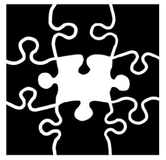
Modelling Multimodality in Human Cognition

Raquel Fernández

UNIVERSITY
OF AMSTERDAM



Institute for Logic,
Language &
Computation

LxMLS 2026

Language & cognition are inherently multimodal

Outlook

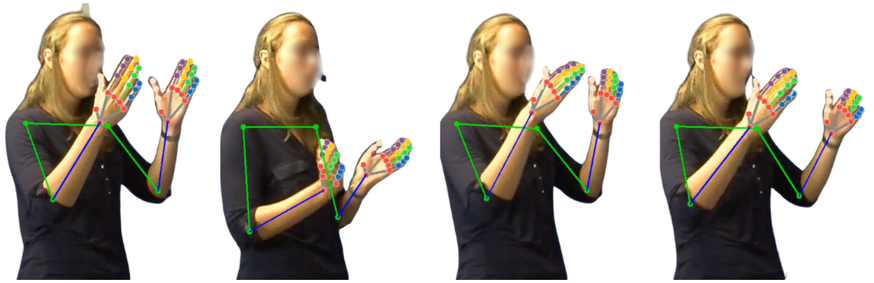
Interest in investigating questions related to multimodal cognition with computational methods

- *To what extent does language interface with perception?*
- *How can we investigate this empirically, using computational methods?*
- *Are deep learning models suitable for this aim, and what insights can they bring?*

Two topics:

- Co-speech gestures in face-to-face dialogue
- Multimodality in the brain

Co-speech gestures



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.

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

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

↳ gesture representation learning

Why learning gesture embeddings?

- A tool to study multimodal interaction from a data-driven perspective, at a larger scale than what symbolic manual annotation allows.
- Potentially useful to power applications such as embodied language understanding, gesture generation, etc.

Why learning gesture embeddings?

Unsupervised learning of representations that capture fundamental properties of gestures, independently from concrete applications.

Why learning gesture embeddings?

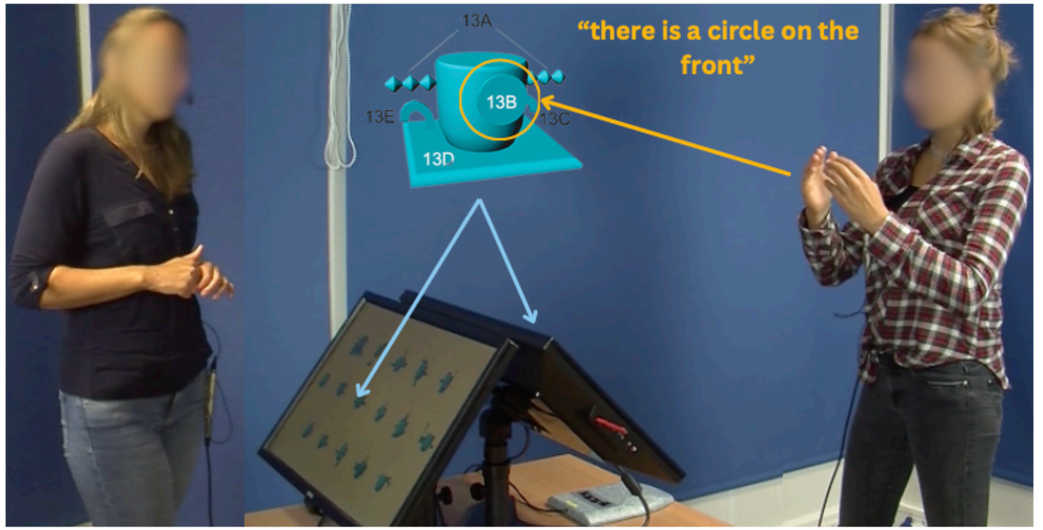
Unsupervised learning of representations that capture fundamental properties of gestures, independently from concrete applications.

Experimenting with two input types aligned with different views of gestures:

- **Unimodal**: only kinematics (body movements)
- **Multimodal**: kinematics + speech, corresponding to a holistic view of co-speech gestures as genuinely multimodal acts (Holler and Levinson, 2019; Özyürek, 2014; Vigliocco et al., 2014; i.a.)

The CABB dataset

Referential task with 16 unconventional objects.
Pairs of Dutch native speakers.



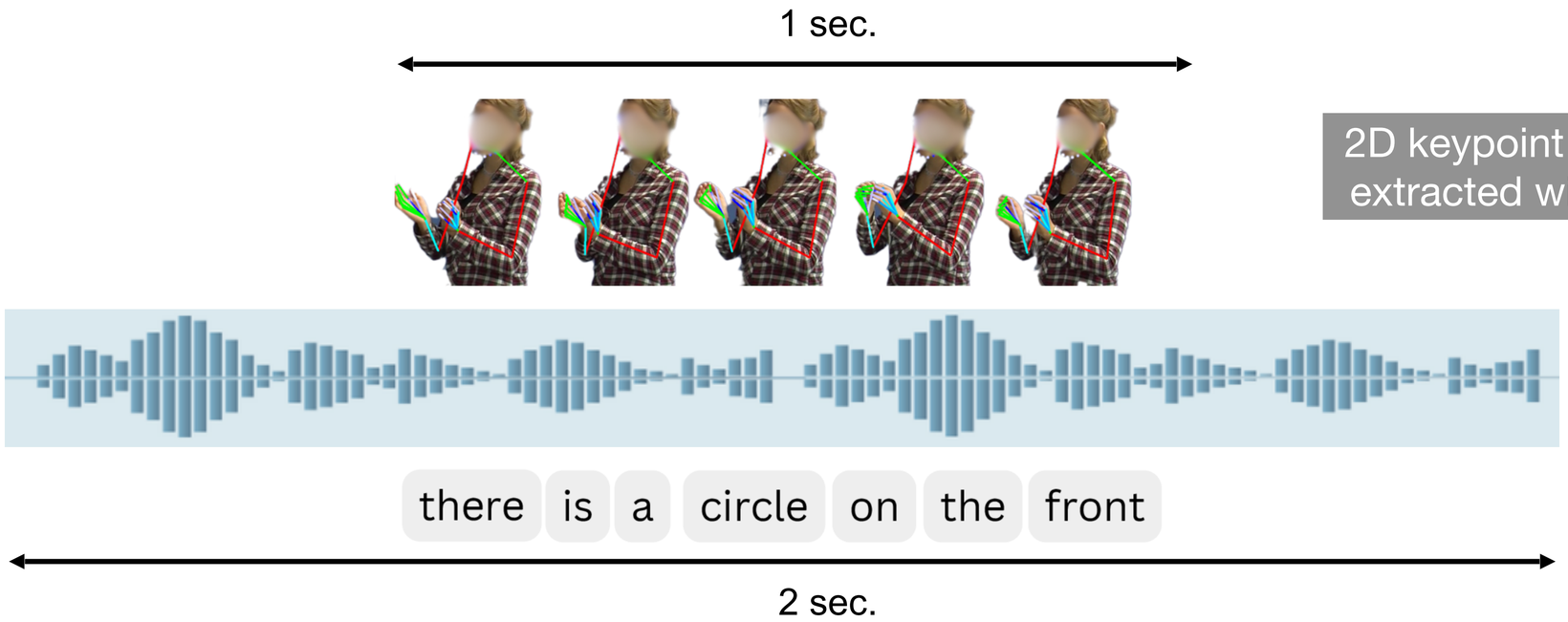
CABB-Small (Rasenberg et al., 2022)

- 19 dialogues (~8 hours), manually transcribed, gesture-segmented (5k), annotated

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

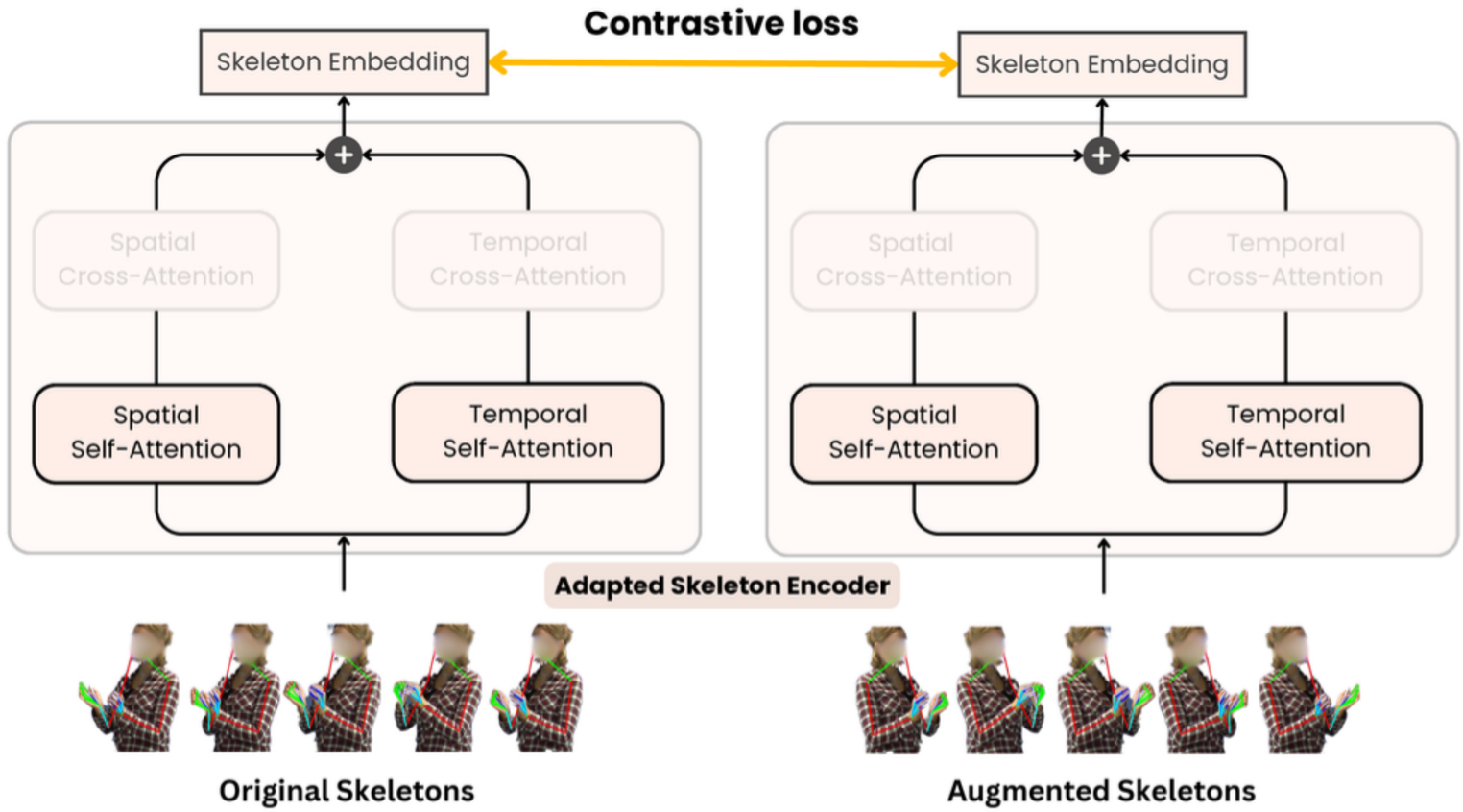
Pre-processing and pre-trained modality encoders



- **Kinematics**: Dual-stream Spatio-temporal Transformer as motion encoder (Zhu et al., 2023)
- **Speech**: Multilingual masked language model wave2vec-2 (Baevski et al., 2020)
- **Text**: Embedding of transcribed speech with BERTje (de Vries et al., 2019)

