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 predictive power. Moreover, reference resolution accuracy further improves when (1) using multimodal gesture representations, even when speech is unavailable at inference time, and (2) leveraging dialogue history. Overall, our findings highlight the complementary roles of gesture and speech in reference resolution, offering a step towards more naturalistic models of human-machine interaction.
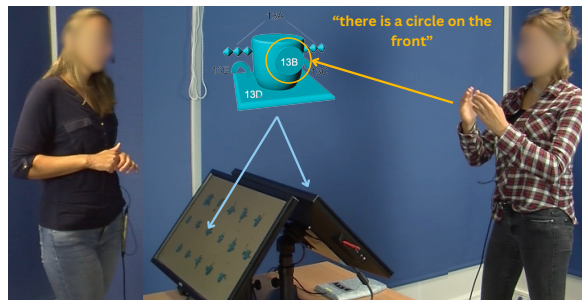
Figure 1: Example from the CABB dataset (Rasenberg et al., 2022), illustrating how participants resolve references through speech and gestures in face-to-face dialogue. The speaker on the right says "there is a circle on the front" while performing a representational gesture that resembles a circle. The object discussed is shown for illustration but not visible to the listener; the orange arrow points to the referent as annotated by experts. Our work draws on these interactions to model *multimodal reference resolution*.

1 Introduction

Referring to objects is common in everyday communication. In face-to-face interaction, when we need to collaborate on new tasks or refer to new objects, we rely on verbal (*i.e.*, speech) and non-verbal (*e.g.*, gestures and gaze) cues to describe salient object features and direct the listener's attention. Among the non-verbal cues are *representational co-speech gestures*, *i.e.*, iconic hand movements semantically and pragmatically related to co-occurring speech (Kendon, 2004). Studies have shown that representational gestures facilitate language comprehension (Drijvers and Özyürek, 2017; Arbona et al., 2023) and help listeners identify referents more quickly than speech alone (Campana

et al., 2005). Along with speech, gestures are used to refer to novel objects and build shared understanding (Rasenberg et al., 2022; Akamine et al., 2024), as shown in Figure 1. These multimodal abilities are inherent ingredients of our communicative interactions (Özyürek, 2014). Hence, developing computational approaches that can interpret such cues is important for naturalistic human-machine collaboration: in situated dialogue, an artificial agent must recognize multimodal inputs and speaker references to meet human needs (Moon et al., 2020; Kontogiorgos et al., 2018).

However, how to computationally represent and interpret co-speech gestures remains an understudied problem, particularly within Natural Language Processing. Other gestural forms, such as deictic gestures, *i.e.*, pointing (Gatt and Paggio, 2013; Kennington and Schlangen, 2017a; Chen et al., 2021; Kontogiorgos et al., 2018) or beat gestures, *i.e.*, rhythmic movements without semantic content (e.g., Sinclair and Schneider, 2021; Abzaliev et al.,

2022), have received some attention. In contrast, the challenges posed by iconic, representational gestures and their contribution to reference resolution have hardly been tackled from a data-driven perspective.

In this paper, we propose an approach to learning embeddings for representational gestures that exploits not only the body movements that make up a gesture but also the semantically related speech that is typically produced simultaneously with it. We combine contemporary self-supervised learning techniques to train a Transformer-based gesture encoder and ground it in information from features extracted by text or speech large language models. Then, we test the effectiveness of the pre-trained gesture embeddings in the downstream task of reference resolution in face-to-face dialogue, showing that gestures—as learned by our proposed multimodal approach—have significant predictive power that complements the verbal modality. More concretely, we make the following contributions:

- We propose three model architectures for gesture representation learning that exploit a version of the motion encoder DSTFormer (Zhu et al., 2023), which we adapt to allow for the integration of speech through cross-modal attention.

- We show that the resulting pre-trained gesture embeddings are aligned with expert knowledge present in manual annotations, clearly surpassing earlier approaches to gesture representation learning (Ghaleb et al., 2024b).

- We introduce a novel multimodal reference resolution task and demonstrate that learning gesture representations by jointly exploiting body movements and the semantics of the concurrent speech results in more accurate models, even when speech is not available at prediction time.

- Our reference resolution experiments also show that leveraging dialogue history improves model prediction and that, when speech is present at test time, gestures provide complementary information that enhances reference resolution accuracy.

- Our experiments make use of the CABB dataset (Rasenberg et al., 2022; Eijk et al., 2022), collected by cognitive scientists. We make available the pre-processed data and the code to reproduce all our results via a public GitHub repository, providing valuable resources to the community.[1]

---

[1]https://github.com/EsamGhaleb/ReferenceResolution

2 Related Work

2.1 Learning Multimodal Representations

Despite the importance of gestures in multimodal communication, learning gesture representations remains challenging and understudied in both computer vision and NLP. Some existing work has used formal approaches to integrate gestures into discourse semantics (Lascarides and Stone, 2009; Lai et al., 2024), while a few other works have employed data-driven methods. For example, Abzaliev et al. (2022) jointly learned gesture and word embeddings from TED talks using contrastive learning, and showed that function words, discourse markers, and the language of the speaker can be predicted from non-representational gestures. Self-supervised contrastive learning techniques (Chen et al., 2020; Radford et al., 2021) have been widely adopted in the field of multimedia to learn representations of human movements from skeletal joint coordinates unimodally (Thoker et al., 2021; Zhu et al., 2023) and in combination with other data modalities (Brinzea et al., 2022; Liu et al., 2024), while (Lee et al., 2021) used self-supervised learning to learn gesture embeddings as a pre-training stage for gesture generation.

Our approach to learning gesture representations is most closely related to the preliminary work of Ghaleb et al. (2024b), who proposed to learn embeddings for representational gestures by grounding them in co-occurring speech. We substantially extend this work by replacing their skeleton encoder with a Transformer-based encoder, allowing us to integrate not only speech but also text-based semantic embeddings with higher temporal granularity and using a much larger amount of data samples. Furthermore, unlike this work, we exploit the learned gesture embeddings for the downstream task of reference resolution, here formulated as the problem of identifying the object referred to by a gesture in face-to-face dialogue.

2.2 Reference Resolution in Dialogue

Reference resolution in dialogue has mostly been modelled as the task of identifying the referent of text-based linguistic expressions, ignoring non-verbal cues. For example, Skantze and Willemsen (2022) proposed COLLIE, a continual learning method that adjusts language embeddings to accommodate new language use for new referents; in an earlier study, Shore and Skantze (2018) found that leveraging dialogue history in the form of pre-

vious referring expressions improves model prediction, similarly to Takmaz et al. (2020). Resolving linguistic referring expressions in the visual modality has also been studied in the field of computer vision thanks to datasets such as ReferIt (Kazemzadeh et al., 2014), Flicker30k Entities (Plummer et al., 2015), and Visual Genome (Krishna et al., 2017), which map referring expressions to regions in an image.

In this work, we focus on reference resolution in face-to-face communication, where linguistic expressions interact with non-verbal signals like gestures. The large majority of work in this domain has been concerned with deictic pointing gestures. For instance, Kennington and Schlangen (2017b) combined linguistic information with gaze and deictic gestures by treating them as separate resolution models and then fusing their predictions via interpolation. Similarly, Kontogiorgos et al. (2018) used multisensory input in a collaborative assembly task to assess the contribution of various cues–such as eye gaze, head direction, and pointing gestures–to reference resolution. They found that deictic gestures, when combined with speech, reliably located objects, while gaze and head direction were only useful for approximating the general location of the intended object when paired with speech. More recently, within the computer vision community, Chen et al. (2021) found that referential expressions were more discriminative when both visual context and pointing gestures were considered, compared to using visual context alone.

In this paper, we tackle reference resolution by means of iconic representational gestures rather than pointing, calling attention to the importance of modelling such gestures to identify objects in multimodal interaction.

