META4: semantically-aligned generation of metaphoric gestures using self-supervised text and speech representation

Mireille Fares, Catherine Pelachaud, Nicolas Obin

META4: SEMANTICALLY-ALIGNED GENERATION OF METAPHORIC GESTURES USING SELF-SUPERVISED TEXT AND SPEECH REPRESENTATIONS

Mireille Fares
ISIR, STMS
Sorbonne University
Paris, France

Catherine Pelachaud
ISIR, CNRS
Sorbonne University
Paris, France

Nicolas Obin
STMS
Sorbonne University
Paris, France

ABSTRACT

Image Schemas are repetitive cognitive patterns that influence the way we conceptualize and reason about various concepts present in speech. These patterns are deeply embedded within our cognitive processes and are reflected in our bodily expressions including gestures. Particularly, metaphoric gestures possess essential characteristics and semantic meanings that align with Image Schemas, to visually represent abstract concepts. The shape and form of gestures can convey abstract concepts, such as extending one's forearm and hand or tracing a line with hand movements to visually represent the image schema of "PATH". Previous behavior generation models have primarily focused on utilizing speech (acoustic features and text) to drive the generation model of virtual agents. They have not considered key semantic information as those carried by Image Schemas to effectively generate metaphoric gestures. To address this limitation, we introduce *META4*, a deep-learning approach that generates metaphoric gestures from both speech and Image Schemas. Our approach has two primary goals: (1) computing Image Schemas from input text to capture the underlying semantic and metaphorical meaning, and (2) generating metaphoric gestures driven by speech and the computed image schemas. We make two key contributions: (1) BERTIS, a model that computes Image Schema tags from text input, and (2) META4, which makes use of BERTIS to model and synthesize the corresponding metaphoric gestures from speech and the generated Image Schemas. To the best of our knowledge, our approach is the first method for generating speech-driven metaphoric gestures while leveraging the potential of Image Schemas. We present evaluation studies to demonstrate the effectiveness of our approach and highlight the importance of both speech and image schemas in modeling metaphoric gestures. Link to code, data and videos: *https://github.com/mireillefares/META4/*.

1 Introduction

In the realm of human communication, our cognitive processes play a vital role in shaping the way we conceptualize and reason about various ideas and concepts. One significant aspect of this cognitive influence is manifested through repetitive patterns known as "*Image Schemas*". These patterns, as described by Johnson et al. [Johnson(2005)], exert a profound impact on our understanding and expression of concepts within speech. They are deeply ingrained in our cognitive processes and find expression in our bodily movements and gestures, as observed by Cienki [Cienki and Müller(2008)]. Of particular interest are *metaphoric gestures*, which are gestures that symbolically represent a concept, object, or event [McNeill and Levy(1982)]. These gestures possess inherent characteristics and semantic meanings that align with *Image Schemas*. They serve as visual representations of abstract concepts [Kendon(2004)], employing specific shapes and forms to convey complex ideas. The conveyed image is a visual representation that is associated with something concrete and actionable in the world. For example, one can sweep his/her flat hand through space, or trace a surface, to visually represent the image schema "SURFACE". When a speaker discusses the promotion of an individual in an organization, he/she may use a metaphoric gesture such as raising his/her hand upward to symbolize the promotion. Despite the significant role that Image Schemas and metaphoric gestures play in conveying abstract concepts, previous gesture generation models [Ravenet et al.(2018), Kucherenko et al.(2020),

Yoon et al.(2019), Yoon et al.(2020), Fares(2020a), Fares et al.(2022a)] for virtual agents have primarily focused on using speech acoustic features alone, or in combination with text semantics, to drive the gesture generation process. However, these generative models do not compute and capture the gesture shapes, nor make use of the rich semantic information conveyed by Image Schemas [Grady(2005)] to synthesize metaphoric gestures.

To address this limitation, we leverage recently developed self-supervised representations in text (BERT [Devlin et al.(2019)]) and audio (AST [Gong et al.(2021)]) to enhance the input representations for multimodal neural networks for gesture synthesis, including metaphorical gestures. More specifically, we propose *META4* (short for *METAPHOR*), a deep-learning approach that generates metaphoric gestures using both speech and Image Schemas. Our approach encompasses two primary objectives: (1) computing Image Schemas from input text to capture the underlying semantic and metaphoric meaning, and (2) generating metaphoric gestures driven by both speech and the computed Image Schemas. Motivated by these objectives, we make two key contributions. First, we present *BERTIS*, a model designed to compute Image Schema tags from textual input. Second, we propose *META4*, which builds upon *BERTIS* to model and synthesize metaphoric gestures based on speech input and the generated *Image Schemas*.

To the best of our knowledge, our approach represents the first method for generating speech-driven metaphoric gestures while leveraging the potential of *Image Schemas*. We conduct evaluation studies to demonstrate the effectiveness of our approach and show the importance of considering both *speech* and *Image Schemas* in modeling metaphoric gestures. Our contributions can be listed as follows:

1. We utilize the advancements in self-supervised representations in text (BERT [Devlin et al.(2019)]) and audio (AST [Gong et al.(2021)]) to enhance the input representations of multimodal neural networks for synthesizing gestures including metaphorical gestures.
2. We propose ***META4***, a novel speech-driven and semantically-aligned multimodal Transformer-based approach that combines speech and Image Schemas modalities to generate metaphoric gestures.
3. We propose ***BERTIS***, an Image Schema computational model that classifies an input text into an Image Schema class.

This paper is organised as follows. We start by discussing some background and a review of the existing gesture generation approaches works. We then explain and introduce the architecture of ***META4*** and ***BERTIS***. Finally, we present objective evaluations and discuss the results.

