MITSUBISHI ELECTRIC RESEARCH LABORATORIES
https://www.merl.com

GPT Sonography: Hand Gesture Decoding from Forearm Ultrasound Images via a Large Vision-Language Model

Bimbraw, Keshav; Wang, Ye; Liu, Jing; Koike-Akino, Toshiaki

TR2026-054 May 16, 2026

Abstract

Large vision-language models (LVLMs), such as the Generative Pre-trained Transformer 4-omni (GPT-4o), are emerging multi-modal foundation models which have great potential as powerful artificial-intelligence (AI) assistance tools for a myriad of applications, including healthcare, industrial, and academic sectors. Although such foundation models perform well in a wide range of general tasks, their capability without fine-tuning is often limited in specialized tasks. However, full fine-tuning of large foundation models is challenging due to enormous computation/memory/dataset requirements. Ultrasound data from the forearm has been shown to be used for hand gesture estimation. However, this typically requires training deep learning models with a large quantity of labeled data. We show that GPT-4o can decode hand gestures from forearm ultrasound data even with no fine-tuning, and improves with few-shot, retrieval augmented in-context learning. In our experiments, the average classification accuracy improved from 19.3% (0-shot) to 74.0% (2-shot) for within-session testing, and from 20.0% (0-shot) to 61.3% (3-shot) for cross-session testing. This demonstrates the potential of LVLMs for ultrasound-based gesture recognition by enabling an alternative to prior ultrasound gesture pipelines that require dedicated model training and large labeled datasets. Additionally, we show that few-shot in-context learning and retrieval-augmented selection can substantially improve performance without any model fine-tuning.

IEEE Access 2026

Mitsubishi Electric Research Laboratories, Inc.
201 Broadway, Cambridge, Massachusetts 02139

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2024.0429000

GPT Sonography: Hand Gesture Decoding from Forearm Ultrasound Images via a Large Vision-Language Model

KESHAV BIMBRAW[1, 2], (Member, IEEE), YE WANG[2], (Senior Member, IEEE), JING LIU[2], (Member, IEEE), and TOSHIAKI KOIKE-AKINO[2] (Senior Member, IEEE)

[1]Worcester Polytechnic Institute, Worcester, MA 01609, USA (e-mail: kbimbraw@wpi.edu, bimbrawkeshav@gmail.com)
[2]Mitsubishi Electric Research Laboratories (MERL), 201 Broadway, Cambridge, MA 02139, USA (e-mail: {bimbraw, yewang, jiliu, koike}@merl.com)

Corresponding author: Keshav Bimbraw (e-mail: bimbrawkeshav@gmail.com).

This work was conducted while K.B was an intern at MERL.

ABSTRACT Large vision-language models (LVLMs), such as the Generative Pre-trained Transformer 4-omni (GPT-4o), are emerging multi-modal foundation models which have great potential as powerful artificial-intelligence (AI) assistance tools for a myriad of applications, including healthcare, industrial, and academic sectors. Although such foundation models perform well in a wide range of general tasks, their capability without fine-tuning is often limited in specialized tasks. However, full fine-tuning of large foundation models is challenging due to enormous computation/memory/dataset requirements. Ultrasound data from the forearm has been shown to be used for hand gesture estimation. However, this typically requires training deep learning models with a large quantity of labeled data. We show that GPT-4o can decode hand gestures from forearm ultrasound data even with no fine-tuning, and improves with few-shot, retrieval augmented in-context learning. In our experiments, the average classification accuracy improved from 19.3% (0-shot) to 74.0% (2-shot) for within-session testing, and from 20.0% (0-shot) to 61.3% (3-shot) for cross-session testing. This demonstrates the potential of LVLMs for ultrasound-based gesture recognition by enabling an alternative to prior ultrasound gesture pipelines that require dedicated model training and large labeled datasets. Additionally, we show that few-shot in-context learning and retrieval-augmented selection can substantially improve performance without any model fine-tuning.

INDEX TERMS GPT, AI, LMM, LLM, LVLM, VLM, Ultrasound Imaging, Human-Machine Interface, Gesture Recognition.

