Large Language Models for Virtual Human Gesture Selection

Parisa Ghanad Torshizi
Northeastern University
Boston, United States
ghanadtorshizi.p@northeastern.edu

Laura B. Hensel
University of Glasgow
Glasgow, United Kingdom
laura.hensel@glasgow.ac.uk

Ari Shapiro
Flawless
Los Angeles, United States
ariyshapiro@gmail.com

Stacy C. Marsella
Northeastern University
Boston, United States
s.marsella@northeastern.edu

ABSTRACT

Co-speech gestures convey a wide variety of meanings and play an important role in face-to-face human interactions. These gestures have been shown to significantly influence the addressee's engagement, recall, comprehension, and attitudes toward the speaker. Similarly, they have been shown to impact human and embodied virtual agent interaction. The process of selecting and animating meaningful gestures has thus become a key focus in designing embodied virtual agents. However, the automation of this gesture selection process poses a significant challenge. Prior gesture generation techniques have attempted to address this challenge in varied ways from fully automated, data-driven techniques – which often struggle to produce contextually meaningful gestures – to more manual approaches of crafting gesture expertise, which are time-consuming and lack generalizability. In this paper, we leverage the semantic capabilities of Large Language Models to realize a gesture selection approach that suggests meaningful, appropriate co-speech gestures. We first illustrate the information on gestures encoded into GPT4. Then we perform a study to specifically evaluate alternative prompting approaches for their ability to select meaningful, contextually relevant gestures and to align them appropriately to the co-speech utterance. Finally, we detail and demonstrate how this approach has been implemented within a virtual agent system, automating the selection and subsequent animation of the selected gestures for human-agent interactions.

KEYWORDS

gesture selection, virtual humans, large language models

ACM Reference Format:
Parisa Ghanad Torshizi, Laura B. Hensel, Ari Shapiro, and Stacy C. Marsella. 2025. Large Language Models for Virtual Human Gesture Selection. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Detroit, Michigan, USA, May 19 – 23, 2025*, IFAAMAS, 9 pages.

Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Y. Vorobeychik, S. Das, A. Nowé (eds.), May 19 – 23, 2025, Detroit, Michigan, USA. © 2025 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

1 INTRODUCTION

People now regularly interact with embodied facsimiles of people. Graphics-based virtual humans and social robots with anthropomorphic features and behaviors engage users using the same verbal and non-verbal behaviors that people use when interacting with each other. These technologies exploit that the nonverbal behaviors of participants powerfully influence face-to-face interaction.

In this paper, we focus on an integral part of such social interactions: co-speech gestures. Gestures convey a range of meanings and have a powerful impact on face-to-face interaction [2, 12, 21, 31], impacting the speaker's persuasiveness as well as an addressee's comprehension, recall, engagement, and trust in the speaker [1, 20, 41].

However, these impacts are dependent on the particular gestures being used and the context in which they occur [e.g., 19]. Clinicians, politicians, and comedians use different gestures because they seek to achieve different goals in different contexts [10, 16, 38]. Beyond these differences in contexts and goals, there are also considerable cultural and individual differences in the use of gestures [23, 35].

This richness makes gesture selection and animation a fundamental challenge in embodied agent research. To solve this challenge, researchers have focused on various automated approaches to select and generate virtual human gestures, often relying on data-driven or analysis-driven approaches [for an overview, see 34, 37].

With any approach, there are design factors that are critical to the realization of the virtual agent's gestures. Among these are requirements for the gestures to effectively convey the speaker's communicative intent. Further, we assume the virtual human should use gestures consistent with its role in an interaction. Thus, a virtual human who is taking on the role of a labor union executive should not gesture like a comedian or clinician, nor should it use gestures based on some statistical average over differing roles and situations.

In this paper, we thus assume one is designing a virtual human to perform a specific role in some application. Leveraging recent work on exploring the ability of Large Language Models (LLMs) to predict gestures [18], we investigate the use of LLMs, in particular the standard GPT-4, in gesture selection with an additional focus on implementing this approach within a virtual human architecture. What makes LLMs particularly promising as a component in gesture selection is that they are trained on very large and highly varied corpora. As we demonstrate, that leads to LLMs encoding representations of both the linguistic properties critical to conveying meaning in gesture performances, as well as the relation of those phenomena to actual gestures. Furthermore, we explore and evaluate LLMs' ability to represent rheme and theme [11, 31], metaphor [7, 13], image schemas [6] and rhetorical structures [29]. GPT4's ability is assessed in the context of crafting a specific individual, a skilled presenter from a labor organization. The gesture selection approach is also implemented into an embodied virtual agent architecture allowing automation of gesture selection.

Our contributions include exploring GPT-4's capabilities in selecting appropriate gestures, as well as using it to do discourse analysis to determine when to gesture. We also evaluated different prompting approaches. Furthermore, we automated the mapping from GPT-4's output into a behavior specification required by the character animation system to select the gesture animation. Finally, we implemented this within a virtual human framework.

2 BACKGROUND

Research on gestures and their generation in virtual humans has a rich history, with gesture studies spanning millennia and virtual gesture generation predating the turn of the century (e.g., [3]). In this section, we touch on some key aspects relevant to how we approached the problem of gesture generation. Research in human co-speech gestures has identified different categorizations of gestures, related to the kinds of information they convey. For example, the work of McNeil (e.g., [31]) identifies four gesture categories—deictic, beat, metaphoric, and iconic. This work will largely focus on metaphoric gestures for two primary reasons: First, our research often focuses on skilled presenters, and such professional speakers tend to use these gestures frequently. Second, metaphoric gestures can require deep analysis of the utterance to uncover relations to physical imagery, called image schemas, that underlie metaphoric gestures. Uncovering these relations presents a challenge for automated gesture selection.

Specifically, image schemas [14] are recurring spatiotemporal relationships grounded in our body's interaction in the world that are argued to underlie common patterns of physical and abstract reasoning, as well as motivate linguistic metaphors and metaphoric gestures. Consider this example of the image schema of container from our data set: "with workers and employers putting a little bit of contribution from each into a trust fund". Note container is not only the collection, the "trust fund", but there is also a pattern of reasoning associated with container, putting things into the container. The speaker used a pattern of metaphoric gestures associated with container, a gesture denoting the container, along with gestures depicting placing things into it.

