



Gestura: A LVLM-powered System Bridging Motion and Semantics for Real-time Free-form Gesture Understanding

ZHUOMING LI*, Institute of Artificial Intelligence (TeleAI) of China Telecom, China

AITONG LIU*, Institute of Artificial Intelligence (TeleAI) of China Telecom, China

MENGXI JIA[†], Institute of Artificial Intelligence (TeleAI) of China Telecom, China

YUBO LU, Institute of Artificial Intelligence (TeleAI) of China Telecom, China

TENGXIANG ZHANG, Goertek Inc., China

CHANGZHI SUN, Institute of Artificial Intelligence (TeleAI) of China Telecom, China

DELL ZHANG[#], Institute of Artificial Intelligence (TeleAI) of China Telecom, China

XUELONG LI[†], Institute of Artificial Intelligence (TeleAI) of China Telecom, China

Free-form gesture understanding is highly appealing for human-computer interaction, as it liberates users from the constraints of predefined gesture categories. However, the sole existing solution—GestureGPT—suffers from limited recognition accuracy and slow response times. In this paper, we propose Gestura, an end-to-end system for free-form gesture understanding. Gestura harnesses a pre-trained Large Vision-Language Model (LVLM) to align the highly dynamic and diverse patterns of free-form gestures with high-level semantic concepts. To better capture subtle hand movements across different styles, we introduce a Landmark Processing Module that compensates for LVLMs' lack of fine-grained domain knowledge by embedding anatomical hand priors. Further, a Chain-of-Thought (CoT) reasoning strategy enables step-by-step semantic inference, transforming shallow knowledge into deep semantic understanding and significantly enhancing the model's ability to interpret ambiguous or unconventional gestures. Together, these components allow Gestura to achieve robust and adaptable free-form gesture comprehension. Additionally, we have developed the first open-source dataset for free-form gesture intention reasoning and understanding with over 300,000 annotated QA pairs. Experimental results show that Gestura achieves the accuracy of 84.73% (closed-set) / 64.14% (open-set) in the exocentric (third-person) setting and 66.14% (closed-set) / 21.71% (open-set) in the egocentric (first-person) setting, achieving approximately **20% and 40% higher accuracy** on closed-set and open-set tasks, respectively, compared to GestureGPT. Moreover, Gestura achieves over a **100× speedup** in response time (1.6 seconds vs. 227 seconds) on an 8B-sized model deployed on a single NVIDIA A100 40GB GPU, and has been validated through real-device experiments with an edge-cloud collaborative setup, bringing free-form gesture understanding markedly closer to practical, real-world deployment. Both the dataset and code about the project can be accessed at <https://evans-lx.github.io/Gestura/>.

*Project Lead.

[†]Both authors contributed equally to this research.

[#]Corresponding authors.

Authors' Contact Information: **Zhuoming Li**, Institute of Artificial Intelligence (TeleAI) of China Telecom, China, lx516789442@gmail.com; **Aitong Liu**, Institute of Artificial Intelligence (TeleAI) of China Telecom, China, liuaitong@sjtu.edu.cn; **Mengxi Jia**, Institute of Artificial Intelligence (TeleAI) of China Telecom, Shanghai, China, jiamx1@chinatelecom.cn; **Yubo Lu**, Institute of Artificial Intelligence (TeleAI) of China Telecom, Shanghai, China, ggu46261@gmail.com; **Tengxiang Zhang**, Goertek Inc., Beijing, China, zhangtengxiang@ict.ac.cn; **Changzhi Sun**, Institute of Artificial Intelligence (TeleAI) of China Telecom, Shanghai, China, czsun@chinatelecom.cn; **Dell Zhang[#]**, Institute of Artificial Intelligence (TeleAI) of China Telecom, Shanghai, China, dell.z@ieee.org; **Xuelong Li**, Institute of Artificial Intelligence (TeleAI) of China Telecom, Shanghai, China, xuelong_li@ieee.org.



This work is licensed under a Creative Commons Attribution 4.0 International License.

© 2025 Copyright held by the owner/author(s).

ACM 2474-9567/2025/12-ART193

<https://doi.org/10.1145/3770709>

CCS Concepts: • **Human-centered computing** → **Ubiquitous and mobile computing systems and tools**; • **Information systems** → *Language models*.

Additional Key Words and Phrases: Free-form Gesture, Gesture Understanding, Gesture Intention, Large Vision Language Model, Multi-Modal

ACM Reference Format:

Zhuoming Li, Aitong Liu, Mengxi Jia, Yubo Lu, Tengxiang Zhang, Changzhi Sun, Dell Zhang[#], and Xuelong Li. 2025. Gestura: A LVLM-powered System Bridging Motion and Semantics for Real-time Free-form Gesture Understanding. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 9, 4, Article 193 (December 2025), 29 pages. <https://doi.org/10.1145/3770709>

1 INTRODUCTION

Gesture-based interaction serves as a critical interface for wearable devices such as smart glasses [7, 33, 45], enabling intuitive human-device communication. However, real-world deployment faces a significant challenge: spontaneous gesture understanding in open-world environments. As these devices become more pervasive in daily life, users expect to interact naturally—using spontaneous, personalized gestures instead of memorizing predefined commands. This shift demands gesture recognition systems that are not only accurate but also adaptable to individual styles, cultural variations, and situational context. Unlike controlled settings, users exhibit diverse, context-driven gestures that carry implicit intentions, demanding systems capable of parsing both motion patterns and latent semantics. Traditional approaches relying on predefined gesture libraries or specialized hardware (e.g., gloves, depth sensors) fail to address this complexity, creating barriers to natural interaction and scalability.

Current gesture recognition methods fall into two categories. *i.e.*, 1) Conventional deep learning methods (e.g., CNNs, LSTMs) excel at classifying fixed gesture vocabularies, but lack semantic reasoning to infer intentions from novel motions. 2) Emerging Large Language Models (LLMs) such as GestureGPT [44], though attempting open-world interpretation, face dual constraints: Firstly, they depend on intensive computational resources and introduce significant latency, which hinders real-world deployment; more critically, existing LLMs lack understanding of fine-grained hand dynamics, which fundamentally limits their capacity to capture nuanced gestural semantics. This presents a crucial research challenge: how to adaptively integrate fine-grained motion cues with general reasoning capabilities to effectively bridge low-level motion tracking and high-level intent inference. This integration constitutes a critical pathway for free-form gesture understanding.

To address this challenge, we turn to Large Vision-Language Models (LVLMs)—an emerging extension of LLMs that incorporate both visual and textual modalities. While early LLMs were limited to text-only inputs, recent advances have enabled them to process images and videos through aligned visual encoders and MLP projector. These LVLMs retain the strong reasoning capabilities of LLMs, while gaining the capacity to perceive complex visual contexts.

This paper explores a novel synergy: leveraging robust hand landmark priors to "teach" a LVLM to resolve ambiguities in free-form gesture understanding. Our key insight is that while traditional hand keypoint detection models offer geometrically accurate but semantically shallow representations, they can provide crucial contextual information to LVLM that lack domain-specific knowledge but possess strong reasoning abilities. Our framework enables kinematic precision to inform contextual semantics through innovative cross-modal interaction.

Specifically, we propose a hierarchical framework centered around **two tightly integrated components**: (1) a Landmark Processing Module, which incorporates hand keypoint information extracted from MediaPipe [24] to enrich visual features with anatomical structure and spatial cues—crucial for distinguishing subtle gesture variations; (2) a LVLM backbone that employs a dual-stream encoder and a large language model to align dynamic visual features with high-level semantic concepts, enabling robust intent inference beyond superficial motion patterns; and a **two-stage training paradigm**: First, Gestura learns general visual-semantic mappings through

a multi-view semantic enhancement strategy to activate the model’s potential for free-form generalization. Subsequently, in Stage 2, Gestura leverages a well-trained landmark processing module to transmit anatomical and spatial contextual signals to the LVLM backbone. Meanwhile, it internalizes advanced reasoning capabilities through Chain-of-Thought (CoT) tuning. This enables Gestura to push the model’s reasoning limits in open-world scenarios and achieve superior generalization.

To evaluate the effectiveness of our model, we conducted comprehensive experiments across both benchmark and real-world scenarios. Specifically, Gestura achieves a top-1 intent accuracy of 84.73% in the exocentric view and 66.14% in the egocentric view, substantially outperforming the previous state-of-the-art, GestureGPT, which achieves 72.08% (closed-set) / 44.46% (open-set) and 40.07% (closed-set) / 17.38% (open-set) under the corresponding settings. This highlights a significant advancement in accurate gesture intent interpretation, particularly under the challenging egocentric perspective. Moreover, on gesture description tasks, Gestura also demonstrates a substantial improvement: achieving BLEU-4 scores of 49.83 (exocentric, closed-set) and 43.67 (egocentric, closed-set), compared to GestureGPT’s 15.05 and 10.75, respectively—representing over 3× higher description quality. Even under open-set conditions, Gestura maintains strong performance, with BLEU-4 scores of 14.17 (exocentric) and 9.93 (egocentric), further validating its robustness in free-form gesture understanding. The relatively lower open-set accuracy in the egocentric setting can be attributed to the smaller scale of egocentric data compared to exocentric data in our dataset, which limits the model’s ability to generalize to novel gestures from first-person perspectives. Beyond benchmarks, we deployed Gestura in practical egocentric smart home settings, where it reached a top-5 accuracy of 69.23% across eight free-form intent categories, proving its robustness to gesture variability and its suitability for real-time, context-aware human-computer interaction. Notably, the average response time in real-device experiment settings is 7.83 seconds, including communication and TTS latency, while pure model inference takes only 1.6 seconds—significantly faster than prior models like GestureGPT (227 seconds)—making Gestura highly efficient for interactive applications.

Our contributions can be summarized as follows:

- (1) **A novel dataset for free-form gesture intent understanding:** We present a pioneering dataset that redefines gesture data through a question-answering (QA) paradigm, specifically designed for free-form, user-defined gestures. Unlike traditional datasets constrained by fixed vocabularies, ours captures the diversity of real-world gestures by including rich annotations at three levels: *action description*, *gesture meaning*, and *contextual intent*. This enables robust benchmarking for intent reasoning under open-world conditions.
- (2) **A LVLM-based inference architecture tailored for free-form gesture comprehension:** We extend a LVLM to support fine-grained gesture reasoning by aligning motion patterns with latent semantics. Our architecture bridges low-level visual encoding and high-level intention inference, allowing it to generalize across unseen gestures. Crucially, our system achieves this while maintaining real-time performance (7.83 s per gesture), making it possible for deployment in wearable and edge-computing scenarios.
- (3) **Semantic grounding through anatomical priors:** To better differentiate between various hand gestures, we utilize Mediapipe as an auxiliary tool by integrating the gesture ground truth detected by Mediapipe with video data. This approach proves beneficial in distinguishing similar gestures, particularly those with visually similar features. In essence, our idea is to leverage the prior knowledge of a specialized gesture recognition model to enhance the understanding of gestures by our language model (LM). This fusion enables the LM to interpret gestures more accurately, facilitating the model to better distinguish fine-grained differences between different gestures.