Model architectures

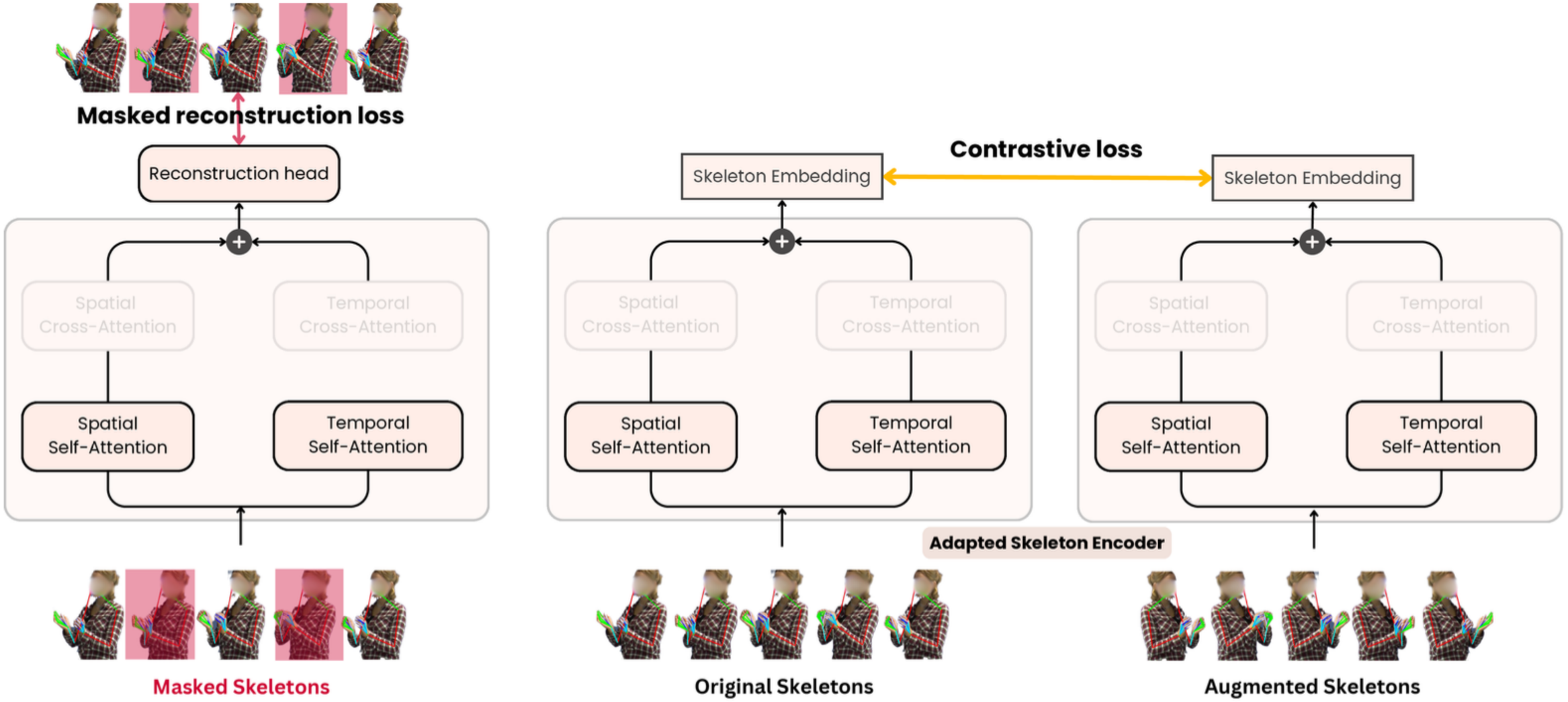
Unimodal: only body movements, with contrastive and masking objectives.



Adaptation of the DSTFormer motion encoder by Zhu et al. (2023)

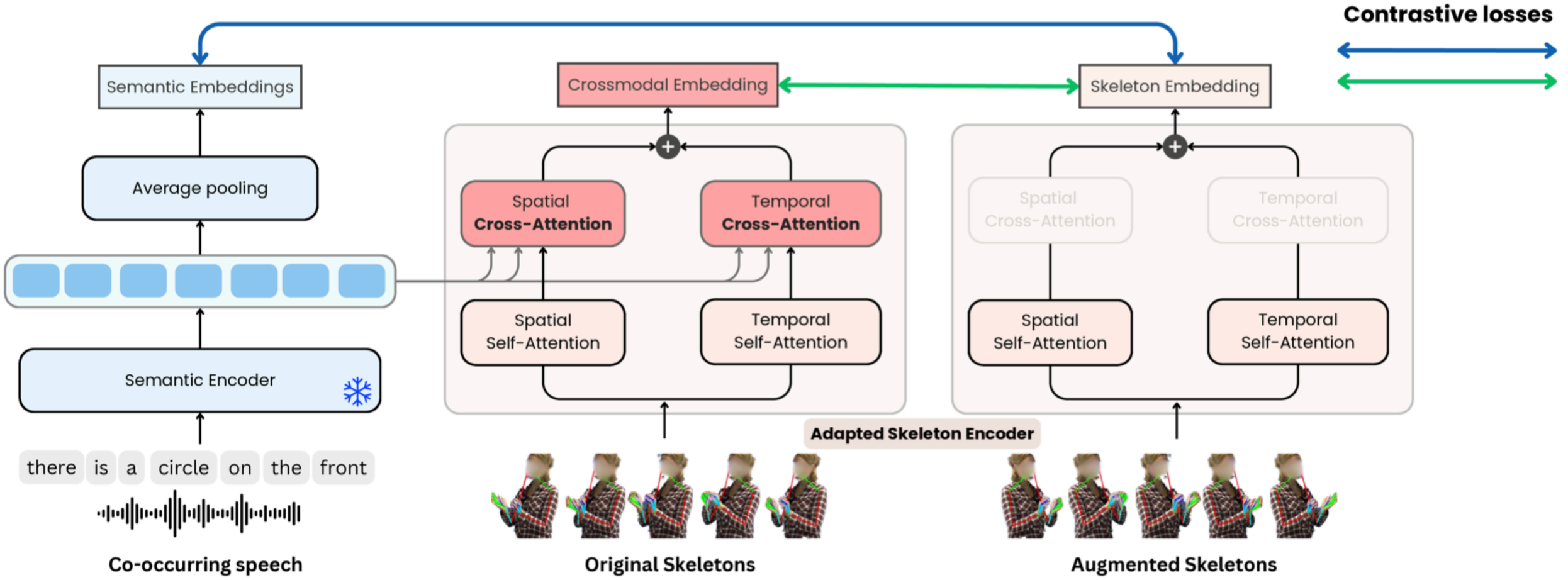
Model architectures

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



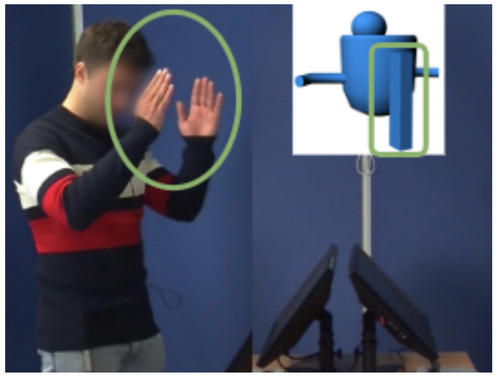
Model architectures

Multimodal: kinematics grounded in co-occurring speech



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

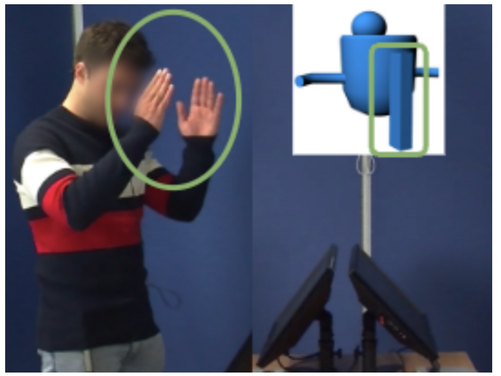
How aligned are the learned representations with theoretically-motivated hypotheses?



Hypothesis 1

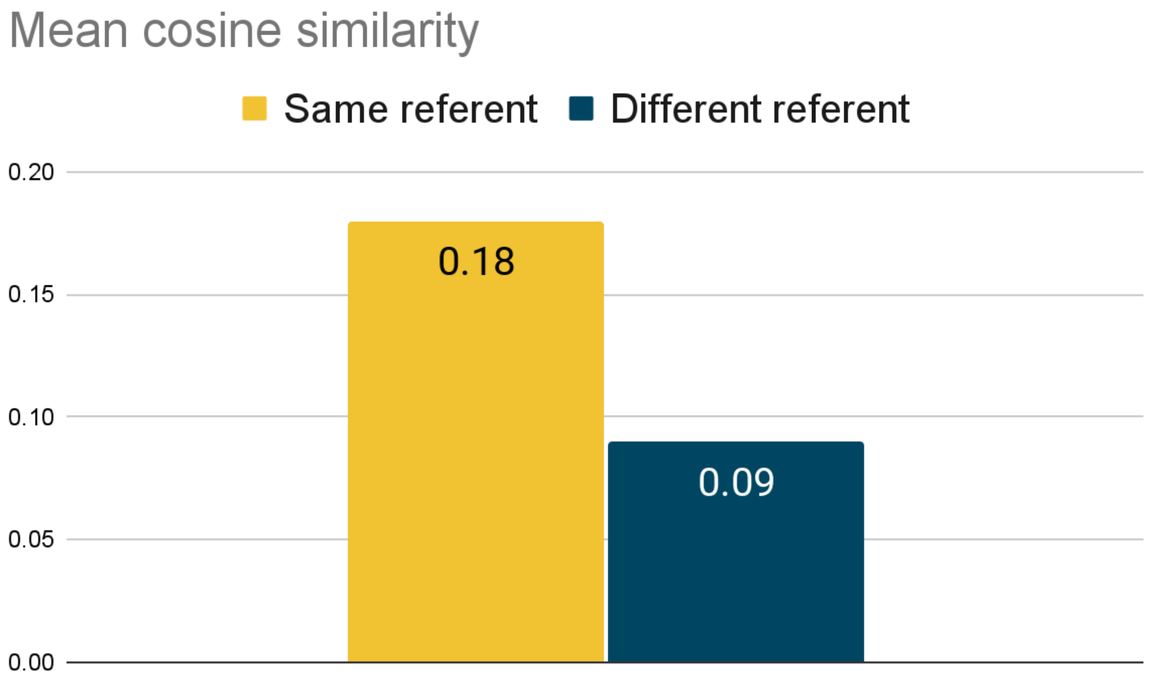
Given their **iconic nature**, gestures with the **same referent** will be more similar than gestures that refer to different objects.

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



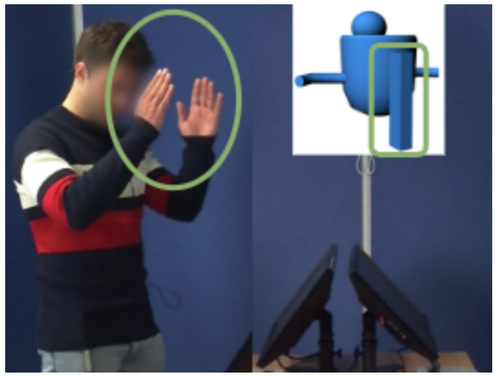
Hypothesis 1

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 architecture; same patterns hold for unimodal; all differences are statistically significant)

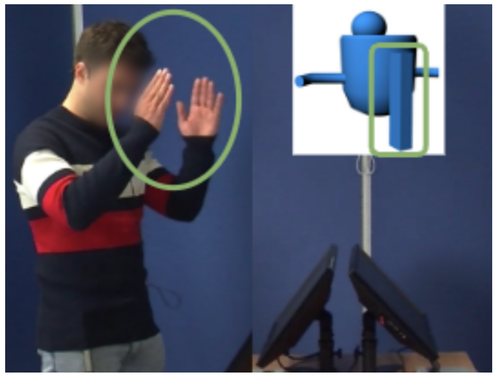
How aligned are the learned representations with theoretically-motivated hypotheses?



Hypothesis 2

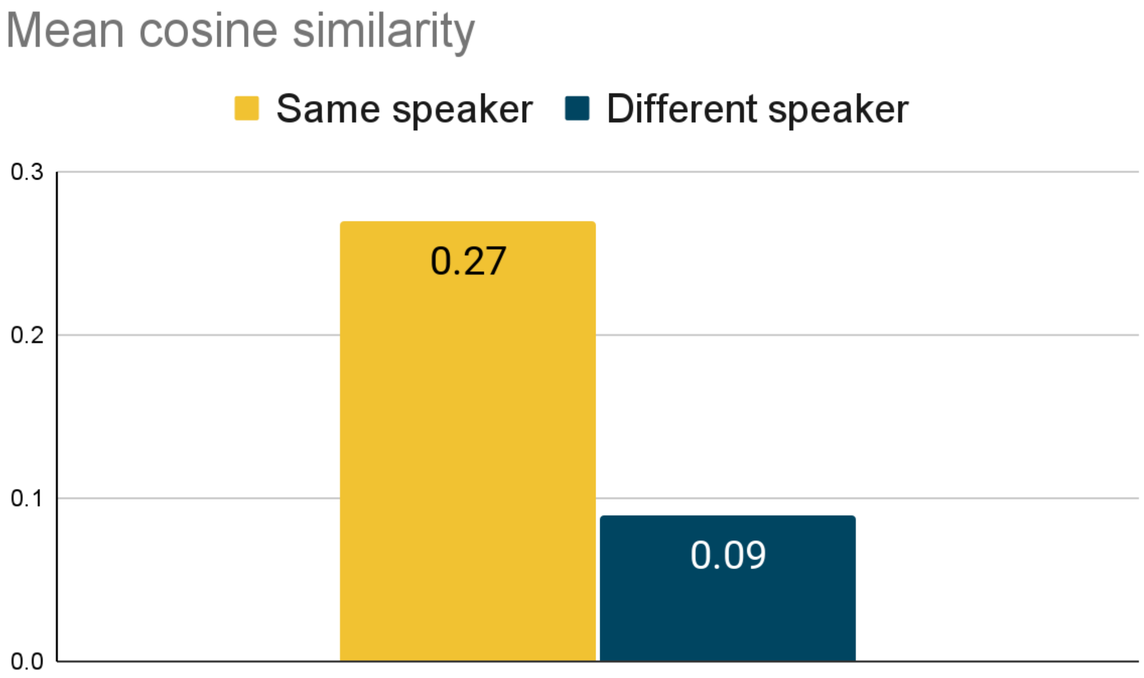
Given individual **speaker idiosyncrasies**, same-referent gestures by the **same speaker** will be more similar than gestures by different speakers.