I. INTRODUCTION

LARGE language models (LLMs) [1], such as generative pre-trained transformers (GPTs) [2], have recently emerged as powerful general assistance tools and exhibited tremendous capabilities in a wide range of applications. LLMs are often configured with billions of parameters to capture linguistic patterns and semantic relationships in natural language processing, enabling text generation, summarization, translation, reasoning, question-answering, etc.

More recently, large multi-modal models (LMMs) [3] with the capability to understand both natural language and other modalities, such as images and sounds, have offered new opportunities for biomedical applications. For example, it was demonstrated that large vision-language models (LVLMs) such as GPT-4o [4] and LLaVa [5] could be a viable tool for medical applications [6], including surgical oncology [7] and radiology diagnosis [8]–[10]. We examine the capabilities of GPT-4o for sonography [11], to analyze and decode ultrasound images.

Musculoskeletal ultrasound is a non-invasive and non-radiative imaging technique that uses ultrasound waves to visualize muscles, tendons, ligaments, and joints. For instance, ultrasound measurements can be used to visualize the anatomical aspects of the forearm, to estimate hand gestures [12], [13]. This is applicable to several domains, such as control of prosthetic hands [14], teleoperation of robotic grippers [15], and controlling virtual reality interfaces [16]. In particular, modern deep learning methods have shown improved performance to estimate different hand gestures [17]. Prior sonomyography work has also reported real-time multi-

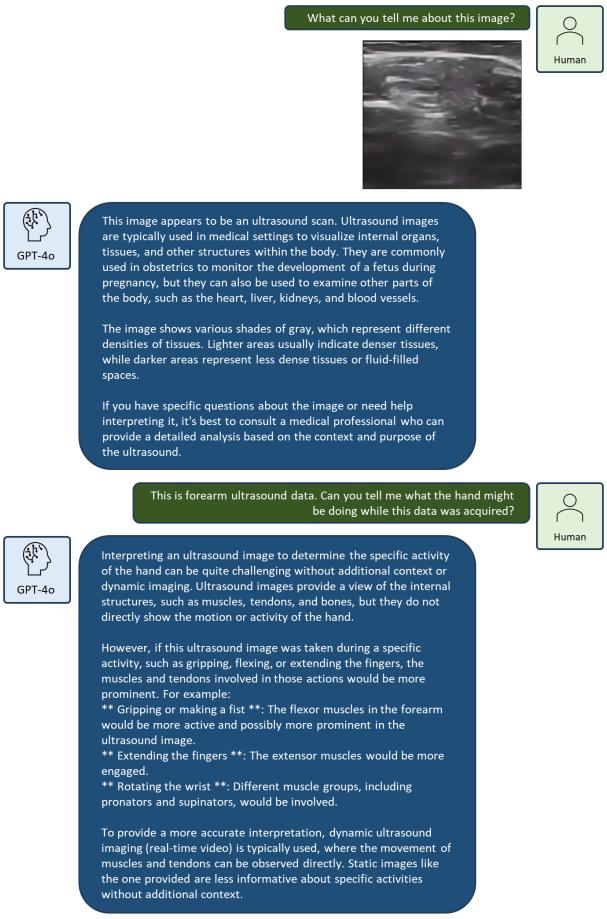


FIGURE 1. Conversation with GPT-4o that motivated us to use the LVLM for ultrasound image decoding.

motion classification from forearm ultrasound and comparative evaluations of ultrasound feature sets and classifiers for prosthesis control [18], [19]. It is highly expected that the use of LVLMs like GPT-4o to classify ultrasound images can provide a lot more information through human readable explanations of the model's predictions, which aids understanding of the reasoning behind gesture recognition. In addition, contextual information can be potentially leveraged to improve the classification performance.

