




GesGPT: Speech Gesture Synthesis With Text Parsing From ChatGPT

Nan Gao , Zeyu Zhao , Zhi Zeng , Shuwu Zhang , Dongdong Weng , and Yihua Bao 

Abstract—Gesture synthesis has gained significant attention as a critical research field, aiming to produce contextually appropriate and natural gestures corresponding to speech or textual input. Although deep learning-based approaches have achieved remarkable progress, they often overlook the rich semantic information present in the text, leading to less expressive and meaningful gestures. In this letter, we propose GesGPT, a novel approach to gesture generation that leverages the semantic analysis capabilities of large language models, such as ChatGPT. By capitalizing on the strengths of LLMs for text analysis, we adopt a controlled approach to generate and integrate professional gestures and base gestures through a text parsing script, resulting in diverse and meaningful gestures. Firstly, our approach involves the development of prompt principles that transform gesture generation into an intention classification problem using ChatGPT. We also conduct further analysis on emphasis words and semantic words to aid in gesture generation. Subsequently, we construct a specialized gesture lexicon with multiple semantic annotations, decoupling the synthesis of gestures into professional gestures and base gestures. Finally, we merge the professional gestures with base gestures. Experimental results demonstrate that GesGPT effectively generates contextually appropriate and expressive gestures.

Index Terms—Gesture synthesis, human robot interaction, large language model.

I. INTRODUCTION

GESTURE synthesis is a research field that aims to create natural and contextually appropriate gestures that correspond to given speech or textual input. The majority of existing methods adopt deep learning-based approaches, which primarily focus on audio features extracted from speech signals to model and generate gestures [1], [2]. Several of these approaches treat

gesture synthesis as a regression problem and utilize various architectures, such as convolutional neural networks [3], [4], sequence networks [5], [6], and generative models [7], [8], to learn relationships between speech and gestures.

However, these existing methods have certain limitations. Firstly, they often overlook the rich semantic information embedded in textual input, which could contribute to the generation of more expressive and meaningful gestures. For instance, an analysis of the GENE Challenge 2022 [9] has shown that the generated gestures exhibit superior human-like similarity compared to motion capture data. However, enhancing the semantic expressiveness of these gestures still requires further exploration. Secondly, the deep learning-based methods for gesture synthesis tend to yield average results, often failing to generate nuanced and intricate hand movements [10]. Lastly, they may not adequately address the inherent intention in gesture expressions, potentially leading to the generation of less plausible or contextually inappropriate gestures. Given these limitations, it is crucial to explore alternative research approaches that can leverage the wealth of information available in textual input and effectively model gesture-speech associations.

Large Language Models (LLMs), such as GPT-3 [11] and BERT [12], have demonstrated remarkable capabilities in semantic analysis, contextual understanding, and extracting meaningful information from text. These models have been applied to various natural language processing tasks with impressive results [13]. Recently, several studies have successfully employed ChatGPT in the field of robotics [14], [15] [16]. Given the potential applications of gestures in both agent and robotic domains [17], [18], it is essential to explore the utilization of LLMs in the gesture generation field.

By leveraging the robust semantic analysis capabilities of LLMs, we introduce GesGPT, a novel approach to gesture generation that focuses on text parsing using ChatGPT. While previous studies using deep learning techniques have successfully generated realistic gestures from text or speech input, these gestures often lack semantic coherence. To address this limitation, our approach incorporates prompt engineering design principles with ChatGPT to generate expressively coherent gestures from text input. Additionally, deep learning methods have proven effective in modeling the rhythmic attributes between speech and gestures [19]. Therefore, we explore the integration of existing deep learning frameworks with LLMs to mutually enhance gesture generation. Following a ‘human-on-the-loop’ paradigm, we generate semantic scripts through prompt engineering and combine them with gestures derived from deep learning methodologies.

Manuscript received 18 September 2023; accepted 15 January 2024. Date of publication 29 January 2024; date of current version 8 February 2024. This letter was recommended for publication by Associate Editor P. Carreno and Editor G. Venture upon evaluation of the reviewers’ comments. This work was supported by the National Key R&D Program of China under Grant 2022YFF0902202. (Corresponding author: Yihua Bao.)

Nan Gao is with the Institute of Automation, Chinese Academy of Sciences, Beijing 100045, China (e-mail: nan.gao@ia.ac.cn).

Zeyu Zhao is with the University of Chinese Academy of Sciences, Beijing 101408, China, and also with the Institute of Automation, Chinese Academy of Sciences, Beijing 100045, China (e-mail: zhaozeyu2019@ia.ac.cn).

Zhi Zeng and Shuwu Zhang are with the Beijing University of Posts and Telecommunications, Beijing 100876, China (e-mail: zhi.zeng@bupt.edu.cn; shuwu.zhang@bupt.edu.cn).

Dongdong Weng and Yihua Bao are with the Beijing Engineering Research Center of Mixed Reality and Advanced Display, Beijing 100081, China, and also with the Institute of Technology Beijing, Beijing 100081, China (e-mail: crgj@bit.edu.cn; boye1900@outlook.com).

This letter has supplementary downloadable material available at <https://doi.org/10.1109/LRA.2024.3359544>, provided by the authors.