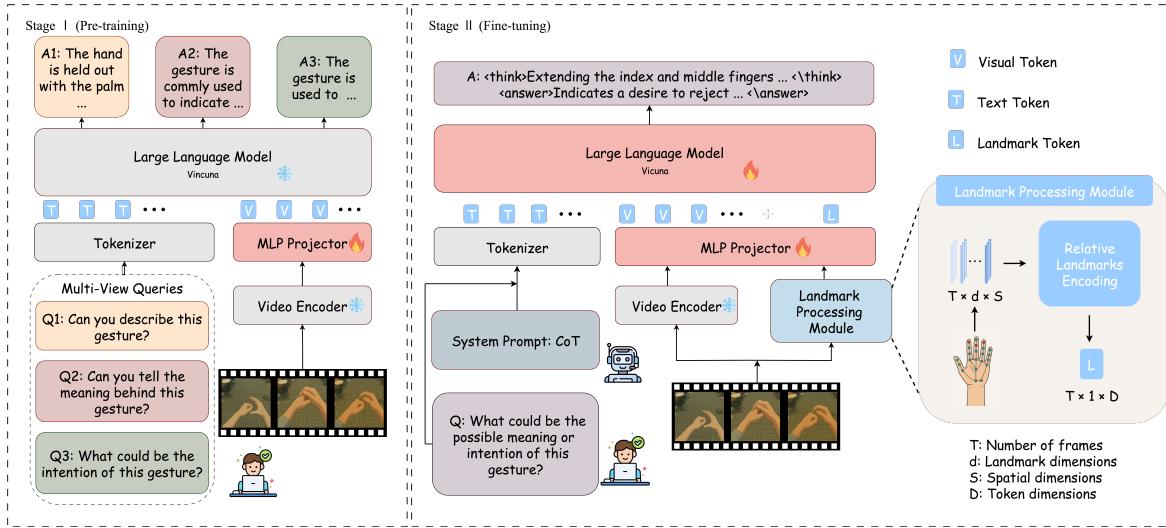


Fig. 1. Overview of the proposed framework of Gestura. Gestura introduces a hierarchical framework with two-phase training for free-form gesture understanding. First, the pre-training stage activates the model’s potential for free-form generalization by using a multi-view semantic enhancement strategy. In Stage 2, Gestura leverages well-trained landmark processing module to deliver anatomical and spatial contextual signals to the LViLM backbone, while internalizing advanced reasoning capabilities through Chain-of-Thought (CoT) tuning. This dual mechanism expands the model’s reasoning boundaries, enabling superior generalization in free-form gesture understanding.

2 RELATED WORK

Gesture recognition has garnered significant attention in both human-computer interaction (HCI) and human-robot interaction (HRI) domains [5, 32]. Recent advances in hand gesture recognition, particularly in dynamic settings, have leveraged different techniques and data modalities such as RGB images, depth sensors, and 3D skeleton-based data [20, 28]. The use of 3D hand skeleton data, particularly hand landmarks, has become an important aspect of gesture recognition due to its robustness and ability to capture rich spatial-temporal features. Many systems rely on spatial-temporal attention mechanisms to model dependencies between hand joints, making them effective for real-time recognition tasks.

In recent years, significant progress has been made in dynamic hand gesture recognition (DHGR) using deep learning models such as convolutional neural networks (CNNs) [2, 39], graph-based neural networks, and long short-term memory (LSTM) networks [34]. For instance, graph-based architectures have been proposed to better handle the spatial and temporal dependencies between joints for improved performance on skeleton-based hand gesture datasets, such as SHREC’17 [35], achieving high accuracy in recognizing complex gestures. These approaches focus on extracting spatial-temporal attention features to capture joint dependencies, allowing for real-time and efficient gesture recognition.

More recently, the emergence of large-scale pre-trained models—particularly transformer-based architectures has further propelled the development of HCI and gesture recognition systems [15, 16, 38, 43]. These models, originally popularized in natural language processing, have demonstrated remarkable capacity in capturing long-range dependencies and multi-level abstractions, making them highly suitable for tasks involving sequential and multi-dimensional inputs, such as gesture sequences. Vision Transformers [10], for example, have been

adapted to process RGB frames or skeletal heatmaps [21], showing competitive performance in modeling global gesture contexts.

At the same time, the growing integration of wearable devices and ubiquitous computing technologies has significantly enhanced gesture recognition capabilities [13, 37, 41]. Wearables such as smart glasses, motion sensors, and inertial measurement units (IMUs) provide continuous, fine-grained motion data that complements traditional visual modalities [14]. These devices enable gesture recognition systems to operate in unconstrained environments with improved responsiveness and contextual awareness. When combined with transformer-based models, the rich multimodal signals from wearable platforms can be effectively aligned and interpreted, paving the way for more natural, adaptive, and personalized interaction experiences across a wide range of real-world applications.

However, while significant advancements have been made in recognizing predefined gestures, most existing systems require users to learn and perform gestures from a fixed set, which can result in a less natural interaction experience. Recent work, such as GestureGPT, proposes a solution for free-form gesture recognition by enabling automatic understanding of spontaneous gestures. This approach eliminates the need for users to learn predefined gestures, allowing for intuitive interaction without explicit training. GestureGPT utilizes LLM to interpret hand gestures from natural language descriptions, mapping them directly to interface functions, thus overcoming the rigid nature of traditional gesture recognition systems. The framework achieves strong performance in zero-shot gesture recognition.

Moreover, a recent exploration into Zero-Shot Learning (ZSL) for dynamic hand gesture recognition proposes a novel multi-modal ZSL framework [34]. This method combines deep learning based features with skeleton-based representations to perform gesture recognition without annotated data. By leveraging transformer models and BERT-based semantic mapping, the system achieves significant performance improvements in recognizing hand gestures without requiring specific gesture annotations. This approach is particularly relevant when handling open-world tasks where the gesture set is not fixed, and new gesture categories need to be recognized on the fly.

In the realm of open-world gesture recognition, methods have also been proposed to tackle the challenges of catastrophic forgetting and incremental learning [36]. The work on Data-Free Class-Incremental Learning [1] introduces a method for dynamic hand gesture recognition that doesn't require access to previously seen data, making it suitable for real-world applications where privacy constraints prevent data storage. By employing a boundary-aware prototypical sampling mechanism, the approach improves model inversion and recognition performance without compromising on accuracy or computational efficiency. This method demonstrates its effectiveness in 3D skeleton gesture recognition by addressing both the challenges of incrementally adding new gestures and avoiding performance degradation on old gestures.

Additionally, the synthesis of stylized gestures for human-robot interaction has gained attention, with systems like GestureDiffuCLIP [3] offering a flexible control mechanism for generating gestures with varying levels of style. By using the CLIP model for style conditioning and a latent diffusion model for gesture generation, this framework allows for high-quality stylized gestures tailored to different interactive scenarios. While this research focuses on synthesizing co-speech gestures, its underlying mechanism for gesture generation and style control could be beneficial in the context of interactive gesture recognition systems, where natural and intuitive user interactions are crucial.

Together, these studies contribute to the growing body of work on dynamic and free-form gesture recognition. They highlight the challenges in achieving real-time performance, maintaining model generalizability, and adapting to open-world environments [12]. Our work contributes a unified framework that leverages structured hand landmark priors to enhance the semantic reasoning capabilities of LVLM, achieving fine-grained and interpretable recognition of free-form gestures in open-world scenarios through hierarchical alignment and progressive training.

3 APPROACH

3.1 Overview

Our framework introduces a hierarchical approach for gesture understanding, combining landmark-enhanced visual processing, cross-modal alignment, and progressive training strategies to bridge gesture dynamics with semantic intent.

Landmark Processing Module augments raw video features with structural hand keypoints extracted via MediaPipe. By encoding spatial relationships between 21 hand landmarks, it enriches visual representations with geometric cues, enabling finer distinction between subtle gesture variations.

LVLM integrates a dual-stream video encoder (capturing static poses and dynamic motions) with a MLP projector that maps visual features into a shared language space. A LLM then understand these features, leveraging its reasoning capabilities to infer gesture meanings beyond surface-level motions.

Training Pipeline follows a two-stage paradigm:

Stage 1. In the initial pre-training stage, we freeze the parameters of the video encoder and the LLM and introduce a trainable MLP projector that serves as a bridge, aligning video features with textual captions. The reason of freezing the parameters of the video encoder is to preserve its spatiotemporal pattern recognition capabilities. This design is motivated by a deep understanding of gestural motion regularities—such as the rhythm of waving or the trajectory of finger flexion—which are fundamental and transferable across tasks. Direct fine-tuning could risk disrupting these fine-grained motion representations. Holding the LLM weights fixed prevent premature drift, so the model first learns a clean alignment between vision tokens and the LLM’s embedding space. This stabilises autoregressive convergence and lays a solid cross-modal foundation[19, 23]. The relevant ablation experiments could be seen in Appendix B.

Specifically, the MLP is trained to map different semantic expressions of the same gesture sample into a shared embedding space while pushing apart unrelated gestures. To enable such fine-grained alignment, we adopt a multi-view semantic enhancement strategy: (1) Motion Description (e.g., “palm extended”), (2) Gesture Semantics (e.g., “inviting”), and (3) Intent Inference (e.g., “beckon”). This hierarchical structure guides the model in learning layered representations and decoupling low-level motion patterns from high-level semantics.

Stage 2. Once the model has acquired a baseline capacity for semantic alignment, the second stage focuses on resolving ambiguity between gestures with similar meanings or similar appearances in open-world scenarios. Here, MediaPipe landmark features are introduced and integrated via an anatomically constrained attention mechanism. This dynamic feature selection allows the model to distinguish between confusable gestures based on subtle joint angle variations and coherent motions.

Furthermore, the incorporation of Chain-of-Thought (CoT) tuning breaks the limitations of traditional end-to-end models by explicitly decomposing the reasoning process into a structured sequence: gesture description → semantic analogy → intent hypothesis → final decision. This enforces logical traceability and encourages the model to form robust associations rather than overfitting to surface-level patterns. As a result, the model achieves improved accuracy in open-ended scenarios and gains interpretability through intermediate reasoning steps.

This two-stage framework mirrors the biological plausibility of human cognition. Just as humans build a repertoire of motion patterns through repeated observation (analogous to Stage 1: pattern recognition), and then infer intent by incorporating contextual cues (analogous to Stage 2: semantic reasoning), our system enables a qualitative leap from motion decoding to cognitive-level interpretation. In this process, MediaPipe landmarks act as a form of “proprioception”—providing the model with physical constraints akin to the human sense of joint position—while CoT-based reasoning simulates expert-like inference grounded in domain knowledge (e.g., anatomical constraints on elbow movement), enabling more accurate intent disambiguation.

3.2 Landmark Processing Module

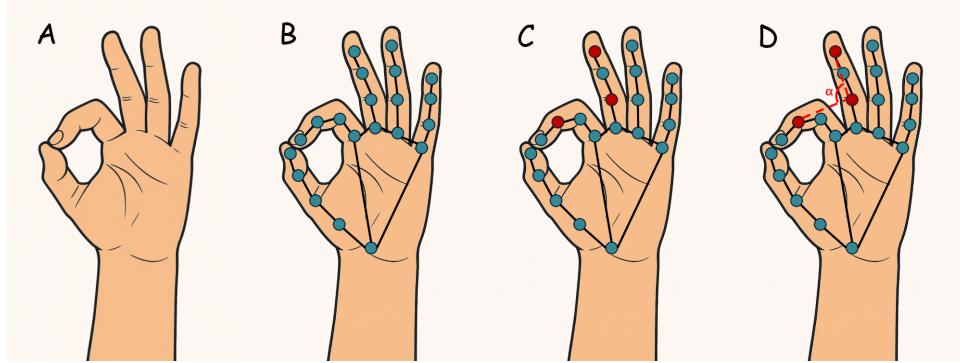


Fig. 2. The process of deriving encoded features from hand video data. [R1, R4] A. Raw RGB frame → B. 21 MediaPipe landmarks → C. One landmark triplet (red) case from the 1,330 possible → D. Measure the enclosed angle. Blue circles mark the 21 hand landmarks detected by MediaPipe. One triplet of landmarks (red) is chosen out of the $\binom{21}{3} = 1,330$ possible combinations each time to form two vectors whose enclosed angle α is measured. The cosine of this angle constitutes a single element of the final feature vector.

