

GestureGPT: Toward Zero-shot Free-form Hand Gesture Understanding with Large Language Model Agents

Xin Zeng*

Institute of Computing
Technology, Chinese Academy of
Sciences.
Beijing, China

Xiaoyu Wang*†

Institute of Computing
Technology, Chinese Academy of
Sciences.
Beijing, China

Tengxiang Zhang‡

Institute of Computing
Technology, Chinese Academy of
Sciences.
Beijing, China
ztxseuthu@gmail.com

Chun Yu

Computer Science and Technology,
Tsinghua University.
Beijing, China

Shengdong Zhao

Synteraction Lab, City University
of Hong Kong.
Hong Kong, China

Yiqiang Chen

Institute of Computing
Technology, Chinese Academy of
Sciences.
Beijing, China

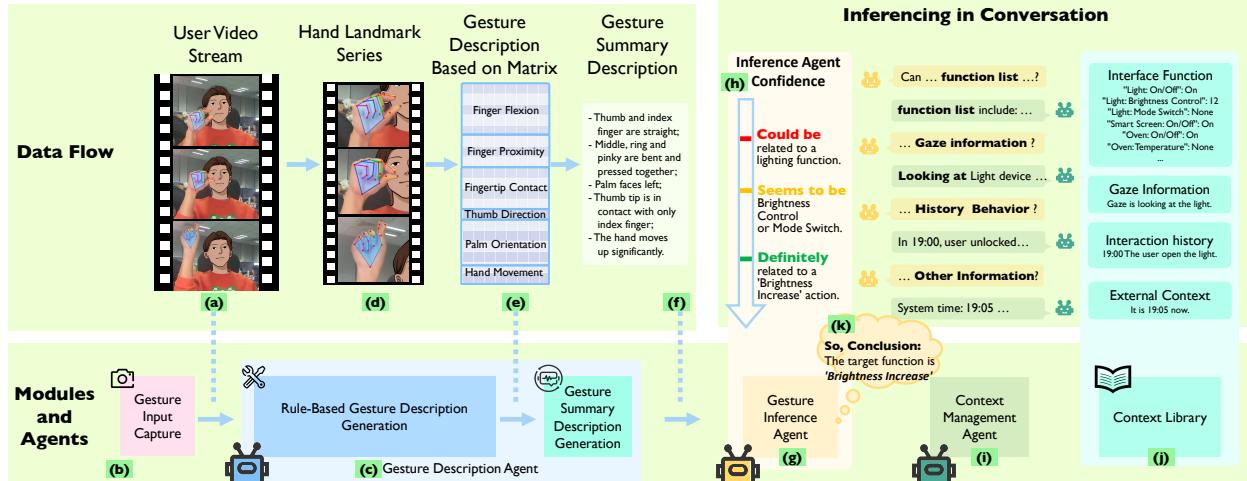


Figure 1: A user aims to adjust the light's brightness, assuming GestureGPT is successfully implemented, so he (a) triggers the system with a raised right hand above the chest, followed by looking at the light and performing a 'pinch and move up' gesture. The (b) input camera captures this action in real-time video streaming. (c) The Gesture Inference Agent then processes the gesture by (d) filtering key frames from the video and extracting hand landmarks for (e) generating the gesture state matrix. Subsequently, (f) the agent generates a description of the gesture's pose and movement. Upon activated by the description, the (g) Gesture Inference Agent analyzes this description to understand which function the gesture maps to and (h) assesses the confidence level of the result. If the confidence is deemed insufficient, the agent requests context information from the (i) Context Management Agent, which retrieves relevant data from the (j) context library containing all available context in the current environment, to inform the inquiry. According to the inquiry, as context information such as "Gaze: user is looking at the Light" and "History: the user opened the light at 19:00" is incorporated, the Gesture Inference Agent (k) concludes that the function designated by the gesture is 'Brightness Increase' of the light. The function could then be triggered to finish the interaction process.

*Both authors contributed equally to this research.

†Also with The Hong Kong University of Science and Technology.

‡Corresponding author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this

work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY'

© 2024 Association for Computing Machinery.

ACM ISBN 978-1-4503-XXXX-X/18/06...\$15.00

<https://doi.org/XXXXXX.XXXXXXXX>

Abstract

Current gesture interfaces typically demand users to learn and perform gestures from a predefined set, which leads to a less natural experience. Interfaces supporting user-defined gestures eliminate the learning process, but users still need to demonstrate and associate the gesture to a specific system function themselves. We introduce GestureGPT, a free-form hand gesture understanding framework that does not require users to learn, demonstrate, or associate gestures. Our framework leverages the large language model's (LLM) astute common sense and strong inference ability to understand a spontaneously performed gesture from its natural language descriptions, and automatically maps it to a function provided by the interface. More specifically, our triple-agent framework involves a Gesture Description Agent that automatically segments and formulates natural language descriptions of hand poses and movements based on hand landmark coordinates. The description is deciphered by a Gesture Inference Agent through self-reasoning and querying about the interaction context (e.g., interaction history, gaze data), which a Context Management Agent organizes and provides. Following iterative exchanges, the Gesture Inference Agent discerns user intent, grounding it to an interactive function. We validated our conceptual framework under two real-world scenarios: smart home controlling and online video streaming. The average zero-shot Top-5 grounding accuracies are 83.59% for smart home tasks and 73.44% for video streaming. We also provided an extensive discussion of our framework including model selection rationale, generated description quality, generalizability etc.

CCS Concepts

- Human-centered computing → Gestural input; User interface management systems.

Keywords

Free-Form Gesture, Zero-Shot, Gesture Recognition, Interaction Context

ACM Reference Format:

Xin Zeng, Xiaoyu Wang, Tengxiang Zhang, Chun Yu, Shengdong Zhao, and Yiqiang Chen. 2024. GestureGPT: Toward Zero-shot Free-form Hand Gesture Understanding with Large Language Model Agents. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 24 pages. <https://doi.org/XXXXXX.XXXXXXX>

1 Introduction

Gestures express human intent in an intuitive and immediate manner, enables natural human computer interaction with low cognitive load [46, 66]. Most existing gesture interfaces only support pre-defined gestures. Within the pre-defined gesture set, such interfaces can achieve a very high gesture classification accuracy [3, 59]. However, it takes great efforts to learn and memorize different gestures and their mappings to different system functions [11, 15], which is especially true

with a large gesture set. The mapping between the gesture and the function is also fixed, which cannot be easily expanded and may contradict the user's preference [8, 41, 58].

To address such limitations of pre-defined gestures, researchers propose gestural interfaces that support user-defined gestures [34, 61, 67]. Users can define their own gestures for each function, and they only need to demonstrate the gesture several times, which eliminates the gesture learning effort and offers greater system flexibility. However, each user needs to define and demonstrate her own gesture, as well as associate it with a system function [38, 53]. Users also still need to memorize their own gestures for different functions, which may lead to frustration using the interface. Such constraints deteriorate the natural interaction experience and hinder the wide adoption of gestural interfaces, which leads to Norman's famous argument that "Natural User Interfaces are Not Natural" [42]. While significant advances have been made in gesture recognition technologies over the years, this naturalness challenge of gestural interfaces persists.

A natural gesture interface requires end-to-end **gesture understanding** rather than merely **gesture recognition**. Users do not need to learn, memorize, or demonstrate gestures. They can just perform gestures as they see fit based on their common sense, which involves the semantics of the function and their daily interaction experience with humans and machines. The association between the gesture and the function is also automatically executed by the interface by considering both the gesture and the interaction context. However, gesture understanding is inherently more challenging than gesture recognition.

Towards a natural gesture understanding interface, we propose GestureGPT, an LLM-based hand gesture interface framework that understands natural free-form hand gestures and automatically associates the gesture with the intended function. The core idea of GestureGPT is to leverage LLM's rich common sense to recognize gestures and understand the interaction context, as well as its strong inference capability to map the gesture to the intended function. However, it still involves plenty of efforts to build such an LLM-based gesture understanding framework. For example, gestures and context information need to be transformed and formatted for LLMs to understand. There also need to be a mechanism for robust and thorough synthesizing of all relevant information without forgetting any of it.

To that end, we design a triple-agent framework to generate gesture description, manage the context, and infer the intended function, which involves a *Gesture Description Agent*, a *Gesture Inference Agent* and a *Context Management Agent*, all based on LLMs. 1) *Gesture Description Agent* describes hand gestures in natural languages to facilitate the powerful text-based LLMs reading and understanding of gestures. A set of specially designed and tuned rules transform hand poses and movements into discreet states (e.g., the bending state of each finger) based on visually-extracted hand landmarks, triggered by raising hand to the chest level. The agent then synthesizes a general description of the gesture based on a matrix

formed with the states. 2) *Gesture Inference Agent* analyzes the description and initiates conversation with *Context Management Agent* to inquire about the context information. After multi-round dialogues, it infers which interactive function the user intents to activate. 3) *Context Management Agent* answers questions of *Gesture Inference Agent* by referring to a context library, which includes different types of context information like gaze point, interaction history, *etc.* The context library is organized in the form of JSON files.

GestureGPT offers a number of advantages that address the challenges and limitations of previous gestural interfaces. 1) GestureGPT allow users to perform natural free-form hand gestures, which do not need to resemble those in a pre-defined gesture set or those previously defined by users. This eliminates any learning, memorizing, or demonstration effort. 2) GestureGPT automatically associates gestures with corresponding interface functions by step-by-step inference based on both the gesture description and the interaction context. Such a novel LLM-based structure successfully addresses gesture's ephemeral nature and close relationship with context [42]. Additionally, the integration of context further enhances the accuracy of gesture mapping. 3) GestureGPT analyzes spatial coordinates of hand landmarks to generate gesture descriptions, making it independent of view angles and even modalities. For example, hand landmarks can also be reconstructed by wearable sensors [27, 75]. Such flexibility makes it easy to adapt GestureGPT for a wide range of applications. The natural language descriptions also preserve user privacy and reduce data transmission load, which are important for interaction interfaces.

Describing nuanced states and movements of fingers are key to accurately understand hand gestures. Even though it is a difficult task for current large vision models, our hand landmarks based method captures finger states accurately and thoroughly by first calculating a set of discreet finger states based on rules, then use an LLM to synthesize a gesture summary. The parameters for the rules are tuned on third-view public gesture image datasets, and test on both third-view dataset (HaGRID [28], 38576 test samples) and first-view dataset (EgoGesture [74], around 5000 test samples). The overall error rate is 2.3% on the third-view dataset and 6.3% on the first-view dataset. The synthesized summary is rated by two gesture experts and achieved a 3.51(std = 1.14) on a 5-point Likert scale (1 being almost none and 5 being almost all of the description is correct and relevant to the gesture).

To evaluate our entire triple-agent framework, we ran experiments under two realistic interaction scenarios to test boundaries of our framework: smart home IoT device control with AR headset (first-person view with 18 functions) and online video streaming on a desktop PC (third-person view with 66 functions). The highest zero-shot Top-1/Top-5 gesture grounding accuracy is 44.79%/83.59% for smart home control and 37.50%/73.44% for video streaming. We report accuracy at different levels to better understand and demonstrate the framework's performance and boundaries. The results shows

great potential for our framework, which is **the first zero-shot free-form gesture understanding framework** that does not require any learning, memorization, demonstration, and association efforts as to our best knowledge.

GestureGPT does not yet offer a practical interface that can be used today—primarily due to the slow inference speeds of LLM systems, which in our evaluation average 227 seconds per task. As a new gesture interface paradigm, however, GestureGPT framework serves as a foundation for further innovations. For example, future advances on edge-side large language models or large multi-modal models fine-tuned for gesture understanding tasks can greatly reduce system response delay, approaching a practical interface in the end.

Our contribution is three-fold:

- (1) We proposed and evaluated the first framework for automatic free-form hand gesture understanding, as to our best knowledge. Our framework leverages LLMs to mimic human gesture understanding process.
- (2) We designed a set of gesture description rules based on hand landmarks to thoroughly and accurately capture states and movements of fingers, which has similar performance with SOTA large multi-modal modal GPT-4o.
- (3) We carefully crafted prompts for each agent and evaluated the effects of different contexts on our framework's performance. Such insights are invaluable for future context-aware agent-based gesture understanding work.

2 Background and Related Work

2.1 LLM as Autonomous Agent

Large language models have displayed an exceptional ability to understand and execute a broad spectrum of tasks [2, 43]. LLM has the potential to emulate human-level intelligence, accurately perceive generalized environments, act accordingly, and iterate to enhance outcome [51] when faced with diverse situations [60]. Agents based on LLMs have shown potential in various domains, from web browsing [9, 68, 77], strategic planning [69] to robotic control [4, 10]. This paper, on the other hand, addresses a notable gap in existing literature and utilizes LLMs for free-form gesture understanding and interaction.