Digital Object Identifier 10.1109/LRA.2024.3359544

© 2024 The Authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see <https://creativecommons.org/licenses/by-nc-nd/4.0/>

In summary, our contributions are as follows:

- We introduce GesGPT, a gesture generation framework that employs ChatGPT for text parsing to generate gestures with semantic meaning. Drawing inspiration from gesture cognition research, we break down text parsing into several tasks, such as intent classification, emphasis word identification, and semantic word recognition. This approach enables us to derive parsing scripts that guide gesture generation, resulting in a controllable and editable approach for generating gestures.
- Co-speech gesture can be considered as composed of a series of basic gesture units [20]. Inspired by this, we decompose gesture generation into two components: professional gestures and base gestures related to rhythm. Professional gestures refer to complete sequences of gesture units. Based on this, we have constructed a comprehensive specialized gesture lexicon with semantic annotations derived from video data. By integrating the gesture lexicon into the gesture generation framework, along with text parsing and utilizing a script-based search module, our method generates expressive gestures.
- We conducted extensive experiments on both English and Chinese datasets. The experimental results reveal that GesGPT effectively capitalizes on the strengths of LLMs in understanding the inherent meaning and context of text input. As a result, it produces contextually appropriate and expressive gestures that enrich the overall communication experience.

II. RELATED WORK

A. Gesture Synthesis

Recent research has shown that modeling gestures using a diffusion model [8], [21] or VQ-VAE [7], [22] can produce natural and diverse gestures. In the field of cognitive psychology [23], researchers have discovered that speech and gesture can enhance each other’s comprehension when they convey information that is semantically consistent. For modeling semantic gestures, manually designed rules have been shown to effectively preserve semantics within limited domains [24], [25]. Various deep learning approaches have been combined with rule-based methods [26], [27] or integrated with predefined gesture dictionaries [28], to produce more natural and semantically rich gestures. Additionally, Seeg [29] explored the use of a gesture semantic classifier to obtain semantic labels for gestures and ensure coherency in generated semantics. Teshima et al. [30] extensively categorized gesture types into beat, imagistic, and no-gesture following McNeill’s work [31] and modeled them separately. In our approach, we refer to McNeill’s [31] classification of gesture functions to classify text into multiple intents, and utilize text parsing scripts derived from ChatGPT to generate expressive gestures with communicative functionality.

B. LLMs in Embodied Artificial Intelligence

ChatGPT for Robotics [15] discusses the application of the ChatGPT model, which is based on natural language processing

in robotics. By incorporating prompt engineering design principles and creating a sophisticated function library, ChatGPT can be adapted to various robotic tasks and simulators, demonstrating its potential in the robotics domain. However, while LLMs possess strong analytical capabilities for existing knowledge, they face challenges in explaining non-linguistic environments such as physical settings. To address this, Huang et al.’s work [32] utilize semantic knowledge from language models and consider external environmental factors by constructing action sequences. This foundational model integrates language models and environmental factors to achieve action-based reasoning. VoxPoser [33] proposes a framework for mapping language instructions to trajectories for robot operations, achieving impressive results across multiple robot manipulation tasks. These works highlight the immense potential of introducing LLMs like ChatGPT into the field of embodied intelligence. Additionally, MotionGPT [34] proposes a unified motion-language model, which is pre-trained based on an motion vocabulary. This approach demonstrates promising results in various tasks, including motion prediction. This paper aims to explore a method for synthesizing co-speech gestures using text parsing results based on LLMs.

III. GESGPT

A. Gesture Synthesis Formulation

Deep learning-based methods typically perform gesture synthesis by dividing the process into multiple segments. Given the segmentation of the original speaker’s videos, we obtain N video segments, each comprising K frames. Body landmarks are extracted to form gesture segments $G_{i=1}^N = \{G_1, \dots, G_N\}$, where $G_i \in \mathbb{R}^{K \times D_G}$, and D_G represents the corresponding dimension. Similarly, the audio and text are temporally segmented to form audio segments $S_{i=1}^N = \{S_1, \dots, S_N\}$ and text segments $T_{i=1}^N = \{T_1, \dots, T_N\}$, where $S_i \in \mathbb{R}^{K \times D_S}$ and $T_i \in \mathbb{R}^{K \times D_T}$. Consequently, the deep learning-based gesture generation method, denoted as M , can be expressed as $G_i = M(S_i, T_i)$.

Current deep learning methods for gesture generation predominantly treat gesture synthesis as a regression problem, primarily focusing on audio input. Nevertheless, text encompasses rich semantic information. In this paper, we propose to approach gesture generation as a text classification and recognition problem. We further explore the application of LLMs for text parsing. We generate gestures grounded on the parsing script derived from purposefully designed prompts.

B. Our Method

1) *Overall Framework*: As shown in Fig. 1, our approach consists of three modules: the text parsing module that leverages LLMs such as ChatGPT, the professional gesture lexicon module with semantic annotations, and the gesture integration module based on the text parsing results. Initially, we analyze the text by applying parsing principles derived from the literature on gesture cognition and our observations of speech-style videos. This step is facilitated by utilizing pre-designed prompts in ChatGPT, which enables us to acquire text parsing results. Following the

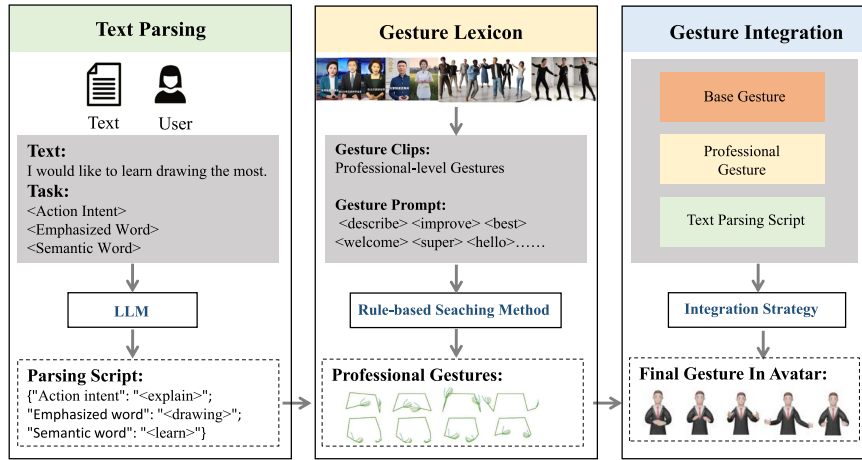


Fig. 1. Pipeline of GesGPT. Firstly, the Text Parsing module is employed to generate a gesture script from the input text. Then, the Gesture Lexicon module is used to search and retrieve corresponding professional gestures based on the script. Next, the Gesture Integration module combines the semantically professional gestures from the gesture lexicon with rhythmic base gestures.