Considering that merely encoding videos through a video encoder is insufficient for the LLM to effectively distinguish the subtle differences in vision features between similar gestures, we have added ground information extracted by Mediapipe to the original vision features. The ground information here refers to the 21 key points of the hand extracted by Mediapipe, with each key point represented by its x, y, and z spatial positions.

For each gesture frame, n keypoints p_1, p_2, \dots, p_n are extracted, where $p_i = (x_i, y_i, z_i)$. Derived features include pairwise distances:

$$d_{ij} = \|p_i - p_j\|_2 = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2}$$

and angles between keypoints:

$$\theta_{ijk} = \cos^{-1} \left(\frac{(p_j - p_i) \cdot (p_k - p_i)}{d_{ij} d_{ik}} \right)$$

which is treated as the ground truth feature.

To prevent the scale differences between videos from affecting the ground information, we chose to use the cosine values formed by every three points as the final ground feature. Among the 1,330 possible combinations of selecting three points from the 21 key points, we used only 1,024 combinations for feature extraction to facilitate subsequent concatenation with vision features. This is because selecting any three points from the hand keypoints results in a total of 1,330 combinations. Considering that the hidden dimension of the vision token is 1,024, and for ease of concatenation, we chose the first 1,024 combinations to compute the cosine similarity values, which are used as the hidden dimension for the landmark tokens. To avoid distortion of the encoded features, we refrained from selecting all combinations and mapping them to 1,024. Since 1,024 combinations cover nearly all possibilities, we selected the first 1,024 combinations for the hands in all frames, ensuring that the features remain intact and consistent.

After adding the ground information, the number of feature tokens corresponding to each frame increased from 257 to 258, enriching the feature representation of gesture videos. The more comprehensive feature representation enables the model to better distinguish fine-grained differences between different gestures.

3.3 Large Vision Language Model

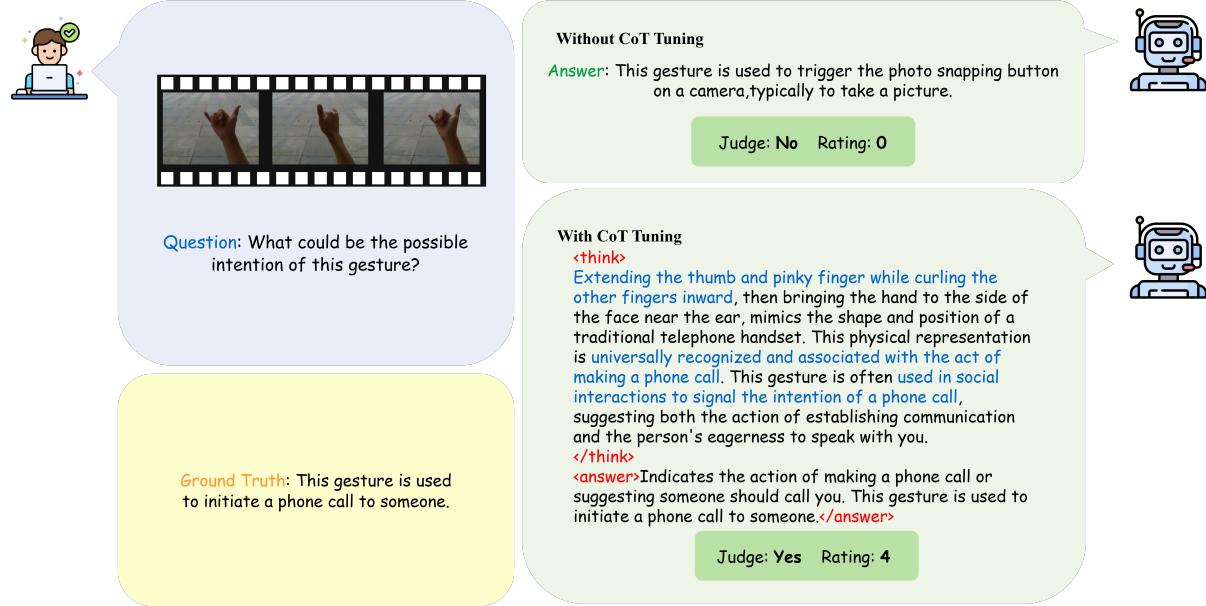


Fig. 3. Score rating and comparison of answers with and without CoT tuning. Certain portions of the generated output are highlighted to underscore their critical role in the inference process leading to the final answer.

To align gesture videos with their textual descriptions, we employ a video encoder that captures both static hand poses and dynamic motion patterns, ensuring comprehensive spatial-temporal representations. These gesture features are then projected into a shared semantic space using an MLP projector. In parallel, gesture-related text is tokenized into structured tokens. Both visual and textual embeddings are fed into an LLM, leveraging its reasoning and contextual understanding to bridge motion and semantics—allowing the model to infer gesture intent beyond surface-level motion.

3.4 Training Pipeline

3.4.1 STAGE 1: Multi-view Semantic Enhancement. In the pre-training stage, we align gesture video features with textual representations in a unified feature space. The video encoder extracts both static spatial features and dynamic motion patterns from gesture videos, which are then transformed by a trainable MLP projector. The video encoder remains frozen, ensuring that its pre-trained weights are preserved, while the MLP projector adapts the video features optimally.

The raw video feature vector from the frozen video encoder is denoted as $\mathbf{v} \in \mathbb{R}^{d_v}$. The MLP projector transforms this into the unified feature space as:

$$\mathbf{z} = \mathbf{W}_3 \cdot \text{GELU} \left(\mathbf{W}_2 \cdot \text{GELU} \left(\mathbf{W}_1 \mathbf{v} + \mathbf{b}_1 \right) + \mathbf{b}_2 \right) + \mathbf{b}_3 \quad (1)$$

where:

- $\mathbf{v} \in \mathbb{R}^{d_v}$: Input video features
- $\mathbf{W}_1 \in \mathbb{R}^{d_h \times d_v}, \mathbf{b}_1 \in \mathbb{R}^{d_h}$: First layer parameters

- $W_2 \in \mathbb{R}^{d_h \times d_h}$, $\mathbf{b}_2 \in \mathbb{R}^{d_h}$: Hidden layer parameters
- $W_3 \in \mathbb{R}^{d_z \times d_h}$, $\mathbf{b}_3 \in \mathbb{R}^{d_z}$: Output layer parameters
- GELU: Gaussian Error Linear Unit activation
- $\mathbf{z} \in \mathbb{R}^{d_z}$: Projected output features

To improve semantic alignment, we employ a multi-view semantic enhancement strategy, analyzing gestures from three distinct perspectives before passing them into the MLP projector. This process refines the representation by capturing nuanced aspects of gesture execution. The resulting video tokens, together with text tokens obtained via a tokenizer, are then fed into a frozen LLM. By maintaining a shared latent space and freezing the LLM parameters, we preserve its strong language reasoning capabilities while enabling the MLP projector to bridge the modality gap between vision and text. Stage 1 relies on a curated subset of approximately 400 000 question–answer pairs paired with 160 000 videos to establish robust descriptive and semantic grounding

The resulting video tokens, combined with text tokens from a tokenizer, are input into a frozen LLM. Freezing the LLM preserves its language reasoning capabilities, while the MLP projector bridges the vision–text modality gap.

Algorithm 1 Multi-view Semantic Enhancement (Pre-training)

Require: video dataset \mathcal{V} with three dimension annotation, pretrained VideoEncoder f_{video} , LLM, epochs E_1 , batch size B

Ensure: trained MLP projector P_{proj}

- 1: initialize P_{proj} parameters $\{W_1, b_1, W_2, b_2, W_3, b_3\}$
- 2: freeze VideoEncoder, freeze LLM
- 3: **for** $e = 1$ to E_1 **do**
- 4: **for** each batch $\{v_i\}_{i=1}^B$ from \mathcal{V}_1 **do**
- 5: $v_{\text{raw}} \leftarrow f_{\text{video}}(v_i)$
- 6: views \leftarrow generate_three_views(v_i)
- 7: $z_{\text{views}} \leftarrow \emptyset$
- 8: **for** each $view$ in views **do**
- 9: $v_{\text{view}} \leftarrow f_{\text{video}}(view)$
- 10: $h_1 \leftarrow \text{GELU}(W_1 v_{\text{view}} + b_1)$
- 11: $h_2 \leftarrow \text{GELU}(W_2 h_1 + b_2)$
- 12: $z_{\text{view}} \leftarrow W_3 h_2 + b_3$
- 13: append z_{view} to z_{views}
- 14: **end for**
- 15: $z_{\text{video}} \leftarrow \text{combine}(z_{\text{views}})$
- 16: tokens_text \leftarrow LLM.tokenize(t_i)
- 17: tokens_video \leftarrow map_to_embeddings(z_{video})
- 18: logits \leftarrow LLM.forward(tokens_video || tokens_text)
- 19: loss \leftarrow AlignmentLoss(logits, tokens_text)
- 20: loss.backward()
- 21: optimizer.step(); optimizer.zero_grad()
- 22: **end for**
- 23: **end for**
- 24: **return** P_{proj}

3.4.2 STAGE 2: Cross-modal Interaction with CoT Tuning. During pre-training with individually separated questions, we observed strong performance on closed-world test data but significant drops in accuracy under open-world scenarios. To address this issue, we aim to equip the model with the ability to reason from gesture descriptions to their intended meanings. We modified the system prompt to encourage causal reasoning across gesture, context, and interpretation while fine-tuned the model using the CoT-formed data to cultivate a habit of chain-of-thought style responses similarly to the cold-start phase of DeepSeek-R1[9]. Furthermore, we incorporated hand landmark information to improve the model’s ability to distinguish between different gestures.

With these modifications, the increased accuracy of open-world test scenarios demonstrating significant improvements in generalization and reasoning capabilities. The training details of the second stage are as follows:

Let $H(t)$ be the hand landmark coordinates at time t extracted by MediaPipe, $V(t)$ be the raw video data at time t , and P_θ be the projector network. The ground truth feature encoding is then computed as:

$$G(t) = f_{\text{rle}}(f_{\text{mp}}(V(t))) \quad (2)$$

where:

- $f_{\text{mp}}(\cdot)$ is the MediaPipe Landmark Extraction function
- $f_{\text{rle}}(\cdot)$ is the Relative Landmarks Encoding function

$$z(t) = P_\theta([G(t); f_{\text{video}}(V(t))]) \quad (3)$$

where:

- $f_{\text{video}}(\cdot)$ extracts deep video features
- $[\cdot; \cdot]$ denotes feature concatenation
- P_θ is the multi-layer projector network with GELU activations

In this stage, the system simultaneously processes two input streams: (1) ground truth features containing annotated gesture information, and (2) raw video features extracted from the visual encoder. These features are combined and transformed through the multi-layer projector to produce modality-aligned embeddings.

In addition to feature augmentation, the fine-tuning stage also modifies the training paradigm by structuring the data used in stage 1 into a Chain-of-Thought (CoT) format to facilitate step-by-step reasoning over gesture intention, yielding a curated set of 110,000 examples. This transformation systematically links gesture video features, fine-grained keypoint information, and the underlying intent in a step-by-step reasoning framework. By integrating structured reasoning chains, the model is encouraged to progressively infer gesture intent rather than relying on surface-level pattern recognition. This procedure is similar to process supervision[11], which further endows the model with enhanced reasoning capabilities.