Though LLM agents have shown promising intelligence, a single agent implementation may suffer from performance degradation in long context scenarios [26]. Instead, multi-agent systems have shown superiority to accomplish more complex tasks in collaboration, reducing hallucination [62], and information exchange [55]. For instance, Park et al. [45] designed a multi-agent system that simulates human behavior in a virtual environment. Park et al. [44] relies on conversational interactions among multiple agents to aid online decision-making, which shows multi-agent's great potential in reasoning under unfamiliar scenarios. Other applications software development [47], span reasoning [32], evaluation [5, 73], and a myriad of intricate tasks [44, 79]. So, GestureGPT chooses a triple-agent architecture to better handle the complicated

context-aware gesture understanding task. The goals of the three agents are clearly defined and isolated, so that they can be optimized individually while collaborate seamlessly to achieve accurate gesture comprehension.

2.2 Natural Free-form Gesture Understanding

Gestural interfaces working with pre-defined gesture set demand large annotated dataset, and the confinement to pre-defined gestures also hampers the naturalness of interaction [64]. The conflicts in different system designs further exacerbate user adaptation challenges across platforms. User-defined gestural interfaces mitigates the learning burden by allowing users to define their own gestures. Only several demonstrations of the gesture are necessary for the system to learn new gestures with the help of advanced few-shot learning algorithms [67]. Gesture Coder [34] allows the user to demonstrate a gesture, and, instead of defining a gesture name for it, the user directly associate it to a designated function with auto-generated recognizer. But in these works, users still need to demonstrate and memorize the gesture while ensuring its distinctiveness from existing gesture commands. Recent advances in zero-shot learning make it possible to recognize gesture classes unseen during training, but it requires manually introducing prior knowledge [36, 37] or representative samples [65] of unseen class.

Nevertheless, all aforementioned approaches are still restricted by a finite set of distinguishable gestures defined either by the interface designers or users. This results in a mismatch between the fixed gesture set and the flexible gestures human performs understand different scenarios, which greatly restricts the inherent expressiveness of gesture. Thus, both pre-defined and user-defined gestural interfaces actually require users to adapt to it, rather than the reverse [64]. Free-from gestural interfaces, on the other hand, do not have such restrictions. For example, Gesture Avatar [33] allows any form of drawing gestures on a screen. The input gesture is associated with a UI element based on their appearance resemblance, providing intuitive interaction experience. However, there are still gaps in research of free-form hand gesture understanding. Compared with screen gestures, hand gestures are more complicated in form, do not necessarily resemble UI elements in appearance, and can change under different scenarios even for the same function. To tackle such challenges, our framework relies on LLMs to understand the semantic meanings of both the hand gestures and the system functions to ensure correct gesture-function mapping. An overview of current gesture-based interaction systems and the placement of our work is shown in Figure 2.

	Related Work		User Efforts			
	References	Gesture Type	Learning when using the system for the first time	Memorization when using the system later on	Demonstration for recording the gesture example	Association for mapping the input to a function
Pre-defined Gestures	Montero and Sucar, 2006 Taranta & et al., 2015 Madapana and Wachs, 2018 ... Wu et al., 2021	Hand Gesture / Screen Gesture / Body Gesture / ...	+	+	-	-
	Gesture Coder (Li and Li, 2012) Xu et al., 2022	Hand Gesture / Screen Gesture	-	+	+	+
	Gesture Avatar (Li and Li, 2011)	Screen Gesture	-	-	-	-
Free-form Gestures	GestureGPT	Hand Gesture	-	-	-	+
	Weixelblat, 1995 ^a	Hand Gesture	-	-	-	+

Figure 2: Overview of current gestural interaction systems.

2.3 Gesture Understanding with Context Information

As Norman pointed out, gestures are ephemeral in nature and highly context related [42]. They inherently possess diverse semantics across different contexts, and may embody social metaphors [54]. Contextual information such as spatial distance [39], gaze [6], speech [64], user history [7, 40] and domain-specific knowledge [56] have been integrated into gestural interaction systems to improve gesture recognition accuracy. However, most previous work only leverage gaze information for simplified tasks [1, 16, 23, 50, 52]. The handling of different types of context typically require highly specialized models [7, 24, 52], which limits the generalizability of such systems.

LLM agents' promising ability to solve complex problems provides an alternative solution for inference based on different types of context information. LLM agents can retrieve higher-level context like semantic meaning of interaction elements [21] and users' profile and preferences [48]. The way it utilizes context and the intentions it can deal with are also greatly enlarged. For example, [20, 29, 30, 48] have shown that LLM agents can turn loosely-constrained commands into appropriate actions. Inspired by existing research, we use a *Context Management Agent* to manage context information stored in the context library, and a *Gesture Inference Agent* to infer the intention behind a gesture based on context cues.

3 Method

GestureGPT system has a triple-agent framework to efficiently manage gesture description, context, and inference tasks, thereby enhancing system flexibility and scalability while addressing the shortcomings of single or dual agent systems, consists of: (1) ***Gesture Description Agent***, generating descriptions of gestures from videos; (2) ***Context Management Agent***, handling interaction context; and (3) ***Gesture Inference Agent***, synthesize information to infer gesture intentions through dialogue and reasoning.

The workflow initiates with the transformation of user gestures, captured by an RGB camera, into natural language description. The *Gesture Description Agent* first uses a set of rules to transform the gesture into a "Gesture State Matrix", which delineates hand and finger states over time. The agent

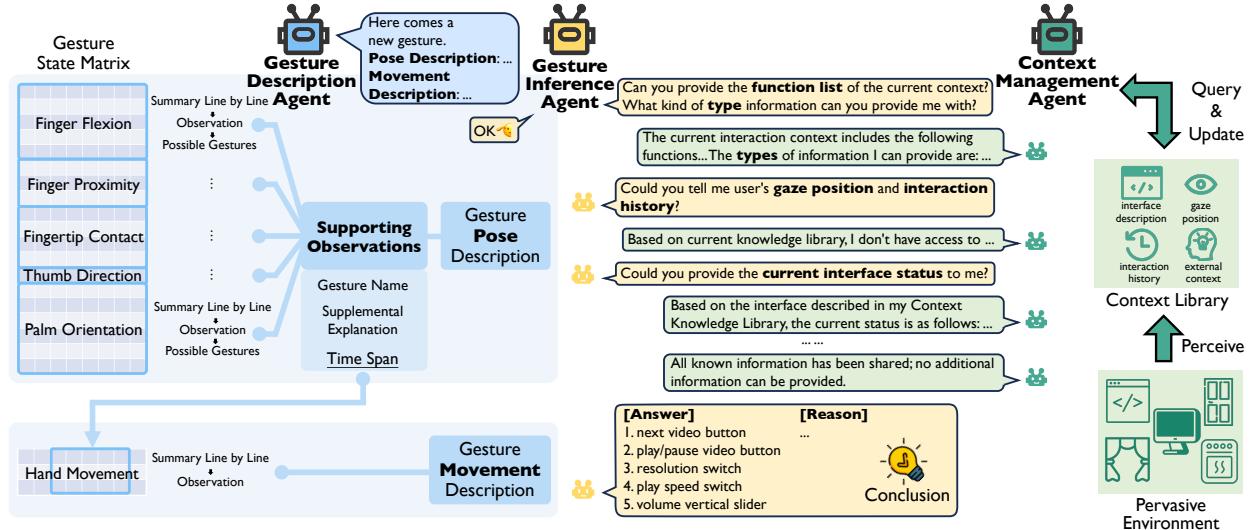


Figure 3: Agents Collaboration Workflow.

then relies on this matrix to produce a summary of gestures, which is forwarded to the *Gesture Inference Agent*. Upon receiving the gesture description, the *Gesture Inference Agent* engages in multi-round dialogue with the *Context Management Agent*. This involves identifying and soliciting relevant context information from a context library managed by the *Context Management Agent*. Through iterative dialogue, the *Gesture Inference Agent* gains a comprehensive understanding of the interactive scenario and completes a dynamic mapping between the gesture and possible functions.

3.1 Gesture Description Agent

The Gesture Description Agent is crucial for translating video-captured gestures into natural language descriptions that are understandable by LLMs. We choose LLM since it has richer common knowledge and stronger inference than existing LMM that we can access, which are vital to deal with context-aware free-form gesture understanding tasks.

3.1.1 Rule-Based Gesture Description Generation

We design a set of 6 rules to encode the hand gesture in terms of finger flexion, proximity, contact, direction, as well as palm orientation and hand position (Table 1). Each state is calculated based on the coordinates of hand landmarks generated by MediaPipe [71] with videos captured by a RGB camera. For each rule, parameters (such as the distance between fingers to determine their proximity) are trained on third-view public gesture dataset, and tested on first-view and third-view datasets. Test results show an error rate of 2.3% for third-view and 6.3% for first-view samples, indicating the rules' efficacy in retaining accurate information from gestures. The detailed process of rule definition and training, as well as the evaluation metrics for these rules, are thoroughly described in Appendix A.

Table 1: Summary of rules, their meanings, and corresponding values.

Rule Name	Applicable to	Value
Finger Flexion	Thumb, Index, Middle, Ring, Pinky	1: Straight; 0: Between; -1: Bent
Finger Proximity	Index-Middle, Middle-Ring, Ring-Pinky	1: Pressed Together; 0: Between; -1: Apart
Thumb Fingertip Contact	Thumb-Index, Thumb-Middle, Thumb-Ring, Thumb-Pinky	1: Contact; 0: Between; -1: No Contact
Pointing Direction	Thumb	1: Upward; -1: Downward; 0: Other Directions/Bent
Palm Orientation	Palm	One-hot encoding: [Left, Right, Down, Up, Inward, Outward]; All zeros: Unknown
Hand Position	Hand	Float Coordinates

The gesture period of interest starts when users raises their hands to or above their chest level. Frames within a gesture period are processed at 0.2-second intervals, with each sampled frame undergoing rule-based calculations. However, the natural language descriptions concatenated from each rule are too long to fit into a prompt of an LLM. So we introduce

a gesture state matrix that contains vectors of each frame, which capitalizes on LLMs' proficiency with code-formatted content [14].

3.1.2 Rules Calculation Method

Flexion of a finger is calculated as the total bending angle of each joint. For thumb it is the bending angle of the ip joint, and for other fingers, it is the bending angle of the pip and dip joint. Then, two parameters *straight_threshold* and *bent_threshold* are set to determine if the finger is straight, bent, or ‘unsure’ if the result falls between them. Since thumb has a different joint structure compared with other fingers, a new pair of thresholds are specially set for thumb.

Proximity of two fingers A and B is calculated as the average minimal distance from each finger’s joint to the other finger. Two thresholds i.e. *together_threshold* and *separated_threshold* are set to determine if the two fingers are pressed together, separated, or ‘unsure’ if the result falls between them.

Contact of thumb and another finger is computed as the distance between their fingertips. Then, two thresholds, i.e. *contact_threshold* and *not_contact_threshold* are set to determine if the two fingers’ fingertips are in contact or not, or ‘unsure’ if the result falls between them.

Point direction of thumb is computed as the direction from thumb’s mcp joint to tip joint. Then, it is compared with two reference vectors representing upward and downward. If a reference vector has the minimal angle with the palm orientation vector and the angle is below *angle_threshold*, the reference vector would be the thumb’s pointing direction. Note that it is only applicable when thumb is straight. If thumb is bent, the result is set to ‘unsure’. This is especially useful when discriminating between gestures like ‘thumb up’ and ‘thumb down’.

Palm orientation is computed as the direction to which the palm is facing. It is computed by the cross product of two vectors within the plane of the palm. (Fig 4). Then, the direction is compared with six reference vectors representing upward, downward, left, right, inward and outward. If a reference vector has the minimal angle with the palm orientation vector and the angle is below *angle_threshold*, the reference vector would be the palm orientation. Otherwise the orientation of palm is set as ‘unsure’.

Hand position is computed as the geometrical center of a hand by taking average of all 21 landmarks’ coordinates. No parameter or threshold is applied here.

The pseudocode for the rules are given in Appendix A.2., The rule-based module evaluation and the matrix can be found in Appendix A.3.

3.1.3 Gesture Summary Description Generation

The system needs to extract the rich information embedded in the matrix and generate a concise summary of gesture descriptions to pass to the *Gesture Inference Agent*. A significant challenge encountered by the agent is the novelty of the state matrix data type; the pre-trained data of the LLM likely does not contain exactly similar data. We employ a

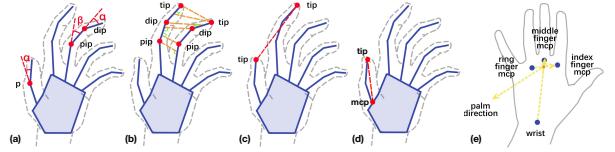


Figure 4: Illustration of gesture description rules. (a) The flexion of a finger is calculated from the sum of bending angle of ip joint (for thumb) or pip and dip joint (for other fingers). (b) The proximity of two fingers is calculated from the average distance from each finger’s pip/dip/tip joint to the other finger. (c) The contact of thumb and another finger is calculated from the distance of their fingertips. (d) The pointing direction of thumb is calculated from the vector from thumb’s mcp to tip. (e) The palm orientation is calculated from the dot product of the two vectors on the hand, pointing towards the reader.

chain-of-thought [63] process to guide the LLM step-by-step, supplemented by high-quality expert examples that demonstrate how the analysis should be conducted. We choose to use two prompts to generate the hand **pose** (finger flexion, finger proximity, fingertip contact, thumb direction and palm orientation) and hand **movement** description separately. As is shown in Figure 3, the *Gesture Description Agent* first converts pose-related rows of the matrix to descriptions, along with the time span during which this gesture happened. The time span is then used to find the corresponding part of movement-related rows, and then another prompt is used to convert it to movement description. As long as the agent correctly finds the pose movement, irrelevant movements are naturally filtered out.

