SARGes: Semantically Aligned Reliable Gesture Generation via Intent Chain

Nan Gao*
nan.gao@ia.ac.cn
Institution of Automation, Chinese Academy of Sciences
Beijing, China

Yihua Bao*
boye1900@outlook.com
Beijing Institute of Technology
Beijing, China

Dongdong Weng
crgj@bit.edu.cn
Beijing Institute of Technology
Beijing, China

Jiayi Zhao
zjyjy@bit.edu.cn
Beijing Institute of Technology
Beijing, China

Jia Li
lijia35@lenovo.com
Lenovo Research
Beijing, China

Yan Zhou
zhouyan03@kuaishou.com
Kuaishou Technology
Beijing, China

Pengfei Wan
wanpengfei@kuaishou.com
Kuaishou Technology
Beijing, China

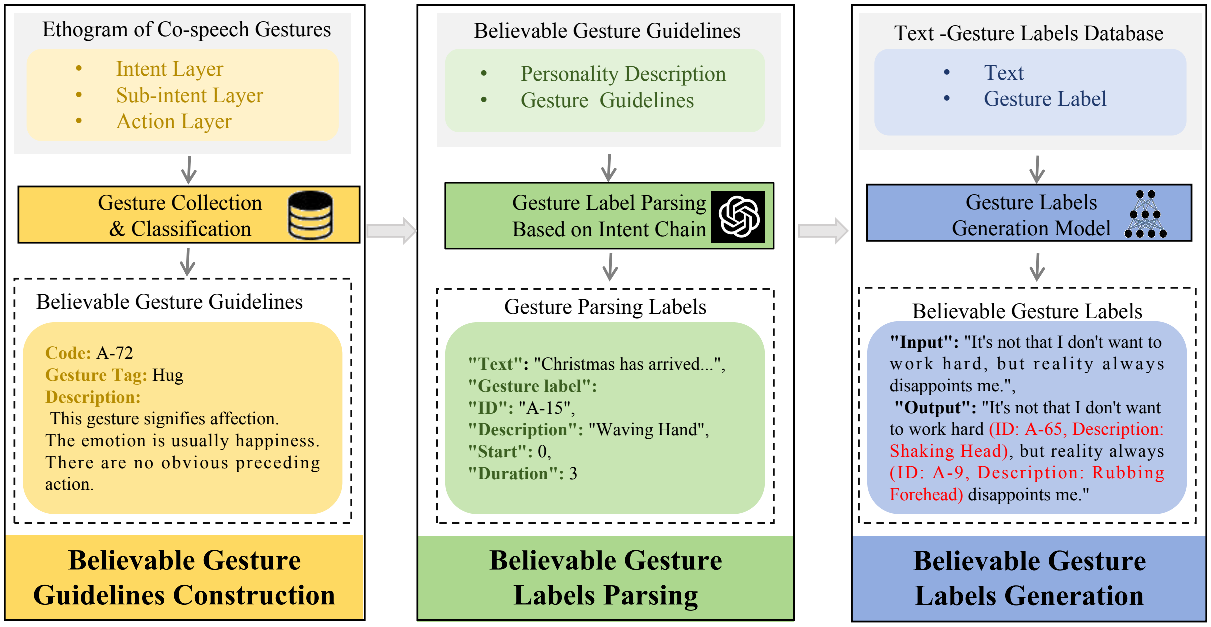

Ethogram of Co-speech Gestures
• Intent Layer
• Sub-intent Layer
• Action Layer
Gesture Collection & Classification
Believable Gesture Guidelines
Code: A-72
Gesture Tag: Hug
Description:
This gesture signifies affection. The emotion is usually happiness. There are no obvious preceding action.
Believable Gesture Guidelines Construction
Believable Gesture Guidelines
• Personality Description
• Gesture Guidelines
Gesture Label Parsing Based on Intent Chain
Gesture Parsing Labels
"Text": "Christmas has arrived...",
"Gesture label":
"ID": "A-15",
"Description": "Waving Hand",
"Start": 0,
"Duration": 3
Believable Gesture Labels Parsing
Text -Gesture Labels Database
• Text
• Gesture Label
Gesture Labels Generation Model
Believable Gesture Labels
"Input": "It's not that I don't want to work hard, but reality always disappoints me.",
"Output": "It's not that I don't want to work hard (ID: A-65, Description: Shaking Head), but reality always (ID: A-9, Description: Rubbing Forehead) disappoints me."
Believable Gesture Labels Generation

Figure 1: Pipeline for Semantic Gesture Label Generation. We propose an intent-chain mechanism based on LLMs to parse input text and its underlying semantic labels. A lightweight model is trained to generate structured semantic labels that effectively guide the subsequent process of Semantically Aligned Reliable Gesture Generation (SARGes).

Abstract

Co-speech gesture generation enhances human-computer interaction realism through speech-synchronized gesture synthesis. However, generating semantically meaningful gestures remains a challenging problem. We propose SARGes, a novel framework that leverages large language models (LLMs) to construct an intent

*Both authors contributed equally to this research.

Conference acronym 'XX, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

chain for parsing speech content and generating reliable semantic gesture labels, which subsequently guide the synthesis of meaningful co-speech gestures. First, we constructed a comprehensive co-speech gesture ethogram and developed an LLM-based intent chain reasoning mechanism that systematically parses and decomposes gesture semantics into structured inference steps following ethogram criteria, effectively guiding LLMs to parse context-aware gesture labels. Subsequently, we constructed a text-to-gesture label dataset and trained a lightweight gesture label generation model, which then guides the generation of credible and semantically coherent co-speech gestures. Experimental results show that SARGes achieves gesture labeling performance comparable to GPT-4 in intent interpretation, with efficient single-pass inference (0.4 seconds), and significantly improves the semantic expressiveness of gesture generation.

Keywords

Co-speech Gesture Generation, Large Language Models, Gesture Ethogram, Intent Chain

ACM Reference Format:

Nan Gao, Yihua Bao, Dongdong Weng, Jiayi Zhao, Jia Li, Yan Zhou, and Pengfei Wan. 2018. SARGes: Semantically Aligned Reliable Gesture Generation via Intent Chain. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX).* ACM, New York, NY, USA, 9 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Gestures in human communication significantly enhance semantic transmission, emotional expression, and conversational flow [1]. Co-speech gesture generation focuses on enabling virtual characters or robots to produce synchronized natural gestures alongside speech, advancing human-computer interaction toward greater naturalness and intelligence [2] [3]. This is crucial in applications requiring enhanced realism and immersion, such as virtual agents [4] and humanoid robots [5], where natural and context-appropriate gestures greatly improve interaction quality.