Ideational Units: The above container example leads to an even more challenging issue. Gestures also occur in sequences where the individual gestures convey inter-related meanings. These gestural sequences are called ideational units [2] or gesture units [21]. This coupling plays important demarcative functions as well as helping to convey related meanings. These ideational units can be associated with image schemas as above but they can also establish relations between parts of an utterance, giving them a rhetorical structure. Consider a communicative intent to convey a rhetorical contrast between two abstract concepts such as strongly different political views. Metaphorically, abstract concepts can be viewed as physical objects, with physical properties such as locations and physical distance conveying conceptual differences. Thus the rhetorical contrast between political views can be conveyed metaphorically by abstract deictic gestures that convey this separation by pointing to disparate regions in space. Such gestures pose difficult challenges for gesture animation because they set up the physical space of multiple gestures in a consistent fashion that relates physical space and motion to meaning as the above two examples illustrate.

2.1 Gesture Selection

The process of automating the selection of co-speech gestures for an utterance can be broken into two tasks. The first task is determining which segments of an utterance have co-speech gestures. Since gestures emphasize as well as convey meaning for parts of the utterance, identifying these segments is fundamental. We characterize this as *when to gesture*. Then there is the question of *what gesture to use*. In other words, what the gesture should convey, given the utterance and the context of the interaction.

When to gesture: Gestures are tied to what the speaker seeks to emphasize [8], which in turn transforms the meaning conveyed. Therefore, to determine when to gesture, it is critical for co-speech gesture generation to derive emphasis information about the utterance. A common approach is to use prosodic cues [9, 15, 30]. However, this presumes that the spoken utterance that is driving gesture generation includes prosodic cues appropriate for the meaning the speaker seeks to convey. An alternative is a discourse analysis [32] that can identify the *theme* or topic of a sentence and *rheme* or focus of the sentence, which is what is being said about the topic. The rheme provides new information about the topic and, critically, tends to be associated with gesture co-occurence [5, 31].

Gesture generation systems have used rheme analysis [4] as well as prosodic analysis [30] to determine when to gesture. Assuming we choose to use rheme analysis to automate when to gesture, the question becomes how to quickly and efficiently do the analysis. In the context of the current work, we can specifically ask whether LLMs can do this kind of discourse analysis as part of a gesture selection process.

What gesture to use: There is a considerable variation in gesture types and usage frequency across individuals and contexts [2, 22].

As just one example of metaphoric gestures, consider the linguistic metaphor that abstract concepts can be physical objects [13]. All the physical properties and actions one takes on physical objects can be used to convey abstract meanings gesturally. Thus, concepts, as physical objects, can be gesturally grasped (understood), they can be thrown away (rejected), and they can be big (important). Two concepts, such as political orientation, as physical objects can have separate locations to contrast them or convey their differences. Sets of abstract concepts can have cardinality (large or small) and openness(closed or open sets). Set operations are apparent in gestural forms, adding elements, deleting, and union of sets.

This richness is not surprising given the richness of metaphors represented in our language [13] and argued to be integral to our thought processes [25]. It also represents a central challenge to approaches to gesture generation.

3 RELATED WORK

As noted earlier, a key challenge in designing embodied conversational agents involves selecting appropriate gestures for the character, with various approaches proposed to address this challenge [33].

Rule-based methods [5, 26] rely on predefined sets of rules that select from a predefined set of gesture animations. For instance, the Behavior Expression Animated Toolkit (BEAT) is a rule-based system that generates body movements and vocal intonations based on linguistic and contextual features of the text [5].

Mixed approaches [28, 30, 36] integrate machine learning and ontologies to perform syntactic, semantic, and prosodic analysis of the utterance to infer communicative intent and then use hand-crafted knowledge to map that intent to non-verbal behaviors such as head movements, gestures, and gaze.

Although these rule-based and mixed systems offer designer control over gesture selection, they are limited to a fixed relationship between the properties of the speech they infer and gestures. Essentially, there is a trade-off between the designer's more explicit control over a character's use of gesture and both the burden placed on the developer or designer and the flexibility of the approach.

Data-driven approaches aim to learn a mapping between utterances and gestural motion from annotated corpora of gestural performances. Work in this area relies on, for example, deep learning techniques. A recent review of this deep learning work [34] has identified several key limitations, including a lack of designer control over the performance and the limited ability of current approaches to realize semantically meaningful gestures such as metaphoric gestures. Recent work has become more focused on the use of Transformer-based and diffusion-based generative models for the selection of gestures. DiffMotion [42] generates gestures by integrating an LSTM with a diffusion model. DiM-Gesture [43] integrates a Mamba-based fuzzy feature extractor with Mamba- 2 diffusion architecture, to generate personalized full-body gestures.

Recent advancements have shifted towards Transformer-based and diffusion-based generative models for gesture selection. DiffMotion [42] combines an LSTM with a diffusion model to generate gestures. Meanwhile, DiM-Gesture [43] utilizes a Mamba-based fuzzy feature extractor alongside a Mamba-2 diffusion architecture, enabling the creation of personalized full-body gestures.

These alternatives are solving different problems in some respects. Rule-based approaches continue to be used in applications, such as health applications, where there are significant ethical concerns and the designer seeks to retain control over the interaction, whereas machine learning-based approaches allow for more general, natural-looking behavior but with less control over what the nonverbal behavior conveys.

3.1 Use of LLMs

In this paper, we explore a different approach. Unlike an end-to-end machine learning approach that maps from utterance to animation, we explore the use of LLMs for gesture selection. We therefore go beyond traditional mixed and rule-based approaches in actually proposing gestures based on a more thorough analysis of the utterance by the LLM but stop short of end-to-end machine learning-based approaches that also generate the animation. This is both a limitation and a potential benefit in that it allows for carefully crafted gestures that effectively convey desired communicative intent.