For pose description, we developed a prompt structured into three parts shown in Figure 5: introduction, procedure, examples. The introduction outlines the task, describes the input matrix, and explains interactive gestures to distinguish between types of gestures and understand their life cycle. The procedure section teaches the agent to decompose the matrix, encouraging a human-like process of understanding and summarizing data, while suggesting transcription to counter LLM’s forgetfulness and improve accuracy. The last part provides examples of static and dynamic gestures and a structured prompt to enhance LLM performance by leveraging its knowledge base for complex task execution. The well-structured prompt ensures the LLM model correctly processes the matrix data type. The hand movement description prompt shares a similar structure.

```

Gesture Description Agent - Pose (Prompt Overview)
# Introduction:
## Task Introduction:
Your task is to analyze hand poses ... identifying the gesture ...
## Input Data Introduction:
A 2D array ... represents hand and finger states. Each time step
is 0.2 seconds ...
## Gesture Introduction:
Users can use static or dynamic gestures to interact ...

# Procedure:
## step 1:
Decompose the input array ...
## step 2:
Synthesize observations and guesses gesture ...
## Behavior Guide:
things should and should not do ... Common Mistakes ...

# Examples
## Example 1
(A specific example of Static Gesture 'peace')
## Example 2
(A specific example of Dynamic Gesture 'Zoom In with Two Fingers')

Gesture Inference Agent (Prompt Overview)
# Introduction:
## Task Introduction:
you will understand an interactive gesture from its description
map it with a specific function ...
## Gesture Description Introduction:
The init input contains content related to hand posed and hand
movements...
## Context Management Agent Introduction:
The **Context Agent** manages all type of interaction context
information...

# Behavior Guidelines:
## Requirements:
Things need to do like 'asking for the function list' and
'requesting one context at a time'...
## Prohibitions:
Things can not do like 'ask **Context Agent** to align the
gesture to a function' ...
## Output block:
The output includes 'Thought', 'Gesture Guess', and 'Question for
Context Agent' when the information is insufficient.
It contains 'Thought' and 'Results' once a decision has been made.
All outputs are requested in JSON format.

```

Figure 5: Formatted Prompt For Gesture Description Agent.

3.2 Gesture Inference Agent

Upon receiving gesture descriptions from *Gesture Description Agent*, *Gesture Inference Agent* engages in dialogue with *Context Management Agent*. This conversational exchange is pivotal for the Gesture Inference Agent to discern the user's actual intent in the context, *i.e.*, associating the gesture with a target system function from a list of potential functions supplied by *Context Management Agent*. This setup epitomizes a collaborative conversation aimed at overcoming the challenges of gesture ambiguities under different interactive scenarios and context.

We outline the agent's tasks and some behavioral guidelines to better demonstrate the semantic nature of the agent task revolves around thinking, summarizing, and inferring. The overall prompt is structured into two parts shown in Figure 6.

The introduction part includes a Task Introduction, a Gesture Description Introduction, and a Context Management Agent Introduction (its conversational counterpart). A critical aspect to convey in this part is that the initial gesture description may contain errors. Therefore, the Gesture Inference Agent must utilize context information to help with the inference, correct potential mistakes, and make decisions.

Behavioral guidelines for the agent consist of seven directives and five prohibitions, acquired by iteratively updating the prompts during implementation. When new error tendencies are observed, corrections are introduced into the prompt, akin to a teacher's guidance. Then, we define the output format. We opt for JSON due to its ease of parsing, allowing for efficient extraction of useful information for subsequent conversational rounds. Importantly, we require the agent to first articulate its **thoughts** before posing **questions** or drawing **conclusions**. This approach has been proven to make the agent's behavior more reasonable. Once the agent deems that it is confident to decide or that no further context can be gleaned, it proceeds

Figure 6: Formatted Prompt For Gesture Inference Agent.

to make a final decision, listing the top-5 possible functions for the current interface, ranked from most to least likely. The Gesture Inference Agent emphasizes the importance of precise communication, contextual reasoning, and analytical capabilities in interpreting and responding to gesture-based interactions.

3.3 Context Management Agent

The Context Management Agent plays a crucial role in our system, offering intelligent management of various types of context information. Unlike static, rule-based context management system, this agent dynamically adapts to unfamiliar context thanks to LLM's vast and intricate knowledge base, ensuring a flexible and responsive interaction environment.

Context management is fundamental for several reasons. First, it facilitates fast interpretation of and dynamic adaptation to diverse types of context information, empowering the system to make informed decisions beyond the limitations of predefined rules. This adaptability significantly enhances system responsiveness and allows for a more precise tailoring of responses to user gestures. Additionally, it streamlines the handling of complex contextual interactions, making the system more accessible to both users, developers and agents.

The overall prompt is structured into two parts shown in Figure 7.

The introduction part includes a Task Introduction and a Context Library Introduction. The Context Library Introduction introduces the currently accessible contexts and is designed to be adjustable. Similar to the gesture inference agent, iterative implementation guides the Context Management Agent's requirements and prohibitions. There are 4 requirements and 2 prohibitions. The output is structured in JSON format for ease of parsing, allowing for efficient extraction of useful information and facilitation of subsequent conversation rounds.

```

Context Management Agent (Prompt Overview)

# Introduction:
## Task Introduction:
Your task is to manage and understand all interaction context information in the current interaction environment, and answer **Gesture Inference Agent**'s question...
## Context Library Introduction:
Includes the currently accessible context introduction. ...

# Behavior Guidelines:
## Requirements:
Things need to do ...
## Prohibitions:
Things can not do ...
## Output block:
The output includes 'Thought', 'Answer'

```

Figure 7: Formatted Prompt For Context Management Agent.

3.3.1 Context Library Operations

The Context Management Agent supports three main operations within the context library: **Adding** new context types, **retrieving** context information based on queries and **calculating** specific context values.

- **Adding Context Types:** Context types and their values are organized using markdown for text and JSON for structured data. New context types are introduced with natural language descriptions, and associated values are formatted in JSON. This facilitates easy system expansion and automatic incorporation of new contexts into the operational prompt.
- **Retrieving Context Information:** The agent automates retrieval operations, streamlining the process by organizing data in a standardized format (e.g., JSON), which the LLM can autonomously interpret.
- **Calculating Context Value:** The agent can calculate specific context values based on predefined methods in the context's description. These methods can be implemented in a Python script. Upon activation of a calculation operation, the agent generates a unique placeholder. This placeholder triggers the system to execute the Python script, automating the retrieval of context values and producing the desired output. The placeholder is then replaced with this output in the final response.

The design of *Context Management Agent* offers significant benefits, such as the ease of adding new context types using simple natural language descriptions and the automated retrieval and understanding of context information. This enhances the system's versatility and usability. Our design also informs the design of context management components of future interactive systems.

4 Evaluation

We designed two experiments to assess GestureGPT's adaptability and effectiveness under different interaction scenarios with varying context environments and camera perspectives.

In the first interaction scenario, a user controls smart home appliances through an AR headset. In this setup, gestures are

captured from the user's own viewpoint, offering a first-person perspective. It is a key interaction scenario in the future when users interact with environment using head-mounted devices.

Our second interaction scenario mimics the case when a user is watching online videos. The user watches the video on a monitor with a camera capturing gestures from the third-person perspective. This setup is prevalent in a variety of settings, including smart TVs, interactive public displays, educational environments, gaming, etc. The complexity increases in this scenario, as there are plenty of functions and interactive elements on a webpage.

We selected four contexts in our experiments to evaluate GestureGPT performance in different context setting:

- **Interface Function List:** Crucial for mapping gestures to interface functions, this context includes interface name and a list of functions with their names, locations, and unique IDs, key for navigating the user's interaction environment.
- **Gaze Information:** Data on the user's gaze, given in 3D (in home scenario) or 2D (in video scenario) coordinates.
- **Interaction History:** Insights into the user's recent interactions.
- **External Context Information:** Information from other devices or sensors. We introduce this type of context to explore how other factors impact gesture understanding and whether the agent can leverage this information.

We informed participants that both video and eye movement data would be collected during the study, and we assured them of the confidentiality and safety of their data. The study is IRB approved by our local institution. All participants provided informed consent. In both experiments, we asked participants to perform the gesture as they see fit to finish the task. Both static hand poses and dynamic gestures involving hand movements are permitted.

We employed the OpenAI GPT-4 model as the underlying architecture for our triple-agent system. Specifically, we utilized the gpt-4-1106-preview version for our evaluations. We configured the request temperature to 0 so that the model's output is as consistent as possible. Apart from this, we adhered to OpenAI's default settings for other parameters. To mitigate the effects of randomness and enhance the reliability of our findings, we run the experiment repeatedly for three times. The aggregated results from these iterations were used to substantiate our conclusions.

4.1 Experiment 1: Augmented Reality-Based Smart Home IoT Control

This study explores the interaction between users and smart home devices through augmented reality (AR), specifically using gesture controls in a simulated kitchen environment. Participants perform gestures as they see fit to control various IoT devices, triggering changes in device state accordingly.

4.1.1 Experiment Setting and Procedure

- **Experimental Platform and Data Collection** - The Microsoft HoloLens 2 served as the primary experimental platform, offering APIs to capture user gaze and hand gesture data accurately. The experimental environment was developed using Unity (version 2020.3.24f1), the Mixed Reality Toolkit (MRTK 2.8.0), and the OpenXR Plugin (1.7.0). The devices were represented with 3D models anchored in the space: a light, a smart cabinet, a smart screen, an oven, and an air cleaner.
- **Context Library Setup** - The experiment implemented context library as follows:
 - *Interface Function List*: Drawing from the XiaoMi SmartHome API¹, five devices and their corresponding functions were synthesized to form a function list. Each device has 3-5 functions with a total of 18 functions in this scenario.
 - *Gaze Data*: User gaze data was captured using the HoloLens 2 Gaze API and saved as 3D spatial coordinates.
 - *Interaction History*: Interaction history was extracted from the task sequence.
 - *External Context*: There might be context information that is external to our system, which can significantly impact the grounding reasoning. We defined several external contexts corresponding to different tasks to understand if our system can correctly leverage those.
- **Task Descriptions** - Specific eight tasks assigned to participants related to smart device control.
- **Participants** - We recruited 16 participants from three local schools, compensating them at a rate of \$12 per hour. Their ages ranged from 15 to 35 years (MEAN = 26.625, SD = 5.325), comprising 13 males and 3 females.
- **Task Procedure** - Upon their arrival, participants were briefed about the scenario and the devices involved. They are asked to make any gesture deemed most intuitive using the right hand after raising their hand above their chest to initiate the trigger. A preliminary warm-up session was conducted to familiarize the participants with the AR devices and gesture control operations. Following this, they were instructed to complete the eight tasks. Feedback from the devices was simulated to enhance the interaction experience and realism of the study.

Detailed information about the experiment simulation interface, the list of devices and their functions, and the task list is provided in Appendix B.1.

A total of 16 participants \times 8 tasks = 128 gestures were collected.

4.1.2 Results Analysis

We run the collected data through our system with four different context settings to evaluate GestureGPT's performance,

respectively: 1) the baseline without any context other than the function list(*Baseline*), 2) with gaze information (*Only Gaze*), 3) with interaction history and external information (*Only History and External*), and 4) all contexts are available (*All*). We also provide the results of *Random Guess* for comparison, which randomly selects from all candidate functions. We run the results three times under each setting to provide robust conclusions. The main results are listed in the left part of Table 2. We also provide a corresponding illustration in Figure 8.

Our first observation is that GestureGPT can effectively utilize the context information to determine the exact intention of the users. The accuracy is effectively improved to an impressive extent when either gaze or history and external is incorporated. When all contexts are combined together, the overall framework achieves the best Top-1 performance at 44.79%.

When comparing the specific benefits between gaze and history and external information, we empirically find that gaze provides much more significant benefits than history and external information. As can be seen from Table 2, the Only Gaze setting achieves 35.16%/83.59% at Top-1/Top-5 accuracy, outperforming the 23.18%/49.48% achieved by the Only History and External setting by a considerable margin. We attribute this to the semantic similarity of candidate functions shared by different home devices as well as their spatial distribution across the entire room space. These two factors significantly increase the reasoning difficulty under the Smart Home scenario. Facing such challenging situations, GestureGPT can fully exploit the gaze information as well as its own common-sense knowledge to finally locate and zoom in on possible candidates through multi-step, collaborative spatial reasoning. As explained in Section 3.3.1, when the Gesture Inference Agent requests gaze information, the Context Management Agent not only outputs the gaze coordinates but also identifies the relevant device within the gaze path using external tool-augmented Python scripts.

