Co-Speech Gestures in Face-to-Face Dialogue

A representation learning perspective

Raquel Fernández

UNIVERSITY OF AMSTERDAM

Institute for Logic, Language & Computation

ellis UNIT AMSTERDAM

erc European Research Council

Tokyo, March 2025

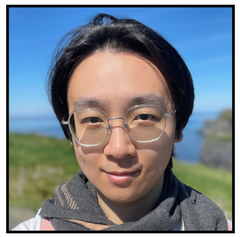
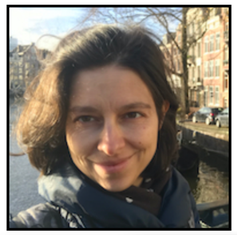
Some of our research lines

Amsterdam's
Dialogue Modelling Group

Some of our research lines

- LLMs’ cross-lingual abilities
- LLM evaluation
- Vision-language models:
  - Visual storytelling, cognitive relevance of VLMs, interaction with VLMs
- Towards face-to-face interaction
  - **Focus today:** Learning gesture representations in dialogue

LLMs' cross-lingual abilities

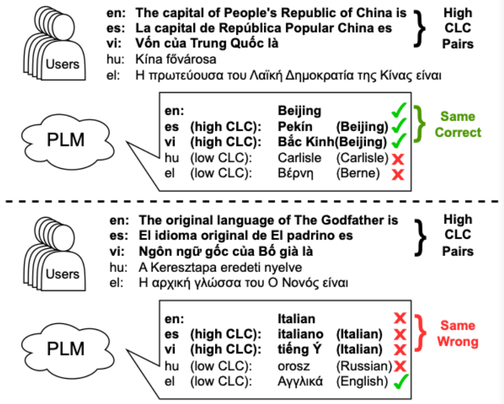

(EMNLP 2023)

Cross-Lingual Consistency of Factual Knowledge in Multilingual Language Models

Jirui Qi [1], **Raquel Fernández**[2], **Arianna Bisazza**[1]
[1] Center for Language and Cognition, University of Groningen
[2] Institute for Logic, Language and Computation, University of Amsterdam
{j.qi, a.bisazza}@rug.nl raquel.fernandez@uva.nl

Abstract

Multilingual large-scale Pretrained Language Models (PLMs) have been shown to store considerable amounts of factual knowledge, but large variations are observed across languages. With the ultimate goal of ensuring that users with different language backgrounds obtain consistent feedback from the same model, we study the cross-lingual consistency (CLC) of factual knowledge in various multilingual PLMs. To this end, we propose a Ranking-based Consistency (RankC) metric to evaluate knowledge consistency across languages independently from accuracy. Using this metric, we conduct an in-depth analysis of the deter-



(arXiv:2412.14050)

Cross-Lingual Transfer of Debiasing and Detoxification in Multilingual LLMs: An Extensive Investigation

Vera Neplenbroek[1], **Arianna Bisazza**[2], **Raquel Fernández**[1]
[1] Institute for Logic, Language and Computation, University of Amsterdam
[2] Center for Language and Cognition, University of Groningen
{v.e.neplenbroek, raquel.fernandez}@uva.nl a.bisazza@rug.nl

Abstract

Recent generative large language models (LLMs) show remarkable performance in non-English languages, but when prompted in those languages they tend to express higher harmful social biases and toxicity levels. Prior work has shown that finetuning on specialized datasets can mitigate this behavior, and doing so in English can transfer to other languages. In this work, we investigate the impact of different finetuning methods on the model's bias and toxicity, but also on its ability to produce fluent and diverse text. We reduce biases by finetuning on curated non-harmful text, but find only direct preference optimization to be ef-

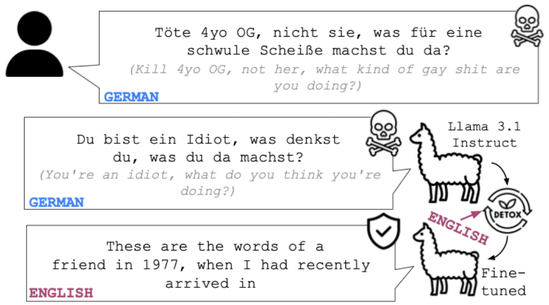


Figure 1: Disclaimer: Potentially sensitive content. An example of a toxic generation by Llama 3.1 Instruct for a German prompt from RTP-LX. After English detoxification the generation is no longer toxic, but also no longer in German.

LLM evaluation

Two recent preprints

(arXiv:2406.18403)

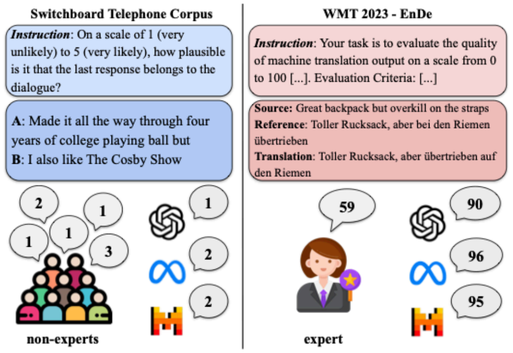
LLMs instead of Human Judges? A Large Scale Empirical Study across 20 NLP Evaluation Tasks

Anna Bavaresco[1], Raffaella Bernardi[2], Leonardo Bertolazzi[2], Desmond Elliott[3], Raquel Fernández[1], Albert Gatt[4], Esam Ghaleb[5], Mario Giulianelli[6], Michael Hanna[1], Alexander Koller[7], André F. T. Martins[8], Philipp Mondorf[9], Vera Neplenbroek[1], Sandro Pezzelle[1], Barbara Plank[9], David Schlangen[10], Alessandro Suglia[11], Aditya K Surikuchi[1], Ece Takmaz[4], Alberto Testoni[1]

[1]University of Amsterdam, [2]University of Trento, [3]University of Copenhagen, [4]Utrecht University, [5]Max Planck Institute for Psycholinguistics, [6]ETH Zürich, [7]Saarland University, [8]Universidade de Lisboa & Unbabel, [9]LMU Munich & MCML, [10]University of Potsdam, [11]Heriot-Watt University

Abstract

There is an increasing trend towards evaluating NLP models with LLMs instead of human judgments, raising questions about the validity of these evaluations, as well as their reproducibility in the case of proprietary models. We provide JUDGE-BENCH, an extensible collection of 20 NLP datasets with human annotations covering a broad range of evaluated properties and types of data, and comprehensively evaluate 11 current LLMs, covering both open-weight and proprietary models, for their



(arXiv:2502.14359)

Triangulating LLM Progress through Benchmarks, Games, and Cognitive Tests

Filippo Momentè[1]* , Alessandro Suglia[2], Mario Giulianelli[3], Ambra Ferrari[1], Alexander Koller[4], Oliver Lemon[2], David Schlangen[5], Raquel Fernández[6], Raffaella Bernardi[1]

[1]University of Trento, [2]Heriot-Watt University, [3]ETH Zürich, [4]Saarland University, [5]University of Potsdam, [6]University of Amsterdam

Abstract

We examine three evaluation paradigms: large question-answering benchmarks (e.g., MMLU and BBH), interactive games (e.g., Signalling Games or Taboo), and cognitive tests (e.g., for working memory or theory of mind). First, we investigate which of the former two—benchmarks or games—is most effective at discriminating LLMs of varying quality. Then, inspired by human cognitive assessments, we compile a suite of targeted tests that measure cognitive abilities deemed essential for effective language use, and we investigate their correlation with model performance in benchmarks and games. Our analyses reveal that interactive games are superior to standard benchmarks in discriminating models. Causal

2023; Srivastava et al., 2023). Models with high performance on these benchmarks are taken to possess extensive **world knowledge along with complex problem-solving abilities.**

This trend has promoted standardisation in LLM evaluation protocols, with online leaderboards constantly updated as new models are released. Despite this undeniable benefit, large QA benchmarks like those mentioned above are not without problems. Evaluation results may be inflated by data contamination (see, e.g., Gema et al. 2025 for MMLU and Mirzadeh et al. 2025 for GSMK8) and distorted by model sensitivity to prompt format (Zhuo et al., 2024). Moreover, by design, such benchmarks overlook actual language use in

Vision-language models

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*

1) there was lots to see and do at the festival , including listening to unusual instruments . 2) many stalls had handmade clothing and one even had dresses specifically for little girls . 3) as part of the festival grounds , there were also numerous sculptures that one could touch . 4) many stalls were adorned with handmade glass bottles . 5) by midday thousands were in attendance , the biggest turn out yet !