Although pre-trained LVLMs work well for general tasks, their performance is often limited for specialized tasks, such as those involving biomedical datasets. Given such a dataset, fine-tuning can greatly improve performance on downstream tasks. Nevertheless, fine-tuning LVLMs is challenging due to the substantial amount of labelled data required [20]. Additionally, it demands significant computational resources and time. Therefore, it is more practical and cost-effective to consider using the pre-trained LVLMs without fine-tuning but with prompt tuning [21] or in-context learning (ICL) [22]. ICL does not modify the pre-trained LVLMs, but instead adds some task-specific examples to the input context to improve the performance of generating the desired responses. Compared with ICL, prompt tuning still needs significantly more training samples for tuning, and also demands significant more computational resources, memory, and time, which are often prohibitive on edge devices. The prompt tuning also suffers from generalization issues if it's not tuned on the testing subject. In contrast, ICL just needs few-shot examples (could be few-shot demos of each gesture provided by the user during the calibration phase) and does not need any tuning. Further, ICL offers more flexibility in the sense that the user can easily add new classes by providing few-shot examples, while prompt tuning needs re-training/tuning. ICL has less generalization issues if the testing subject can provide few-shot examples during the calibration. Though fine-tuning the whole model may lead to better performance than ICL, it's very costly, and may not be available for closed-source models.

In this work, we show that we can leverage GPT-4o to classify ultrasound images using a few-shot ICL strategy. We demonstrate that providing some labelled examples to the LVLM significantly improves its performance for forearm ultrasound-based gesture recognition. This opens up exciting applications for LVLMs in medical imaging. Specifically, the key departure from prior work is replacing gradient-based training of an ultrasound-specific classifier with a prompt and example driven adaptation of a frozen LVLM, including retrieval-augmented selection. The contributions of this paper are summarized as follows.

- We examine the capability of LVLMs for sonography analysis.
- We use GPT-4o to analyze forearm ultrasound images for hand gesture decoding.
- We demonstrate that GPT-4o can achieve high accuracy of over 70% for cross-sub-session experiments to classify hand gesture even without any fine-tuning.
- We show that the few-shot ICL strategy is substantially effective to improve the classification accuracy.
- We show that retrieval augmented generation (RAG) can help improve the ICL performance.

The remainder of this paper is structured as follows. Section II provides the motivation behind our study, describing initial explorations with GPT-4o for ultrasound-based classification. Section III details the methodology, including data acquisition, model prompting strategies, and in-context learning (ICL). Section IV outlines the experimental setup and evaluation metrics, followed by Section V, which presents the results of within-session and cross-session classification experiments. Section VI discusses key findings, including the impact of different prompting strategies and model reasoning ability. Finally, Section VII concludes the paper.

II. MOTIVATION

LVLMs have the capability to handle tasks that involve both images and texts. They have been utilized for tasks involving image classification, object detection and semantic segmentation with zero or limited in-context learning examples [23]. For image classification, LVLMs have been used for visual question answering [24] and reasoning [25] among other such