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

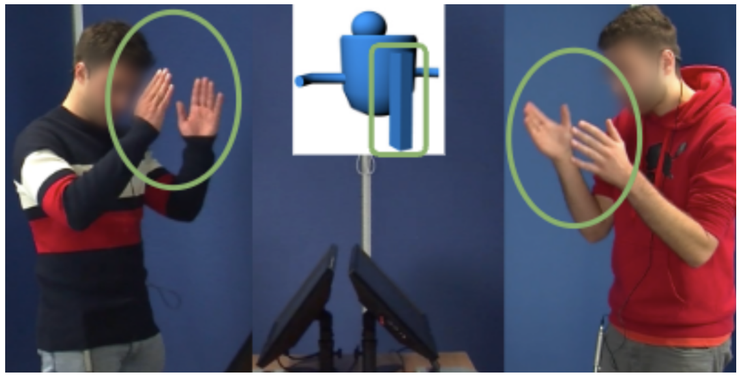


Hypothesis 2

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



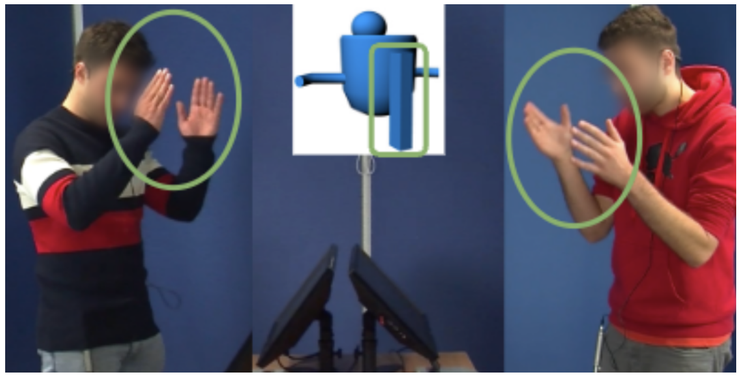
How aligned are the learned representations with theoretically-motivated hypotheses?



Hypothesis 3

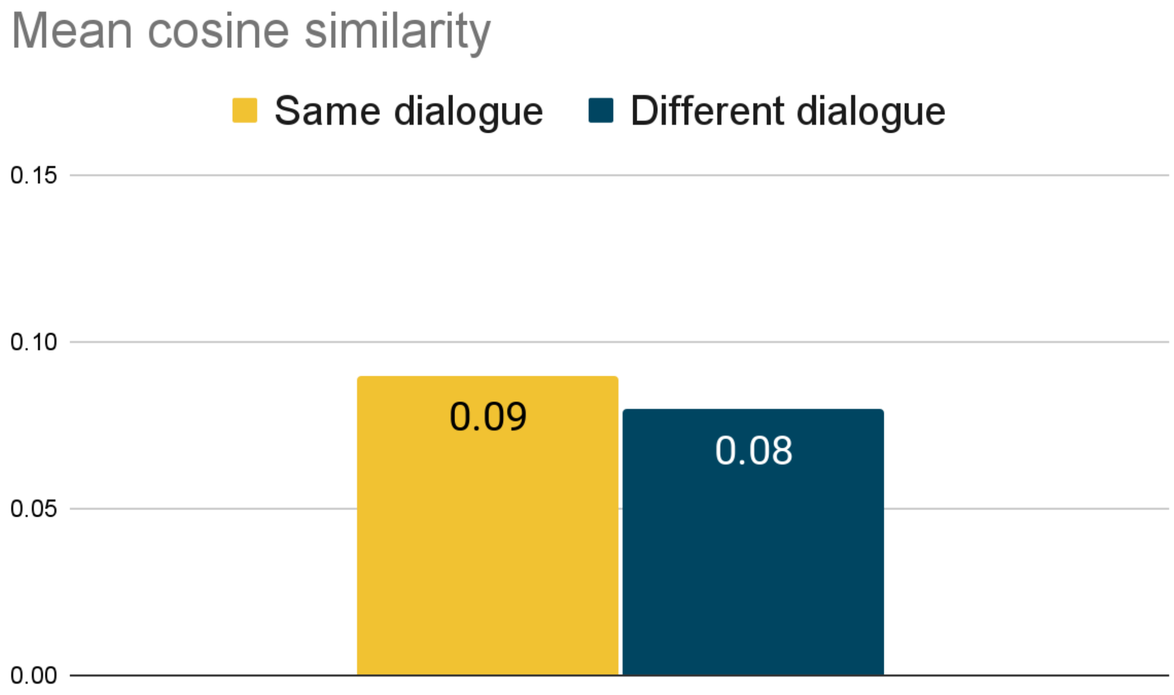
Given that participants **entrain through interaction**, same-referent gestures by two speakers **within a dialogue** will be more similar than from different dialogues.

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

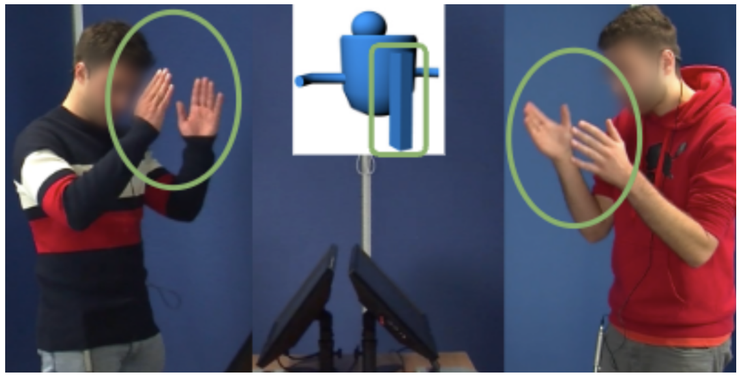


Hypothesis 3

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

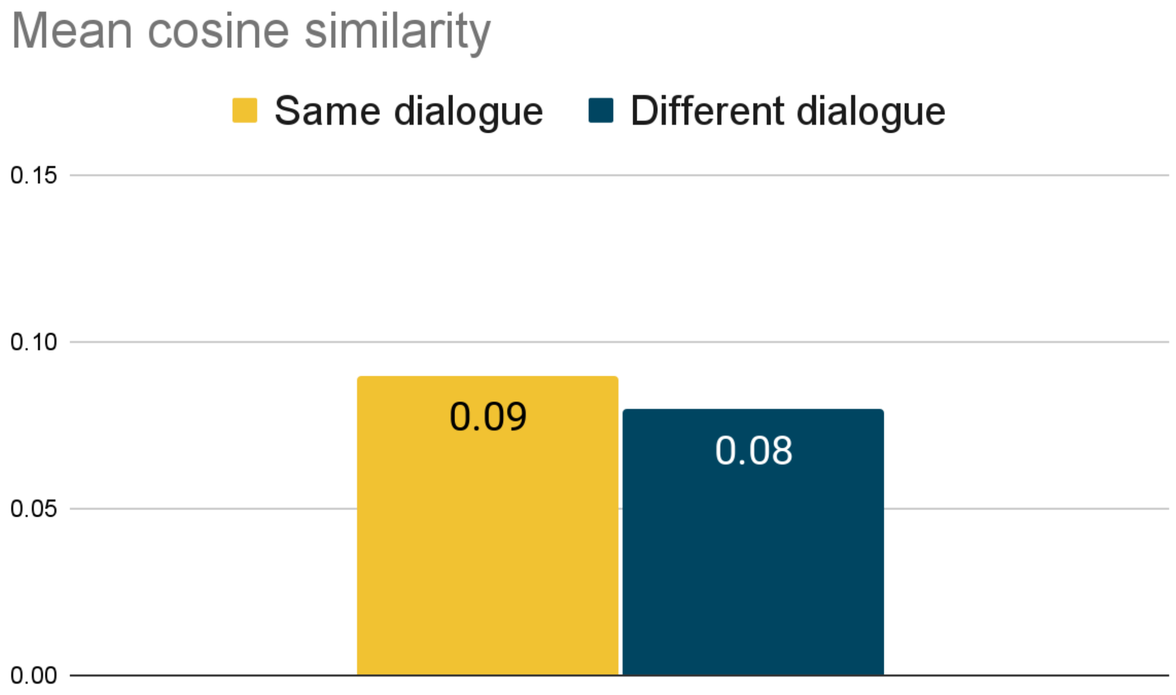


How aligned are the learned representations with theoretically-motivated hypotheses?



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.



Replication experiment with a corpus of computer-mediated interaction:

- same patterns with audio and video
- no gesture entrainment with only audio

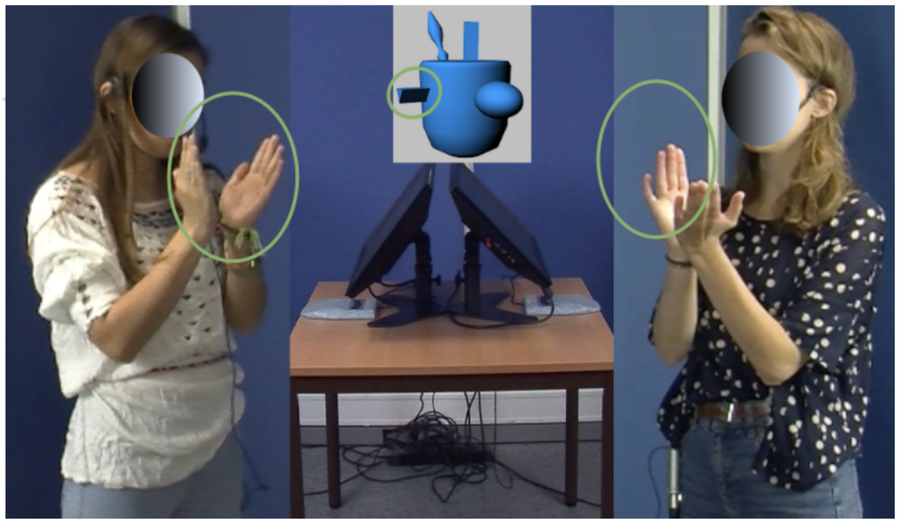
What properties of gestures are encoded in the learned embeddings?

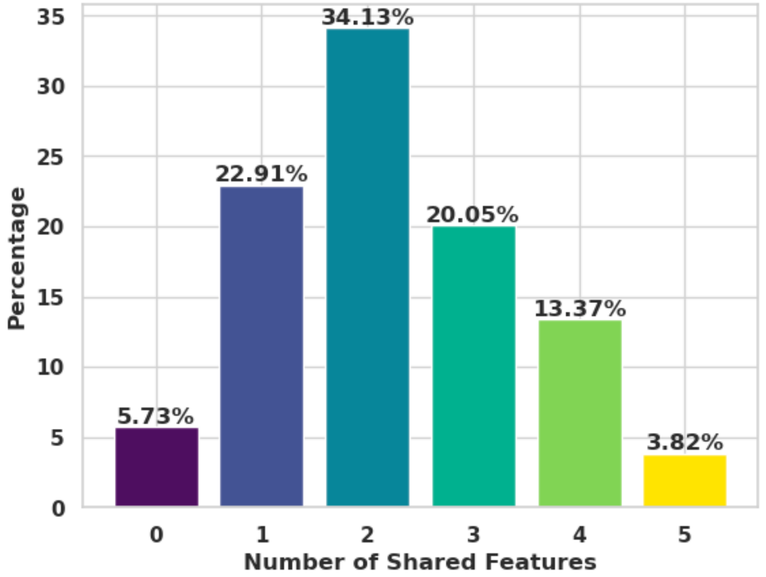
What properties of gestures are encoded in the learned embeddings?