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

Cognitive relevance of vision-language models

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

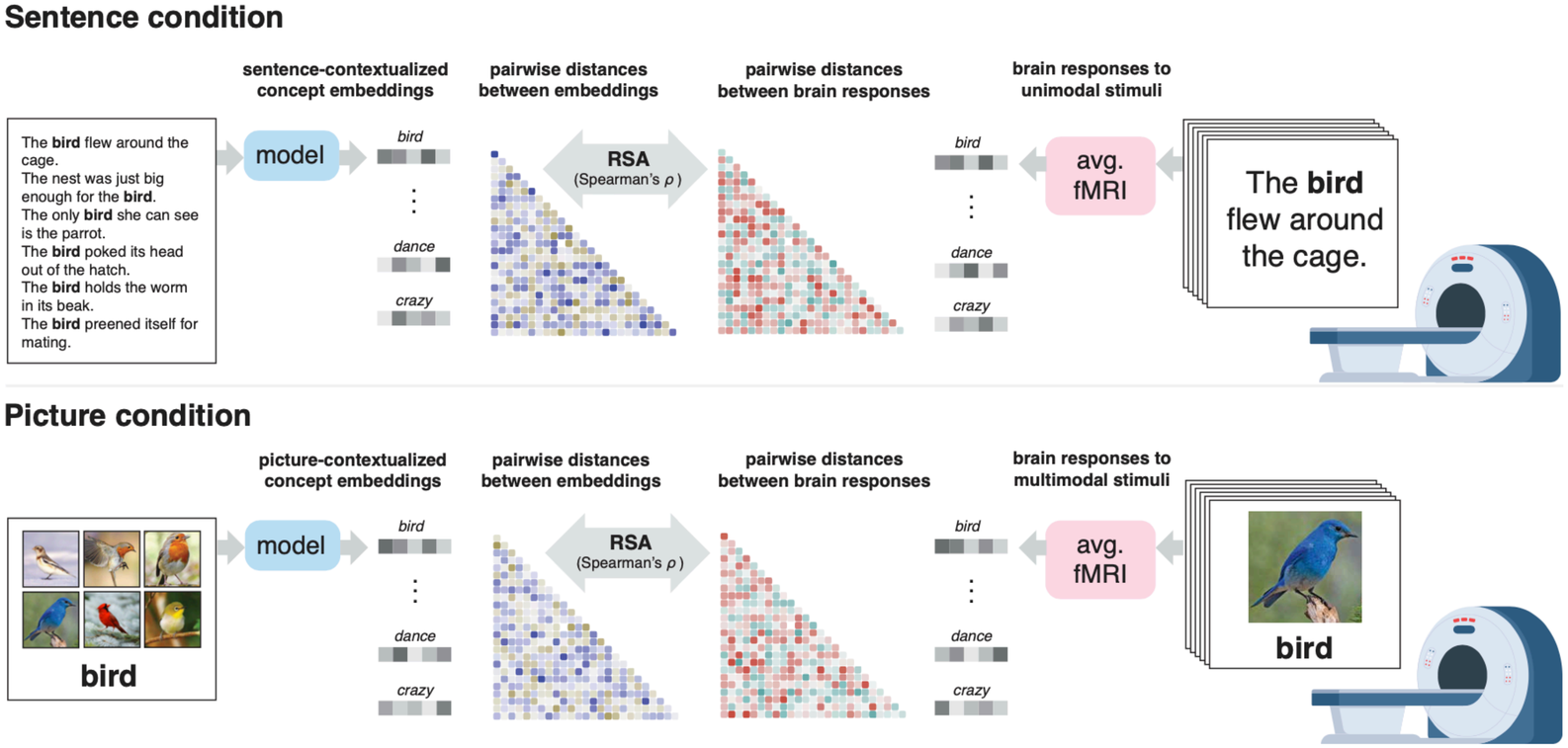

Max: 7.07 sec

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.

Cognitive relevance of vision-language models

"Modelling Multimodal Integration in Human Concept Processing with Vision-Language Models"

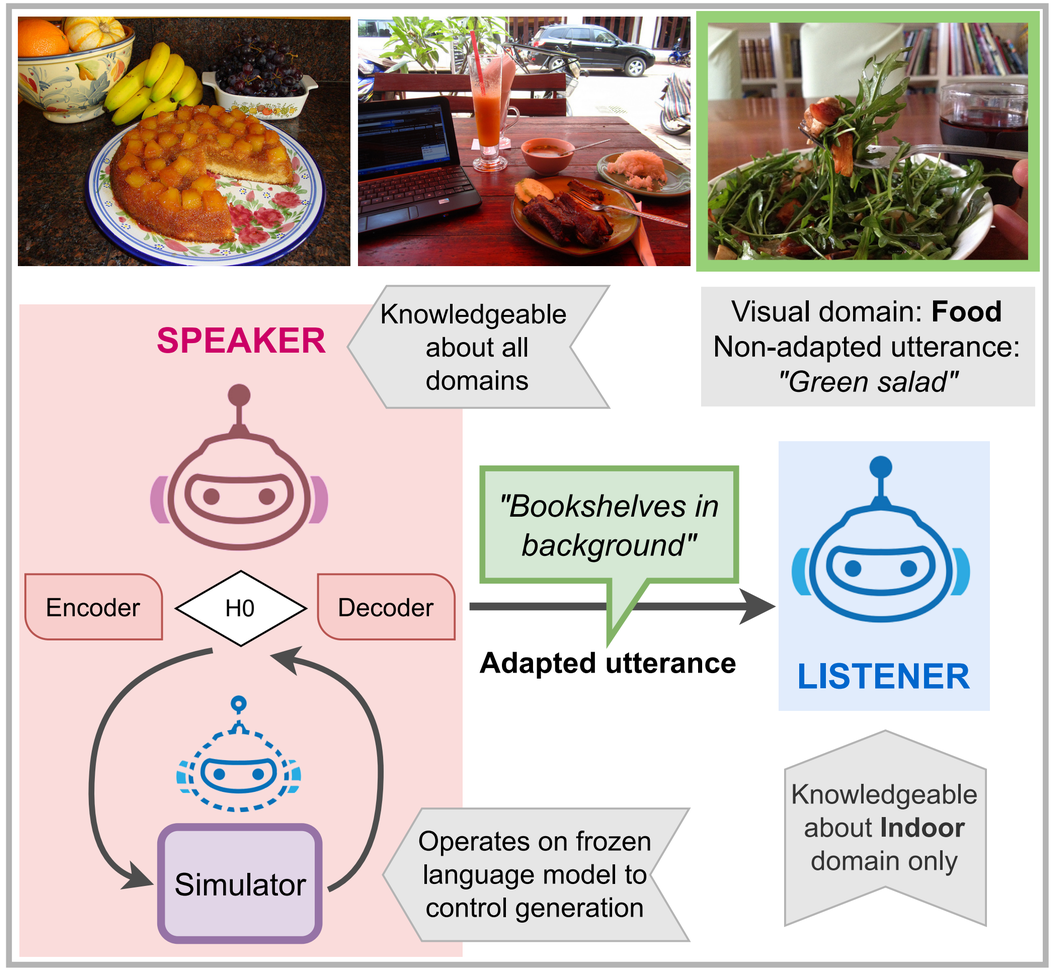
with Anna Bavaresco, Marianne De Heer Kloots, Sandro Pezzelle



Given their multimodal pre-training, do VLMs learn representations that are more aligned with how the brain represents conceptual knowledge?