3 Data

For our study, we use the CABB dataset (Eijk et al., 2022; Rasenberg et al., 2022), which consists of face-to-face conversations in Dutch between two dialogue participants who play a reference game. The setup is shown in Figure 1. The participants' task is to identify 16 objects without conventional names that are made up of different geometrical parts (see Appendix A). Each dyad plays the game for six rounds, exchanging the roles of 'director' (who describes one of the target objects) and 'matcher' (who attempts to identify the director's intended referent among the 16 candidate objects displayed

on a screen). The participants are free to communicate as they like, which elicits spontaneous speech and gestures. Speakers were video recorded from different angles and we make use of the semifrontal views shown in Figure 1, as well as the audio recordings.

We use two different subsets of this data, which we refer to as CABB-S and CABB-L, plus an extension of the latter which we call CABB-XL:

CABB-S (Rasenberg et al., 2022) consists of 19 dialogues by 38 individuals, corresponding to over 8 hours of recordings. The dataset includes manual speech transcriptions and manual segmentation of gesture strokes, with 4949 gesture segments in total. Approximately 97% of these segments are accompanied by concurrent speech. CABB-S also includes manual annotations of gesture strokes with two types of information:[2]

- **Referent:** The object subpart referred to by the gesture. The candidate objects and their sub-parts are shown in Appendix A, Figure 9.

- **Form similarity:** 419 pairs of gestures with the same referent are annotated with five low-level binary features indicating whether two semantically related gestures are similar regarding shape, movement, rotation, position, and use of hands.

We use CABB-S for evaluating our pre-training approach to gesture representation learning (Section 4.3) and for the experiments on reference resolution (Section 5).

CABB-L (Eijk et al., 2022) contains an additional 49 dialogues by 98 different subjects, with about 42.5 hours of recordings. It is therefore much larger than CABB-S. Only 42 dialogues are manually transcribed and no manual annotations regarding gestures are present. To identify gestures, we use the segmentation model by Ghaleb et al. (2024a), which has been shown to achieve a mean Average Precision (mAP) of 76% on the CABB-S dataset. Applying this model to CABB-L results in 30k automatically segmented gestures.[3]

To increase the amount of data available for pre-training, we oversample by selecting 1-second time windows overlapping more than 50% with the automatically segmented gestures. This results in approximately 400k data samples, which we refer to as **CABB-XL**. We use Whisper-X (Bain et al.,

---

[2]For further details, see Rasenberg et al. (2022).

[3]Some qualitative results on segmentation performance can be found in Appendix B.

2023) to automatically generate speech transcriptions when manual transcriptions are unavailable. 83% of the gestures are accompanied by speech. We use CABB-L and CABB-XL for pre-training the models introduced in Section 4.

Pre-processing To process body movements, we apply the procedure used by Ghaleb et al. (2024b) to CABB-S, CABB-L, and CABB-XL. Concretely, we sample 1-second time windows centered around each segmented gesture and use ViTPose (Xu et al., 2022) to extract skeletal information, *i.e.*, 2D keypoint coordinates for 27 upper body and hand joints. Holler and Levinson (2019a) showed that related speech is often produced before or after gesture strokes (the most meaning-bearing segment of a gesture) by a few hundred milliseconds. To account for this temporal asynchrony between speech and gestures, when processing the verbal modality we extract 2-second windows centered around the sampled gestures, and use both the raw speech and the transcriptions as described in the next section.

4 Gesture Representation Learning

In this section, we present our approach to learning robust gesture representations in a self-supervised fashion. To do so in the context of *multimodal* communication, we experiment with three types of input: gestures themselves (*i.e.*, skeletal information corresponding to body movement), raw speech, and text-based semantics (Section 4.1). We propose three pre-training model architectures that use these input types to different degrees and with different multimodal integration strategies (Section 4.2). We train these models on CABB-L/XL and evaluate them against expert annotations using the unseen gestures in CABB-S (Section 4.3). In Section 5, we then test the effectiveness of our pre-training approach for the task of reference resolution.

4.1 Modality Encoders

We use three encoders to extract representations of speech, text, and body movements, respectively.

Speech. As speech encoder, we use multilingual wav2vec-2 (version `wav2vec2-xlsr-300`), a masked-language model pre-trained on a large number of speech datasets in multiple languages (Baevski et al., 2020; Conneau et al., 2020). Similarly to Pepino et al. (2021), we aggregate the embeddings across all Transformer layers using a learnable weighted average and pass the output through two point-wise CNN layers to fuse signals along the temporal dimension.

Semantics. Although wav2vec-2 representations may capture diverse linguistic properties including prosody, phonetics, and to some extent semantics (Tsai et al., 2022; Zaiem et al., 2025), they are less semantically rich than word embeddings learned from text. Therefore, we also experiment with the word embeddings from a pre-trained Dutch BERT-based model (BERTje; de Vries et al., 2019).

Skeleton. We adapt DSTFormer (Zhu et al., 2023) to encode sequences of body movements. The original model has two parallel branches: one applies temporal self-attention followed by spatial self-attention, and the other one spatial followed by temporal. To reduce overhead, in each encoder, we keep only one temporal layer and one spatial layer in each branch and replace the second layer with an optional cross-attention module. This optional cross-attention takes semantic or speech embeddings as keys and values, as schematically illustrated in Figure 2.

4.2 Model Architectures

We propose three pre-training strategies to learn gesture representations. The first one is unimodal, in the sense that it learns representations only considering the body movements that make up a gesture. The other two (multimodal and multimodal-X) are motivated by a more holistic conception of co-speech gestures as multimodal acts (Holler and Levinson, 2019b; Özyürek, 2014), and therefore exploit both skeletal and concurrent verbal input. We describe the gist of each architecture here and provide further technical details in Appendix C.

Unimodal architecture. This model only leverages skeletal information. It jointly optimizes a masked reconstruction loss and a unimodal contrastive loss. For the former, we follow the original procedure for pre-training DSTFormer (Zhu et al., 2023) by randomly masking portions of the 2D keypoint skeletal input and learning to reconstruct them. The unimodal contrastive loss pulls representations of two views of augmented skeletal data closer while pushing them away from other negative samples in a batch. A detailed diagram of this architecture is shown in Figure 14, Appendix C.

Multimodal architecture. This model combines the two losses from the unimodal architecture with
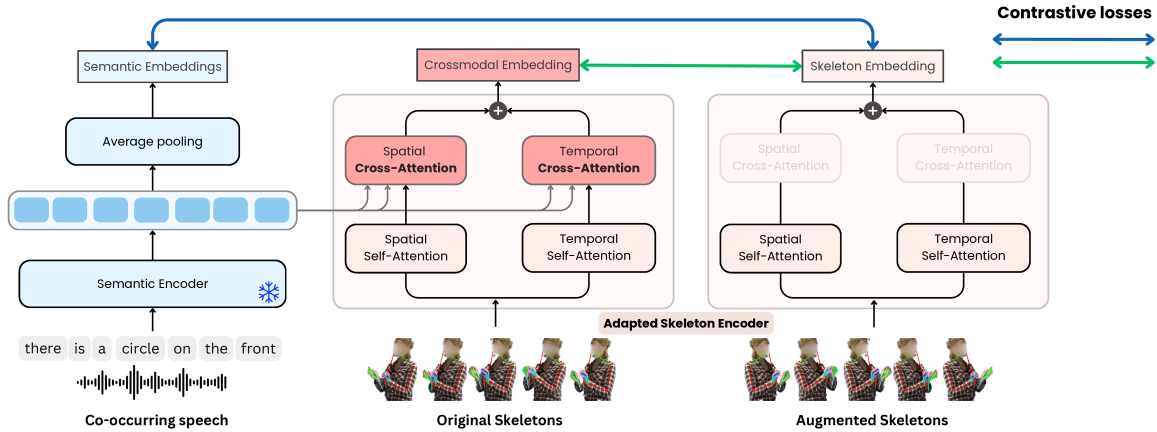
Figure 2: Our multimodal-X architecture. The left branch encodes semantic information (text or speech) and fuses it with skeleton embeddings via the proposed cross-attention blocks in our adapted skeleton encoder. The architecture is trained by minimizing contrastive losses.

a *multimodal* contrastive loss that integrates skeletal information with either speech or semantics. For the multimodal contrastive loss, we use a CLIP-like contrastive objective (Radford et al., 2021) mapping global representations of gestures and co-occurring utterances (as either raw speech or semantics) into a joint feature space.