In addition, LLMs encode massive amounts of information across domains. Unlike the simpler word-based ontologies like Wordnet used previously in mixed approaches, or even the data used to train end-to-end approaches, the far more extensive data used to train LLMs allows them to encode contextual, semantic, and structural relations at the phrase, utterance and situational levels. LLMs also encode extensive amounts of information across scientific domains. GPT4, when asked, can define rheme and theme, metaphors, image schema, and metaphoric gestures. As one example:

```
GPT4: Image schemas are recurring, dynamic patterns
of our perceptual interactions and motor programs ...
basic cognitive structures that arise from bodily
interactions with the world ... are understood to be
one of the fundamental building blocks of human thought
```

```
GPT4: Container:|definition|, Path:|definition|,
Source-Path-Goal:|definition|, Force:|definition|...
```

LLMs also have representations of different gestural forms and styles used by speakers in different roles. They have for example knowledge of clinicians, their role, and their goals, and can even capture the variations in gestural styles among well-known individuals within these professions. Moreover, they can be prompted with examples that prime and integrate with their existing knowledge to improve their ability to select gestures. We asked GPT4 to report a repertoire of gestures that a clinician and a politician would use.

```
GPT4: Here's a repertoire of hand gestures that
clinicians often use:
Palms Open and Upward: Indicates openness and
transparency, used when explaining procedures or
discussing options.
Hand to Chest: Expresses sincerity and empathy,
particularly in sensitive conversations.
Enclosed Hands or Fingers Interlocked: Demonstrates
patience and focus, typically seen during attentive
listening.
```

```
GPT4: Here's a repertoire of hand gestures that
politicians often use:
Fist Pumping: Used to express determination or
victory, energizing the audience.
Thumb and Index Finger Pinch: Known as the
"precision grip," it emphasizes specific and
important points, suggesting control.
Hand Chopping: Aligns with making strong,
decisive points, conveying clarity and decisiveness.
```

Detailing the extent of all these representations is beyond the scope of the current paper. Rather, the focus of this paper is to lay out an integrated approach to using LLMs to automate gesture selection and evaluate the ability to determine *when to gesture* and *what gestures to use*. Our approach builds on the work of Hensel et al. [18] but goes further in exploring different prompting techniques, having the LLM also determine when to gesture, develop a prompting approach that allows the automated mapping of LLM output to a behavior specification for a character animation system and finally implementing the approach. The implementation takes utterance text as input, generates the utterance audio, and drives a virtual human's spoken dialog and gesturing.

4 APPROACH

To explore a gesture generation framework using GPT-4, we lay out a set of research questions relating to gesture generation more broadly and the selection of individual gestures more specifically:

RQ1: Selection of Gestures. What are the impacts of alternative prompting approaches on the appropriateness of gesture selection and the speed of inference?

- *Appropriateness* How appropriate are the selected gestures, with regards to the context of the speech and the speaker?
- *Speed of inference* Can GPT-4 select gestures with a speed sufficient to enable real-time inference in animation systems?

RQ2: Selection of when to gesture. Does the rheme identified through GPT-4-enabled rheme and theme analysis correspond accurately to the actual gestural timings of the speaker in the data? To answer these research questions, we first annotated hand gestures in a specific speech. We then used these annotations as ground truths to evaluate gestures suggested by GPT-4 on the same utterances. We used GPT-4 as its performance is superior to that of GPT-3 or GPT-3.5, as shown by [18]. Specifically, we used the GPT-4 Chat Completion API as it is faster and more versatile in terms of parameter settings, with the following parameters: (Temperature: 0.2, Max tokens: 256, Frequency penalty: 0). The transcribed speech was split into utterances, using end of sentence punctuations.

4.1 Data

To derive a set of ground truth gestures and utterances, we selected a video featuring Elizabeth Shuler, a labor activist, speaking at the Working Families Summit, which is available through the National Archives and Records Administration.[1] To collate Elizabeth Schuler's gesture repertoire, we annotated three segments of the video in which she was the primary speaker, resulting in a total of six minutes of annotated video. We then split these annotations into a training set (21 utterances) and a test set (20 utterances).

4.2 Classes of Gestures: Gestural Intents

Based on the training data set, we created a list of gestural classes based on image schemas and metaphors that were apparent in the speaker's performance. Each of these classes conveys a particular intent and we thus refer to these as *gestural intents.* Note that alternative gestures with different physical properties can realize one of these gestural intents. For example, the speaker may convey the gestural intent of progress either through a forward circling of the hands or a sweep showing the direction of progress. Below, we list the speaker's repertoire of *gestural intents*, including brief explanations.

- Progress: This gesture represents progress, advancement, or moving forward. It is part of the path group image schema.
- Regress: This gesture represents moving backward, regressing, or returning to a previous point. It is part of the path group image schema.
- Cycle: This gesture represents actions or processes that repeat in a continuous loop or follow a recurring pattern. It is an image schema.
- Collect: This gesture represents gathering, collecting, or bringing things together, into one entity. It is an image schema.
- Container: This gesture represents a boundary, a sweep, or an imaginary box holding a collection of items. This is an image schema and basis for the container metaphoric gesture.
- Oscillation: This gesture represents alternation, uncertainty, indecision, or items being out of balance. It is part of the balance group image schema.
- Temporal: There are many, culture-specific time metaphors. Here we refer to representing time as a line, with different points on the line representing past, present, and future.

4.3 Experimental Design: Gesture Selection

To evaluate the selection of gestures, we designed a 2X2 factorial design of alternative prompting approaches: *gestural intents repertoire explained/not explained* **X** *annotated examples given/not given.* For each of those approaches, we asked GPT-4 to produce gestures based on the utterances in the test set. Below is a detailed description of each approach.

Approach 0: In this approach, the model is not prompted with any prior information, on gestural intents or annotations. The model is just asked to report a gesture for each of the utterances, describe the physical properties of its suggested gesture, and the specific phrase in that utterance that the gesture is trying to illustrate.