Form features

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

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

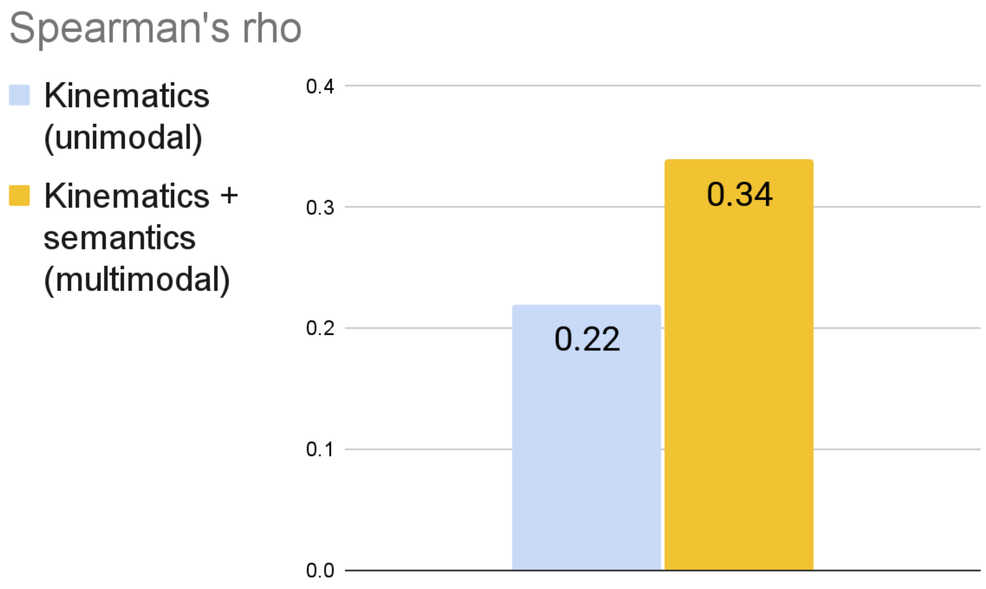




What properties of gestures are encoded in the learned embeddings?

Form features

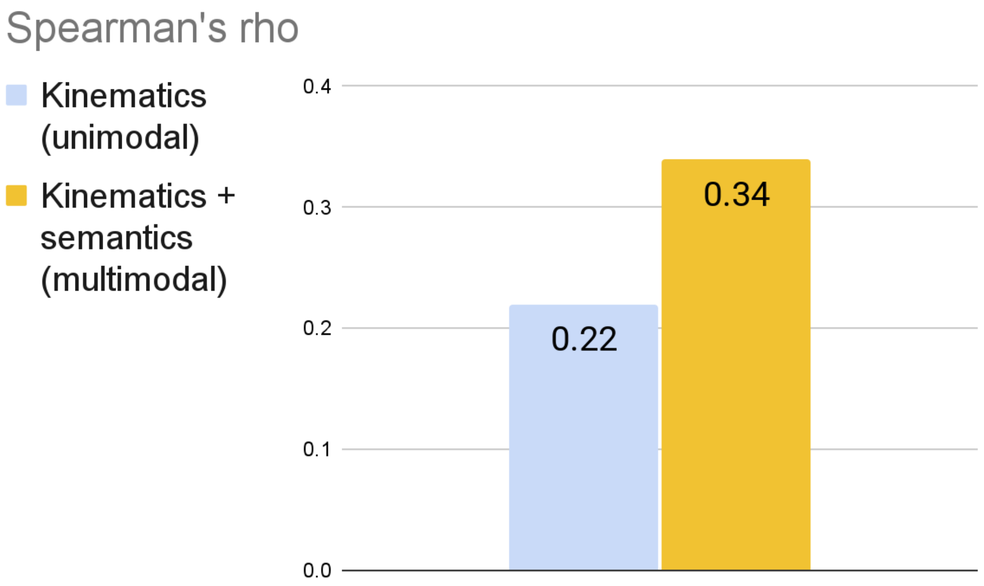
Significant positive **correlation** between number of manually coded form similarity features and cosine similarity of learned gesture embeddings.



What properties of gestures are encoded in the learned embeddings?

Form features

Significant positive **correlation** between number of manually coded form similarity features and cosine similarity of learned gesture embeddings.



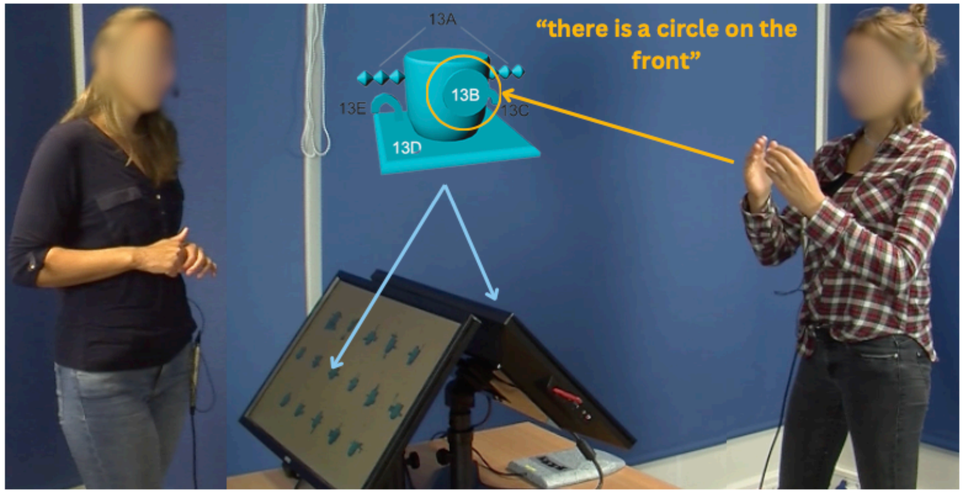
Linear **probing classifiers** for each feature (100 experiments, with random embeddings as baseline)

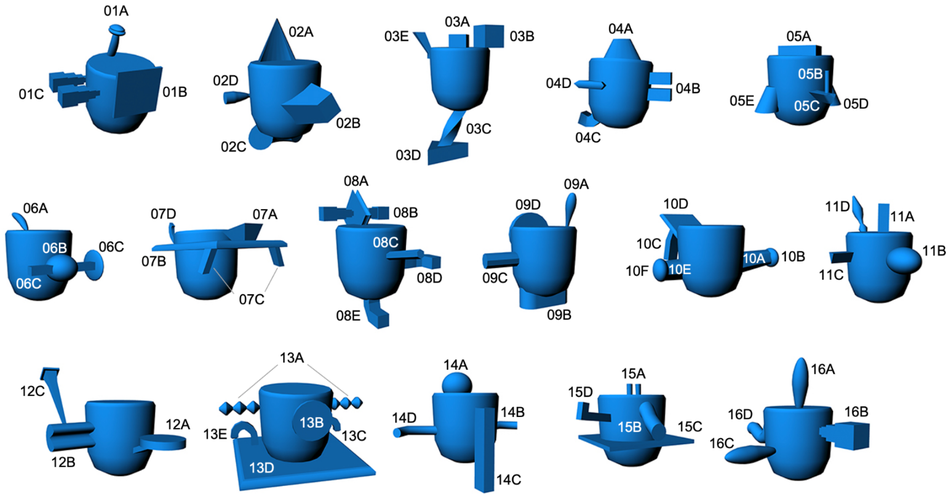
- All features can be recovered significantly above baseline.
- Similarity with respect to position and handedness are more accurately recovered from gesture embeddings learned multimodally.

What properties of gestures are encoded in the learned embeddings?

Referents

CABB-Small includes referent annotations per gesture

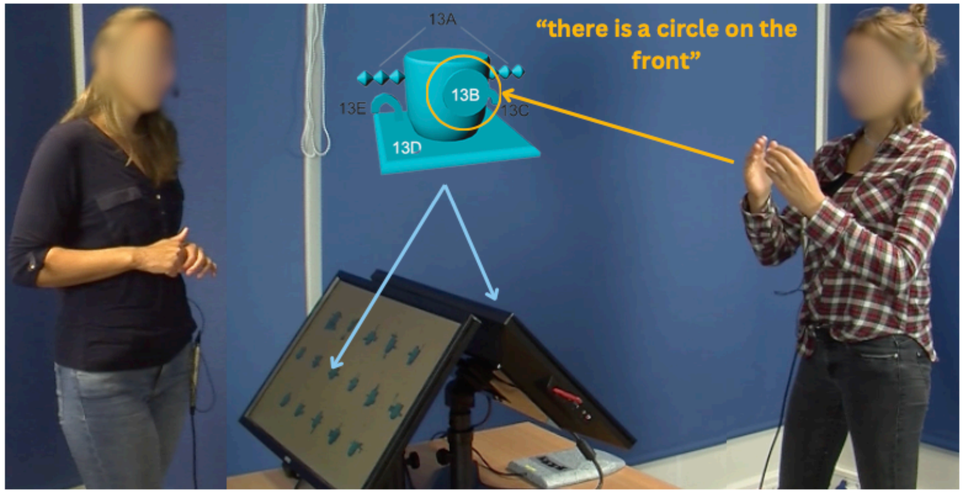


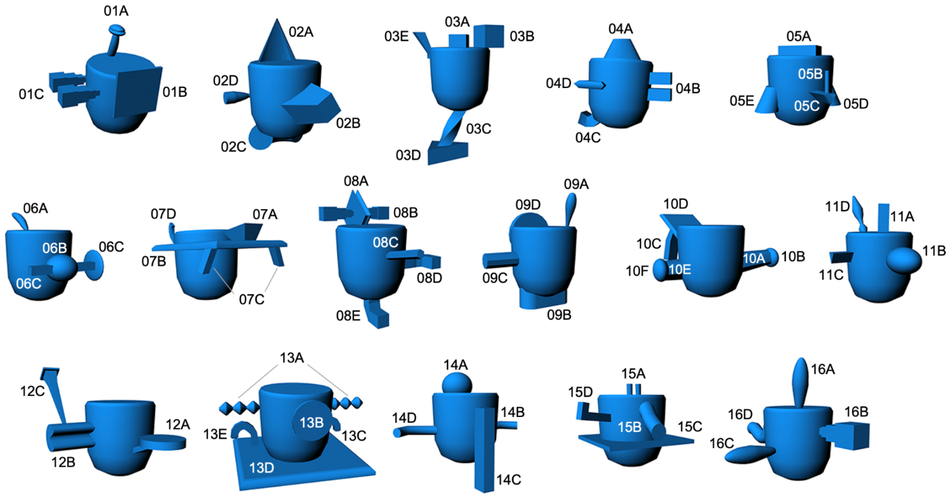


What properties of gestures are encoded in the learned embeddings?

Referents

CABB-Small includes referent annotations per gesture



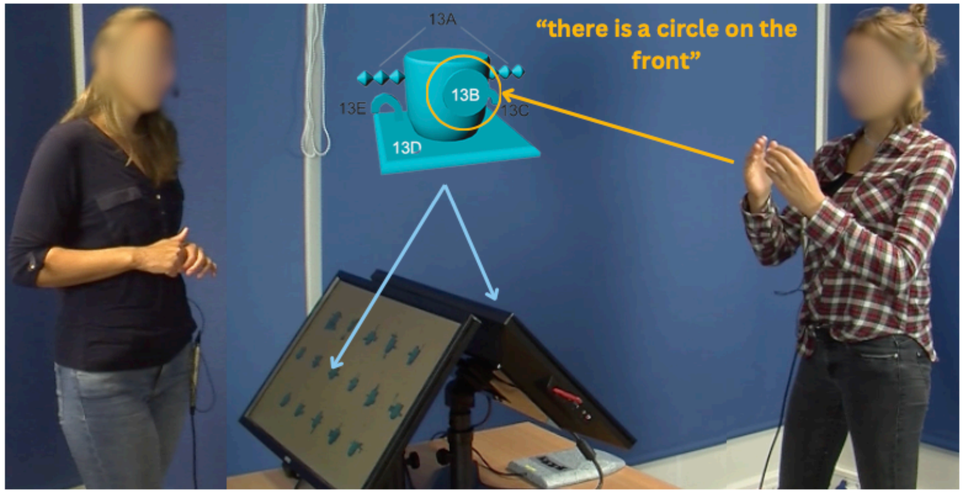


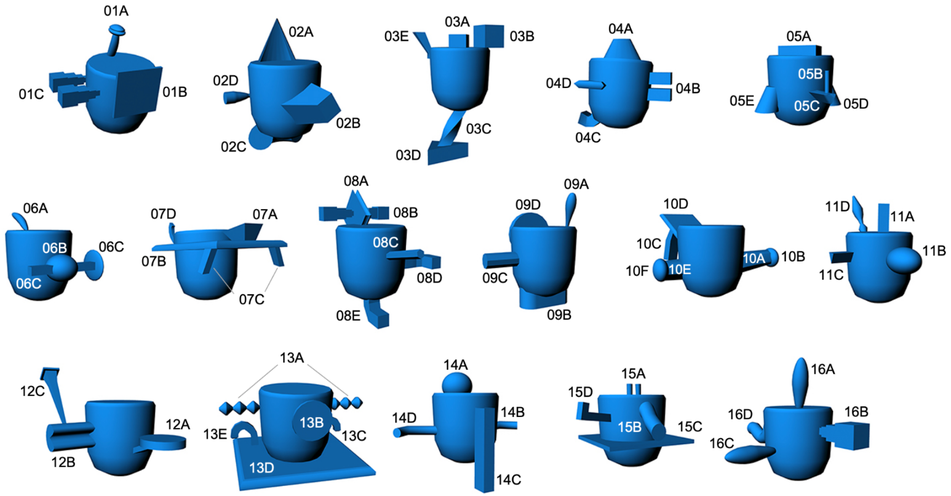
- MLP probing classifier trained to predict one referent among 70 possible object sub-parts.

What properties of gestures are encoded in the learned embeddings?