Furthermore, the system prompt used in the large language model (LLM) is carefully redesigned to stimulate its reasoning capabilities[42], guiding it to establish logical connections between gesture actions, contextual cues, and inferred meanings. This refined approach not only improves accuracy in distinguishing similar gestures but also enhances the model’s interpretability, making it better suited for real-world applications where precise gesture understanding is crucial.

4 EXPERIMENT

4.1 Training Details

In our work, we adopt the Vicuna-7b v1.5 model as the language backbone. Vicuna is an instruction-following large language model fine-tuned from Llama 2 using user-shared conversations collected from ShareGPT. In the training process, we uniformly sample 8 frames from each video, and each frame undergoes preprocessing where

Algorithm 2 Cross-modal Interaction with CoT Tuning (Fine-tuning)

Require: video dataset \mathcal{V} with CoT, pretrained VideoEncoder f_{video} , LLM, epochs E_2 , batch size B
Ensure: trained ground-MLP projector P_{ground}

```

1: initialize  $P_{\text{ground}}$  parameters  $\theta$ 
2: freeze VideoEncoder, unfreeze LLM
3:  $mp\_extractor \leftarrow \text{MediaPipeLandmarkExtractor}()$ 
4:  $rle\_encoder \leftarrow \text{RelativeLandmarksEncoder}()$ 
5: for  $e = 1$  to  $E_2$  do
6:   for each batch  $\{v_i\}_{i=1}^B$  from  $\mathcal{V}_2$  do
7:      $H_i \leftarrow mp\_extractor.extract(v_i)$ 
8:      $G_i \leftarrow rle\_encoder.encode(H_i)$ 
9:      $V_{\text{feat}} \leftarrow f_{\text{video}}(v_i)$ 
10:     $C_i \leftarrow [G_i \parallel V_{\text{feat}}]$ 
11:     $h_1 \leftarrow \text{GELU}(\theta \cdot W_1 C_i + \theta \cdot b_1)$ 
12:     $h_2 \leftarrow \text{GELU}(\theta \cdot W_2 h_1 + \theta \cdot b_2)$ 
13:     $z_{\text{fine}} \leftarrow \theta \cdot W_3 h_2 + \theta \cdot b_3$ 
14:     $cot\_prompt \leftarrow \text{build\_CoT\_prompt}(z_{\text{fine}})$ 
15:     $\text{tokens\_prompt} \leftarrow \text{LLM.tokenize}(cot\_prompt)$ 
16:     $\text{tokens\_target} \leftarrow \text{LLM.tokenize}(\text{intent}_i)$ 
17:     $\text{output} \leftarrow \text{LLM.generate}(\text{tokens\_prompt})$ 
18:     $\text{loss} \leftarrow \text{CrossEntropy}(\text{output}, \text{tokens\_target})$ 
19:     $\text{loss.backward}()$ 
20:     $\text{optimizer.step}(); \text{optimizer.zero\_grad}()$ 
21:  end for
22: end for
23: return  $P_{\text{ground}}$ 

```

it is resized and cropped to a size of 224×224 . In the first stage, we train for one epoch with a batch size of 256, keeping both the video encoder and language model (LM) frozen, and only training the MLP projector between them. During this stage, we use simple single-turn dialogues to align the vision token with its corresponding description, meaning, or intention text token. In the second stage, we reduce the batch size to 128, train the LM along with the MLP projector, while keeping the encoder frozen. Regarding data usage, we train the model with a constructed long-thought-chain dataset to help the model develop reasoning abilities for gesture descriptions and abstract meanings. At the same time, we use Mediapipe-extracted landmarks as auxiliary information, added as extra tokens to help the model distinguish fine-grained gestures. The initial learning rates for the two stages are set to 1e-3 and 2e-5, with a warmup ratio of 0.03, and the AdamW optimizer is used with a cosine learning rate schedule. We conduct the training on four 80GB A800 GPUs, with the first stage taking about 7 hours and the second stage taking approximately 14 hours. Throughout both stages, we employ full-parameter fine-tuning to ensure maximal model expressivity and performance.

4.2 Dataset

We introduce GestureInt, the first dataset specifically designed for gesture meaning and intention understanding. It comprises gesture videos captured from both egocentric (first-person) and exocentric (third-person) viewpoints organized in the form of QA pairs.

Table 1. Comparison of gesture recognition datasets. Egocentric/Exocentric columns indicate support for egocentric/exocentric viewpoints. Our dataset provides broader coverage with 3 caption types.

Dataset	#Samples	#Classes	#Caption Types	Egocentric	Exocentric
Chalearn LAP IsoGD [40]	47,933	249	1	✗	✓
HMDB-51 [17]	6,766	51	1	✗	✓
Microsoft Kinect & Leap Motion [26]	1,400	10	1	✗	✓
NVIDIA Dynamic Hand Gesture [29]	1,532	25	1	✗	✓
EgoGesture	24,161	83	1	✓	✗
Briareo [25]	120	12	1	✗	✓
IPN Hand [6]	5,649	14	1	✗	✓
JESTER	148,092	27	1	✗	✓
GestureInt (Ours)	159,561	110	3	✓	✓

GestureInt is build on top of two existing datasets—Jester[27] and Egogesture[46], totaling over 150,000 clips across exocentric and egocentric views. GestureInt is constructed via a two-stage annotation pipeline that combines large language model (LLM) generation with expert verification. In Stage 1, all video clips and categories are used to ensure broad coverage. In Stage 2, we focus on 73 intent-bearing categories and construct chain-of-thought (CoT) traces to support intent reasoning.

We structure the dataset in two distinct formats:

1. A three-dimensional separation—where data is explicitly categorized into description, meaning, and intention
2. A fully integrated format—where these dimensions are logically connected to provide a structured and comprehensive reasoning framework.

The first format enables the model to adapt to the new task while the second format helps it learn the implicit patterns linking visual features to semantic meanings.

Our dataset includes approximately 130K third-person gesture videos and 30K first-person gesture videos along with over 400K high-quality textual annotations. A critical flaw in previous gesture datasets is that their annotations are often too brief, lacking the depth necessary for LLMs to accurately learn semantic meanings. Additionally, existing annotations frequently mix descriptions (e.g., “thumbs up”), meanings (e.g., “stop sign”), and intentions (e.g., “make a phone call”) leading to inconsistencies and ambiguity.

To address these issues, we extended the original labels into detailed depictions and generated possible meanings using ChatGPT [31]. Both steps were reviewed and refined by gesture experts to ensure accuracy and consistency. Furthermore, we expanded the dataset by generating possible gesture intentions related to smart device control and smart interface interactions making it more applicable to real-world use cases. Several examples of dataset are available in Appendix A.1.

All annotations were produced through a two-stage, LLM-assisted pipeline. In Stage 1, GPT-4o generated a detailed motion description and a plausible semantic meaning for every gesture clip; two domain experts independently verified and amended these outputs and then added the intention label, with disagreements resolved by discussion. In Stage 2, GPT-4o was prompted again to generate a CoT trace that logically connects description, meaning, and intention, after which the same experts checked the trace for logical soundness and refined it when necessary. For completeness, the full prompt used during the data annotation stage is provided in Appendix A.2.

We split the data as follows: 10 % of the gesture classes are withheld entirely for open-set testing, while from each of the remaining classes we set aside 10 % of the videos for closed-set testing and use the rest for training. For

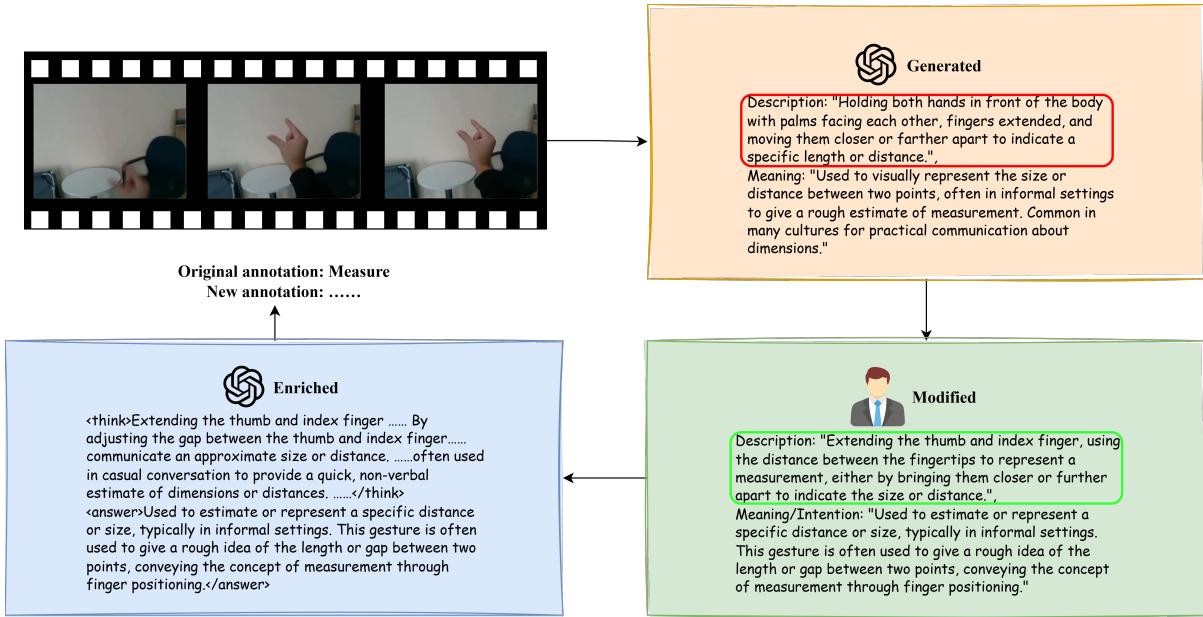


Fig. 4. The annotation pipeline.

instance, the "thumb upward" gesture was held out entirely for open-set evaluation, whereas "thumb downward" videos were included both in training and in closed-set testing.

4.3 Evaluation Metrics

To comprehensively evaluate the performance of our gesture intent understanding system, we adopt a set of complementary metrics that measure both the lexical overlap and the semantic consistency between the model predictions and the reference descriptions. Specifically, we report BLEU-1 to BLEU-4, SPICE, and a GPT-assisted semantic accuracy (ACC) as defined below.

BLEU-1 to BLEU-4. BLEU (Bilingual Evaluation Understudy) is a precision-based metric commonly used in machine translation and text generation. It measures the n-gram overlap between a candidate sentence and one or more reference sentences. BLEU-n refers to the precision of n-grams, and is formally defined as:

$$\text{BLEU-n} = \text{BP} \cdot \exp \left(\sum_{i=1}^n w_i \log p_i \right), \quad (4)$$

where p_i is the modified precision for i -gram, w_i is typically set to $\frac{1}{n}$, and BP is the brevity penalty to penalize overly short predictions. BLEU-1 measures unigram precision (e.g. BLEU-1 score of 66.65 means 66.65 % words shared), BLEU-2 tightens the criterion to bigrams such as "thumb up", and so on up to BLEU-4.

SPICE. SPICE (Semantic Propositional Image Caption Evaluation) evaluates the semantic content of generated sentences by parsing both candidate and reference sentences into scene graphs consisting of objects, attributes, and relations. It computes an F1 score based on the overlap of these semantic tuples. Unlike BLEU, which relies on surface-level n-gram matching, SPICE is more aligned with human judgments in capturing meaning and conceptual correctness, so paraphrases like "lifts thumb in agreement" still earn high marks.