Approach 1: In this approach, the model is only prompted with the above-mentioned list of gestural intents, and their descriptions. The model is then asked to report the gestural intents in the utterance, suggest a gesture for each gestural intent, the physical properties of that gesture, and the associated phrase.

Approach 2: In this approach, the model is prompted with only the annotations from the training set. The model is then asked to report the gestural intent, suggest a gesture for that gestural intent, the physical properties of that gesture, and the associated phrase.

Approach 3: In this approach, the model is prompted with the above-mentioned list of gestural intents, and the annotations from the training set. The model is then asked to report the gestural intent, suggest a gesture for that gestural intent, the physical properties of that gesture, and the associated phrase.

4.4 Rheme/Theme Analysis: When to Gesture

We used GPT-4 driven discourse analysis to identify the rheme and theme of each utterance. We then explored whether the rheme and theme can serve as criteria to decide on when to gesture, or which part of the utterance is most likely to be accompanied by a gesture. Specifically, we prompted GPT-4 to identify the rheme and theme in each utterance. Such a rheme and theme analysis could be used on top of the gesture selection system, to prioritize which gesture/s to use in the animation system; with gestures associated with the rheme part of the utterance being given higher priority. To evaluate, we analyzed whether they correspond to the part of the utterance that the speaker gestured on.

[1] https://youtu.be/-6NA1xl32uY?si=lOjTOkyIrQhgnsKJ

5 ANALYSIS

This section details the results of our experiments on GPT-4's ability to select appropriate gestures and determine when to gesture.

5.1 Selection of Gestures

In this section, we investigated our RQ1, examining the appropriateness of gestures selected by alternative prompting approaches and comparing their speed of inference.

5.1.1 Appropriateness. To assess the semantic appropriateness of the gestures selected by the model, two experts jointly evaluated each proposed gesture for every utterance in the test set across all four approaches. Evaluators made their judgments for each gesture and then immediately compared them. If there was disagreement, the two discussed where this disagreement arose and resolved it through further discussion. It should be noted, however, that there was very little disagreement between the two experts. In order to do the evaluation, we divided all the selected gestures by GPT-4 into the following two categories.

Category 1:

This category consists of gestures that the model generates for specific parts of an utterance, where there is a corresponding gesture on that (or close) part of the utterance in the actual speech (i.e., ground truth is available). Experts have assigned one of the following tags to these gestures:

- **Appropriate**: The gesture does convey the semantic meaning that the speaker is trying to convey. Regarding the definition of appropriateness, we defined this as gestures that are not only appropriate for the utterance (i.e., 'fit' with the intended meaning) but also for the context of the speaker.
- **Inappropriate**: The gesture does not convey the semantic meaning that the speaker is trying to convey.
- **No corresponding gesture**: The speaker produced a gesture but the model did not propose a gesture at the corresponding part of the utterance.

Figure 1a shows the number of gestures that belong to each label across the proposed prompting approaches. The results indicate that providing more information to the prompts yields more appropriate gestures and less inappropriate gestures. Specifically, approach 3 is better than approaches 1 and 2, and approaches 1,2, and 3 are better than approach 0. Moreover, providing the model with examples (approach 2) is slightly more effective in selecting more appropriate gestures compared to providing the model with solely the explanation of the gestures (approach 1).

Category 2:

This category consists of gestures that the model generates for specific parts of an utterance, yet there is no corresponding gesture on that (or close) part of the utterance in the actual speech (i.e., ground truth is not available). The experts labeled each gesture with either the appropriate or the inappropriate tag.

The results in Figure 1b suggest that approach 1 has the highest number of appropriate gestures and the lowest percentage of inappropriate gestures.

Combining all the gestures in the two categories listed above, we analyzed the overall appropriateness, regardless of whether a corresponding ground truth was present or not. The results, shown in Figure 1c, indicate that there is a slight difference between approaches 1,2 and 3, but they all outperform approach 0. Additionally, the performance of approach 1, where the model was not prompted with any annotations, is promising. This suggests that LLMs can minimize the need for extensive annotations.

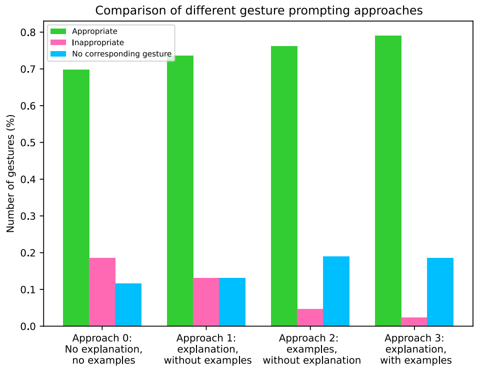
Furthermore, we analyzed how often the gestures generated by GPT-4, regardless of their appropriateness, were aligned with the gestures performed by the speaker, meaning how often the speaker and GPT-4 gesture on the same phrase. Figure 2a and 2b depicts the frequency of gestures being aligned, or not aligned. In not-aligned cases, GPT-4 either produced gestures where the speaker did not gesture or the speaker gestured but GPT-4 did not produce a gesture. Approach 2 has a higher alignment of gestures between GPT-4 and the speaker. On the other hand, approach 0 produced the highest misalignment. These results suggest that limiting the gesture classes to the speaker's gesture repertoire helps to constrain gesture selection to produce more accurate gesture timings. In conclusion, the approaches that prompted the model with a defined set of gestural intents generated more appropriate gestures and less inappropriate gestures. Moreover, they helped in selecting gestures that were happening concurrently with the speaker's gestures and thus were more aligned.

5.1.2 Inference time. As mentioned earlier, a potential use case for LLMs in gesture selection is to suggest gestures in real-time. Therefore, we explored the inference times of gesture selection across different prompting approaches. In this work, the inference times are actually recorded as network latencies, to evaluate real-time suitability for human-agent interactions. In these approaches, where GPT4 was queried to report gesture name, its properties, gesture description, and the associated phrase; the average inference time per utterance across different approaches ranges from 6.51 to 9.29 seconds. Specifically, approach 0 showed the highest inference time, and approach 1 had the lowest inference time. However, for the purpose of animating gestures in the virtual agent, since the animations are already defined, we truncated the query into asking only the gesture type and the associated phrase. We calculated different inference times for different models, GPT4 1.20 (s), GPT4-O-mini 1.08(s), and Llama 3.1 1.23(s), Llama 2.7b fine-tuned on our dataset 1.8(s). With a target inference time of less than one second for online, real-time, interactions, our next step is to adapt a smaller, faster model that can achieve this speed.