Referents

CABB-Small includes referent annotations per gesture

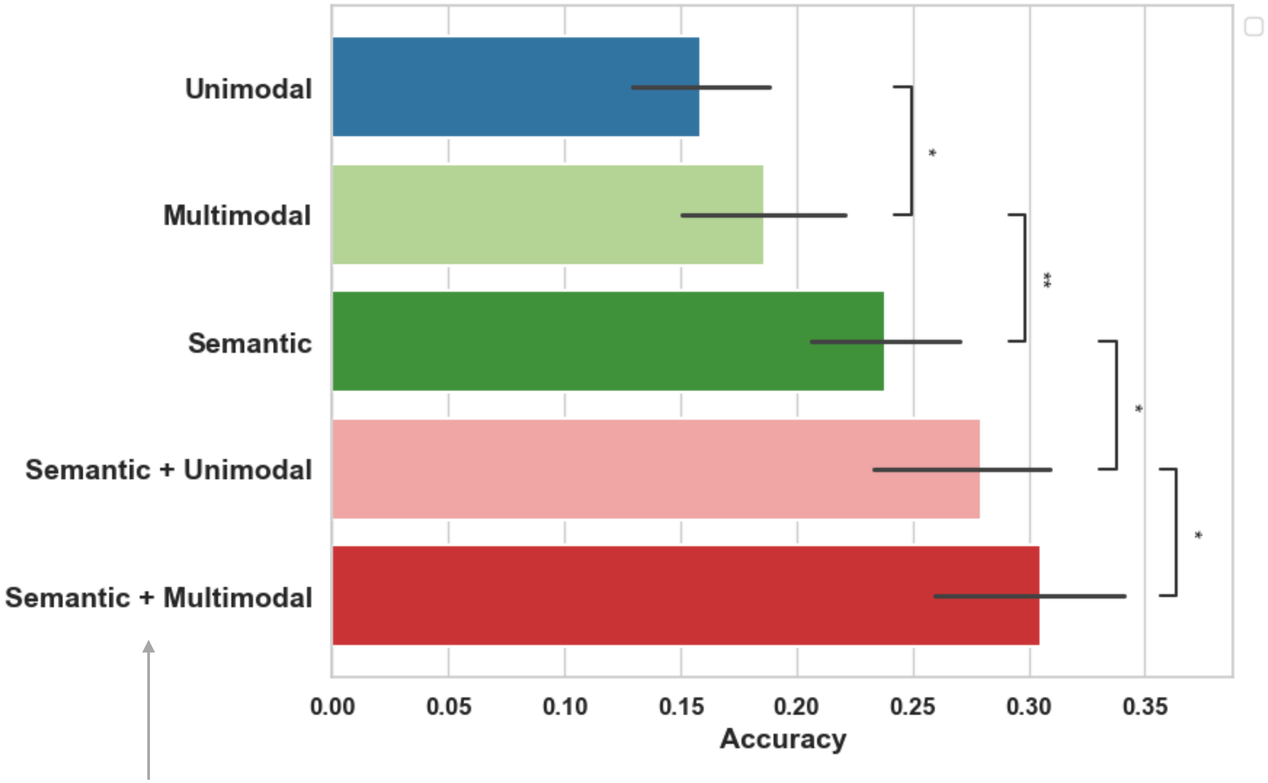




- MLP probing classifier trained to predict one referent among 70 possible object sub-parts.
- Chance accuracy < 2%; accuracy with random embeddings ~3\%.

What properties of gestures are encoded in the learned embeddings?

Referents



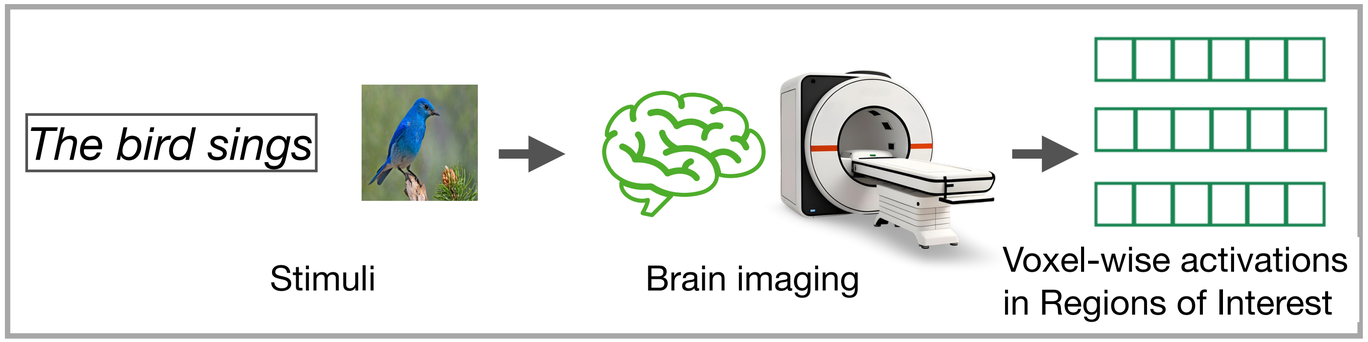
(just concatenation)

- Reference resolution accuracy significantly above baseline with unimodal and multimodal gesture embeddings.
- 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.)

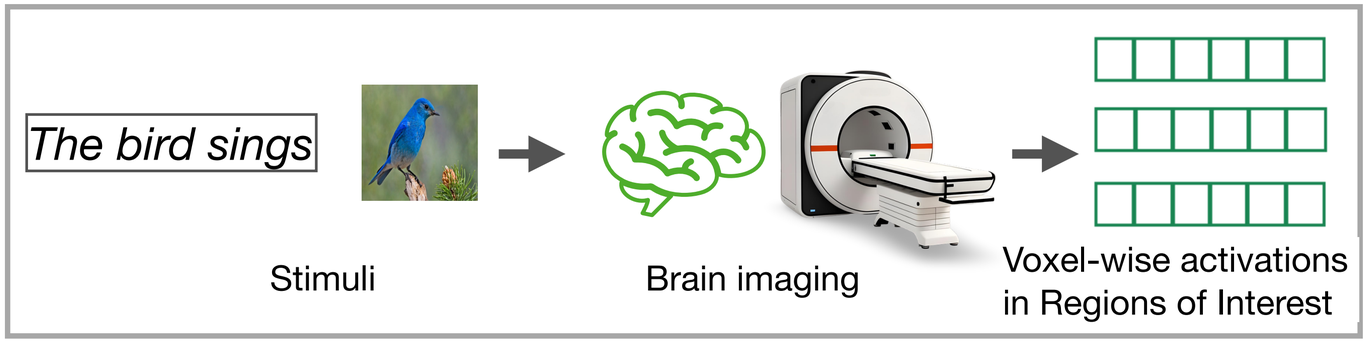
Interim summary

- A self-supervised learning approach aimed at capturing fundamental properties of **gestures from a multimodal perspective** (kinematics + vocal).
- Modelling gestures by grounding them in speech leads to embeddings that **comply with theoretical expectations**.
- Many open challenges:
  - Generalisation to other languages? other domains?
  - Continuous video, without segmented gestures as starting point?
  - Using raw images rather than skeleton joints?
  - Referent grounding?

Multimodality in the brain

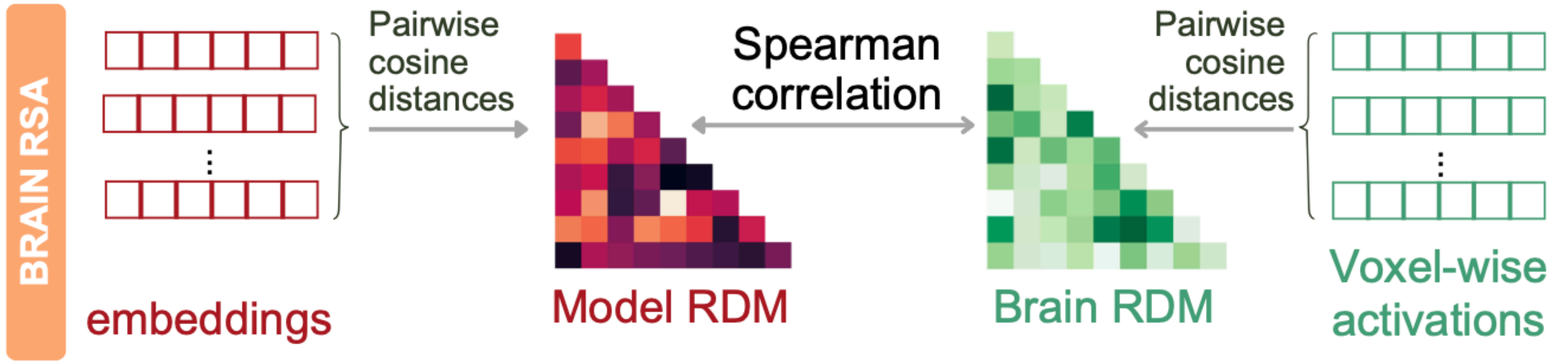


Given a stimulus, fMRI gives a voxel activation map that can be represented as a vector.

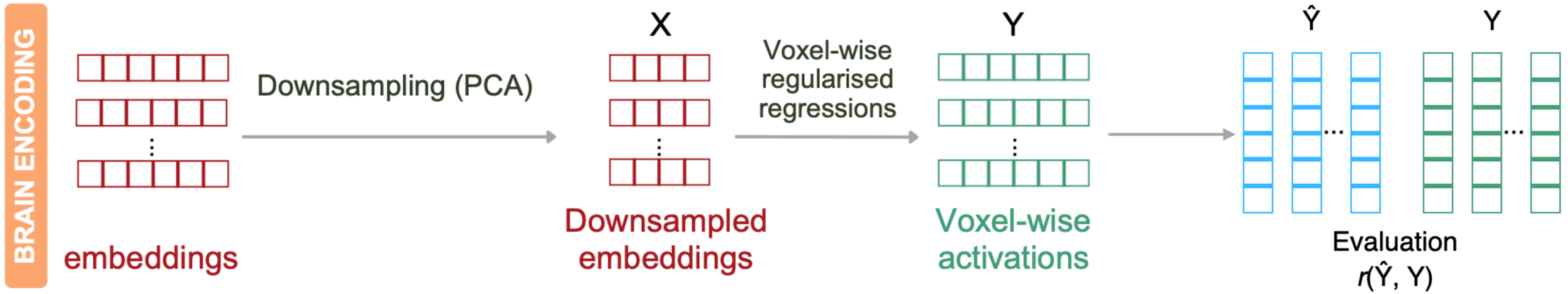


Given a stimulus, fMRI gives a voxel activation map that can be represented as a vector.

Representational Similarity Analysis



Encoding Models:



Do vision-language models approximate human concept understanding better than language-only models?

(CoNLL 2025)

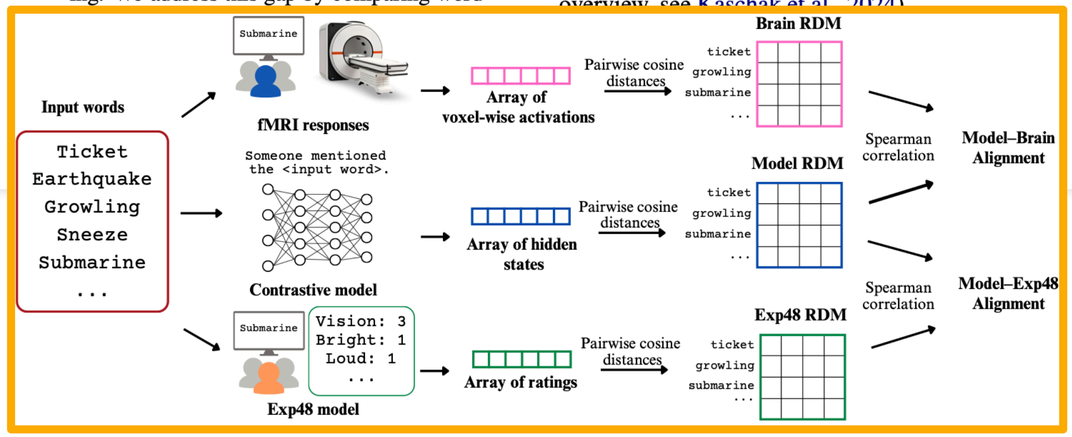
Experiential Semantic Information and Brain Alignment: Are Multimodal Models Better than Language Models?