Semantic Accuracy (ACC). To evaluate the semantic correctness of model outputs beyond lexical similarity, we adopt a GPT-4o-assisted evaluation protocol Evaluation and justification of LLM-as-Judge is available in Appendix C. Specifically, we provide GPT-4o with both the model-generated output and the corresponding ground truth label and prompt it to score the semantic similarity on a 0–5 scale:

- 0: completely unrelated
- 1–3: weak or partial match
- 4–5: semantically accurate

We treat a prediction as **correct** if the score assigned by GPT-4o is greater than or equal to 4. The final ACC score is computed as the percentage of samples classified as correct:

$$\text{ACC} = \frac{N_{\text{score} \geq 4}}{N_{\text{total}}} \times 100\%. \quad (5)$$

4.4 Comparative Experiments

We evaluate the performance of several state-of-the-art LVLMs on our GestureInt dataset, targeting gesture intention recognition under open-ended, real-world conditions. As shown in Table 2, our model Gestura consistently surpasses all baselines across both exocentric and egocentric views in terms of intent classification accuracy (ACC) and BLEU scores.

Specifically, Gestura achieves a top-1 intent accuracy of 84.73% in the exocentric setting and 66.14% in the egocentric view, substantially outperforming the previous state-of-the-art, GestureGPT, which achieves 72.08% (closed-set) / 44.46% (open-set) and 40.07% (closed-set) / 17.38% (open-set) under the corresponding settings. This highlights a significant advancement in accurate gesture intent interpretation, particularly under the challenging egocentric perspective.

We further assess each model’s ability to generate descriptive text for gestures using BLEU-1 to BLEU-4 scores. Gestura achieves a BLEU-4 score of 49.83 (exocentric) and 43.67 (egocentric), which are magnitudes higher than those of previous models—e.g., Qwen2.5VL-7B (0.46 and 2.67), or LLaVA-Next-Onevision-7B (0.36 and 0.24). These results highlight Gestura’s superior capacity to translate visual signals into coherent and semantically rich descriptions.

Beyond numeric metrics, Gestura also demonstrates strong qualitative performance. Compared to top-tier models like LLaVA-Next and Qwen2.5VL, which often produce generic or semantically ambiguous outputs, our model generates more context-aware and intention-sensitive descriptions, accurately capturing not just physical movement, but the communicative purpose behind the gesture. This is particularly important in free-form gesture understanding, where similar gestures may convey distinct intents based on subtle variations or situational context.

Overall, our results underscore the effectiveness of combining structured hand landmark information with vision-language pretraining. The substantial gains across both classification and generation tasks validate our two-stage training strategy and landmark-enhanced representation, establishing Gestura as a new state-of-the-art for free-form gesture understanding.

Table 2. Comparison between our Gestura and the currently state-of-the-art LVLM models on our Test sets of GestureInt dataset. Metrics include BLEU-1~4 and ACC (where ACC reports both closed-set / open-set results).

Method	Exocentric					Egocentric				
	ACC	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ACC	BLEU-1	BLEU-2	BLEU-3	BLEU-4
Internvl2.5-8B [8]	8.29	12.07	3.95	1.12	0.34	10.55	14.18	3.76	1.03	0.25
LLaVA-Next-Video-7B-DPO [22]	8.20	13.01	4.72	1.20	0.34	8.41	13.59	4.76	1.26	0.36
LLaVA-Next-Onevision-7B [18]	10.97	16.84	3.86	1.04	0.36	12.28	13.40	3.66	1.00	0.24
LLaVA-Video-7B [47]	11.77	18.13	4.18	1.14	0.38	13.90	20.43	5.24	1.58	0.51
Qwen2.5VL-7B [4]	6.60	17.33	5.85	1.57	0.46	5.42	19.41	9.53	4.80	2.67
Video-LLaVA-7B [19]	2.89	12.50	1.95	0.24	0.04	6.50	13.93	2.35	0.33	0.07
GestureGPT(close)*	72.08	23.53	19.10	16.63	15.05	40.07	21.26	15.14	12.33	10.75
GestureGPT(open)*	44.46	18.65	12.21	9.11	6.79	17.38	21.11	11.74	6.95	4.39
Gestura(close)	84.73	53.94	51.87	50.61	49.83	66.14	52.33	47.72	45.15	43.67
Gestura(open)	65.65	34.88	25.36	19.37	14.17	21.71	33.73	22.10	14.60	9.93

Table 3. Ablation analysis of Qwen2.5VL and Gestura under **exocentric** settings.

Method	Exocentric					
	ACC	BLEU-1	BLEU-2	BLEU-3	BLEU-4	SPICE
Qwen2.5VL-7B (Same data)	close	73.46	45.16	42.13	40.58	39.68
	open	66.76	29.82	21.19	16.16	11.57
Gestura	close	84.73	53.94	51.87	50.61	49.83
	open	65.65	34.88	25.36	19.37	14.17

Table 4. Ablation analysis of Qwen2.5VL and Gestura under **egocentric** settings.

Method	Egocentric					
	ACC	BLEU-1	BLEU-2	BLEU-3	BLEU-4	SPICE
Qwen2.5VL-7B (Same data)	close	41.39	39.11	31.41	27.16	25.06
	open	21.23	31.41	18.55	10.81	5.93
Gestura	close	66.14	52.33	47.72	45.15	43.67
	open	21.71	33.73	22.10	14.60	9.93

4.5 Ablation Study

To better understand the contributions of each component in our system, we conduct a comprehensive ablation study under both exocentric and egocentric test settings. Specifically, we evaluate three configurations: (1) baseline, (2) fine-tuning with only Chain-of-Thought (CoT), (3) full pipeline with both CoT and LPM used during both training and inference, and (4) fine-tuning with both CoT and Landmark Processing Module (LPM) but testing without it. The performance is compared in both closed-set and open-set scenarios.

Ablation Study Results. Table 5 presents a comprehensive ablation study evaluating the contributions of Chain-of-Thought (CoT) and the Landmark Processing Module (LPM) under both exocentric and egocentric settings, across closed-set and open-set scenarios.

We observe that incorporating both CoT and LPM achieves the best overall performance. Specifically, the full model configuration (row 3) reaches the highest intent recognition accuracy in both views: 84.73% (exocentric

Table 5. Ablation Study on ACC results under four ablation experimental conditions across both **Exocentric** and **Egocentric** scenarios.

Methods	Components				Exocentric		Egocentric	
	<i>Multi-View.</i>	<i>CoT.</i>	<i>LPM fine-tuning.</i>	<i>LPM testing.</i>	<i>close</i>	<i>open</i>	<i>close</i>	<i>open</i>
1	-	-	-	-	2.89	/	6.50	/
2	✓	-	-	-	70.13	25.34	53.14	12.25
3	✓	✓	-	-	80.27	63.52	60.34	17.78
4	✓	✓	✓	✓	84.73	65.65	66.14	21.71
Gestura	✓	✓	✓	-	82.73	66.48	66.20	21.06

closed-set) and 66.14% (egocentric closed-set). In comparison, the CoT-only setup (row 2) yields lower accuracy (80.27% and 60.34%, respectively), while the baseline without any of these components performs significantly worse (70.13% and 53.14%). These results demonstrate that LPM offers complementary structural cues that enhance the semantic grounding facilitated by CoT reasoning.

In the open-set setting, where gesture categories are unseen during training, overall performance drops due to the increased challenge of generalization. Nevertheless, our model remains robust: the full configuration still outperforms all alternatives, achieving 65.65% (exocentric) and 21.71% (egocentric).

Considering that the extraction and encoding of Landmark information in practical applications may negatively affect latency, we compare the performance during the test phase with and without using MediaPipe information. Notably, even when LPM is removed at test time (row 4), the model maintains competitive accuracy, highlighting that gesture-aware fine-tuning with LPM imparts lasting benefits even when the module is no longer available during inference. Our experiments demonstrate two key findings: (1) Training with landmark information improves model robustness; (2) The model retains high performance even when landmark data is excluded during testing.

Additionally, to assess the effectiveness of our proposed framework, we conducted an ablation study by fine-tuning the latest multimodal model Qwen2.5VL-7B on the same data used by Gestura. As shown in Tables 8 and 9, despite being trained on identical datasets, Gestura significantly outperforms Qwen2.5VL across both exocentric and egocentric settings. In both closed and open scenarios, Gestura achieves higher accuracy, BLEU scores, and SPICE metrics, demonstrating its superior capability in understanding free-form gestures and inferring intent.

These findings confirm that:

- (1) CoT reasoning can better connect gesture descriptions with their meanings or intentions by providing an objective and logical inference process. This approach enables the model to make reasonable predictions for novel gestures by leveraging existing prior knowledge.
- (2) Gesture landmarks offer both direct (test-time input) and indirect (training-time regularization) advantages, improving robustness and generalizability.
- (3) The combination of CoT and LPM provides the most reliable and interpretable results across diverse perspectives and settings, underscoring the importance of integrating both symbolic and spatial priors for free-form gesture understanding.
- (4) Compared to state-of-the-art LVLM fine-tuned on the same dataset, our framework consistently achieves superior performance across exocentric and egocentric views. This highlights the effectiveness of our framework and the advantage of a structured, intent-aware pipeline.

5 IMPLEMENTATION

In this section, we explore the deployment of Gestura in a device-server hybrid architecture, where the model is hosted on a backend server and communicates with a wearable AI glasses prototype. This setup balances computational efficiency and real-time responsiveness, while maintaining low latency and user mobility.

We describe our system implementation, including the hardware setup of the AI glasses and the interaction pipeline from gesture capture to audio feedback. To evaluate real-world usability, we conducted a user study in which participants performed intuitive, free-form gestures based on given intent prompts. We further assess the system's performance through metrics such as intent recognition accuracy and response time, demonstrating that Gestura achieves reliable on-the-fly inference. Finally, feature-level analysis reveals how semantic reasoning helps unify diverse gesture expressions, reinforcing the importance of integrating visual and language understanding in gesture-intent recognition.

5.1 System Setup

The implementation of Gestura for on-device deployment is mainly on our prototype AI glasses. We implement the model first, and using the AI glasses as the wearable device to handle the visual input with its monocular camera in the middle of the frame, and the audio input with its microphone. After receiving, the inputs go directly to the model for processing, the output text go straight to TTS tool and the transformed audio sent to the AI glasses and played by the embedded speaker. Specifically, Gestura was deployed on a NVIDIA A100 GPU for inferencing. The AI glasses equipped Qualcomm Hexagon™ NPU, with a lightweight Linux-based operating system run on it. Video input was captured via an embedded camera with a resolution of 4K resolution at 30 FPS, ensuring compatibility with real-world gesture capture scenarios.

Task Description: Specific eight tasks assigned to participants related to smart device and interface control. Participants are required to execute gesture freely based on provided key words about tasks.

Task Procedure: Participants were first given a brief introduction to the smart device and interface control task without any prior guidance on specific gestures. They then wore our device and proceeded with the experiment. Each participant was provided with eight independent keyword phrases and was instructed to spontaneously perform the most natural and intuitive gesture corresponding to each phrase, without prior contemplation. The system recorded the experimental data based on the feedback received from the device.

5.2 Data Collection and Participants

To evaluate the on-device deployment, we conducted an additional user study involving 13 volunteers. We recruited 13 adult volunteers (9 men, 4 women) from an institute of a company. Their ages ranged from 20 to 30 years (MEAN = 23.8, SD = 4.2). Gesture-interaction experience was assessed with a pre-study questionnaire, which shows that seven participants have prior gesture-based interface experience while others don't.

Participants were presented with a set of 8 intent-related gesture keywords and were instructed to perform free-form gestures based on their intuitive understanding of each keyword, without receiving any predefined demonstrations or constraints. This setup aimed to capture naturalistic and diverse gesture expressions across individuals.

In total, the study yielded 104 gesture samples (13 participants \times 8 keywords), each recorded as a first-person perspective video with an average duration of approximately 2 seconds. All recordings were collected under consistent lighting and environmental conditions to ensure data quality and comparability. This participant-generated dataset serves as a valuable complement to our open-source gesture-intent corpus, offering realistic and unconstrained gesture expressions for evaluating real-time recognition performance in wearable scenarios.