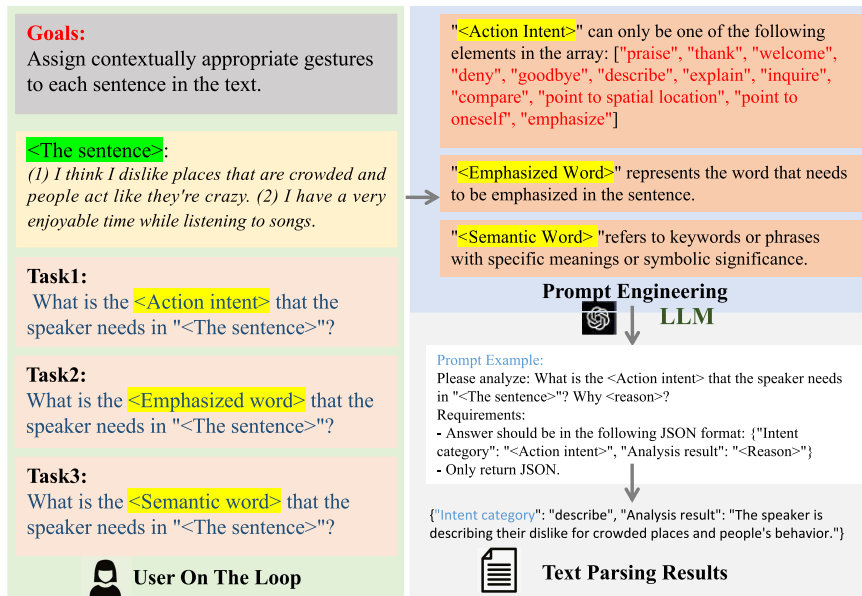


Fig. 2. Pipeline of text parsing. It involves three constituent subtasks: action intent classification, emphasis word recognition, and semantic word recognition. Finally, the results are integrated into a parsing script.

parsing of the text, the pre-defined search rules are employed to match the corresponding professional gestures from the gesture lexicon. Subsequently, within the gesture integration module, diverse gestures are integrated based on the parsing script. This framework allows for the controlled generation of meaningful and visually comprehensive gestures in accordance with the text parsing script.

2) *Text Parsing From ChatGPT:* Speech gestures can enhance speakers' ability to effectively convey information and emphasize essential points during presentations. Our research primarily focuses on exploring the utilization of LLMs to assist in generating gestures with enriched semantic expression. This process is depicted in Fig. 2. Firstly, we process one sentence

at a time and divide it into three subtasks: (i) determining the intention expressed in the sentence, (ii) identifying the words that require emphasis within the sentence, and (iii) recognizing the words that carry semantic significance. We design specific prompts for these three objectives, with expressing intention defined as a classification task, emphasizing words and semantic words defined as recognition tasks. Leveraging ChatGPT, we can analyze the deep semantic information embedded within the sentence. The obtained text parsing script using this approach enables the controlled generation of co-speech gestures that exhibit significant expressiveness.

Action Intention: To generate expressive and meaningful gestures, we classify them based on their intended meaning

using McNeill’s work [31] as a framework. McNeill categorizes gestures into five types: emblematic, iconic, metaphoric, deictic, and beat gestures. Building on this, we further divide these gesture types into several categories based on different speaking intents, including praise, thank, welcome, deny, goodbye, describe, explain, inquire, compare, point to spatial location, point to oneself, and emphasize. For instance, the iconic gesture is used to convey what is being said, and we define the corresponding textual intention as ‘describe’. Metaphoric gestures that illustrate abstract concepts are categorized as ‘explain,’ ‘inquire,’ or ‘compare’ in terms of textual intention. Subsequently, we treat the determination of textual intent as a classification task. When given a sentence as input, we prompt the LLMs to infer the most appropriate intention category, denoted as $label_1$ from this set of classification labels. The LLMs then provide a response in the format of and provide a response in the form of $\langle intention : label_1 \rangle$.

Emphasized Word: The timing of co-speech gestures varies among individuals. In this article, we make the assumption that each sentence corresponds to a specific gesture. We prompt the LLMs to identify the keyword, denoted as $\langle stroke : label_2 \rangle$, from the sentence. This allows us to insert gestures from the gesture lexicon based on the position of emphasized words, ensuring the coherence and appropriateness of generated gestures.

Semantic Word: We have observed a correlation between certain words and gestures. For example, when the word ‘excellent’ is mentioned, a thumbs-up gesture is often made. We refer to these words as semantic words and prompt the LLMs to identify the semantic word, denoted as $\langle semantic : label_3 \rangle$. Semantic gestures are characterized by specific meanings or symbolic significance. Parsing out semantic words from the text will aid in generating gestures that serve as effective auxiliary expressions.