2 Background and Related works

2.1 Image Schemas

The concept of Image Schema was first introduced in 1987 in Johnson's book "The body in the mind"[Johnson(2013)]. The original definition was set as follows: "An image schema is a recurring, dynamic pattern of our perceptual interactions and motor programs that gives coherence and structure to our experience" [Johnson(2013)]. The idea is to define a parallel between how the body perceives and acts, and how the mind understands and knows; implying that sensory-motor capacities are recruited for abstract thinking [Johnson(2005)]. Such patterns regroup all types of cues accumulated since childhood that are then flexibly categorized by language, allowing to put words on general and metaphoric concepts [Langacker(1987)]. To illustrate Image Schemas, let us take an example: CONTAINMENT that can be found in multiple expressions, such as "Putting an idea inside someone's head". CONTAINMENT emerges structurally from the bodily experience of either seeing something going into something else, putting something into something else, or going oneself into something. Studies in cognitive linguistics suggest that over two dozen different image schemas appear regularly in people's everyday thinking, reasoning, and imagination [Johnson(2013), Lakoff(1987), Gibbs Jr and Colston(2006)]. For example, Cienki [Cienki(2005)] suggests that the image schemas CONTAINER, CYCLE, FORCE, OBJECT, PATH are reliably used to categorize gestures observed from natural conversations [Cienki(2013)].

Several studies have highlighted the links between image schemas and gestures. Gestures and speech arise from a common cognitive process [Kendon(2004)]. Gestures embody thoughts. They do not always convey the same information; they may complement each other, or even one may substitute for the other one. Gesture features, such as hand shape or wrist motion, are commonly found to illustrate image schemas. For example, the image schema PATH is often marked by the linear trajectory of the hand [Williams(2008), Cienki(2013)], while the CYCLE Image Schema can be aligned with a repeated circular movement of the hand [Ladewig(2011)]. Image schemas can also be combined. For example, a container can be lifted, filled in, or put aside; it can be augmented with adjectives (small, hard, large, etc) that will be reflected in the gesture shape and trajectory [Antonova(2020)]

2.2 Gesture Generation Approaches

Previously, there have been various approaches proposed for gesture generation. The earliest approaches[Pelachaud et al.(1996), Cassell et al.(2001), Pelachaud et al.(2002), Kopp et al.(2006)] were rule-based, relying on predefined correspondences between patterns of human communication and behavior. However, these rule-based approaches have limitations in terms of requiring significant human effort to determine the rules, resulting in limited and repetitive gestures.

To overcome these limitations, researchers turned to statistical approaches [Kipp(2005), Kipp(2001), Neff et al.(2008), Bergmann and Kopp(2009), Marsella et al.(2013)], which synthesize gestures based on statistics from a corpus of human non-verbal behavior. While statistical models mitigated some of the limitations of rule-based approaches, they still suffered from a lack of diversity and variability in the generated gestures.

More recently, learning-based models, often referred to as data-driven models, have been proposed. These models are trained on large amounts of data and utilize machine learning algorithms. Gesture generative models based on sequential generative parametric models like Hidden Markov Models (HMMs) [Hofer and Shimodaira(2007)], Recurrent Neural Networks (RNNs) [Wang et al.(2021), Haag and Shimodaira(2016)], and Dynamic Bayesian Networks (DBNs) [Mariooryad and Busso(2012), Sadoughi and Busso(2019)] have been used to generate head motion from speech. Generative Adversarial Networks (GANs) [Karras et al.(2017), Vougioukas et al.(2019)] have also been employed to produce facial gestures from speech.

However, a common limitation of many of these works is that they primarily rely on a single modality of human communication, typically speech, as input. Some approaches use text transcriptions of language to synthesize gestures, but gestures are influenced by both speech prosody and language, including facial, hand, and body gestures. For instance, Ishi et al. [Ishi et al.(2018)] proposed a text-based approach for generating gestures to control a humanoid robot. Their method translated text into gesture motions by associating words with concepts, concepts with gesture categories (iconic, metaphoric, deictic, beat, emblem, and adapter), and finally, gesture categories with specific gesture motions. Other works [Chiu et al.(2015), Kucherenko et al.(2020), Yoon et al.(2020), Ahuja et al.(2020a), Fares(2020b), Fares et al.(2022b), Fares et al.(2023)] have attempted to fuse both modalities, incorporating speech prosody and language, for gesture synthesis. While such approaches capture well the rhythm and fluidity of gesture motion they lack of semantic expressivity.

An earlier work by Ravenet et al. [Ravenet et al.(2018)] developed an approach for generating metaphoric gestures by extracting metaphorical properties from input speech using Behavior Markup Language (BML) [Kopp et al.(2006)]. Their method involved extracting image schemas using WordNet. Then speech audio and gestures are synchronized through BML annotations; BML configures as well various aspects of gestures, such as hand shape, movement, and wrist orientation. Such a method allows conveying effectively the intended representational meaning during behavior realization. However this model was not trained on real data.

Thus, a common limitation of the latest works is that they do not compute and capture the shape of gestures nor utilize the rich semantic information conveyed by Image Schemas [Grady(2005)] to synthesize metaphoric gestures. Our aim is to cover this gap.

3 META4: Metaphoric Gesture Generation

In this work, and based on existing neural architectures for generating gestures from speech, we focus on more complex tasks such as generating metaphorical gestures. We introduce ***META4***, a speech-driven and semantically-aligned multimodal Transformer-based approach that combines speech and Image Schemas modalities to generate communicative gestures paying attention to their shape and motion. ***META4*** is designed as a multimodal system that leverages both speech and Image Schemas modalities to capture the underlying semantic and metaphorical meaning of the input speech and text more effectively. The Image Schemas modality is computed from the input text by employing ***BERTIS***, a pre-trained Image Schema computation model that we have independently developed and trained. Figure 1 illustrates the overall architecture of ***META4***.