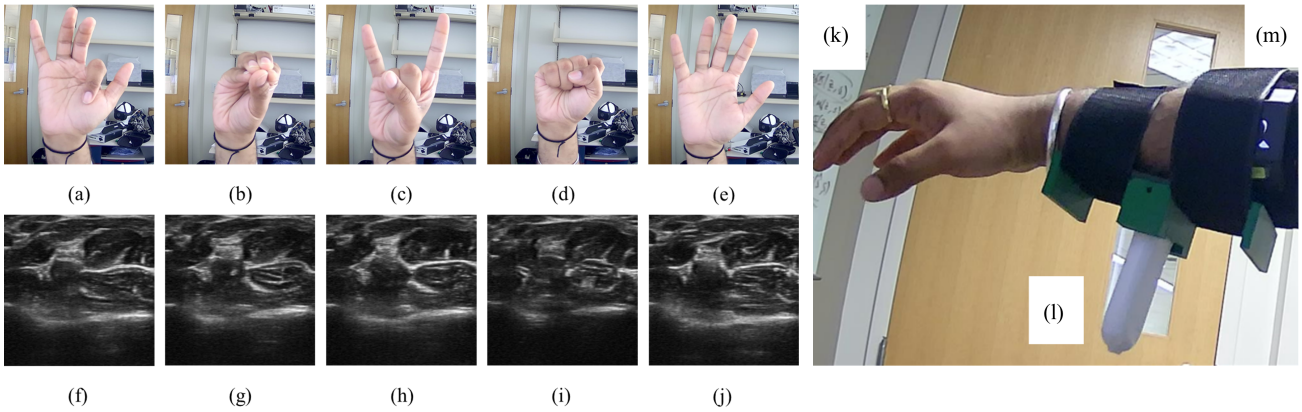


FIGURE 2. **Hand gestures (a through e) and the corresponding forearm ultrasound image (f through j) from subject 1. (a) and (f): Index flexion; (b) and (g): all pinch; (c) and (h) hand horns; (d) and (i) fist; (e) and (j): open hand; The ultrasound probe shown strapped to the forearm with the hand (k), ultrasound probe enclosed within a 3D printed casing (l), and a strap (m).**

TABLE 1. **Subject Information**

	Subject 1	Subject 2	Subject 3
Age	37	29	39
Gender	M	M	M
Height (cm)	176	180	175

applications [3]. They have proven to be useful for understanding medical image data, especially with extensive fine-tuning [8].

Since full fine-tuning of LVLMs requires substantial computational resources, we first examined to see how GPT-4o would perform without fine-tuning. GPT-4o was provided a forearm ultrasound image, and asked a simple question ''What can you tell me about this image?''. The LVLM was able to identify that it is an ultrasound image, and gave some additional information about generic ultrasound images and their visual properties. We then examined whether it could infer some additional information when it is given some context. To this end, a follow-up question was asked: ''This is forearm ultrasound data. Can you tell me what the hand might be doing while this data was acquired?''. The LVLM gave some more information about physiology of hand movement and how different hand movements would lead to different ultrasound images. The full conversation can be seen in Fig. 1. This motivated us to experiment with GPT-4o to see if it could classify forearm ultrasound images corresponding to different hand movements. We are also interested in evaluating its performance while varying the amount of data and context that it is exposed to.

III. METHODOLOGY

For this study, ultrasound data was acquired from 3 subjects. The demographic information is provided in Table 1 (Age: 35 $\pm$ 4.32 years; Height: 177 $\pm$ 2.16 cm; Sex: Male). The study was approved by the institutional research ethics committee (IRB reference number 23001). Written informed consent was given by the subjects before data acquisition. Per subject, data was acquired for 5 hand gestures as shown in Figs. 2: (1) index flexion; (2) all pinch; (3) hand horns; (4) fist; and (5) open hand. These are based on activities of daily living and the chosen gestures are a subset of the dataset in [12]. The five gestures were selected to represent a diverse range of hand movements, since these included a finger flexion, a pinch gesture, an indicative gesture (hand horns), open hand and closed fist gestures, relevant to daily activities. These are the mainstay of daily human interaction, and human manipulation of objects [26], [27]. Limiting the selection to these 5 enabled us to effectively perform 1, 2, and 3-shot learning experiments, given the model's constraint of using up to 19 images for ICL.

A. DATA ACQUISITION

The ultrasound data was acquired using a Sonostar 4L linear palm Doppler ultrasound probe [28]. A custom-designed 3D-printed wearable was strapped onto the subject's forearm. The probe streams brightness mode (B-Mode) ultrasound data from the ultrasound probe. The probe is positioned perpendicular to the forearm positioned in its upper portion. The ultrasound probe (Figure 2(l)) was strapped to the forearm using a custom made 3D printed casing and Velcro straps (Figure 2(m)). The data from the probe was streamed to a Windows system over Wi-Fi, and screenshots of the ultrasound images were captured using a custom Python script. The 4L linear probe has 80 channels of ultrasound data, and the post-processed beamformed B-mode data is obtained, from which 350 $\times$ 350-pixel images are acquired. Beyond the acquisition pipeline that outputs these images, we did not apply additional image preprocessing (e.g., denoising, filtering, segmentation) prior to Base64 encoding.

In this study, a *session* refers to a full data collection procedure conducted on a participant in a single sitting. Each session is further divided into *sub-sessions*, where each sub-session represents a specific segment of the session corresponding to a repeated instance of the gesture sequence. This structure allows for controlled variability in the dataset and facilitates cross-session and cross-sub-session comparisons.