5.3 Evaluation

The performance of the on-device *Gestura* system was evaluated by incorporating a language-based intent inference approach. Specifically, we utilized ChatGPT as the intent recognition intermediary: for each gesture video, our model first generated a textual output, which was subsequently fed into ChatGPT to infer the most likely corresponding function.

For each sample, we collected the Top-1, Top-3, and Top-5 intent predictions based on ChatGPT’s semantic ranking against a predefined intent pool. Recognition accuracy was then computed as the percentage of cases in which the ground-truth intent appeared within the top- k predictions ($k \in \{1, 3, 5\}$).

In addition to intent recognition accuracy, we assessed the system using the following metrics:

- **Top- k Accuracy:** The proportion of test samples where the correct intent appeared in the top-1, top-3, or top-5 predicted intents.
- **Response Time:** The average latency (in milliseconds) from video input to final intent prediction, capturing the total runtime of the vision-to-language-to-intent pipeline.

These evaluation metrics reflect a comprehensive balance between recognition accuracy, responsiveness, and deployment efficiency, addressing the practical considerations of real-world edge computing scenarios.

5.4 Overall Performance

Table 6. Top-1 / Top-3 / Top-5 accuracy per intent category in an open-world experiment.

Intent Category	Samples	Top-1 Acc	Top-3 Acc	Top-5 Acc
come over / beckon	13	0.0000	0.3846	0.7692
confirm / agree	13	0.7692	1.0000	1.0000
decrease temp (AC)	13	0.1538	0.3846	0.3846
go to next page	13	0.3077	0.5385	0.6154
go to previous page	13	0.4615	0.7692	0.8462
increase temp (AC)	13	0.0000	0.0769	0.1538
make a phone call	13	0.6154	0.7692	0.9231
take a photo	13	0.6923	0.6923	0.8462
Overall	104	0.3750	0.5769	0.6923

Based on the results presented in Table 6, our system demonstrates promising performance on egocentric gesture-intent recognition, achieving an overall Top-1 accuracy of 37.5%, Top-3 accuracy of 57.7%, and Top-5 accuracy of 69.2% across all eight intent categories. These results highlight the system’s effectiveness in capturing intent-relevant semantics, even when users perform free-form and visually diverse gestures. In terms of latency, the system achieves a practical response time ranging from 5 to 10 seconds, with an average of **7.83 seconds**. Notably, when excluding communication delays and TTS overhead, the core model processes inputs and produces outputs within an average of just **1.6 seconds**—a substantial improvement over the multi-agent-based GestureGPT, which requires 227 seconds per gesture. This dramatic reduction in response time is largely attributed to our lightweight, end-to-end framework design, which eliminates the need for complex inter-agent communication and enables much faster and more efficient inference.

Intent categories such as “confirm / agree” and “make a phone call” achieve particularly high recognition rates (Top-1 accuracy of 76.9% and 61.5% respectively), suggesting that these gestures tend to be more consistent across users. In contrast, categories like “increase temp (AC)” and “come over / beckon” exhibit lower Top-1 performance, reflecting greater variability in individual expression and ambiguity in visual cues. However, the Top-3 and Top-5

accuracies indicate that the correct intent is still frequently ranked among the model’s top predictions, which is valuable for downstream interactive applications.

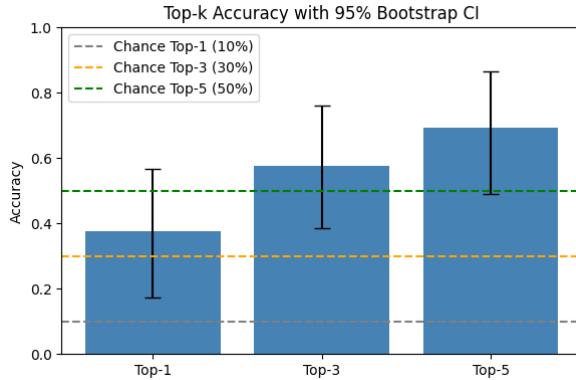


Fig. 5. Top-k Accuracy with 95% Bootstrap CI.

We provide an error-bar plot 5 contrasting our Top-1/3/5 accuracies with their chance baselines. One-sample t-tests show that Top-1 ($p = 0.0386 < 0.05$, $d = 0.90$) and Top-3 ($p = 0.0309$, $d = 0.95$) outperform chance significantly, whereas Top-5 accuracy showed a positive trend ($p = 0.1036 \approx 0.10$).

To strengthen the statistical interpretation, we additionally performed bootstrap analysis with 10,000 resamples, yielding consistent results. For Top-5, the bootstrap delta was still positive (+19.2 pp), and the effect size remained moderate ($d = 0.66$), the 95% CI marginally included zero [-1.0 pp, +36.5 pp]. This suggests a positive trend despite the lack of statistical significance, reinforcing the overall robustness of the model under looser evaluation thresholds. The Top-1 improvement over chance was +27.5 pp (95% BCa CI = [7.3 pp, 46.7 pp], $d = 0.90$), and Top-3 was +27.7 pp (CI = [8.5 pp, 46.0 pp], $d = 0.95$). These results confirm that the model is statistically and practically effective in producing correct Top-1 and Top-3 predictions.

Overall, these results demonstrate the robustness and adaptability of our approach in realistic, unconstrained egocentric scenarios, and underscore the benefits of integrating visual and semantic understanding for gesture-based intent inference.

5.5 Analysis

The variability of real-world gesture performance can introduce a gap between controlled experimental data and practical applications. Categories with semantic uniqueness tend to achieve better performance.

Misdirections in the open-world test arise mainly from three factors. (1) Cultural and personal variation: volunteers often select different gestures to express the same intent. (2) Motion symmetry: visually similar trajectories such as “hand moving up” versus “hand moving down” remain hard to disambiguate. This situation appears as well when we generating enriched motion description using GPT-4o (3) Gesture repetition: users frequently repeat dynamic motions for emphasis, which obscures directionality and confuses opposite pairs like “scroll up vs. scroll down” or “swipe left vs. swipe right.”

In real-world experiments, we observed that participants often performed different gestures to express the same intent. As illustrated in Figure 6, the visual features extracted by the video encoder – represented as tensors of shape (8, 4096) – tend to exhibit clustering only when the performed gestures are visually similar, such as

those corresponding to Gesture Type 1 and Gesture Type 7. However, for other gesture types, due to individual variations in motor expression, the extracted features are more dispersed, leading to less distinct clusters.

In contrast, as shown in Figure 7, the semantic features derived from LLM output tokens – typically in the shape of (56, 4096) – demonstrate clearer clustering patterns even when the visual gestures differ. By leveraging our proposed pipeline, gestures that differ significantly in visual appearance yet share the same underlying intent are effectively brought closer together in the semantic embedding space. These findings highlight the importance of gesture-intent recognition, as it enables more robust understanding beyond surface-level gesture differences.

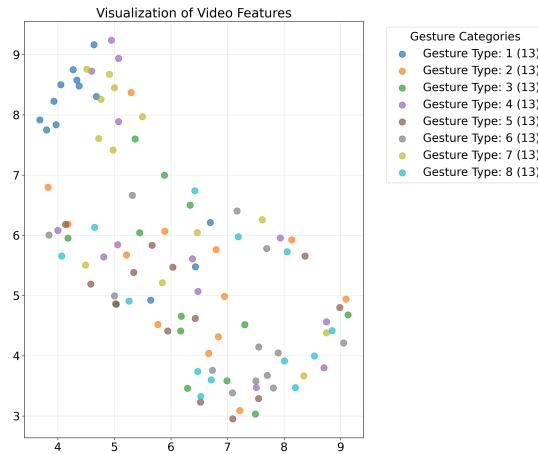


Fig. 6. Visualization of Video Features.

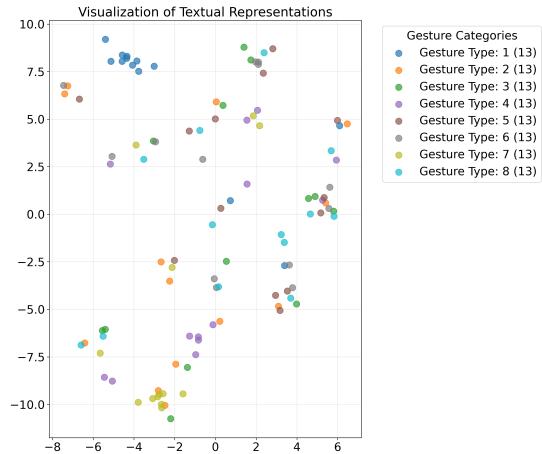


Fig. 7. Visualization of Textual Representations.

6 DISCUSSION

Despite the progress made in free-form gesture comprehension, a persistent challenge remains: how can open-set gesture recognition systems approach the performance of closed-set counterparts? In closed-set settings, models operate under the assumption that all gesture categories are predefined and known during training, leading to high accuracy but limited generalization. Conversely, open-set systems must generalize to unseen or user-defined gestures, which inherently introduces ambiguity and performance degradation.

The performance gap between exocentric closed-set and egocentric open-set scenarios arises from several factors. First, the exocentric subset contains fewer categories but more training samples per class, making it easier for the model to generalize. In contrast, the egocentric set exhibits greater category diversity and dynamic hand motions, increasing ambiguity. Second, in open-world settings, the model may force predictions into seen categories even when visual matches are poor—particularly problematic for egocentric gestures with more varied semantics. Third, egocentric videos are often affected by environmental noise such as lighting or cluttered backgrounds. Lastly, human performance variability (e.g., due to cultural factors or directional ambiguity) further compounds the recognition challenge.

While Gestura is currently deployed on a single NVIDIA A100 40GB GPU for benchmarking, its architecture is designed with efficiency in mind. Given its 8B parameter scale and streamlined end-to-end design, there is strong potential for edge deployment. Notably, with optimization techniques such as model quantization, pruning, and distillation, Gestura can be significantly compressed without substantial performance loss. For instance, platforms like the NVIDIA Jetson AGX Orin[30], which provides up to 64GB of shared memory and 2048-core Ampere GPU, offer a feasible hardware base for deploying a compressed version of our model. This opens the door for real-time,

free-form gesture understanding on edge devices, enabling applications in smart homes, AR/VR systems, and robotics without reliance on cloud resources.

To bridge this gap, we propose several future directions:

1) Enriching high-quality, semantically-annotated gesture data: Our current framework benefits from a structured, QA-formatted gesture-intent dataset. Expanding this dataset with more diverse users, environments, and cultural contexts would significantly enhance model robustness. Incorporating multimodal annotations—e.g., combining gesture videos with eye gaze, speech cues, or physiological signals—could provide richer context for disambiguation.

2) Adaptive inference through test-time scaling: Instead of relying solely on fixed training representations, test-time adaptation techniques (e.g., retrieval-augmented inference, few-shot adaptation, or gradient-based prompt tuning) may allow models to dynamically adjust to novel gestures. These approaches enable the system to scale semantically at inference time without retraining, better capturing the latent intent behind unfamiliar hand motions.

3) Contrastive grounding with weak supervision: Leveraging large-scale, weakly labeled video data (e.g., instructional videos, sign language corpora) via contrastive pretraining may help align free-form gestures with high-level semantics in the absence of dense annotation. This could reduce the reliance on curated datasets while enhancing open-set generalization.