3.1 Problem Positioning

The goal is to generate the upper body gestures, represented by 2D poses, by utilizing both speech and Image Schemas. More specifically, we propose a novel Transformer-based architecture that effectively captures the interplay between speech, Image Schemas, and gestures. Our approach is based on the following hypothesis:

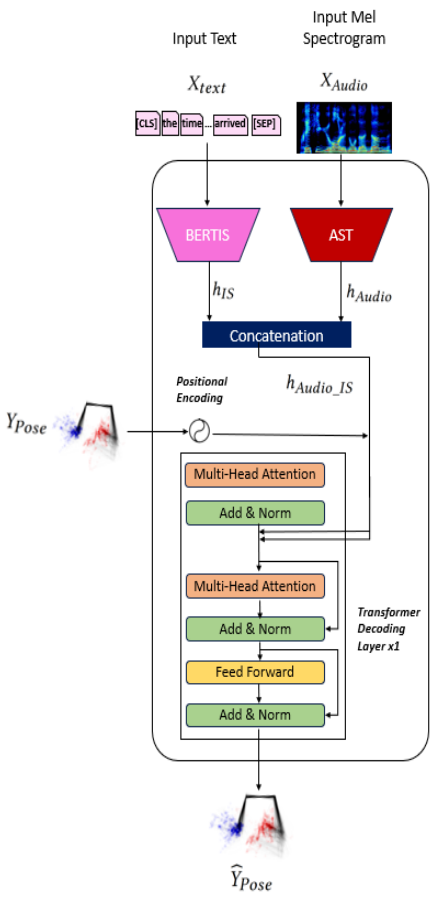

Figure 1: Overview of ***META4***, a speech-driven and semantically-aligned multimodal approach for metaphoric gesture synthesis. ***META4*** exploits both speech and image schema modalities as inputs, to capture the semantic meaning in speech and Image Schemas.

1. Image Schemas are powerful conveyors of semantic information and metaphorical meaning, which share common meanings with metaphoric gestures. Complex and abstract concepts can be conveyed by employing specific gestural shapes and forms which are directly linked to the characteristics and semantic meanings conveyed by Image Schemas.

2. By jointly modeling Image Schemas and speech acoustic features, we can enhance the generation of metaphoric gestures by capturing more precisely gestures' shape and motion, allowing a better representation of the relationships between acoustic, metaphorical, and semantic features.

To implement these assumptions, we propose an architecture for encoding a speaker's speech represented by his/her audio and text information and synthesizing the corresponding upper body gestures, including metaphoric gestures. Our approach includes two primary objectives:

1. Computing Image Schemas from input text to capture the underlying semantic and metaphorical meaning.

2. Generating metaphoric gestures driven by both speech and the computed Image Schemas.

The network processes input speech and text data at the segment level ***S***, where each segment ***S*** consists of 64 frames representing 4.266 seconds that also includes silence. For each ***S*** the network takes as input the speaker's Mel spectrogram (X_{audio}) and the corresponding text (X_{text}, sequence of words corresponding to the segment ***S***). For each ***S***, the output of the network is the generation of the corresponding upper-body gestures represented by 2D poses ($\widehat{Y}_{Pose}$).

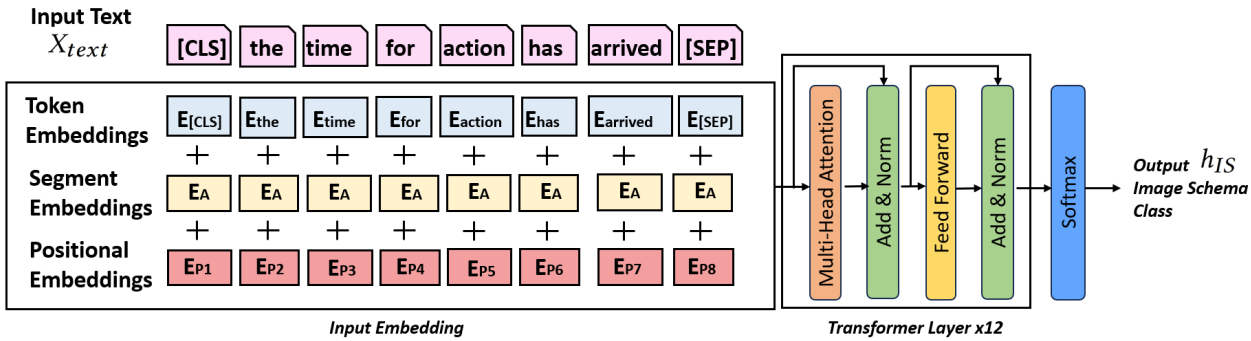

Figure 2: Overview of **BERTIS** architecture, a model for Image Schema Computation. We fine-tune the *BERT Base Cased model* to classify the input text X_{text} into an Image Schema class h_{IS}.

3.2 Neural Formulation

In the subsequent subsections, we provide a detailed explanation of each module in ***META4***.