(in preparation)

Interaction with vision-language models



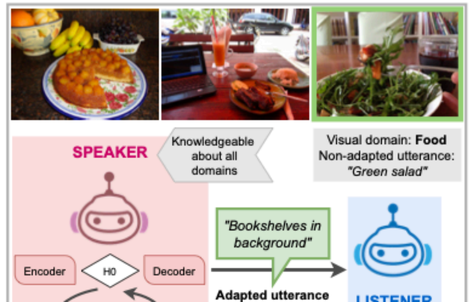
(ACL Findings 2023)

Speaking the Language of Your Listener: Audience-Aware Adaptation via Plug-and-Play Theory of Mind

Ece Takmaz[◁*], **Nicolo' Brandizzi**[◇*],
Mario Giulianelli[◁], **Sandro Pezzelle**[◁], **Raquel Fernández**[◁]
[◁]University of Amsterdam [◇]Sapienza University of Rome
{ece.takmaz|m.giulianelli|s.pezzelle|raquel.fernandez}@uva.nl
brandizzi@diag.uniroma1.it

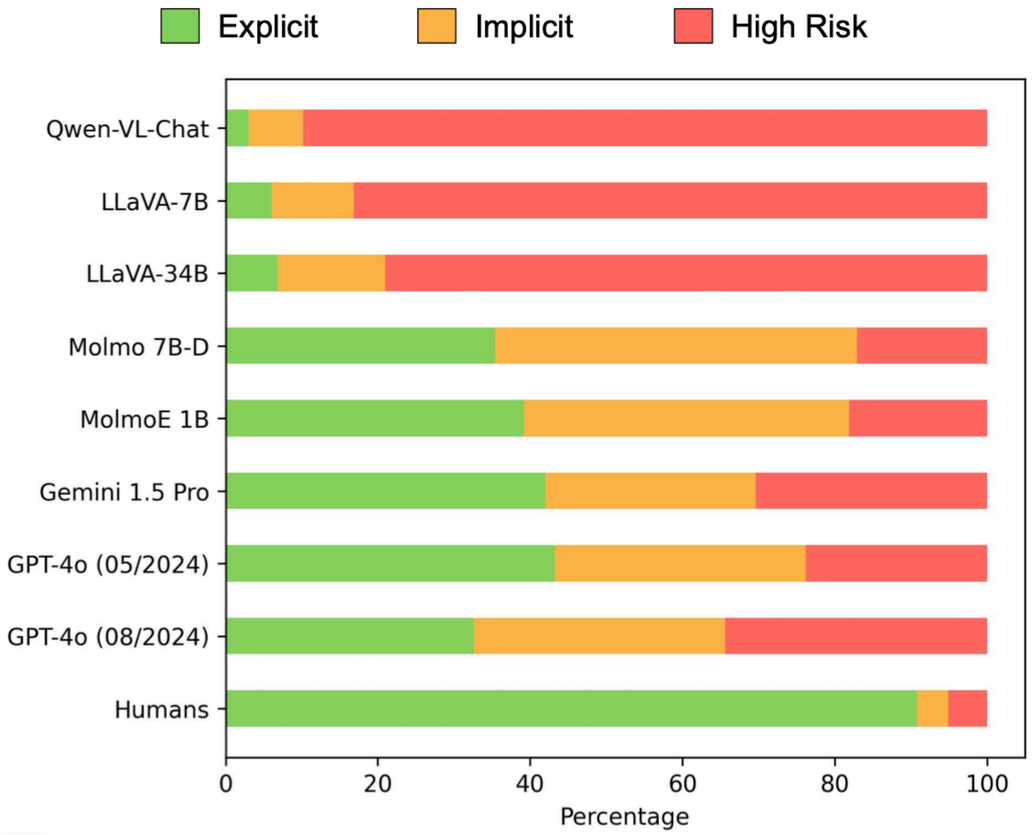
Abstract

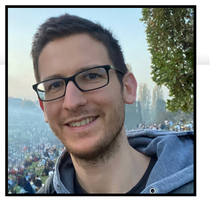
Dialogue participants may have varying levels of knowledge about the topic under discussion. In such cases, it is essential for speakers to adapt their utterances by taking their audience into account. Yet, it is an open question how such adaptation can be modelled in computational agents. In this paper, we model a visually grounded referential game between a knowl-



Interaction with vision-language models

Due to their instruction fine-tuning, models assume grounding and don't ask for clarification, even in ambiguous contexts.





(arXiv:2412.13835)

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 am-

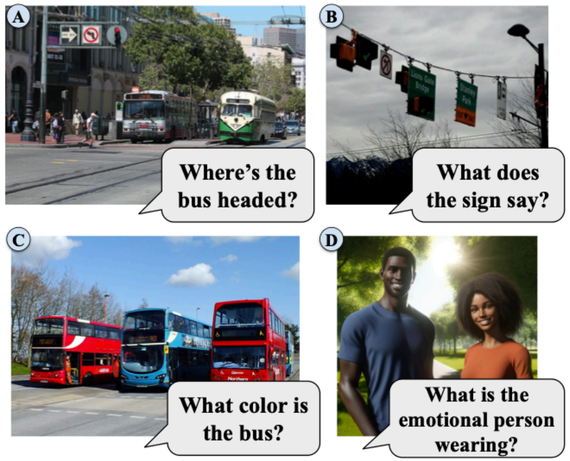


Figure 1: Examples of ambiguous question-image pairs from RACQUET-GENERAL (A,B,C) and RACQUET-BIAS (D).

Towards modelling face-to-face interaction

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

Multimodality also refers to the rich array of signals we exploit in this setting: gestures, gaze, facial expressions — and their interplay with speech.

Different kinds of **gesture**

- Emblems
- Beat or rhythmic
- Pointing or deictic
- **Representational or iconic co-speech gestures**

Two recent papers 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 Psycholinguistics

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

ABSTRACT

In face-to-face dialogues, the form-meaning relationship of co-speech gestures varies depending on contextual factors such as what the gestures refer to and the individual characteristics of speakers. These factors make co-speech gesture representation learning challenging. How can we learn meaningful gestures representations considering gestures' variability and relationship with speech? This paper tackles this challenge by employing self-supervised contrastive learning techniques to learn gesture representations from skeletal and speech information. We propose an approach that includes both unimodal and multimodal pre-training to ground gesture representations in co-occurring speech. For training, we utilize a face-to-face dialogue dataset rich with representational iconic gestures. We conduct thorough intrinsic evaluations of the learned representations through comparison with human-annotated pairwise gesture similarity. Moreover, we perform a diagnostic probing analysis to assess the possibility of recovering interpretable gesture features from the learned representations. Our results show 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

Gesture Representations in Dialogue through Contrastive Learning: An Intrinsic Evaluation. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '24), November 4–8, 2024, San Jose, Costa Rica.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3678957.3685707

1 INTRODUCTION

Co-speech hand gestures are intentionally used along with speech to convey meaning [43]. For instance, representational iconic gestures depict objects, events or actions through various representational techniques such as enacting, tracing, and hand-shaping [31]. Gesture analysis is an active research area in fields such as Human-Computer Interaction (HCI) [39], Sign Language Recognition (SLR) [29, 32], and human behavior analysis [17, 34], where sensory data collected through wearable sensors [22] or, more commonly, through passive sensors like RGB or depth cameras are widely used for studying gestures [50, 59, 60].