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

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)



(EACL 2026)

Vision-Language Models Align with Human Neural Representations in Concept Processing

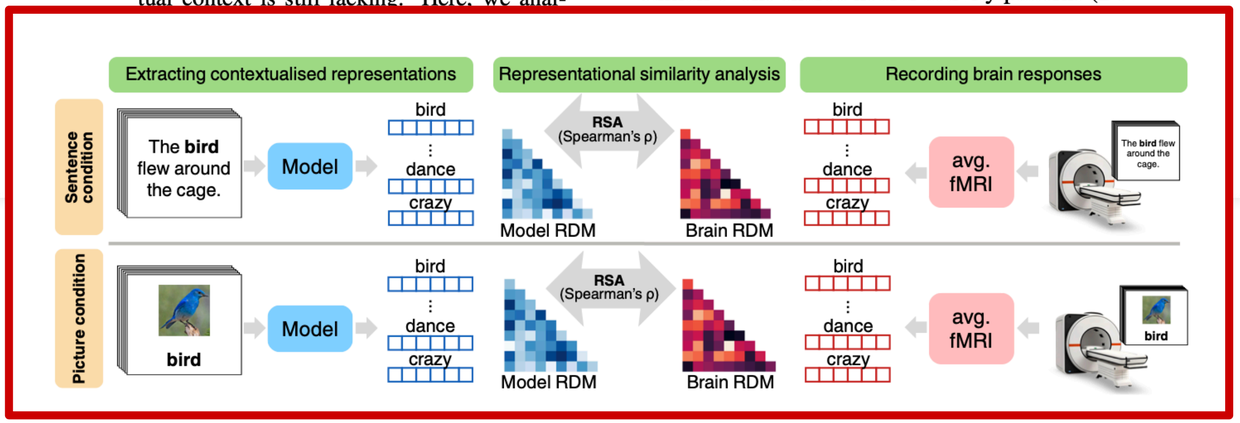
Anna Bavaresco, Marianne de Heer Kloots, Sandro Pezzelle, Raquel Fernández
Institute for Logic, Language and Computation
University of Amsterdam
{a.bavaresco, m.l.s.deheerkloots, s.pezzelle, raquel.fernandez}@uva.nl

Abstract

Recent studies suggest that transformer-based vision-language models (VLMs) capture the multimodality of concept processing in the human brain. However, a systematic evaluation exploring different types of VLM architectures and the role played by visual and textual context is still lacking. Here, we anal-

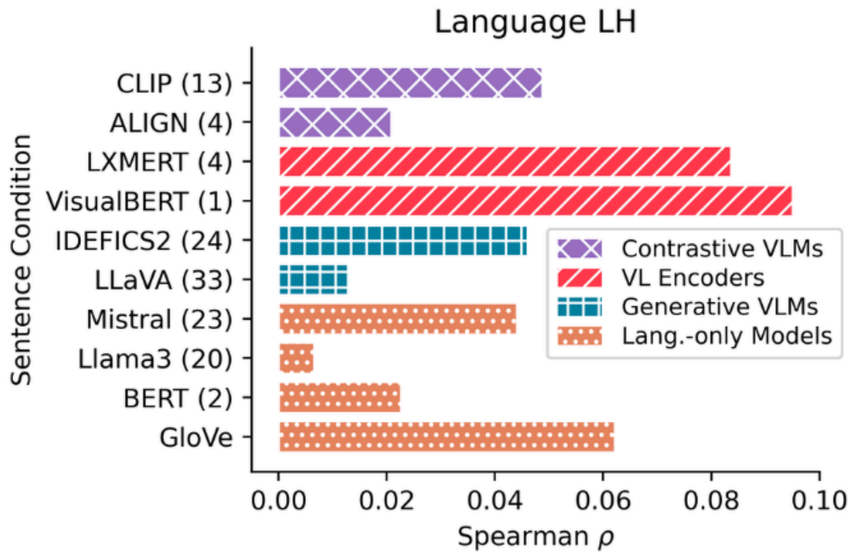
of human language modelling are not yet fully understood.

Insights from cognitive and neuroscientific work suggest that human semantic representations are deeply grounded in multimodal sensory experiences (Louwerse, 2011; Barsalou, 1999; Harnad, 1990; Bergen, 2012), and that visual and linguistic stimuli evoke shared neural activity patterns (Dev-



Do vision-language models approximate human concept understanding better than language-only models?

The **bird** flew around the cage.



(EACL 2026)

Vision-Language Models Align with Human Neural Representations in Concept Processing

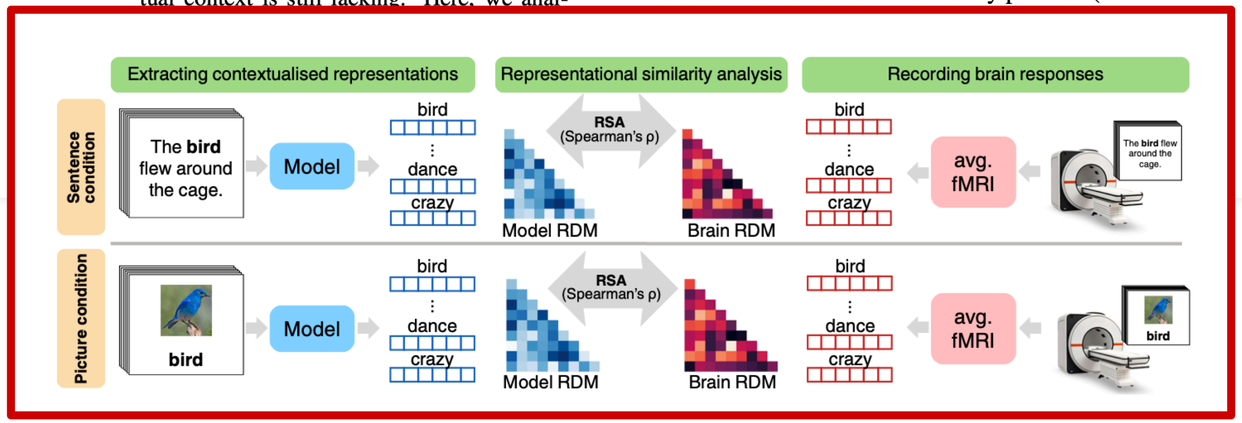
Anna Bavaresco, Marianne de Heer Kloots, Sandro Pezzelle, Raquel Fernández
Institute for Logic, Language and Computation
University of Amsterdam
{a.bavaresco, m.l.s.deheerkloots, s.pezzelle, raquel.fernandez}@uva.nl

Abstract

Recent studies suggest that transformer-based vision-language models (VLMs) capture the multimodality of concept processing in the human brain. However, a systematic evaluation exploring different types of VLM architectures and the role played by visual and textual context is still lacking. Here, we anal-

of human language modelling are not yet fully understood.

Insights from cognitive and neuroscientific work suggest that human semantic representations are deeply grounded in multimodal sensory experiences (Louwerse, 2011; Barsalou, 1999; Harnad, 1990; Bergen, 2012), and that visual and linguistic stimuli evoke shared neural activity patterns (Dev-



- **Previous studies:** mostly focus on modelling brain activation in language processing areas elicited by linguistic stimuli, using visually grounded model representations.

- **Previous studies:** mostly focus on modelling brain activation in language processing areas elicited by linguistic stimuli, using visually grounded model representations.

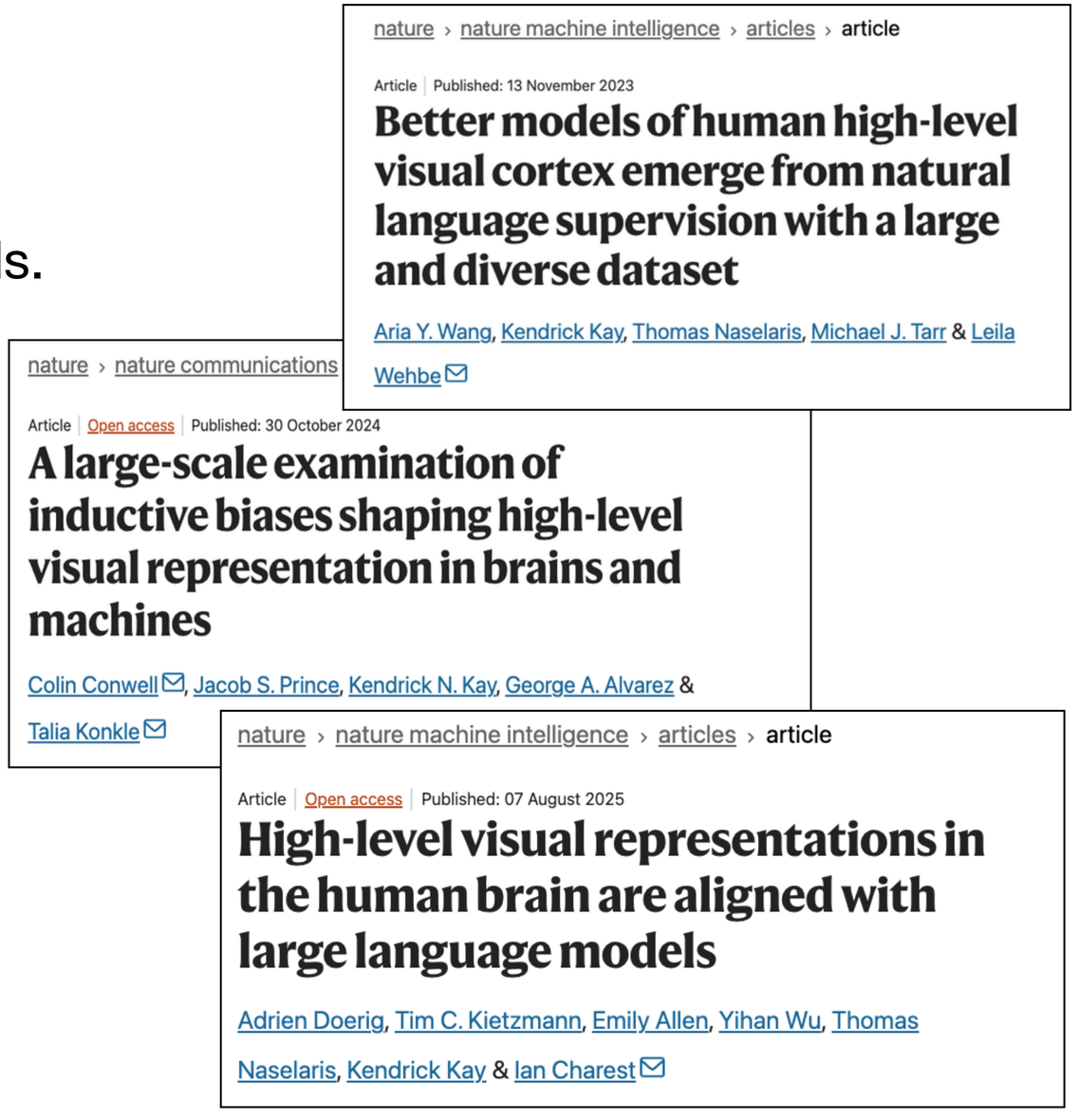
- **Next study:** focus on modelling brain activations in high-level vision processing areas elicited by visual stimuli, using language model representations.

Predicting high-level visual representations

Previous work has found that:

- Image features by VLMs (CLIP) are more brain-predictive than image features by vision-only models.
- Captions embedded by LLMs are as brain-predictive as image features by vision-only models.

Why?



Predicting high-level visual representations

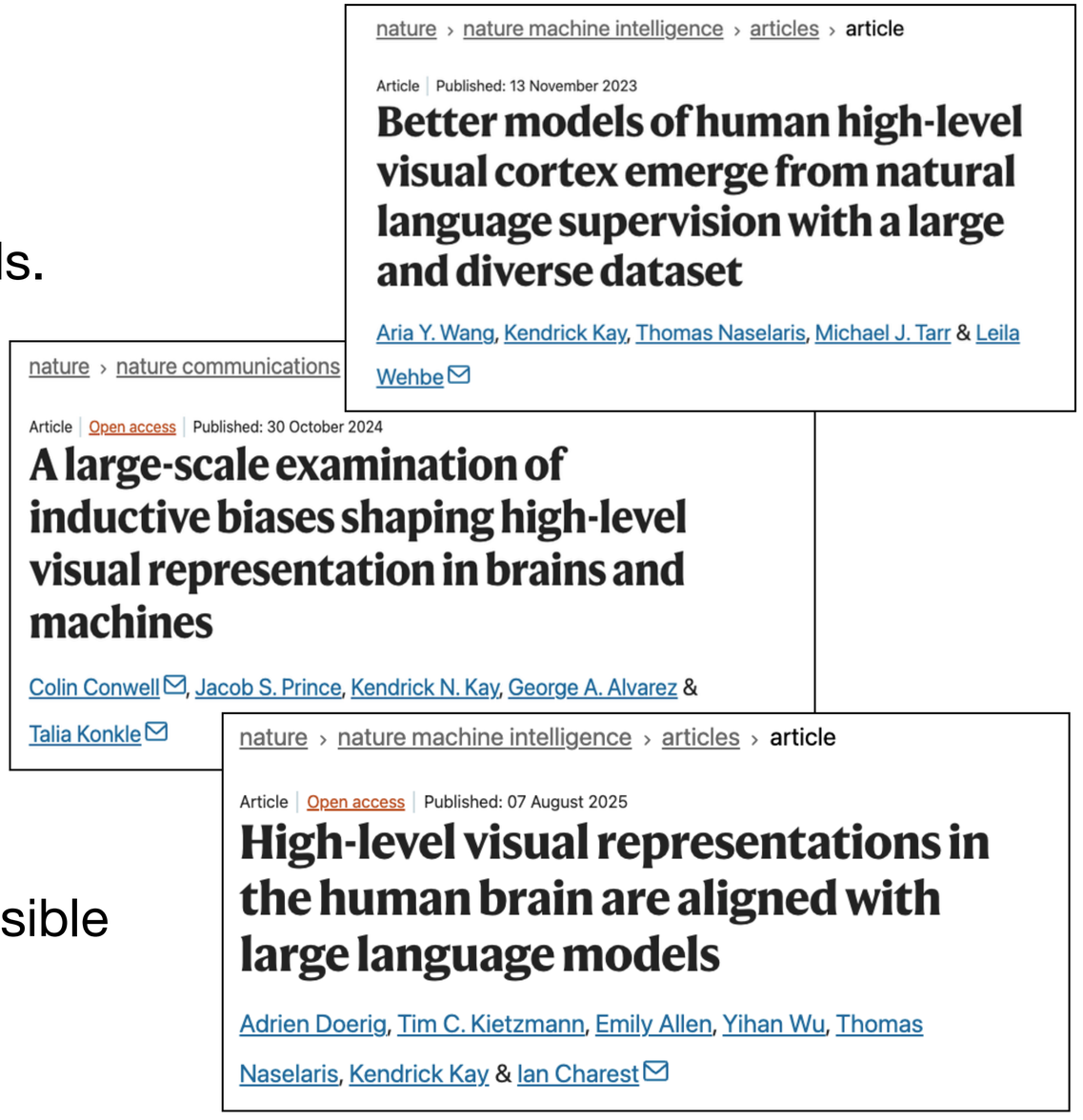
Previous work has found that:

- Image features by VLMs (CLIP) are more brain-predictive than image features by vision-only models.
- Captions embedded by LLMs are as brain-predictive as image features by vision-only models.