3.2.1 Background - BERT

BERT (Bidirectional Encoder Representations from Transformers) [Kenton and Toutanova(2019)] is a language representation model that has significantly advanced the field of natural language processing. It is built on the Transformer encoder [Vaswani et al.(2017)] architecture, allowing it to capture contextual information from both the preceding and following words or tokens. In addition to standard tokens, BERT incorporates special tokens like [CLS] (classification) and [SEP] (separator) for specific purposes. The [CLS] token represents the overall representation of the input sequence and is commonly used for classification tasks. It enables BERT to generate a fixed-size representation that encapsulates the entire meaning of the input. The [SEP] token is used to separate different sentences within a single input sequence, facilitating sentence boundary understanding and inter-sentence relationship modeling. To construct the input representation, BERT sums three types of embeddings: token embeddings, segment embeddings, and position embeddings. Token embeddings capture the semantic meaning of individual words or tokens, segment embeddings differentiate between different segments or sentences, and position embeddings provide positional information within the input sequence. BERT comes in two versions: *BERT base* and *BERT large*. In this study, we utilize *BERT base*, which consists of 12 Transformer blocks, 12 attention heads, and a hidden layer size of 768.

3.2.2 Background - BERT Fine-Tuning

During the fine-tuning process of BERT, the parameters of the Transformer blocks, attention heads, and hidden layers, along with additional task-specific layers, are fine-tuned end-to-end. This allows the model to adapt its learned representations to the specific downstream task, leveraging both the pre-training on large-scale language data and the task-specific data during fine-tuning.

3.2.3 Image Schema Computation Model ($BERTIS$)

To compute the Image Schema class h^i_{IS} for a given input text X_{text}, we employ *BERTIS*, a pre-trained model that we developed and trained independently from *META4*. Specifically, we fine-tune the *BERT Base Cased model* [Devlin et al.(2018)] for the task of classifying the input text X_{text} into an Image Schema class h^i_{IS}, where i represents the class label. The set of image schema labels considered in this study consists of 14 classes, as originally proposed by [Wachowiak and Gromann(2022)]. We refer the readers to the appendix for a comprehensive list of these Image Schemas. The fine-tuning is done by adding a fully connected dense layer on top of the output layer of *BERT* and re-training the entire model using 80% of the Image Schema Corpus introduced by [Wachowiak and Gromann(2022)]. 10% of this corpus was used for validation, and 10% for testing. As illustrated in Figure 2, *BERTIS* takes the input text X_{text} corresponding to the segment level ***S***. It computes an input embedding by adding up the token, segment, and position embeddings of each token. The computed input embedding is then given as input to 12 Transformer layers. A SoftMax activation is added on top of the model to predict the likelihood of the output class h^i_{IS}, which can therefore be written as follows:

$$h^i_{IS} = E_{BERTIS}(X_{text}) \tag{1}$$

3.2.4 Audio Encoder (E_{audio})

The speech modality is encoded using E_{audio}, the pre-trained Audio Spectrogram Transformer (AST) *base384* model proposed by Gong et al. [Gong et al.(2021)]. AST operates by first splitting the input Mel spectrogram X_{audio}, which has 128 frequency bins, into a sequence of $16x16$ patches with overlap. These patches are then linearly projected into a sequence of 1D patch vectors, which are augmented with positional embeddings. A special [CLS] token is appended to this sequence. The resulting sequence is then input to a Transformer Encoder. Originally designed for audio classification, we modify the AST for our purposes by removing the linear layer with a sigmoid activation function at the output of the Transformer Encoder, as we do not require a classification task. Instead, we use the output of the Transformer Encoder's [CLS] token as the representation of the Mel spectrogram. The Transformer Encoder in our modified AST has an embedding dimension d_{model} equals to 64, 12 encoding layers, and 12 attention heads. The output vector h_{audio} can therefore be written as follows:

$$h_{audio} = E_{audio}(X_{audio}) \tag{2}$$

3.2.5 2D Upper Body Gesture Generator (G_{Pose})

The generated audio and image schema encoding vectors h_{audio} and h^i_{IS} are then concatenated together. The resulting vector h_{Audio_IS} can therefore be written as follows:

$$h_{Audio\_IS} = [E_{audio}(X_{audio}), E_{BERTIS}(X_{text})] \tag{3}$$

The vector h_{Audio_IS} is subsequently provided as input to a Transformer decoder with a single decoding layer. To incorporate positional information, a positional encoding function is applied to the ground truth sequence of upper-body gestures $\widehat{Y}_{Pose}$, which is then fed into the Transformer decoder during training time. The Transformer decoder generates a probability distribution over the sequence of upper-body gestures $\widehat{Y}_{Pose}$ that corresponds to the segment ***S***. The resulting output vector $\widehat{Y}_{Pose}$ can therefore be written as follows:

$$\widehat{Y}_{Pose} = G_{pose}(h_{Audio\_IS}) \tag{4}$$

4 Experimental Evaluations

We describe the experimental evaluations in this section, and more specifically the datasets used for training, testing, and validating both ***BERTIS*** and ***META4***; as well as the objective evaluation we conduct to assess our approach objectively, and post-hoc visualization of the generated 2D upper body poses.

4.1 Material and Model setups.

4.1.1 PATS Corpus

We train ***META4*** on the *PATS Corpus* [Ahuja et al.(2020b), Ginosar et al.(2019)] which comprises various modalities, including 2D upper-body joint keypoints, aligned with Mel spectrogram, and Bert embeddings, from 25 speakers. The speakers in the corpus are categorized into different groups, namely 15 talk show hosts, 5 lecturers, 3 YouTubers, and 2 televangelists. Each speaker exhibits a unique communication style, contributing to lexical and gesture diversity within the corpus. The corpus consists of a total of 251 hours of data, with $84,000$ intervals, and an average duration of 10.7 seconds per interval. The standard deviation for interval duration is 13.5 seconds. Each interval corresponds to an utterance containing 64 timesteps and corresponding to ***S***=4.666 seconds. Despite the inclusion of finger data in the PATS Corpus, we have made the decision not to incorporate finger modeling in our work. This choice is based on the observation that the quality of the extracted finger data is highly noisy and lacks accuracy. Instead, our focus lies in modeling and predicting the 2D joints of the upper body and arms, utilizing 11 joints specifically to represent these regions.