In face-to-face interaction, the form-meaning relationship of co-speech gestures is influenced by various situational and contextual 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 [41, 62, 64]. Thus, the employed objectives are focused on the task-

(arXiv:2503.00071)

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
[4]Institute for Logic, Language and Computation, University of Amsterdam
[1]**Correspondence:** esam.ghaleb@mpi.nl

Abstract

In face-to-face interaction, we use multiple modalities, including speech and gestures, to communicate information and resolve references to objects. However, how representational co-speech gestures refer to objects remains understudied from a computational perspective. In this work, we address this gap by introducing a multimodal reference resolution task centred on representational gestures, while simultaneously tackling the challenge of learning 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

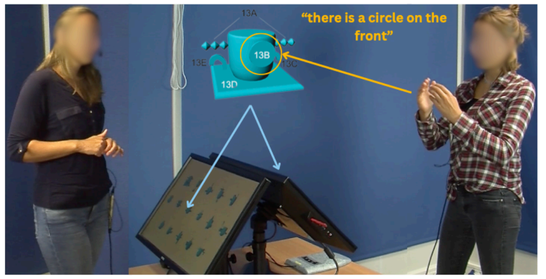


Figure 1: Example from the CABB dataset (Rasenberg et al., 2022), illustrating how participants resolve references through speech and gestures in face-to-face dialogue. The speaker on the right says "there is a circle on the front" while performing a representational gesture. The object is shown for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

Why learning gesture representations?

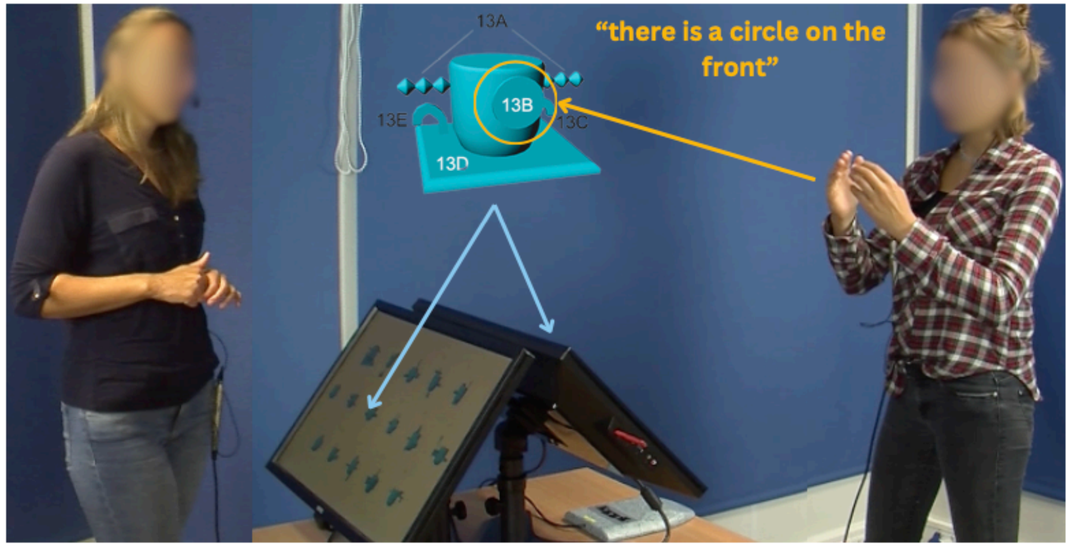
- To study multimodal interaction, e.g., compute gesture similarity to measure shared understanding and alignment in conversation
- To interpret gestures in downstream tasks, e.g., reference resolution

Data-driven representations that can be used to study gestures at a large scale, and to power downstream tasks

The CABB dataset

Referential task, Dutch native speakers

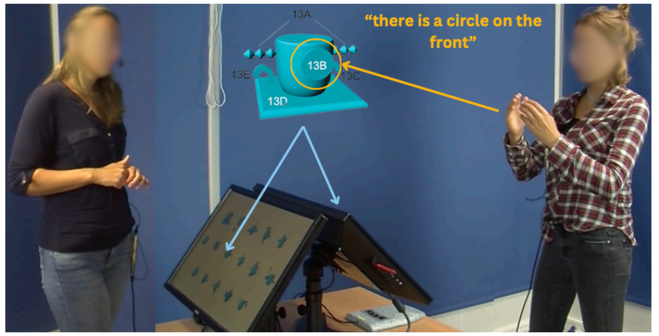
- Director and matcher roles
- 16 objects without conventional names
- Each dyad plays the game for 6 rounds



Classic setup to study shared understanding and cross-speaker alignment

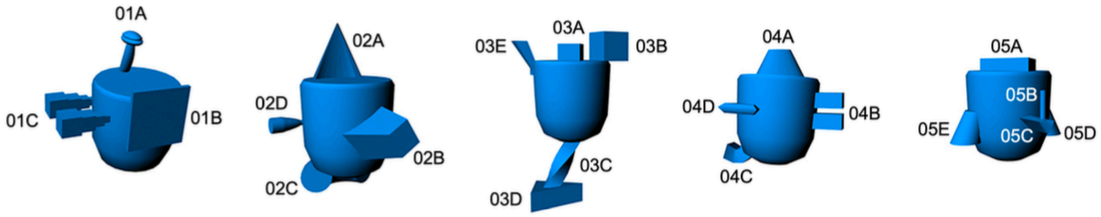
- Entrainment and conceptual pacts with linguistic expression (Ghaleb et al., 2024)
- Alignment in the used of representational gestures (Akamine et a., 2024)

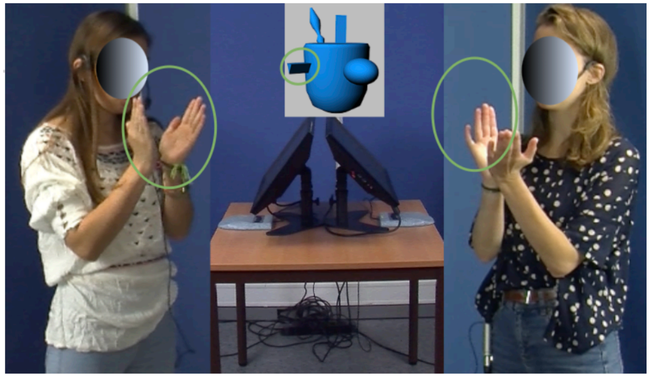
The CABB dataset



CABB-Small (Rasenberg et al., 2022)

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





- 419 pairs of gestures are manually annotated with form features indicating similarity wrt *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