Recent research on gesture generation has increasingly focused on leveraging large datasets and models. Alexanderson et al. [6] showed that diffusion models are effective for synthesizing human motions, such as dancing and co-speech gestures. DiffMotion [7] integrates an autoregressive temporal encoder with a denoising diffusion model to produce high-fidelity speech-synchronized gestures. Building on this, Diffsheg [8] combines diffusion models with Transformers for the real-time generation of 3D gestures and facial expressions. In multimodal applications, GestureDiffuCLIP [9] uses a CLIP encoder and Adaptive Instance Normalization (AdaIN) to generate a variety of gestures flexibly. MotionGPT [10] employs a pre-trained language model to interpret human motion as a foreign language, facilitating gesture generation tasks. While methods in [6] and [10] can produce semantically meaningful actions like "running" and "jumping," there is still a gap in generating semantic gestures specifically for co-speech scenarios. Moreover, despite advances in aligning gestures with speech rhythm, diffusion models and other learning-based approaches still lack semantic understanding. Therefore, our research aims to generate meaningful gesture labels in co-speech scenarios to enhance the semantic richness of gesture generation.

Large Language Models (LLMs) are proficient at extracting semantic information and have been leveraged in gesture generation research to improve the semantic consistency of generated gestures. For instance, GesGPT [11] utilizes the semantic analysis capabilities of GPT to generate meaningful gestures from text by categorizing predefined gesture intentions. However, systematic research on defining and categorizing semantic gestures for co-speech generation is limited, and much of the existing work relies on individually defined semantic gesture libraries [12]. Furthermore, LLMs are prone to hallucinations [13], which can reduce the reliability of text-based semantic gesture generation. In ethology, ethograms are systematically organized according to specific standards to describe and understand behavior patterns [14]. Given that co-speech gestures are human behaviors, we adopt the concept of ethograms as a tool for gesture classification. We construct a co-speech gesture ethogram with defined guidelines and develop an LLM-based intent chain method, employing these ethogram criteria as constraints. This approach helps reduce model hallucinations and enhances the reliability of the generated gesture labels.

In summary, our contributions are as follows:

- Inspired by research in animal behavior [14] and gesture cognition [1], we developed a hierarchical ethogram for co-speech gestures, categorizing them based on their intentions. This ethogram serves as a systematic tool for managing semantic gestures, providing detailed descriptions of the meaning and usage guidelines of each gesture.
- We designed an intent chain reasoning mechanism to interact with LLMs and parse gesture labels from text, utilizing techniques like chain-of-thought and self-reflection when generating gesture labels with GPT. By leveraging gesture guidelines from the ethogram as auxiliary information, we enhance the model's semantic parsing capabilities, resulting in more reliable gesture labels and reduced hallucinations.
- We constructed a dedicated dataset for gesture label generation and fine-tuned a language model on textual inputs. Experimental results demonstrate the effectiveness of this approach in producing accurate and semantically meaningful gesture labels, which serve as reliable control signals for subsequent gesture synthesis, significantly improving the semantic coherence and expressiveness of the generated gestures.

2 RELATED WORKS

2.1 Semantic Gesture Generation

Deep learning methods using multimodal inputs, like text and speech, have been employed to generate gestures by learning features from the text, providing richer semantic information [15][16]. However, these methods struggle to generate semantic gestures due to the limited examples of semantic gestures in the datasets. To further enhance the quality of semantic gesture generation, Seeg [17] requires the predicted results to express the same semantics as the ground truth labels. The method uses semantic labels that are categorized into five classes based on the semantics and emotional characteristics of gestures. Teshima et al. [18], based on McNeill's work [1], classify gestures into beats, imagistic, and no-gesture types, mapping input text words to corresponding gestures to generate specific ones. The Text2Gestures [19] model posits that each text sentence is associated with an expected emotion, primarily represented in the VAD (Valence, Arousal, Dominance) space. Methods using motion graphs [20][21] and retrieval-based subsystems [22] retrieve motion segments from a predefined action library that best match the text's semantics and speech rhythm. GesGPT [11] is the first to use LLMs for text intent classification, linking intent labels with semantic gestures. This method greatly improves the flexibility and accuracy of semantic gesture generation. Semantic Gesticulator [12] introduces a semantic-aware co-speech gesture synthesis system that utilizes a GPT-based generator along with a semantic alignment mechanism to ensure the quality and coherence of the generated semantic gestures. However, current studies often rely on custom categories or specific databases, limiting coverage. To address this, we plan to use the ethogram to create a gesture set for co-speech scenarios and explore LLM-based gesture label generation.

2.2 LLMs in Embodied Agents

Story-to-Motion [23] employs LLMs to extract action sequences from extended texts. This method integrates action retrieval with semantic and trajectory constraints to generate character motions that are natural and controllable. In the Digital Life Project [24], researchers proposed a framework that uses language to construct autonomous 3D characters. This approach extends action generation beyond Story-to-Motion by filtering action candidates with high-level textual semantics and refining them with kinematic features, such as joint positions, to align with the scene and character state. Another study [25] translates LLM-generated action plans into executable actions by using pretrained models to match action phrases with permissible actions, ensuring semantic consistency. The Generative Agents framework [26] extends LLMs to store comprehensive experience records of agents, enabling them to form higher-level reflections based on these memories and dynamically retrieve them for behavior planning. This approach allows agents to exhibit realistic social behaviors in simulated environments, enhancing the authenticity of interactions. Collectively, these works demonstrate that LLMs can effectively parse human behavior-related information from text and convert it into executable actions, thereby serving as a viable approach for achieving semantically gesture generation.

3 PROPOSED METHOD

This study proposes a method using LLMs to generate gesture labels from text. Given a text x, the method generates the corresponding gesture label y, where y belongs to a predefined gesture ethogram G. This relationship is modeled by $y = M_{\Phi}(x)$, with $y \in G$.

3.1 Pipeline

Firstly, as shown in Fig.1, we established a gesture ethogram for co-speech scenarios, which includes a gesture classification hierarchy and usage guidelines. These guidelines enhance the reliability of gesture labels by providing clear guidance for GPT text parsing. We designed an intent chain reasoning method based on LLMs for behavior intention inference, employing chain-of-thought [27] and self-reflective [28] strategies to generate gesture label parsing results aligned with the text. Subsequently, based on the parsed labels, we constructed a dataset to train a gesture label generation model. These labels can then be used in conjunction with motion matching or motion generation techniques to enhance semantic gesture generation in applications such as virtual digital humans and social robots.

3.2 Believable Gesture Guidelines Construction

This section utilizes the ethogram construction methodology to categorize and collate gestures commonly observed in co-speech scenarios. We then provide comprehensive descriptions of each categorized gesture, along with specific usage guidelines.