4.1.2 Image Schemas Corpus.

We fine-tune the ***BERTIS*** model using a corpus proposed by Wachowiak et al. [Wachowiak and Gromann(2022)]. This corpus is specifically designed to aid researchers in classifying natural language expressions into image schemas, and it contains examples from the image schema literature. The annotation data encompass samples in various languages, but in this study, we focus solely on the English samples. The English subset consists of 1994 utterance samples along with their corresponding image schema labels. During the training of ***BERTIS***, we perform data oversampling to address the class imbalance and compensate for the limited number of samples used for fine-tuning ***BERT***. This oversampling technique helps improve the classification accuracy of the model.

4.1.3 *META4* Testing.

We evaluate the performance of ***META4*** following a testing protocol that includes two conditions: the "Seen Speaker" condition and the "Unseen Speaker" condition. In the "Seen Speaker" condition, we assess how accurately our model can generate gestures when presented with speakers that were included in the training data. The "Unseen Speaker" condition evaluates our model's ability to generalize its predictions to speakers that were not encountered during the training phase. PATS database already provides predefined train, validation, and test sets for each speaker. In our experiments, we train our model using data from 16 PATS speakers. For the "Seen Speaker" condition, we use the test sets specifically designated for the 16 PATS speakers as our evaluation dataset. For the "Unseen Speaker" condition, we select 6 speakers and utilize their respective test sets to conduct our experiments.

4.2 Objective Evaluation

We conduct an objective evaluation of the performance of our ***META4*** model by considering two crucial aspects: (1) the *distance* between the predicted gestures $\widehat{Y}_{Pose}$ and the ground truth gestures Y_{Pose}, and (2) the *similarity* between $\widehat{Y}_{Pose}$ and Y_{Pose}. To assess the effectiveness of the Image Schema input modality, we perform an ablation study where we compare two variations of our model: (1) the full model of ***META4***, and (2) a modified version of ***META4*** with the image schema input modality ablated. Additionally, we conduct a sensitivity analysis to evaluate the model's reliance on the image schema input and its robustness in handling variations. To assess the reliability of the ***BERTIS*** component within the overall framework of ***META4***, we evaluate it separately, by assessing the robustness and accuracy of this module in classifying text into Image Schemas.

4.2.1 Metrics

For our ablation studies, we use the following metrics to assess the performance of ***META4***:

1. **Root Mean Squared Error (*RMSE*)** and **Mean Absolute Error (*MAE*)** are employed to measure the *distance* between $\widehat{Y}_{Pose}$ and Y_{Pose}.
2. **Pearson Correlation Coefficient (*PCC*)** and **Cosine Similarity** are used to evaluate the *similarity* between $\widehat{Y}_{Pose}$ and Y_{Pose}.

To assess the classification accuracy of ***BERTIS*** module in classifying text into image schemas, we use the following metrics:

1. **Precision**: measures the proportion of correctly predicted positive instances out of all instances predicted as positive. It focuses on the *accuracy* of positive predictions.
2. **Recall**: measures the proportion of correctly predicted positive instances out of all actual positive instances. It focuses on the ability to capture positive instances.
3. **F1-score**: the harmonic mean of precision and recall, providing a balanced measure that combines both metrics

5 Results and Discussion

Table 1 reports the objective evaluation results of the evaluation conducted on ***BERTIS*** for assessing its reliability in the framework of ***META4***, and its accuracy in classifying text into Image Schemas. For all Image Schema classes, *F1-score* is between 0.77 and 1, indicating that ***BERTIS*** performs well in terms of both accurately predicting the correct Image Schema classes (*precision*) and capturing many correct Image Schema classes as possible (*recall*). Indeed these results are also reflected in the *precision* and *recall* scores for all Image Schema classes. We observe a high *precision* (between 0.74 and 1) for all classes, which indicates the number of correct predicted Image Schema classes. The *recall* scores are above 0.74 for most of the classes, with one Image Schema class having a *recall* score equal to 0.67. As reported in Table 1, the overall *accuracy* of ***BERTIS*** with respect to all Image Schema classes is equal to 0.93, indicating a high accuracy and good performance in classifying the input text into the correct Image Schema classes.

Class	Precision	Recall	F1-score
"CENTER-PERIPHERY"	0.98	0.94	0.96
"CONTACT"	1	1	1
"CONTAINMENT"	0.74	0.67	0.7
"COVERING"	1	1	1
"FORCE"	0.8	0.91	0.85
"LINK"	1	1	1
"OBJECT"	0.81	0.83	0.82
"PART-WHOLE"	1	1	1
"SCALE"	1	1	1
"SOURCE_PATH_GOAL"	0.81	0.74	0.77
"SPLITTING"	1	1	1
"SUBSTANCE"	1	1	1
"SUPPORT"	1	1	1
"VERTICALITY"	0.88	0.93	0.90
Overall Accuracy	0.93		

Table 1: Objective evaluation results of ***BERTIS***. Precision, Recall, and F1-score are computed for each of the 14 Image Schema classes. The overall accuracy is also computed. Higher scores indicate better performance for all three metrics.