While Gestura is currently deployed on a single NVIDIA A100 40GB GPU for benchmarking, its architecture is designed with efficiency and scalability in mind. Given its 8B parameter scale and streamlined end-to-end design, there is strong potential for edge deployment. Notably, with optimization techniques such as model quantization, pruning, and knowledge distillation, Gestura can be significantly compressed without substantial performance loss. For instance, platforms like the NVIDIA Jetson AGX Orin, which offers up to 64GB of shared memory and a 2048-core Ampere GPU, provide a viable hardware foundation for running a lightweight version of the model. This opens the door for real-time, free-form gesture understanding on edge devices, enabling applications in smart homes, AR/VR systems, and robotics without reliance on cloud resources.

Ultimately, we envision a future where open-set gesture recognition systems not only narrow the performance gap with their closed-set counterparts but also excel in adaptability, efficiency, and deployability. Realizing this vision requires synergistic advancements in robust gesture encoding, semantic grounding, and context-aware reasoning, alongside hardware-aware model design. With frameworks like Gestura, which balance performance and efficiency, we are one step closer to enabling real-time, intuitive, and accessible gesture interaction across diverse real-world environments.

7 CONCLUSION

In this work, we introduce Gestura, a free-form hand gesture understanding system that harnesses the power of LVLMs to bridge the gap between motion patterns and semantic intent. Our framework is built upon a novel hierarchical architecture that combines landmark processing with advanced cross-modal alignment techniques. This enables the system to not only interpret subtle hand movements but also to infer the underlying user intent with remarkable accuracy. Through extensive experiments in real-world scenarios, Gestura demonstrates exceptional performance, achieving a top-1 intent accuracy of 84.73% in exocentric settings and 66.14% in egocentric settings. These results highlight the system’s ability to handle the complexity and variability of open-world environments. Moreover, Gestura’s average response time of just 7.83 seconds, with pure model inference taking only 1.6 seconds, underscores its suitability for real-time applications. By eliminating the need for users to learn predefined gestures, Gestura sets a new standard for intuitive and natural human-computer interaction, opening avenues for future research in creating more adaptive and user-friendly interfaces.

8 Acknowledgments

We thank the AI Glasses product team at China Telecom for their enthusiastic support and valuable assistance.

References

- [1] Shubhra Aich, Jesus Ruiz-Santaquiteria, Zhenyu Lu, Prachi Garg, K J Joseph, Alvaro Fernandez Garcia, Vineeth N Balasubramanian, Kenrick Kin, Chengde Wan, Necati Cihan Camgoz, Shugao Ma, and Fernando De La Torre. 2023. Data-Free Class-Incremental Hand Gesture Recognition. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, Paris, France, 20901–20910. [doi:10.1109/ICCV51070.2023.01916](https://doi.org/10.1109/ICCV51070.2023.01916)
- [2] Naif Al Mudawi, Hira Ansar, Abdulwahab Alazeb, Hanan Aljuaid, Yahay AlQahtani, Asaad Algarni, Ahmad Jalal, and Hui Liu. 2024. Innovative Healthcare Solutions: Robust Hand Gesture Recognition of Daily Life Routines Using 1D CNN. *Frontiers in Bioengineering and Biotechnology* 12 (July 2024), 1401803. [doi:10.3389/fbioe.2024.1401803](https://doi.org/10.3389/fbioe.2024.1401803)
- [3] Tenglong Ao, Zeyi Zhang, and Libin Liu. 2023. GestureDiffuCLIP: Gesture Diffusion Model with CLIP Latents. *ACM Transactions on Graphics* 42, 4 (Aug. 2023), 1–18. [doi:10.1145/3592097](https://doi.org/10.1145/3592097)
- [4] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. 2025. Qwen2.5-VL Technical Report. arXiv:2502.13923 [cs.CV] <https://arxiv.org/abs/2502.13923>
- [5] Andrea Bandini and José Zariffa. 2022. Analysis of the Hands in Egocentric Vision: A Survey. [doi:10.48550/arXiv.1912.10867](https://doi.org/10.48550/arXiv.1912.10867) arXiv:1912.10867
- [6] Gibran Benitez-Garcia, Jesus Olivares-Mercado, Gabriel Sanchez-Perez, and Keiji Yanai. 2021. IPN Hand: A Video Dataset and Benchmark for Real-Time Continuous Hand Gesture Recognition . In *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE Computer Society, Los Alamitos, CA, USA, 4340–4347. [doi:10.1109/ICPR48806.2021.9412317](https://doi.org/10.1109/ICPR48806.2021.9412317)
- [7] Yuhu Chang, Yingying Zhao, Mingzhi Dong, Yujiang Wang, Yutian Lu, Qin Lv, Robert P. Dick, Tun Lu, Ning Gu, and Li Shang. 2021. MemX: An Attention-Aware Smart Eyewear System for Personalized Moment Auto-capture. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 2, Article 56 (June 2021), 23 pages. [doi:10.1145/3463509](https://doi.org/10.1145/3463509)
- [8] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yimin Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, Conghui He, Botian Shi, Xingcheng Zhang, Han Lv, Yi Wang, Wenqi Shao, Pei Chu, Zhongying Tu, Tong He, Zhiyong Wu, Huipeng Deng, Jiaye Ge, Kai Chen, Kaipeng Zhang, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhui Wang. 2025. Expanding Performance Boundaries of Open-Source Multimodal Models with Model, Data, and Test-Time Scaling. arXiv:2412.05271 [cs.CV] <https://arxiv.org/abs/2412.05271>
- [9] DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghai Lu, Shangyan Zhou, Shanhuan Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. arXiv:2501.12948 [cs.CL] <https://arxiv.org/abs/2501.12948>

- [10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *Corr* abs/2010.11929 (2020). arXiv:2010.11929 <https://arxiv.org/abs/2010.11929>
- [11] Guhao Feng, Bohang Zhang, Yuntian Gu, Haotian Ye, Di He, and Liwei Wang. 2023. Towards revealing the mystery behind chain of thought: a theoretical perspective. In *Proceedings of the 37th International Conference on Neural Information Processing Systems* (New Orleans, LA, USA) (NIPS '23). Curran Associates Inc., Red Hook, NY, USA, Article 3100, 42 pages.
- [12] Mark Higger, Polina Rygina, Logan Daigler, Lara Ferreira Bezerra, Zhao Han, and Tom Williams. [n. d.]. Toward Open-World Human-Robot Interaction: What Types of Gestures Are Used in Task-Based Open-World Referential Communication? ([n. d.]).
- [13] Zhizhang Hu, Amirmohammad Radmehr, Yue Zhang, Shijia Pan, and Phuc Nguyen. 2024. IOTeeth: Intra-Oral Teeth Sensing System for Dental Occlusal Diseases Recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, 1 (March 2024), 1–29. doi:10.1145/3643516
- [14] Zeyuan Huang, Cangjun Gao, Haiyan Wang, Xiaoming Deng, Yu-Kun Lai, Cuixia Ma, Sheng-feng Qin, Yong-Jin Liu, and Hongan Wang. 2024. SpeciFingers: Finger Identification and Error Correction on Capacitive Touchscreens. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, 1 (March 2024), 1–28. doi:10.1145/3643559
- [15] Yonchanok Khaokaew, Hao Xue, and Flora D. Salim. 2024. MAPLE: Mobile App Prediction Leveraging Large Language Model Embeddings. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, 1 (March 2024), 1–25. doi:10.1145/3643514
- [16] 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. doi:10.1145/3643505
- [17] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. 2011. HMDB: a large video database for human motion recognition. In *Proceedings of the International Conference on Computer Vision (ICCV)*.
- [18] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024. LLaVA-OneVision: Easy Visual Task Transfer. arXiv:2408.03326 [cs.CV] <https://arxiv.org/abs/2408.03326>
- [19] Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. 2024. Video-LLaVA: Learning United Visual Representation by Alignment Before Projection. arXiv:2311.10122 [cs.CV] <https://arxiv.org/abs/2311.10122>
- [20] Manousos Linardakis, Iraklis Varlamis, and Georgios Th Papadopoulos. 2025. Survey on Hand Gesture Recognition from Visual Input. doi:10.48550/arXiv.2501.11992 arXiv:2501.11992 [cs]
- [21] Chen Liu, Zixuan Dong, Li Huang, Wenlong Yan, Xin Wang, Dingyi Fang, and Xiaojiang Chen. 2024. TagSleep3D: RF-based 3D Sleep Posture Skeleton Recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, 1 (March 2024), 1–28. doi:10.1145/3643512
- [22] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge. <https://llava-vl.github.io/blog/2024-01-30-llava-next/>
- [23] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems* (New Orleans, LA, USA) (NIPS '23). Curran Associates Inc., Red Hook, NY, USA, Article 1516, 25 pages.
- [24] Camillo Lugaressi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Ubweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. 2019. MediaPipe: A Framework for Building Perception Pipelines. arXiv:1906.08172 [cs.DC] <https://arxiv.org/abs/1906.08172>
- [25] Fabio Manganaro, Stefano Pini, Guido Borghi, Roberto Vezzani, and Rita Cucchiara. 2019. Hand Gestures for the Human-Car Interaction: the Briareo dataset. In *International Conference on Image Analysis and Processing*. Springer, 560–571.
- [26] Giulio Marin, Fabio Dominio, and Pietro Zanuttigh. 2014. Hand gesture recognition with leap motion and kinect devices. In *2014 IEEE International Conference on Image Processing (ICIP)*. 1565–1569. doi:10.1109/ICIP.2014.7025313
- [27] Joanna Materzynska, Guillaume Berger, Ingo Bax, and Roland Memisevic. 2019. The Jester Dataset: A Large-Scale Video Dataset of Human Gestures. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*. 2874–2882. doi:10.1109/ICCVW.2019.00349
- [28] Abu Saleh Musa Miah, Md. Al Mehedi Hasan, and Jungpil Shin. 2023. Dynamic Hand Gesture Recognition Using Multi-Branch Attention Based Graph and General Deep Learning Model. *IEEE Access* 11 (2023), 4703–4716. doi:10.1109/ACCESS.2023.3235368
- [29] Pavlo Molchanov, Xiaodong Yang, Shalini Gupta, Kihwan Kim, Stephen Tyree, and Jan Kautz. 2016. Online Detection and Classification of Dynamic Hand Gestures with Recurrent 3D Convolutional Neural Networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 4207–4215. doi:10.1109/CVPR.2016.456
- [30] NVIDIA. [n. d.]. NVIDIA Jetson Orin Next-level AI performance for next-gen robotics and edge solutions. <https://www.nvidia.com/en-us/autonomous-machines/embedded-systems/jetson-orin/>
- [31] 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 Babuschkin, 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 Decareaux, 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, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jom, 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 Kiros, Matt Knight, Daniel Kokotajlo, Lukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai 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 Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, 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, Mario 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 Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, 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. arXiv:2303.08774 [cs.CL] <https://arxiv.org/abs/2303.08774>

[32] Jing Qi, Li Ma, Zhenchao Cui, and Yushu Yu. 2024. Computer Vision-Based Hand Gesture Recognition for Human-Robot Interaction: A Review. *Complex & Intelligent Systems* 10, 1 (Feb. 2024), 1581–1606. doi:10.1007/s40747-023-01173-6

[33] Xiangyao Qi, Qi Lu, Wentao Pan, Yingying Zhao, Rui Zhu, Mingzhi Dong, Yuhu Chang, Qin Lv, Robert P. Dick, Fan Yang, Tun Lu, Ning Gu, and Li Shang. 2023. CASES: A Cognition-Aware Smart Eyewear System for Understanding How People Read. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 7, 3, Article 115 (Sept. 2023), 31 pages. doi:10.1145/3610910