3.2.1 Ethogram of Co-speech Gestures. In ethological research, ethograms are commonly used to identify patterns and regularities in animal behavior [14][29]. We plan to construct an ethogram for co-speech gestures, systematically classifying and organizing gestures that occur in co-speech scenarios. Following the hierarchical construction strategy used in ethological research [29], we organize the ethogram of co-speech gestures into three layers: the Intent Layer, the Sub-intent Layer, and the Action Layer. This layered structure effectively transitions from behavioral motivations to specific gestures.

Intent Layer: This layer provides a high-level abstraction of the motivations behind co-speech gestures. Based on McNeill's classification of gesture functions [1], we categorize gesture motivations into four types: *Information Display*, *Concrete Reinforcement*, *Tone Reinforcement*, and *Comfort Behaviors*.

Information Display gestures, similar to what McNeill describes as emblematic gestures, convey meanings on their own, separate from speech, like showing emotions or depicting certain looks. *Concrete Reinforcement* includes iconic, metaphoric, and pointing gestures, which help verbal communication by showing attributes like direction, shape, or position. Beat gestures fit under *Tone Reinforcement*, as they express emotions through emphasis, rhythm, or a questioning tone. *Comfort Behaviors* are gestures that help alleviate emotions such as tension, anxiety, or sadness. These gestures are also common in human-computer interaction; therefore, we have included this category in the intent layer.

Sub-intent Layer: This layer refines the Intent Layer by breaking down broader intents into more specific sub-intents. For instance, within the *Comfort Behaviors* category, sub-intents can include alleviating tension, fear, anxiety, excitement, boredom, sadness, embarrassment, and anger, among others.

Action Layer: This layer focuses on specific gestures that directly reflect the intents of the corresponding layers. For example, gestures like *spreading arms wide*, *waving hands*, and *two-handed applause*. The complete gesture spectrum can be found at https://github.com/gesture-label/ethogram.

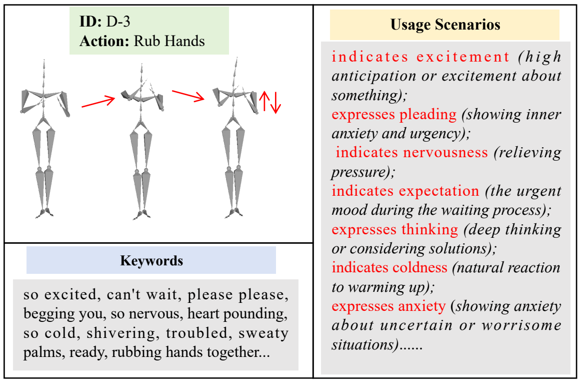
By collecting and classifying a wide range of gestures from communication scenarios according to the defined ethogram, we can systematically organize gestures in co-speech contexts. New gestures can be categorized within the ethogram based on their definitions, ensuring a clear rationale for classification. This approach allows for continuous refinement as more data and new gestures are added. Additionally, we assign an index to gestures based on their distinct motivations, enabling precise management and retrieval of gesture data, as illustrated in Table 1.

Table 1: Ethogram of Co-speech Gestures

Intent Layer	Sub-intent Layer	Action Layer
Information Display	*Display Appearance* *Display Special Meaning* ...	*A-1: Stretch Shoulders* *A-2: Thumbs Down* ...
Concrete Reinforcement	*Concrete Direction* *Concrete Shape* ...	*B-1: Point Finger in Target Direction* *B-2: Form Hands into a Circle* ...
Tone Reinforcement	*Questioning Tone* *Emphasizing Tone* ...	*C-1: Wave Palm Upwards* *C-2: Shake Interlocked Fists* ...
Comfort Behaviors	*Soothe Nervousness* *Soothe Fear* ...	*D-1: Rub or Pinch Fingers* *D-2: Cover Eyes with Hands* ...

3.2.2 Believable Gesture Guidelines . Based on the co-speech gesture ethogram, we collected over 200 co-speech gestures. We will establish guidelines by analyzing their usage and identifying keywords frequently associated with each gesture. These guidelines will serve as auxiliary information in subsequent LLM-based gesture label parsing, enhancing the reliability of generating gesture labels corresponding to text.

We initially employed prompt engineering with ChatGPT to generate usage descriptions for all gestures within the ethogram, focusing on contextual and emotional correlations. To enhance output consistency, example texts were provided as inputs. Following the generation of descriptions, we conducted manual reviews and corrections to ensure accuracy and uniformity. This process culminated in the development of the Believable Gesture Guidelines, which provide a robust theoretical foundation for gesture label generation. Fig. 2 offers a partial illustration of the gesture guideline.



indicates excitement *(high anticipation or excitement about something);*
expresses pleading *(showing inner anxiety and urgency);*
indicates nervousness *(relieving pressure);*
indicates expectation *(the urgent mood during the waiting process);*
expresses thinking *(deep thinking or considering solutions);*
indicates coldness *(natural reaction to warming up);*
expresses anxiety (*showing anxiety about uncertain or worrisome situations)......*

so excited, can't wait, please please, begging you, so nervous, heart pounding, so cold, shivering, troubled, sweaty palms, ready, rubbing hands together...

Figure 2: Guidelines Illustration for the 'Rub Hands' Gesture

3.3 Believable Gesture Labels Parsing

By applying gesture guidelines as constraints, we ensure that the labels parsed from the intent chain based on LLMs are appropriate and believable for the given text.

3.3.1 Chain-of-Thought Gesture Prompting. Chain-of-thought (CoT) prompting improves model reasoning by encouraging intermediate steps in problem-solving [27]. We designed prompts using the CoT approach to interact with ChatGPT. Our aim is for the model to process text segments step-by-step, identify relevant keywords, and select appropriate gestures from the ethogram based on predefined guidelines.

First, we establish the virtual agent's character profile as a foundation for LLM interactions. Then, the LLM clarifies the conversation theme and identifies the speaker's primary intent. Based on this intent, the model analyzes keywords from the text and, considering the character's personality, selects appropriate gestures from the defined gesture guidelines. These gestures are then inserted before the identified keywords. This approach allows the language model to select the most suitable gestures, guided by character traits and context. By clearly defining each step and decision criterion, it helps reduce the occurrence of model hallucinations and enhances the credibility of text parsing.

3.3.2 Gesture Prompting with Self-reflection. Self-reflection can significantly enhance decision-making capabilities in agent interaction tasks [28]. In our research, we integrate this mechanism into prompt engineering. Utilizing the CoT prompting strategy, we introduce multiple rounds of gesture selection and reflection, allowing LLMs to re-evaluate and refine their results based on established rules. The reflection rules encompass semantic relevance and action relevance.