		Distance Metrics		*Similiarity Metrics*	
Condition		**RMSE**	**MAE**	**PCC**	**Cosine Similiarity**
Speaker Dependent	**Full Model**	**0.01627**	**0.01197**	**0.98292**	**0.9985**
	IS Ablation	0.02004	0.01425	0.97645	0.997775
	Mismatched	0.01742	0.01222	0.96311	0.98415
Speaker Independent	**Full Model**	**0.02100**	**0.01543**	**0.97736**	**0.997497**
	IS Ablation	0.02520	0.01824	0.96779	0.99647
	Mismatched	0.02591	0.0161	0.97311	0.97313

Table 2: Objective evaluation results of ***META4*** for both conditions Speaker dependent and Speaker Independent. Results are reported for three conditions: the full model, the model with Image Schema (IS) modality ablation, and the mismatched condition.

Table 2 reports the ablation study as well as the sensitivity analysis results for both conditions *Speaker Dependent (SD)* and *Speaker Independent (SI)*. The full model performs best for both conditions in terms of errors produced w.r.t the Ground Truth (GT), and in terms of ***META4*** predictions' correlation w.r.t to GT. When ablating the Image Schema modality, the ***RMSE*** and ***MAE*** errors increase for both *SD* and *SI* conditions. We notice a decrease in **PCC** and *Cosine Similarity*, which indicate that the predictions became less correlated with the GT. These results confirm our hypothesis that Image Schema modality influences the predictions and therefore constitutes an important modality to consider when designing generative models for gesture synthesis. The sensitivity analysis results are also reported in Table 2 under the condition "mismatched", which represents the error condition, and serves as a control condition to evaluate the robustness of ***META4*** in handling Image Schema variations, allowing us to understand further the reliance of ***META4*** on Image Schemas modality. For both conditions *SD* and *SI*, we notice an increase in terms *RMSE* and

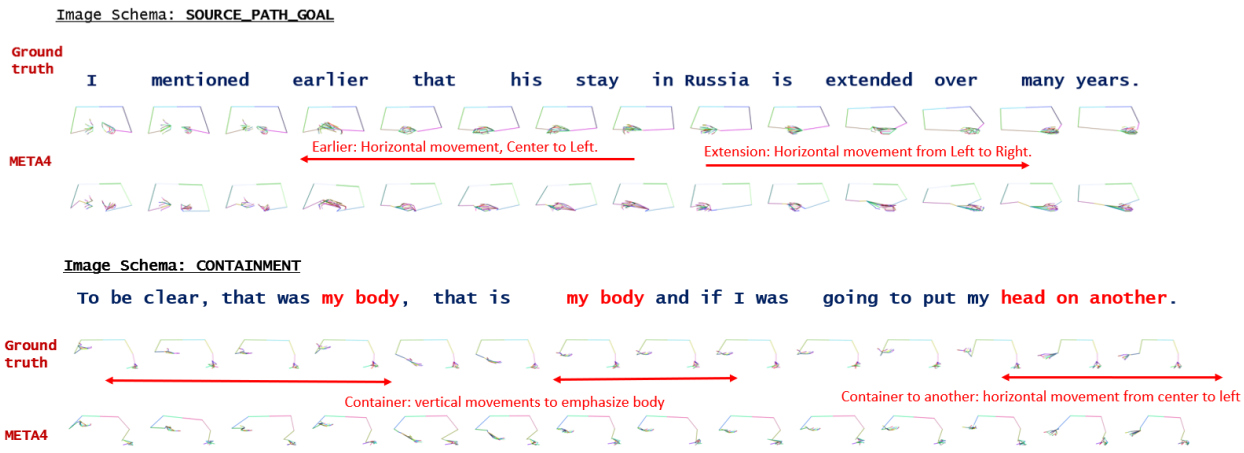

Figure 3: Visualizations of the Ground Truth Vs. META4 gestures for two utterances, one having a "SOURCE_PATH_GOAL" image schema, and the other a "CONTAINMENT" image schema. Fingers are added for sake of visualisation.

MAE errors, indicating that distorting the input Image Schema modality generates errors. The same conclusion can be drawn by looking at the *PCC* and *Cosine Similarity* between predictions and the ground truth, which decreased. Furthermore, looking at the results of *SI* condition, we can validate ***META4's*** ability to generalize its predictions to speakers that were not encountered during the training phase.

6 Post-Hoc Animation Visualization

Figure 3 illustrates the motion of a speaker saying two different utterances, each having a different Image Schema. The first key-frame animation corresponds to a text with a "SOURCE_PATH_GOAL" Image Schema. The speaker moves both of his arms from left to right. The same motion is reproduced by ***META4***. The second key-frame animation corresponds to a speaker saying an utterance with the image schema "CONTAINMENT". The speaker performs a vertical movement with his left hand to visually illustrate the concept of his body which he emphasizes in his utterance. He visually illustrate the metaphor of putting his head on another, he performs a horizontal movement. ***META4*** synthesizes the same behavior. This post-hoc visualization allows us to validate that ***META4*** can reproduce the motion in the Ground Truth for different Image Schemas.

7 Conclusion

In this work, we propose an approach for synthesizing speech-driven metaphoric gestures while leveraging the potential of *Image Schemas*. We conduct evaluation studies to demonstrate the effectiveness of our approach and show the importance of considering both speech and Image Schemas for generating *metaphoric* gestures. In future work, we plan to conduct other evaluations to validate our approach subjectively on an Embodied Conversational Agent.

Appendix

Image Schema Classes

The image schema classes that we consider in our work are 14, and are listed in Table 3.

Image Schema Class	Example
CENTER-PERIPHERY	She brushed the thought away.
CONTACT	That blew me away.
CONTAINMENT	Keep it in the back of your mind
COVERING	His judgement is clouded.
FORCE	They are attracted to each other.
LINK	breaking social ties
OBJECT	Seize the opportunity.
PART-WHOLE	They assembled a theory.
SCALE	This class is bigger than that one.
SOURCE_PATH_GOAL	The time for action has arrived.
SPLITTING	What separates the men from the boys?
SUBSTANCE	Emotions are tinged with suffuse
VERTICALITY	No known spoken language uses the lateral axis for time.
SUPPORT	The poor in our country need a boost up.

