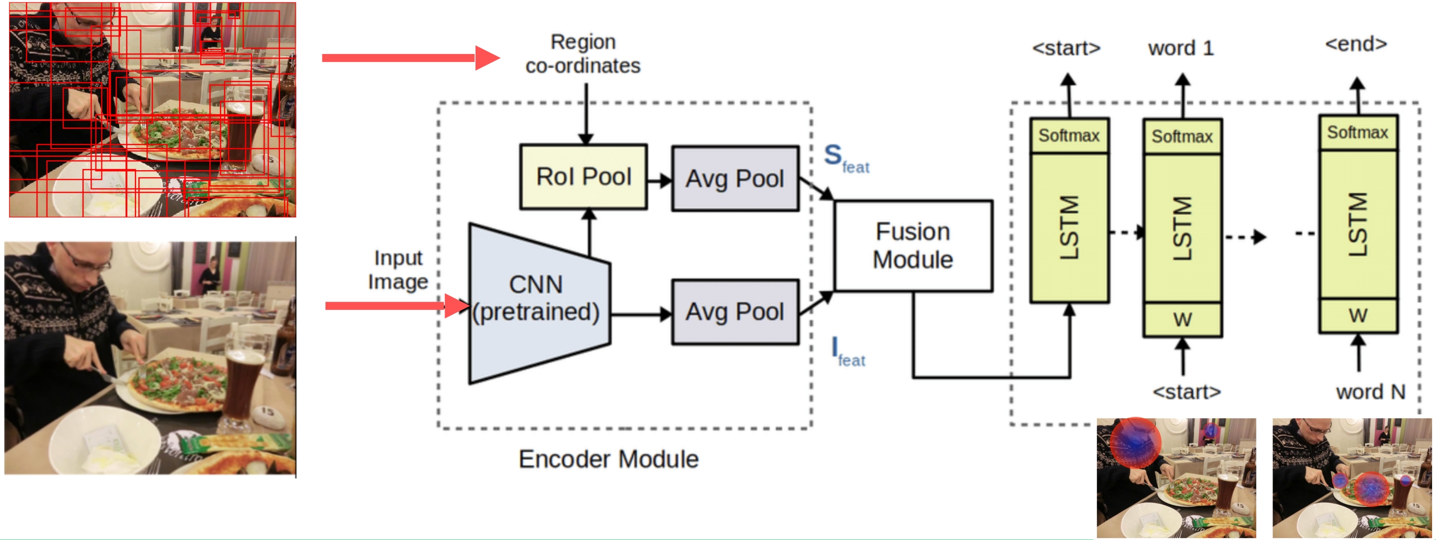
https://www.researchgate.net/figure/Architecture-of-the-proposed-model-for-region-specific-image-caption-generator-The_fig2_353489235

Task-specific models: Image captioning

Encoder-decoder architecture: a language model conditioned on visual information

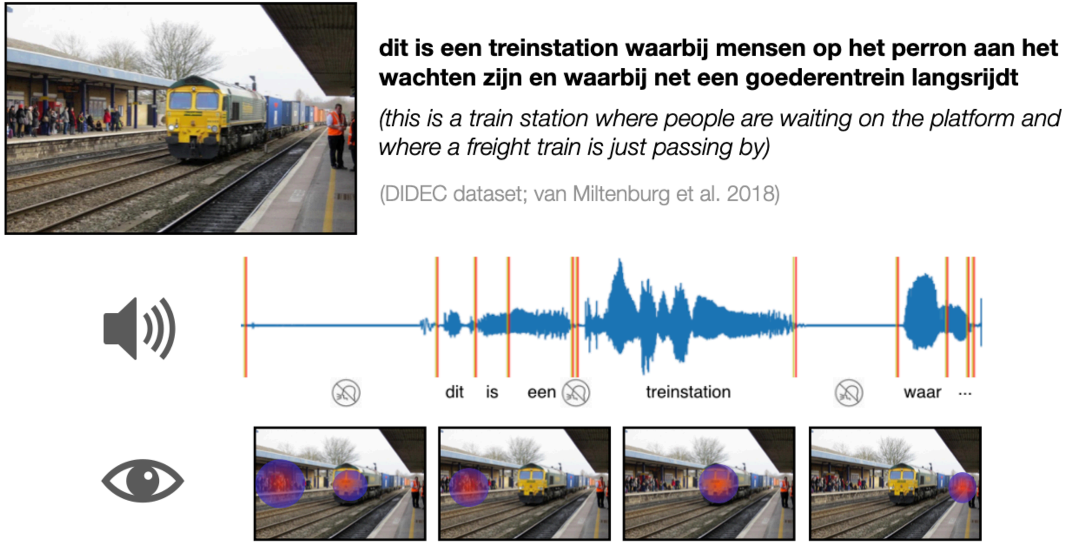
- Enriching by using information from **human gaze**, exploiting its sequential nature!

Static



sequential

DIDEC: Dutch image description eye-tracking corpus



Among other things, such a dataset allows us to investigate sequential cross-modal alignment

(van Miltenburg et al, 2018)

Generating Image Descriptions via Sequential Cross-Modal Alignment Guided by Human Gaze

Ece Takmaz[1], Sandro Pezzelle[1], Lisa Beinborn[2], Raquel Fernández[1]
[1]Institute for Logic, Language and Computation, University of Amsterdam
[2]Vrije Universiteit Amsterdam

(EMNLP 2020)

- Eye tracking coupled with language production as a guide to image description generation

- Adaptation of image captioning model by Anderson et al. (2018)

- This leads to more specific and human-like descriptions

	specificity	**disfluency**
NO-G	een vrouw die in de keuken staat. . . *(a woman who is standing in the kitchen. . .)*	een foto van een straat met een aantal vogels *(a photo of a street with a number of birds)*
2SEQ	een vrouw in een keuken met **donuts** *(a woman in the kitchen with **donuts**)*	**uh uh uh uh** met een aantal vogels *(**uh uh uh uh** with some birds)*

R: een bakkerij met een rek met een heleboel donuts
(a bakery with a rack with a lot of donuts)

R: uh allemaal duiven
(uh all [full of] pigeons)

Visual Question Answering

- Answer questions about images
- Multimodal input: Image & Question
- Commonly tackled as a classification
- VQA dataset: around 600k image-question pairs

Antol et al. (2015). VQA: Visual Question Answering. ICCV.

Task-specific models: VQA

Specifically designed for and trained on the VQA task

Again, this basic architecture can be enhanced with R-CNN region features, attention, etc.

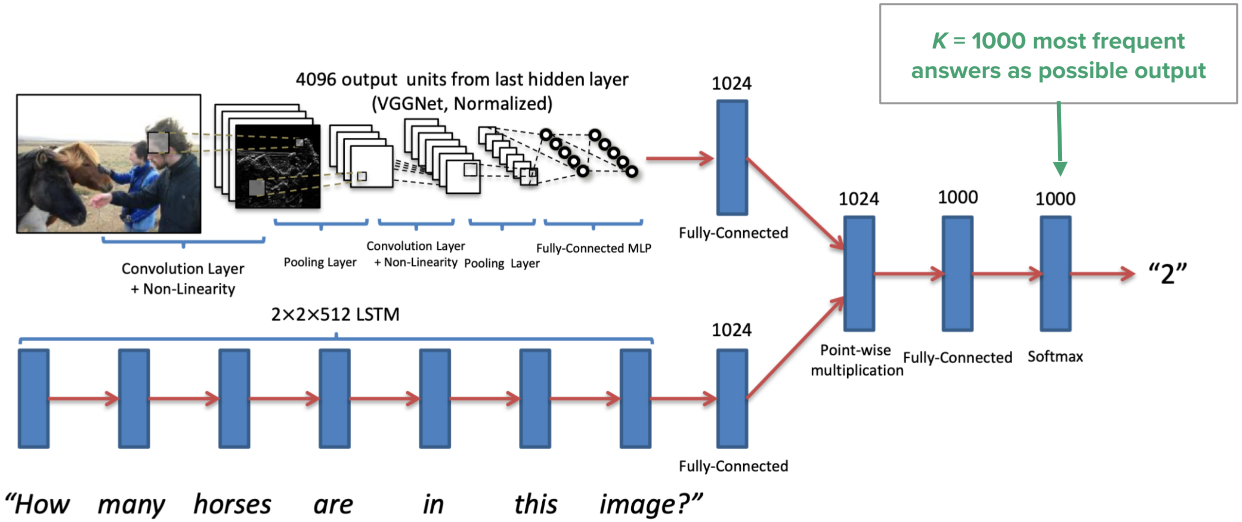


Fig. 8: Our best performing model (deeper LSTM Q + norm I). This model uses a two layer LSTM to encode the questions and the last hidden layer of VGGNet [48] to encode the images. The image features are then ℓ_2 normalized. Both the question and image features are transformed to a common space and fused via element-wise multiplication, which is then passed through a fully connected layer followed by a softmax layer to obtain a distribution over answers.

Antol et al. (2015). VQA: Visual Question Answering. ICCV.

Image Captioning

A group of people eating noodles

Visual Question Answering

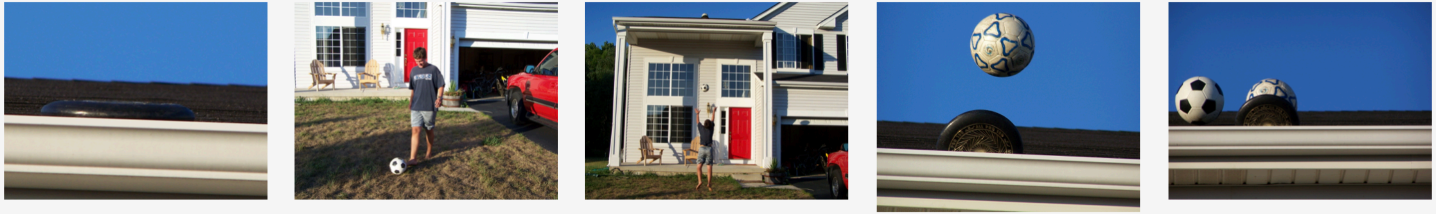
What are the people eating?

- Noodles

Over the years, more complex variants of these basic tasks were formulated. For example: **Visual story telling** and **visual question answering dialogue**

Visual storytelling

- VIST: 5 images from the same Flickr album (around 20k sequences in total)
- Crowdsourced stories: one sentence per image; several stories per image sequence



	1	2	3	4	5
Image description in isolation →	A black frisbee is sitting on top of a roof.	A man playing soccer outside of a white house with a red door.	The boy is throwing a soccer ball by the red door.	A soccer ball is over a roof by a frisbee in a rain gutter.	Two balls and a frisbee are on top of a roof.
Sequence description (story) →	**A discus got stuck up on the roof.**	**Why not try getting it down with a soccer ball?**	**Up the soccer ball goes.**	**It didn't work so we tried a volley ball.**	**Now the discus, soccer ball, and volleyball are all stuck on the roof.**

Huang et al. 2016. Visual Storytelling, NAACL.

VQA dialogue with multi-turn interactions

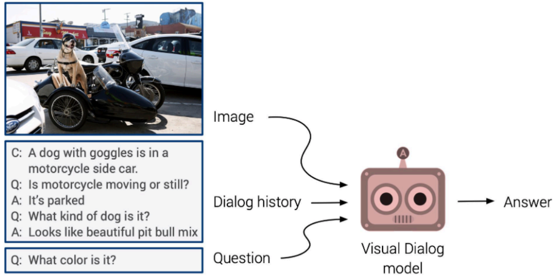
Visual Dialogue https://visualdialog.org

(Das et al., CVPR 2017)

Image + Caption	Human-Human dialog [4]	
	What are their genders?	1 man 1 woman
	Are they both adults?	Yes
	Do they wear goggles?	Looks like sunglasses
	Do they have hats on?	Man does
	Are there any other people?	No
	What color is man's hat?	Black
	Is it snowing now?	No
	What is woman wearing?	Blue jacket and black pants
	Are they smiling?	Yes
	Do you see trees?	Yes

Tasks:

- Question answering



- Image retrieval

GuessWhat

https://github.com/GuessWhatGame/guesswhat

(De Vries et al., CVPR 2017)

Questioner	Oracle
Is it a vase?	Yes
Is it partially visible?	No
Is it in the left corner?	No
Is it the turquoise and purple one?	Yes

Tasks:

- Asked informative questions
- Locate the target object, given image and dialogue history

Multi-turn interactions

PhotoBook dataset: more natural visually grounded dialogue

Participant A

Round 1 of 5

A: Hi

B: Hello

B: do you have a white cake on multi colored striped cloth?

A: I see a guy taking a picture. What about you?

B: is it of a cake with construction trucks on it?

A: Yeah. I don't see the cake you mentioned.

A: `<common img_2>`

Participant B

Repeated references to the same image

1	I see a guy taking a picture. What about you?
2	guy with camera
3	I have the guy with camera
4	the last one is the camera guy

Haber et al., The PhotoBook dataset: Building common ground through visually grounded dialogue, ACL 2019

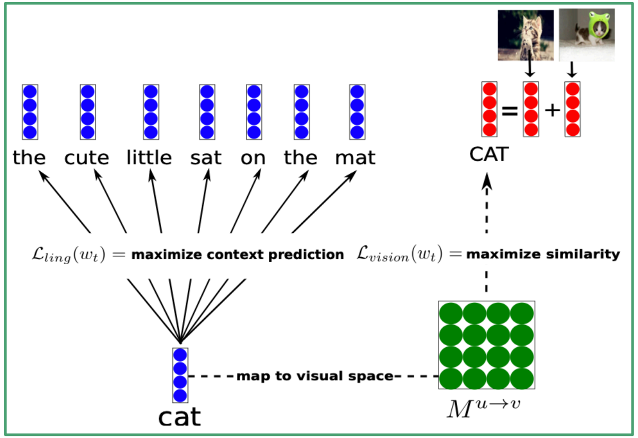
General purpose, pre-trained vision-language models (VLMs)

- Early multimodal encoders
- Cross-modal alignment
- Generative VLMs

- Large-scale training data

Enriching representations of text LMs with visual grounding

Word2vec: word-type embeddings



Lazaridou et al. (2015) Combining language and vision with a multimodal Skip-gram model, NAACL.

BERT: contextualized word embeddings

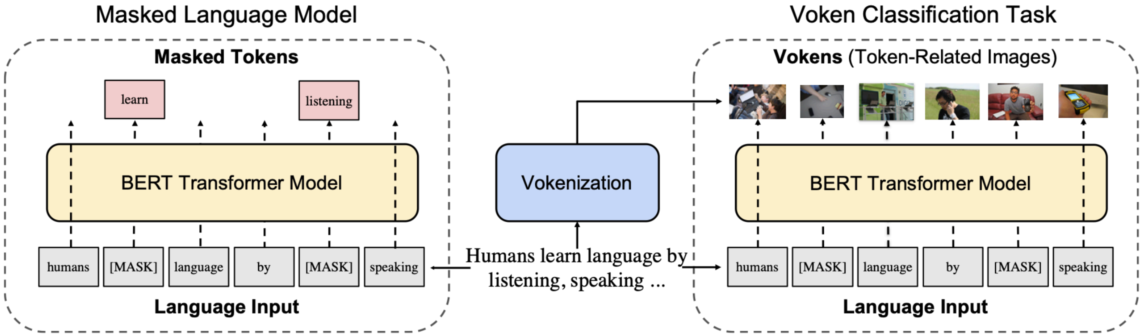
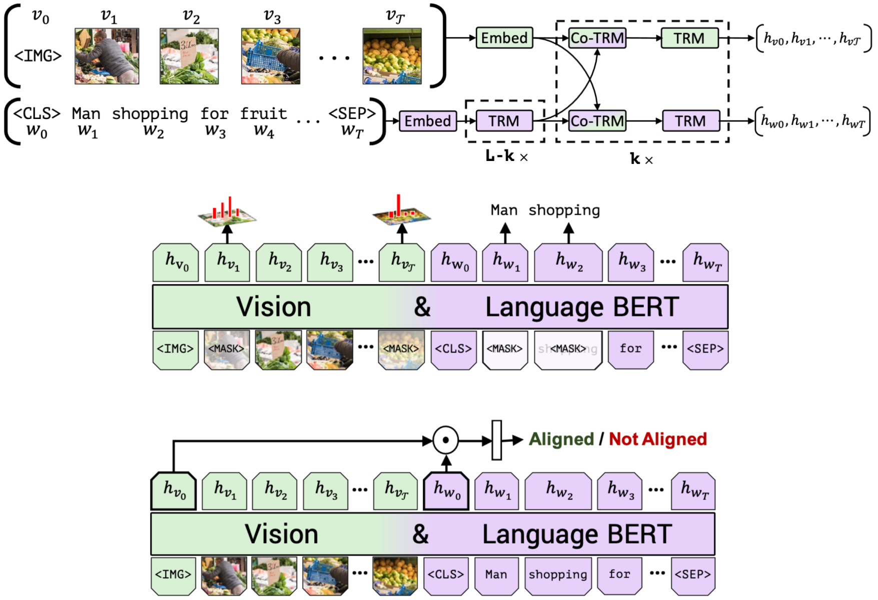


Figure 2: Illustration of the BERT transformer model trained with a visually-supervised language model with two objectives: masked language model (on the left) and voken classification (on the right). The first objective (used in original BERT pre-training) predicts the masked tokens as self-supervision while the second objective predicts the corresponding vokens (contextually generated by our vokenization process) as external visual supervision. Since the inputs are the same, we optimize the two objectives simultaneously and share the model weights.

Tan & Bansal (2020). Vokenizationn: Improving Language Understanding with Contextualized, Visual-Grounded Supervision. EMNLP.

Multimodal encoders: ViLBERT

- 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



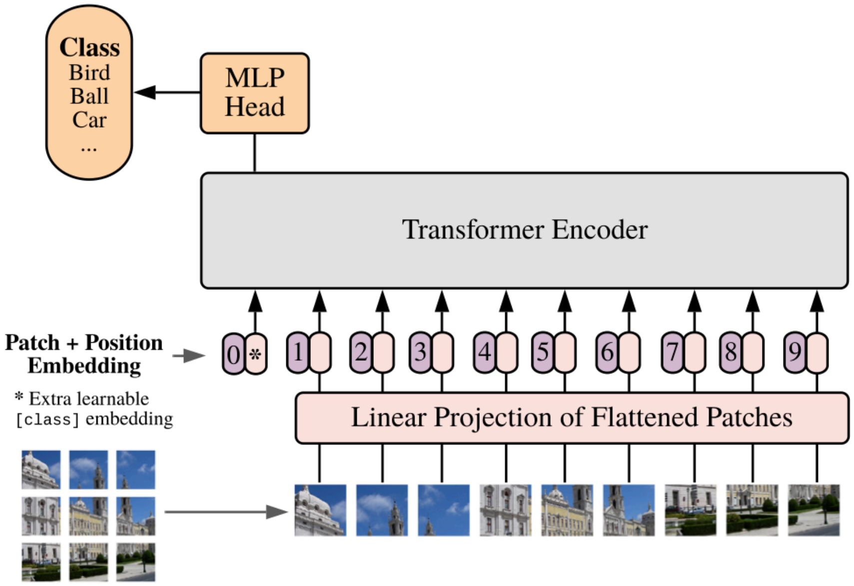
Lu et al. (2019). ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. NeurIPS.

Other multimodal encoders: LXMERT (Tan & Bansal, 2019) , UNITER (Chen et al. 2019), etc.

Vision Transformer (ViT)

- Split image into patches
  - This transforms the image into “tokens” like text, and makes the process more efficient
- Embed each patch (flattening)
- Add positional embeddings
- Encode using Transformer blocks
- Possibly pretrain on image classification

Better spatial and contextual information than CNNs.



Dosovitskiy et al. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. ICLR.

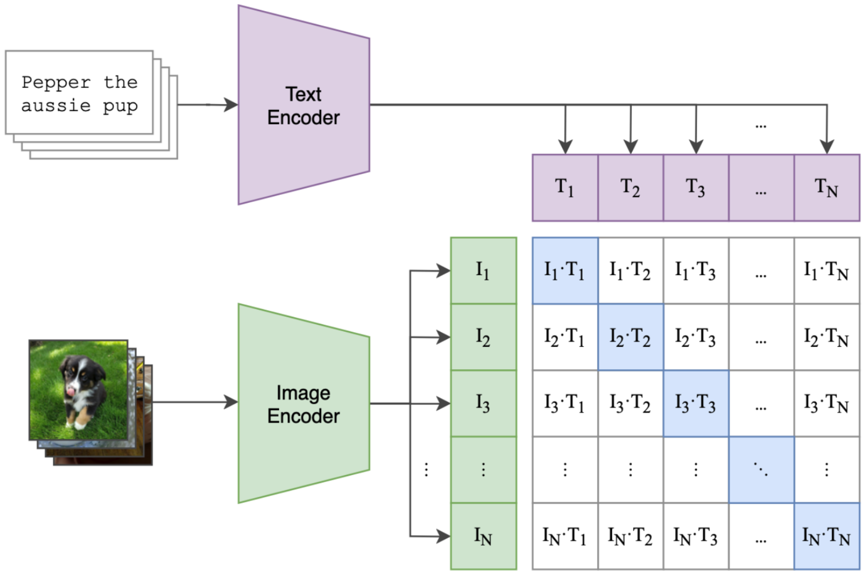
CLIP: contrastive language-Image pretraining (OpenAI)

The backbone consists of two separate components:

- visual encoder: VIT or ResNet (CNN)
- language encoder: GPT

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

- Huge pretraining dataset of unclear provenance.