The semantic relevance evaluation includes: (i) Context Matching: Checking if body movements align with the speaker's identity, theme, and setting, such as using open gestures in positive contexts or composed postures in serious discussions. (ii) Keyword Matching: Assessing if gestures semantically match the keywords. (iii) Consistency of Emotional Expression: Evaluating if gestures align with the emotional tone, like using faster movements for exciting or joyful topics.

The action relevance evaluation includes: (i) Positional Consistency: Evaluating whether the placement of gestures aligns with the position of the keywords. (ii) Moderation: Evaluating the frequency of gestures to ensure they are not too frequent, with no more than two gestures in a single sentence.

3.3.3 Gesture Parsing Labels. Each gesture label includes: a unique Gesture ID, such as 'A-15' for the 15th gesture in *Information Display*; the gesture's name; its start position in the text, measured in characters (e.g., '0' for the beginning); the duration, also in characters (e.g., '5' for five characters). Table 2 presents the key prompt used for gesture label parsing.

3.4 Believable Gesture Labels Generation

We employed GPT to parse gesture labels and build a text-gesture labels dataset. This dataset facilitated the training of a model focused on reliable, cost-efficient gesture label generation.

3.4.1 Text-Gesture Labels Database. We collected various speech videos from the internet and utilized speech-to-text technology to extract the corresponding text, forming a corpus in which each sentence is treated as a distinct unit. Using the gesture label parsing

Table 2: LLM Prompt Structure for Gesture Labeling

Component	Description
Role	You are a speech expert who deeply understands the meaning and timing of body language, designing gestures for a speech sentence-by-sentence.
Goal	(1) Analyze the semantics of the input sentence. (2) Assign appropriate gestures to different parts of the sentence.
Constraints	You may only choose gestures from the provided *Ethogram of Co-speech Gestures* (each with a numeric ID).
Reasoning Steps	1) Extract the speaker's main intent. 2) Identify keywords in the sentence. 3) Select gestures according to the theme and keywords. 4) Insert the gesture before the keyword's position. 5) Return the selected gestures in the format `(id: gesture ID, description: gesture name)` and output the final annotated text.
Example Input	Hello! I'm very glad you could come today. You are truly amazing!
Example Output	Hello! I'm very (id: A-10, description: open arms) glad you could come today. You are truly (id: A-9, description: applause) amazing!

method previously discussed, we annotated this corpus with gesture labels. The labels are formatted as follows:

Input: "Hello, it's great to have you here today. You are truly amazing!"
Output: "Hello, it's great (id: A-97, description: spreading arms wide) to have you here today. You are truly (id: A-6, description: clapping) amazing!"

Through this approach, we constructed the training dataset for the gesture label generation model.

3.4.2 Gesture Labels Generation Model. LLMs trained on vast datasets have demonstrated exceptional performance in various natural language processing tasks, including translation [30]. We conceptualize the training of the gesture label generation model as a text translation task, where the input consists of the current dialogue text, and the output is the text annotated with gesture labels. We fine-tune open-source LLMs to exploit their robust semantic processing capabilities, thereby enhancing their adaptability to the gesture label generation task.

For the gesture label generation task, the training dataset is defined as: $Z = \{(x_i, y_i)\}_{i=1,..,N}$, where x_i is the natural query input, and y_i is the text with gesture labels. We fine-tune a language model Qwen based on the transformer architecture [30], represented as $M_{\Phi_0}(y|x)$, with parameters Φ_0. LoRA (Low-Rank Adaptation) is a method that achieves model compression and acceleration through low-rank decomposition [31]. This approach uses low-rank matrices to replace or adjust the weight matrices of the original model, achieving efficient parameterization and significantly reducing resource consumption in training and inference. We apply LoRA technique to the q_proj and v_proj modules of the Qwen model by inserting low-rank matrices to adjust their outputs. The goal is to maximize the conditional log-likelihood of the language model, enabling it to generate gesture labels. The optimization problem can be expressed as (1):

$$\max_{\Theta} \sum_{(x,y)\in Z} \sum_{i=1}^{|y|} \log\left(M_{\Phi_0+\Delta\Phi(\Theta)}(y_i \mid x, y_{<i})\right) \tag{1}$$

where $M_{\Phi_0+\Delta\Phi(\Theta)}(y_t \mid x, y_{<i})$ represents the conditional probability computed with the current parameters $\Phi_0 + \Delta\Phi(\Theta)$. To improve computational and memory efficiency, $\Delta\Phi(\Theta)$ can utilize the low-rank representation proposed by the LoRA method, significantly reducing the number of parameters and computational load during training.

4 Experiments

4.1 Gesture Label Parsing

Test Dataset: We designed a specific text test dataset based on gestures from the created ethogram. Specifically, we selected 77 representative gestures from the ethogram and constructed relevant text examples according to the semantic descriptions of each gesture. For example, the gesture 'rubbing hands,' described as 'expressing excitement or pleading,' is paired with the example sentence, *'I can hardly wait, this news has me so excited!'* which is semantically aligned with the action of 'rubbing hands.' Similarly, we designed three highly related example sentences for each gesture, resulting in a total of 231 sentences that make up the test dataset. These sentences are paired with their corresponding gesture labels and are used to evaluate the performance of the prompt-based method in parsing the alignment between text and gesture labels.

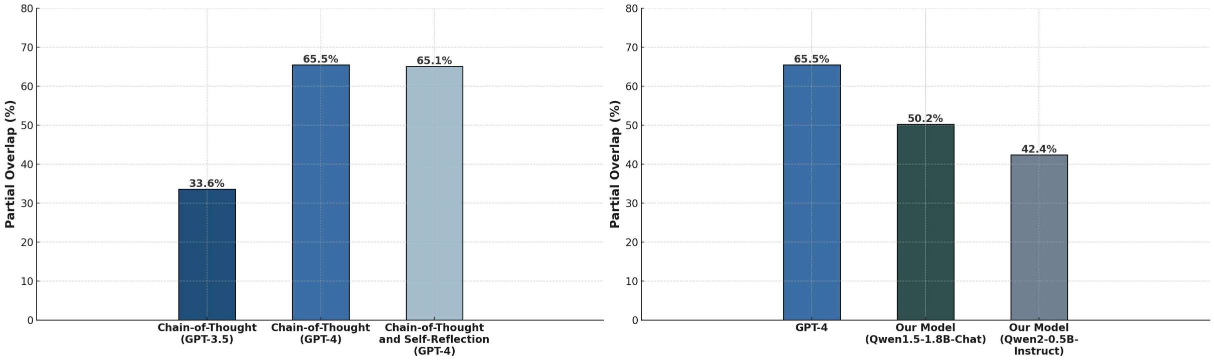
Evaluation Metric: We designed the *Partial Overlap* metric to measure the accuracy of the gesture labels $\hat{y}$ in the parsing results relative to the true labels y, as defined by Equation (1). Given that we include 77 gesture labels, with a sentence potentially corresponding to multiple labels, we addressed this complexity by clustering the labels into five emotion-based categories: 'joy,' 'anger,' 'sorrow,' 'fear,' and 'special meaning.' The 'special meaning' category includes gestures with unique meanings, such as 'bunny ears.' In accuracy evaluation, we map both the gesture label parsing results and the test data labels to these five categories and then calculate the Partial Overlap on the mapped results using (2).