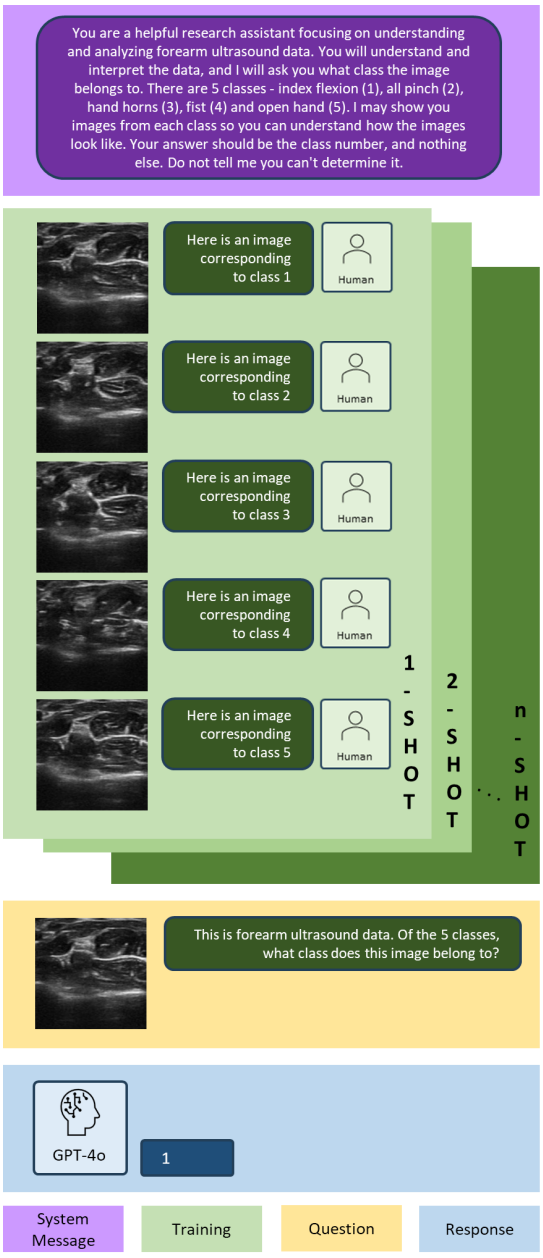


FIGURE 3. Conversation with GPT-4o for forearm ultrasound image classification based on n-shot ICL. For 1, 2, and 3 shot ICL strategy, 1, 2 and 3 samples per class are a part of the prompt, respectively.

For each subject, 5 sessions of data were collected. In each session, subjects performed a sequence of 5 gestures. Within each session, this sequence was repeated 4 times, resulting in 20 sub-sessions. For our study, we analyzed 10 frames per sub-session, resulting in a total of 1000 images (i.e., 20 sub-sessions, 10 frames/sub-session, 5 gestures) per subject.

B. LVLM FOR ULTRASOUND IMAGE CLASSIFICATION

We used GPT-4o [4], a state-of-the-art large vision–language model capable of jointly processing image and text inputs, to perform ultrasound image classification via in-context learning. For inference, Azure OpenAI module within OpenAI's Python library was used [29]. Azure cloud computing was used within a Linux system with Python 3.11 for scripting. Ultrasound images were encoded using Base64 formatting [30] prior to submission to GPT-4o, enabling transmission of image data through the model's multimodal API.

C. GPT-4O PROMPTS

Prior work has shown that large multimodal models can be adapted to new computer vision tasks through prompt-based conditioning and in-context learning, without task-specific fine-tuning [31]–[34]. In this study, we adopt a standard few-shot in-context prompting strategy, where a small number of labeled ultrasound image–class pairs are provided to condition GPT-4o for gesture classification on unseen samples.