Moreover, while integrating All context setting achieves the highest Top-1 performance, it unexpectedly underperforms in the Top-5 metric compared to the Gaze Only setting (79.69% vs 83.59%). Through further case examination, we hypothesize that an excess of context can sometimes disrupt the agent's analysis, leading to the following behaviors: 1) Irrelevant context information can at times mislead the agent, for example, the mention of "fingerprint unlocking"(in task 1) within external information might lead the agent to infer fingerprint recognition as the unlocking method, rather than a gesture. 2) An abundance of context may cause the agent to overlook certain information, for example, in the 'Open the air purifier' task (in task 5) where the agent incorrectly assumes the device is on based on the external information 'The air purifier's sensor detected that the current environment has heavy cooking fumes', disregarding its actual OFF status. These insights will be further discussed in Section 5.3 regarding optimal context selection.

Finally, we observe that Tasks 2, 3, and 7 exhibit relatively high baseline performance, all of which involve operations

¹<https://iot.mi.com/new/doc/design/spec/xiaomi>

Table 2: Main Results of GestureGPT in the Two Experiments

	Smart Home Scenario				Video Streaming Scenario			
	Top 1 (↑)	Top 3 (↑)	Top 5 (↑)	Negative (↓)	Top 1 (↑)	Top 3 (↑)	Top 5 (↑)	Negative (↓)
Random Guess	5.56%	16.67%	27.78%	72.22%	3.15%	9.46%	15.76%	84.24%
Baseline	10.94% \pm 3.38	24.48% \pm 3.51	35.16% \pm 2.92	64.84% \pm 2.92	19.53% \pm 3.55	38.28% \pm 1.10	54.43% \pm 1.84	45.57% \pm 1.84
Only Gaze	35.16% \pm 1.28	70.05% \pm 1.33	83.59% \pm 1.10	16.41% \pm 1.10	25.78% \pm 0.64	47.66% \pm 3.19	60.42% \pm 2.88	39.58% \pm 2.88
Only History and External	23.18% \pm 4.25	37.50% \pm 4.47	49.48% \pm 4.34	50.52% \pm 4.34	26.30% \pm 3.01	47.14% \pm 5.12	63.28% \pm 3.38	36.72% \pm 3.38
All	44.79% \pm 3.21	67.45% \pm 4.10	79.69% \pm 1.69	20.31% \pm 1.69	37.50% \pm 4.18	59.90% \pm 4.83	73.44% \pm 1.91	26.56% \pm 1.91

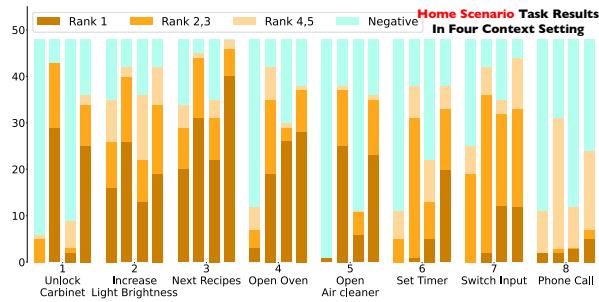


Figure 8: Illustration of results from Experiment 1: Home Scenario, encompassing all 8 tasks (across the x-axis) and 4 settings (indicated by hue). For each task, the 4 settings are arranged from left to right as follows: Baseline, Only Gaze, Only History and External, and All. Each specific bar is further divided into Top 1 / Top 3 / Top 5 accuracy (Top 1, Top 3, Top 5 in gradient brown) and Negative outcomes (in cyan), totaling 48 data points.

related to "switching". For example, in Task 2 'Increasing Light Brightness', GestureGPT first recognizes that the gesture involves significant 'movement', leading the agent to consider that the gesture may be related to some type of value control function. Then, by considering the current status of devices—specifically that the lights were on while other devices with similar switch and control functionalities were mostly off—the agent was able to infer that the gesture is related to Light device control. Similarly, for Tasks 3 and 7, GestureGPT employed a comparable analytical approach to discern the correct answer. This illustrates GestureGPT's strength in analyzing gestures in conjunction with the current states of devices, thereby accurately interpreting user intentions even in the baseline context.

In conclusion, our investigation into GestureGPT's performance across different contexts underscores the significant enhancements brought about by gaze data and the contributions of history and external information. Further refinement in how external information is presented and queried will unlock greater performance of our system.

4.2 Experiment 2: Online Video Streaming on PC

This scenario is set when a user is watching online video on a PC monitor. Grounding gestures to the correct function is much more challenging in this scenario compared to the previous one. The video streaming interface contains a considerably broader range of functions, with numbers up to 66 in some tasks from the previous 18 functions. Moreover, many functions have similar semantic meanings, such as the "vlog channel" button and the "anime channel" button, making them difficult to distinguish solely through gestures. Furthermore, the interactive buttons and elements on the screen are much smaller than the smart appliances in the previous experiment, which reduces the performance of gaze-based function differentiation. The size of function elements ranges from 0.27 to 20.71 cm² (MEAN = 2.73, SD = 3.67). By navigating through an interface rich in functions and semantic complexities, we intend to explore the boundary of our system's performance and understand whether it can differentiate user intentions with only minor differences, which is essential for real-world applications.

4.2.1 Experiment Setting and Procedure

- **Experimental Platform and Gesture Data Collection** - Our experimental framework utilizes Python and Selenium to interact with a video streaming platform, specifically targeting the website "[China] From the Spring and Autumn Period to the Prosperous Tang Dynasty (Season 1, 12 Episodes)" on Bilibili². The platform automates video control operations via Selenium. User gestures are captured using a 1080P resolution webcam.
- **Context Library Configuration** - The study incorporates a comprehensive context library comprising four distinct aspects same as in previous scenario:
 - *Interface Function List:* The function list is automatically extracted from the webpage and organized. Functions on the website included their position and raw HTML code are identified via JavaScript. The code for each function is then extracted and fed into GPT-4 to generate function names, aiding in the compilation of the interface function list. For tasks 4, 5,

²<https://www.bilibili.com/video/BV1sh41j7A4/>

- and 6, the function count is 17; for all other tasks, it is 66.
- *Gaze Data*: The Tobii Eye Tracker 5 is used to collect gaze data from participants.
 - *Interaction History and External Context*: Interaction history was extracted from the task sequence, while the external contexts were predefined.
 - **Task Descriptions** - Eight tasks were designed to simulate common operations performed while watching videos.
 - **Participants** - We recruited 16 participants from four local schools, compensating them at a rate of \$12 per hour. Their ages ranged from 18 to 35 years (MEAN = 26.875, SD = 4.689), comprising 10 males and 6 females.
 - **Task Procedure** - Participants were briefed on the experiment's aims and the devices utilized upon their arrival. They were also asked to make any gesture deemed most intuitive with their right hand, similarly triggered by raising their hand above the chest. There was a warm-up phase for the participant to familiarize with gesture controls. Following this, the participant sequentially completes the eight tasks. Specific gestures triggered predefined responses on the website, simulating real-time interaction for a more realistic experience.

As a result, a total of 16 participants \times 8 tasks = 128 data points were collected.

Detailed information about the experiment simulation interface, the list of devices and their functions, and the task list is provided in Appendix B.2.

4.2.2 Results Analysis

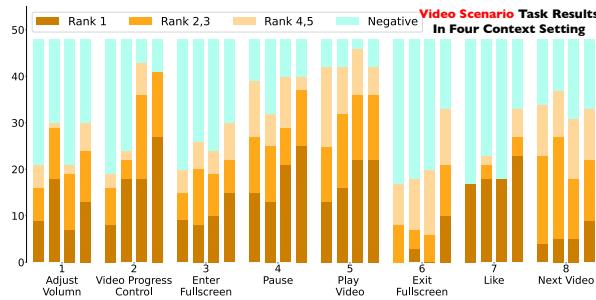


Figure 9: Illustration of results from Experiment 2: Online Video Scenario, encompassing all 8 tasks (across the x-axis) and 4 settings (indicated by hue). For each task, the 4 settings are arranged from left to right as follows: Baseline, Only Gaze, Only History and External, and All. Each specific bar is further divided into Top 1 / Top 3 / Top 5 accuracy (Top 1, Top 3, Top 5 in gradient brown) and Negative outcomes (in cyan), totaling 48 data points.

In this scenario, we evaluate GestureGPT under the same 4 context settings. The results are presented in the right part of Table 2 as well as Figure 9.

The main conclusion that GestureGPT can effectively incorporate various context information to reason and predict user intentions remains consistent. Specifically, the Top-1 accuracy improved from 19.53% to 37.50% when all types of context are available, and also significantly outperforms each single context setting: +11.20% compared to only the history and external setting, and +11.72% compared to the only gaze setting. Both contexts contribute to the final performance, and the performance greatly drops if either of them is excluded. Compared with Experiment 1, where gaze information brings relatively more utility, these results further demonstrate that it requires joint reasoning over all available contexts to achieve the best performance under such a more complicated scenario.

An intriguing observation from our results was that the baseline performance was superior to that observed in the home scenario, thereby necessitating a task-specific exploration. Tasks such as pausing, playing videos, and selecting the next video (Tasks 4, 5, and 8), which require less context, exhibited strong performance across all tested context settings. This phenomenon suggests that the agent's inherent knowledge of video streaming interfaces makes it more probable to infer actions that are more often seen in such a video streaming interface. To some extent, the agent's success in identifying these operations can be attributed more to an educated guess rather than a calculated match of gesture and context, indicating a form of intuition derived from the model's extensive training data. This intuition helps our system's performance in certain tasks, but also acts as bias in other tasks. This phenomenon is discussed further in Section 5.4.

Another special case is Task 7, which is semantic specific task, highlights the essential role of accurate gesture description in achieving precise gesture grounding. Out of 16 participants, 12 performed the 'thumbs-up' gesture for the task. In such cases, if the agent accurately recognizes the gesture, the outcome is correct, notably reflecting in high Top-1 accuracy results. The failures predominantly occurred due to incorrect gesture segmentation. For example, because users formed a fist after lifting their hand, leading the agent to mistakenly focus on the action of making a fist, which leads to map the "Full Screen" or "volume" function. We conducted a pilot validation exercise on the gesture segmentation rule used in our system with a human labeler splitting gesture frames from the whole video on data from eight participants as the ground truth in gesture segmentation. Our results, compared to the ground truth, revealed a high recall rate of 95.28%, yet a precision of only 43.91%, reflecting the inclusion of non-gesture-related frames. Future enhancements could involve advanced object detection techniques to improve segmentation precision.

In cases where gesture descriptions were poorly articulated, incorporating additional context proved beneficial. For example, considering interaction history context, such as users exiting full-screen mode, allowed the agent to infer that the user might want to perform operations unrelated to video control, like "liking" a video. This various context adaptation significantly improved this task performance in Top-5 accuracy metrics in this task from 35.42% to 68.75%.

Despite the challenges we designed in the video streaming scenario, GestureGPT still demonstrates commendable performance, largely attributed to the LLM’s robust common sense and contextual interpretation capabilities.

5 Discussion

5.1 Language Input vs. Visual Input for Agents

GestureGPT is currently built with a Large *Language Model* for its superior understanding and reasoning capability. For the Gesture Description Agent, an alternative implementation could employ a multimodal model, which seems more intuitive. So we tested replacing the rule-based Gesture Description Agent with GPT-4o’s vision in a video streaming scenario. With all context considered, the Top-1 and Top-5 accuracy results were 41% and 72%. These results demonstrate performance comparable to our rule-based module which is 37.5% and 73%, thereby validating the effectiveness of the rule-based approach.

On the other hand, LMMs require users to upload gesture recordings, which raises further privacy concerns. By contrast, our solution can process the gesture input with end-affordable devices and only send anonymized skeleton signals to the data center, where LLMs can perform dense computing. Nevertheless, it is anticipated that with further development of LMMs, GestureGPT could integrate both the visual recognition capabilities of LMMs and the prominent commonsense understanding, contextual reasoning capabilities of LLMs, to ultimately better serve its purpose.

5.2 Gesture Description Quality Assessment

The quality of the gesture description agent was evaluated through an expert questionnaire. For this evaluation, we randomly sampled descriptions generated from three repetitions for 256 gestures across two scenarios, resulting in a compilation of 256 gestures and their descriptions. The same questionnaire was then distributed to two gesture experts, who rated each gesture description on a 5-point Likert scale. Both experts have published gesture-related research articles on premier conferences. One expert’s score was 3.28 (std=1.41), and the other’s was 3.74 (std=0.73). The positive ratings show that our description agent can capture key information of the gesture.

5.3 Context Selection in Complex Systems

Our experiments revealed that although adding more contexts is supposed to be informative and useful, for certain circumstances, it might also bring irrelevant noise and accordingly distract the reasoning process of LMM agents (See Top-5 accuracy of all setting vs only gaze setting in Home scenario). In the home scenario, most informative contexts are concentrated on gesture itself and gaze, while history and external provide relatively less contribution to the final performance.

On one hand, the inclusion of more heterogeneous contexts increases the complexity and difficulty of context organization

and management, which then requires more competent agents to process and reason over them. On the other hand, contexts with less information entropy inevitably brings irrelevant implications, and may result in misleading of agents for specific cases. As a result, this underscores the importance of making informed trade-offs between context incorporation and agent capability to optimize the system performance.

5.4 LLM Common Sense Bias

This phenomenon pertains to the cognitive biases of LLM, which has been recognized as a common issue that influences LLM outputs [49].