Multimodal-X architecture. Finally, this model is optimized with two complementary contrastive losses, as illustrated in Figure 2: the multimodal contrastive loss (blue arrows) described as part of the multimodal architecture above and an additional *crossmodal* contrastive loss (green arrows), which aligns the unimodal skeleton representation with the fused gesture-semantic (or gesture-speech) embedding—the latter is obtained by injecting text tokens (or speech frames) into our DSTFormer backbone via the cross-attention layers described in Section 4.1.

The unimodal and multimodal architectures can thus be seen as ablations of the arguably more powerful multimodal-X architecture: the multimodal architecture does not include cross-attention layers, and the unimodal architecture omits multimodal alignment altogether.

Training and implementation details. We train the three architectures above using CABB-L and CABB-XL, which allows us to test the impact of increasing the size of the training data. In Appendix D, we provide further details about the backbone models, projection heads, and the parameters used in the learning objectives, along with the implementation details.
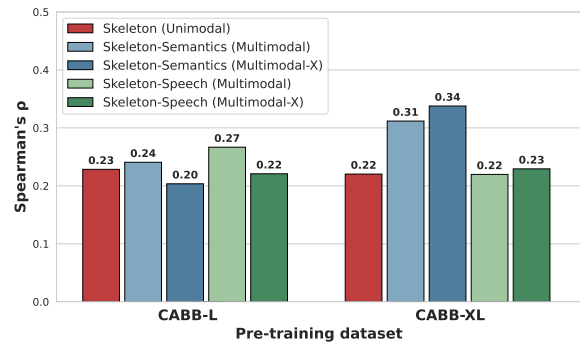

Figure 3: Spearman correlation between the number of form features shared by a pair of gestures and their cosine similarity using embeddings from skeleton-speech, skeleton-semantics, and unimodal models. Pre-training was conducted on CABB-L and CABB-XL, while the correlation scores were computed on CABB-S. All coefficients are statistically significant ($p \ll 0.05$).

4.3 Evaluation

We evaluate the gesture representations learned with our pre-training architectures using the CABB-S dataset, which contains manually annotated information on gestures unseen during model pre-training. To monitor pre-training progress and save the best model variants across epochs, we conducted evaluations using form similarity as correlation. Following Ghaleb et al. (2024b), we compute Spearman's correlation between the number of form features a pair of gestures share according to experts' annotations and cosine similarity between the gestures' learned representations.

Figure 3 shows the correlation results, yielded by the best models obtained during pre-training. The figure shows that the variants with the high-
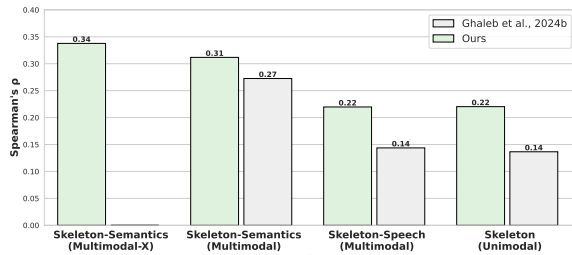
13195

Figure 4: Pre-training on CABB-XL: comparing best models against (Ghaleb et al., 2024b). All Spearman correlation coefficients ρ are statistically significant ($p \ll 0.05$).
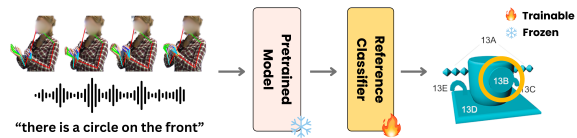

Figure 5: Gesture-based reference resolution. The reference resolution classifier leverages gesture representations encoded with our pre-trained models.

est performance are the multimodal-X and multimodal architectures where gesture representations are jointly learned with text-based semantics from concurrent speech, using the large CABB-XL as training data. The other model variants—unimodal, multimodal (with raw speech), and multimodal-x with raw speech—do not benefit as much from an increase in the amount of training data. In fact, when exploiting raw speech, the best correlation coefficient is obtained with the multimodal architecture and CABB-L.

Comparison with related work. We compare our models against the framework by Ghaleb et al. (2024b). This work proposed a pre-training approach to gesture representation learning based on Spatio-Temporal Graph Convolutional Networks (ST-GCN) using unimodal and multimodal contrastive learning with co-occurring raw speech. For comparability with our approach, when reproducing this framework, in addition to raw speech we extend it to also use text-based semantics and pre-train it on CABB-XL.[4] The comparison of correlation coefficients with form similarity is shown in Figure 4. As can be observed, the gesture representations learned by our Transformer-based encoder are more aligned with form-based expert annotations as evidenced by higher correlation values across the board.

Overall, the correlation analysis indicates that the best pre-training strategies combine skeletal data with semantic information—using either multimodal or multimodal-X approaches—specifically when trained on a large dataset like CABB-XL. In the next section, we focus on model variants trained on CABB-XL that use semantic embeddings, with

unimodal models serving as a baseline.

5 Reference Resolution

An important question in situated interactions is to what extent representational gestures complement or supplement speech in reference resolution. Here we shed light on this question by leveraging our pre-trained gesture models for the downstream task of reference resolution. We investigate whether models that have learned gestures by exploiting multimodal information (from body movements and concurrent speech) have more predictive power than models that represent gestures exclusively from body movements. Moreover, we test whether gestures contribute complementary information to the verbal modality when identifying referents.

5.1 Resolution Model and Evaluation Setup

The resolution model leverages our model architectures pre-trained on CABB-XL without any fine-tuning. This is schematically shown in Figure 5. The model is implemented as a multi-class MLP classifier with two hidden layers of size 300 and 150, respectively, and it is trained on CABB-S. Given a gesture encoded with our pre-trained models, we train the MLP to predict one referent among 70 possible object sub-parts (see Appendix A for details) using a batch size of 32 and a learning rate of 10^{-4} with the Adam optimizer for 200 epochs. Recall from Section 3 that each dialogue consists of six rounds. We use leave-one-round-out cross-validation, holding out the gestures in one round as a test set and training on the gestures in the remaining rounds across all dialogues in CABB-S. We use accuracy as an evaluation metric. Since speakers can refer to any of 70 possible sub-parts (across the 16 objects), randomly guessing results in a rate of 1/70, leading to approximately 1.4% accuracy. A model using random gesture embeddings (without access to our pre-trained models) achieves around 3% accuracy.

Given a gesture unseen during pre-training, we investigate two scenarios: In Section 5.2, we mea-

---

[4]Due to the architecture of ST-GCN, it is not possible to combine it with the multimodal-X architecture introduced in our work (Figure 2).
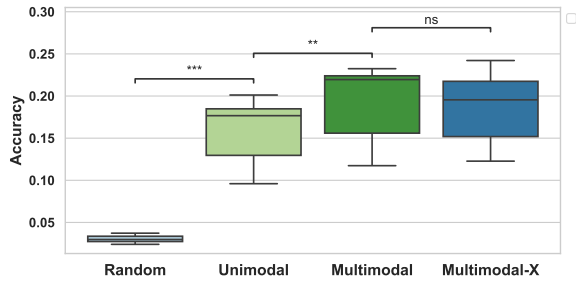
Figure 6: Average reference resolution accuracies for gesture-only embeddings are shown for unimodal, multimodal, and multimodal-x representations. The multimodal and multimodal-x models are pre-trained with text-based semantic input, and the embeddings are derived only from skeletal data. Asterisks indicate statistically significant differences, with ** $p < 0.01$ and *** $p < 0.001$; 'ns' stands for 'not significant'.

sure how accurate a reference resolution system that only has access to the gesture (*i.e.*, to skeletal input) is at predicting the gesture's referent. Here the gesture embedding is extracted zero-shot with our models, some of which exploited raw speech or text semantics during pre-training—but importantly verbal input is not provided at inference time in this scenario. In Section 5.3, we consider a second scenario, where at inference time the reference resolution system has access to both the unseen gesture to be resolved and any concurrent speech. We operationalise this by concatenating the gesture embedding extracted with our models and a semantic embedding of the concurrent speech, and then measure whether this leads to higher reference resolution accuracy than only exploiting the semantics of concurrent speech.