$$\text{Partial Overlap} = \frac{\sum_{i=1}^{N} |\hat{y}_i \cap y_i|}{\sum_{i=1}^{N} |y_i|} \tag{2}$$

We employed CoT and self-reflection prompt engineering methods for text-based gesture label parsing and compared the text parsing performance between GPT-3.5 and GPT-4 models. In the self-reflection prompting method, the results were iterated three times based on the reflection criteria mentioned above. The results are shown in Fig. 3.

Based on the experimental results, it is clear that the accuracy of gesture label parsing using GPT-3.5 is substantially lower compared to GPT-4, highlighting the necessity of a more advanced model for enhanced semantic comprehension in this task. Specifically, using GPT-4 for annotation yielded an accuracy of 65.5%. However, incorporating self-reflection resulted in a marginal decrease in accuracy compared to the CoT prompt method.

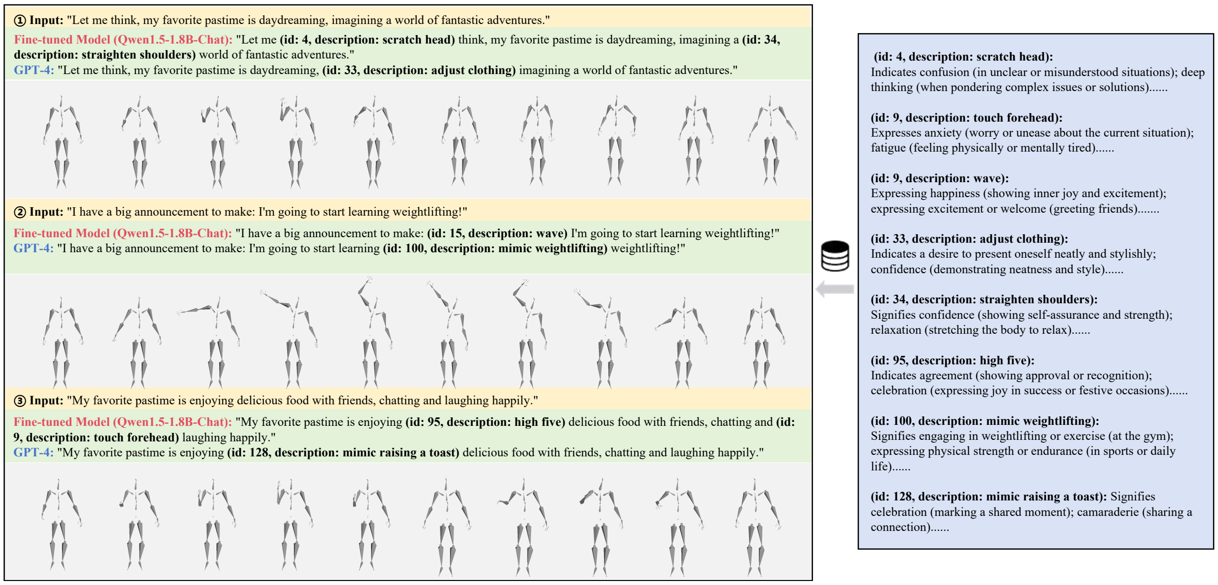
In tasks with objective evaluation criteria, such as programming (e.g., successful code compilation) [33], self-reflection can effectively enhance the model's decision-making process. However, for gesture generation, an ill-posed problem lacking clear evaluation standards, the reflection mechanism can introduce additional uncertainty, potentially leading to increased variability in the results and limiting improvements in decision-making accuracy.



(a) Objective Evaluation of Gesture Label Parsing

(b) Objective Evaluation of Gesture Label Generation

Figure 3: Objective Evaluation Results



① **Input:** "Let me think, my favorite pastime is daydreaming, imagining a world of fantastic adventures."

Fine-tuned Model (Qwen1.5-1.8B-Chat): "Let me **(id: 4, description: scratch head)** think, my favorite pastime is daydreaming, imagining a **(id: 34, description: straighten shoulders)** world of fantastic adventures."
GPT-4: "Let me think, my favorite pastime is daydreaming, **(id: 33, description: adjust clothing)** imagining a world of fantastic adventures."

② **Input:** "I have a big announcement to make: I'm going to start learning weightlifting!"

Fine-tuned Model (Qwen1.5-1.8B-Chat): "I have a big announcement to make: **(id: 15, description: wave)** I'm going to start learning weightlifting!"
GPT-4: "I have a big announcement to make: I'm going to start learning **(id: 100, description: mimic weightlifting)** weightlifting!"

③ **Input:** "My favorite pastime is enjoying delicious food with friends, chatting and laughing happily."

Fine-tuned Model (Qwen1.5-1.8B-Chat): "My favorite pastime is enjoying **(id: 95, description: high five)** delicious food with friends, chatting and **(id: 9, description: touch forehead)** laughing happily."
GPT-4: "My favorite pastime is enjoying **(id: 128, description: mimic raising a toast)** delicious food with friends, chatting and laughing happily."

(id: 4, description: scratch head):
Indicates confusion (in unclear or misunderstood situations); deep thinking (when pondering complex issues or solutions)......

(id: 9, description: touch forehead):
Expresses anxiety (worry or unease about the current situation); fatigue (feeling physically or mentally tired)......

(id: 9, description: wave):
Expressing happiness (showing inner joy and excitement); expressing excitement or welcome (greeting friends).......

(id: 33, description: adjust clothing):
Indicates a desire to present oneself neatly and stylishly; confidence (demonstrating neatness and style)......

(id: 34, description: straighten shoulders):
Signifies confidence (showing self-assurance and strength); relaxation (stretching the body to relax)......

(id: 95, description: high five):
Indicates agreement (showing approval or recognition); celebration (expressing joy in success or festive occasions)......