Such a bias was also observed in our system, notably within the video streaming scenario. When the agent learned that this is a video streaming scenario, there is a bias towards predicting video control functions that are commonly used on such an interface, such as play, pause, and volume control. On one hand, even in the absence of context, the agent can make accurate guesses if the intended function is one of such functions. On the other hand, it leads to the consistent inclusion of these functions within the candidate options, detrimentally impacting the top-1 accuracy.

One way to address this bias could involve employing Modular Debiasing Networks [13] to mitigate bias or utilizing sophisticated prompt engineering to diminish its effects, which we plan to investigate in the future.

5.5 System Scalability

By substituting the Gesture Description Agent with specially designed counterparts, our system can adapt to more input modalities and more generalized form of gestures.

In our implementation, gesture feature extraction is solely based on a RGB camera and *MediaPipe*. Yet this approach is susceptible to lighting conditions and finger occlusion issues. Wearable devices provide an alternative robust solution for hand reconstruction [22, 67, 72], which even works with subtle gestures like thumb-tip gestures [19] and wrist gestures [17, 18]. Hand landmarks reconstructed from wearable sensors can then be used to generate gesture description.

Our system can also be extended to general gestures, beyond the scope of merely hand gestures. For example, touch-screen gestures can be extracted as traces and pressure intensity, which differ significantly from the form of hand gestures. But a specially designed Gesture Description Agent can extract the features of such gestures (either using rule-based methods or leveraging large language or vision models) and describe it to *Gesture Inference Agent*, thus easily integrated into our framework.

6 Limitation and Future Work

6.1 Recognition of Non-Interactive Gestures

Currently, GestureGPT identifies meaningful gestures with the system triggered by the action of raising a hand above chest to start a gesture. However, this approach may lead to

false activation, as many gestures not intended for interaction could inadvertently trigger the system. In real-world scenarios, where various non-interactive movements occur, this can interfere with the LLM model's reasoning process, consequently reducing system performance.

There are already several studies addressing this issue. Researchers have developed frameworks using body pose, gaze, and gesture data [50] or the multimodal feedback mechanism [12] to predict interaction intent, while others have integrated EEG and gaze patterns to differentiate meaningful attention during tasks [16, 31]. Our system allows easy integration of advanced gaze pattern analysis algorithms and new input channels, which will greatly help in discerning the meaningful interactive gesture.

6.2 Integration and Utilization of Top-N Predictions

GestureGPT currently outputs a Top-5 candidate function list, which can be readily augmented with list selection interaction optimizations, as studied in many existing works. Specifically, Isomoto et al. [25] developed a dwell selection system that employs machine learning to predict user intent solely based on eye movement data. Furthermore, statistical design of dynamic menus could also expedite user selection particularly in AR/VR contexts. G-Menu [57] utilizes gestures and keywords to design a dynamic menu, aiding in swift function selection. These techniques and designs can be readily integrated into GestureGPT to achieve efficient user selection of the target function.

6.3 System Delay

The overall time cost of the framework is predominantly occupied by the LLM-related costs, accounting for up to 90% of the current time expenditure. These costs include LLM API request costs. For each gesture understanding task, GestureGPT typically requires 1 request for the gesture description and 6 rounds of conversation between the *Gesture Inference Agent* and *Context Management Agent*, which indicates great impact of response delay on the realtimeness of our system. Without focusing on practicality, our conceptual framework averages 227 seconds and 38,785 tokens per task. Possible directions for improvements:

- **KV-cache:** API calls require concatenating long system prompts each round, exponentially increasing time costs. Our current implementation adds the entire context library (11,000 tokens) to each query. KV-cache saves computed attention values after each round, significantly reducing time. The task execution time drops to 20 seconds after removing the context from latter rounds.
- **Inference speed:** Groq's fast model inference services have an output speed of over 700 tokens per second. Combining Groq (without rate limits) and KV-cache, the first context round will take 2-3 seconds, while the remaining rounds taking a total of about 1-2 seconds

(averaging 6 dialogue rounds with input of 200 tokens and output of 100 tokens each round). This would reduce the theoretical optimal latency to under 5 seconds.

Even though we do not have access to KV-cache-ready high-performance LLMs nor computing acceleration hardware as used by Groq, we anticipate GestureGPT will attract interest from other research domains, fostering collaborative efforts towards a practical and accessible free-form gesture interface.

On the other hand, since each agent of our system has specialized and clear goals, we believe a properly fine-tuned LLM could mirror GPT-4's effectiveness with fewer resources as demonstrated in NLP tasks [78]. Besides, advances in model distillation and acceleration hardware suggest that lightweight LLMs could be deployed directly on devices with fast inference [35, 76, 76] for real-time performance in the future. Under which circumstances, the API request cost can also be removed.

7 Conclusion

In this work, we propose GestureGPT, a free-form hand gesture understanding framework that leverages the large language models' capability to automatically map a spontaneously performed free-form gesture to its targeted interface function. The framework is designed and implemented in the form of triple-agent collaboration, which interprets gesture poses and movements, manages context information, and reason over common sense, to finally discern user intent. In two real-world scenarios, we achieved the highest zero-shot Top-5 gesture grounding accuracy of 83.59% for smart home control and 73.44% for video streaming. The inspiring results showcase, for the first time, the great potential of a free-form hand gesture understanding system, eliminating the need for users to learn, memorize, demonstrate, or manually associate gestures with specific functions. GestureGPT paves way for future research on natural interaction interfaces including but not limited to gestures.

References

- [1] Deepak Akkil and Poika Isokoski. 2016. Gaze Augmentation in Egocentric Video Improves Awareness of Intention. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, San Jose California USA, 1573–1584. <https://doi.org/10.1145/2858036.2858127>
- [2] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosius, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schieber, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. 2022. Constitutional AI: Harmlessness from AI Feedback. <http://arxiv.org/abs/2212.08073> arXiv:2212.08073 [cs].
- [3] Sigal Berman and Helman Stern. 2012. Sensors for Gesture Recognition Systems. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 42, 3 (May 2012), 277–290. <https://doi.org/10.1109/TSMCC.2011.2161077> Conference Name: IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews).
- [4] Anthony Brohan, Noah Brown, Justice Carbalaj, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, Pete Florence, Chuyuan Fu, Montse Gonzalez Arenas, Keerthana Gopalakrishnan, Kehang Han, Karol Hausman, Alexander Herzog, Jasmine Hsu, Brian Ichter, Alex Irpan, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Lisa Lee, Tsang-Wei Edward Lee, Sergey Levine, Yao Lu, Henryk Michalewski, Igor Mordatch, Karl Pertsch, Kanishka Rao, Krista Reymann, Michael Ryoo, Grecia Salazar, Pannag Sanketi, Pierre Sermanet, Jaspal Singh, Anikait Singh, Radu Soricut, Huong Tran, Vincent Vanhoucke, Quan Vuong, Ayaan Wahid, Stefan Welker, Paul Wohlhart, Jialin Wu, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. 2023. RT-2: Vision-Language-Action Models Transfer Knowledge to Robotic Control. <http://arxiv.org/abs/2307.15818> arXiv:2307.15818 [cs].
- [5] Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. ChatEval: Towards Better LLM-based Evaluators through Multi-Agent Debate. <http://arxiv.org/abs/2308.07201> arXiv:2308.07201 [cs].
- [6] Ishan Chatterjee, Robert Xiao, and Chris Harrison. 2015. Gaze+Gesture: Expressive, Precise and Targeted Free-Space Interactions. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. ACM, Seattle Washington USA, 131–138. <https://doi.org/10.1145/2818346.2820752>
- [7] Weihao Chen, Chun Yu, Huadong Wang, Zheng Wang, Lichen Yang, Yukun Wang, Weinan Shi, and Yuanchun Shi. 2023. From Gap to Synergy: Enhancing Contextual Understanding through Human-Machine Collaboration in Personalized Systems. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23)*. Association for Computing Machinery, New York, NY, USA, 1–15. <https://doi.org/10.1145/3586183.3606741>
- [8] William Delamare, Chaklam Silpasuwanchai, Sayan Sarcar, Toshiaki Shiraki, and Xiangshi Ren. 2019. On Gesture Combination: An Exploration of a Solution to Augment Gesture Interaction. In *Proceedings of the 2019 ACM International Conference on Interactive Surfaces and Spaces*. ACM, Daejeon Republic of Korea, 135–146. <https://doi.org/10.1145/3343055.3359706>
- [9] Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Samuel Stevens, Boshi Wang, Huan Sun, and Yu Su. 2023. Mind2Web: Towards a Generalist Agent for the Web. <http://arxiv.org/abs/2306.06070> arXiv:2306.06070 [cs].
- [10] Danny Driess, Fei Xia, Mehdil S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. 2023. PaLM-E: An Embodied Multimodal Language Model. <http://arxiv.org/abs/2303.03378> arXiv:2303.03378 [cs].
- [11] Afshan Ejaz, Maria Rahim, and Shakeel Ahmed Khoja. 2019. The Effect of Cognitive Load on Gesture Acceptability of Older Adults in Mobile Application. In *2019 IEEE 10th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*. IEEE, New York City, NY, USA, 0979–0986. <https://doi.org/10.1109/UEMCON47517.2019.8992970>
- [12] Euan Freeman, Stephen Brewster, and Vuokko Lantz. 2016. Do That, There: An Interaction Technique for Addressing In-Air Gesture Systems. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. Association for Computing Machinery, New York, NY, USA, 2319–2331. <https://doi.org/10.1145/2858036.2858308>
- [13] Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2024. Bias and Fairness in Large Language Models: A Survey. <http://arxiv.org/abs/2309.00770> arXiv:2309.00770 [cs].
- [14] Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. PAL: Program-aided Language Models. <https://doi.org/10.48550/arXiv.2211.10435> arXiv:2211.10435 [cs].
- [15] Qi Gao, Zheng Ma, Quan Gu, Jiaofeng Li, and Zaifeng Gao. 2023. Working Memory Capacity for Gesture-Command Associations in Gestural Interaction. *International Journal of Human-Computer Interaction* 39, 15 (Sept. 2023), 3045–3056. <https://doi.org/10.1080/10447318.2022.2091213>
- [16] Xianliang Ge, Yunxian Pan, Sujie Wang, Linze Qian, Jingjia Yuan, Jie Xu, Nitish Thakor, and Yu Sun. 2023. Improving Intention Detection in Single-Trial Classification Through Fusion of EEG and Eye-Tracker Data. *IEEE Transactions on Human-Machine Systems* 53, 1 (Feb. 2023), 132–141. <https://doi.org/10.1109/THMS.2022.3225633>
- [17] Jun Gong, Zheer Xu, Qifan Guo, Teddy Seyed, Xiang 'Anthony' Chen, Xiaojun Bi, and Xing-Dong Yang. 2018. WrisText: One-handed Text Entry on Smartwatch using Wrist Gestures. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, Montreal QC Canada, 1–14. <https://doi.org/10.1145/3173574.3173755>
- [18] Jun Gong, Xing-Dong Yang, and Pourang Irani. 2016. WristWhirl: One-handed Continuous Smartwatch Input using Wrist Gestures. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*. ACM, Tokyo Japan, 861–872. <https://doi.org/10.1145/2984511.2984563>
- [19] Jun Gong, Yang Zhang, Xia Zhou, and Xing-Dong Yang. 2017. Pyro: Thumb-Tip Gesture Recognition Using Pyroelectric Infrared Sensing. In *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology*. ACM, Québec City QC Canada, 553–563. <https://doi.org/10.1145/3126594.3126615>
- [20] Izzeddin Gur, Hiroki Furuta, Austin Huang, Mustafa Safdari, Yutaka Matsuo, Douglas Eck, and Aleksandra Faust. 2024. A Real-World WebAgent with Planning, Long Context Understanding, and Program Synthesis. <http://arxiv.org/abs/2307.12856> arXiv:2307.12856 [cs].
- [21] Izzeddin Gur, Ofir Nachum, Yingjie Miao, Mustafa Safdari, Austin Huang, Aakanksa Chowdhery, Sharan Narang, Noah Fiedel, and Aleksandra Faust. 2023. Understanding HTML with Large Language Models. <http://arxiv.org/abs/2210.03945> arXiv:2210.03945 [cs].
- [22] Fang Hu, Peng He, Songlin Xu, Yin Li, and Cheng Zhang. 2020. FingerTrak: Continuous 3D Hand Pose Tracking by Deep Learning Hand Silhouettes Captured by Miniature Thermal Cameras on Wrist. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 2 (June 2020), 1–24. <https://doi.org/10.1145/3397306>
- [23] Chien-Ming Huang, Sean Andrist, Allison Sauppé, and Bilge Mutlu. 2015. Using gaze patterns to predict task intent in collaboration. *Frontiers in Psychology* 6 (July 2015). <https://doi.org/10.3389/fpsyg.2015.01049>
- [24] Lili Idan. 2022. A Network-Based, Multidisciplinary Approach to Intention Inference. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. ACM, New Orleans LA USA, 1–7. <https://doi.org/10.1145/3491101.3519754>
- [25] Toshiya Isomoto, Shota Yamanaka, and Buntarou Shizuki. 2022. Dwell Selection with ML-based Intent Prediction Using Only Gaze Data. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 3, Article 120 (sep 2022), 21 pages. <https://doi.org/10.1145/3550301>
- [26] Huiqiang Jiang, Qianhui Wu, Xufang Luo, Dongsheng Li, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. 2023. LongLLMLingua: Accelerating and Enhancing LLMs in Long Context Scenarios via Prompt Compression. (2023). <https://doi.org/10.48550/ARXIV.2310.06839> Publisher: [object Object] Version Number: 1.
- [27] Shuo Jiang, Ling Li, Haipeng Xu, Junkai Xu, Guoying Gu, and Peter B. Shull. 2020. Stretchable e-Skin Patch for Gesture Recognition on the Back of the Hand. *IEEE Transactions on Industrial Electronics* 67, 1 (Jan. 2020), 647–657. <https://doi.org/10.1109/TIE.2019.2914621>
- [28] Alexander Kapitanov, Karina Kvanchiani, Alexander Nagaev, Roman Kraynov, and Andrei Makhilarchuk. 2024. HaGRID - HAnd Gesture Recognition Image Dataset. <http://arxiv.org/abs/2206.08219> arXiv:2206.08219 [cs].
- [29] Evan King, Haoxiang Yu, Sangsu Lee, and Christine Julien. 2023. "Get ready for a party": Exploring smarter smart spaces with help from large language models. <http://arxiv.org/abs/2303.14143> arXiv:2303.14143 [cs].
- [30] Evan King, Haoxiang Yu, Sangsu Lee, and Christine Julien. 2024. Sasha: Creative Goal-Oriented Reasoning in Smart Homes with Large Language