5.2 Gesture-Only Reference Resolution

We first evaluate the resolution model when it only has access to gestural information (*i.e.*, skeletal data) as input. As shown in Figure 6, when a gesture is encoded with our unimodal model, the average resolution accuracy is 16%, significantly above the random baselines. Using embeddings from models that were pre-trained jointly with text-based semantics significantly increases resolution accuracy to around 19%, with no statistically significant difference between the multimodal and multimodal-X approaches. These results show that our pre-trained gesture representations capture information that is useful to identify referents.[5] Moreover, they
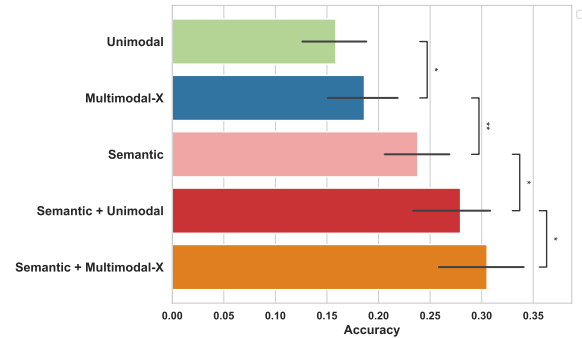

Figure 7: Average accuracies for reference resolution using gestures and co-occurring verbal information are reported for unimodal gesture embeddings, multimodal-x gesture embeddings, semantic embeddings, and their concatenated representations. The multimodal-x gesture embeddings are learned through pre-training with semantic embeddings. Asterisks indicate statistically significant differences, with * $p < 0.05$, ** $p < 0.01$ and *** $p < 0.001$.

indicate that learning gesture representations by jointly exploiting body movements and the semantics of co-occurring speech enhances their reference resolution potential, even when information about concurrent speech is not provided at prediction time.

5.3 Reference Resolution with Gestural and Co-occurring Verbal Information

Next, we assess the resolution model when it has access to gestures and the speech co-occurring with them. As mentioned above, we operationalise this scenario by concatenating a gesture embedding (extracted with either unimodal or multimodal-X pre-trained models)[6] with a semantic embedding derived from the transcribed co-occurring speech using BERTje (de Vries et al., 2019). The results are shown in Figure 7, where we also include a condition where the reference resolution model exclusively uses the concurrent speech in the form of a semantic embedding. In that condition, resolution accuracy is 24%. That is, the concurrent information present in the verbal modality has stronger predictive power to identify referents than body movements alone, which is to be expected in spoken conversations. Importantly, when both the verbal and gestural modalities are combined, we observe a significant increase in reference res-

---

[5]In Appendix E, we present a supplementary experiment showing that this is the case even in the presence of noisy

skeletal data, which further emphasizes the robustness of our gesture representations.

[6]Given the lack of statistically significant difference between multimodal and multimodal-X in Section 5.2 Fig. 6, we focus on the multimodal-X model for this experiment.

olution accuracy, reaching 31% when gestures are encoded with our multimodal-X model, pre-trained with text-based semantics. These findings confirm the complementary roles of gesture and speech in reference resolution and highlight the benefits of exploiting such complementarity for gesture representation learning.

5.4 Impact of Dialogue History

It is well known that, in referential communication tasks, participants tend to reuse the same referential expressions over the course of the dialogue, creating dialogue-specific conventions (Clark and Wilkes-Gibbs, 1986; Brennan and Clark, 1996). Such 'alignment' has been observed for both speech and gestures (Akamine et al., 2024). Hence, a system tasked with identifying the referent of a gesture is expected to achieve higher accuracy if it has access to other gestures previously used within the same dialogue than if such dialogue history is not available. To test whether our approach to gesture representation learning gives rise to this pattern, we train two versions of our reference resolution model: a *baseline* model and a *dialogue-specific* model. The baseline model is trained on all dialogues in CABB-S, except the target dialogue—thus, referent prediction for the gestures in the target dialogue is carried out without dialogue history. In contrast, the dialogue-specific model is progressively adapted over the dialogue rounds: *i.e.*, in round 1 it is identical to the baseline model, but by round n, it has additionally seen data from all previous dialogue rounds up to $n - 1$. To keep the amount of training data comparable between the baseline and dialogue-specific models, when new round data is added, we proportionally reduce the amount of data drawn from other dialogues during the re-training of the dialogue-specific model. As a result, both models are trained on an identical number of samples in every round.

To isolate the impact of dialogue history on gesture-driven reference resolution, in this experiment we focus on identifying referents with only gestural information as input (as in Section 5.2),[7] comparing our unimodal and multimodal-x pre-trained models. Figure 8 shows that as the conversation unfolds over the rounds, the dialogue-specific reference resolution models outperform
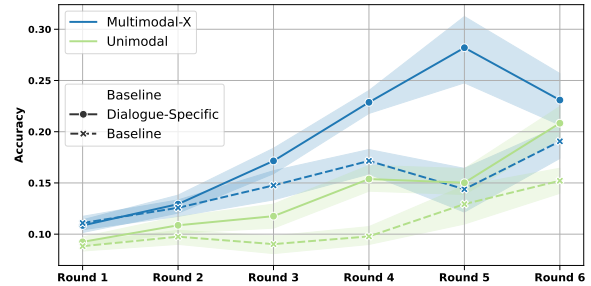

Figure 8: Average reference resolution accuracy over dialogue rounds using gesture embeddings from the unimodal (green) and multimodal-X (blue) pre-trained models. Dotted lines are used for the corresponding baseline models.

the baselines (dotted lines). Our statistical analysis shows that there is a significant difference in accuracy between the two (independent t-test yielding $t = 2.9$, $p \ll 0.05$ for both the unimodal and multimodal-x models). It also indicates that the pattern of increased accuracy, as more dialogue history becomes available, is more pronounced when the gestures are encoded with the multimodal-x pre-trained model (Spearman correlation between accuracy values and dialogue round numbers: $\rho = 0.32$ for the unimodal model and $\rho = 0.35$ for the multimodal-x model, with $p \ll 0.05$ in both instances).[8]

Overall, the results indicate that our gesture representations, particularly when learned via multimodal-x pre-training, encode features that capture the subtle increase in similarity between gestures referring to the same object within a given dialogue. In other words, to some extent the models capture gesture entrainment, which results in an advantage for the task of reference resolution. From a practical point of view, this suggests that access to dialogue history can be an asset to agents deployed with a gesture resolution model.

6 Conclusion

In this work, we have studied representational co-speech gestures in collaborative dialogue, using an existing dataset of face-to-face interactions collected by cognitive scientists. We introduced a novel reference resolution task formulated as the problem of identifying the intended referent of a co-speech gesture, while addressing key challenges in gesture representation learning. We proposed a self-supervised Transformer-based approach to learning

---

[7]The impact of dialogue history on text-based reference resolution has already been extensively studied, *e.g.*, Shore and Skantze (2018); Haber et al. (2019); Takmaz et al. (2020); Hawkins et al. (2020).

[8]Note that there is no statistically significant difference in accuracy between rounds 5 and 6, despite the apparent drop.

pre-trained gesture embeddings by jointly exploiting skeletal information and concurrent language encoded with text or speech large language models. Our experiments showed that the resulting gesture embeddings effectively contribute to reference resolution. Representing gestures by exclusively exploiting skeletal information has significant predictive power, and grounding body movements in concurrent speech during pre-training further improves resolution accuracy, even when speech is not provided at test time. An interesting avenue for future work could be to ground the reference resolution models in the visual properties of the referents, in order to learn mappings between iconic gestures and the objects they represent, which might lead to further improvements. Moreover, we showed that reference resolution from representational gestures can benefit from having access to gestures previously used within a dialogue, thus providing empirical support for the presence of gestural entrainment in face-to-face interaction.