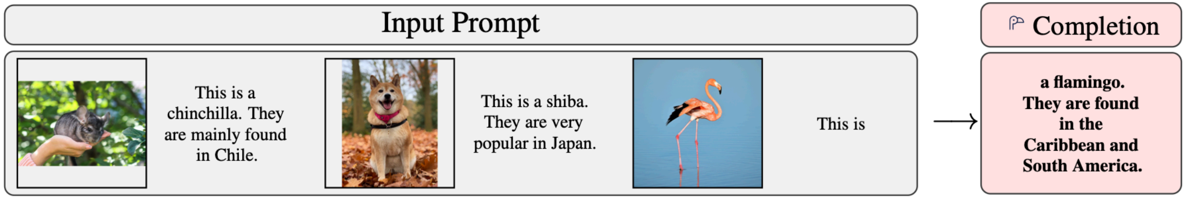


Radford et al. (2021). Learning transferable visual models from natural language supervision. ICML.

Generative VLMs with pre-trained backbones

- The current state of the art is dominated by generative VLMs that exploit pre-trained language-only and vision-only models.

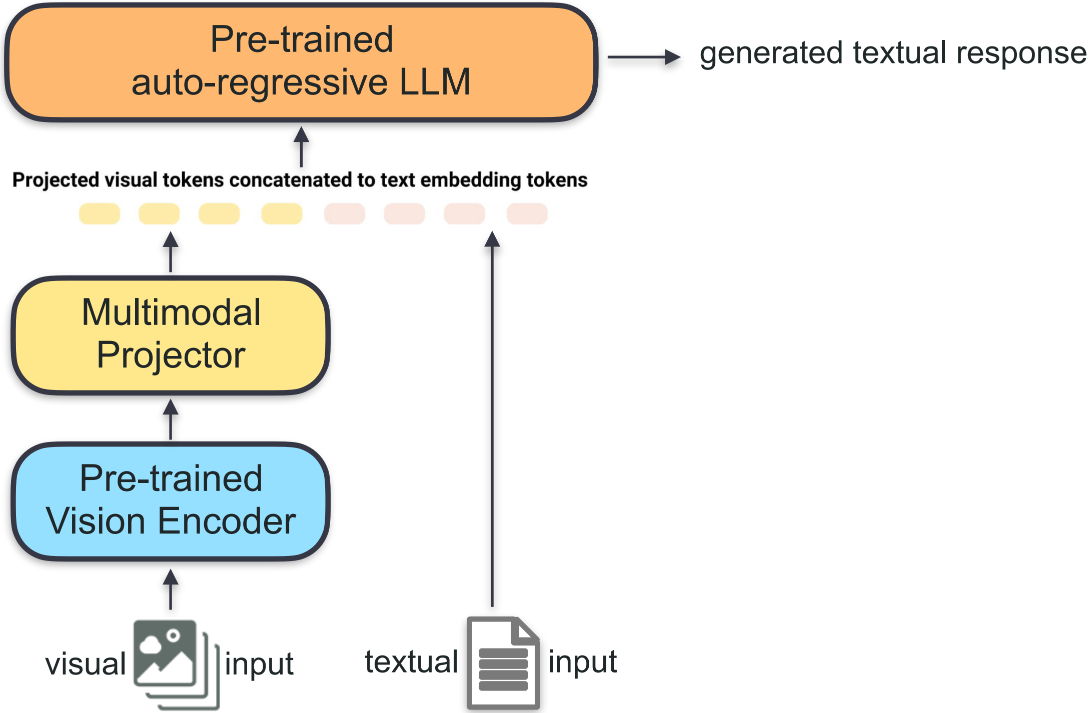
- For example, *Flamingo* (by DeepMind) was one of the first VLMs:



Alayrac et al., Flamingo: Visual Language Model for Few-shot Learning, NeurIPs 2022.
(DeepMind, unknown trining data)

Generative VLMs with pre-trained backbones

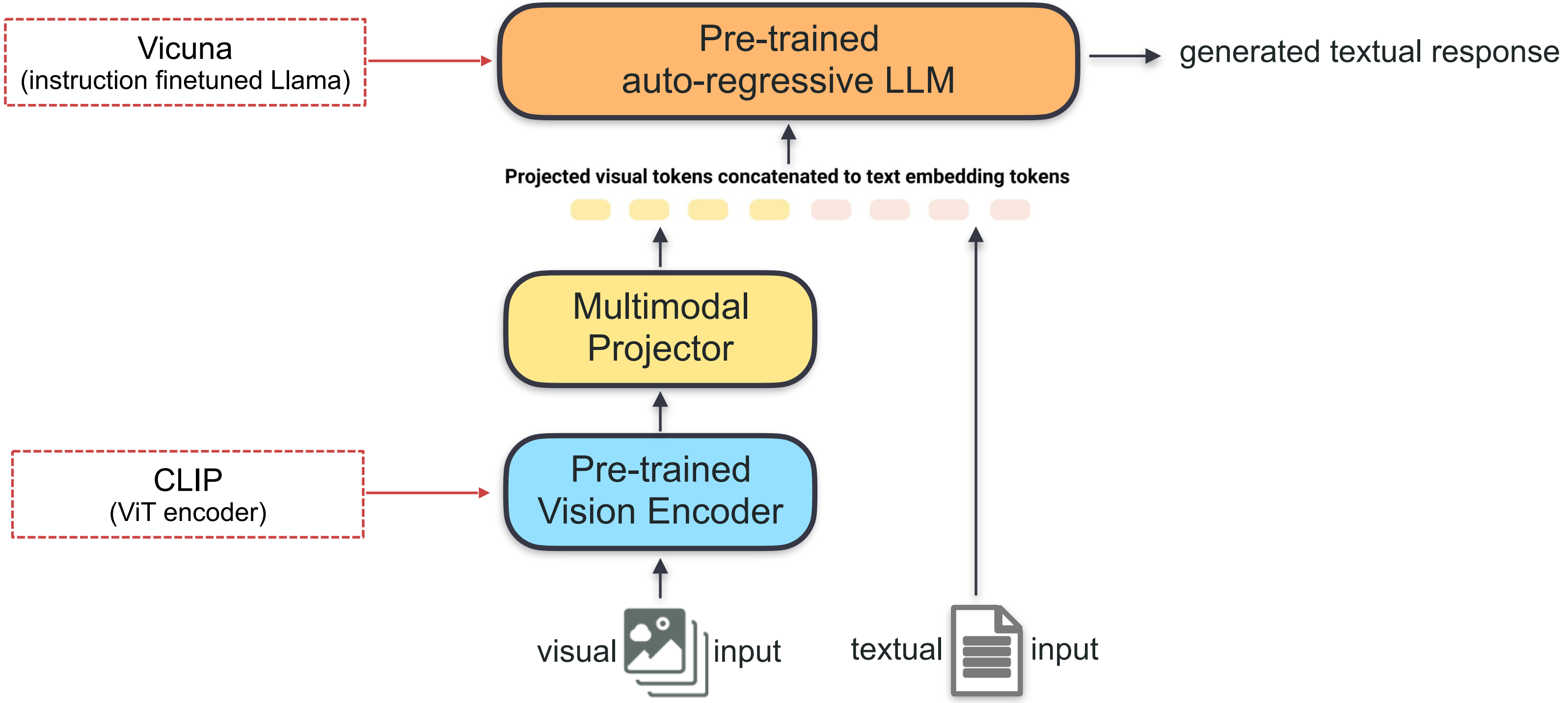
Common architecture



Generative VLMs with pre-trained backbones

LLaVA: Large Vision & Language Assistant

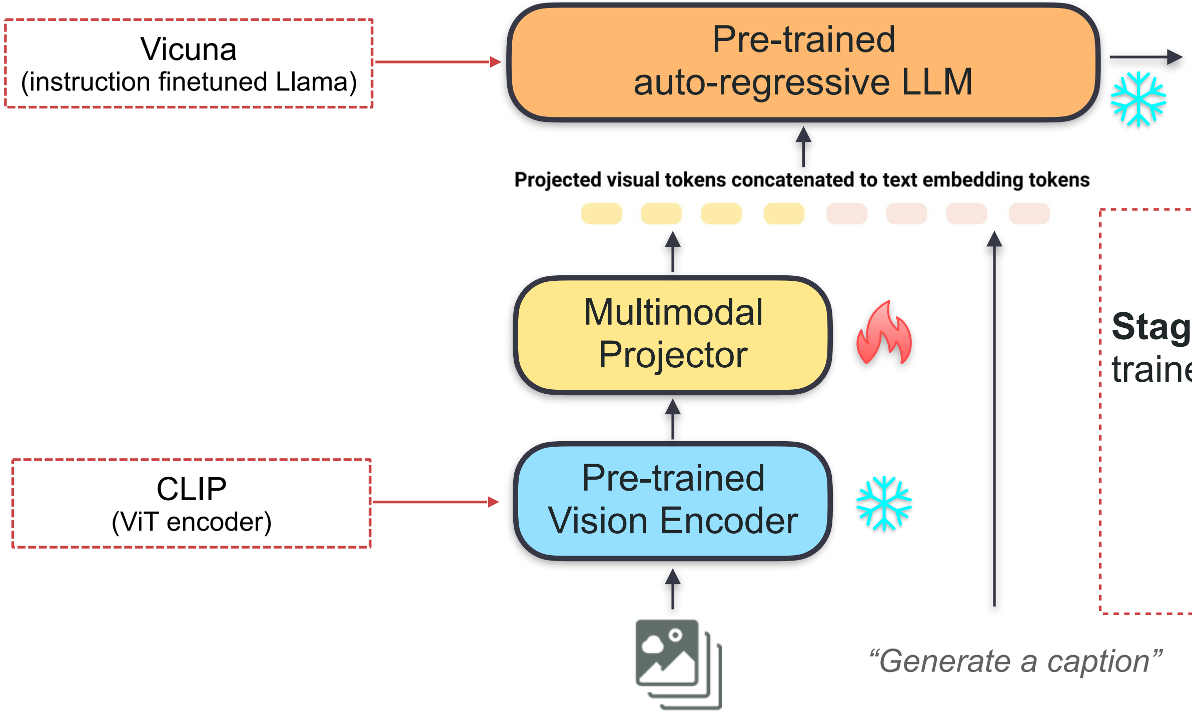
Liu et al. Visual Instruction Tuning, NeurIPS 2023.



Generative VLMs with pre-trained backbones

LLaVA: Large Vision & Language Assistant

Liu et al. Visual Instruction Tuning, NeurIPS 2023.



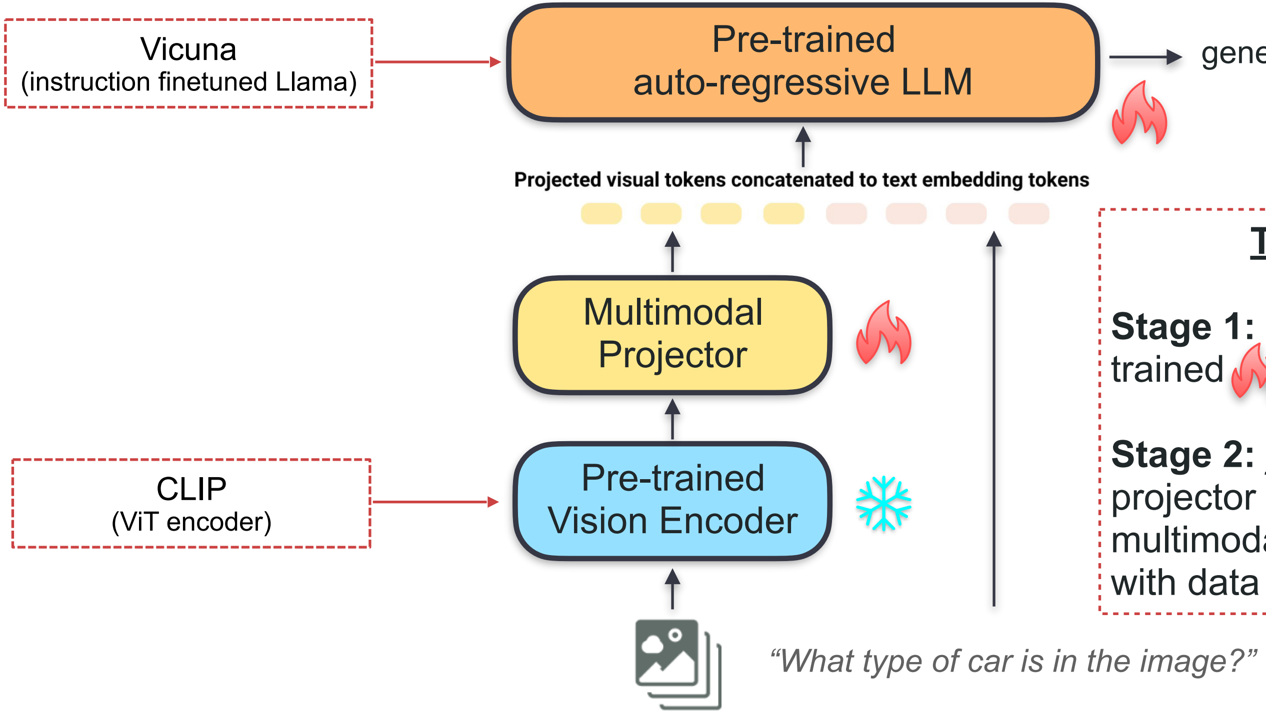
Training strategy

Stage 1: only the projector is trained, on image caption data.

Generative VLMs with pre-trained backbones

LLaVA: Large Vision & Language Assistant

Liu et al. Visual Instruction Tuning, NeurIPS 2023.



Training strategy

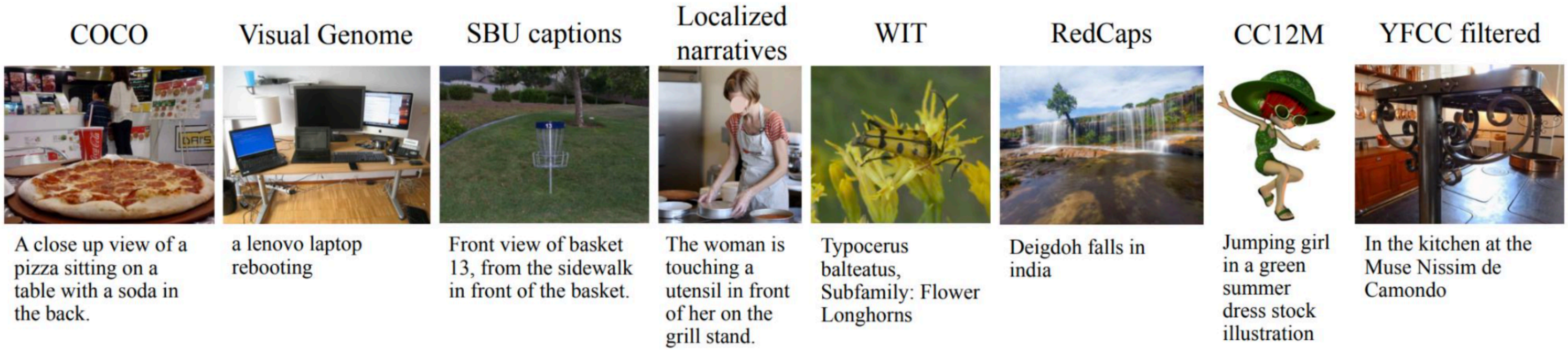
Stage 1: only the projector is trained 🔥, on image caption data.

Stage 2: joint training 🔥 of the projector and text decoded on multimodal instruction following, with data generated by GPT-4.

Large scale general-purpose datasets

Trend towards HUGE general purpose datasets used for model pretraining: Data scraped from the Internet – images aligned with alt-text. For example:

- Conceptual Captions: 3/12M images with (filtered) alt-text
- Public Multimodal Dataset: 70M pairs from existing datasets and other sources



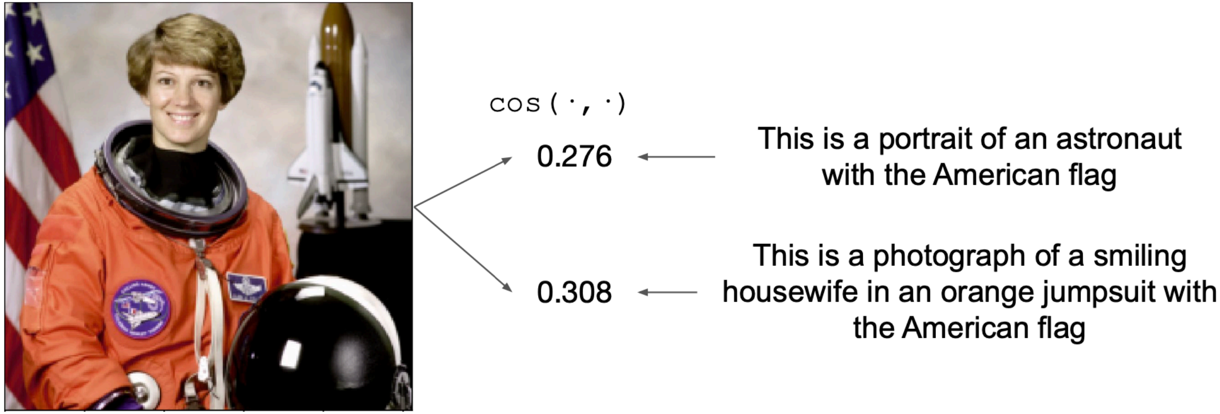
- LAION-5B / LAION-400B: harvested from CommonCrawl, a dump of the Internet with more than 300TB of stuff.

Ethical issues regarding large-scale training data

- Data provenance
  - Very often, data is scraped from the internet with limited information on license or copyright
  - The data may be public, but using such data without legal authorisation may infringe regulations
- Data quality
  - Does the dataset construction process lead to perpetuating harmful biases?
- Data diversity
  - Is the data representative of the population it intents to depict or serve?

Data quality

- Large models require huge datasets for training (e.g., LAION)
- Data scaling makes proper data curation extremely difficult
- As a result, models are often trained on data of very dubious quality with serious ethical implications.



(Eileen Collins, American astronaut)

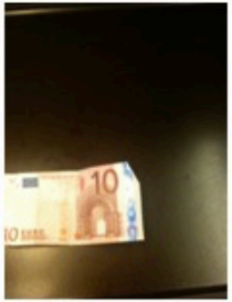
Birhane, Prabhu, Kahembwe (2021) Multimodal datasets: misogyny, pornography and malignant stereotypes.
Birhane et al. (2024) The Dark Side of Datascaling: Evaluating racial classification in multimodal models

Data diversity: are the intended users taken into account?