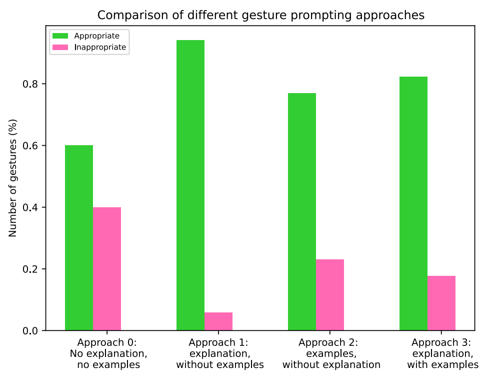
5.2 Selection of When to Gesture: Rheme and Theme

Next, in order to answer our RQ2, we sought to analyze how often the speaker's gestures occur within the rhemes identified by GPT-4. To do so, we considered all the gestures made by the speaker, including deictics, beats, metaphorics, and iconics; and counted the number of times these gestures happened within the rheme of the utterances identified by GPT-4.

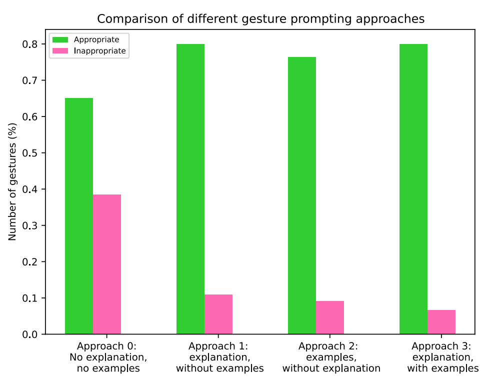
The results show that out of 49 gestures made by the speaker, 43 of them occur within the identified rheme and 6 of them occur outside the identified rheme (they occurred on the identified theme). Among these six gestures, were one deictic, two beat, and three metaphoric gestures.



(a) Selected gestures whose utterances have ground truth



(b) Selected gestures whose utterances do not have ground truth



(c) all gestures

Figure 1: comparison of different prompting approaches in terms of their appropriateness

6 IMPLEMENTATION

This section details the implementation of our proposed LLM-based nonverbal behavior generation system and explains its integration within SIMA, the Socially Intelligent Multimodal Agent. Figure 3 illustrates the architecture for selecting gestures, assuming the virtual human uses text-to-speech. Upon initiation, the LLM processes a System Prompt that includes conversational context and, potentially, examples of utterance-gesture pairings relevant to that context. The prompt may be tailored for various roles, such as a labor union representative at a panel or a presidential candidate at a rally. Furthermore, this prompting stage explores different prompting approaches to assess their impact on the creativity and accuracy of gestures proposed by the LLM. Within our architecture, the text-to-speech and behavior scheduling processes are standard components commonly found in virtual human architectures, as described in [17]. The text-to-speech system generates a schedule for the behavior scheduler, which aligns nonverbal behaviors (gestures, visemes) with the corresponding audio. This behavior schedule is then sent to the animation engine in Behavior Markup Language (BML) [24], which manages co-articulation and blending of gestures. Here, we focus on how the LLM can provide specific gesture information to guide the behavior scheduler and animation engine, including labels or physical properties for real-time animation. The full implementation, and information on all prompting approaches is available on GitHub[2].

6.1 SIMA Gesture Generation

The gesture generation component in SIMA takes speech in textual format and generates a BML file specifying the gestures that the virtual human should perform. The model used in SIMA is GPT-4, which is prompted with prompting approach 3, as specified earlier.

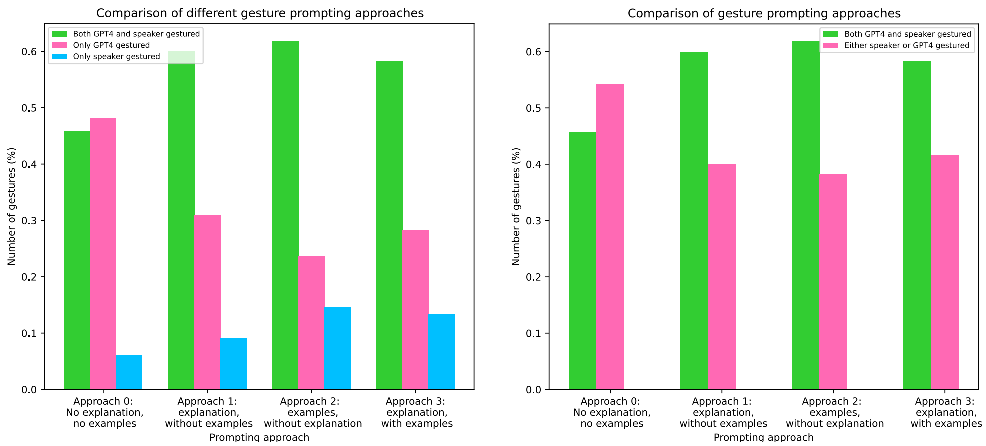
6.1.1 Input Processing. The input dialogue text is tokenized and marked with <mark name=""> tags in XML format. These <mark> tags allow the text-to-speech engine to replace markers with precise BML timings, enabling synchronization between spoken words and gestures. In the current implementation, GPT-4 directly performs the structural and semantic analyses of the dialogue text relevant to its proposed gestures.