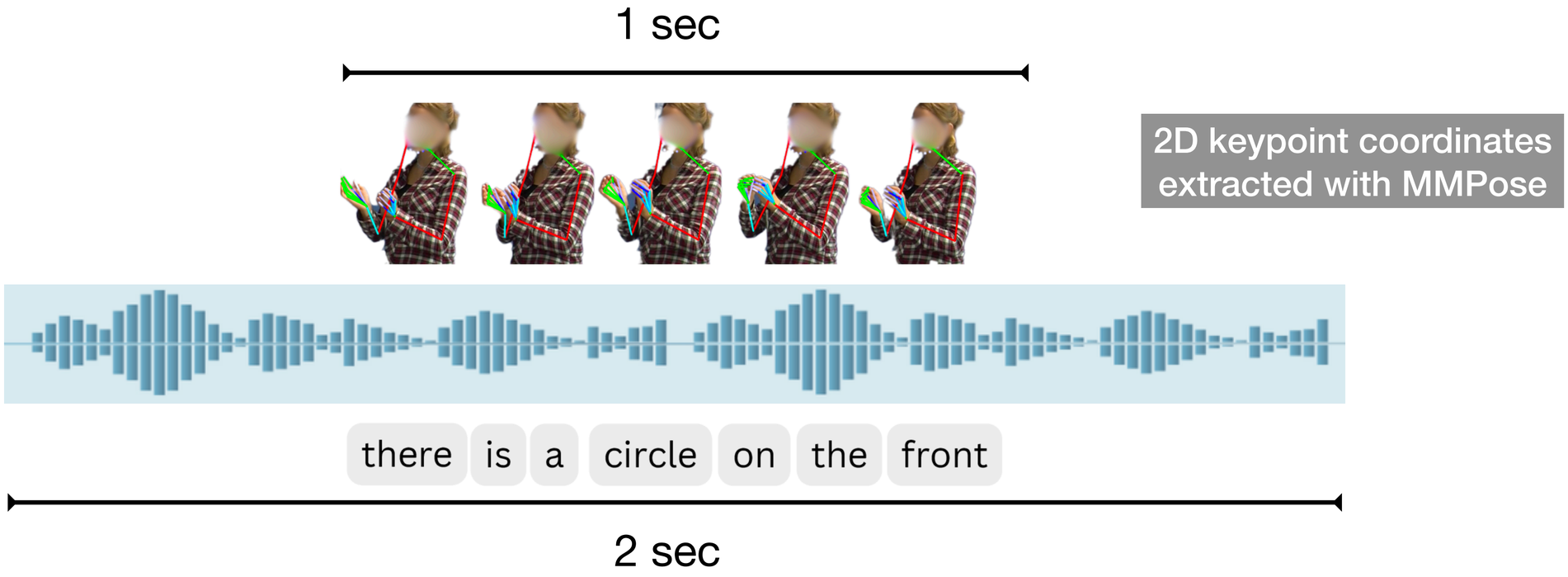
Self-supervised pre-training for gesture representation learning using **CABB-LX**

- Input: kinematics (only body movements) *vs.* kinematics + speech
- Different model architectures that exploit contrastive learning objectives and differ in technical complexity

Evaluation using **CABB-S**

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

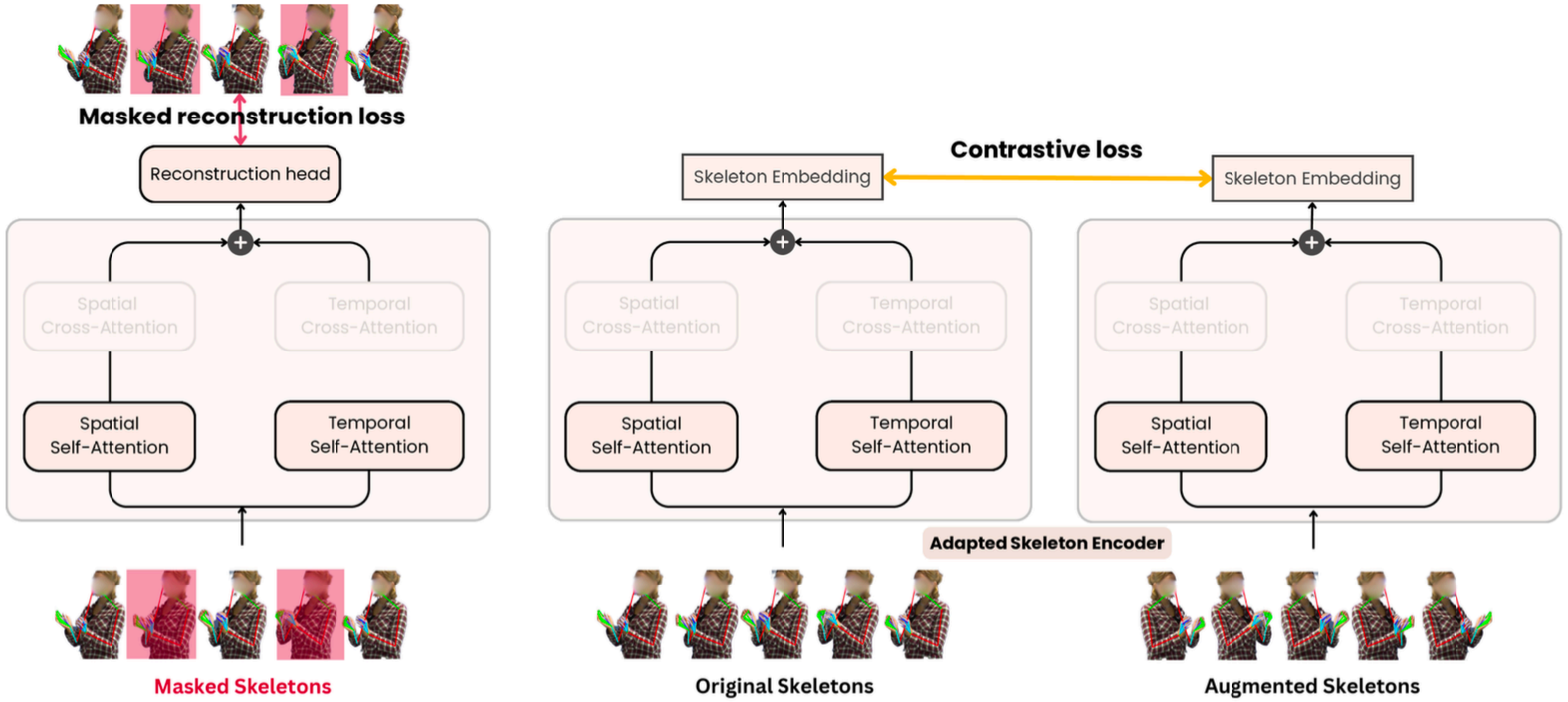
Pre-processing and modality encoders



- **Kinematics**: Transformer encoder for sequences of body movements (Zhu et al., 2023)
- **Speech**: Multilingual marked speech language model `wave4vec-2` (Baevski et al., 2020)
- **Semantics**: embedding of transcribed speech with Dutch BERT (de Vries et al., 2019)