Why?

High-level visual areas selectively respond to semantically coherent categories (faces, bodies, place) — LLM representations capture semantics.

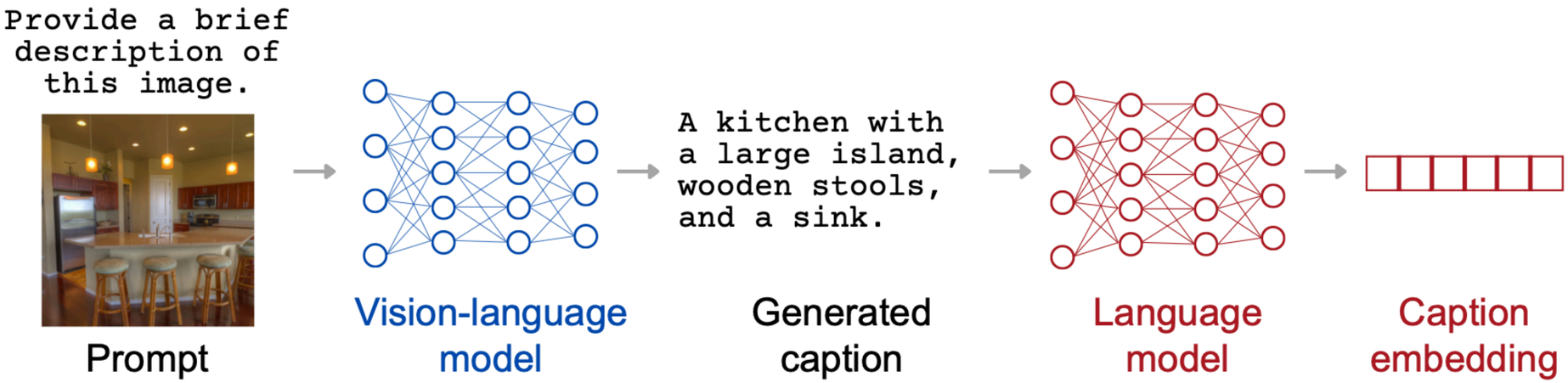
Yet, there is little understanding of the factors responsible for the brain productivity of image captions.



Caption embeddings

We evaluate the brain predictivity of:

- 6 different types of captions per image (only MS COCO captions in previous work), varying in source: human- vs. machine-generated (Phi-4, Qwen2.5-VL, LLaVA-OV, Pixtral, Molmo).
- 5 different language models to generate embeddings, including decoder-only, masked, and state-of-the-art sentence embedders (GPT-2, Llama3, BERT, Qwen3 and KaLM).



Caption sources: human & machine generated

COCO: A train coming down the tracks in the country.

LLaVA: A train is traveling down the tracks with a long line of cars behind it.

Phi-4: A train is traveling down the tracks in a rural area.

Pixtral: A train with multiple tanker cars travels through a rural landscape with hilly terrain in the background. The sky is overcast, and power lines run parallel to the train tracks.

Qwen2.5-VL: A freight train, painted in a distinctive yellow and black livery, travels through a rural landscape under a cloudy sky. The train consists of a locomotive pulling a series of cargo cars.

Molmo: A freight train with a black locomotive featuring a yellow front, pulling multiple blue tanker cars along rusty tracks. The train is set against a backdrop of green hills, trees, and a cloudy sky, with power lines visible overhead.

Caption type	*#words*	*PPL*	*Visualness*	LD_{nn}	LD_{adj}	LD_{vb}	*LD*
MS COCO	11 (2)	134 (130)	28 (7)	35 (9)	7 (8)	10 (7)	54 (10)
LLaVA	13 (3)	39 (23)	37 (8)	34 (8)	8 (8)	8 (5)	52 (7)
Phi-4	11 (2)	50 (32)	29 (6)	34 (7)	6 (7)	8 (6)	48 (7)
Pixtral	31 (5)	15 (6)	74 (12)	30 (5)	12 (6)	11 (4)	55 (6)
Qwen2.5-VL	28 (7)	19 (9)	70 (16)	33 (6)	12 (6)	11 (4)	58 (6)
Molmo	37 (8)	13 (5)	93 (18)	31 (5)	16 (5)	10 (3)	57 (5)

- Perplexity computed with Ministral3.
- Visualness: sum of visual ratings from the Lancaster sensorimotor norms.
- Lexical density as % of content words.

Brain encoding

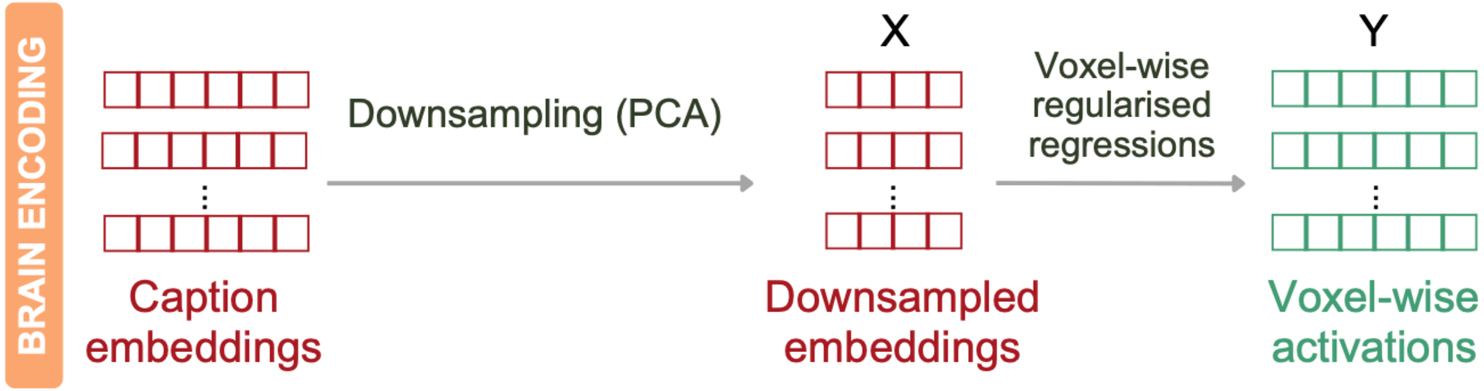
Natural Scene Dataset (NSD)

- **Stimuli:**
  - 1000 images from MS COCO, centre-cropped.
- **fMRI experiment:**
  - 8 participants; high-resolution fMRI scanner (7T);
  - continuous recognition task; ~3 presentations of each stimulus, visible for 3s;

Brain encoding

Natural Scene Dataset (NSD)

- **Stimuli:**
  - 1000 images from MS COCO, centre-cropped.
- **fMRI experiment:**
  - 8 participants; high-resolution fMRI scanner (7T);
  - continuous recognition task; ~3 presentations of each stimulus, visible for 3s;
  - We consider 3 high-level visual regions of interest (functionally localised ROIs), selective for Faces, Bodies, and Places.

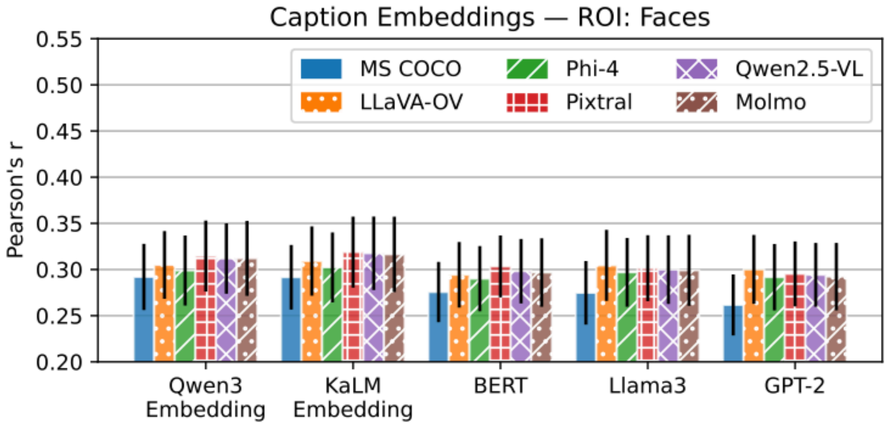


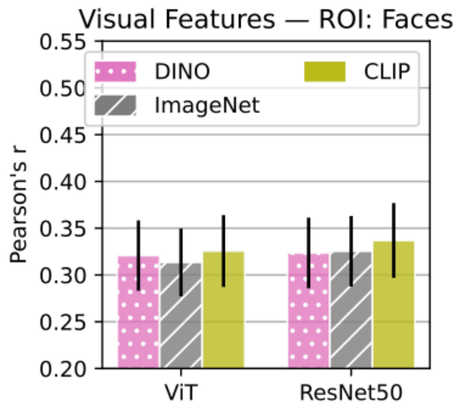
Encoding results

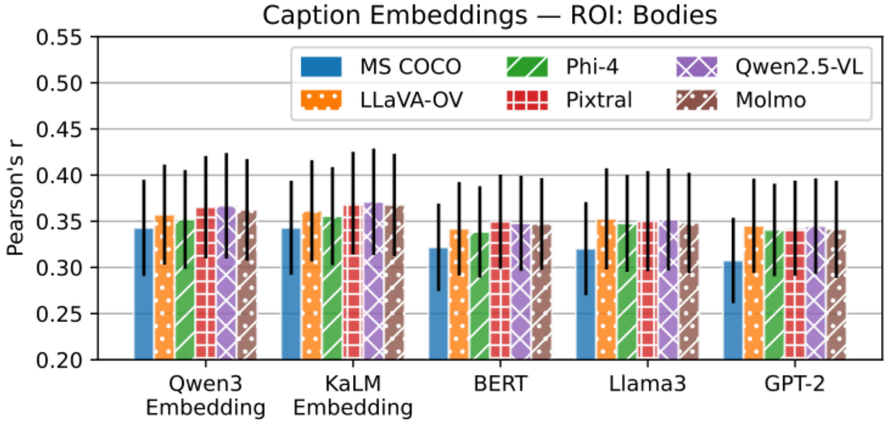
Pearson’s correlation between predicted and observed neural activations, with the most predictive model layer.

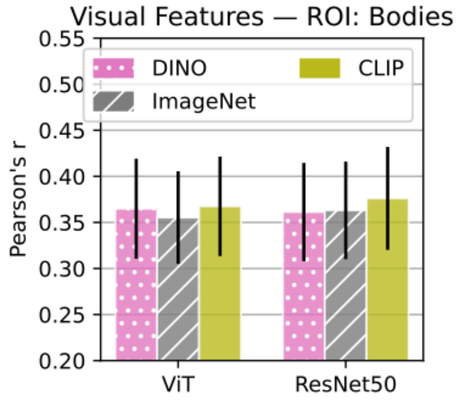
Average over voxels with significant predictions; error bars show sd across participants.