- Models. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, 1 (March 2024), 1–38. <https://doi.org/10.1145/3643505>
- [31] Fatemeh Koochaki and Laleh Najafizadeh. 2018. Predicting Intention Through Eye Gaze Patterns. In *2018 IEEE Biomedical Circuits and Systems Conference (BioCAS)*. IEEE, Cleveland, OH, 1–4. <https://doi.org/10.1109/BIOCAS42018.8584665>
- [32] Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Zhaopeng Tu, and Shuming Shi. 2023. Encouraging Divergent Thinking in Large Language Models through Multi-Agent Debate. <http://arxiv.org/abs/2305.19118> arXiv:2305.19118 [cs].
- [33] Hao Lü and Yang Li. 2011. Gesture avatar: a technique for operating mobile user interfaces using gestures. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, Vancouver BC Canada, 207–216. <https://doi.org/10.1145/1978942.1978972>
- [34] Hao Lü and Yang Li. 2012. Gesture coder: a tool for programming multi-touch gestures by demonstration. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12)*. Association for Computing Machinery, New York, NY, USA, 2875–2884. <https://doi.org/10.1145/2207676.2208693>
- [35] Ruilong Ma, Jingyu Wang, Qi Qi, Xiang Yang, Haifeng Sun, Zirui Zhuang, and Jianxian Liao. 2023. Poster: PipeLLM: Pipeline LLM Inference on Heterogeneous Devices with Sequence Slicing. In *Proceedings of the ACM SIGCOMM 2023 Conference (ACM SIGCOMM '23)*. Association for Computing Machinery, New York, NY, USA, 1126–1128. <https://doi.org/10.1145/3603269.3610856>
- [36] Naveen Madapana and Juan Wachs. 2019. Database of Gesture Attributes: Zero Shot Learning for Gesture Recognition. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*. IEEE, Lille, France, 1–8. <https://doi.org/10.1109/FG2019.8756548>
- [37] Naveen Madapana and Juan P. Wachs. 2018. Hard Zero Shot Learning for Gesture Recognition. In *2018 24th International Conference on Pattern Recognition (ICPR)*. IEEE, Beijing, 3574–3579. <https://doi.org/10.1109/ICR.2018.8545869> zero shot.
- [38] George B. Mo, John J Dudley, and Per Ola Kristensson. 2021. Gesture Knitter: A Hand Gesture Design Tool for Head-Mounted Mixed Reality Applications. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, 1–13. <https://doi.org/10.1145/3411764.3445766>
- [39] José Antonio Montero and L. Enrique Sucar. 2006. Context-Based Gesture Recognition. In *Progress in Pattern Recognition, Image Analysis and Applications*, David Hutchison, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Friedemann Mattern, John C. Mitchell, Moni Naor, Oscar Nierstrasz, C. Pandu Rangan, Bernhard Steffen, Madhu Sudan, Demetri Terzopoulos, Dough Tygar, Moshe Y. Vardi, Gerhard Weikum, José Francisco Martínez-Trinidad, Jesús Ariel Carrasco Ochoa, and Josef Kittler (Eds.). Vol. 4225. Springer Berlin Heidelberg, Berlin, Heidelberg, 764–773. http://link.springer.com/10.1007/11892755_79
- [40] Louis-Philippe Morency and Trevor Darrell. 2006. Head gesture recognition in intelligent interfaces: the role of context in improving recognition. In *Proceedings of the 11th international conference on Intelligent user interfaces*. ACM, Sydney Australia, 32–38. <https://doi.org/10.1145/1111449.1111464>
- [41] Miguel A. Nacenta, Yemliha Kamber, Yizhou Qiang, and Per Ola Kristensson. 2013. Memorability of pre-designed and user-defined gesture sets. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*. Association for Computing Machinery, New York, NY, USA, 1099–1108. <https://doi.org/10.1145/2470654.2466142>
- [42] Donald A. Norman. 2010. Natural user interfaces are not natural. *Interactions* 17, 3 (May 2010), 6–10. <https://doi.org/10.1145/1744161.1744163>
- [43] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuskin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Deareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhusn Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Lukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kirov, Matt Knight, Daniel Kokotajlo, Lukasz Kondraciuk, Andrew Kondrich, Ari Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Iki Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makajnu, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavy, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emry Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Maria Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cérion Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, C. J. Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willnall, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. GPT-4 Technical Report. <http://arxiv.org/abs/2303.08774> [cs].
- [44] Jeongeon Park, Bryan Min, Xiaojuan Ma, and Juho Kim. 2023. Choice-Mates: Supporting Unfamiliar Online Decision-Making with Multi-Agent Conversational Interactions. <https://doi.org/10.48550/arXiv.2310.01331> arXiv:2310.01331 [cs].
- [45] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative Agents: Interactive Simulacra of Human Behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. ACM, San Francisco CA USA, 1–22. <https://doi.org/10.1145/3586183.3606763>
- [46] V.I. Pavlovic, R. Sharma, and T.S. Huang. 1997. Visual interpretation of hand gestures for human-computer interaction: a review. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 19, 7 (July 1997), 677–695. <https://doi.org/10.1109/34.598226> Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.
- [47] Chen Qian, Xin Cong, Wei Liu, Cheng Yang, Weize Chen, Yusheng Su, Yufan Dang, Jiahao Li, Juyuan Xu, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2023. Communicative Agents for Software Development. <http://arxiv.org/abs/2307.07924> arXiv:2307.07924 [cs].
- [48] Dmitriy Rivkin, Francois Hogan, Amal Feriani, Abhishek Konar, Adam Sigal, Steve Liu, and Greg Dudek. 2024. SAGE: Smart home Agent with Grounded Execution. <http://arxiv.org/abs/2311.00772> arXiv:2311.00772 [cs].
- [49] Siddharth Samsi, Dan Zhao, Joseph McDonald, Baolin Li, Adam Michaleas, Michael Jones, William Bergeron, Jeremy Kepner, Devesh Tiwari, and Vijay Gadepally. 2023. From Words to Watts: Benchmarking the Energy Costs of Large Language Model Inference. <http://arxiv.org/abs/2310.03003> arXiv:2310.03003 [cs].
- [50] Julia Schwarz, Charles Claudio Marais, Tommer Leyvand, Scott E. Hudson, and Jennifer Mankoff. 2014. Combining body pose, gaze, and gesture to determine intention to interact in vision-based interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, Toronto Ontario Canada, 3443–3452. <https://doi.org/10.1145/2556288.2556989>

- [51] Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language Agents with Verbal Reinforcement Learning. <http://arxiv.org/abs/2303.11366> [cs].
- [52] Ronal Singh, Tim Miller, Joshua Newn, Liz Sonenberg, Eduardo Velloso, and Frank Vetere. 2018. Combining Planning with Gaze for Online Human Intention Recognition. (2018).
- [53] Maximilian Speicher and Michael Nebeling. 2018. GestureWiz: A Human-Powered Gesture Design Environment for User Interface Prototypes. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–11. <https://doi.org/10.1145/3173574.3173681>
- [54] Zhida Sun, Sitong Wang, Chengzhong Liu, and Xiaojuan Ma. 2022. Metaphoration: Support Gesture-based Interaction Design with Metaphorical Meanings. *ACM Transactions on Computer-Human Interaction* 29, 5 (Oct. 2022), 1–33. <https://doi.org/10.1145/3511892> [meaning].
- [55] Yashar Talebirad and Amirhossein Nadiri. 2023. Multi-Agent Collaboration: Harnessing the Power of Intelligent LLM Agents. <http://arxiv.org/abs/2306.03314> arXiv:2306.03314 [cs].
- [56] Eugene M. Taranta Ii, Thaddeus K. Simons, Rahul Sukthankar, and Joseph J. Laviola Jr. 2015. Exploring the Benefits of Context in 3D Gesture Recognition for Game-Based Virtual Environments. *ACM Transactions on Interactive Intelligent Systems* 5, 1 (March 2015), 1–34. <https://doi.org/10.1145/2656345>
- [57] Jean Vanderdonckt and Éric Petit. 2019. G-Menu: A Keyword-by-Gesture Based Dynamic Menu Interface for Smartphones. In *Human-Computer Interaction, Recognition and Interaction Technologies: Thematic Area, HCI 2019, Held as Part of the 21st HCI International Conference, HCII 2019, Orlando, FL, USA, July 26–31, 2019, Proceedings, Part II*. Springer, 99–114.
- [58] Radu-Daniel Vatavu. 2012. User-defined gestures for free-hand TV control. In *Proceedings of the 10th European Conference on Interactive TV and Video (EuroITV '12)*. Association for Computing Machinery, New York, NY, USA, 45–48. <https://doi.org/10.1145/2325616.2325626>
- [59] Chan Wah Ng and Surendra Ranganath. 2002. Real-time gesture recognition system and application. *Image and Vision Computing* 20, 13 (Dec. 2002), 993–1007. [https://doi.org/10.1016/S0262-8856\(02\)00113-0](https://doi.org/10.1016/S0262-8856(02)00113-0)
- [60] Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, and Ji-Rong Wen. 2024. A Survey on Large Language Model based Autonomous Agents. <http://arxiv.org/abs/2308.11432> arXiv:2308.11432 [cs].
- [61] Tianyi Wang, Xun Qian, Fengming He, Xiyun Hu, Yuanzhi Cao, and Karthik Ramani. 2021. GesturAR: An Authoring System for Creating Freehand Interactive Augmented Reality Applications. In *The 34th Annual ACM Symposium on User Interface Software and Technology (UIST '21)*. Association for Computing Machinery, New York, NY, USA, 552–567. <https://doi.org/10.1145/3472749.3474769>
- [62] Zhenhailong Wang, Shaoguang Mao, Wenshan Wu, Tao Ge, Furu Wei, and Heng Ji. 2023. Unleashing the Emergent Cognitive Synergy in Large Language Models: A Task-Solving Agent through Multi-Persona Self-Collaboration. (2023). <https://doi.org/10.48550/ARXIV.2307.05300> Publisher: [object Object] Version Number: 4.
- [63] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. <http://arxiv.org/abs/2201.11903> arXiv:2201.11903 [cs].
- [64] Alan Wexelblat. 1995. An approach to natural gesture in virtual environments. *ACM Transactions on Computer-Human Interaction* 2, 3 (Sept. 1995), 179–200. <https://doi.org/10.1145/210079.210080>
- [65] Jinting Wu, Yujia Zhang, and Xiaoguang Zhao. 2021. A Prototype-Based Generalized Zero-Shot Learning Framework for Hand Gesture Recognition. In *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, Milan, Italy, 3435–3442. <https://doi.org/10.1109/ICPR48806.2021.9412548>
- [66] Haijun Xia, Michael Glueck, Michelle Annett, Michael Wang, and Daniel Wigdor. 2022. Iteratively Designing Gesture Vocabularies: A Survey and Analysis of Best Practices in the HCI Literature. *ACM Transactions on Computer-Human Interaction* 29, 4 (2022), 37:1–37:54. <https://doi.org/10.1145/3503537>
- [67] Xuhai Xu, Jun Gong, Carolina Brum, Lilian Liang, Bongsoo Suh, Shivam Kumar Gupta, Yash Agarwal, Laurence Lindsey, Runchang Kang, Behrooz Shahsavari, Tu Nguyen, Heriberto Nieto, Scott E Hudson, Charlie Maalouf, Jax Seyed Mousavi, and Gierad Laput. 2022. Enabling Hand Gesture Customization on Wrist-Worn Devices. In *CHI Conference on Human Factors in Computing Systems*. ACM, New Orleans LA USA, 1–19. <https://doi.org/10.1145/3491102.3501904>
- [68] Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. 2022. WebShop: Towards Scalable Real-World Web Interaction with Grounded Language Agents. In *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Eds.), Vol. 35. Curran Associates, Inc., 20744–20757. https://proceedings.neurips.cc/paper_files/paper/2022/file/82ad13ec01f9fe44c01cb91814fd/b8c-Paper-Conference.pdf
- [69] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing Reasoning and Acting in Language Models. <http://arxiv.org/abs/2210.03629> arXiv:2210.03629 [cs].
- [70] Yiyu Yao. 2012. An Outline of a Theory of Three-Way Decisions. In *Rough Sets and Current Trends in Computing*, David Hutchison, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Friedemann Mattern, John C. Mitchell, Moni Naor, Oscar Nierstrasz, C. Pandu Rangan, Bernhard Steffen, Madhu Sudan, Demetri Terzopoulos, Doug Tygar, Moshe Y. Vardi, Gerhard Weikum, JingTao Yao, Yan Yang, Roman Słowiński, Salvatore Greco, Huaxiong Li, Sushmita Mitra, and Lech Polkowski (Eds.), Vol. 7413. Springer Berlin Heidelberg, Berlin, Heidelberg, 1–17. https://doi.org/10.1007/978-3-642-32115-3_1 Series Title: Lecture Notes in Computer Science.
- [71] Fan Zhang, Valentin Bazarevsky, Andrey Vakunov, Andrei Tkachenko, George Sung, Chuo-Ling Chang, and Matthias Grundmann. 2020. MediPipe Hands: On-device Real-time Hand Tracking. <http://arxiv.org/abs/2006.10214> arXiv:2006.10214 [cs].
- [72] Qiang Zhang, Yuqiao Lin, Yubin Lin, and Szymon Rusinkiewicz. 2023. UltraGlove: Hand Pose Estimation with Mems-Ultrasonic Sensors. <http://arxiv.org/abs/2306.12652> arXiv:2306.12652 [cs].
- [73] Xinghua Zhang, Bowen Yu, Haiyang Yu, Yangyu Lv, Tingwen Liu, Fei Huang, Hongbo Xu, and Yongbin Li. 2023. Wider and Deeper LLM Networks are Fairer LLM Evaluators. <http://arxiv.org/abs/2308.01862> arXiv:2308.01862 [cs].
- [74] Yifan Zhang, Congqi Cao, Jian Cheng, and Hanqing Lu. 2018. EgoGesture: A New Dataset and Benchmark for Egocentric Hand Gesture Recognition. *IEEE Transactions on Multimedia* 20, 5 (May 2018), 1038–1050. <https://doi.org/10.1109/TMM.2018.2808769>
- [75] Yu Zhang, Tao Gu, Chu Luo, Vassilis Kostakos, and Aruna Seneviratne. 2018. FinDroidHR: Smartwatch Gesture Input with Optical Heartrate Monitor. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 1 (2018), 56:1–56:42. <https://doi.org/10.1145/3191788>
- [76] Junchen Zhao, Yurun Song, Simeng Liu, Ian G. Harris, and Sangeetha Abdu Jyothi. 2023. LinguaLinked: A Distributed Large Language Model Inference System for Mobile Devices. <http://arxiv.org/abs/2312.00388> arXiv:2312.00388 [cs].
- [77] Boyuan Zheng, Boyu Gou, Jihyung Kil, Huan Sun, and Yu Su. 2024. GPT-4Vision) is a Generalist Web Agent, if Grounded. <http://arxiv.org/abs/2401.01614> arXiv:2401.01614 [cs].
- [78] Qihuang Zhong, Liang Ding, Juhua Liu, Bo Du, and Dacheng Tao. 2023. Can ChatGPT Understand Too? A Comparative Study on ChatGPT and Fine-tuned BERT. <http://arxiv.org/abs/2302.10198> arXiv:2302.10198 [cs] version: 2.
- [79] Mingchen Zhuge, Haozhe Liu, Francesco Faccio, Dylan R. Ashley, Róbert Csordás, Anand Gopalakrishnan, Abdullah Hamdi, Hasan Abed Al Kader Hammoud, Vincent Herrmann, Kazuki Irie, Louis Kirsch, Bing Li, Guohao Li, Shuming Liu, Jinjie Mai, Piotr Piękos, Aditya Ramesh, Imanol Schlag, Weimin Shi, Aleksandar Stanić, Wenyi Wang, Yuhui Wang, Mengmeng Xu, Deng-Ping Fan, Bernard Ghanem, and Jürgen Schmidhuber. 2023. Mindstorms in Natural Language-Based Societies of Mind. <http://arxiv.org/abs/2305.17066> arXiv:2305.17066 [cs].