Taken together, our findings emphasize the multimodal character of conversation (Holler and Levinson, 2019b; Özyürek, 2014) and the importance of capturing the complementarity between gestures and speech in naturalistic human-machine interaction. Further work is needed to test the extent to which the proposed pre-training approach would transfer to other referential domains and other tasks—a step we leave to future research.

Limitations

The current work focuses on Dutch-speaking task-oriented dialogues, thus contributing to linguistic diversity in the current English-centric NLP landscape. We nevertheless acknowledge that it is an open question how well the proposed models may generalise to other languages, cultural contexts, tasks, as well as open-domain dialogues. On the methodological front, while we employ and adapt a state-of-the-art motion encoder and show that our pre-training objectives and architecture choices are effective, further optimisation and integration with more advanced speech and semantic encoders may give additional improvements. Finally, our method is agnostic as to whether concurrent speech and gesture are semantically congruent (i.e., express compatible content). We leverage both information streams and observe that this yields stronger representations and higher performance. Although the nature of the collaborative referential task in the CABB datasets makes it likely that the two modalities align in content, whether the performance gains stem from true cross-modal congruence remains an open empirical question that could be explored in the future. This, however, requires manual annotation of the linguistic referential expressions, which is currently not available.

Acknowledgements

References

2024. ELAN. [Computer software].

Artem Abzaliev, Andrew Owens, and Rada Mihalcea. 2022. Towards understanding the relation between gestures and language. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5507–5520.

Sho Akamine, Esam Ghaleb, Marlou Rasenberg, Raquel Fernández, Antje Meyer, and Aslı Özyürek. 2024. Speakers align both their gestures and words not only to establish but also to maintain reference to create shared labels for novel objects in interaction. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 46.

Jason Ansel, Edward Yang, Horace He, Natalia Gimelshein, Animesh Jain, Michael Voznesensky, Bin Bao, Peter Bell, David Berard, Evgeni Burovski, Geeta Chauhan, Anjali Chourdia, Will Constable, Alban Desmaison, Zachary DeVito, Elias Ellison, Will Feng, Jiong Gong, Michael Gschwind, Brian Hirsh, Sherlock Huang, Kshiteej Kalambarkar, Laurent Kirsch, Michael Lazos, Mario Lezcano, Yanbo Liang, Jason Liang, Yinghai Lu, CK Luk, Bert Maher, Yunjie Pan, Christian Puhrsch, Matthias Reso, Mark Saroufim, Marcos Yukio Siraichi, Helen Suk, Michael Suo, Phil Tillet, Eikan Wang, Xiaodong Wang, William Wen, Shunting Zhang, Xu Zhao, Keren Zhou, Richard Zou, Ajit Mathews, Gregory Chanan, Peng Wu, and Soumith Chintala. 2024. PyTorch 2: Faster Machine Learning Through Dynamic Python Bytecode Transformation and Graph Compilation. In *29th ACM International Conference on*

Architectural Support for Programming Languages and Operating Systems, Volume 2 (ASPLOS '24). ACM.

Eléonore Arbona, Kilian G Seeber, and Marianne Gullberg. 2023. Semantically related gestures facilitate language comprehension during simultaneous interpreting. *Bilingualism: Language and cognition*, 26(2):425–439.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. 2023. Whisperx: Time-accurate speech transcription of long-form audio. *INTERSPEECH 2023*.

Susan E Brennan and Herbert H Clark. 1996. Conceptual pacts and lexical choice in conversation. *Journal of experimental psychology: Learning, memory, and cognition*, 22(6):1482.

Razvan Brinzea, Bulat Khaertdinov, and Stylianos Asteriadis. 2022. Contrastive learning with cross-modal knowledge mining for multimodal human activity recognition. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 01–08. IEEE.

Ellen Campana, Laura Silverman, Michael K Tanenhaus, Loisa Bennetto, and Stephanie Packard. 2005. Real-time integration of gesture and speech during reference resolution. In *Proceedings of the 27th annual meeting of the Cognitive Science Society*, pages 378–383.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

Yixin Chen, Qing Li, Deqian Kong, Yik Lun Kei, Song-Chun Zhu, Tao Gao, Yixin Zhu, and Siyuan Huang. 2021. Yourefit: Embodied reference understanding with language and gesture. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1385–1395.

Herbert H Clark and Deanna Wilkes-Gibbs. 1986. Referring as a collaborative process. *Cognition*, 22(1):1–39.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2020. Unsupervised cross-lingual representation learning for speech recognition. *arXiv preprint arXiv:2006.13979*.

Wietse de Vries, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim. 2019. BERTje: A Dutch BERT Model. arXiv:1912.09582.

Linda Drijvers and Asli Özyürek. 2017. Visual context enhanced: The joint contribution of iconic gestures and visible speech to degraded speech comprehension. *Journal of Speech, Language, and Hearing Research*, 60(1):212–222.

Lotte Eijk, Marlou Rasenberg, Flavia Arnese, Mark Blokpoel, Mark Dingemanse, Christian F. Doeller, Mirjam Ernestus, Judith Holler, Branka Milivojevic, Asli Özyürek, Wim Pouw, Iris van Rooij, Herbert Schriefers, Ivan Toni, James Trujillo, and Sara Bögels. 2022. The CABB dataset: A multimodal corpus of communicative interactions for behavioural and neural analyses. *NeuroImage*, 264:119734.

William Falcon and The PyTorch Lightning team. 2019. PyTorch Lightning.

Albert Gatt and Patrizia Paggio. 2013. What and where: An empirical investigation of pointing gestures and descriptions in multimodal referring actions. In *Proceedings of the 14th European Workshop on Natural Language Generation*, pages 82–91, Sofia, Bulgaria. Association for Computational Linguistics.

Esam Ghaleb, Ilya Burenko, Marlou Rasenberg, Wim Pouw, Ivan Toni, Peter Uhrig, Anna Wilson, Judith Holler, Aslı Özyürek, and Raquel Fernández. 2024a. Leveraging speech for gesture detection in multimodal communication. *arXiv:2404.14952v1*.

Esam Ghaleb, Bulat Khaertdinov, Wim Pouw, Marlou Rasenberg, Judith Holler, Asli Ozyurek, and Raquel Fernández. 2024b. Learning co-speech gesture representations in dialogue through contrastive learning: An intrinsic evaluation. In *Proceedings of the 26th International Conference on Multimodal Interaction*, pages 274–283.

Esam Ghaleb, Marlou Rasenberg, Wim Pouw, Ivan Toni, Judith Holler, Aslı Özyürek, and Raquel Fernández. 2024c. Analysing cross-speaker convergence in face-to-face dialogue through the lens of automatically detected shared linguistic constructions. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 46.

Janosch Haber, Tim Baumgärtner, Ece Takmaz, Lieke Gelderloos, Elia Bruni, and Raquel Fernández. 2019. The photobook dataset: Building common ground through visually-grounded dialogue. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1895–1910.

Robert Hawkins, Minae Kwon, Dorsa Sadigh, and Noah Goodman. 2020. Continual adaptation for efficient machine communication. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 408–419, Online. Association for Computational Linguistics.

Dan Hendrycks, Mantas Mazeika, Saurav Kadavath, and Dawn Song. 2019. Using self-supervised learning can improve model robustness and uncertainty. *Advances in neural information processing systems*, 32.

Judith Holler and Stephen C Levinson. 2019a. Multimodal language processing in human communication. *Trends in Cognitive Sciences*, 23(8):639–652.

Judith Holler and Stephen C Levinson. 2019b. Multimodal language processing in human communication. *Trends in Cognitive Sciences*, 23(8):639–652.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798.

Adam Kendon. 2004. Gesture units, gesture phrases and speech. In *Gesture: Visible Action as Utterance*, chapter 7, page 108–126. Cambridge University Press.

Casey Kennington and David Schlangen. 2017a. A simple generative model of incremental reference resolution for situated dialogue. *Computer Speech & Language*, 41:43–67.