In the text parsing module, for a sentence T , we obtain the parsing results of the three corresponding tasks, namely $\{T : \langle intention : label_1 \rangle, \langle stroke : label_2 \rangle, \langle semantic : label_3 \rangle\}$. Subsequently, we organize all the parsed results of the text into a JSON-based parsing script, which serves as the foundation for gesture generation and integration based on the script.

3) **Gesture Lexicon:** According to Kendon’s research [35], a gesture can be divided into five stages in the temporal dimension, namely rest position, preparation, stroke, hold, and retraction. Based on this, we define the basic unit in the gesture lexicon as a gesture with the aforementioned initial and final stages. Specifically, a gesture in the lexicon starts from a rest position, goes through a series of hand movements, and ends at the rest position, as illustrated in Fig. 3.

Gesture Clips: The deep learning-based methods are trained in an end-to-end manner and do not consider the completeness of generated gestures, as mentioned above, particularly in terms of complete stages. However, complete gestures have the potential to enhance the professional and detailed auxiliary expressive effect. To address this, GesGPT model generates gestures with complete stages and professionalism. We primarily extract gesture clips through two approaches: utilizing a publicly available

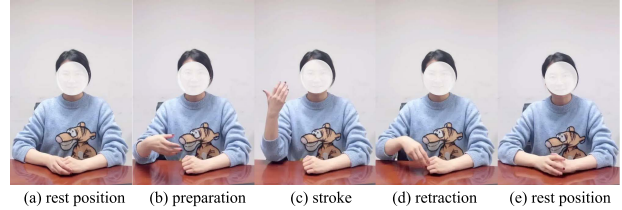


Fig. 3. Illustration of gesture clip in gesture lexicon. Gestures typically go through complete stages, including rest pose, preparation, stroke, retraction, and return to rest pose. However, not all Gesture clips have all three stages of preparation, stroke, and retraction.

dataset BEAT [36] and collecting the ZHUBO dataset obtained from internet videos. The specific introduction of these datasets will be presented in Section IV.

To extract gesture clips, we first simplify the representation of gestures using the positions of four skeletal points: the left elbow E_L , right elbow E_R , left wrist W_L , and right wrist W_R of the human body. This simplified representation is denoted as $D \in \mathbb{R}^{4 \times 3}$, where in the ZHUBO dataset, the keypoints are represented using three-dimensional spatial positions denoted as (x, y, z) coordinates. In the BEAT and motion capture datasets, the keypoints are represented using euler angles denoted as roll, pitch, and yaw. Next, we calculate the distance between adjacent frames i and $i - 1$ as $D_i - D_{i-1}$, and based on a pre-defined threshold, we filter out the starting and ending points of the gestures, thus obtaining the gesture clip. We found that this detection method performs well in extracting gesture clips in situations where gestures and static rest pose occur intermittently, such as in the ZHUBO dataset. However, it yields unsatisfactory results for continuous motion of a person’s gestures, as in the case of the BEAT dataset. Thus, we further optimized the quality of the extracted gesture clips by incorporating partial manual annotation.

Gesture Prompt: Based on our analysis of text-gesture pairs in the ZHUBO dataset, we have identified some commonalities between the visual appearance of gestures and their intended meanings. For example, explanatory gestures are often accompanied by upward or outward movements of the palm, while emphatic gestures are often accompanied by downward chopping movements of the hand. In order to enhance the functionality of our gesture lexicon, we applied text parsing methods to annotate each gesture unit with its corresponding action intention and semantic word, based on the accompanying text. Subsequently, we enlisted the expertise of a professional speaker to manually refine the original gesture labels in our dataset, thus creating our final semantically enriched gesture library. Each gesture clip is assigned an intention label, and some clips are associated with semantic words, collectively serving as prompts for the gesture clips. The example of the gesture lexicon is shown in Fig. 4.

Rule-based Searching Method: After obtaining the text parsing results, represented as $\{T : \langle intention : label_1 \rangle, \langle stroke : label_2 \rangle, \langle semantic : label_3 \rangle\}$, we employ a linear search algorithm to search for matching keywords in the gesture lexicon based on the $label_3$ in $\langle semantic : label_3 \rangle$. If no corresponding gesture prompt is found, we perform a search based on the $label_1$ in $T : \langle intention :$

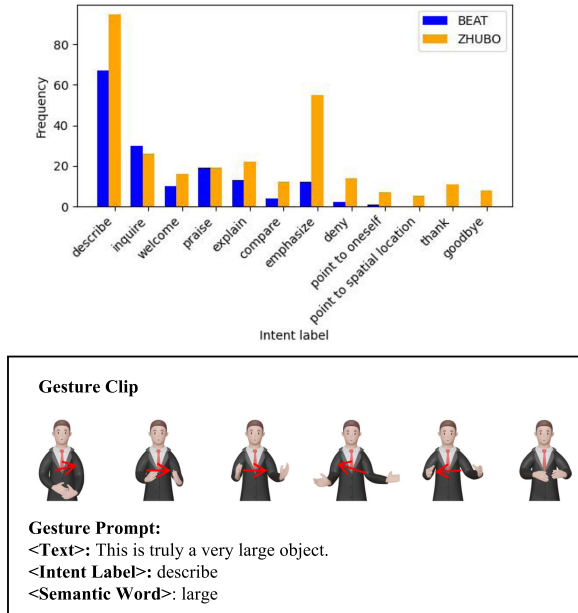


Fig. 4. Professional Gesture Lexicon. The upper graph represents the quantity of gestures belonging to different intent categories. Below is an example of a gesture clip from the lexicon.