6.1.2 Gesture Prompting. Using predefined prompts, GPT-4 identifies potential gestures for corresponding parts of the utterance, specifying gestural intents (e.g., container) and their associated phrases (part of the speech where the gesture occurs). Due to how the animation system handles gestures, this step does not include the physical properties of gestures. The GPT4's output is in JSON format, which is then converted into BML. The gestural intent identified by the model is set as the gesture lexeme in the BML, and the first word in the associated phrase (identified by its word number within the complete sentence) in the output of the LLM is set as the stroke-start point in the BML. All gestures generated by GPT-4 are embedded in a single BML file, which is then passed as input to the animation engine. The animation engine then executes the provided gestures according to the provided timings (the stroke-start points). The formatting of the behavior BML block is as follows: <gesture stroke-start="T3" lexeme="Container" type="METAPHORIC" emotion="neutral" />

6.2 Animation Realization

SIMA uses SmartBody [39, 40] as its animation engine, an open-source framework for real-time animation of conversational agents. SmartBody converts the BML generated by the SIMA gesture generation into character animation aligned with the speech audio.

[2] https://github.com/pariesque/SIMA



(a) Both speaker and GPT-4 gestured, GPT-4 gestured, or speaker gestured

(b) Both speaker and GPT-4 gestured, either GPT-4 gestured or speaker gestured

Figure 2: comparison of different prompting approaches in terms of their alignment with the speaker

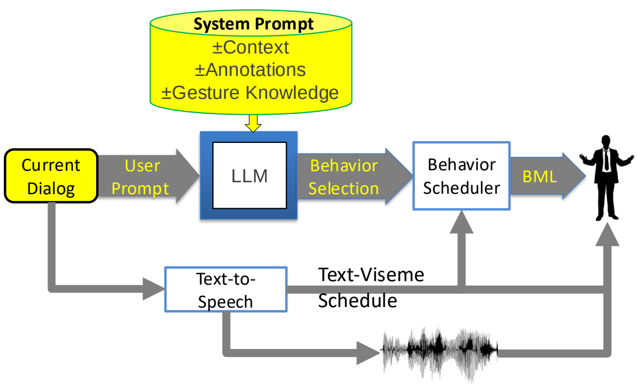


Figure 3: LLM Approach to selecting gestures.
Based on the type of approach, the input of the LLM can contain the context, annotation (examples), or gesture knowledge (gesture description)

SmartBody's behavior processing consists of a behavior & and schedule manager and a motion controller engine. The behavior & schedule manager parses the BML, extracting behavior requests and their synchronization points. These behaviors encompass gestures, speech visemes, and other nonverbal cues. This system retrieves the timing of the speech markers to synchronize the speech with the behaviors. Then, the scheduler assigns the absolute timing to the behaviors' time markers. Meanwhile, the motion controller manipulates skeletal joints, coordinating movements like gaze, gestures, and head nods to ensure fluid, synchronized animation.

The overall pipeline is automated, from the utterance to the crafting of speech audio, gesture selection by LLM, mapping to the BML specification of the animation, and realization in the SmartBody animation system.

7 DISCUSSION

In this paper, we evaluated the use of LLMs to automate gesture selection, examining various prompting approaches that yield diverse results suitable for different applications. As to be expected, Approach 0, where the model was not prompted with any prior information, had the most flexibility in the gestures it recommended and could generate creative variations of gestures. However, this approach had the highest number of inappropriate gestures, the lowest appropriate gestures, and the lowest alignment with the speaker's gestural performance and gestural timings. One might imagine it being useful for suggesting gestures to a designer, as opposed to automating gesture selection. In other approaches, adding information to the prompts, including the explanation of gestural intents and examples from the speaker, increased the number of appropriate gestures and decreased the number of inappropriate gestures. Also, gestures tended to be more aligned with original speaker, with slight variations between approaches 1, 2 and 3.

Looking at the results provided by these different approaches, several qualitative distinctions were apparent. Note that all approaches provided, in addition to the gesture, a description of the physical properties. In general, this was very surprising, as it represented deep relationships between meaning in the utterance and the physical motion, as argued by gesture researchers [2, 22].

In Approach 0, which lacked gestural intents or annotation examples, the model sometimes produced shallow responses, such as self-referential deictics triggered by words like 'I,' 'we, and 'our.' Additionally, this approach frequently generated inappropriate iconic gestures that detracted from speech. Approach 1, which was prompted with gestural intents and descriptions, had fewer gestures than one might expect. Also sometimes the gestural intent was off while the description of the physical motion was good. Approach 2, which was prompted with annotations, but no list or description of gestural intents, would come up with novel gestural intents not in the annotations. Additionally, our findings on the rheme and theme analysis suggest that this linguistic construct can be used to narrow down the selection of gestures from an LLM, focusing on those most likely to match the original speaker's actions.

Looking at these results from the perspective of building a virtual agent, the animation system Figure 3 requires a mapping between gesture labels (intents) and a defined set of animations. Therefore, limiting the model's choice of gestures as was done is aligned with such an animation system. Accordingly, we used approach 3 in our implementation. The generated gestures are in this case constrained to the gestural intents and annotations in our sample of the speakers' style of gesturing, which in some cases is a positive outcome in terms of personality/role consistency in behavior but in others may be viewed as limiting, especially compared to end-to-end machine learning based approaches to gesture animation. Though to be clear, what is limited here is the space of gestural types, not the language that triggers those gestures which is very general due to the use of an LLM. Given the fact that there is a many-to-many mapping between language and gesture, especially when considering an individual's gestural style, this arguably is not as limiting as it might seem.

7.1 Open Issues

While GPT-4 has demonstrated promising capabilities in gesture selection, its latency is too high for real-time applications in face-to-face conversations, where inference times need to be under a second, considering additional latencies in the system. To address this, we plan to explore and evaluate the fine-tuning of smaller language models to determine their suitability for integration into a virtual human framework, aiming for real-time gesture generation.

Another issue arises when there are multiple gestures in an utterance. The spatial location of gestures, and connecting one to the other becomes an important factor in ensuring the consistency of the gestures and making sure the right information is being conveyed. For example, GPT-4 sometimes identified multiple gestures where motion of one gesture is meaningfully tied to the preceding gesture, such as a container gesture along with a gesture suggesting something being removed or placed in the container [27].