Table 3: Image Schema classes with examples taken from [Wachowiak and Gromann(2022)].

References

[Ahuja et al.(2020a)] Chaitanya Ahuja, Dong Won Lee, Ryo Ishii, and Louis-Philippe Morency. 2020a. No Gestures Left Behind: Learning Relationships between Spoken Language and Freeform Gestures. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*. 1884–1895.

[Ahuja et al.(2020b)] Chaitanya Ahuja, Dong Won Lee, Yukiko I Nakano, and Louis-Philippe Morency. 2020b. Style transfer for co-speech gesture animation: A multi-speaker conditional-mixture approach. In *European Conference on Computer Vision*. Springer, 248–265.

[Antonova(2020)] Marina Antonova. 2020. The container image schema as the conceptual basis of English adjectives' semantics. *Journal of Language and Education* 6, 1 (2020).

[Bergmann and Kopp(2009)] Kirsten Bergmann and Stefan Kopp. 2009. GNetIc–Using bayesian decision networks for iconic gesture generation. In *International workshop on intelligent virtual agents*. Springer, 76–89.

[Cassell et al.(2001)] Justine Cassell, Hannes Högni Vilhjálmsson, and Timothy Bickmore. 2001. Beat: the behavior expression animation toolkit. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*. 477–486.

[Chiu et al.(2015)] Chung-Cheng Chiu, Louis-Philippe Morency, and Stacy Marsella. 2015. Predicting co-verbal gestures: A deep and temporal modeling approach. In *International Conference on Intelligent Virtual Agents*. Springer, 152–166.

[Cienki(2005)] Alan Cienki. 2005. Image schemas and gesture. *From perception to meaning: Image schemas in cognitive linguistics* 29 (2005), 421–442.

[Cienki(2013)] Alan Cienki. 2013. Image schemas and mimetic schemas in cognitive linguistics and gesture studies. *Review of Cognitive Linguistics. Published Under the Auspices of the Spanish Cognitive Linguistics Association* 11, 2 (2013), 417–432.

[Cienki and Müller(2008)] Alan Cienki and Cornelia Müller. 2008. Metaphor, gesture, and thought. *The Cambridge handbook of metaphor and thought* 483 (2008), 501.

[Devlin et al.(2018)] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[Devlin et al.(2019)] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,*

Volume 1 (Long and Short Papers). Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[Fares(2020a)] Mireille Fares. 2020a. Towards multimodal human-like characteristics and expressive visual prosody in virtual agents. In *Proceedings of the 2020 International Conference on Multimodal Interaction*. 743–747.

[Fares(2020b)] M Fares. 2020b. Towards Multimodal Human-Like Characteristics and Expressive Visual Prosody in Virtual Agents. In *Int Con on Multimodal Interaction*. 743–747.

[Fares et al.(2022a)] Mireille Fares, Michele Grimaldi, Catherine Pelachaud, and Nicolas Obin. 2022a. Zero-Shot Style Transfer for Gesture Animation driven by Text and Speech using Adversarial Disentanglement of Multimodal Style Encoding. *arXiv preprint arXiv:2208.01917* (2022).

[Fares et al.(2022b)] Mireille Fares, Catherine Pelachaud, and Nicolas Obin. 2022b. Transformer Network for Semantically-Aware and Speech-Driven Upper-Face Generation. In *European Signal Processing Conference (EUSIPCO)*.

[Fares et al.(2023)] Mireille Fares, Catherine Pelachaud, and Nicolas Obin. 2023. Zero-shot style transfer for gesture animation driven by text and speech using adversarial disentanglement of multimodal style encoding. *Frontiers in Artificial Intelligence* 6 (2023), 1142997.

[Gibbs Jr and Colston(2006)] Raymond W Gibbs Jr and Herbert L Colston. 2006. Image schema. *Cognitive linguistics: Basic readings* 6, 4 (2006).

[Ginosar et al.(2019)] Shiry Ginosar, Amir Bar, Gefen Kohavi, Caroline Chan, Andrew Owens, and Jitendra Malik. 2019. Learning individual styles of conversational gesture. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3497–3506.

[Gong et al.(2021)] Yuan Gong, Yu-An Chung, and James Glass. 2021. Ast: Audio spectrogram transformer. *arXiv preprint arXiv:2104.01778* (2021).

[Grady(2005)] Joseph E Grady. 2005. Image schemas and perception: Refining a definition. *From perception to meaning: Image schemas in cognitive linguistics* 29 (2005), 35.

[Haag and Shimodaira(2016)] K Haag and H Shimodaira. 2016. Bidirectional LSTM networks employing stacked bottleneck features for expressive speech-driven head motion synthesis. In *Intelligent Virtual Agents*. 198–207.

[Hofer and Shimodaira(2007)] G Hofer and H Shimodaira. 2007. Automatic head motion prediction from speech data. In *Interspeech*.

[Ishi et al.(2018)] Carlos T Ishi, Daichi Machiyashiki, Ryusuke Mikata, and Hiroshi Ishiguro. 2018. A speech-driven hand gesture generation method and evaluation in android robots. *IEEE Robotics and Automation Letters* 3, 4 (2018), 3757–3764.

[Johnson(2005)] Mark Johnson. 2005. The philosophical significance of image schemas. *From perception to meaning: Image schemas in cognitive linguistics* (2005), 15–33.

[Johnson(2013)] Mark Johnson. 2013. *The body in the mind: The bodily basis of meaning, imagination, and reason*. University of Chicago press.