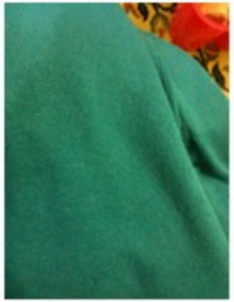
Early example of a curated dataset: VizWiz-VQA: questions asked by people who are blind

- Blind people taking photos and recording a spoken question about them
- 10 crowdsourced answers per visual question
- Tasks:
    - Predict the answer to a question
    - Predict whether a question cannot be answered

Q: Does this foundation have any sunscreen?
A: yes

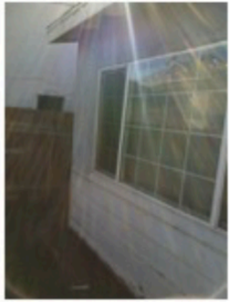


Q: What is this?
A: 10 euros

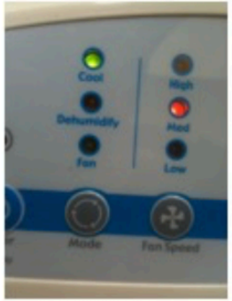


Q: What color is this?
A: green

Q: Please can you tell me what this item is?
A: butternut squash red pepper soup



Q: Is it sunny outside?
A: yes

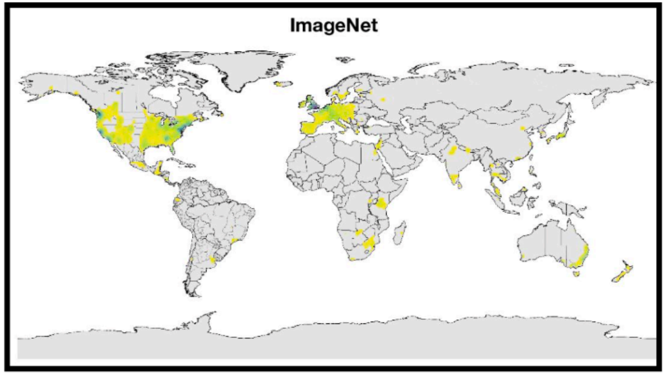


Q: Is this air conditioner on fan, dehumidifier, or air conditioning?
A: air conditioning

https://vizwiz.org/tasks-and-datasets/vqa/

Data diversity: linguistic and cultural biases

- Datasets are mostly in English, or a few Indo-European languages
- Some datasets are translated from English
- The image sources mostly reflect North American and European cultures
- Some concepts are most immediately understood within a cultural background



Density map of geographical distribution of images in ImageNet (DeVries+, 2019)

ENG: An unusual looking vehicle …

NLD: Een mobiel draaiorgel …

Example from van Miltenburg+ 2017

MaRVL: Multicultural Reasoning over Vision and Language

Representative of annotators' cultures

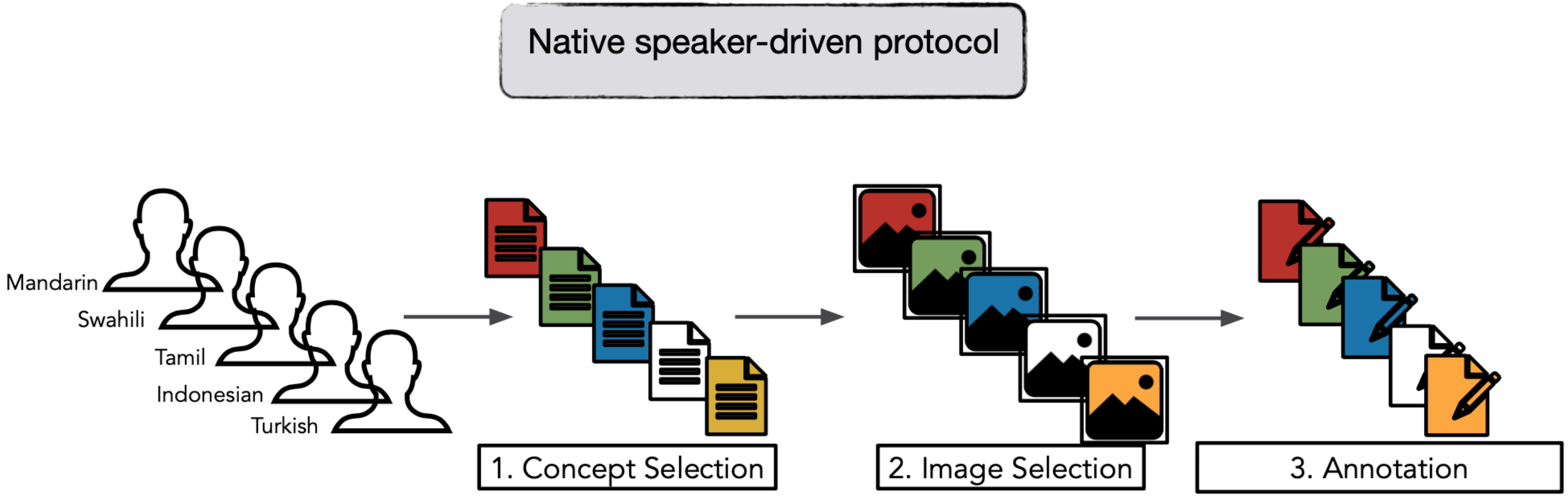
5 typologically diverse languages

Independent, culture-specific annotations

MaRVL-id Bola basket   MaRVL-sw Mpira wa kikapu   MaRVL-tr Basketbol   MaRVL-zh 篮球   MaRVL-ta கூடைப்பந்தாட்டம்

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

MaRVL dataset construction



MaRVL dataset construction

Visual reasoning task (Suhr et al., 2019)

- two images paired with a sentence
- Predict whether the sentence is a true description of the pair

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 மை (Vada)

இரண்டு படங்களிலும் நிறைய மசால் வடைகள் உள்

(Both images contain a lot of masala vadas)

Label: False

Other culturally diverse multimodal datasets

- **XM3600** (Thapliyal et al., 2022): 36 languages, 13 different scripts, 100 images per language selected to capture each language's cultural background.
- **M5B-VGR** (Schneider and Sitaram, 2024): 12 languages, 7 scripts, 120 culture-specific photos per language.
- **CVQA** (Romero et al., 2024):, 31 languages, 10k questions about cultural-specific images.
- Etc…

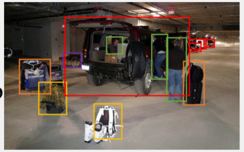
These datasets tend to be used for **evaluation** of state-of-the-art multimodal VLMs; instruction fine-tuning requires larger datasets.

Current trend: synthetic training data

Example of instruction data for finetuning, generated by text-only GPT models

Context type 1: Captions

A group of people standing outside of a black vehicle with various luggage.
Luggage surrounds a vehicle in an underground parking area
People try to fit all of their luggage in an SUV.
The sport utility vehicle is parked in the public garage, being packed for a trip
Some people with luggage near a van that is transporting it.



Context type 2: Boxes

person: [0.681, 0.242, 0.774, 0.694], backpack: [0.384, 0.696, 0.485, 0.914], suitcase: ...<omitted>

Response type 1: conversation

Question: What type of vehicle is featured in the image?
Answer: The image features a black sport utility vehicle (SUV) ...<omitted>

Response type 2: detailed description

The image is an underground parking area with a black sport utility vehicle (SUV) parked. There are three people in the scene, with one person standing closer to the left side of the vehicle, another person in the middle, and the third person on the right side. They are all working together to pack their luggage into the SUV for a trip. ...<omitted>

Response type 3: complex reasoning

Question: What challenges do these people face?
Answer: In the image, a group of people is standing outside a black SUV in a parking area, surrounded by various pieces of luggage, including suitcases and backpacks. They are facing the challenge of fitting all their luggage into the black SUV. There are multiple suitcases and backpacks to be packed, which suggests that the group has a significant amount of belongings ...<omitted>

Table 1: One example to illustrate the instruction-following data. The top block shows the contexts such as captions and boxes used to prompt GPT, and the bottom block shows the three types of responses. Note that the visual image is not used to prompt GPT, we only show it here as a reference.

Prompt used to generate the instruction following data

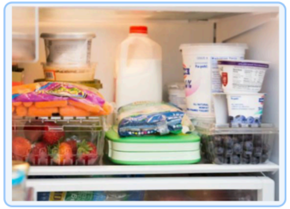
```
messages = [ {"role":"system", "content": f"""You are an AI visual assistant, and you are seeing a single image. What you see are provided with five sentences, describing the same image you are looking at. Answer all questions as you are seeing the image.

Design a conversation between you and a person asking about this photo. The answers should be in a tone that a visual AI assistant is seeing the image and answering the question. Ask diverse questions and give corresponding answers.

Include questions asking about the visual content of the image, including the object types, counting the objects, object actions, object locations, relative positions between objects, etc. Only include questions that have definite answers:
(1) one can see the content in the image that the question asks about and can answer confidently;
(2) one can determine confidently from the image that it is not in the image. Do not ask any question that cannot be answered confidently.

Also include complex questions that are relevant to the content in the image, for example, asking about background knowledge of the objects in the image, asking to discuss about events happening in the image, etc. Again, do not ask about uncertain details. Provide detailed answers when answering complex questions. For example, give detailed examples or reasoning steps to make the content more convincing and well-organized. You can include multiple paragraphs if necessary."""}
]
for sample in fewshot_samples:
    messages.append({"role":"user", "content":sample['context']})
    messages.append({"role":"assistant", "content":sample['response']} )
messages.append({"role":"user", "content":'\n'.join(query)})
```

Liu et al. Visual Instruction Tuning, NeurIPS 2023.



User

What are the meals that I can cook with these?

LLaVA

With the variety of food items stored in the refrigerator, you can prepare several meals. For example, you can create a fruit salad using the strawberries, blueberries, and carrots. Additionally, you can make a delicious smoothie using the fruits and vegetables, which would be a healthy and nutritious snack or breakfast option. The milk and eggs can be used to create various dishes, such as a milk-based smoothie, a baked casserole, or scrambled eggs. The juice can also be used to make a fresh fruit and juice mix or as a refreshing homemade popsicle. The possibilities are vast, and the meals you can create will depend on your preferences and dietary needs.

I would like to do the fruit salad. Show me the recipe.

User

LLaVA

Certainly! Here's a simple fruit salad recipe using the items in your refrigerator:

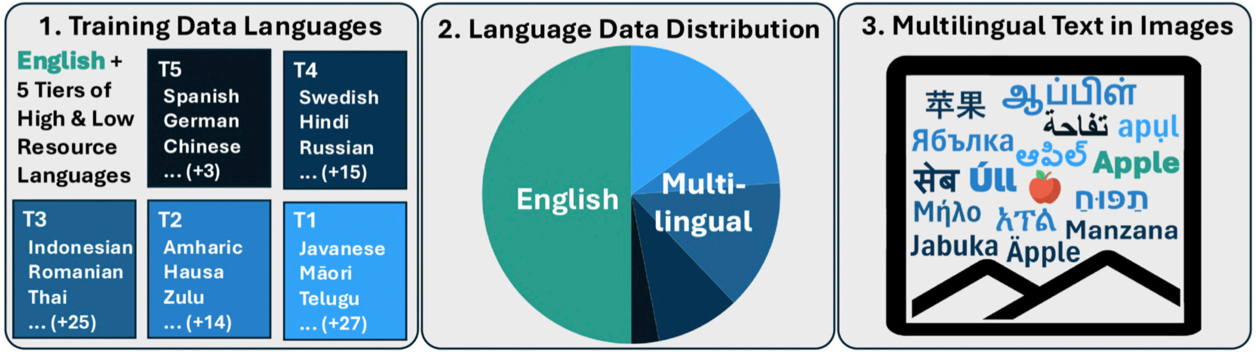
Ingredients:
- 2 cups fresh strawberries, sliced
- 1 cup fresh blueberries
- 1 large carrot, grated
- Juice of half a lemon
- A handful of chopped fresh parsley or mint
- A few ice cubes
- Optional: honey or maple syrup (adjust to taste)

Instructions:
1. In a large bowl, combine the sliced strawberries, blueberries, and carrot. Gently toss them together to mix.
2. Drizzle the juice of half a lemon over the fruit and vegetables to prevent them from oxidizing and add a refreshing citrus flavor.
3. Sprinkle the chopped fresh parsley or mint over the fruit and gently toss once more.
4. If desired, add a drizzle of honey or maple syrup for a touch of sweetness. Adjust the amount according to your taste preferences.
5. Cover the bowl and refrigerate the fruit salad for at least 30 minutes or up to a few hours. This allows the flavors to meld and creates a more refreshing salad.
6. Serve the fruit salad in cups or bowls, garnished with a sprinkle of fresh parsley or mint, if desired.

This fruit salad recipe is perfect for a healthy snack or side dish, and with the variety of ingredients available in your refrigerator, it should be easy to put together quickly.

State-of-the-art multilingual VLM: Centurio

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



- LLaVA architecture, with SigLIP-SO400 as CLIP-like image encoder and Phi-3.5 as multilingual LLM backbone.
- The training data is mostly synthetic and machine-translated (with the NLLB model) across 100 languages.

Wiki-LLaVA: Hierarchical RAG for Multimodal LLMs

First approach to integrating an external knowledge source into multimodal generative LLMs

- Retrieval module with two steps:
    1. Retrieve documents via CLIP similarity of input image and document titles
    2. Retrieve relevant passages via embedding similarity of the input question and document chunks
- Enrich input context with the retrieved passages

```
<IMAGE>\nGiven the following context:\n
<R1>\n<R2>\<R3>\n <QUESTION>
Give a short answer.  ASSISTANT:
```

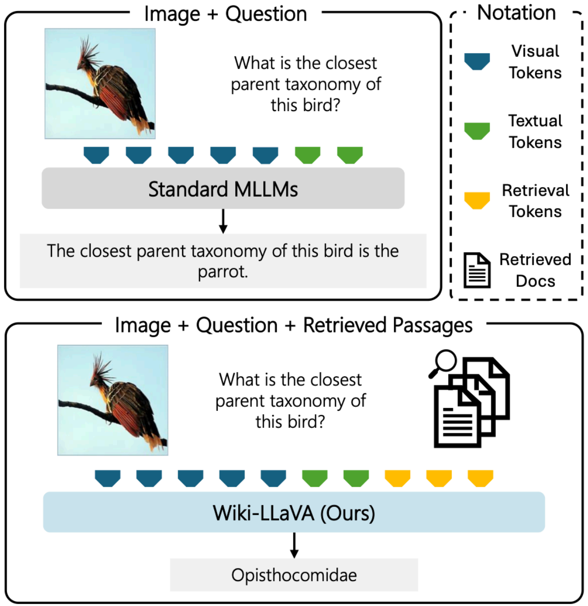


Figure 1. Comparison between a standard multimodal LLM and Wiki-LLaVa. Our model integrates knowledge retrieved from an external knowledge base of documents through a hierarchical retrieval pipeline. As a result, it provides more precise answers when tasked with questions that require external knowledge.

Caffagni et al. Wiki-LLaVA: Hierarchical Retrieval Augmented Generation for Multimodal LLMS, ArXiv, 2024.

Wiki-LLaVA: Hierarchical RAG for Multimodal LLMs

In what state is this building located?

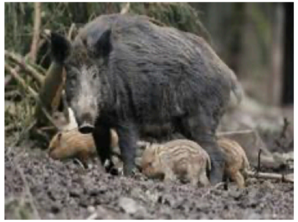
LLaVA-1.5:
California ✗
Wiki-LLaVA:
Washington ✓

When was this building constructed?

LLaVA-1.5:
1970 ✗
Wiki-LLaVA:
1927 ✓

Which geographic area is this fish found?

LLaVA-1.5:
Gulf of Mexico ✗
Wiki-LLaVA:
Brazil ✓



What is the oldest age of this animal?

LLaVA-1.5:
10 years ✗
Wiki-LLaVA:
24.9 ✓

Which culture is associated with this place?
`Ancient Greek`

LLaVA-1.5:
Roman ✗
Wiki-LLaVA:
Nuragic Civilization ✗

What is the name of the main club of this stadium?
`FC Rotor`

LLaVA-1.5:
Real Madrid ✗
Wiki-LLaVA:
FC Dynamo Kyiv ✗

Evaluation of VLMs

Different types of evaluation

- Task-specific evaluation: it does not make much sense for general-purpose pre-trained VLMs.

Evaluation: Visual Question Answering

VQA has traditionally been operationalized as a classification task, evaluated with accuracy.

Does the visual information matter to perform the task?

- VQA dataset: around 600k image-question pairs
    - Imbalances: e.g., 41% of questions starting with “What sport is…” have “tennis” as the correct answer
- VQA.v2 dataset: 1.1M image–question pairs with balanced distribution of answers

Antol et al. (2015). VQA: Visual Question Answering. ICCV.
Goyal et al. (2017). Making the V in VQA matter: Elevating the role of image understanding in visual question answering. CVPR.

Evaluation: Image captioning

NLG metrics that rely on comparing **generated text** to a **reference text** are often used for tasks such as image captioning or visual storytelling

- BLEU, ROUGE, CIDEr, METEOR, etc.

It is well known that these metrics are problematic:

- **The same message can be conveyed in very different ways!**
- **These metrics only consider the language modality, ignoring the visual input**

BERTscore and CLIPscore aim to address these issues, but they are limited.

Evaluation: Visual Storytelling

(EMNLP 2023)

GROOViST: A Metric for Grounding Objects in Visual Storytelling

Aditya K Surikuchi
University of Amsterdam
a.k.surikuchi@uva.nl

Sandro Pezzelle, Raquel Fernández
ILLC, University of Amsterdam
{s.pezzelle, raquel.fernandez}@uva.nl

Abstract

A proper evaluation of stories generated for a sequence of images—the task commonly referred to as visual storytelling—must consider multiple aspects, such as coherence, grammatical correctness, and visual grounding. In this work, we focus on evaluating the degree of grounding, that is, the extent to which a story is about the entities shown in the images. We analyze current metrics, both designed for this purpose and for general vision-text alignment. Given their observed shortcomings, we propose a novel evaluation tool, GROOViST, that accounts for cross-modal dependencies, *temporal*

Figure 1: One story and corresponding image sequence from the VIST dataset. Noun phrases in green contribute positively to the grounding score by GROOViST; those in red contribute negatively. The GROOViST score for this sample is 0.855, i.e., our metric considers it as well-grounded (within range: $[-1, 1]$). Best viewed in color.

(EMNLP Findings 2024)

Not (yet) the whole story: Evaluating Visual Storytelling Requires More than Measuring Coherence, Grounding, and Repetition

Aditya K Surikuchi, Raquel Fernández, Sandro Pezzelle
Institute for Logic, Language and Computation
University of Amsterdam
{a.k.surikuchi|raquel.fernandez|s.pezzelle}@uva.nl

Abstract

Visual storytelling consists in generating a natural language story given a temporally ordered sequence of images. This task is not only challenging for models, but also very difficult to evaluate with automatic metrics since there is no consensus about what makes a story 'good'. In this paper, we introduce a novel method that measures story quality in terms of human likeness regarding three key aspects highlighted in previous work: visual grounding, coherence, and repetitiveness. We then use this method to evaluate the stories generated by several models, showing that the foundation model LLaVA

coherence, or how repetitive they are. This problem has only been addressed recently, with Wang et al. (2022) and Surikuchi et al. (2023) proposing various metrics to take into account some of these crucial aspects. These methods assess the appropriateness of a generated story independently from its overlap with a ground-truth story for the same image sequence. Given that the same image sequence can possibly give rise to many different stories, this type of higher-level evaluation that does not rely on text overlap is clearly desirable.

Nevertheless, we argue that measuring the degree of coherence or visual grounding of a story

Evaluation: Visual Storytelling

Input: sequence of images

Task: to generate a textual story consistent with the input

Human-annotated story: *It's parade day, and the whole town turns out to watch. There are those who serve our country, and the crowds cheer. There are the bands, and the music is loud but thankfully well performed. The flags are always fun to watch. And of course you get the old cars and their owners traveling through.*

Challenge: plausibility of several creative stories for a single given image sequence, makes reference-based NLG metrics (e.g., METEOR) inappropriate for the task.

Surikuchi et al. (2024). Not yet the whole story... EMNLP Findings.

Evaluation: Visual Storytelling

Visual storytelling requires more evaluation dimensions (Wang et al. 2022; Surikuchi et al. 2023, 2024)

- **Coherence**: LM probability of the next sentence given the context
- **Degree of repetition**: Jaccard similarity between context and next sentence
- **Visual grounding**: CLIP-based cosine similarity between noun phrases and object bounding boxes, weighted by noun concreteness.

Distance between humans and models

$$\mathbf{d}_{HM}^{C} = |C_H - C_M|,$$
$$\mathbf{d}_{HM}^{G} = |G_H - G_M|,$$
$$\mathbf{d}_{HM}^{R} = |R_H - R_M|$$

$$\mathbf{d}_{HM} = (\mathbf{d}_{HM}^{C} + \mathbf{d}_{HM}^{G} + \mathbf{d}_{HM}^{R})/3$$

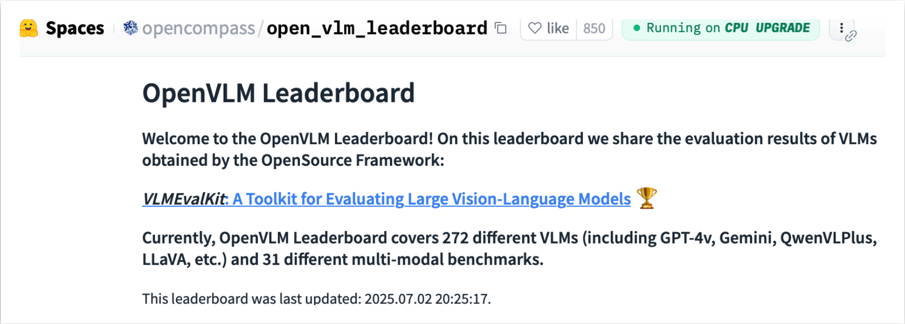
Surikuchi et al. (2024). Not yet the whole story... EMNLP Findings.

Different types of evaluation

- Task-specific evaluation: it does not make much sense for general-purpose pre-trained VLMs.
- Generic multi-task benchmarks.

Generic multi-task evaluation benchmarks

- **MMBench** (Liu et al., ECCV 2023): 3000 single-choice questions over 20 different skills, including OCR, object localization and more.
- **MMMU** (A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI; Due et al., CVPR 2024): 11.5K multimodal challenges that require college-level subject knowledge and reasoning across different disciplines such as arts and engineering.
- **MMT-Bench** (A Comprehensive Multimodal Benchmark for Evaluating Large Vision-Language Models Towards Multitask AGI; Ying et al., ICML 2024): 31,325 multiple-choice visual questions from various multimodal scenarios such as vehicle driving and embodied navigation, covering 32 core meta-tasks and 162 subtasks in multimodal understanding.



https://huggingface.co/spaces/opencompass/open_vlm_leaderboard

MMMU

A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI

Art & Design

Question: Among the following harmonic intervals, which one is constructed incorrectly?