$label_1 >$ and return the corresponding gestures. It is important to note that multiple gestures may be found during the search process. In this case, we randomly select one and integrate it based on the position of $label_2$ in the text provided by $<stroke : label_2 >$. Additionally, the text and audio are aligned, and the time of the audio can be found based on the position of the text.

4) *Gesture Integration*: In the field of gesture linguistics, the work by [20] suggests that gestures typically comprise a set of basic gesture units and random perturbations. In this study, we adopt the professional gesture defined in the gesture lexicon as the fundamental gesture unit. Additionally, we incorporate base gestures to introduce random fluctuations and enhance the authenticity of the generated gestures. For example, even in the absence of explicit gestures, the body may exhibit rhythmic oscillations. Subsequently, professional gestures with base gestures are merged to generate the final results.

Base Gesture: Kucherenko et al. [19] have shown that deep learning methods can effectively capture the association between audio and movement rhythms. In this study, we adopt a learning-based framework to model rhythmic body sway as a base gesture. We employ a one-dimensional convolutional model in the form of a U-NET for base gesture modeling. The base gesture G^B is generated under supervision by minimizing the L1 distance between the ground truth G_{GT}^B and the predicted gesture G_i^B . To reduce the jitter in the learned gestures, we incorporate a loss function based on higher-order derivatives, as proposed in [37] in (1). We utilize one-Euro algorithm filter [38] to smooth the generated results, obtaining the final base gesture.

$$L_{Base} = \frac{1}{N} \sum_{i=1}^N |G_{GT}^B - G_i^B| + |G_{GT}^{B'} - G_i^{B'}| + |G_{GT}^{B''} - G_i^{B''}| \quad (1)$$

For the blending between professional gestures G^P and base gestures G^B , we employ linear interpolation. Assuming the total number of frames for the current G^P is N , we conduct linear interpolation between the j frames preceding G^P and the j frames following its end. Here, j represents the interpolation step size, which is set to 6 for our experiment. For the front section fusion of G^P , the equation is represented as in (2):

$$G_{i-m} = \frac{1}{m} G_1^P + \left(1 - \frac{1}{m}\right) G_i^B \quad (2)$$

where $m = (1, 2, \dots, j)$ and G_1^P represents the first frame of the current G^P , with i denoting the current frame number.

For the back section fusion of G^P , it is expressed as shown in (3):

$$G_{i+m} = \frac{1}{m} G_N^P + \left(1 - \frac{1}{m}\right) G_{i+m}^B \quad (3)$$

where G_N^P represents the last frame of the current G^P . We only blend the base gesture with the frames before and after the professional gesture, and do not blend it during the frames of the professional gesture. For motion capture data such as BERT [36], we convert euler angles to quaternions for linear interpolation.

IV. EXPERIMENTS

A. Experimental Data

We conducted experiments using the BEAT dataset [36] and our proposed ZHUBO dataset, focusing on English and Chinese languages, respectively. For the BEAT dataset, we selected a segment featuring the character ‘wayne’ from the English data and divided it into 35 videos for the training set and 8 videos for the testing set. We obtained the corresponding transcripts for these videos and utilized text parsing and gesture lexicon modules to obtain the parsed text results and annotated professional gestures. Additionally, we used the MFA tool [39] to align the audio and text in the dataset, enabling us to obtain temporal information of the text, which facilitated the fusion of gestures with the speech content.

We collected Chinese host speech videos from the internet, which includes speech videos from multiple persons accompanied by professional gestures. In this paper, we selected segments featuring the character ‘kanghui’ and divided them into 68 videos for the training set and 18 videos for the testing set. We used Mediapipe [40] to obtain the skeletal positions of human bodies in the videos. Following the same methodology mentioned earlier, we obtained text with time annotations from the videos, along with parsed text results and annotated professional gestures. It is worth noting that the annotations for the ZHUBO dataset are in Chinese, and the text input from the ZHUBO dataset used in ChatGPT is also in Chinese.

B. Experimental Evaluation

1) *Objective Evaluation*: In our study, we performed an evaluation of the GesGPT and benchmarks on two datasets, BEAT and ZHUBO, respectively. For evaluation metrics, we utilized

TABLE I
OBJECTIVE EVALUATION RESULTS

Dataset	Method	BeatAlign \uparrow	FGD \downarrow
BEAT	CaMN(baseline)[36]	0.788	183.3
	DiffuseStyleGesture[21]	0.789	178.19
	QPGesture[22]	0.790	195.14
	GesGPT	0.838	173.09
ZHUBO	Seq2Seq(baseline)[42]	0.859	/
	Trimodal[41]	0.862	/
	Template-BP[3]	0.870	/
	GesGPT	0.882	/

the Beat Alignment Score introduced in CaMN [36], which is designed to evaluate the association between gesture motion and speech. Specifically, the score is defined by computing the average distance between the motion beat of each gesture B_i^G and the closest speech beat B_i^A , as in (4). We adopted the parameter $\sigma = 0.3$ setting described in [36]. In the ZHUBO dataset, the speech data is accompanied by background music.

$$BeatAlign = \frac{1}{N} \sum_{i=1}^N \exp\left(\frac{-\min |B_i^G - B_i^A|^2}{2\sigma^2}\right) \quad (4)$$

Furthermore, we utilized the Fréchet Gesture Distance (FGD) distance proposed in [41], which stands for Fréchet distance for the Gaussian mean and covariance of latent features, to quantify the similarity between two gestures' distribution features. These distribution features are defined by their mean and variance denoted by (μ_1, σ_1) and (μ_2, σ_2) , respectively, where the gesture distribution features were extracted using a pre-trained model. Specifically, we utilized the FDG distance to evaluate the similarity between generated gestures and real gestures, as in (5). We extracted the distributions using the model pre-trained on the BEAT dataset in [36], which was only tested on the BEAT dataset and not evaluated on the ZHUBO dataset.

$$FGD = |\mu_1 - \mu_2|^2 + Tr\left(\mu_1 + \mu_2 - (\mu_1\mu_2)^{\frac{1}{2}}\right) \quad (5)$$

For the BEAT dataset, we utilized the CaMN approach [36] as our baseline. This approach utilizes a sequential model to generate gestures. Regarding the ZHUBO dataset, we employed a baseline network that was trained on a bidirectional gated recurrent unit model [42]. The objective evaluation results are presented in Table I. GesGPT outperforms both baselines in terms of rhythmic consistency and realism. This method mitigates the issue of averaging effects that can arise from end-to-end network training.