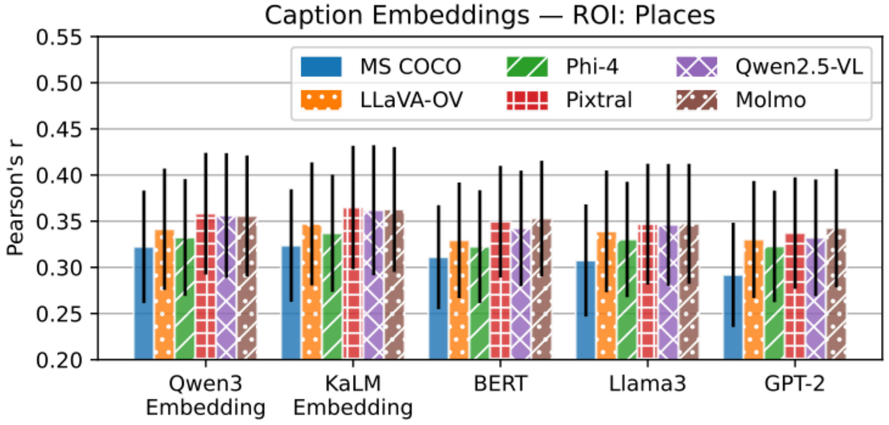
- Captions significantly predict brain activity in visual processing areas, to a similar extent as image features.

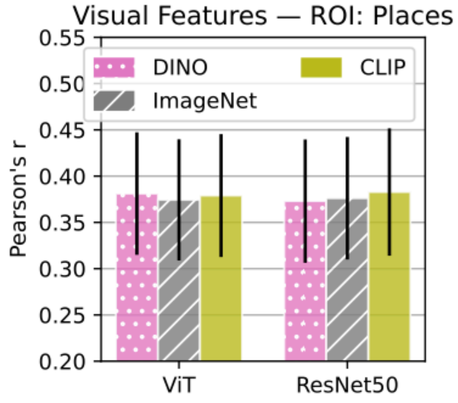








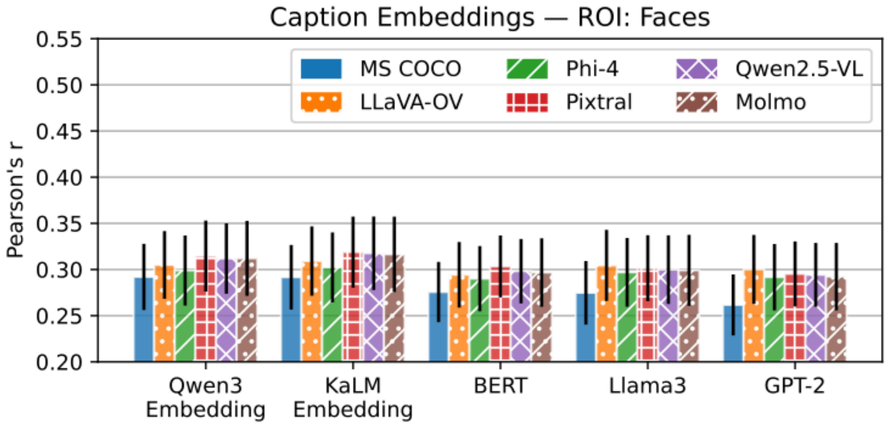


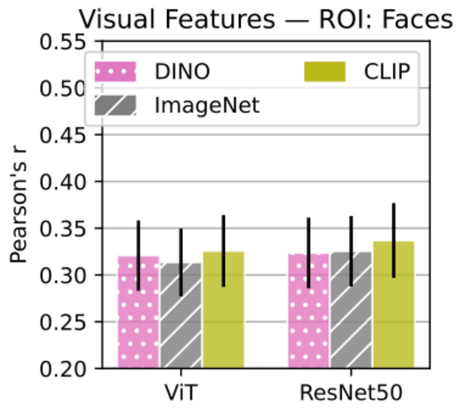


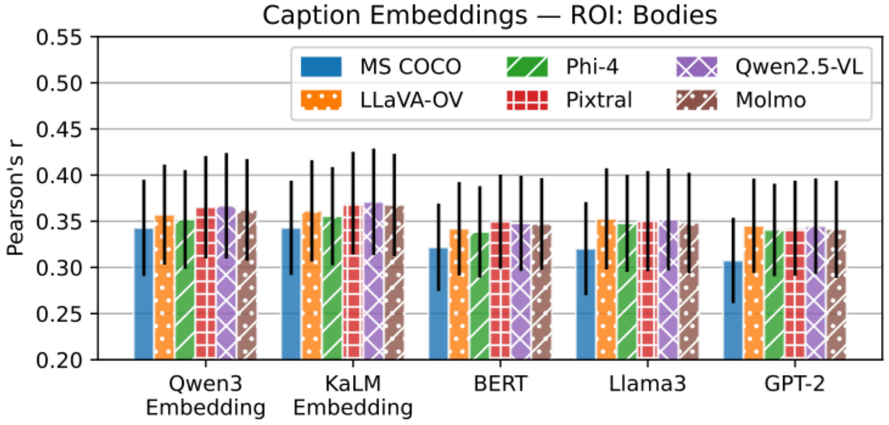
Encoding results

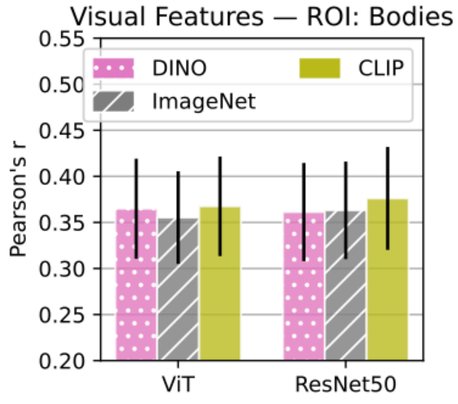
Linear mixed-effects models with caption source and embedding LM as fixed effects show that:

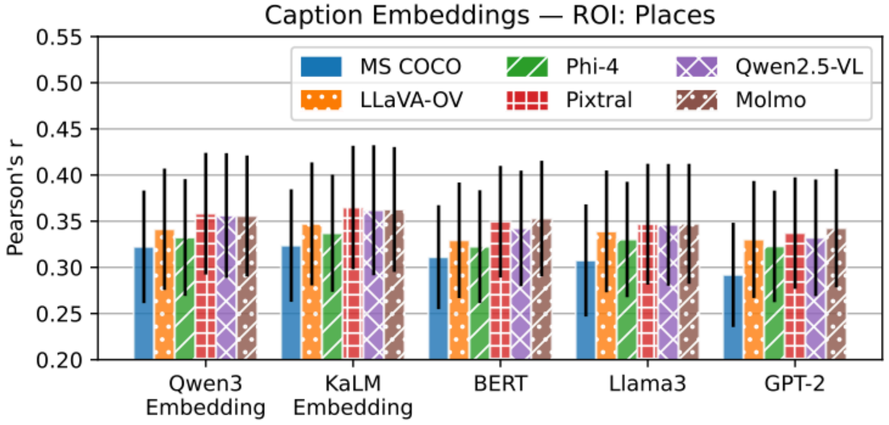
- MS COCO captions are the least predictive.
- Different machine-generated caption types perform best for different ROIs: captions with higher visualness and lexical density.
- In all cases, recent text embedder (Qwen3 and KaLM) have an advantage over earlier masked and autoregressive LMs.

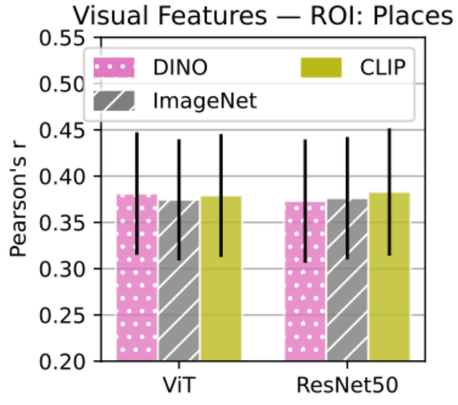












Layer-wise analysis

What information is responsible for brain predictivity?

LMs transform the input into useful information by compressing it into a few dimensions (intrinsic dimensionality).

A peak in high intrinsic dimensionality signals abstract linguistic information: fair performance in semantic/syntactic tasks and downstream task transfer (Cheng et al. 2025).

Cheng et al. Emergence of a high-dimensional abstraction phase in language transformers. ICLR 2025.

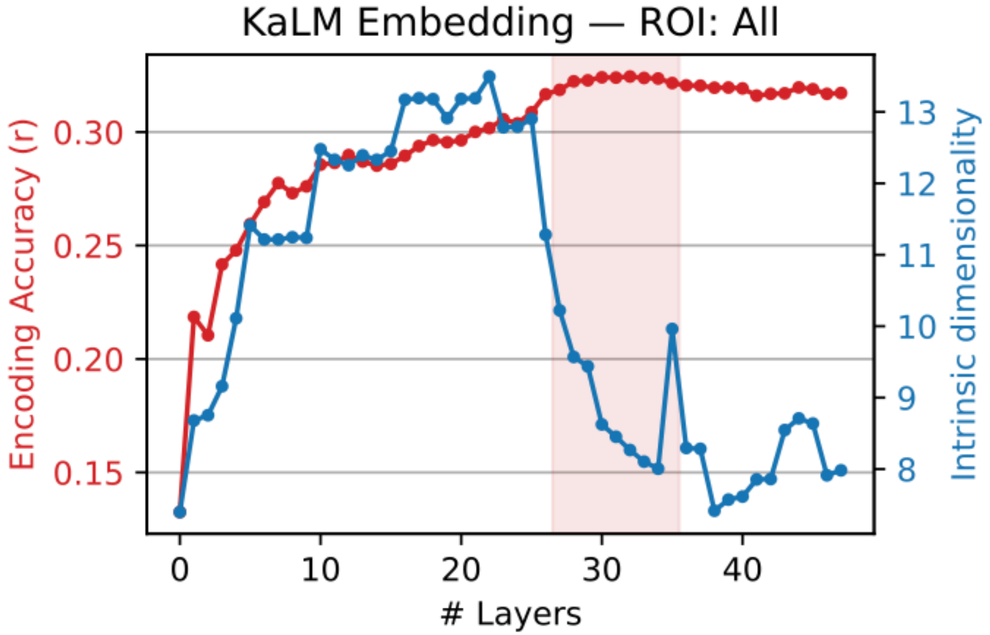
Layer-wise analysis

What information is responsible for brain predictivity?

LMs transform the input into useful information by compressing it into a few dimensions (intrinsic dimensionality).

A peak in high intrinsic dimensionality signals abstract linguistic information: fair performance in semantic/syntactic tasks and downstream task transfer (Cheng et al. 2025).

Brain productivity peaks right after a phase of high intrinsic dimensionality.



Estimated ID per layer. The original dimensionality of KaLM Embeddings is 3840.

Cheng et al. Emergence of a high-dimensional abstraction phase in language transformers. ICLR 2025.

Interim summary

- LM embeddings (of machine-generated captions) significantly predict brain activity in high-level visual brain areas.
- The LM used to embed the caption matters, more so than surface level differences in caption types.
- Brain encoding peaks after a phase of high intrinsic dimensionality.
- More experiments and results in paper

Wrapping Up

Interest in investigating questions related to multimodal cognition with computational methods

- *To what extent does language interface with perception?*
- *How can we investigate this empirically, using computational methods?*
- *Are deep learning models suitable for this aim, and what insights can they bring?*

- Co-speech gestures in face-to-face dialogue
- Multimodality in the brain

Wrapping Up

Interest in investigating questions related to multimodal cognition with computational methods

- *To what extent does language interface with perception?*
- *How can we investigate this empirically, using computational methods?*
- *Are deep learning models suitable for this aim, and what insights can they bring?*

- Co-speech gestures in face-to-face dialogue
- Multimodality in the brain

Current machine learning methods now make it possible to study these questions from a data-driven computational perspective at a scale never seen before.

An additional tool complementing experimental and theoretical research.

References






- Esam Ghaleb, Bulat Khaertdinov, Marlou Rasenberg, Wim Pouw, Judith Holler, Aslı Özyürek, & Raquel Fernández. Learning Co-Speech Gesture Representations in Dialogue through Contrastive Learning: An Intrinsic Evaluation. In *Proceedings of the 26th ACM International Conference on Multimodal Interaction*: ICMI 2024.

- Esam Ghaleb, Bulat Khaertdinov, Aslı Özyürek, & Raquel Fernández. I see what you mean: Co-Speech Gestures for Reference Resolution in Multimodal Dialogue. In *Findings of the Association for Computational Linguistics:* ACL 2025.

- Anna Bavaresco & Raquel Fernández. Experiential Semantic Information and Brain Alignment. Are Multimodal Models Better than Language Models? In *Proceedings of the 29th Conference on Computational Language Learning:* CoNLL 2025.

- Anna Bavaresco, Marianne Heer Kloots, Sandro Pezzelle, & Raquel Fernández. Vision-Language Models Align with Human Neural Representations in Concept Processing. In *Proceedings of the The 19th Conference of the European Chapter of the Association for Computational Linguistics:* EACL 2026.

- Anna Bavaresco, Ina Klaric, Raquel Fernández, & Marie-Francine Moens. What Makes Linguistic Representations Good Models of High-Level Visual Perception in the Human Brain?, 2026 [*under journal submission*].





Thanks!