This raises the issue as to the capabilities of GPT-4 to specify spatial relations between gestures in ideational units. To automatically animate, it is necessary to identify where gestures begin and end in the gestural space. The following explores this issue through a prompt that specifically requests spatial information.

```
Prompt:
In the following utterance, please identify any Image
Schemas evoked by the utterance, the Phrase of the
utterance that comprise the image schema and what
gesture might be used to illustrate that Phrase as well
as where that gesture Spatially Begins and
Spatially Ends, center, left or right.
Please provide the answer in the following format:
[Type of Image Schema:]
 [Phrase : Gesture : Spatially Begins: Spatially Ends]
 [Phrase : Gesture : Spatially Begins: Spatially Ends]
 [Phrase : Gesture : Spatially Begins: Spatially Ends]
Here is the utterance:
"For example, up in New York with SEIU-1199, you know,
coming together with workers and employers putting a
little bit of contribution from each into a trust fund
to provide for care for -- whether it's, you know,
elderly or children or whatever, to subsidize that."

GPT4:
To analyze the utterance for image schemas and propose
gestures that illustrate the phrases, we can identify
several key elements in the text:
[Type of Image Schema: CONTAINER]
["with SEIU-1199" : Cupped hands gesture :
   Spatially Begins:Center : Spatially Ends:Center]
["into a trust fund" : Gesture of placing objects into a
   container (using both hands) :
   Spatially Begins:Left : Spatially Ends:Center]
[Type of Image Schema: PATH]
["coming together" : Gesture of moving hands toward
   each other :
   Spatially Begins:Left and Right : Spatially Ends:Center]
["putting a little bit of contribution" : Gesture mimicking
    placing or dropping something along a path :
    Spatially Begins:Left : Spatially Ends:Center]
[Type of Image Schema: SUPPORT]
["provide for care" : Hands held flat, palms up, as if
     supporting something :
     Spatially Begins:Center : Spatially Ends:Center]
These gestures and their spatial dynamics help to visually
represent the conceptual ideas expressed in the text.
```

Although above illustrations of spatially coordinating gestures in ideational units are promising, more extensive testing is required.

8 CONCLUSIONS AND FUTURE WORK

The assumption underlying this work is that different designers may have different intentions so their approach to automating gesture generation may differ. In this work, we are exploring ways to automate gesture selection assuming that a designer wants to retain a degree of control over the form and the usage of the gestures.

We specifically demonstrated the use of LLMs to automate the selection of semantically rich gestures for virtual humans. As part of the effort, the approach considered ways of tailoring gesture selection to a specific individual's or role's use of gestures. We evaluated alternative prompting approaches that varied to the degree that prompting constrained gesture use. The approaches were evaluated by their semantic appropriateness (RQ1). In general, the semantic appropriateness was impressive regardless of whether the LLM was prompted to restrict its gesture use to those gestures common to a specific speaker. We evaluated as well the approaches in terms of the speed of inference (RQ1), in order to determine the suitability for incremental real-time gesture selection as opposed to offline gesture design, as well the capability of the LLM to assess when to gesture (RQ2). We additionally presented initial explorations in the ability to realize spatially consistent gestures in ideational units. Those results were very promising but need further evaluation of methods for the LLM to control the use of gestural space across gestures. And finally, the overall approach was implemented within a virtual human architecture.

Our future steps would be to explore adapting a smaller language model to enable real-time inference while also compensating for decreases in inference appropriateness.

ACKNOWLEDGMENTS

This material is based upon work supported by NSF Grant #2128743. Any opinions, findings, conclusions, or recommendations expressed are those of the authors and do not necessarily reflect the views of the NSF.

REFERENCES

[1] Janet Beavin Bavelas. 1994. Gestures as part of speech: Methodological implications. *Research on language and social interaction* 27, 3 (1994), 201–221.

[2] Geneviève Calbris. 2011. *Elements of meaning in gesture*. Vol. 5. John Benjamins Publishing.

[3] Justine Cassell, Catherine Pelachaud, Norman Badler, Mark Steedman, Brett Achorn, Tripp Becket, Brett Douville, Scott Prevost, and Matthew Stone. 1994. Animated conversation: rule-based generation of facial expression, gesture & spoken intonation for multiple conversational agents. In *Proceedings of the 21st annual conference on Computer graphics and interactive techniques*. 413–420.

[4] Justine Cassell, Matthew Stone, and Hao Yan. 2000. Coordination and context-dependence in the generation of embodied conversation. In *Proceedings of the first international conference on Natural language generation - Volume 14 (INLG '00)*. Association for Computational Linguistics, USA, 171–178. https://doi.org/10.3115/1118253.1118277

[5] 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.

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

[7] Alan J Cienki and Jean-Pierre Koenig. 1998. Metaphoric gestures and some of their relations to verbal metaphoric expressions. *Discourse and cognition: Bridging the gap* (1998), 189–204.

[8] Sharice Clough and Melissa C. Duff. 2020. The Role of Gesture in Communication and Cognition: Implications for Understanding and Treating Neurogenic Communication Disorders. *Frontiers in Human Neuroscience* 14 (2020). https://www.frontiersin.org/articles/10.3389/fnhum.2020.00323

[9] Mireille Fares, Michele Grimaldi, 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. https://hal.science/hal-03972415

[10] Gretchen N. Foley and Julie P. Gentile. 2010. Nonverbal Communication in Psychotherapy. *Psychiatry (Edgmont)* 7, 6 (June 2010), 38–44. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2898840/

[11] Udo Fries. 1984. Theme and rheme revisited. In *Modes of interpretation: Essays presented to Ernst Leisi on the occasion of his 65th birthday*. 177–192.

[12] Susan Goldin-Meadow and Martha Wagner Alibali. 2013. Gesture's role in speaking, learning, and creating language. *Annual review of psychology* 64 (2013), 257–283.

[13] Joseph Grady. 1997. Foundations of meaning: Primary metaphors and primary scenes. (1997).