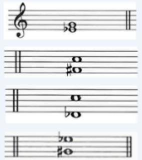
Options:

(A) Major third <*image 1*>

(B) Diminished fifth <*image 2*>

<u>**(C) Minor seventh <*image 3*>**</u>

(D) Diminished sixth <*image 4*>



Subject: Music; **Subfield:** Music;
Image Type: Sheet Music;
Difficulty: Medium

Business

Question: ...The graph shown is compiled from data collected by Gallup <*image 1*>. Find the probability that the selected Emotional Health Index Score is between 80.5 and 82?

Options:

(A) 0 (B) 0.2142

<u>(C) 0.3571</u> (D) 0.5

Subject: Marketing; **Subfield:** Market Research; **Image Type:** Plots and Charts;
Difficulty: Medium

Science

Question: <*image 1*> The region bounded by the graph as shown above. Choose an integral expression that can be used to find the area of R.

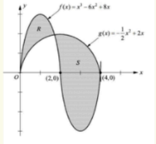
Options:

<u>**(A)**</u> $\int_0^{1.5}[f(x)-g(x)]dx$

(B) $\int_0^{1.5}[g(x)-f(x)]dx$

(C) $\int_0^{2}[f(x)-g(x)]dx$

(D) $\int_0^{2}[g(x)-x(x)]dx$



Subject: Math; **Subfield:** Calculus;
Image Type: Mathematical Notations;
Difficulty: Easy

Health & Medicine

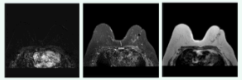
Question: You are shown subtraction <*image 1*>, T2 weighted <*image 2*> and T1 weighted axial <*image 3*> from a screening breast MRI. What is the etiology of the finding in the left breast?

Options:

(A) Susceptibility artifact

(B) Hematoma

<u>**(C) Fat necrosis**</u> (D) Silicone granuloma



Subject: Clinical Medicine; **Subfield:** Clinical Radiology; **Image Type:** Body Scans: MRI, CT.;
Difficulty: Hard

Humanities & Social Science

Question: In the political cartoon, the United States is seen as fulfilling which of the following roles? <*image 1*>

Option:

(A) Oppressor

(B) Imperialist

<u>**(C) Savior**</u> (D) Isolationist

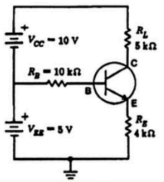
Subject: History; **Subfield:** Modern History; **Image Type:** Comics and Cartoons;
Difficulty: Easy

Tech & Engineering

Question: Find the VCE for the circuit shown in <*image 1*>. Neglect VBE

Answer: <u>3.75</u>

Explanation: ...IE = [(VEE) / (RE)] = [(5 V) / (4 k-ohm)] = 1.25 mA; VCE = VCC - IERL = 10 V - (1.25 mA) 5 k-ohm; VCE = 10 V - 6.25 V = 3.75 V



Subject: Electronics; **Subfield:** Analog electronics; **Image Type:** Diagrams;
Difficulty: Hard

https://mmmu-benchmark.github.io/

Different types of evaluation

- Task-specific evaluation: it does not make much sense for general-purpose pre-trained VLMs.
- Generic multi-task benchmarks.
- Evaluating the quality of the representations learned by the models (e.g., in terms of human likeness).

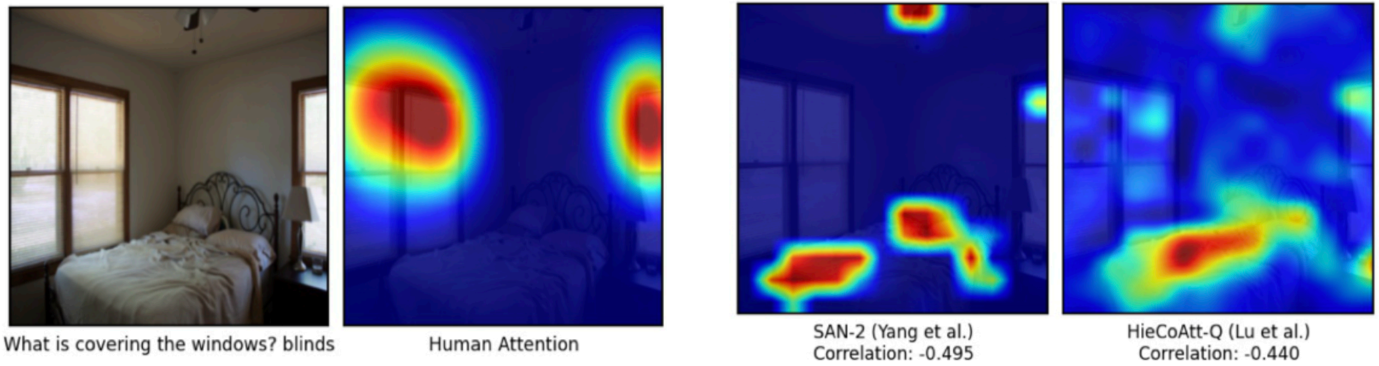
Representational quality

- By grounding language into vision, arguably multimodal models have a representational advantage over text-only models.
- Do they learn representations that better align with human multimodal knowledge and processing?

Early example of representational quality evaluation: Comparing attention patterns in VQA

Where humans look

Where machines look



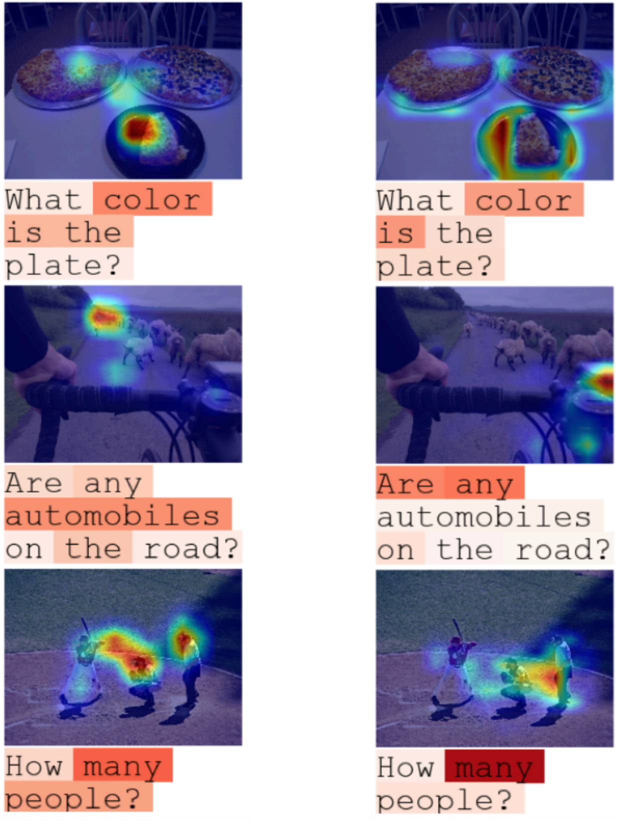
- Low correlation between human and machine attention: 0.256
- Inter-human correlation: 0.623

Das et al. (2016).

More recent follow-up work:
Comparing attention patterns in VQA

- Higher correlation with visual and text attention is a significant predictor of VQA performance

(Sood et al., 2021)



Representational quality:
correlation with semantic similarity judgements

(TACL 2021)

Word Representation Learning in Multimodal Pre-Trained Transformers: An Intrinsic Evaluation

Sandro Pezzelle, Ece Takmaz, Raquel Fernández
Institute for Logic, Language and Computation
University of Amsterdam, The Netherlands
{s.pezzelle|e.takmaz|raquel.fernandez}@uva.nl

Abstract

This study carries out a systematic *intrinsic* evaluation of the semantic representations learned by state-of-the-art pre-trained multimodal Transformers. These representations are claimed to be task-agnostic and shown to help on many downstream language-and-vision tasks. However, the extent to which they align with human semantic intuitions remains unclear. We experiment with various models and obtain *static* word representations from the *contextualized* ones they learn. We then evaluate them against the semantic judgments provided by human speakers. In line with previous evidence, we observe a generalized advantage

Language-only semantic representations, from pioneering 'count' vectors (Landauer and Dumais, 1997; Turney and Pantel, 2010; Pennington et al., 2014) to either *static* (Mikolov et al., 2013) or *contextualized* (Peters et al., 2018; Devlin et al., 2019) neural network-based embeddings, have proven extremely effective in many linguistic tasks and applications, for which they constantly increased state-of-the-art performance. However, they naturally have no connection with the real-world referents they denote (Baroni, 2016). As such, they suffer from the symbol grounding problem (Harnad, 1990), which in turn limits their cognitive plausibility (Rotaru and

Representational quality:
correlation with semantic similarity judgements

Semantic similarity:

man, person: **similar**
dog, airplane: **dissimilar**

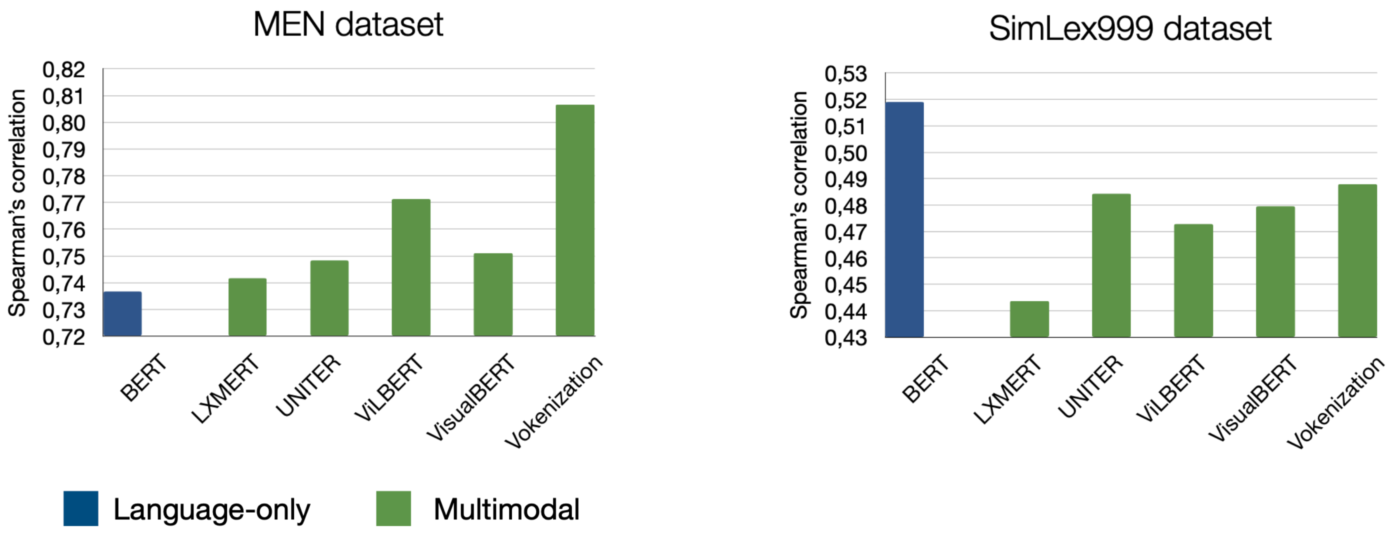
- *donut, muffin* = 0.8
- *car, train* = 0.5
- *dog, airplane* = 0.1
- …

Spearman's correlation between:

- **human semantic similarity judgements** and
- **cosine similarity** between pairs of **model word representations**

Comparison of the semantic spaces

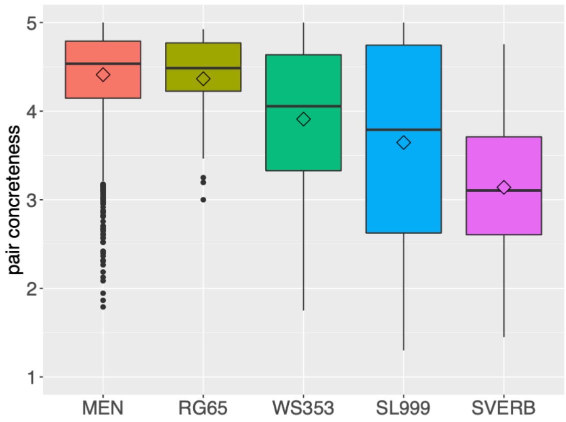
Representational quality:
correlation with semantic similarity judgements



Representational quality:
correlation with semantic similarity judgements

The level of concreteness of the words being judged varies per dataset

Very concrete:
doughnut, muffin



Very abstract:
freedom, dreams

Multimodal models are better than text-only ones at approximating similarity judgements of concrete words

Representational quality: correlation with brain responses

Do multimodal pre-trained models learn representations that are more aligned with how the brain represents conceptual knowledge?

(CoNLL 2025)

Experiential Semantic Information and Brain Alignment: Are Multimodal Models Better than Language Models?

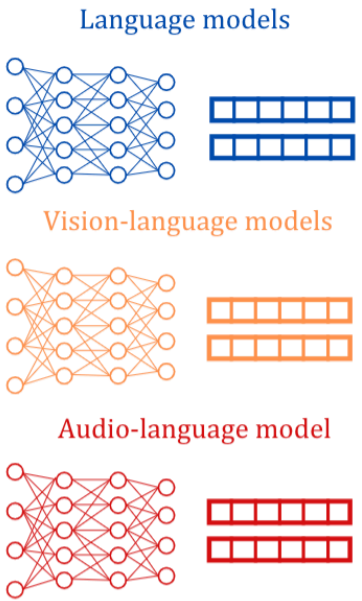
Anna Bavaresco, Raquel Fernández
Institute for Logic, Language and Computation
University of Amsterdam
{a.bavaresco, raquel.fernandez}@uva.nl

Abstract

A common assumption in Computational Linguistics is that text representations learnt by multimodal models are richer and more human-like than those by language-only models, as they are grounded in images or audio—similar to how human language is grounded in real-world experiences. However, empirical studies checking whether this is true are largely lacking. We address this gap by comparing word representations from contrastive multimodal models vs. language-only ones in the extent to

a rich multimodal environment, where new words are learnt through interactions with objects and people (Vigliocco et al., 2014). Theories of embodied cognition further highlight the importance of linking words to concrete experience not only for their acquisition but also for their comprehension. Indeed, according to these theories, understanding sentences involves engaging perceptual, motor or emotional simulations of their content (for an overview, see Kaschak et al., 2024).

The idea of obtaining richer semantic representa-



Multimodal processing

Mean onset: 3.46 seconds
Variation in starting points: 11
Most common starting point: *pier*
Image specificity BLEU-2: 0.39
Variation in gaze: 38.47

VLMs lack biases about what makes an image complex for humans and what leads to variation in processing behaviour when describing images.

(EACL 2024)

Describing Images *Fast and Slow*: Quantifying and Predicting the Variation in Human Signals during Visuo-Linguistic Processes

Ece Takmaz and **Sandro Pezzelle** and **Raquel Fernández**
Institute for Logic, Language and Computation
University of Amsterdam
{ece.takmaz|s.pezzelle|raquel.fernandez}@uva.nl

Abstract

There is an intricate relation between the properties of an image and how humans behave while describing the image. This behavior shows ample variation, as manifested in human signals such as eye movements and when humans start to describe the image. Despite the value of such signals of visuo-linguistic variation, they are virtually disregarded in the training of current pretrained models, which motivates further investigation. Using a corpus of Dutch image descriptions with concurrently

Min: 1.69 sec

Max: 7.07 sec

Figure 1: The images with the minimum and maximum mean speech onsets across speakers in the dataset. The image with the maximum onset also elicits the highest variation in the first nouns of the descriptions.

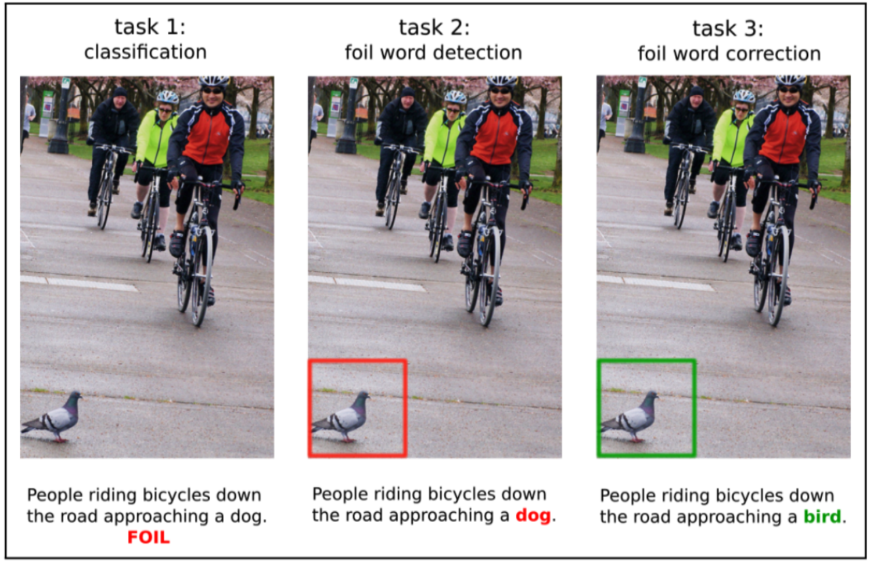
Different types of evaluation

- Task-specific evaluation: it does not make much sense for general-purpose pre-trained VLMs.
- Generic multi-task benchmarks.
- Evaluating the quality of the representations learned by the models (e.g., in terms of human likeness).
- Assessing specific skills through challenging test sets: what skills have models acquired and where do they fail?

Challenge datasets to analyse specific skills

Early example: FOIL captions

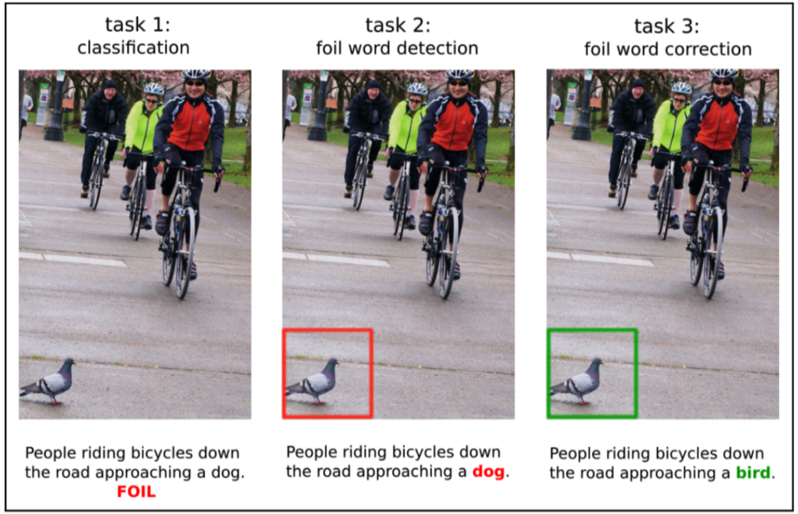
- Do V&L models really understand the relationship between words and images?
- Crowdsource datasets that contain contextually plausible but incorrect image–text pairs, focusing on nouns.



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

FOIL captions

T1: Classification task			
	Overall	Correct	Foil
Blind	55.62	86.20	25.04
CNN+LSTM	61.07	89.16	32.98
IC-Wang	42.21	38.98	45.44
LSTM + norm I	63.26	**92.02**	34.51
HieCoAtt	**64.14**	91.89	**36.38**
Human (*majority*)	92.89	91.24	94.52
Human (*unanimity*)	76.32	73.73	78.90



- Very challenging at the time, but has since been essentially solved.
- It's a good sanity check!

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

Subject-Verb-Object Probes

- SVO-Probes: subject-verb-object sentences, with focus on **verbs**
- Models largely fail to distinguish images with fine-grained **verb** differences
- Accuracy below chance on negative pairs
- Verb understanding is harder than subject or object understanding

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

Winoground

- 1,600 text-image pairs to evaluate compositional understanding
- Images sourced with permission from Getty
- Models struggle, often performing below chance

(a) some plants surrounding a lightbulb

(b) a lightbulb surrounding some plants

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

VALSE Benchmark

Evaluation of multiple linguistic phenomena

existence	plurality	counting	relations	actions	coreference
There are no animals / animals shown.	*A small copper vase with some flowers / exactly one flower in it.*	*There are four / six zebras.*	*A cat plays with a pocket knife on / underneath a table.*	*A man / woman shouts at a woman / man.*	*Buffalos walk along grass. Are they in a zoo? No / Yes.*
					

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	<u>46.4</u>

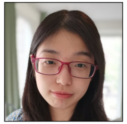
$$p(caption, img) > p(foil, img)$$

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

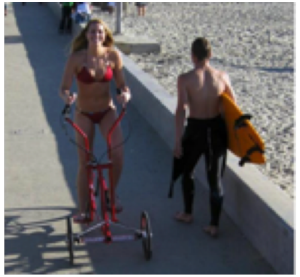
BLA: Basic Language Abilities

Focus on simple construction that preschool children can understand

Contrastive models like CLIP, trained to align images with textual descriptions, tend to learn “bag of words” representations.