[Karras et al.(2017)] Tero Karras, Timo Aila, Samuli Laine, Antti Herva, and Jaakko Lehtinen. 2017. Audio-driven facial animation by joint end-to-end learning of pose and emotion. *ACM Transactions on Graphics (TOG)* 36, 4 (2017), 1–12.

[Kendon(2004)] Adam Kendon. 2004. *Gesture: Visible action as utterance*. Cambridge University Press.

[Kenton and Toutanova(2019)] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, Vol. 1. 2.

[Kipp(2001)] Michael Kipp. 2001. Anvil-a generic annotation tool for multimodal dialogue. In *Seventh European conference on speech communication and technology*.

[Kipp(2005)] Michael Kipp. 2005. *Gesture generation by imitation: From human behavior to computer character animation*. Universal-Publishers.

[Kopp et al.(2006)] Stefan Kopp, Brigitte Krenn, Stacy Marsella, Andrew N Marshall, Catherine Pelachaud, Hannes Pirker, Kristinn R Thórisson, and Hannes Vilhjálmsson. 2006. Towards a common framework for multimodal generation: The Behavior Markup Language. In *International workshop on intelligent virtual agents*. Springer, 205–217.

[Kucherenko et al.(2020)] Taras Kucherenko, Patrik Jonell, Sanne van Waveren, Gustav Eje Henter, Simon Alexanderson, Iolanda Leite, and Hedvig Kjellström. 2020. Gesticulator: A framework for semantically-aware speech-driven gesture generation. In *Proceedings of the ACM International Conference on Multimodal Interaction*.

[Ladewig(2011)] Silva H Ladewig. 2011. Putting the cyclic gesture on a cognitive basis. *CogniTextes. Revue de l'Association française de linguistique cognitive* Volume 6 (2011).

[Lakoff(1987)] George Lakoff. 1987. Image metaphors. *Metaphor and Symbol* 2, 3 (1987), 219–222.

[Langacker(1987)] Ronald W Langacker. 1987. *Foundations of cognitive grammar: Volume I: Theoretical prerequisites*. Vol. 1. Stanford university press.

[Mariooryad and Busso(2012)] S Mariooryad and C Busso. 2012. Generating human-like behaviors using joint, speech-driven models for conversational agents. *IEEE Trans on Audio, Speech, & Language Processing* 20, 8 (2012).

[Marsella et al.(2013)] Stacy Marsella, Yuyu Xu, Margaux Lhommet, Andrew Feng, Stefan Scherer, and Ari Shapiro. 2013. Virtual character performance from speech. In *Proceedings of the 12th ACM SIGGRAPH/Eurographics symposium on computer animation*. 25–35.

[McNeill and Levy(1982)] David McNeill and Elena Levy. 1982. Conceptual representations in language activity and gesture. *Speech, place, and action* (1982), 271–295.

[Neff et al.(2008)] Michael Neff, Michael Kipp, Irene Albrecht, and Hans-Peter Seidel. 2008. Gesture modeling and animation based on a probabilistic re-creation of speaker style. *ACM Transactions on Graphics (TOG)* 27, 1 (2008), 1–24.

[Pelachaud et al.(1996)] Catherine Pelachaud, Norman I Badler, and Mark Steedman. 1996. Generating facial expressions for speech. *Cognitive science* 20, 1 (1996), 1–46.

[Pelachaud et al.(2002)] Catherine Pelachaud, Valeria Carofiglio, Berardina De Carolis, Fiorella de Rosis, and Isabella Poggi. 2002. Embodied contextual agent in information delivering application. In *Proceedings of the first international joint conference on Autonomous agents and multiagent systems: part 2*. 758–765.

[Ravenet et al.(2018)] Brian Ravenet, Catherine Pelachaud, Chloé Clavel, and Stacy Marsella. 2018. Automating the production of communicative gestures in embodied characters. *Frontiers in psychology* 9 (2018), 1144.

[Sadoughi and Busso(2019)] N Sadoughi and C Busso. 2019. Speech-driven animation with meaningful behaviors. *Speech Communication* 110 (2019), 90–100.

[Vaswani et al.(2017)] A Vaswani, N Shazeer, N Parmar, J Uszkoreit, L Jones, A N Gomez, L Kaiser, and I Polosukhin. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762* (2017).

[Vougioukas et al.(2019)] K Vougioukas, S Petridis, and M Pantic. 2019. Realistic speech-driven facial animation with GANs. *Int Journal of Computer Vision* (2019), 1–16.

[Wachowiak and Gromann(2022)] Lennart Wachowiak and Dagmar Gromann. 2022. Systematic analysis of image schemas in natural language through explainable multilingual neural language processing. In *Proceedings of the 29th International Conference on Computational Linguistics*. 5571–5581.

[Wang et al.(2021)] S Wang, L Li, Y Ding, C Fan, and X Yu. 2021. Audio2Head: Audio-driven One-shot Talking-head Generation with Natural Head Motion. *preprint arXiv:2107.09293* (2021).

[Williams(2008)] Robert F Williams. 2008. Gesture as a conceptual mapping tool. *Metaphor and gesture* (2008), 55–92.

[Yoon et al.(2020)] Youngwoo Yoon, Bok Cha, Joo-Haeng Lee, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. 2020. Speech gesture generation from the trimodal context of text, audio, and speaker identity. *ACM Transactions on Graphics (TOG)* 39, 6 (2020), 1–16.

[Yoon et al.(2019)] Youngwoo Yoon, Woo-Ri Ko, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. 2019. Robots learn social skills: End-to-end learning of co-speech gesture generation for humanoid robots. In *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 4303–4309.