Model architectures

Unimodal: only body movements

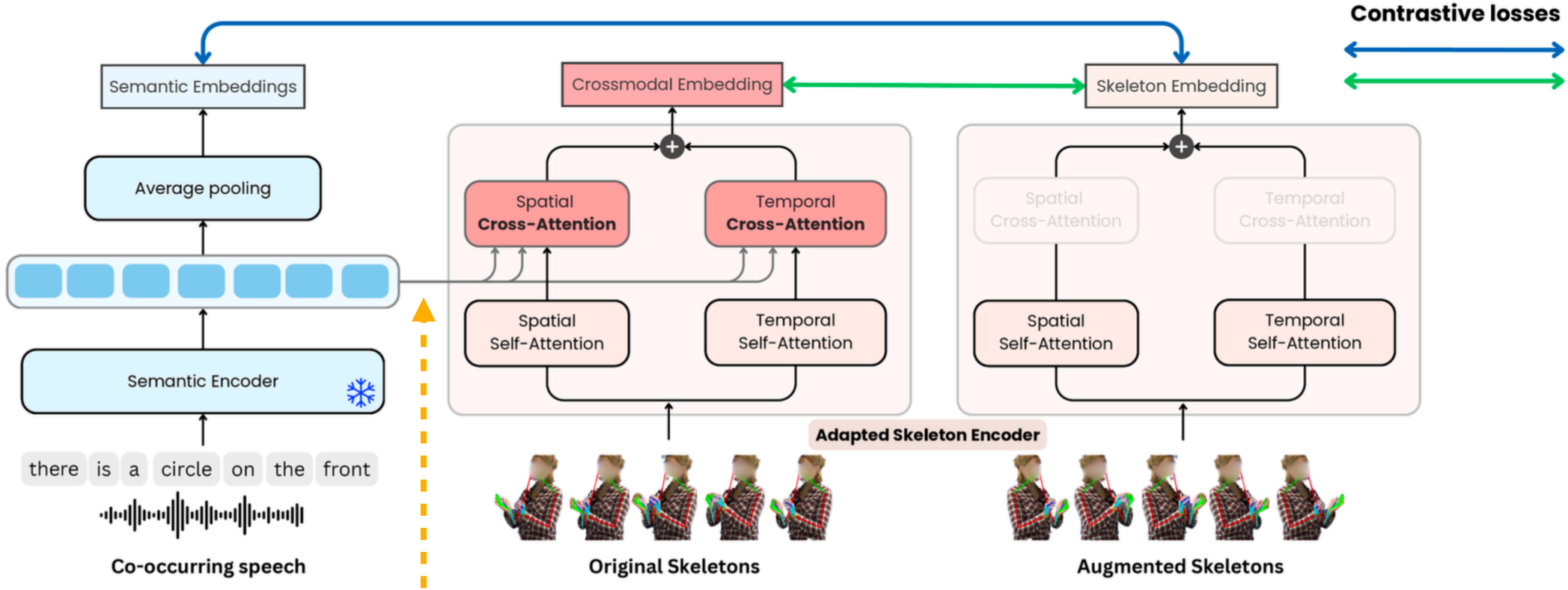


Adaptation of DSTFormer (Zhu et al., 2023)

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)



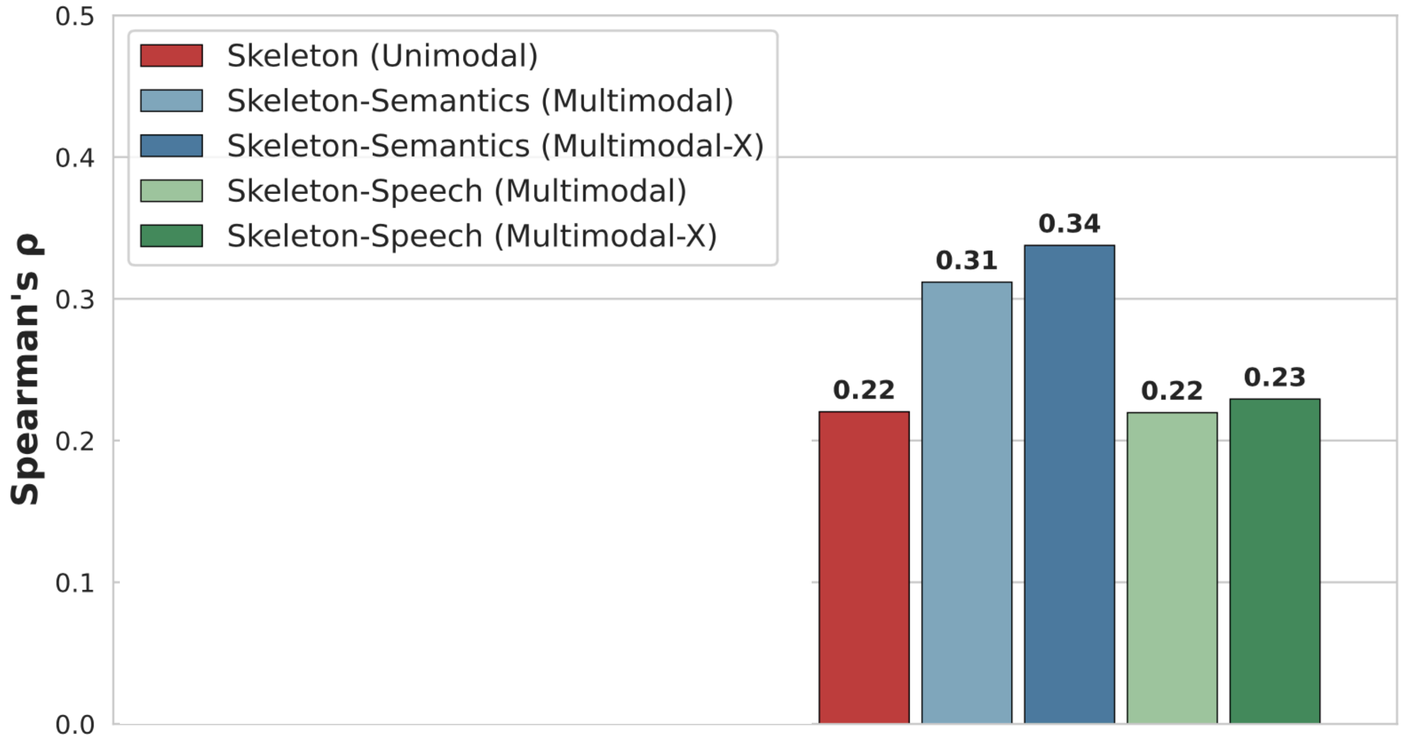
Raw speech or semantic embeddings

Multimodal without or with attention (Multimodal-x)

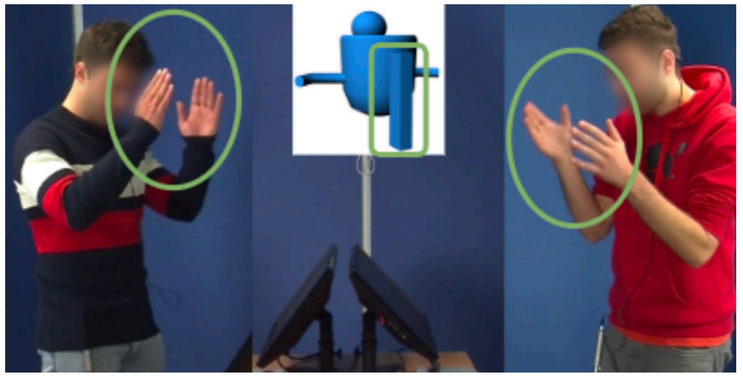
Intrinsic evaluation results

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

We observe a positive correlation with manually annotated gesture similarity:



Intrinsic evaluation results



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

Given the nature of representational gestures, gestures with the **same referent** are expected to be more similar than gestures that refer to the different objects.

- Mean cosine similarity 0.22 vs. 0.12

There is individual variability across speakers. Hence, gestures by the **same speaker** are expected to be more similar than gestures by different speakers.