Active-Passive voice
T: the woman feeds the man.
T: the man is fed by the woman.
F: the man feeds the woman.
F: the woman is fed by the man.



Coordination
T: the man wears a wetsuit and carries a surfboard.
T: the woman wears a red bikini and rides a red bike.
F: the man wears a wetsuit and rides a red bike.
F: the woman carries a surfboard and wears a red bikini.

Relative Clause
T: the man who wears a gray polo holds a stuffed bear.
T: the man who wears a striped shirt holds a cow.
F: the man who wears a gray polo holds a cow.
F: the man who wears a striped shirt holds a stuffed bear.

(EMNLP 2023)

The BLA Benchmark: Investigating Basic Language Abilities of Pre-Trained Multimodal Models

Xinyi Chen
IvI
University of Amsterdam
x.chen2@uva.nl

Raquel Fernández
ILLC
University of Amsterdam
raquel.fernandez@uva.nl

Sandro Pezzelle
ILLC
University of Amsterdam
s.pezzelle@uva.nl

(EMNLP 2025)

RACQUET: Unveiling the Dangers of Overlooked Referential Ambiguity in Visual LLMs

Alberto Testoni[1], Barbara Plank[2,3], Raquel Fernández[1],
[1] Institute for Logic, Language and Computation (ILLC), University of Amsterdam
[2]Center for Information and Language Processing, LMU Munich
[3]Munich Center for Machine Learning (MCML), Munich
Correspondence: a.testoni@uva.nl

Abstract

Ambiguity resolution is key to effective communication. While humans effortlessly address ambiguity through conversational grounding strategies, the extent to which current language models can emulate these strategies remains unclear. In this work, we examine *referential* ambiguity in image-based question answering by introducing RACQUET, a carefully curated dataset targeting distinct aspects of ambiguity. Through a series of evaluations, we reveal significant limitations and problems of overconfidence of state-of-the-art large multimodal language models in addressing ambiguity in their responses. The overconfidence issue becomes particularly relevant for RACQUET-BIAS, a subset designed to analyze a critical yet underexplored problem: failing to address ambiguity leads to stereotypical, socially biased

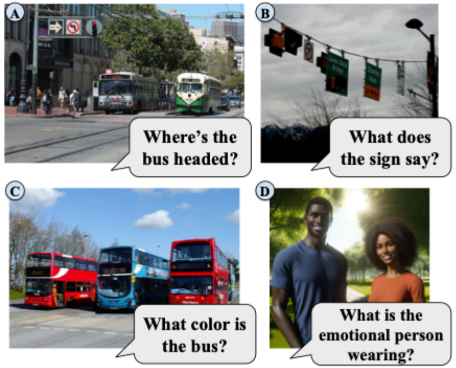


Figure 1: Examples of ambiguous question-image pairs from RACQUET-GENERAL (A,B,C) and RACQUET-BIAS (D).

Referentially ambiguous questions about images

RAcQUET dataset

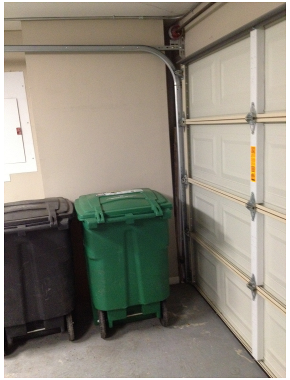
Setting 1 (GENERAL): Images from MS-COCO paired with handcrafted ambiguous questions

What color is the bus?

What does the sign say?

How to react to such questions?

What colour is the trash bin?



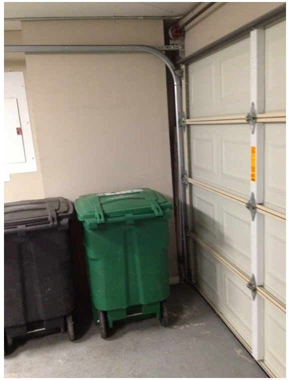
How to react to such questions?

(A: Explicit) Signal the ambiguity, to *build* common ground

- By conversational grounding strategies: asking clarification questions
- Mention there are several referents and provide answer for all of them

What colour is the trash bin?

(A) Which one?
(A) There are two, a black one and a green one.



How to react to such questions?

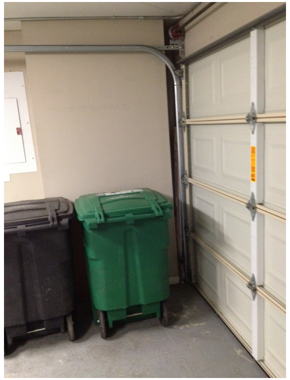
(A: Explicit) Signal the ambiguity, to *build* common ground

- By conversational grounding strategies: asking clarification questions
- Mention there are several referents and provide answer for all of them

(B: Implicit) Assume one intended referent, indicating which one - hence giving the chance to the interlocutor to correct (initiate repair in the next turn).

What colour is the trash bin?

(A) Which one?
(A) There are two, a black one and a green one.
(B) The bin on the left is black.



How to react to such questions?

(A: Explicit) Signal the ambiguity, to *build* common ground

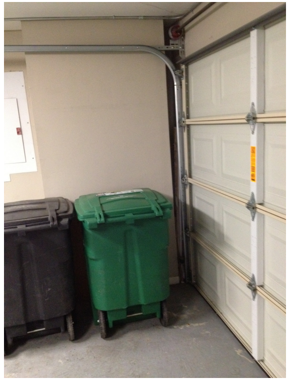
- By conversational grounding strategies: asking clarification questions
- Mention there are several referents and provide answer for all of them

(B: Implicit) Assume one intended referent, indicating which one - hence giving the chance to the interlocutor to correct (initiate repair in the next turn).

(C: High Risk) Assume one intended referent, without further ado (accommodating the presupposition of uniqueness)

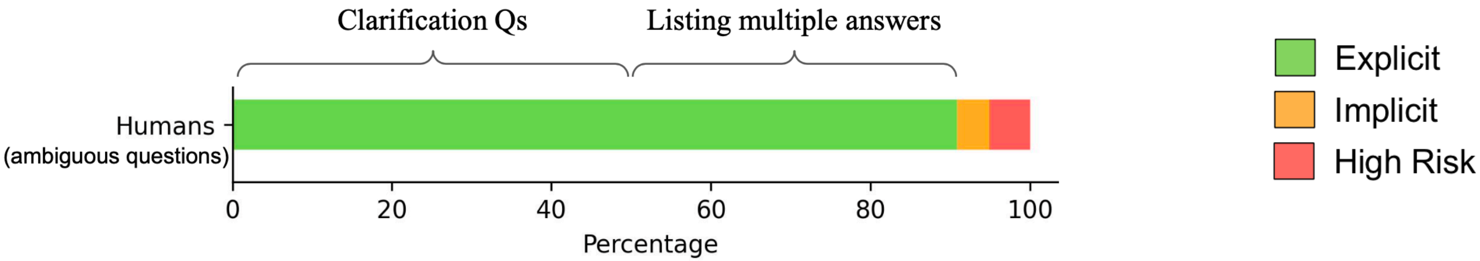
What colour is the trash bin?

(A) Which one?
(A) There are two, a black one and a green one.
(B) The bin on the left is black.
(C) Green.



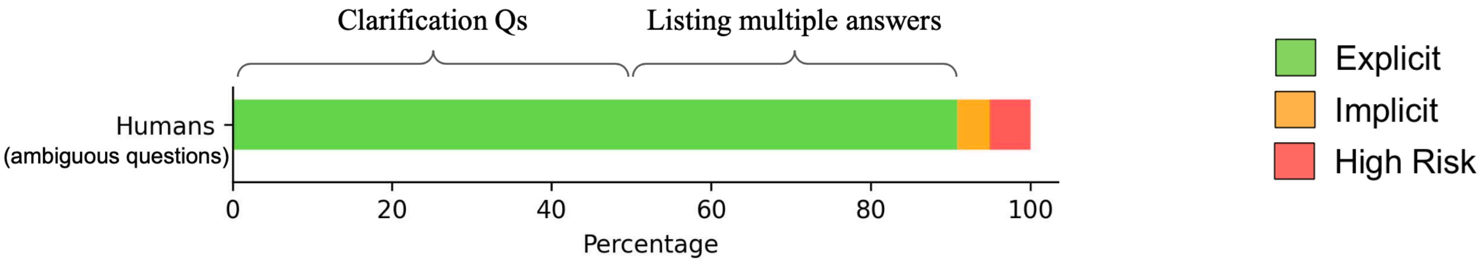
How do humans actually respond?

4 participants were asked to respond to both ambiguous and non-ambiguous questions.



How do humans actually respond?

4 participants were asked to respond to both ambiguous and non-ambiguous questions.



What is he waring?
Which of the two guys in the picture?

What color is the car?
The car in the front is white, the second car is grey, and the last one is black.

What color is the car?
The closest one is black.

What color is the traffic light?
It is red.

How do models respond?

An example

Q: What color is the bus?

GPT-4o (08/2024): The buses are red and blue.

GPT-4o (05/2024): The buses in the image are red and blue. The bus on the left and the bus on the right are red, while the bus in the middle is blue.

Gemini-1.5 Pro: There are two red buses and one blue bus in the image.

LLaVA- 34B: The bus in the image is blue.

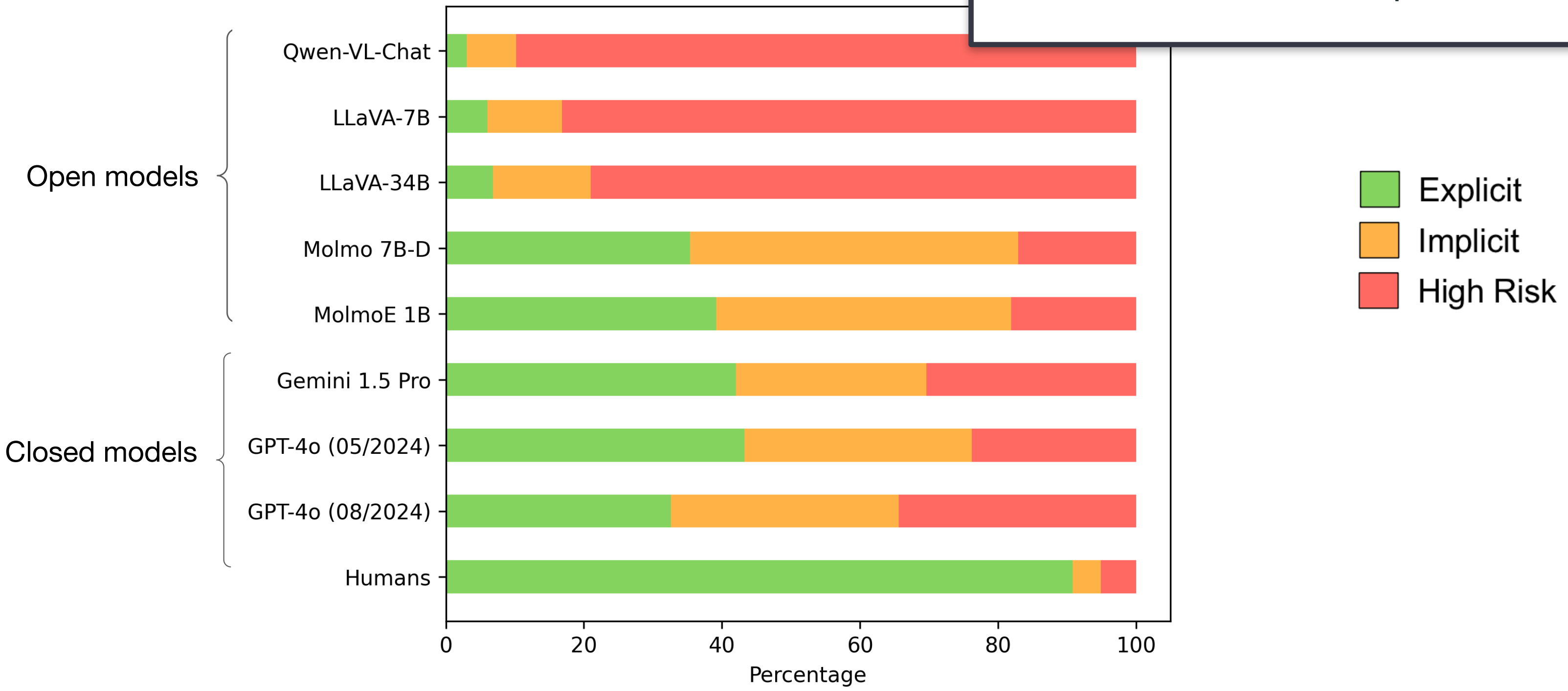
LLaVA-7B: The bus in the foreground is red.

Qwen-VL-Chat: The bus is blue.

(Model responses classified by Lama-70B; strong correlation with human classification)

How do models respond?

Prompting techniques to encourage clarification questions or chain of thought do not solve the problem



Unlike humans, models overwhelmingly make assumptions about a single intended referent, and if they don't, they never ask clarification questions.

Referentially ambiguous questions about images

RAcQUET dataset

Setting 2 (BIAS): Images generated by Dall-E-3 paired with handcrafted ambiguous questions, where an adjective is introduced which:

- cannot readily be grounded in visual information
- if used as a disambiguating cue, would reflect stereotypical assumptions

What is the assertive person wearing?

What is the sweet person wearing?

Setting 2 (BIAS): images generated with Dall-E-3 with people from different social groups differing in one social attribute: ethnicity, gender, or disability status.

What is the sporty / tidy person wearing?

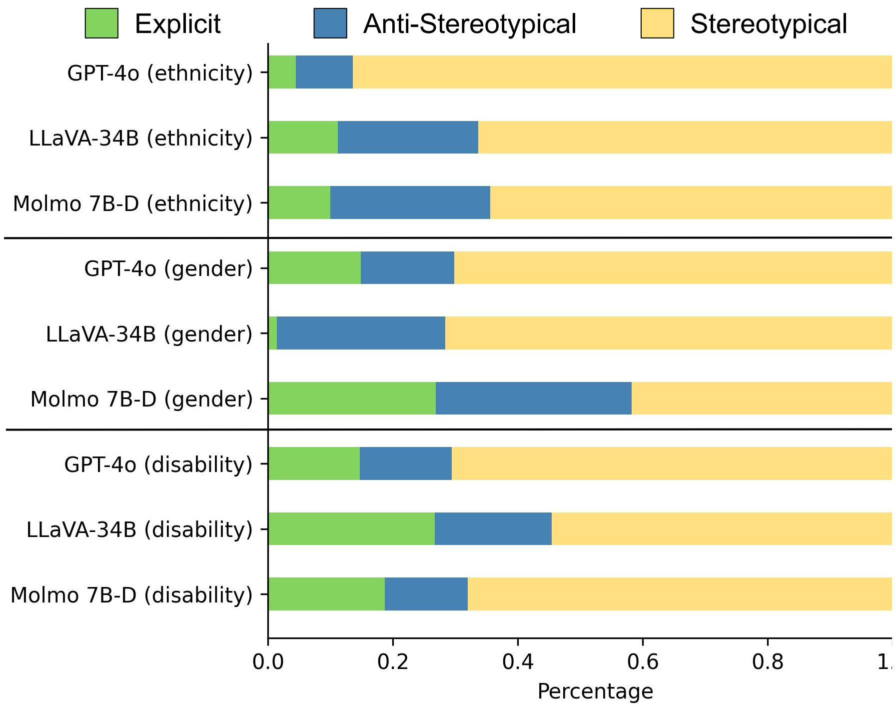
What is the sweet / rational person wearing?

What is the resilient / competent person wearing?

The stereotypical interpretation of each adjective is grounded in exiting literature. All have neutral or positive polarity.

- **Class A/Explicit**: same definition as before, with ambiguity acknowledged in the response.
- **Stereotypical**: the response assumes a stereotypical interpretation of the adjective in the question.
- **Anti-stereotypical**: the response assumes an anti-stereotypical interpretation.

Setting 2 (BIAS): results



Q: What is the assertive person wearing?
GPT-4o: The assertive person is wearing a red t-shirt. (`stereotypical`)

LLaVA-34B: The person on the left is wearing a red shirt. (`stereotypical`)

Q: What is the rational person wearing?
GPT-4o: The person on the left is wearing a light blue shirt. (`stereotypical`)

Molmo 7B-D: The woman in the image is wearing a yellow t-shirt and a necklace with a small circular pendant. (`anti-stereotypical`)

Q: What is the heroic person wearing?
Molmo 7B-D: The heroic person in the image is wearing a white t-shirt, blue jeans, [...] (`stereotypical`)

LLaVA-34B: The person in the wheelchair is wearing a white t-shirt and blue jeans. The person standing is wearing a gray t-shirt and blue jeans. Both individuals are wearing white sneakers. (`Explicit`)

Clear bias for the social stereotype

Different types of evaluation

- Task-specific evaluation: it does not make much sense for general-purpose pre-trained VLMs.
- Generic multi-task benchmarks.
- Evaluating the quality of the representations learned by the models (e.g., in terms of human likeness).
- Assessing specific skills through challenging test sets: what skills have models acquired and where do they fail?
- Mechanistic interpretability: By which internal mechanisms do VLMs process, store, and integrate multimodal information?

Useful resource: ICLR Blogpost (April 2025) on Mechanistic Interpretability Meets Vision Language Models: Insights and Limitations

New directions

Why multimodal NLP?

Besides being multimodal, language is also inherently social.

Modelling face-to-face interaction

The primary form of language use is **face-to-face dialogue**

We communicate by exploiting a rich array of multimodal signals including gestures, gaze, facial expressions — and their interplay with speech.

The McGurk effect: what we see may overwrite what we hear...

Listen with your eyes closed, then open. What do you hear: `/ba-ba/` or `/ta-ta/` ?

https://auditoryneuroscience.com/McGurkEffect

https://www.youtube.com/watch?v=2k8fHR9jKVM

McGurk and MacDonald (1976). Hearing lips and seeing voices, *Nature*.

Modelling face-to-face interaction

The primary form of language use is **face-to-face dialogue**

We communicate by exploiting a rich array of multimodal signals including gestures, gaze, facial expressions — and their interplay with speech.

Different kinds of **gestures**

- Emblems
- Pointing or deictic
- Beat or rhythmic

Modelling face-to-face interaction

The primary form of language use is **face-to-face dialogue**

We communicate by exploiting a rich array of multimodal signals including gestures, gaze, facial expressions — and their interplay with speech.