(id: 100, description: mimic weightlifting):
Signifies engaging in weightlifting or exercise (at the gym); expressing physical strength or endurance (in sports or daily life)......

(id: 128, description: mimic raising a toast): Signifies celebration (marking a shared moment); camaraderie (sharing a connection)......

Figure 4: Gesture Label Generation Visualization. It is important to note that we did not further distinguish different categories (such as A, B, C, and D) within the gesture IDs in order to avoid increasing the difficulty of model training.

4.2 Gesture Label Generation

4.2.1 Objective Evaluation. We employed the CoT prompting technique using GPT-4 to parse text for the construction of a training dataset, resulting in a training dataset comprising 3,242 entries. Each entry includes the original text with embedded gesture labels formatted as *(ID: Action Name)*, facilitating direct integration of gesture annotations within the text.

Experimental Parameters: We utilized Qwen [31] for fine-tuning. The LoRA technique was applied to the q_proj and v_proj layers, with a batch size of 6, a learning rate of 5×10^{-5}, over 100 epochs. Training was performed on a 4090 server, with a total duration of approximately 13 hours. For evaluation, we used the same test dataset and metrics as previously, involving 231 test texts and employing Partial Overlap as the primary evaluation metric. The results of the experiment are presented in Fig. 3.

Compared to the performance of GPT-4 in generating gesture labels (with an overlap rate of 65.5%), our model also demonstrated the capability to produce semantically relevant action labels. Specifically, the fine-tuned models based on Qwen1.5-1.8B-Chat and Qwen2-0.5B-Instruct achieved Partial Overlap rates of 50.2% and 42.4%, respectively, on the test set. Our gesture label generation model, despite utilizing a smaller dataset, still achieves comparable accuracy in generating gesture labels that align with the text semantics, slightly lower to GPT-4.

Moreover, the comparison of response times revealed a significant performance advantage of our model. GPT-4's average response time is about 3 seconds, processing approximately 8,000 tokens to generate a single set of action labels, with a cost of $0.26. In contrast, the response time of our models was significantly reduced to about 0.4 seconds, greatly enhancing generation efficiency. Although the Partial Overlap rate of the fine-tuned models is slightly lower, their faster response time and significantly reduced cost offer advantages in practical applications, particularly in scenarios demanding high response speed and lower resource consumption.

4.2.2 Discussion. We present specific examples of gesture label generation, as illustrated in Fig. 4. The left column shows the generated labels, each accompanied by the corresponding gestures produced by our model using a database retrieval approach. The parsing results from GPT-4 are used as a benchmark for our comparative analysis. In Fig. 4, the left side 172. The input *'Let me think, my favorite pastime is daydreaming, imagining a world of fantastic adventures.'* prompted our model to insert a *'scratch head'* gesture at *'think'*, reflecting thought and contemplation. Both our model and GPT-4 generated gestures indicating comfort in *'imagining a world...'*, demonstrating semantic consistency.

In Fig. 4, the left side 173 shows the input *'I have a big announcement, I am going to learn weightlifting,'* for which GPT-4 accurately generated the *'id: 100: mimic weightlifting'* gesture. In contrast, our model inserted a *'wave'* gesture that, although not semantically aligned, effectively conveyed a joyful emotion. This demonstrates that GPT-4 excels in deep semantic understanding, while our model can achieve similar semantic gesture label recognition at a lower cost.

In hard cases, such as in Fig. 4, the left side 174, where the input conveyed a joyful mood during a meal with friends, the model generated a *'id: 9, description: touch forehead'* gesture, which is typically associated with confusion or anxiety, resulting in a mismatch with the intended context.

Overall, our gesture label generation method demonstrates satisfactory performance, particularly with advantages in cost and efficiency. However, compared to GPT-4, the fine-tuned model shows limitations in semantic depth and detail sensitivity, occasionally resulting in less accurate gesture choices. These issues could be improved by further increasing the diversity of the training data.

4.3 Semantically Aligned Gesture Generation

4.3.1 Visualization Results. We adopt DiffuseStyleGesture[34] as the baseline, which is built upon a diffusion model trained on large-scale datasets and is capable of generating gesture sequences that are highly aligned with speech rhythm. Building on this, we propose a novel method that incorporates semantic modeling: semantic intent labels are first extracted from the input text, and corresponding semantic gesture segments are retrieved accordingly. These semantic gestures are then naturally integrated into the rhythm-aligned gesture sequence through a linear fusion strategy, enabling coordinated expression of both semantics and rhythm.

Figure 5 presents a visualization of the generated gesture sequences, where the semantically aligned gestures produced by our method (SARGes) are highlighted with red bounding boxes. In the first example, when the input text conveys the semantic concept of "frustrating," SARGes generates a foot-stomping gesture indicative of emotional expression. In the second example, in response to the semantic cue of "helpless," the system produces a head-scratching gesture, commonly associated with confusion or lack of control. In the third example, upon encountering the negation "isn't afraid," SARGes generates a side-to-side body sway gesture to convey denial. These semantically aligned gestures closely correspond to the intended meanings conveyed by the input text, demonstrating the effectiveness of our method in semantic expression. Overall, the results indicate that under the guidance of semantic intent labels, our proposed method significantly outperforms the purely learning-based baseline in accurately conveying semantic content. This demonstrates enhanced controllability and expressiveness in the semantic dimension of gesture generation. We have uploaded the visualization results on our website, see video.mp4 at https://github.com/gesture-label/ethogram.

4.3.2 User Study. We segmented the gesture video clips generated by the DiffuseStyleGesture[34] and SARGes methods, with each segment containing 2–3 complete utterances along with their corresponding gestures. A total of 20 segments were selected for subjective evaluation. Subsequently, we conducted a user study involving 15 volunteers, who were asked to rate each video on a 0–5 scale based on three evaluation criteria: gesture naturalness (i.e., whether the gestures resemble natural human motion), temporal coherence (i.e., the alignment between gestures and speech rhythm), and semantic consistency (i.e., whether the gestures appropriately convey the meaning of the spoken text). The evaluation results are summarized in Table 2.

Overall, the SARGes method outperformed the baseline across all three dimensions, with particularly strong performance in semantic consistency. These findings suggest that the proposed approach is capable of generating reliable gesture intent labels in a fast and cost-effective manner, thereby improving the semantic alignment and trustworthiness of gesture generation. Furthermore, the proposed mechanism exhibits good generalizability and can be flexibly integrated with mainstream gesture generation models to enhance their semantic expressiveness.