A Gesture Description Rules

A.1 Training for Threshold and Evaluation for Gesture Description Rules

We use HaGRID [28] dataset to tune and evaluate our rules. HaGRID contains images of 18 kinds of gestures (call, dislike, fist, four, like, mute, ok, one, palm, peace, peace inv., rock, stop, stop inv., three, three 2, two up, two up inv.).

A.1.1 Training Process

A subsample split of HaGRID has 100 images per gesture, and it is used to tune the rule parameters.

To tune the rule parameters, we first manually annotate the ground truth label for each rule on each gesture class. For each gesture, we assume that most people perform it in the same way, and thus the ground truth label is obtained by common knowledge. (For example, for ‘thumb up’ gesture, we label thumb as straight, index to pinky fingers as bent, index and middle finger are pressed together, etc.)

However, there exist ambiguous cases, in which more than one labels are acceptable. (For example, in ‘peace’ gesture, it is reasonable for thumb to be either straight or bent.) In this case, the label is more than one, and consequently cases like this are not used for parameter tuning.

For most rules are tuned using the whole subsample split. One exception is **Finger Flexion** on thumb, because in most gesture classes, the ground truth of thumb is either [‘straight’] or [‘straight’, ‘bent’] (thus not used for training), and only one class is [‘bent’]. To address the issue of severe imbalance for thumb flexion, we specially select the training set, which is composed of gesture ‘like’ (thumb is straight) and gesture ‘ok’ (thumb is bent), and each image in this set is manually checked to ensure the quality of the training data).

As is shown above, for those cases used to tune the rules, the ground truth label is one of the two possible states (e.g., ‘straight’ or ‘bent’). But during prediction, we use Three-Way Decision Making method [70]. When the rule cannot clearly determine the state, it will predict it as ‘unsure’, to avoid wrong conclusions that may easily mislead LLMs. Consequently, different from the classical binary classification, some modifications have been made to the correctness assessment and tuning criteria to accommodate our specific setting.

Correctness Assessment. For each (prediction, label) pair, there are three possible circumstances:

- Unsure: Prediction = ‘unsure’.
- Error: Prediction ≠ ‘unsure’, and prediction = label.
- Correct: Prediction ≠ ‘unsure’, and prediction ≠ label.

Tuning criteria. To find the optimal parameters, we define the loss in different cases as follows:

- Unsure: Loss = 0.2
- Error: Loss = 1
- Correct: Loss = 0

In ‘unsure’ cases, the loss is between 0 and 1. This is because if we set the loss to be 0 (same as ‘correct’ case), then the rules tend to predict every case to ‘unsure’; if we set the loss to be 1

(same as ‘error’ case), then the rules tend to not predict any ‘unsure’, which may lead to misleading predictions.

A grid search is conducted to find out the optimal parameter with minimal average loss. The optimal parameters for each rule is shown in Table 3

A.1.2 Test Performance

We test the generalization ability of our rules on two dataset: HaGRID test set (third-view) and EgoGesture dataset (first view) [74].

The performance on HaGRID test set (38576 images) are shown in Table 3. For most rules, the average error rate among all gestures is below 4.2% and the overall accuracy is above 88.2%. One exception is flexion rule for thumb, in which the output ‘unsure’ takes about 45% of all cases. This may be attributed to the unique shape of thumb: the topmost segment of the thumb is curved by nature, so when the thumb is extended, the landmark seems to be slightly bent, which may affect the train and test process. But by Three-Way Decision, we use ‘unsure’ to avoid the misleading information that may potentially confuse the *Gesture Description Agent*.

When checking the algorithm’s performance for each gesture, it is shown that the error rates for most gestures are below 5% and none of them are above 16%. The error could be attributed to landmark mistakes made by MediaPipe, the shortcomings of our algorithm, and a small number of people just perform the gesture differently from most people (i.e., different from the ground truth we established before). The results show good generalizability since the parameters are only tuned on a small number of samples.

Table 3: Rule Parameters, and Their Performance on HaGRID Test Set

Rule	Parameters	Performance on HaGRID Test Set		
		error	unsure	correct
Flexion - thumb	(16, 38)	0.036	0.457	0.507
Flexion - other fingers	(57, 74)	0.019	0.049	0.932
Proximity	(0.024, 0.029)	0.031	0.067	0.902
Contact	(0.046, 0.055)	0.020	0.024	0.956
Pointing Direction - thumb	40	0.047	0.239	0.714
Palm Orientation	41	0.042	0.075	0.882
Overall	-	0.023 ± 0.018	0.062 ± 0.039	0.916 ± 0.053

To evaluate if the parameters trained on third-view dataset can adapt to first-view images, we tested them on a first-view gesture dataset EgoGesture. We choose 20 gesture classes (fist, measure, zero, one, two, three, four, five, six, seven, eight, nine, ok, three2, C, thumb down, thumb right, thumb left, thumb backward, thumb forward) and label the ground truth for each class. Each gesture has around 250 testing samples. The results are shown in Table 4. (Thumb pointing direction is not evaluated on this dataset because there is no adequate gesture for testing. Only the ‘thumb down’ gesture has a clearly downward pointing direction, yet the images are not pointing strictly downward.)

On this first-view dataset, the error rate of the rules only increase from 2.3% to 6.3% (though the correct rate decrease

by around 19% because more ‘unsure’ are predicted), showing that our rules works across different views.

Table 4: Performance of Rules on EgoGesture Test Set

Rule	Performance on EgoGesture		
	error	unsure	correct
Flexion - thumb	0.046	0.532	0.422
Flexion - other fingers	0.060	0.218	0.722
Proximity	0.098	0.194	0.708
Contact	0.044	0.159	0.797
Pointing Direction - thumb	-	-	-
Palm Orientation	0.108	0.250	0.642
Overall	0.063 ± 0.043	0.216 ± 0.096	0.720 ± 0.123

More detailed error analysis of our rules’ performance on EgoGesture Dataset can be found in Table 5. Only those with error rate above 15% is shown.

Table 5: Analysis of Error Cases on EgoGesture Dataset

Rule	Gesture	Finger	Error Rate	Observed Reasons
Flexion - other fingers	seven	ring	0.171	MediaPipe's mistake for occluded fingers.
	C	ring	0.595	The finger in this gesture is slightly bent by nature, hard to predict precisely.
		pinky	0.360	Same as above.
	thumb down	index	0.177	MediaPipe's mistake for occluded fingers.
		ring	0.326	Same as above.
		ring	0.281	Same as above.
Proximity	C	middle-ring	0.255	The rule does not generalize very well.
	three	middle-ring	0.154	Same as above.
	four	index-middle	0.260	Landmarks mistake; some people perform it differently; the fingers are slightly separated by nature, hard to predict precisely, but it doesn't influence the recognition of the gesture very much.
		middle-ring	0.536	Same as above.
		ring-pinky	0.353	Same as above.
	five	index-middle	0.289	Same as above.
		middle-ring	0.767	Same as above.
		ring-pinky	0.488	Same as above.
	ok	middle-ring	0.404	Same as above.
		ring-pinky	0.578	Same as above.
	nine	index-middle	0.244	Landmarks mistake.
Contact	seven	thumb-ring	0.202	Some people perform it differently; MediaPipe's mistake for occluded fingers.
		thumb-pinky	0.151	Same as above.
	measure	thumb-index	0.198	Landmark mistake; in some gestures thumb and index finger are close so it is hard to discriminate.
	nine	thumb-index	0.191	Landmark mistake.
		thumb-pinky	0.244	MediaPipe's mistake for occluded fingers.
	thumb down	thumb-index	0.223	Landmark mistake.
	thumb back-ward	thumb-index	0.249	Same as above.
	thumb for-ward	thumb-index	0.186	Same as above.