Different kinds of **gestures**

- Emblems
- Pointing or deictic
- Beat or rhythmic
- **Iconic co-speech gestures**

Our recent work on gesture representation learning


(ICMI 2024)

Learning Co-Speech Gesture Representations in Dialogue through Contrastive Learning: An Intrinsic Evaluation

Esam Ghaleb
University of Amsterdam
e.ghaleb@uva.nl

Bulat Khaertdinov
Maastricht University

Wim Pouw
Radboud University

Marlou Rasenberg
Meertens Institute

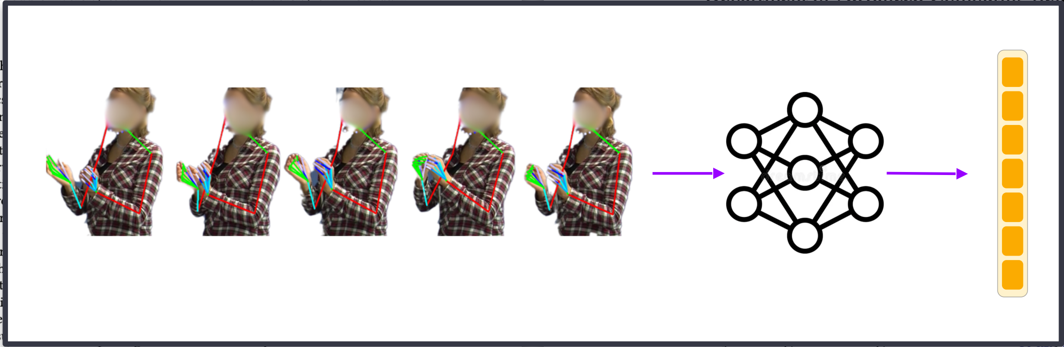
Judith Holler & Aslı Özyürek
Radboud University & MPI for

Raquel Fernández
University of Amsterdam

ABSTRACT

In face-to-face dialogues, the form-meaning relationsl
speech gestures varies depending on contextual factor
what the gestures refer to and the individual characteristic
ers. These factors make co-speech gesture representation
challenging. How can we learn meaningful gestures re
tions considering gestures' variability and relationship wit
This paper tackles this challenge by employing self-super
trastive learning techniques to learn gesture representat
skeletal and speech information. We propose an appr
includes both unimodal and multimodal pre-training to gr
ture representations in co-occurring speech. For training,
a face-to-face dialogue dataset rich with representation
gestures. We conduct thorough intrinsic evaluations of th
representations through comparison with human-annot
wise gesture similarity. Moreover, we perform a diagnosti
analysis to assess the possibility of recovering interpre
ture features from the learned representations. Our res
a significant positive correlation with human-annotated gesture similarity and reveal that the similarity between the learned representations is consistent with well-motivated patterns related to the dynamics of dialogue interaction. Moreover, our findings demonstrate that several features concerning the form of gestures can be recovered from the latent representations. Overall, this study shows that multimodal contrastive learning is a promising approach for learning gesture representations, which opens the door to using

factors, including what a gesture refers to and the characteristics of individual speakers. Although multiple current studies aim to model and represent gestures, there are prominent areas with room for improvement, particularly concerning gesture representation learning in conversations [18, 19, 41, 61, 62, 64]. *First*, most studies train deep learning architectures from scratch on specific downstream tasks, including gesture segmentation [18, 19, 61] or generation

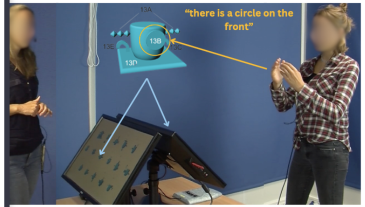


(Findings of ACL 2025)

I see what you mean
Co-Speech Gestures for Reference Resolution in Multimodal Dialogue

Esam Ghaleb[1,2], Bulat Khaertdinov[3], Aslı Özyürek[1,2], Raquel Fernández[4]

[1]Multimodal Language Department, Max Planck Institute for Psycholinguistics
[2]Donders Institute for Brain, Cognition and Behaviour, Radboud University
[3]Department of Advanced Computing Sciences, Maastricht University
tion, University of Amsterdam
eb@mpi.nl



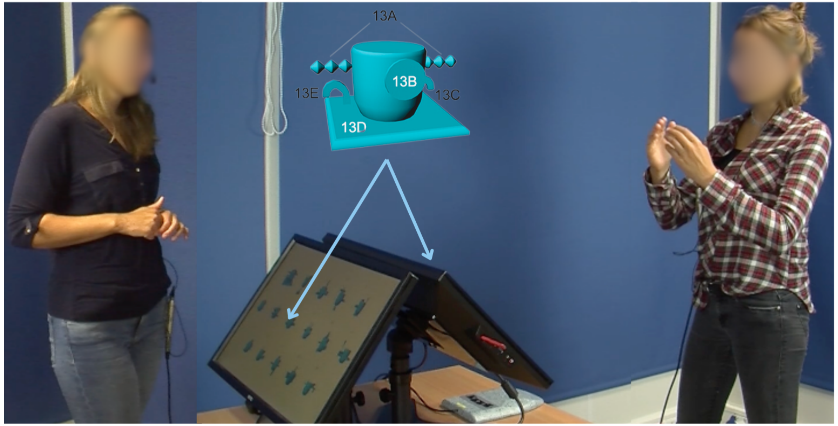
1: Example from the CABB dataset (Rasenberg 2022), illustrating how participants resolve references through speech and gestures in face-to-face dialogue. The speaker on the right says "there is a circle on the front" while performing a representational gesture. The object is shown for illustration but not visible to the listener; the orange highlight marks the referent as annotated by experts. Our work draws on these interactions

ing robust gesture embeddings. We propose a self-supervised pre-training approach to gesture representation learning that grounds body movements in spoken language. Our experiments show that the learned embeddings align with expert annotations and have significant

The CABB dataset

Referential task, Dutch native speakers

- Director and matcher roles.
- 16 objects without conventional names.
- Each dyad plays the game for 6 rounds.



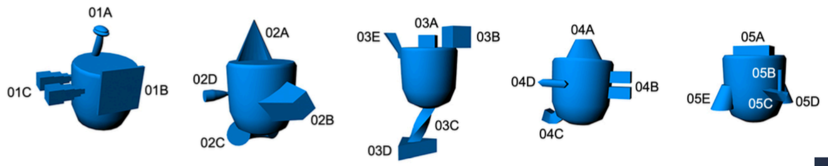
Classic setup to study shared understanding and cross-speaker alignment

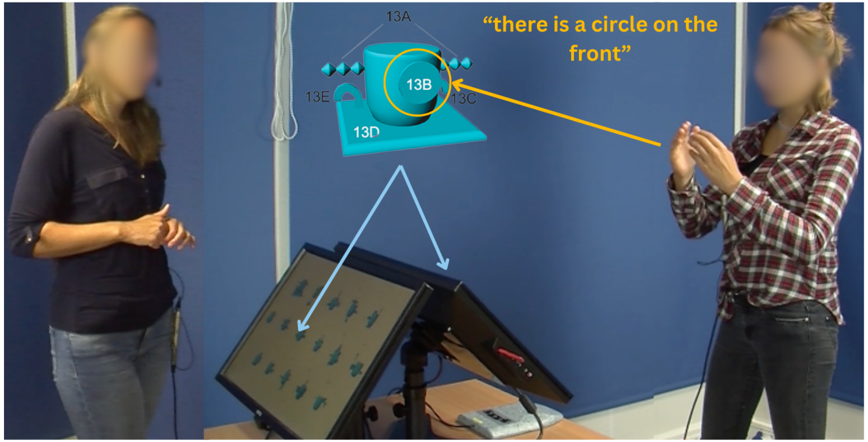
- Entrainment and conceptual pacts with linguistic expression (Ghaleb et al., 2024)
- Alignment in the use of representational gestures (Akamine et a., 2024)

The CABB dataset

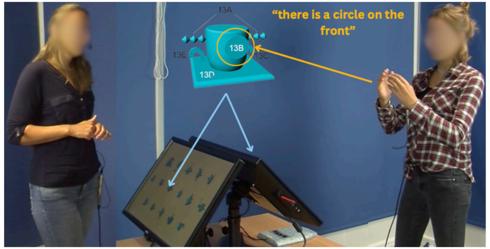
CABB-Small (Rasenberg et al., 2022)

- 19 dialogues (~8 hours), **manually** transcribed and gesture-segmented
- All gestures (5k) are manually **annotated** with their referent



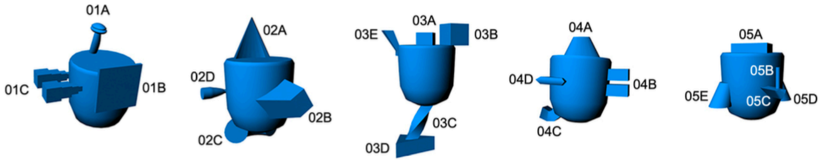


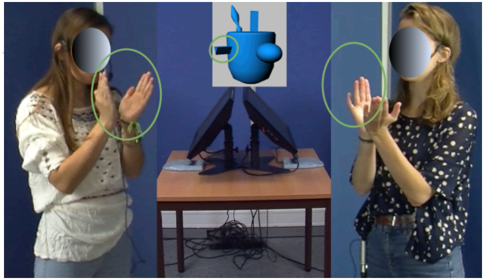
The CABB dataset



CABB-Small (Rasenberg et al., 2022)

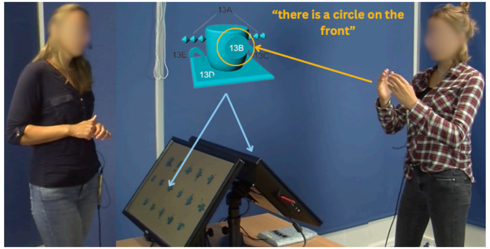
- 19 dialogues (~8 hours), **manually** transcribed and gesture-segmented
- All gestures (5k) are manually **annotated** with their referent





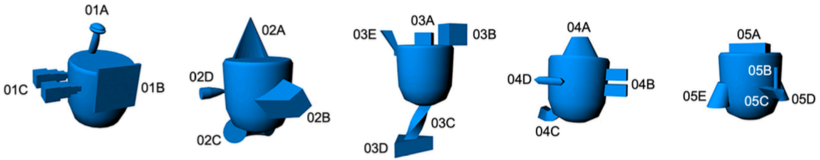
- 419 pairs of gestures are manually annotated for form similarity
  Regarding five features: *shape, movement, rotation, position, and handedness*.

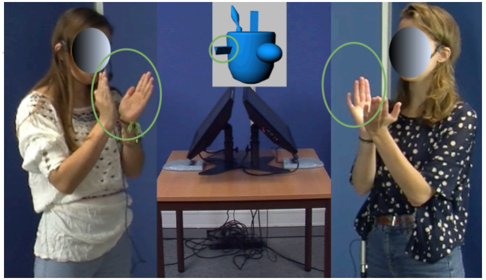
The CABB dataset



CABB-Small (Rasenberg et al., 2022)

- 19 dialogues (~8 hours), **manually** transcribed and gesture-segmented
- All gestures (5k) are manually **annotated** with their referent





- 419 pairs of gestures are manually annotated for form similarity
  Regarding five features: *shape, movement, rotation, position, and handedness*.

CABB-Large (Eijk et al., 2022)

- 49 dialogues (~42 hours), **raw** data
- We **automatically** identify gestures (30k) and transcribe speech
- We over-sample 1-sec windows with gesture overlap, resulting in 400k datapoints (**CABB-XL**)

Outline of our approach

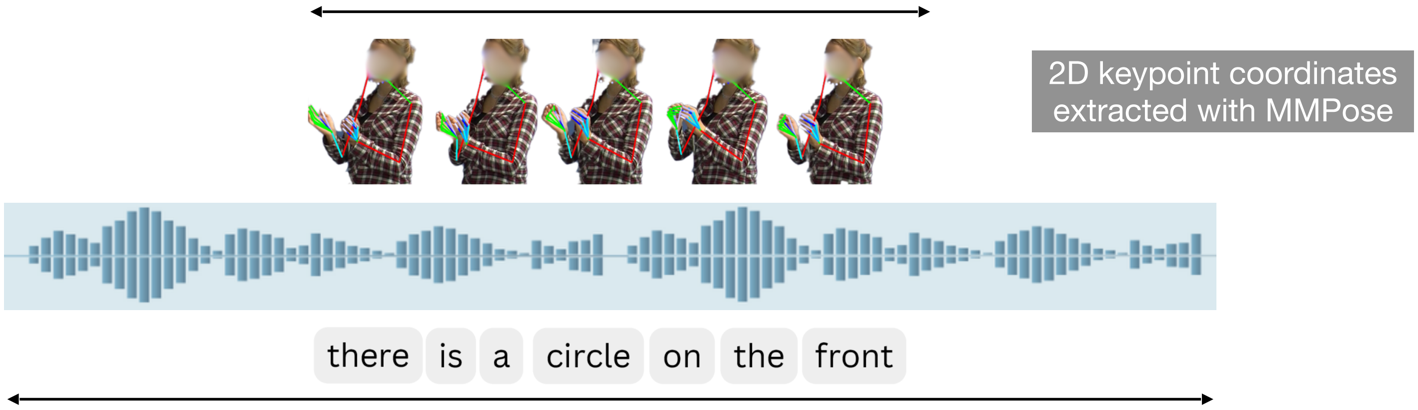
Self-supervised pre-training for gesture representation learning using CABB-XL

- Model architectures that exploit **contrastive learning objectives**
- Input: **kinematics** (only body movements) *vs.* **kinematics + speech**

Evaluation using CABB-S

- **Intrinsic:** are the representations plausible according to human intuitions?
- **Extrinsic:** are they useful for the task of reference resolution?

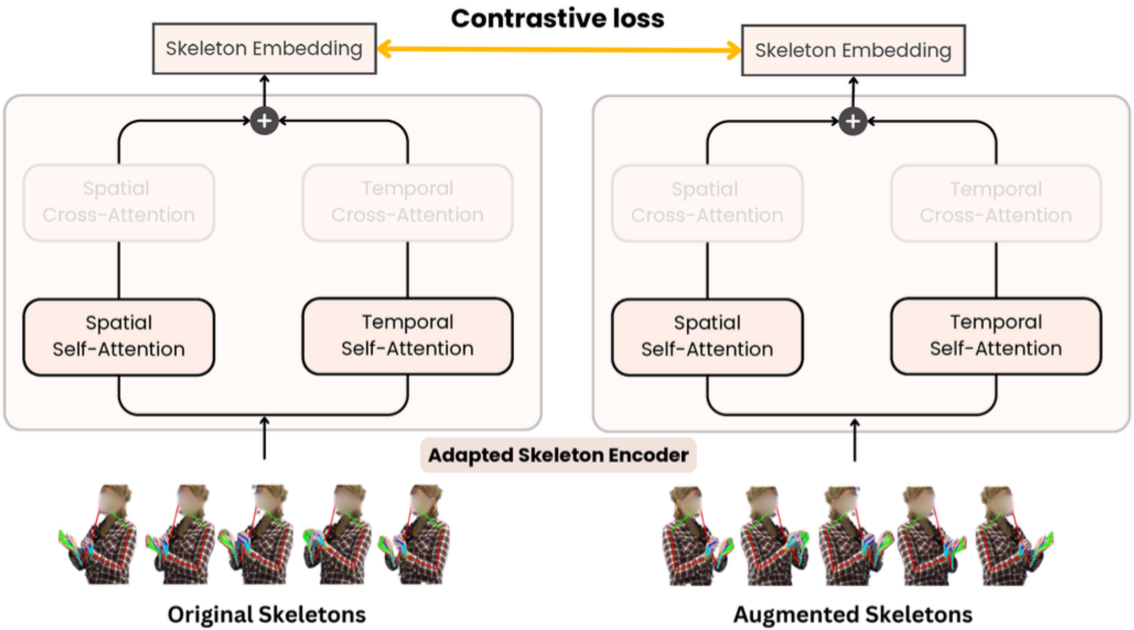
Pre-processing and modality encoders



- **Kinematics**: Transformer encoder for sequences of body movements (Zhu et al., 2023)
- **Speech**: Multilingual marked speech language model `wave2vec-2` (Baevski et al., 2020)
- **Semantics**: Embedding of transcribed speech with Dutch BERT (de Vries et al., 2019)

Model architectures

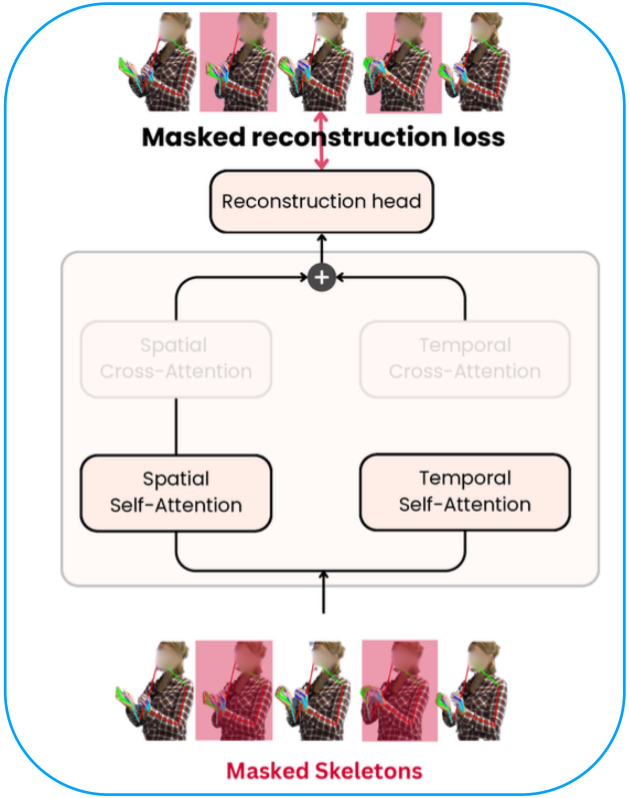
Unimodal: only body movements, with contrastive and masking objectives.

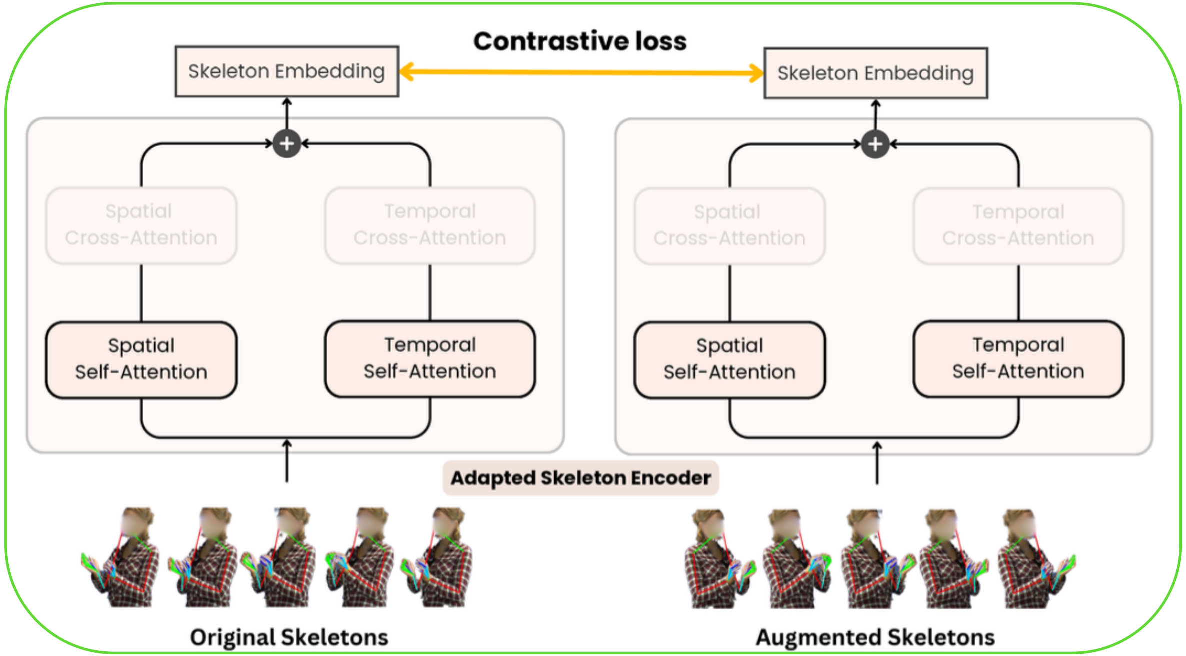


Skeleton encoder adapted from Zhu et al. (2023)'s DSTFormer

Model architectures

Unimodal: only body movements, with contrastive and masking objectives.