Table 3: User Study Results on Gesture Quality (0–5 Scale)

Evaluation Dimension	DiffuseStyleGesture	SARGes
Naturalness	*3.73*	*4.13*
TemporalCoherence	*4.23*	*4.30*
SemanticConsistency	*3.63*	*4.23*

5 CONCLUSION

In this paper, we propose a novel framework for generating gesture labels from text, laying the foundation for semantic gesture generation. We developed a gesture ethogram tailored for co-speech scenarios, systematically classifying gesture patterns. We designed an intent chain behavior intention parsing method based on LLMs by integrating reliable gesture guidelines and advanced prompting techniques, allowing us to parse text and generate corresponding gesture labels effectively. Using these labels, we constructed a dataset and fine-tuned a language model, enabling the effective generation of semantically meaningful gesture labels. Although our model's accuracy is slightly lower than that of GPT-4, it demonstrates significant advantages in response speed and cost efficiency, making it suitable for real-time applications in virtual agents and social robots. Experimental results further demonstrate that the

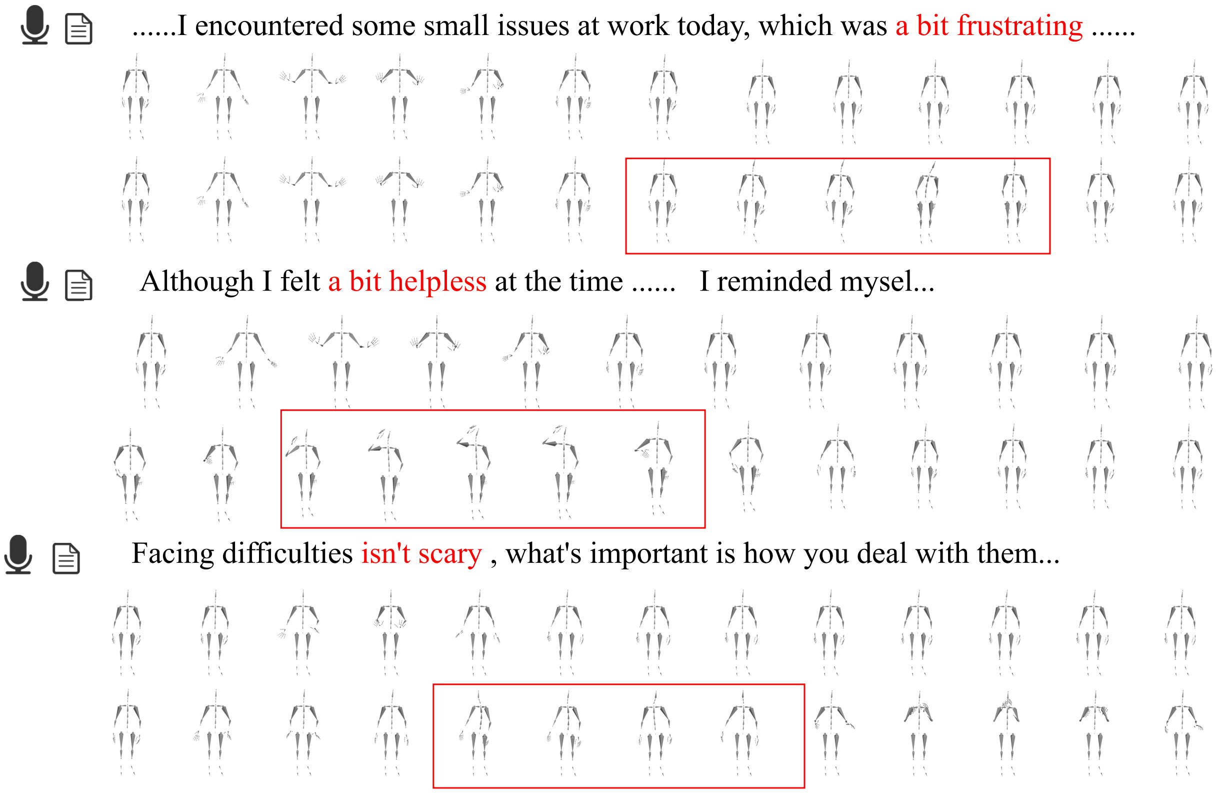

......I encountered some small issues at work today, which was a bit frustrating ......
Although I felt a bit helpless at the time ...... I reminded mysel...
Facing difficulties isn't scary , what's important is how you deal with them...

Figure 5: Semantically Aligned Gesture Generation Visualization. Both the baseline method (DiffuseStyleGesture [34]) and our proposed method SARGes, take audio and its corresponding text as input. Each example in the figure presents a textual input followed by two rows of generated gesture sequences: the first row beneath the text shows the gesture sequence produced by the baseline method, while the second row displays the gesture sequence generated by SARGes.

proposed method can generate gesture labels that are highly consistent with the intended semantics, thereby significantly improving the accuracy and expressiveness of semantic gesture generation. Moreover, this approach can be integrated with existing generative gesture synthesis methods as a complementary semantic module, enhancing their capability and controllability in the semantic expression dimension.

Our approach enables the efficient generation of reliable, low-latency, and cost-effective gesture labels, while offering substantial potential for further performance enhancement. In future work, we plan to enrich the diversity of training data and refine the labeling accuracy to further improve the robustness and generalizability of the system.

ACKNOWLEDGMENT

This work was supported by the open research fund of Pengcheng Laboratory (No.2025KF1A0030).