[14] Joseph E. Grady. 2005. *Image schemas and perception: Refining a definition*. De Gruyter Mouton, Berlin, New York, 35–56. https://doi.org/doi:10.1515/9783110197532.1.35

[15] Bahia Guellaï, Alan Langus, and Marina Nespor. 2014. Prosody in the hands of the speaker. *Frontiers in Psychology* 5 (2014). https://www.frontiersin.org/articles/10.3389/fpsyg.2014.00700

[16] Kira Hall, Donna M. Goldstein, and Matthew Bruce Ingram. 2016. The hands of Donald Trump: Entertainment, gesture, spectacle. *HAU: Journal of Ethnographic Theory* 6, 2 (Sept. 2016), 71–100. https://doi.org/10.14318/hau6.2.009 Publisher: The University of Chicago Press.

[17] Arno Hartholt, Ed Fast, Zongjian Li, Kevin Kim, Andrew Leeds, and Sharon Mozgai. 2022. Re-architecting the virtual human toolkit: towards an interoperable platform for embodied conversational agent research and development. In *Proceedings of the 22nd ACM International Conference on Intelligent Virtual Agents*. 1–8.

[18] Laura Birka Hensel, Nutchanon Yongsatianchot, Parisa Torshizi, Elena Minucci, and Stacy Marsella. 2023. Large language models in textual analysis for gesture selection. In *Proceedings of the 25th International Conference on Multimodal Interaction*. 378–387.

[19] Autumn B. Hostetter. 2011. When do gestures communicate? A meta-analysis. *Psychological Bulletin* 137, 2 (2011), 297. https://doi.org/10.1037/a0022128 Publisher: US: American Psychological Association.

[20] Azadeh Jamalian and Barbara Tversky. 2012. Gestures alter thinking about time. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 34. 503–508.

[21] Adam Kendon. 1997. Gesture. *Annual review of anthropology* 26, 1 (1997), 109–128.

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

[23] Sotaro Kita. 2020. Cross-cultural variation of speech-accompanying gesture: A review. *Speech Accompanying-Gesture* (2020), 145–167.

[24] 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 *Intelligent Virtual Agents*, Jonathan Gratch, Michael Young, Ruth Aylett, Daniel Ballin, and Patrick Olivier (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 205–217.

[25] George Lakoff and Mark Johnson. 2008. *Metaphors we live by*. University of Chicago press.

[26] Jina Lee and Stacy Marsella. 2006. Nonverbal behavior generator for embodied conversational agents. In *International Conference on Intelligent Virtual Agents*. Springer, 243–255.

[27] Margot Lhommet and Stacy Marsella. 2014. Metaphoric gestures: towards grounded mental spaces. In *Intelligent Virtual Agents: 14th International Conference, IVA 2014, Boston, MA, USA, August 27-29, 2014. Proceedings 14*. Springer, 264–274.

[28] Margot Lhommet, Yuyu Xu, and Stacy Marsella. 2015. Cerebella: automatic generation of nonverbal behavior for virtual humans. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*. 4303–4304.

[29] William C Mann and Sandra A Thompson. 1987. Rhetorical structure theory: Description and construction of text structures. In *Natural language generation: New results in artificial intelligence, psychology and linguistics*. Springer, 85–95.

[30] 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* (Anaheim, California) *(SCA '13)*. ACM, New York, NY, USA, 25–35. https://doi.org/10.1145/2485895.2485900

[31] David McNeill. 1992. *Hand and mind: What gestures reveal about thought*. University of Chicago press.

[32] David McNeill and Susan Duncan. 2000. Growth points in thinking-for-speaking. *Language and gesture* 1987 (2000), 141–161.

[33] Michael Neff. 2016. Hand gesture synthesis for conversational characters. *Handbook of Human Motion* (2016), 1–12.

[34] Simbarashe Nyatsanga, Taras Kucherenko, Chaitanya Ahuja, Gustav Eje Henter, and Michael Neff. 2023. A Comprehensive Review of Data-Driven Co-Speech Gesture Generation. https://doi.org/10.1111/cgf.14776 arXiv:2301.05339 [cs].

[35] Demet Özer and Tilbe Göksun. 2020. Gesture use and processing: A review on individual differences in cognitive resources. *Frontiers in Psychology* 11 (2020), 573555.

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

[37] Carolyn Saund and Stacy Marsella. 2021. Gesture generation. In *The Handbook on Socially Interactive Agents: 20 years of Research on Embodied Conversational Agents, Intelligent Virtual Agents, and Social Robotics Volume 1: Methods, Behavior, Cognition*. 213–258.

[38] Susan Seizer. 2011. On the Uses of Obscenity in Live Stand-Up Comedy. *Anthropological Quarterly* 84, 1 (2011), 209–234. https://www.jstor.org/stable/41237487 Publisher: The George Washington University Institute for Ethnographic Research.

[39] Ari Shapiro. 2011. Building a character animation system. In *Motion in Games: 4th International Conference, MIG 2011, Edinburgh, UK, November 13-15, 2011. Proceedings 4*. Springer, 98–109.

[40] Marcus Thiebaux, Stacy Marsella, Andrew N Marshall, and Marcelo Kallmann. 2008. Smartbody: Behavior realization for embodied conversational agents. In *Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems-Volume 1*. 151–158.

[41] Barbara Tversky and Bridgette Martin Hard. 2009. Embodied and disembodied cognition: Spatial perspective-taking. *Cognition* 110, 1 (2009), 124–129.

[42] Fan Zhang, Naye Ji, Fuxing Gao, and Yongping Li. 2023. DiffMotion: Speech-driven gesture synthesis using denoising diffusion model. In *International Conference on Multimedia Modeling*. Springer, 231–242.

[43] Fan Zhang, Naye Ji, Fuxing Gao, Bozuo Zhao, Jingmei Wu, Yanbing Jiang, Hui Du, Zhenqing Ye, Jiayang Zhu, WeiFan Zhong, et al. 2024. DiM-Gesture: Co-Speech Gesture Generation with Adaptive Layer Normalization Mamba-2 framework. *arXiv preprint arXiv:2408.00370* (2024).