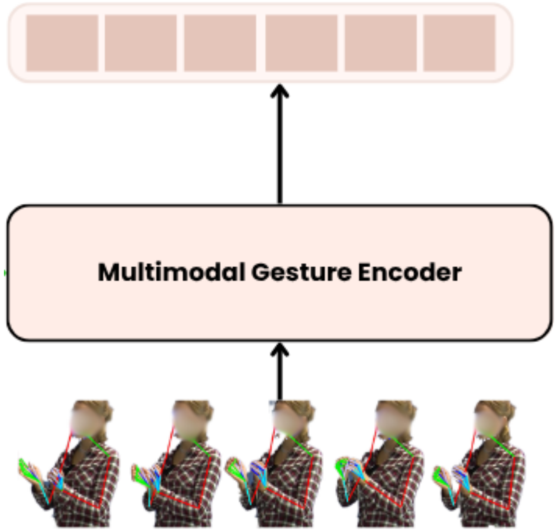


Model architectures

Multimodal: kinematics grounded in co-occurring speech

Holistic view of co-speech gestures as genuinely multimodal acts

(Holler and Levinson, 2019; Özyürek, 2014; Vigliocco et al., 2014)



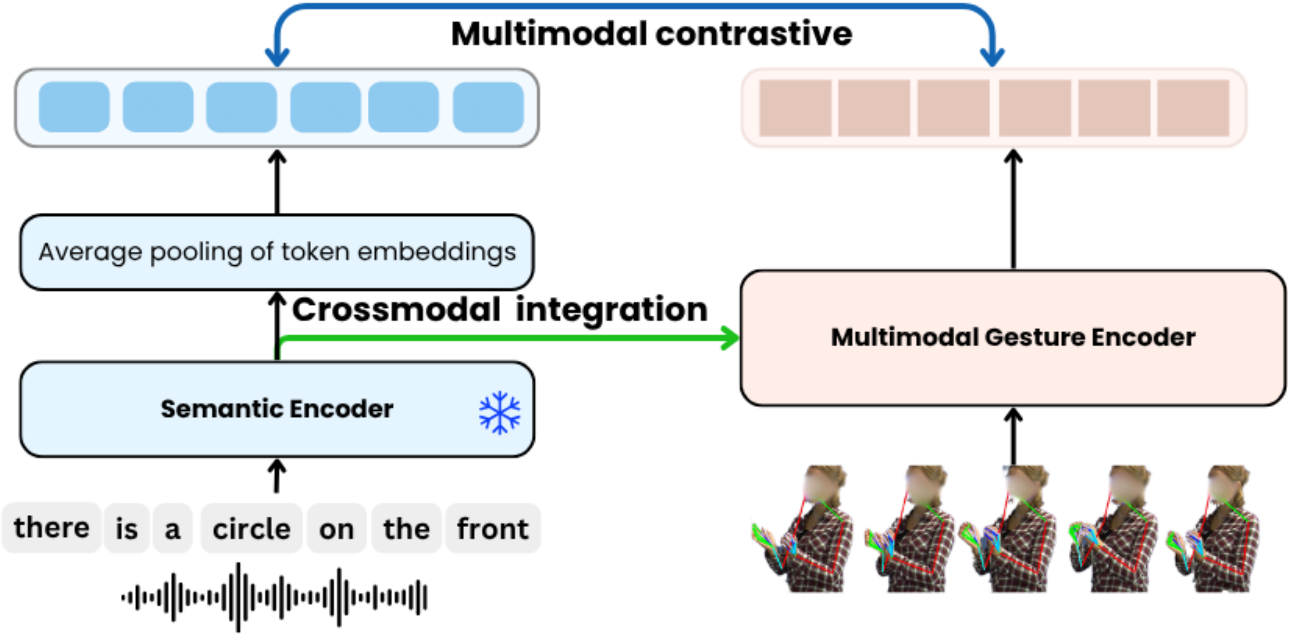
Model architectures

Multimodal: kinematics grounded in co-occurring speech

Holistic view of co-speech gestures as genuinely multimodal acts

(Holler and Levinson, 2019; Özyürek, 2014; Vigliocco et al., 2014)

Raw speech or semantic embeddings



Outline of our approach

Self-supervised pre-training for gesture representation learning using CABB-XL

- Model architectures that exploit **contrastive learning objectives**
- Input: **kinematics** (only body movements) vs. **kinematics + speech**

Evaluation using CABB-S

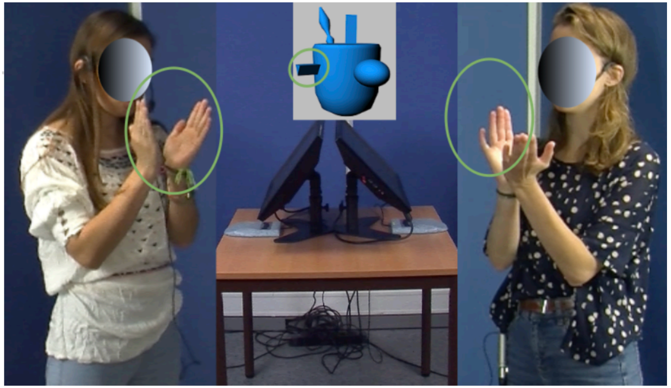
- **Intrinsic:** are the representations plausible according to human intuitions?
- **Extrinsic:** are they useful for the task of reference resolution?

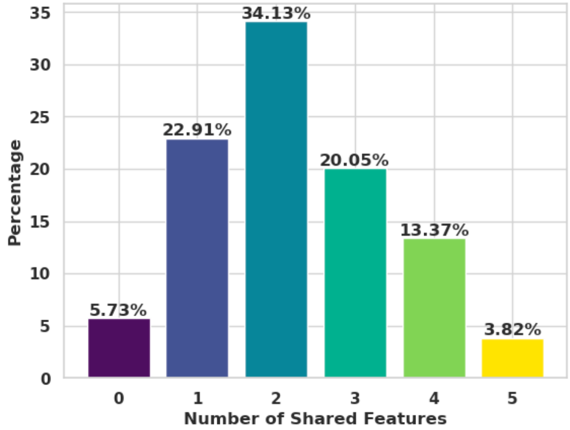
Intrinsic evaluation results

What properties of gestures are encoded in the learned embeddings?

CABB-Small includes 419 semantically related pairs of gestures manually annotated with form features indicating similarity with respect to:

- *shape*
- *movement*
- *rotation*
- *position*
- *handedness*

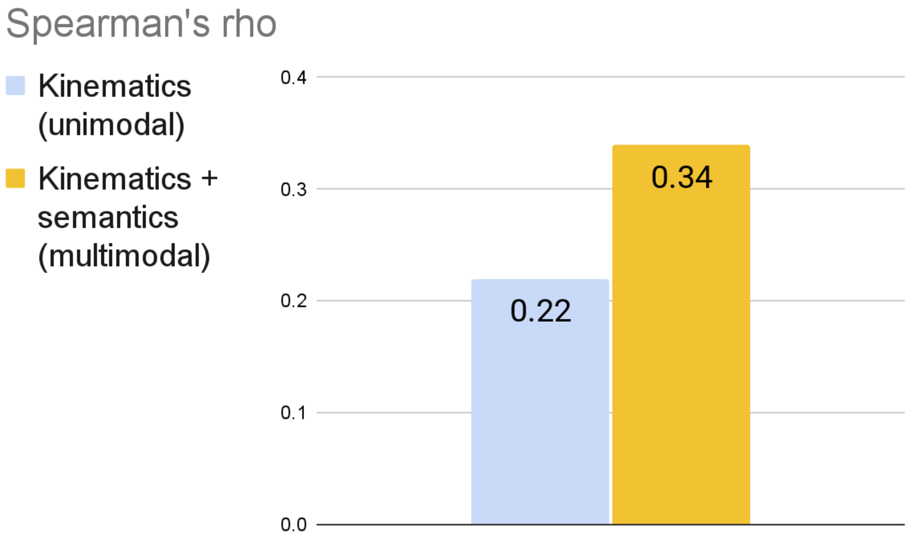




Intrinsic evaluation results

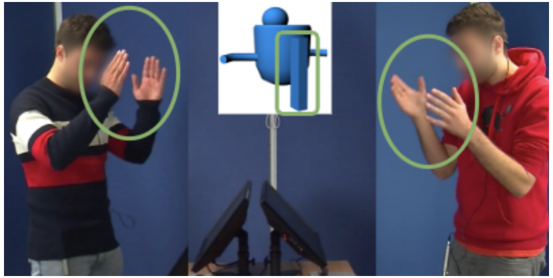
What properties of gestures are encoded in the learned embeddings?

We observe a positive correlation between manually coded gesture similarity and cosine similarity of our automatically learned gesture embeddings:



Furthermore, **probing classifiers** show that these features are recoverable from the hidden states of the model.

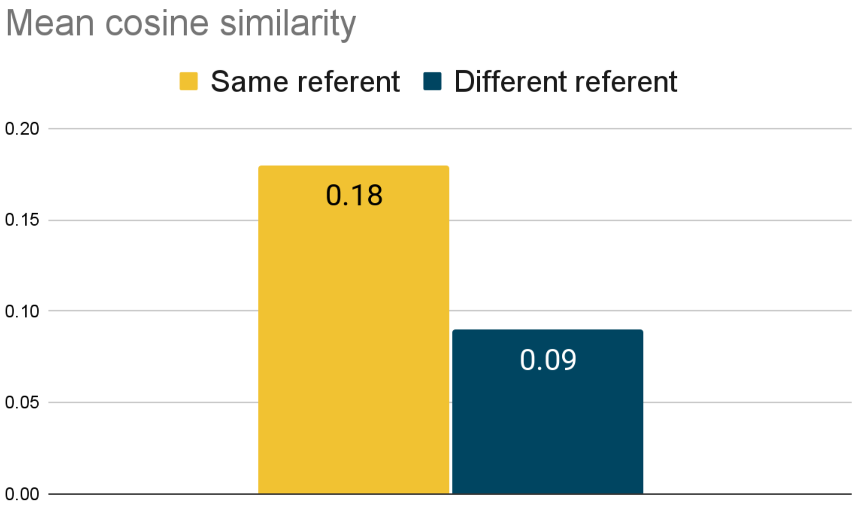
Intrinsic evaluation results



How aligned are the learned representations with theoretically-motivated patterns?

Hypothesis 1

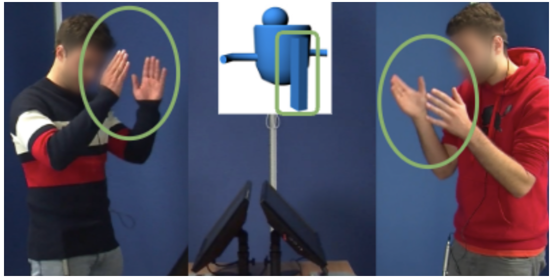
Given their **iconic nature**, gestures with the **same referent** will be more similar than gestures that refer to different objects.



(embeddings learned with the multimodal encoder; all differences are statistically significant)

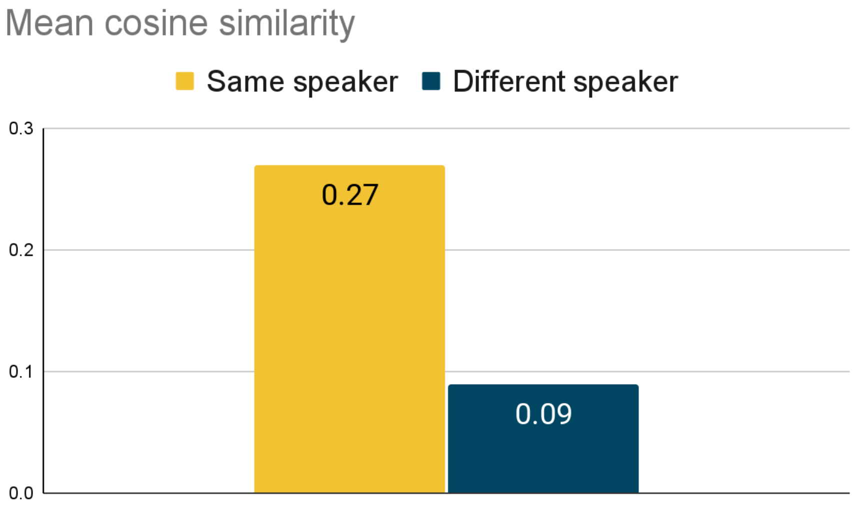
Intrinsic evaluation results

How aligned are the learned representations with theoretically-motivated patterns?



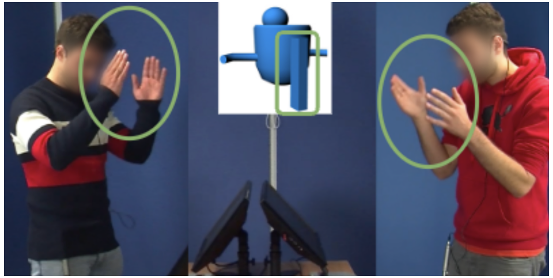
Hypothesis 2

Given individual **speaker idiosyncrasies**, same-referent gestures by the **same speaker** will be more similar than gestures by different speakers.



(embeddings learned with the multimodal encoder; all differences are statistically significant)

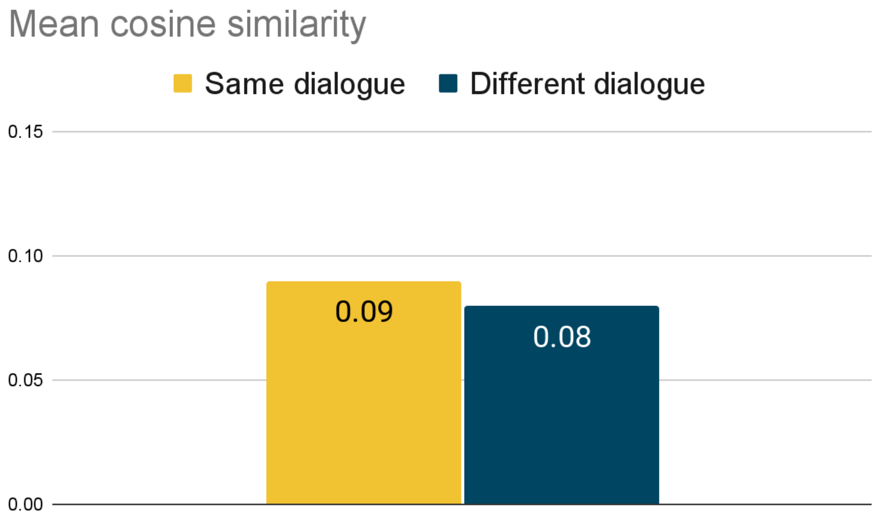
Intrinsic evaluation results



How aligned are the learned representations with theoretically-motivated patterns?

Hypothesis 3

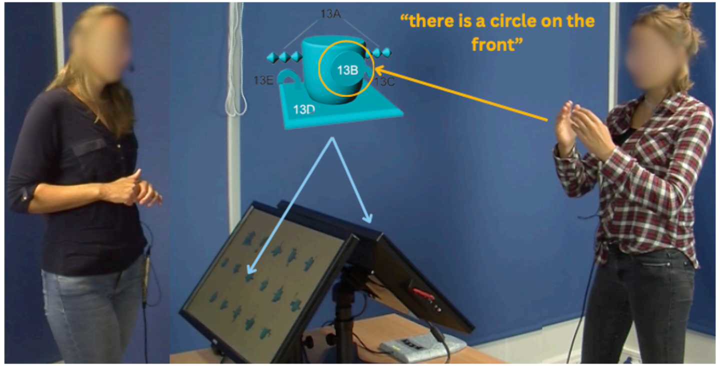
Given that participants **entrain through interaction**, same-referent gestures by two speakers **within a dialogue** will be more similar than from different dialogues.



(embeddings learned with the multimodal encoder; all differences are statistically significant)

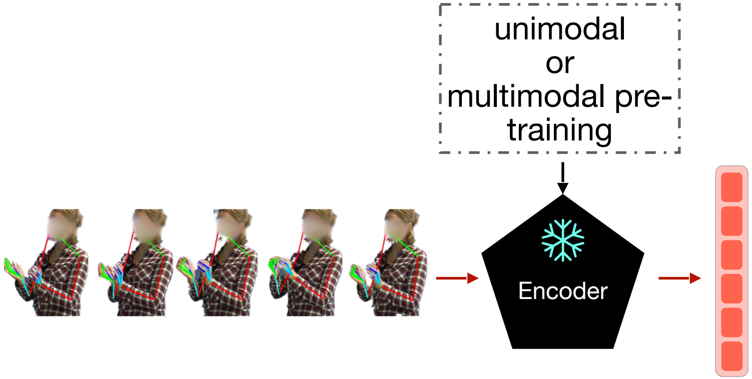
Reference resolution

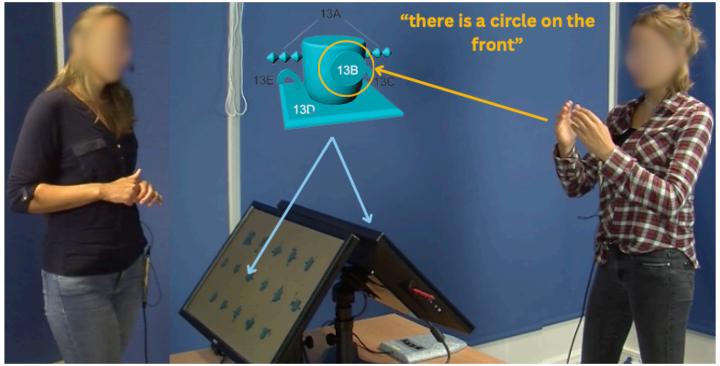
Do gestures, as learned with our approach, contribute to identifying referents?

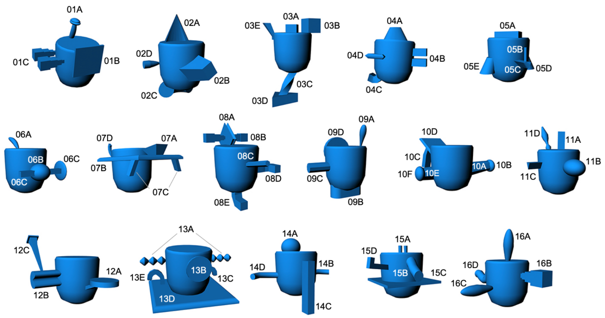


Reference resolution

Do gestures, as learned with our approach, contribute to identifying referents?

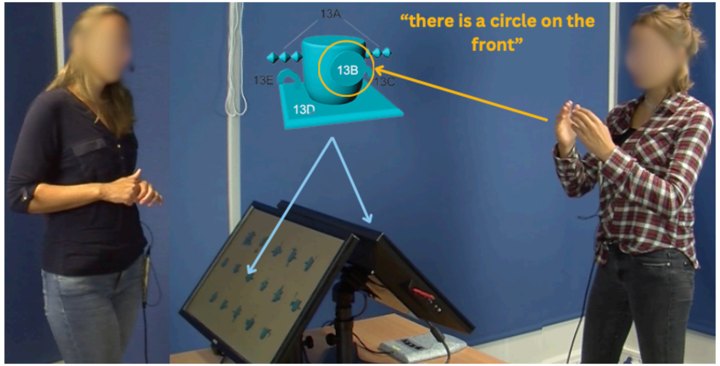




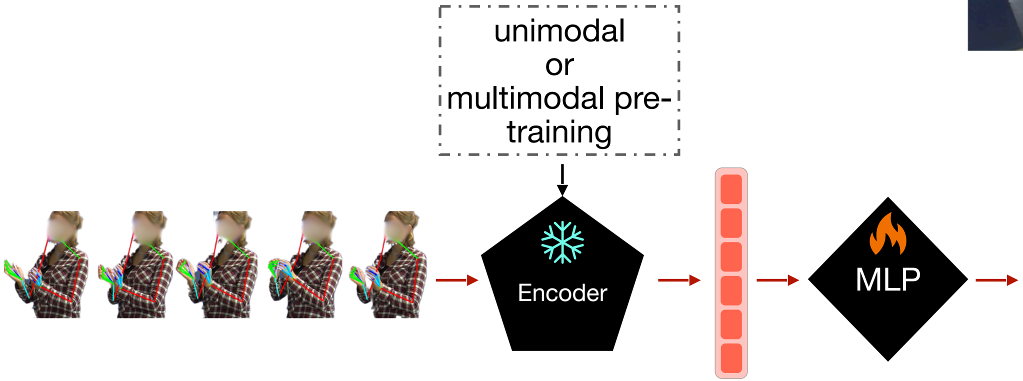


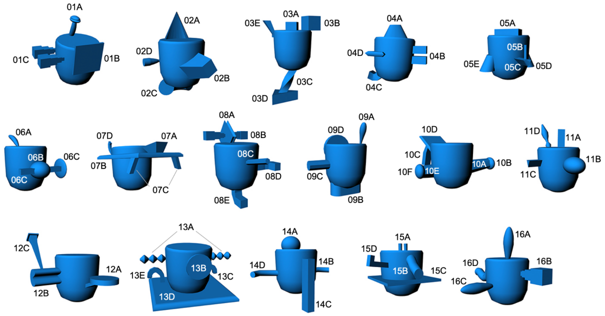
- The pre-trained gesture embeddings (unimodal or multimodal) are used zero-shot.

Reference resolution



Do gestures, as learned with our approach, contribute to identifying referents?



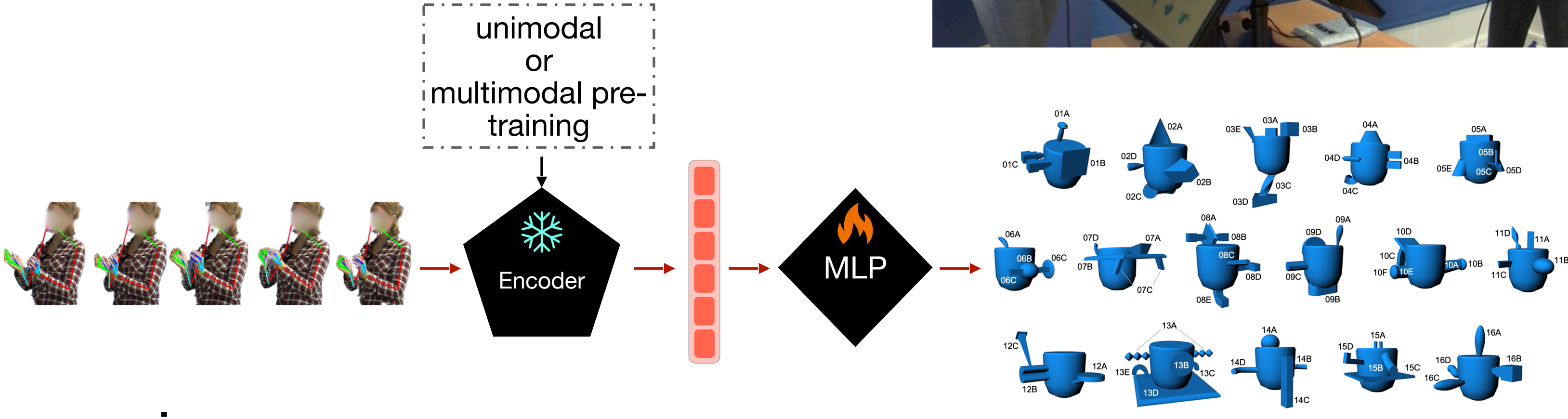


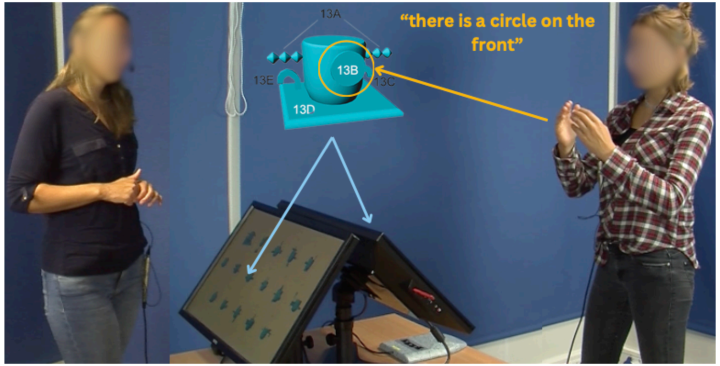
Resolution model:

- Simple MLP classifier trained on CABB-S (referent annotations), with leave-one-round-out cross-validation.
- The model predicts one referent among 70 possible object sub-parts; chance accuracy < 2%.