Casey Kennington and David Schlangen. 2017b. A simple generative model of incremental reference resolution for situated dialogue. *Computer Speech & Language*, 41:43–67.

Dimosthenis Kontogiorgos, Elena Sibirtseva, Andre Pereira, Gabriel Skantze, and Joakim Gustafson. 2018. Multimodal reference resolution in collaborative assembly tasks. In *Proceedings of the 4th International Workshop on Multimodal Analyses Enabling Artificial Agents in Human-Machine Interaction*, pages 38–42.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73.

Kenneth Lai, Richard Brutti, Lucia Donatelli, and James Pustejovsky. 2024. Encoding gesture in multimodal dialogue: Creating a corpus of multimodal amr. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5806–5818.

Alex Lascarides and Matthew Stone. 2009. A formal semantic analysis of gesture. *Journal of Semantics*, 26(4):393–449.

Dong Won Lee, Chaitanya Ahuja, and Louis-Philippe Morency. 2021. Crossmodal clustered contrastive learning: Grounding of spoken language to gesture. In *Companion Publication of the 2021 International Conference on Multimodal Interaction*, pages 202–210.

Jinfu Liu, Chen Chen, and Mengyuan Liu. 2024. Multi-modality co-learning for efficient skeleton-based action recognition. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 4909–4918.

Seungwhan Moon, Satwik Kottur, Paul A Crook, Ankita De, Shivani Poddar, Theodore Levin, David Whitney, Daniel Difranco, Ahmad Beirami, Eunjoon Cho, et al. 2020. Situated and interactive multimodal conversations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1103–1121.

Aslı Özyürek. 2014. Hearing and seeing meaning in speech and gesture: Insights from brain and behaviour. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1651):20130296.

Leonardo Pepino, Pablo Riera, and Luciana Ferrer. 2021. Emotion recognition from speech using wav2vec 2.0 embeddings. *Proc. Interspeech 2021*, pages 3400–3404.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Marlou Rasenberg, Asli Özyürek, Sara Bögels, and Mark Dingemanse. 2022. The primacy of multimodal alignment in converging on shared symbols for novel referents. *Discourse Processes*, 59(3):209–236.

Todd Shore and Gabriel Skantze. 2018. Using lexical alignment and referring ability to address data sparsity in situated dialog reference resolution. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2288–2297.

Arabella J Sinclair and Bertrand Schneider. 2021. Linguistic and gestural coordination: Do learners converge in collaborative dialogue?. *International Educational Data Mining Society*.

Gabriel Skantze and Bram Willemsen. 2022. Collie: Continual learning of language grounding from language-image embeddings. *Journal of Artificial Intelligence Research*, 74:1201–1223.

Ece Takmaz, Mario Giulianelli, Sandro Pezzelle, Arabella Sinclair, and Raquel Fernández. 2020. Refer, reuse, reduce: Generating subsequent references in visual and conversational contexts. In *Proceedings*

of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), page 4350. Association for Computational Linguistics.

Fida Mohammad Thoker, Hazel Doughty, and Cees GM Snoek. 2021. Skeleton-contrastive 3d action representation learning. In *Proceedings of the 29th ACM international conference on multimedia*, pages 1655–1663.

Hsiang-Sheng Tsai, Heng-Jui Chang, Wen-Chin Huang, Zili Huang, Kushal Lakhotia, Shu-wen Yang, Shuyan Dong, Andy Liu, Cheng-I Lai, Jiatong Shi, Xuankai Chang, Phil Hall, Hsuan-Jui Chen, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed, and Hung-yi Lee. 2022. SUPERB-SG: Enhanced speech processing universal PERformance benchmark for semantic and generative capabilities. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8479–8492, Dublin, Ireland. Association for Computational Linguistics.

Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. 2022. Vitpose: Simple vision transformer baselines for human pose estimation. *Advances in neural information processing systems*, 35:38571–38584.

Salah Zaiem, Youcef Kemiche, Titouan Parcollet, Slim Essid, and Mirco Ravanelli. 2025. Speech self-supervised representations benchmarking: a case for larger probing heads. *Computer Speech & Language*, 89:101695.

Wentao Zhu, Xiaoxuan Ma, Zhaoyang Liu, Libin Liu, Wayne Wu, and Yizhou Wang. 2023. Motionbert: A unified perspective on learning human motion representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15085–15099.

Appendix

A Objects in the CABB Dataset

Rasenberg et al. (2022) segmented each gesture stroke in the CABB-S dataset and classified them into four categories: *iconic* (depicting an aspect of the target object), *deictic* (explicitly pointing with an extended finger or hand), *other* (e.g., beat gestures or pragmatic signals like "you go ahead"), and *undecided* when the gesture type was unclear. In our study, we focus exclusively on the iconic gestures, which refer to specific parts of the novel objects in the CABB setup (see Figure 9). The average number of gestures in each round per dialogue is shown in Figure 10. Each iconic gesture was annotated with a sub-part label—such as 06A for a single sub-part or 06A+06B for multiple sub-parts. When a gesture was annotated with multiple parts, we split it into separate samples corresponding to

each sub-part. Additionally, a main label was assigned if the gesture targeted the object's main part, and general is used when the gesture indicates a broad area (e.g., "the left side").

B Segmentation Qualitative Results

Figure 11 compares the duration distributions of manually segmented gestures in CABB-S with those automatically segmented in CABB-L. We can see that both distribution curves peak between 0.8 and 0.9 seconds and share a right skew. The automatic segmentation shows a small portion of gestures lasting longer than two seconds. This is most likely because the segmentation model was trained on extended segments to account for the preparation and retraction phases of gestures. Figure 12 plots the average number of gestures per pair across the six rounds of the referential game. Consistent with Figure 10, gesture frequency is highest in the initial rounds and declines in later ones. The higher gesture count in CABB-L reflects longer interaction time: besides the referential interaction, these pairs also carried out an on-screen object localisation task that required identifying the object's position with respect to the others. We also imported the automatically segmented gestures into ELAN (ela, 2024) for quality check. As shown in Figure 13, we could visually inspect and verify that gestures' onsets were detected and segmented with high reliability.

C Model Details

C.1 Pre-training objectives

In Section 4, we introduced three pre-training architectures, each containing a combination of self-supervised learning objectives. Here, we provide a detailed technical overview of these losses. The proposed architectures exploit three modalities, namely 2D skeletal keypoints and joint prediction confidence scores $X_i^g \in \mathbb{R}^{T_g \times 27 \times 3}$ for gestures, text-based semantics X_i^t, and raw speech signals X_i^s. These inputs are encoded using one of the following encoders: our adaptation of DSTFormer $f_{\Theta_g}(\cdot)$ for skeletons (Section 4.1; Zhu et al. (2023)), Dutch BERT-based model $f_{\Theta_t}(\cdot)$ for text (de Vries et al., 2019), and wav2vec2-xlsr-300 $f_{\Theta_s}(\cdot)$ for raw speech (Conneau et al., 2020).