A.2 Pseudocode for description rules

Algorithm 1 Flexion of a finger

```

1: procedure FLEXION(finger, thresholdLow, thresholdHigh)
2:   curl  $\leftarrow$  0
3:   if finger is thumb then
4:     curl  $\leftarrow$  Angle( $\overrightarrow{\text{MCP-PIP}}$ ,  $\overrightarrow{\text{IP-TIP}}$ )
5:   else
6:     curl  $\leftarrow$  Angle( $\overrightarrow{\text{MCP-PIP}}$ ,  $\overrightarrow{\text{PIP-DIP}}$ )
7:     curl  $\leftarrow$  curl + Angle( $\overrightarrow{\text{PIP-DIP}}$ ,  $\overrightarrow{\text{DIP-TIP}}$ )
8:   end if
9:   if curl  $\leq$  thresholdLow then
10:    return straight
11:   else if curl  $\geq$  thresholdHigh then
12:     return bent
13:   else
14:     return unsure
15:   end if
16: end procedure

```

Algorithm 2 Proximity of two fingers

```

1: procedure PROXIMITY(finger1, finger2, thresholdLow, thresholdHigh)
2:   jointDis  $\leftarrow$  0
3:   polylineF1  $\leftarrow$  Polyline(finger1 PIP, finger1 DIP, finger1 TIP)
4:   polylineF2  $\leftarrow$  Polyline(finger2 PIP, finger2 DIP, finger2 TIP)
5:   distance1  $\leftarrow$  Distance(finger1 PIP, polylineF2)
6:   distance2  $\leftarrow$  Distance(finger2 PIP, polylineF1)
7:   jointDis  $\leftarrow$  jointDis + min(distance1, distance2)
8:   distance3  $\leftarrow$  Distance(finger1 DIP, polylineF2)
9:   distance4  $\leftarrow$  Distance(finger2 TIP, polylineF1)
10:  jointDis  $\leftarrow$  jointDis + min(distance3, distance4)
11:  jointDis  $\leftarrow$  jointDis/3
12:  if jointDis is less than thresholdLow then
13:    return adjacent
14:  else if jointDis is greater than thresholdHigh then
15:    return separated
16:  else
17:    return unsure
18:  end if
19: end procedure

```

Algorithm 3 Contact of two fingers

```

1: procedure CONTACT(finger1, finger2, thresholdLow, thresholdHigh)
2:   distance  $\leftarrow$  Distance(finger1 TIP, finger2 TIP)
3:   if distance  $\leq$  thresholdLow then
4:     return contact
5:   else if distance  $\geq$  thresholdHigh then
6:     return notcontact
7:   else
8:     return unsure
9:   end if
10: end procedure

```

Algorithm 4 Thumb Pointing Direction

```

1: procedure CONTACT(thumb, threshold)
2:   if thumb is not straight then
3:     return unsure
4:   else
5:     Thumb  $\leftarrow$   $\overrightarrow{\text{MCP-TIP}}$ 
6:     minAngle  $\leftarrow$   $+\infty$ 
7:     ThumbDir  $\leftarrow$  None
8:     for dir in [down, up] do
9:       angle  $\leftarrow$  Angle(Thumb, dir)
10:      if angle is less than minAngle then
11:        minAngle  $\leftarrow$  angle
12:        ThumbDir  $\leftarrow$  dir
13:      end if
14:    end for
15:    if minAngle is greater than threshold then
16:      return unsure
17:    else
18:      return ThumbDir
19:    end if
20:  end if
21: end procedure

```

Algorithm 5 Palm orientation

```

1: procedure PALMORIENTATION(hand, threshold)
2:   PalmVec1  $\leftarrow$   $\overrightarrow{\text{pinkyMCP, indexMCP}}$ 
3:   PalmVec2  $\leftarrow$   $\overrightarrow{\text{wrist, middleMCP}}$ 
4:   if hand is left hand then
5:     PalmOriVec  $\leftarrow$  PalmVec1  $\times$  PalmVec2
6:   else if hand is right hand then
7:     PalmOriVec  $\leftarrow$  PalmVec2  $\times$  PalmVec1
8:   end if
9:   minAngle  $\leftarrow$   $+\infty$ 
10:  PalmOri  $\leftarrow$  None
11:  for dir in [right, left, down, up, outward, inward] do
12:    angle  $\leftarrow$  Angle(PalmOriVec, dir)
13:    if angle is less than minAngle then
14:      minAngle  $\leftarrow$  angle
15:      PalmOri  $\leftarrow$  dir
16:    end if
17:  end for
18:  if minAngle is greater than threshold then
19:    return unsure
20:  else
21:    return PalmOri
22:  end if
23: end procedure

```

A.3 Hand State Matrix Details

The matrix generated by the rule-based module is subsequently interpreted by *Gesture Description Agent*'s gesture summary description generation module, and the result is provided to the *Gesture Inference Agent* for further analysis. This matrix encapsulates hand pose and movement status across two distinct channels, each offering a different dimension of gesture representation. This pose-movement split is proven to promote Gesture Description Agent's performance, avoiding omitting important characteristics or overemphasizing certain aspect.

Channel 1: Hand Pose The first channel is a 2D array comprising 19 rows and T columns, where T denotes the number of time steps, with each step representing 0.2 seconds. The rows are indexed starting from 1 and detail the following aspects:

- Rows 1-5 correspond to finger flexion for the thumb, index, middle, ring, and pinky fingers, respectively. The values are encoded as 1 (straight), 0 (between straight and bent), and -1 (bent), describing the extent of finger flexion.
- Rows 6-8 represent finger proximity for adjacent finger pairs (index-middle, middle-ring, ring-pinky) with similar encoding scheme. The aim is to indicate how closely each finger is to its neighbor.
- Rows 9-12 detail thumb fingertip contact with the fingertips of the other fingers, again using similar value encoding.
- Row 13 specifies the pointing direction of the thumb, with 1 (upward), -1 (downward), and 0 (no specific direction or unknown when thumb is bent).
- Rows 14-19 are dedicated to palm orientation, indicating the direction the palm faces from the user's perspective. A specific orientation is marked by a single row set to 1 among these rows, representing left, right, down, up, inward, and outward directions. All rows equal to 0 means no specific direction can be identified.

Channel 2: Hand Movement The second channel consists of a 2D array with 2 or 3 rows (depending on whether we can extract hand position in 3d space or 2d space) and T columns. On each time step, the vector corresponds to the geometric center of the hand at this time:

- Row 1 tracks the horizontal position (0 for leftmost to 1 for rightmost), where increasing values suggest movement from left to right.
- Row 2 follows the vertical position (0 for bottommost to 1 for topmost), where increasing values suggest movement from down to up.

To gauge movement magnitude, the hand's width is also provided. For example, a hand width of 0.05 with a rightward movement of 0.05 in the array suggests a displacement of approximately one hand width.

This detailed matrix representation ensures a comprehensive and nuanced understanding of hand gestures, facilitating advanced processing and interpretation in gesture recognition systems.

Table 6: Tasks And External Context in the Smart Home Scenario

Instruction	External Context
Unlock the Smart Cabinet.	["It is 7:00 PM now.", "The child lock on this cabinet supports fingerprint unlocking."]
Increase the brightness of the light.	["It is 7:05 PM now."]
Show the next recipes on the smart screen.	["It is 7:12 PM now."]
Open the oven.	["It is 7:14 PM now.", "Recipe instructions: now you need to open the oven"]
Open the air cleaner.	["It is 7:20 PM now.", "The air purifier's sensor detected that the current environment has heavy cooking fumes."]
Set a timer on the smart screen.	["It is 7:30 PM now.", "Recipe instructions: now you need to cook on high heat for five minutes."]
Switch input source of the smart screen to the smart bell.	["It is 7:32 PM now.", "The doorbell is ringing."]
Make a phone call through the smart screen.	["It is 7:33 PM now.", "Just now, it was the deliveryman delivering goods; the owner of the goods is the user's roommate, Mark."]

B Detailed Experiment Setting

This section outlines the experimental setup for two experiments carried out within this research.

B.1 Experiment 1: Augmented Reality-Based Smart Home IoT Control

The first experiment focuses on controlling IoT devices within a smart home environment through augmented reality. The setup simulates a scenario where users interact with various home devices using gesture controls.

As illustrated in Figure 10, the experimental platform integrates an augmented reality interface for IoT device control within a home setting.

The experimental setup encompasses a variety of device functions and user tasks to mimic real-world interactions with a smart home environment. Details for this experiment are presented in Table 7 for device functions and Table 6 for tasks and external contexts. The smart home scenario encompasses a total of 18 functions across 5 devices, offering a comprehensive assessment of gesture-based control in an augmented reality context.

B.2 Experiment 2: Online Video Streaming on PC

The second experiment focuses on user interaction with online video content on a PC monitor. To ensure familiarity with the website's interface among participants, we selected a highly

popular video website for the experiment. Figure 11 illustrates the setup of our experimental platform.

Detailed descriptions of the tasks, including the number of functions and external context information for the video streaming environment, are provided in Table 8. Additionally, the list of functions available for tasks 4, 5, and 6 is presented in Table 9, while the function list for the remaining tasks is shown in Table 10. The variation in the function list is attributed to task 3, which involves full-screening the video page, resulting in a reduced number of available functions.

C System Cost for LLM Use

Our system employs the OpenAI API³ gpt-4-1106-preview for experiments, which is one of the best performance LLM and achieving near-human-level common sense and reasoning. The cost of each run is determined by the total token count.

For each gesture, GestureGPT consumes an average of 38785 tokens ($SD = 10432$) for input and an average of 3443 tokens ($SD = 1339$) for output, spanning 6.08 rounds of conversation ($SD = 0.85$). This results in a cost of \$0.389 per gesture.

Although costs are currently elevated due to the premium on model resources and extensive token requirements, we anticipate a reduction in expenses as LLM technology continues to evolve.

³<https://openai.com/pricing#language-models>



Figure 10: The experimental platform utilized in the smart home scenario. (a) IoT device control interface, simulated using Unity. (b) User wearing the Hololens and performing gestures with the right hand for device control.

Table 7: Device Functions In Smart Home Scenario(Totally 18 functions in 5 devices)

Device Name	Function Name
Light	On/Off
	Brightness control
	Mode Switch(Task Lighting/Morning Lighting/Accent Lighting)
Smart Cabinet	Child Lock activated/deactivated
	Temperature Control
	Humidity Control
Smart Screen	On/Off
	Switch Recipes(Recipe 1/Recipe 2/Recipe 3)
	Switch Input Source(Smart Screen/Smart Doorbell/Ipad Video)
	Phone Call
	Settable Timer
Oven	On/Off
	Temperature Control
	Self cleaning on/off
	Mode Switch(Bake Mode/Convection Roast/Bottom Heat Only/Keep Warm/Energy Efficiency)
Air cleaner	On/Off
	Airflow speed control
	Mode Switch(Strong/Silent/Custom)



Figure 11: The experimental platform utilized in the video streaming scenario. (a) and (b) are the video interface and the function distribution, (c) a user was watching the video and do gesture to control.

Table 10: Video Scenario Function List in Task 1,2,3,7,8

ID	Name	ID	Name
0	Redirect to Bilibili Homepage	33	SendButton
1	AnimeSectionLink	34	LikeTheVideo Button
2	LiveBroadcastEntry	35	CoinTipping
3	GameCenterEntry	36	AddToFavorites
4	MemberPurchaseLink	37	Video Share Link Retrieval
5	BilibiliMangaAccess	38	UserProfileLink
6	Match Entry Link	39	GetUserProfile
7	DownloadApp	40	SendMessage
8	SearchButton	41	Follow Button
9	imgOnError	42	AdReportSync
10	OpenVipPage	43	SwitchToggle
11	MessageNotificationLink	44	MarkVideoAsWatched
12	OpenDynamicPage	45	NavigateToEpisode
13	AddToFavorites (first)	46	VideoSelection
14	ViewHistory	47	EpisodeSelection (first)
15	CreativeCenterLink	48	EpisodeNavigation
16	UploadVideoEntry	49	EpisodeSelection (second)
17	VideoProgressControl	50	VideoNavigationLink
18	PlayPauseToggle	51	VideoSeriesNavigation
19	NextButton	52	PlayVideo
20	DisplayPlaybackTime	53	VideoNavigation
21	ChangeVideoQuality	54	OpenVideoLecture
22	ChangePlaybackRate	55	bmgOnLoad (first)
23	SubtitleSettings	56	VideoLink
24	AdjustVolume	57	ProfileLink (first)
25	Video Settings	58	bmgOnLoad (second)
26	PictureInPictureToggle	59	DocumentarySeriesChina
27	ToggleWideScreenMode	60	ProfileLink (second)
28	ToggleWebFullScreen	61	bmgOnLoad (third)
29	ToggleFullscreen	62	VideoPreviewLink
30	DanmakuSwitch	63	DisplayUserProfileLink
31	DanmakuToggleButton	64	VideoLinkNavigation
32	DanmuEtiquetteHint	65	VideoPlayArea

Table 8: Tasks, Function Numbers, and External Context in the Video Scenario

Instruction	Function Numbers	External Context
Turn up the volume.	66	[It is 8:01 PM now.]
Drag the progress bar forward.	66	[It is 8:02 PM now., The user has watched the earlier part of this video.]
Enter full screen mode.	66	[It is 8:04 PM now.]
Pause the video.	17	[It is 8:15 PM now., Right now, the user's phone is actively receiving an incoming call.]
Resume the video.	17	[It is 8:17 PM now., The user hung up the phone]
Exit full screen mode.	17	[It is 8:48 PM now.]
Like the video.	66	[It is 8:49 PM now.]
Go to the next episode.	66	[It is 8:50 PM now.]

Table 9: Video Scenario Function List in Task 4,5,6

ID	Name	ID	Name
0	VideoProgressBarUpdate	9	SelectEpisode
1	PlayPauseButton	10	ChangePlaybackSpeed
2	NextButton	11	SubtitleControl
3	SeekTimeUpdate	12	VolumeControl
4	ToggleDanmakuDisplay	13	VideoSettingsMenu
5	DanmakuToggle	14	PictureInPictureToggle
6	DanmuEtiquetteHint	15	ToggleFullscreen
7	SendMessageButton	16	VideoPlayArea
8	VideoQualitySelection		