References

[1] D. McNeill, "Hand and mind1," *Advances in Visual Semiotics*, vol. 351, 1992.
[2] Y. Wei, D. Hu, Y. Tian, and X. Li, "Learning in audio-visual context: A review, analysis, and new perspective," *arXiv preprint arXiv:2208.09579*, 2022.
[3] S. Nyatsanga, T. Kucherenko, C. Ahuja, G. E. Henter, and M. Neff, "A comprehensive review of data-driven co-speech gesture generation," in *Computer Graphics Forum*, vol. 42, no. 2. Wiley Online Library, 2023, pp. 569–596.
[4] M. Neff, M. Kipp, I. Albrecht, and H.-P. Seidel, "Gesture modeling and animation based on a probabilistic re-creation of speaker style," *ACM Transactions On Graphics (TOG)*, vol. 27, no. 1, pp. 1–24, 2008.
[5] R. M. Holladay and S. S. Srinivasa, "Rogue: Robot gesture engine," in *2016 AAAI Spring Symposium Series*, 2016.
[6] S. Alexanderson, R. Nagy, J. Beskow, and G. E. Henter, "Listen, denoise, action! audio-driven motion synthesis with diffusion models," *ACM Transactions on Graphics (TOG)*, vol. 42, no. 4, pp. 1–20, 2023.
[7] F. Zhang, N. Ji, F. Gao, and Y. Li, "Diffmotion: Speech-driven gesture synthesis using denoising diffusion model," in *International Conference on Multimedia Modeling*. Springer, 2023, pp. 231–242.
[8] J. Chen, Y. Liu, J. Wang, A. Zeng, Y. Li, and Q. Chen, "Diffsheg: A diffusion-based approach for real-time speech-driven holistic 3d expression and gesture generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 7352–7361.
[9] T. Ao, Z. Zhang, and L. Liu, "Gesturediffuclip: Gesture diffusion model with clip latents," *ACM Transactions on Graphics (TOG)*, vol. 42, no. 4, pp. 1–18, 2023.
[10] B. Jiang, X. Chen, W. Liu, J. Yu, G. Yu, and T. Chen, "Motiongpt: Human motion as a foreign language," *Advances in Neural Information Processing Systems*, vol. 36, 2024.
[11] N. Gao, Z. Zhao, Z. Zeng, S. Zhang, D. Weng, and Y. Bao, "Gesgpt: Speech gesture synthesis with text parsing from chatgpt," *IEEE Robotics and Automation Letters*, 2024.
[12] Z. Zhang, T. Ao, Y. Zhang, Q. Gao, C. Lin, B. Chen, and L. Liu, "Semantic gesticulator: Semantics-aware co-speech gesture synthesis," *ACM Transactions on Graphics (TOG)*, vol. 43, no. 4, pp. 1–17, 2024.
[13] S. Tonmoy, S. Zaman, V. Jain, A. Rani, V. Rawte, A. Chadha, and A. Das, "A comprehensive survey of hallucination mitigation techniques in large language models," *arXiv preprint arXiv:2401.01313*, 2024.
[14] L. A. Stanton, M. S. Sullivan, and J. M. Fazio, "A standardized ethogram for the felidae: A tool for behavioral researchers," *Applied Animal Behaviour Science*, vol.

173, pp. 3–16, 2015.

[15] T. Kucherenko, P. Jonell, S. Van Waveren, G. E. Henter, S. Alexandersson, I. Leite, and H. Kjellström, "Gesticulator: A framework for semantically-aware speech-driven gesture generation," in *Proceedings of the 2020 international conference on multimodal interaction*, 2020, pp. 242–250.

[16] Y. Yoon, B. Cha, J.-H. Lee, M. Jang, J. Lee, J. Kim, and G. Lee, "Speech gesture generation from the trimodal context of text, audio, and speaker identity," *ACM Transactions on Graphics (TOG)*, vol. 39, no. 6, pp. 1–16, 2020.

[17] Y. Liang, Q. Feng, L. Zhu, L. Hu, P. Pan, and Y. Yang, "Seeg: Semantic energized co-speech gesture generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 473–10 482.

[18] H. Teshima, N. Wake, D. Thomas, Y. Nakashima, H. Kawasaki, and K. Ikeuchi, "Deep gesture generation for social robots using type-specific libraries," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 8286–8291.

[19] U. Bhattacharya, N. Rewkowski, A. Banerjee, P. Guhan, A. Bera, and D. Manocha, "Text2gestures: A transformer-based network for generating emotive body gestures for virtual agents," in *2021 IEEE virtual reality and 3D user interfaces (VR)*. IEEE, 2021, pp. 1–10.

[20] Y. Zhou, J. Yang, D. Li, J. Saito, D. Aneja, and E. Kalogerakis, "Audio-driven neural gesture reenactment with video motion graphs," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 3418–3428.

[21] Z. Zhao, N. Gao, Z. Zeng, G. Zhang, J. Liu, and S. Zhang, "Gesture motion graphs for few-shot speech-driven gesture reenactment," in *Proceedings of the 25th International Conference on Multimodal Interaction*, 2023, pp. 772–778.

[22] S. Zhang, J. Yuan, M. Liao, and L. Zhang, "Text2video: Text-driven talking-head video synthesis with personalized phoneme-pose dictionary," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 2659–2663.

[23] Z. Qing, Z. Cai, Z. Yang, and L. Yang, "Story-to-motion: Synthesizing infinite and controllable character animation from long text," in *SIGGRAPH Asia 2023 Technical Communications*, 2023, pp. 1–4.

[24] Z. Cai, J. Jiang, Z. Qing, X. Guo, M. Zhang, Z. Lin, H. Mei, C. Wei, R. Wang, W. Yin *et al.*, "Digital life project: Autonomous 3d characters with social intelligence," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 582–592.

[25] W. Huang, P. Abbeel, D. Pathak, and I. Mordatch, "Language models as zero-shot planners: Extracting actionable knowledge for embodied agents," in *International conference on machine learning*. PMLR, 2022, pp. 9118–9147.

[26] J. S. Park, J. O'Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein, "Generative agents: Interactive simulacra of human behavior," in *Proceedings of the 36th annual acm symposium on user interface software and technology*, 2023, pp. 1–22.

[27] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in neural information processing systems*, vol. 35, pp. 24 824–24 837, 2022.

[28] N. Shinn, F. Cassano, B. Labash, A. Gopinath, K. Narasimhan, and S. Yao, "Reflexion: Language agents with verbal reinforcement learning.(2023)," *arXiv preprint cs.AI/2303.11366*, 2023.

[29] P. N. Lehner, "Design and execution of animal behavior research: an overview," *Journal of animal science*, vol. 65, no. 5, pp. 1213–1219, 1987.

[30] H. Xu, A. Sharaf, Y. Chen, W. Tan, L. Shen, B. Van Durme, K. Murray, and Y. J. Kim, "Contrastive preference optimization: Pushing the boundaries of llm performance in machine translation," *arXiv preprint arXiv:2401.08417*, 2024.

[31] J. Bai, S. Bai, Y. Chu, Z. Cui, K. Dang, X. Deng, Y. Fan, W. Ge, Y. Han, F. Huang *et al.*, "Qwen technical report," *arXiv preprint arXiv:2309.16609*, 2023.

[32] D. Nam, A. Macvean, V. Hellendoorn, B. Vasilescu, and B. Myers, "Using an llm to help with code understanding," in *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering*, 2024, pp. 1–13.

[33] D. Nam, A. Macvean, V. Hellendoorn, B. Vasilescu, and B. Myers, "Using an llm to help with code understanding," in *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering*, 2024, pp. 1–13.

[34] S. Yang, Z. Wu, M. Li, Z. Zhang, L. Hao, W. Bao, M. Cheng, and L. Xiao, "DiffuseStyleGesture: Stylized audio-driven co-speech gesture generation with diffusion models," in *Proceedings of the 32nd International Joint Conference on Artificial Intelligence* , pp.1-11, 2023.