Unimodal masked reconstruction loss. We follow the original DSTFormer (Zhu et al., 2023) by randomly masking portions of the 2D keypoint in-
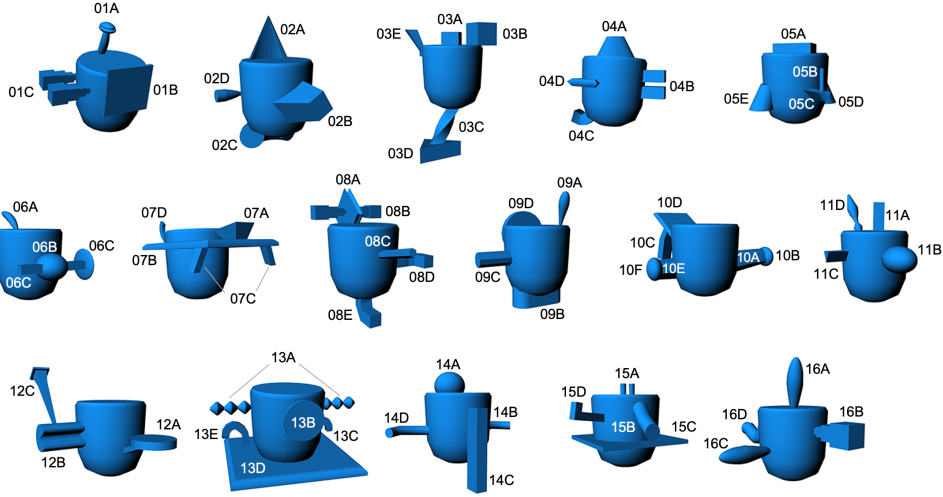
Figure 9: The candidate objects and their sub-parts present in the CABB dataset (Eijk et al., 2022).
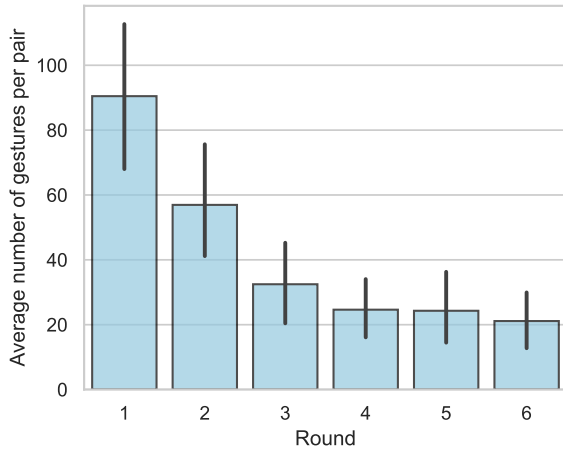

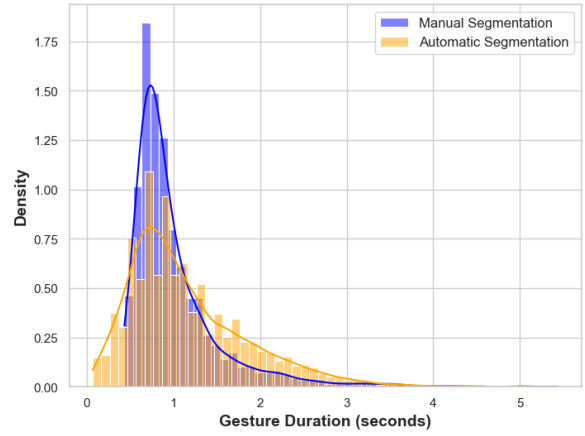Figure 10: Distribution of manually segmented gestures across rounds of interaction.

Figure 11: Durations distribution of manually and automatically segmented gestures.

put and learning to reconstruct them. Specifically, we accumulate a reconstruction loss between the original and predicted coordinates of masked keypoints as follows:

$$\mathcal{L}_k = \sum_{t=1}^{T} \sum_{j=1}^{J} \delta_{t,j} ||\hat{\boldsymbol{x}}_{t,j} - \boldsymbol{x}_{t,j}||^2, \qquad (1)$$

where $\hat{\boldsymbol{x}}_{t,j} \in \mathbb{R}^2$ is the predicted coordinates of keypoint j at timestep t, $\boldsymbol{x}_{t,j}$ is the ground truth keypoint, and $\delta_{t,j}$ is a weighting factor that accounts for confidence or visibility of the keypoints. To enforce spatial and temporal consistency, we introduce two additional reconstruction objectives $\mathcal{L}_b$ and $\mathcal{L}_m$: bone and motion reconstruction losses. The former ensures structural consistency by preserving the distances between adjacent keypoints $||\boldsymbol{x}_{t,j} - \boldsymbol{x}_{t,j-1}||$ across frames, while the latter mini-

mizes the difference between the temporal displacement $||\boldsymbol{x}_{t,j} - \boldsymbol{x}_{t-1,j}||$ of predicted and ground truth keypoints. The overall objective for masked reconstruction is given by the average of $\mathcal{L}_k$, $\mathcal{L}_b$ and $\mathcal{L}_m$. Figure 14 illustrates how this objective is integrated into the unimodal pre-training architecture.

Unimodal contrastive loss. A unimodal contrastive loss is applied to different views of the same skeletal keypoint sequence distorted with simple augmentations, as illustrated in the middle and right branches of Figure 14. Formally, for input skeleton $\boldsymbol{X}_i^g$, we obtain two augmented views $a(\boldsymbol{X}_i^g)$ and $a'(\boldsymbol{X}_i^g)$. These views are then passed through skeleton encoder $f_{\Theta_g}(\cdot)$, namely DSTFormer, and projection layers $g_{\Theta_g}(\cdot)$ to obtain projected features $\boldsymbol{z}_i^g = g_{\Theta_g}(f_{\Theta_g}(a(\boldsymbol{X}_i^g)))$ and $\boldsymbol{z}_j^g = g_{\Theta_g}(f_{\Theta_g}(a'(\boldsymbol{X}_i^g)))$. These representations are treated as a positive pair in a contrastive loss
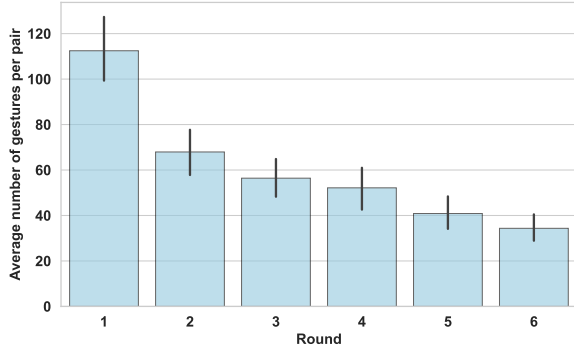
Figure 12: Distribution of automatically segmented gestures across rounds of interaction.
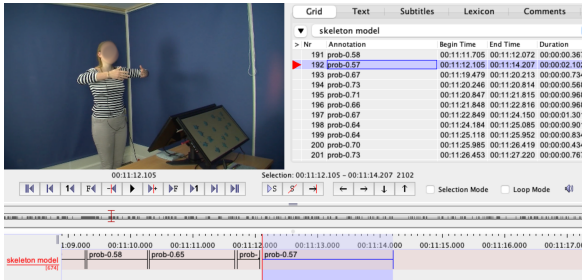

Figure 13: Screenshot of ELAN (ela, 2024), a popular annotation software, which we used to inspect the quality of automatic segmentation.

function, whereas all other views from a training mini-batch are considered negative (Chen et al., 2020):

$$l_{um}(i,j) = -log \frac{exp(\frac{s(\boldsymbol{z}_i^g, \boldsymbol{z}_j^g)}{\tau})}{\sum_{k=1}^{2b} \mathbb{I}_{[k \neq i]} exp(\frac{s(\boldsymbol{z}_i^g, \boldsymbol{z}_k^g)}{\tau})}. \quad (2)$$

The loss maximizes cosine similarity $s(\cdot)$ for the positive pair and minimizes similarity with other views in a mini-batch of size b.

Multimodal contrastive loss. We propose a CLIP-like contrastive objective (depicted with a blue line in Figure 2) mapping global representations of skeletons and co-occurring utterances in a joint feature space (Radford et al., 2021). In detail, given projected representations of skeletons $\boldsymbol{z}_i^g = g_{\Theta_g}(f_{\Theta_g}(\boldsymbol{X}_i^g))$ and co-occurring utterances (*e.g.*, text-based semantics) $\boldsymbol{z}_i^t = g_{\Theta_t}(f_{\Theta_t}(\boldsymbol{X}_i^t))$, the multimodal objective aims to maximize their similarity as follows:

$$l_{mm}^{g \to t}(i) = -log \frac{exp(\frac{s(\boldsymbol{z}_i^g, \boldsymbol{z}_i^t)}{\tau})}{\sum_{k=1}^{b} exp(\frac{s(\boldsymbol{z}_i^g, \boldsymbol{z}_k^t)}{\tau})}. \quad (3)$$

The final error function accumulates losses $l_{mm}^{g \to t}$ and $l_{mm}^{t \to g}$ for each skeleton-utterance pair in a mini-batch.