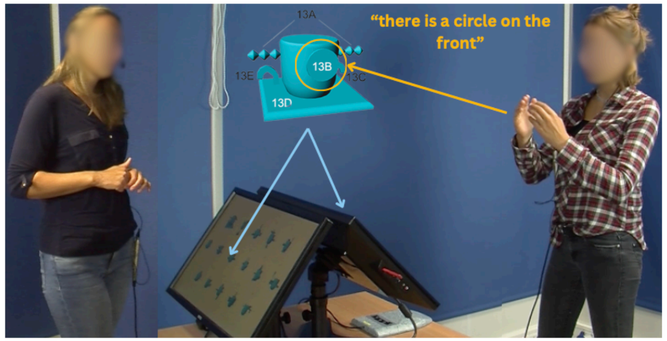
- Mean cosine similarity 0.26 vs. 0.08

Dialogue participants tend to align through **interaction**. Hence, within-dialogue gestures with the same referent will be more similar than across dialogues.

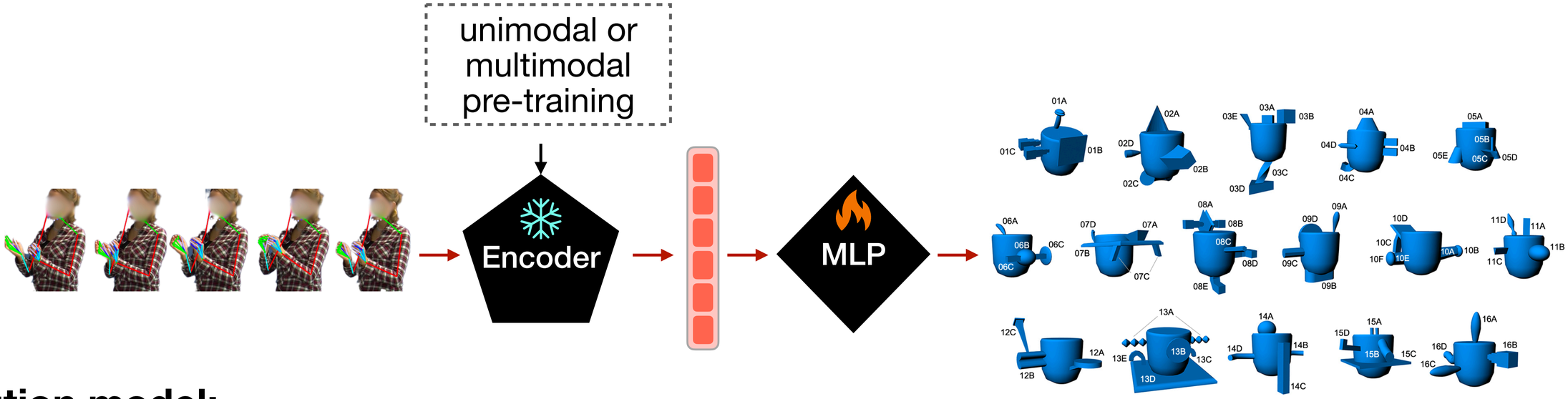
- Mean cosine similarity 0.10 vs. 0.09

(Results from Ghaleb et al. ICMI 2024)

Reference resolution



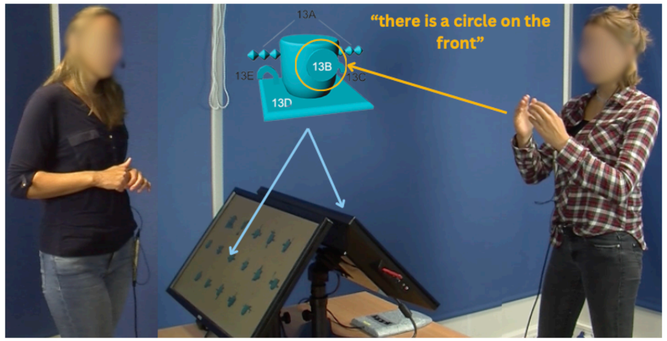
Do gestures, as learned with our approach, contribute to identifying referents?



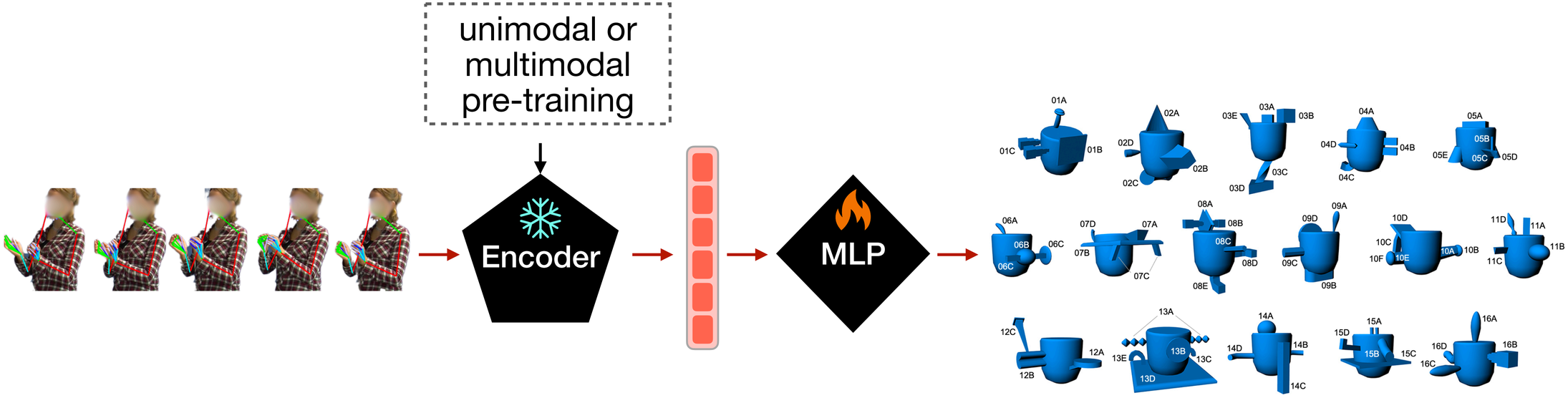
Resolution model:

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

Reference resolution



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

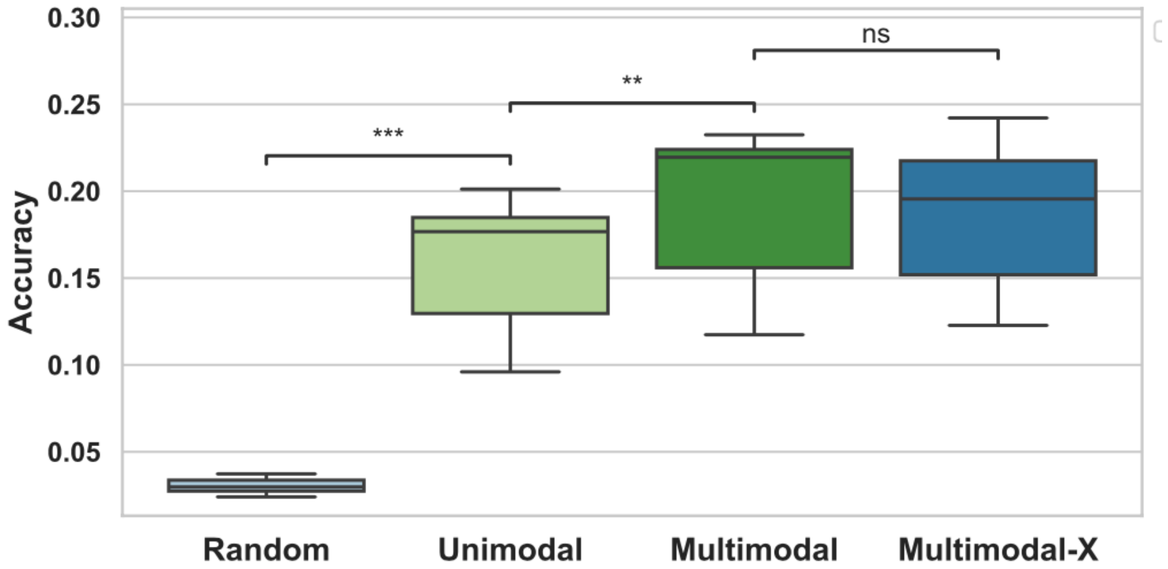


Two scenarios:

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

Reference resolution results

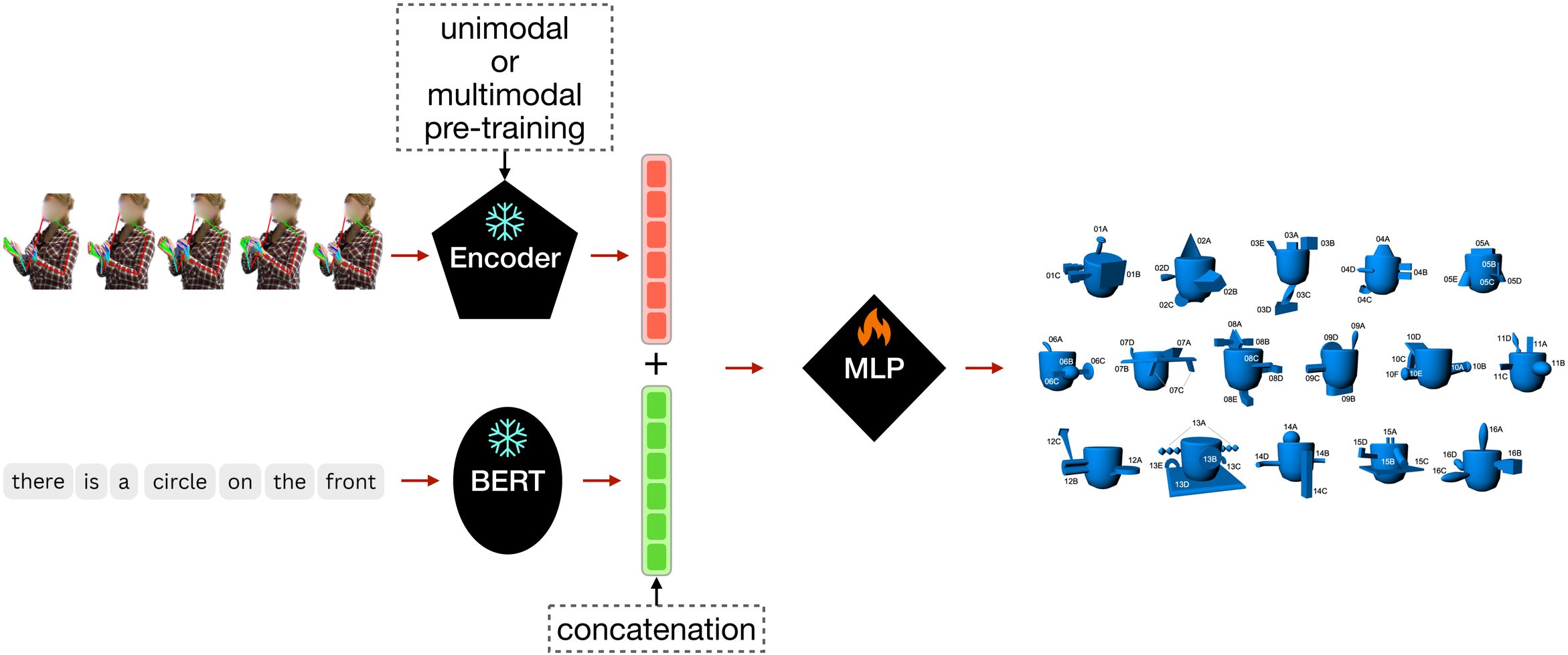
Only body movements at prediction time



- 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

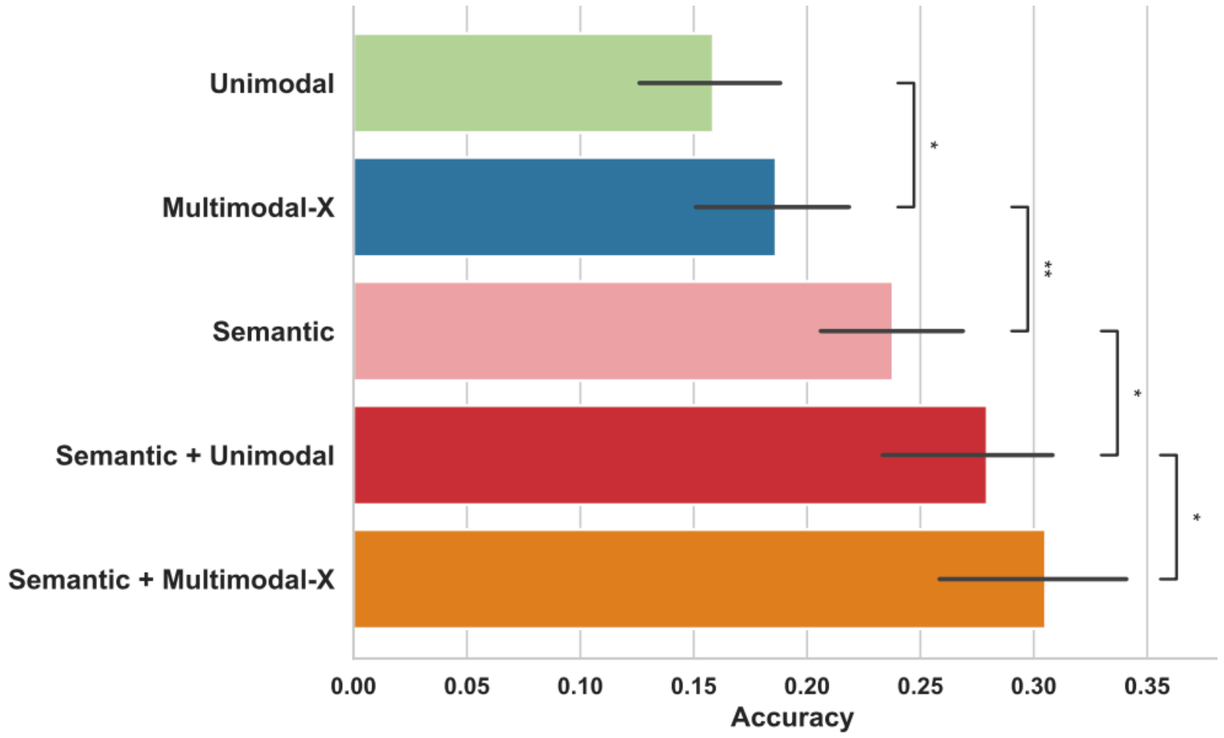
Reference resolution results

Body movements and speech at prediction time



Reference resolution results

Body movements and speech at prediction time



- Information in the vocal modality has more predictive power: 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.)

Conclusion

Modelling gestures by grounding them in speech leads to representations that **comply with theoretical expectations** and contribute to **reference resolution**.

Some directions for future work:

- Impact of dialogue history: initial results show it is useful
- Amount of gestures decreases over the dialogue: what does this tell us about their role?
- Grounding gestures (and speech) on the visual properties of referents

Thanks!

Amsterdam's
Dialogue Modelling Group

UNIVERSITY
OF AMSTERDAM

Institute for Logic,
Language & Computation

ellis
UNIT
AMSTERDAM

erc
European
Research
Council