2) *User Study*: We conducted a user study using the BEAT and ZHUBO datasets. From the test sets, we selected 8 videos, each approximately 20 seconds long and containing complete sentences. Users were assigned to evaluate gesture videos of Ground Truth, baseline, and GesGPT in terms of synchrony, naturalness, and semantic expressiveness of gestures, respectively. Ratings for Synchrony, Naturalness, and Semantic were collected on a scale of 1-5, with 1 indicating the lowest quality and 5 indicating the highest quality. We gathered feedback from 50 users for the BEAT dataset and 35 users for the ZHUBO

TABLE II
SUBJECTIVE EVALUATION RESULTS

Dataset	Method	Synchrony \uparrow	Naturalness \uparrow	Semantic \uparrow
BEAT	GT	3.95	3.83	3.79
	CaMN[36]	3.54	3.50	3.42
	GesGPT	3.74	3.70	3.77
ZHUBO	GT	3.29	3.36	3.21
	Seq2Seq[42]	3.16	3.14	3.02
	GesGPT	3.65	3.45	3.40

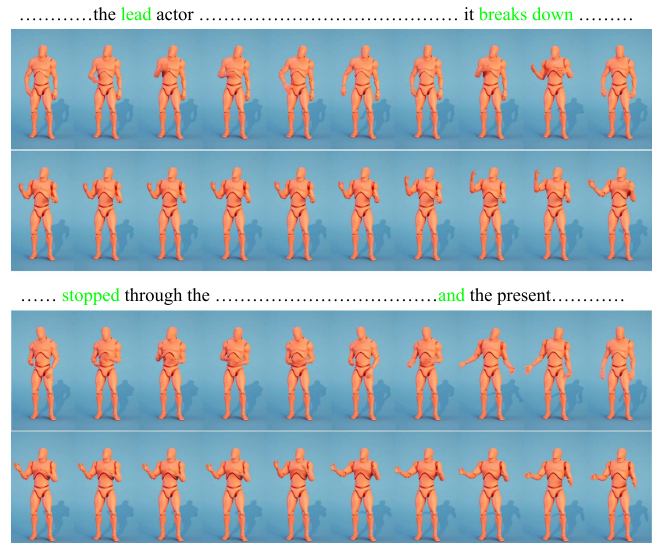


Fig. 5. Visualization of generated examples on the BEAT dataset. The first line below each text represents the generated results of GesGPT, and the second line represents the generated results of the baseline.

dataset. The subjective evaluation results are presented in Table II. We observed that GesGPT performed superiorly across all three aspects. GesGPT achieved semantic scores close to the ground truth on the BEAT dataset, and its superiority was even more pronounced on the ZHUBO dataset. This could be attributed to the fact that the character in the ZHUBO Chinese dataset predominantly remained in a resting pose state, while in the BEAT English dataset, the characters' gestures exhibited continuous variation. Additionally, the lower quality of video pose estimation in ZHUBO also affects the subjective results. Subjective experiment indicate that our method is capable of generating natural gestures that effectively convey intentions.

C. Visualization Results

We presented the visualization results of GesGPT and the baseline model on the BEAT dataset, as depicted in Fig. 5. The figure illustrates the effectiveness of both methods in generating gestures for complete sentences. To highlight the changes in gestures, we performed frame sampling on the original videos. The green segments represent the positions where the emphasized words appear in the parsed script. As research indicates that gestures often precede co-expressive speech [43], we synthesized the selected gestures at 3 frames before the emphasized words using GesGPT, while the baseline model generated