Multimodal-X losses. Two losses are employed to optimize the multimodal-X architecture (Figure 2). First, the contrastive loss is computed between skeleton representations and pooled semantic embeddings in line with Equation 3. Furthermore, we introduce an objective that leverages cross-attention in our adapted DSTFormer (Section 4.1). Specifically, the representations of skeletons in one branch of the architecture are fused with semantic or speech embeddings via cross-attention layers (middle branch in Figure 2), while the other branch remains unimodal (right branch in the figure). We then apply the same contrastive formulation (Equation 2) to align unimodal skeleton representations with the fused skeleton–crossmodal embeddings. This strategy encourages both robust unimodal representations and cross-modal alignment.

C.2 Handling mismatched temporal resolutions.

Multimodal data has different resolutions. Skeletal input and speech raw waveform come with high-frequency motion frames, whereas linguistic information is tokenised into subword units of much coarser granularity. We experiment with two integration strategies.

1. **Multimodal (mean-pool).** All subword embeddings in an utterance or speech waveform representations of 25 milliseconds are mean-pooled into a single vector. Likewise, the T gesture-frame embeddings are concatenated and projected down to one vector. The two global representations are then fused by the cross-modal Transformer.

2. **Multimodal-X (frame-wise attention).** Each gesture frame ($\mathbf{q}_t$) attends to the set of subword embeddings ($\mathbf{k}_j, \mathbf{v}_j$) for the co-occurring utterance through multi-head attention. We apply the same mechanisms when we handle speech frame-level representations, which operate at segments of 25 milliseconds. This enables more fine-grained alignment, as each frame can focus on the specific linguistic context it co-occurs with.
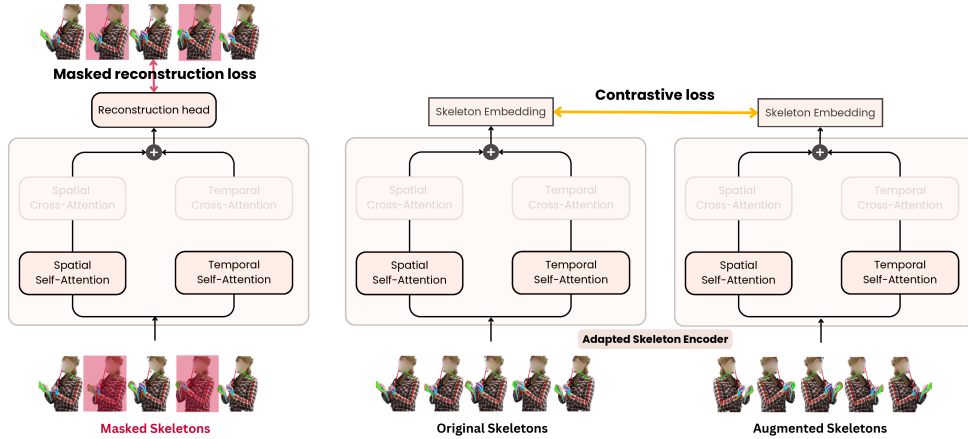
Figure 14: Unimodal architecture jointly optimizes masked reconstruction (left branch) and unimodal contrastive losses (middle and right branches). Note that cross-attention blocks are not utilized in this pre-training approach.

	Unimodal	Multimodal-X		Multimodal	
		Semantic	Speech	Semantic	Speech
Skeleton Encoders params					
+ Projection heads (M)	10.3	22.0	24.2	10.5	10.5
Non-trainable params (M)	–	109	315	109	315
Total params (M)	10.3	131	339	119	325
Model size (GB)	0.041	0.525	1.359	0.478	1.304

Table 1: Parameters for the three architectures. The skeleton encoders (the adapted DSTFormer) and the projection heads are the trainable parameters. The speech (i.e., wav2vec2) and text-based semantic (i.e., BERTje) encoders are frozen during pre-training.

D Implementation Details

All models are implemented using PyTorch (Ansel et al., 2024) and PyTorch-Lightning (Falcon and The PyTorch Lightning team, 2019). Training is performed on nodes with four NVIDIA A100-SXM4-40GB GPUs. The experimented three model types—unimodal, multimodal, and multimodal-X—each trained using its respective objective for a maximum of 100 epochs. We use Adam optimizer with a learning rate of 0.001. For multimodal-X, a per-GPU batch size of 96 (for a total of 384 across four GPUs) strikes a balance between VRAM utilization and achieving reliable convergence. For multimodal models, we could only fit a batch size of 64 per GPU due to using the additional model for masked reconstruction. We also used a batch size of 64 for the unimodal models. For the contrastive objective, we set the temperature to 0.1 by default. Masked reconstruction follows the DSTFormer (Zhu et al., 2023) configuration with a masking probability of 5% and an equivalent amount of noise applied. We randomly split 90% of our generated time windows in CABB-XL for pre-training and use the remaining

10% for validation. The CABB-S dataset is reserved solely to monitor pre-training convergence and agreement with expert annotations; we select final checkpoints based on these performance metrics. Throughout model training, we employ data augmentations similar to those proposed by Ghaleb et al. (2024c). Namely, we apply various skeletal transformations (mirror, shift, scaling, rotation, jitter, shear) to ensure the models generalize to pose variability.

D.1 Implementation of the skeleton encoder

The adapted DSTFormer encoder processes skeletal data with two parallel branches that separately attend to spatial and temporal features. In its unimodal version (without cross-attention), the encoder consists of 4 blocks per branch (8 blocks overall), each containing standard attention and MLP layers with residual connections. The encoder's output is then fed into a projection head—an MLP that maps the encoded features (e.g., from 256 to 128 dimensions)—to produce the final feature representation.

In the multimodal-x variant, each block contains an additional cross-attention module that fuses ei-

Model variant	σ	Accuracy (mean $\pm$ s.d.)
Multimodal-X (clean)	0.00	0.19 $\pm$ 0.05
Multimodal-X (jittered)	0.20	0.18 $\pm$ 0.04
	1.00	0.19 $\pm$ 0.05
	15.0	0.16 $\pm$ 0.04

Table 2: Reference resolution accuracy using the multimodal-X architecture with skeletal information including different degrees of noise (jittered).

ther text-based semantic embeddings or speech embeddings with the skeleton representation. This extension doubles the number of trainable parameters, resulting in about 22.0 million parameters when using semantic inputs and 24.2 million when using speech inputs. The difference in parameter counts is due to the different embedding sizes from the semantic (768 dimensions) and speech (1024 dimensions) backbones and the subsequent projection heads. Similar to the unimodal case, dedicated projection heads then map the features to the shared embedding space. The multimodal-x model pre-trained with text-based semantics takes approximately 15 hours to run, while the multimodal model runs for roughly 12 hours. The unimodal one requires considerably less time (around 8 hours for 100 epochs) since it does not rely on the backbone models of co-occurring speech.

In Table 1, we summarize the number of parameters for each model architecture.

E Impact of Errors on Pose Estimation

There may be errors in pose estimation because we rely on off-the-shelf ViTPose (Xu et al., 2022) and, in the CABB dataset set-up, some joints are occasionally occluded. Several design choices were therefore made to make the models' representations robust. First, we employ self-supervised objectives, which are known to decrease the effect of noisy inputs (Hendrycks et al., 2019). Second, we feed the per-joint confidence scores returned by ViTPose as an additional channel in the skeletal input (see Appendix C), so the model can learn to down-weight unreliable joints.

Additional noise experiment. To test the models' resilience to noisy data, we conduct an additional experiment by adding Gaussian noise with varying jitter (σ) into the 2D skeleton coordinates. Table 2 shows the reference resolution accuracies achieved by the skeleton encoder pre-trained with multimodal-X when different degrees of noise are added. The results show that our model is reasonably robust and only at very high noise levels ($\sigma > 10$, which exceeds the typical error rate of current pose estimators) does performance drop slightly.