[34] Razieh Rastgoo, Kourosh Kiani, Sergio Escalera, and Mohammad Sabokrou. 2024. Multi-Modal Zero-Shot Dynamic Hand Gesture Recognition. *Expert Systems with Applications* 247 (Aug. 2024), 123349. doi:10.1016/j.eswa.2024.123349

[35] Emanuele Rodolà, Luca Cosmo, Or Litany, Michael M. Bronstein, Alexander M. Bronstein, N. Audebert, A. Ben Hamza, Alexandre Boulch, Umberto Castellani, Minh N. Do, Anh Duc Duong, Andrea Gasparetto, Y. Hong, J. Kim, B. L. Saux, Roee Litman, Majid Masoumi, Giorgia Minello, Ryutarou Ohbuchi, Thuyen V. Phan, M. Rezaei, A. Torsello, Minh-Triet Tran, Q. T. Tran, Bao Truong, Lili Wan, and Changqing Zou. 2017. SHREC '17 : Deformable Shape Retrieval with Missing Parts. <https://api.semanticscholar.org/CorpusID:3993253>

[36] Junxiao Shen, Matthias De Lange, Xuhai "Orson" Xu, Enmin Zhou, Ran Tan, Naveen Suda, Maciej Lazarewicz, Per Ola Kristensson, Amy Karlson, and Evan Strasnick. 2024. Towards Open-World Gesture Recognition. doi:10.48550/arXiv.2401.11144 arXiv:2401.11144 [cs]

[37] Shaikh Shawon Arefin Shimon, Ali Neshati, Junwei Sun, Qiang Xu, and Jian Zhao. 2024. Exploring Uni-manual Around Ear Off-Device Gestures for Earables. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, 1 (March 2024), 1–29. doi:10.1145/3643513

[38] Ke Sun, Chunyu Xia, Xinyu Zhang, Hao Chen, and Charlie Jianzhong Zhang. 2024. Multimodal Daily-Life Logging in Free-living Environment Using Non-Visual Egocentric Sensors on a Smartphone. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, 1 (March 2024), 1–32. doi:10.1145/3643553

[39] Qi Sun, Tong Zhang, Shang Gao, Liuqingqing Yang, and Fenghua Shao. 2024. Optimizing Gesture Recognition for Seamless UI Interaction Using Convolutional Neural Networks. doi:10.48550/arXiv.2411.15598 arXiv:2411.15598 [cs]

[40] Jun Wan, Chi Lin, Longyin Wen, Yunan Li, Qiguang Miao, Sergio Escalera, Gholamreza Anbarjafari, Isabelle Guyon, Guodong Guo, and Stan Z. Li. 2022. ChaLearn Looking at People: IsoGD and ConGD Large-Scale RGB-D Gesture Recognition. *IEEE Transactions on Cybernetics* 52, 5 (2022), 3422–3433. doi:10.1109/TCYB.2020.3012092

[41] Shuning Wang, Linghui Zhong, Yongjian Fu, Lili Chen, Ju Ren, and Yaoxue Zhang. 2024. UFace: Your Smartphone Can "Hear" Your Facial Expression! *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, 1 (March 2024), 1–27. doi:10.1145/3643546

[doi:10.1145/3643546](https://doi.org/10.1145/3643546)

- [42] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems* (New Orleans, LA, USA) (*NIPS* '22). Curran Associates Inc., Red Hook, NY, USA, Article 1800, 14 pages.
- [43] Kaijie Xiao, Yi Gao, Fu Li, Weifeng Xu, Pengzhi Chen, and Wei Dong. 2024. ChatCam: Embracing LLMs for Contextual Chatting-to-Camera with Interest-Oriented Video Summarization. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, 4 (Nov. 2024), 1–34. [doi:10.1145/3699731](https://doi.org/10.1145/3699731)
- [44] Xin Zeng, Xiaoyu Wang, Tengxiang Zhang, Chun Yu, Shengdong Zhao, and Yiqiang Chen. 2024. GestureGPT: Toward Zero-shot Interactive Gesture Understanding and Grounding with Large Language Model Agents. [doi:10.48550/arXiv.2310.12821](https://doi.org/10.48550/arXiv.2310.12821) arXiv:2310.12821 [cs]
- [45] Dell Zhang, Yongxiang Li, Zhongjiang He, and Xuelong Li. 2024. Empowering Smart Glasses with Large Language Models: Towards Ubiquitous AGI. In *Companion of the 2024 on ACM International Joint Conference on Pervasive and Ubiquitous Computing* (Melbourne VIC, Australia) (*UbiComp* '24). Association for Computing Machinery, New York, NY, USA, 631–633. [doi:10.1145/3675094.3678992](https://doi.org/10.1145/3675094.3678992)
- [46] 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 (2018), 1038–1050. [doi:10.1109/TMM.2018.2808769](https://doi.org/10.1109/TMM.2018.2808769)
- [47] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. 2024. Video Instruction Tuning With Synthetic Data. arXiv:2410.02713 [cs.CV] <https://arxiv.org/abs/2410.02713>

A Dataset

A.1 Data Subset

Gesture Class	Image	Motion Description	Semantic Meaning / Intention	CoT Reasoning (Excerpt)
Measure		Extending the thumb and index finger	This gesture is often used to give a rough idea of the length or gap between two points, conveying the concept of measurement through finger positioning.	<think>.....where the distance between the fingertips serves as the measurable span.</think><answer>This gesture is often used to give a rough idea of the length or gap between two points, conveying the concept of measurement through finger positioning.</answer>
Phone call		Extending the thumb and pinky finger while curling the other fingers inward	Indicates the action of making a phone call or suggesting someone should call you.	<think>.....mimics the shape and position of a traditional telephone handset.</think><answer>This gesture is used to initiate a phone call to someone.</answer>
Thumb up		Raising the thumb upward while the other fingers are curled into the palm	Represents approval, agreement, or a positive acknowledgment. Commonly used in many cultures to signal that something is good or satisfactory.	<think>.....This gesture is widely recognized and utilized</think><answer>Represents approval, agreement, or a positive acknowledgment. Commonly used in many cultures to signal that something is good or satisfactory.</answer>
Push hand away		Extending the arm forward with the palm facing outward and fingers spread	It can signify refusal, disapproval, or a request for personal space. This gesture signals to stop or prevents entry into a particular area.	<think>.....create a physical or metaphorical boundary.</think><answer>It can signify refusal, disapproval, or a request for personal space. This gesture signals to stop or prevents entry into a particular area.</answer>

Table 7. Examples of GestureInt annotations, including image, semantic breakdown, and reasoning.

A.2 Data Annotation

First, we input the collected gesture videos and prompts that ask to caption the video with description and meaning into ChatGPT-4o. Then we have experts check whether the output from ChatGPT-4o matches the gesture category, and make modifications if there are discrepancies. This generates a dataset of captions that describe the gestures and their meaning, which we use as the training data for the first stage of training. The prompt given to GPT-4o as follow:

```
'role': 'user',
'content': 'Please caption the following hand gesture video by providing a detailed description of hands and its potential meaning:
Gesture Video: {video_data}
Provide your response only as a Python dictionary string with keys, 'description', and 'meaning'.
- 'description' should be a clear, concise explanation of the gesture's physical appearance and common usage.
- 'meaning' should explain the potential interpretation or cultural significance of the gesture.
DO NOT PROVIDE ANY OTHER OUTPUT TEXT OR EXPLANATION. Only provide the Python dictionary string.
For example, your response should look like this: { 'description': 'Raising the thumb upward while other fingers are curled.', 'meaning': 'Represents approval or agreement.'}'
```

For the second phase, we provide ChatGPT-4o with a caption describing a gesture and explicitly inform it of the intended meaning or purpose of the gesture, instructing ChatGPT-4o to reason through the relationship between the description of the gesture and its intended meaning. And ask it to put his thought process in `<think> {reason} </think>`, ending with `<answer> {gesture meaning} </answer>`. The prompt given to ChatGPT-4o as follow:

```
'role': 'user',
'content': 'Given the following gesture description and its intended meaning, please explain the reasoning process that connects the physical appearance of the gesture to its intended interpretation.
- Gesture Description: {description}
- Intended Meaning: {meaning}
Write your reasoning enclosed in <think> ...</think> , and conclude with the inferred gesture meaning enclosed in <answer> ...</answer> .
Only output the reasoning and answer in the specified format. DO NOT include any other text, comments, or explanations.
For example, your response should look like this:
<think> Raising the thumb is commonly used in many cultures to indicate positivity or agreement. Since the gesture matches this form, the intended meaning is approval.</think>
<answer> Represents approval or agreement.</answer> '
```

The output generated by ChatGPT-4o, becomes the target for our training data. The input for this training data consists of two components: the video data (which provides the visual information) and the prompt "Please describe the gesture in detail and provide its meaning or intent." The model is expected to generate a coherent, structured response that reflects both the description of the gesture and its intended purpose.

B Ablation of freezing the LLM’s weights

Freezing the LLM during Stage 1 leads to consistent improvements across all settings (table 8, table 9), especially in egocentric and open-world scenarios. This suggests that preserving the language model’s prior knowledge is crucial for robust gesture understanding.

Table 8. Ablation analysis of Gestura with LLM unfrozen in stage 1 and Gestura under **exocentric** settings.

Method	Exocentric					
	ACC	BLEU-1	BLEU-2	BLEU-3	BLEU-4	SPICE
Gestura(LLM unfrozen)	close	83.01	51.20	50.92	49.57	48.76
	open	64.61	34.58	25.15	19.13	13.97
Gestura	close	84.73	53.94	51.87	50.61	49.83
	open	65.65	34.88	25.36	19.37	14.17
						0.63
						0.30

Table 9. Ablation analysis of Gestura with LLM unfrozen in stage 1 and Gestura under **egocentric** settings.

Method	Egocentric					
	ACC	BLEU-1	BLEU-2	BLEU-3	BLEU-4	SPICE
Gestura(LLM unfrozen)	close	56.38	47.03	41.32	38.21	36.50
	open	20.18	34.21	22.50	14.82	9.85
Gestura	close	66.14	52.33	47.72	45.15	43.67
	open	21.71	33.73	22.10	14.60	9.93
						0.56
						0.20

C LLM as Judge

Table 10. Pairwise inter-rater agreement (Cohen κ) between the four human raters.

	human1	human2	human3	human4
human1	—	0.906	0.906	0.775
human2	0.906	—	0.964	0.883
human3	0.906	0.964	—	0.843
human4	0.775	0.883	0.843	—

Table 11. Agreement between GPT-4o and the human majority vote.

	Cohen κ	MCC	MAE (0/1)
GPT-4o vs. Human	0.982	0.982	0.016

We evaluate the reliability of GPT-4o as an automatic judge on a 200-sample subset independently labelled by four human raters. Using majority voting, we construct a human gold standard \mathbf{H} and compare it against GPT-4o’s binary decisions \mathbf{G} .

To quantify agreement, we report three complementary metrics. The mean absolute error (MAE) between GPT and human labels is $MAE(G, H) = 0.016$, meaning GPT-4o's decisions differ from the human consensus on only 1.6% of cases. Cohen's $\kappa = 0.982$ and Matthews correlation coefficient (MCC) = 0.982 both indicate great agreement, consistent with prior interpretation standards.

We also assess potential bias and instability. The average rating difference $\bar{G} - \bar{H}$ is $+0.020$, suggesting GPT is 2 percentage points more lenient than humans. To test robustness, we prompted GPT-4o three times using paraphrased instructions. The standard deviation of verdicts across these runs was 0.09, indicating low instability and minimal sensitivity to prompt variation.

Overall, these results confirm that GPT-4o produces judgments that are highly consistent with expert human raters, with negligible systematic bias and strong reliability across reformulations. We therefore employ GPT-4o as the primary evaluator for all large-scale experiments in this study.