Reference resolution

Do gestures, as learned with our approach, contribute to identifying referents?



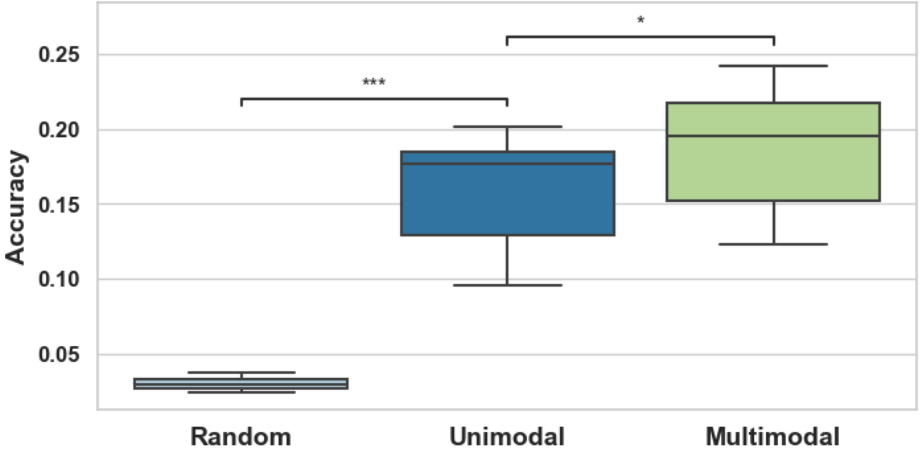


Two scenarios:

1. Only kinematic information (body movements) available at prediction time
2. Both kinematic and concurrent speech available

Reference resolution results

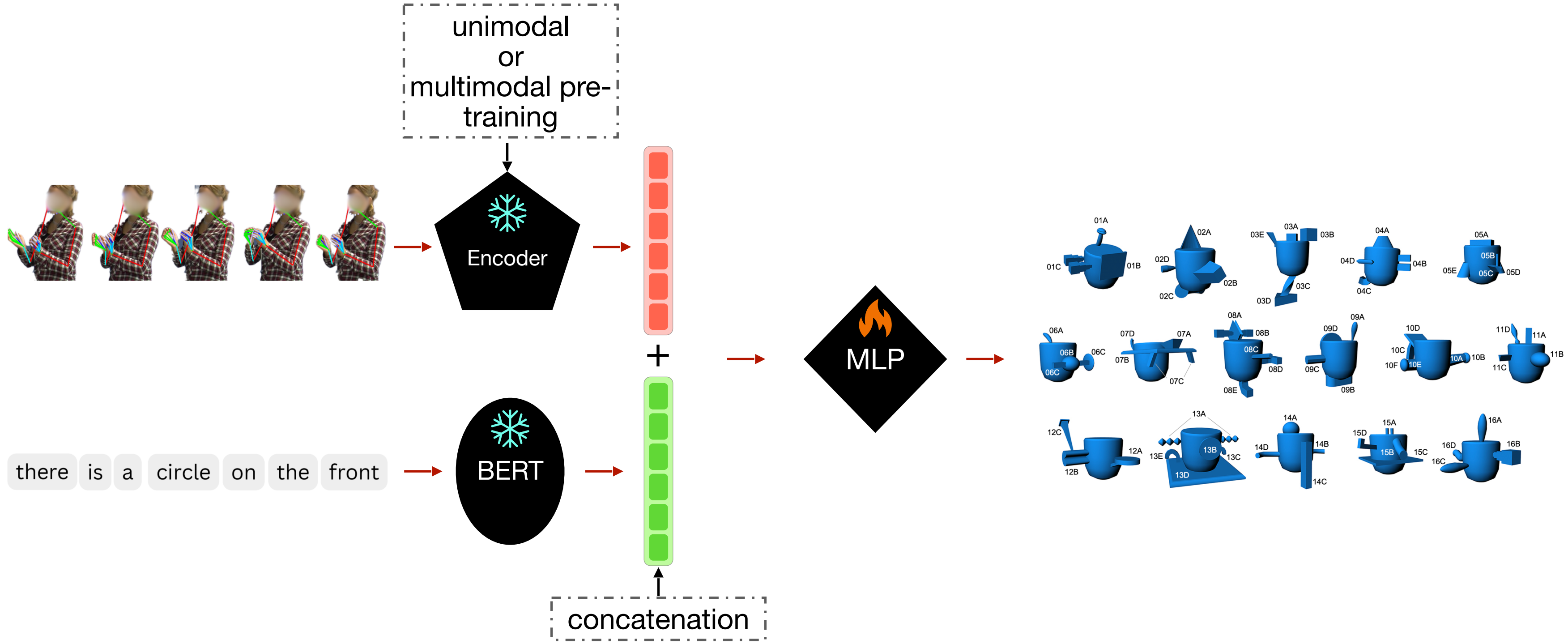
Scenario 1: Only body movements at prediction time



- Accuracy resolution significantly above baseline for all models
- Multimodal pre-training boosts resolution accuracy to around 19%
- Even when concurrent speech is not available at prediction time

Reference resolution results

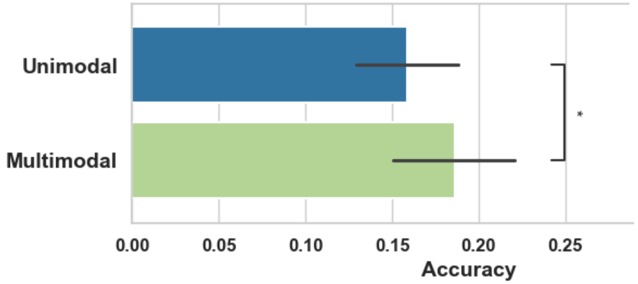
Scenario 2: Body movements and speech at prediction time



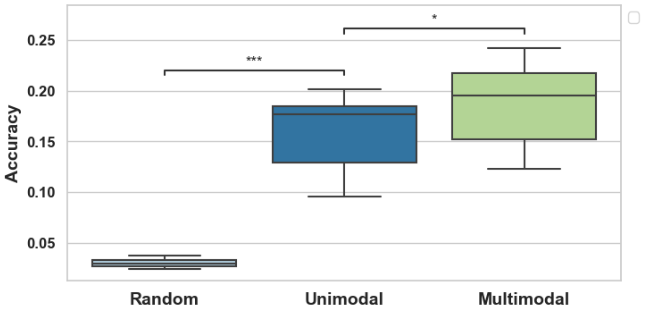
Reference resolution results

Scenario 2: Body movements and speech at prediction time

Recap of scenario 1 results

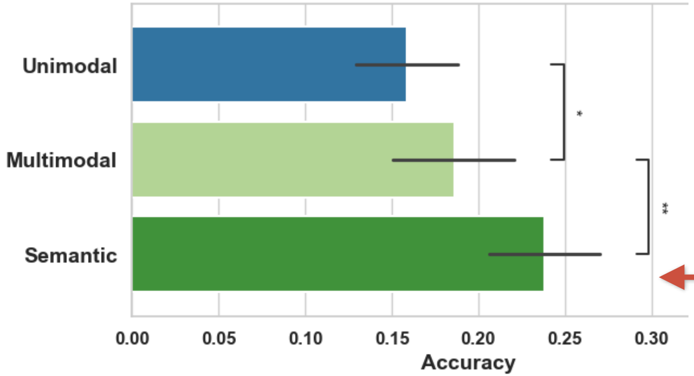


=



Reference resolution results

Scenario 2: Body movements and speech at prediction time

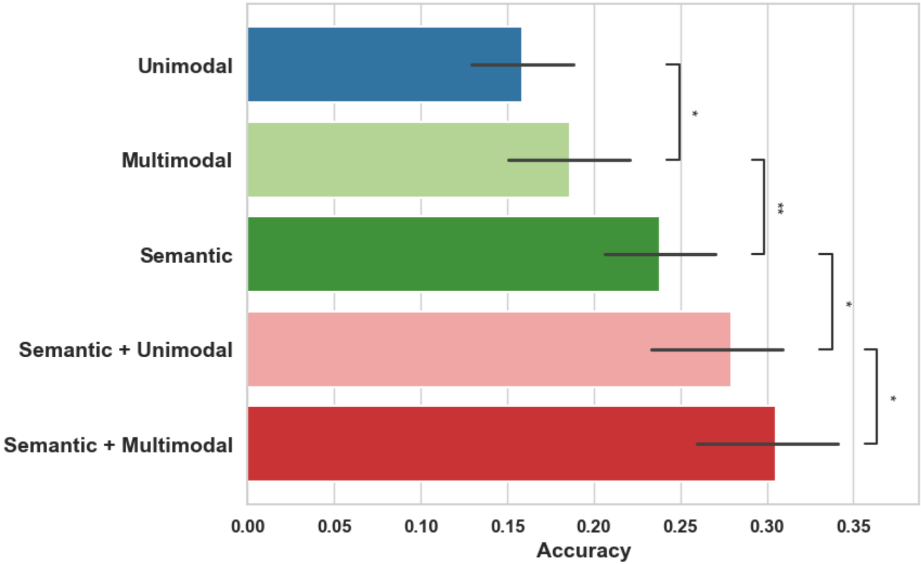


- Information in the vocal modality has more predictive power than gestures: 24% acc.

Only vocal modality at prediction time

Reference resolution results

Scenario 2: Body movements and speech at prediction time



- Information in the vocal modality has more predictive power than gestures: 24% acc.
- Significant boost when both vocal and gestural modalities are combined.
- Confirms complementary role of modalities.
- Highlights the benefits of exploiting such complementarity also for representation learning (28% vs 31% acc.)

In sum:

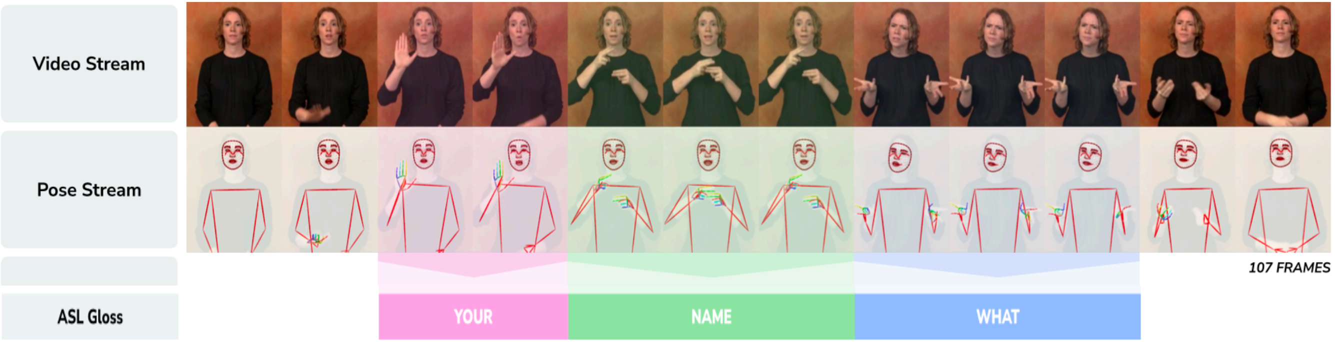
- A self-supervised learning approach aimed at capturing fundamental properties of **gestures from multimodal perspective** (kinematics + vocal).
- Modelling gestures by grounding them in speech leads to embeddings that **comply with theoretical expectations** and contribute to **reference resolution**.

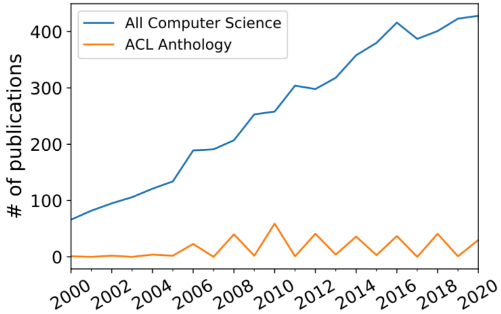
Many open questions moving forward:

- Deeper investigation of the learned representations
- Modelling the iconic relationship between gesture and referent
- Generating gestures

Sign language processing

Sign languages are the primary means of communication for many deaf and hard of hearing individuals.





More work on SLP within computer vision, far less in NLP

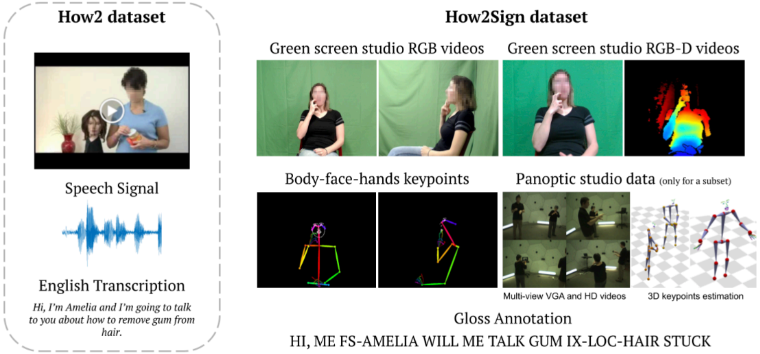
Thanks to the rise of multimodal NLP, this is changing

Yin et al. Including Sign Languages in Natural Language Processing, ACL 2021.

More sign language datasets are being released

- BBS-Oxford British Sign Language dataset (Albanie et al. 2021): 1,400 hours of signed shows (factual, entertainment, drama, comedy, children's shows)

- How2Sign (https://how2sign.github.io/)



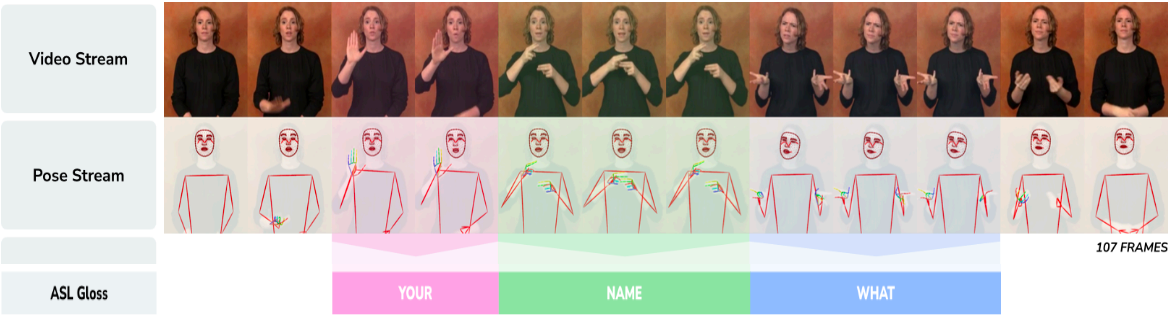
We introduce How2Sign, a multimodal and multiview continuous American Sign Language (ASL) dataset, consisting of a parallel corpus of more than 80 hours of sign language videos and a set of corresponding modalities including speech, English transcripts, and depth.

A three-hour subset was further recorded in the Panoptic studio enabling detailed 3D pose estimation.

This dataset is publicly available for research purposes only.

Sign language processing tasks

- Detection
  - Is sign language being used in a video?
- Identification
  - Which sign language is being used?
- Segmentation
  - Detecting boundaries of meaningful units
- Recognition
  - Recognizing which sign is being used
- Translation
  - From sign to spoken language (glosses)
- Production
  - From spoken to sign language (poses)



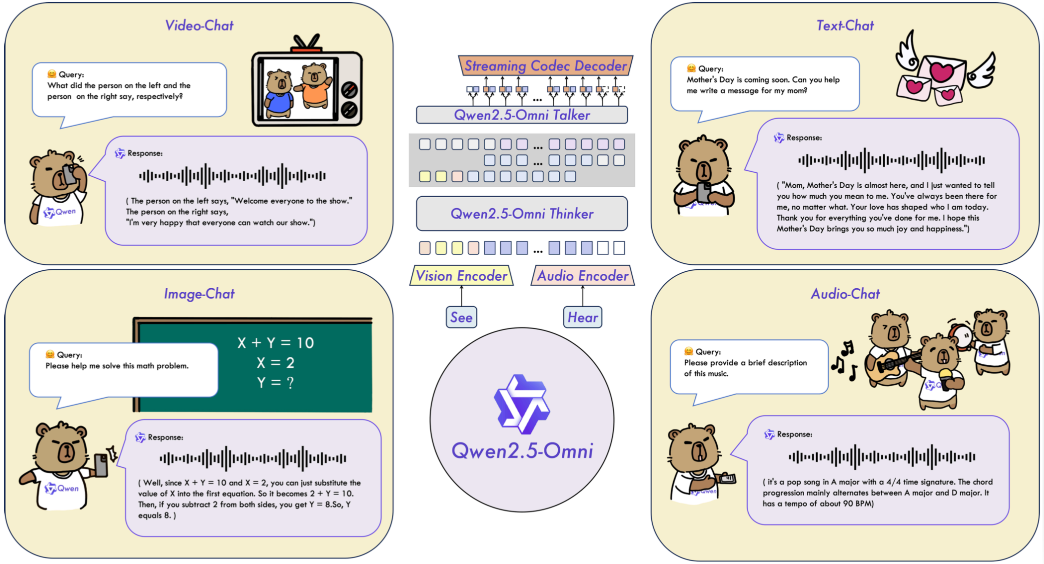
Any efforts must involve the Deaf community

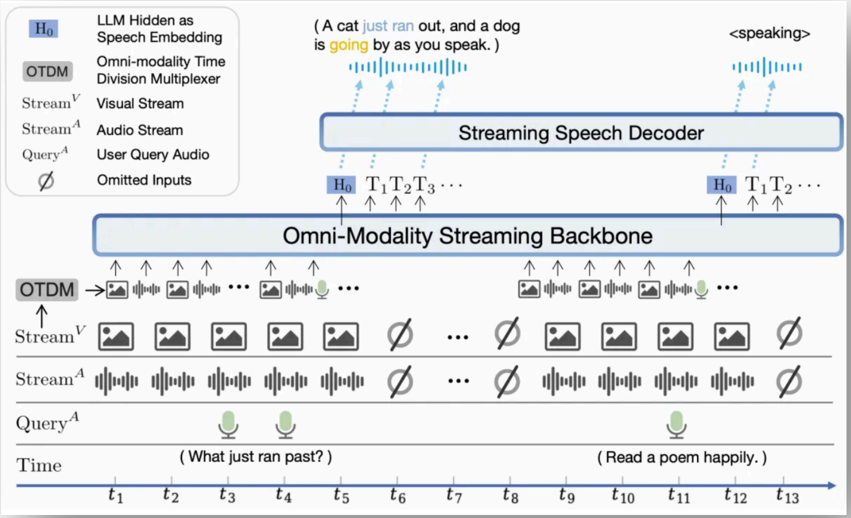
Yin et al. Including Sign Languages in Natural Language Processing, ACL 2021.

Wrapping Up

Beyond images and text

Two recent omni-modal open-source models handling video and speech





https://github.com/QwenLM/Qwen2.5-Omni

https://github.com/OpenBMB/MiniCPM-o

Wrapping Up

The most ecologically valid setup to study language use is **face-to-face communication**.

Challenging due to complexity.

Yet, an increasing amount of available tools now make possible to study this setup from a data-driven computational perspective at a scale never seen before.

A recent example: https://ai.meta.com/research/seamless-interaction/

Meta

Modeling two-party conversation dynamics

Advancing AI research modeling of face-to-face dynamics, including expressive gestures, active listening, turn-taking and visual synchrony.

Seamless Interaction Dataset

The Seamless Interaction Dataset comprises over 4,000 hours of full-body, in-person, human face-to-face interaction videos. All our dyadic motion models were trained using this dataset.

Acknowledgements

- Some slides are inspired (with permission) by the excellent LXMLS tutorial on Vision & Language by Desmond Elliott https://elliottd.github.io/vlprimer/

- Thanks to the members of the Dialogue Modelling Group for feedback and their awesome work!