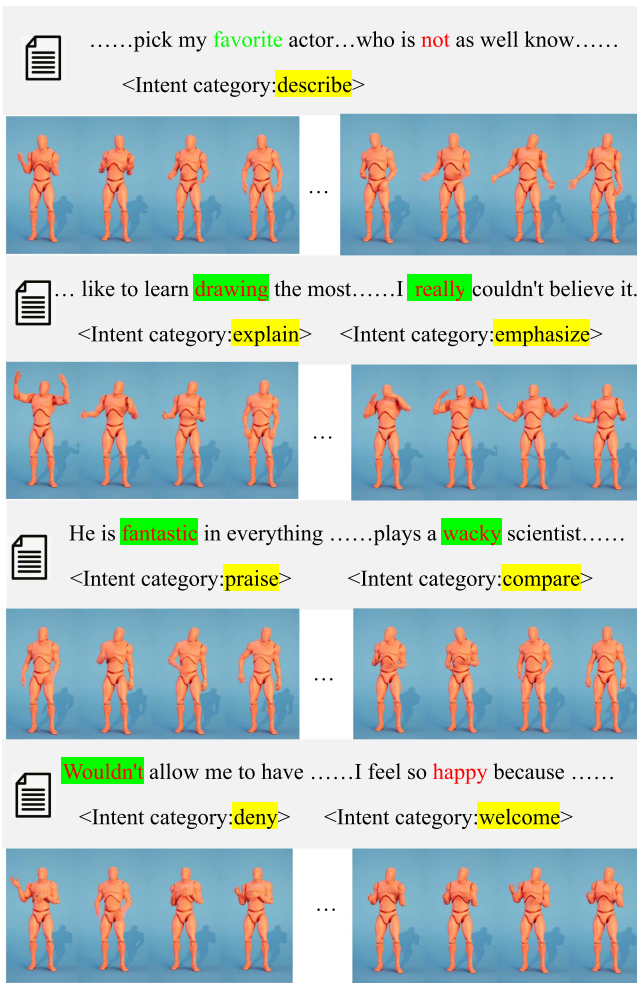


Fig. 6. Visualization results of GesGPT. Under the guidance of the text parsing script, GesGPT enables controlled generation of final gestures. Emphasis words are denoted in green, while semantic words are represented in red in the text. Additionally, the intent categories of the statements are highlighted in yellow.

gestures end-to-end using a sequence model. It is evident that GesGPT is capable of generating more diverse and comprehensive gestures, whereas the baseline model's generated results are relatively monotonous. Language models such as ChatGPT exhibit superior abilities in analyzing sentence expressions, and we propose leveraging these capabilities to generate gestures that offer enhanced auxiliary expressiveness. By creating a comprehensive and specialized gesture lexicon and employing text parsing methods, we can generate high-quality gestures that possess expressive richness.

We utilized BEAT videos as test data and employed text parsing methods to obtain action scripts. The search strategy and learning-based models described in Section III were utilized to generate corresponding gestures. The fused visualization results are shown in Fig. 6, where we assume a one-to-one mapping between sentences and intentions, meaning that each sentence can generate a professional gesture. LLMs possess powerful text semantic analysis capabilities, making them suitable for assisting in generating gestures that are more expressive. However, LLMs currently lack direct audio perception. Therefore,

we propose utilizing deep learning-based models to learn basic rhythmic gestures from a large amount of data and integrate them with meaningful gestures, providing a more optimized approach. Furthermore, this form of gesture generation based on action scripts enhances controllability, allowing targeted edits to be made to the generated results by modifying the annotations in the script.

V. CONCLUSION

In summary, we propose a new method for semantic co-speech gesture generation based on LLMs, called GesGPT. We use ChatGPT for text parsing to obtain the textual intent, emphasis words, and semantic words, and form a parsing script. Subsequently, we generate and fuse gestures based on the parsing script, leveraging a constructed annotated gesture lexicon. Experimental results on the BEAT and ZHUBO datasets demonstrate that our method can generate natural and expressively rich gestures. By effectively utilizing the capabilities of large-scale language models in text analysis and designing appropriate prompts, our method produces gestures with enhanced intentional meaning and adaptability to context. This research highlights the potential of ChatGPT in embodied intelligence and gesture synthesis, showcasing its effectiveness in generating meaningful gestures. We believe that further improvements in gesture generation can be achieved through enhanced semantic analysis of text input.

However, our study has several limitations. Firstly, we made the assumption that each sentence contains only one intent during text parsing, which may restrict the capability to express multiple intents in longer texts or texts without specific intents. Additionally, considering more contextual information beyond a single sentence during text parsing could potentially improve parsing effectiveness. Moreover, the gesture lexicon constructed in this study was based on data from specific speakers, reflecting individual styles. In future research, we aim to expand upon this work by enriching the parsing results and creating a more generalized gesture dataset.

REFERENCES

- [1] G. E. Henter, S. Alexanderson, and J. Beskow, "MoGlow: Probabilistic and controllable motion synthesis using normalising flows," *ACM Trans. Graph.*, vol. 39, no. 6, pp. 1–14, 2020.
- [2] S. Ghorbani et al., "ZeroEGGS: Zero-shot example-based gesture generation from speech," *Comput. Graph. Forum*, vol. 42, no. 1, pp. 206–216, 2023.
- [3] S. Qian et al., "Speech drives templates: Co-speech gesture synthesis with learned templates," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 11077–11086.
- [4] I. Habibie et al., "Learning speech-driven 3D conversational gestures from video," in *Proc. 21st ACM Int. Conf. Intell. Virtual Agents*, 2021, pp. 101–108.
- [5] D. Hasegawa et al., "Evaluation of speech-to-gesture generation using bi-directional LSTM network," in *Proc. 18th Int. Conf. Intell. Virtual Agents*, Sydney, 2018, pp. 79–86.
- [6] F. Yunus, C. Clavel, and C. Pelachaud, "Sequence-to-sequence predictive model: From prosody to communicative gestures," in *Proc. Int. Conf. Hum.-Comput. Interact.*, 2021, pp. 355–374.
- [7] T. Ao et al., "Rhythmic gesticulator: Rhythm-aware co-speech gesture synthesis with hierarchical neural embeddings," *ACM Trans. Graph.*, vol. 41, no. 6, pp. 1–19, 2022.

- [8] L. Zhu et al., "Taming diffusion models for audio-driven co-speech gesture generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 10544–10553.
- [9] Y. Yoon et al., "The GENE challenge 2022: A large evaluation of data-driven co-speech gesture generation," in *Proc. Int. Conf. Multimodal Interact.*, 2022, pp. 736–747.
- [10] S. Nyatsanga et al., "A comprehensive review of data-driven co-speech gesture generation," *Comput. Graph. Forum*, vol. 42, no. 2, pp. 569–596, 2023.
- [11] T. Brown et al., "Language models are few-shot learners," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2020, pp. 1877–1901.
- [12] J. Devlin et al., "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.
- [13] P. Liu et al., "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," *ACM Comput. Surv.*, vol. 55, no. 9, pp. 1–35, 2023.
- [14] N. Wake, A. Kanehira, K. Sasabuchi, J. Takamatsu, and K. Ikeuchi, "ChatGPT empowered long-step robot control in various environments: A case application," *IEEE Access*, vol. 11, pp. 95060–95078, 2023.
- [15] S. Vemprala et al., "ChatGPT for robotics: Design principles and model abilities," *Microsoft Auton. Syst. Robot. Res.*, vol. 2, 2023, Art. no. 20.
- [16] M. Parakh et al., "Human-Assisted continual robot learning with foundation models," 2023, *arXiv:2309.14321*.
- [17] R. M. Holladay and S. S. Srinivasa, "ROGUE: Robot gesture engine," in *Proc. AAAI Spring Symp. Ser.*, Palo Alto, 2016, pp. 127–134.
- [18] Y. Yoon et al., "SGToolkit: An interactive gesture authoring toolkit for embodied conversational agents," in *Proc. 34th Annu. ACM Symp. User Interface Softw. Technol.*, 2021, pp. 826–840.
- [19] T. Kucherenko et al., "Multimodal analysis of the predictability of hand-gesture properties," 2021, *arXiv:2108.05762*.
- [20] M. Kipp, *Gesture Generation by Imitation: From Human Behavior to Computer Character Animation*. Boca Raton, FL, USA: Universal-Publishers, 2005.
- [21] S. Yang et al., "DiffuseStyleGesture: Stylized audio-driven co-speech gesture generation with diffusion models," in *Proc. Int. Joint Conf. Artif. Intell.*, Macao, 2023, pp. 5860–5868.
- [22] S. Yang et al., "QPGesture: Quantization-based and phase-guided motion matching for natural speech-driven gesture generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Vancouver, 2023, pp. 2321–2330.
- [23] A. Özyürek et al., "On-line integration of semantic information from speech and gesture: Insights from event-related brain potentials," *J. Cogn. Neurosci.*, vol. 19, no. 4, pp. 605–616, 2007.
- [24] J. Cassell et al., "Animated conversation: Rule-based generation of facial expression, gesture & spoken intonation for multiple conversational agents," in *Proc. 21st Annu. Conf. Comput. Graph. Interactive Techn.*, Orlando, 1994, pp. 413–420.
- [25] P. Bremner et al., "Beat gesture generation rules for human-robot interaction," in *Proc. 18th IEEE Int. Symp. Robot Hum. Interactive Commun.*, Toyama, 2009, pp. 1029–1034.
- [26] C. Zhou, T. Bian, and K. Chen, "GestureMaster: Graph-based speech-driven gesture generation," in *Proc. Int. Conf. Multimodal Interact.*, 2022, pp. 764–770.
- [27] I. Habibie et al., "A motion matching-based framework for controllable gesture synthesis from speech," in *Proc. ACM SIGGRAPH Conf.*, Vancouver, 2022, pp. 1–9.
- [28] S. Zhang, J. Yuan, M. Liao, and L. Zhang, "Text2Video: Text-driven talking-head video synthesis with personalized phoneme-pose dictionary," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Singapore, 2022, pp. 2659–2663.
- [29] Y. Liang et al., "SEEG: Semantic energized co-speech gesture generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, New Orleans, 2022, pp. 10473–10482.
- [30] H. Teshima, N. Wake, D. Thomas, Y. Nakashima, H. Kawasaki, and K. Ikeuchi, "Deep gesture generation for social robots using type-specific libraries," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Kyoto, 2022, pp. 8286–8291.
- [31] D. McNeill, *Gesture and Thought*. Chicago, IL, USA: Univ. Chicago Press, 2019.
- [32] W. Huang et al., "Grounded decoding: Guiding text generation with grounded models for robot control," 2023, *arXiv:2303.00855*.
- [33] W. Huang et al., "VoxPoser: Composable 3D value maps for robotic manipulation with language models," 2023, *arXiv:2307.05973*.
- [34] B. Jiang et al., "MotionGPT: Human motion as a foreign language," 2023, *arXiv:2306.14795*.
- [35] A. Kendon, "Some relationships between body motion and speech," *Stud. Dyadic Commun.*, vol. 7, 1972, Art. no. 90.
- [36] H. Liu et al., "BEAT: A large-scale semantic and emotional multi-modal dataset for conversational gestures synthesis," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 612–630.
- [37] S. Li et al., "Bailando: 3D dance generation by actor-critic GPT with choreographic memory," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, New Orleans, 2022, pp. 11050–11059.
- [38] G. Casiez, N. Roussel, and D. Vogel, "1€ Filter: A simple speed-based low-pass filter for noisy input in interactive systems," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, New York, 2012, pp. 2527–2530.
- [39] M. McAuliffe et al., "Montreal forced aligner: Trainable text-speech alignment using Kaldi," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2017, pp. 498–502.
- [40] C. Lugaresi et al., "MediaPipe: A framework for building perception pipelines," 2019, *arXiv:1906.08172*.
- [41] Y. Yoon et al., "Speech gesture generation from the trimodal context of text, audio, and speaker identity," *ACM Trans. Graph.*, vol. 39, no. 6, pp. 1–16, 2020.
- [42] K. Cho et al., "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in *Proc. Int. Conf. Empirical Methods Natural Lang. Process.*, 2014, pp. 1724–1734.
- [43] S. Nobe, "Where do most spontaneous representational gestures actually occur with respect to speech," *Lang. Gesture*, vol. 2, 2000, Art. no. 4.