To optimize few-shot in-context learning (ICL) for ultrasound image classification, we implemented a retrieval-augmented generation (RAG) [35] approach. This method enhances standard ICL by dynamically selecting the most relevant examples from a labeled dataset to construct each prompt, creating a task-specific context for the large vision-language model (LVLM). Let $\mathcal{D} = \{(x_1, y_1), ..., (x_n, y_n)\}$ denote our labeled dataset, where x_i represents an ultrasound image and y_i its corresponding label. For a given test sample x_t, we compute its similarity to each example in $\mathcal{D}$ using cosine similarity on flattened image vectors as below:

$$\text{sim}(\hat{x_t}, \hat{x_i}) = \frac{\hat{x_t} \cdot \hat{x_i}}{\|\hat{x_t}\|\|\hat{x_i}\|}. \tag{1}$$

Here, $\hat{x_t}$ is the flattened vector of the test image, $\hat{x_i}$ is the flattened vector of a dataset image, $\cdot$ denotes the dot product, and $\|\cdot\|$ represents the Euclidean norm. We select the k most similar examples to the test sample as follows:

$$\mathcal{N}_k(x_t) = \text{top-k}_{(x_i,y_i)\in\mathcal{D}} \text{sim}(\hat{x_t}, \hat{x_i}), \tag{2}$$

where $\mathcal{N}_k(x_t)$ is the set of k most similar examples to x_t, and top-k selects the k highest-scoring pairs according to the similarity function. These selected examples are then used to construct the ICL prompt for GPT-4o. The RAG approach enables dynamic construction of task-relevant contexts and has been used for vision language models in medical data context [36]. By leveraging similarity between the test sample and labeled examples, this method can provide more informative contexts for the LVLM, leading to improved hand gesture classification performance in few-shot ICL scenarios.

The conversation flow we use is described in Fig. 3. We used the exact system message and query text shown in the figure. For 0-shot, we omit the in-context examples and directly ask the query. For 1-/2-/3-shot, we prepend the corresponding number of labeled image–class pairs (as defined in Sec. IV) using the same "training example" phrasing shown in Fig. 3. In all settings, GPT-4o is explicitly constrained to output *only* the class number (1–5) and no additional text. To effectively utilize GPT-4o, we designed the conversation as follows.

1) System Message

We began with a system message to set context and guidelines for the conversation. GPT-4o was informed that it would serve as a helpful research assistant and will assist in classifying hand gestures using forearm ultrasound data.

2) In-Context Learning (ICL)

We used an ICL strategy which provides training examples in contexts. We use a few forearm ultrasound image samples along with the class labels for the in-context examples to assist GPT-4o for specialized classification tasks. Note that ICL does not involve any 'learning' procedure such as fine-tuning, adaptation, or post-training.

3) Query for Classification

The GPT-4o was then asked to predict the hand gesture class based on the given ultrasound image. It was explicitly instructed to provide just the class number, which can be saved for further analysis.

IV. EXPERIMENTAL SETUP

The overall workflow can be visualized in Fig. 4. The workflow begins with data acquisition and preprocessing, followed by in-context learning (ICL) and retrieval-augmented ICL (RAG-ICL) for classification using GPT-4o. Additionally, a nearest neighbor classifier is trained for comparison, and the results are evaluated to assess performance across different strategies.

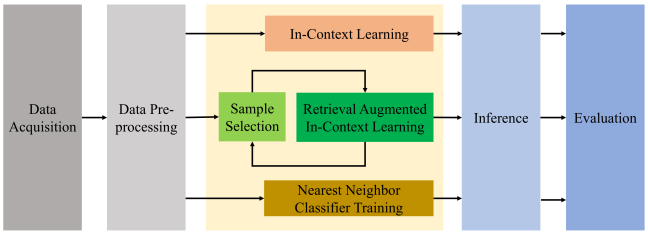


FIGURE 4. Workflow of the proposed ultrasound-based gesture classification approach using GPT-4o, illustrating data acquisition, preprocessing, retrieval augmented in-context learning, inference, and evaluation.

The performance was evaluated with few-shot in-context strategies: 0-shot; 1-shot; 2-shot; and 3-shot ICL. Two experiments were carried out: within-session analysis and cross-session analysis. The within-session analysis was conducted to evaluate the model's performance when there is minimal time difference between the acquisition of samples used for training and evaluation, reflecting a more controlled environment. In contrast, the cross-session analysis aimed to assess the model's robustness and generalization ability when there is a greater time gap between the acquisition of samples, simulating more realistic and varied conditions. For the former, for a given subject, of the 40 images per class in session 1, the last sub-session (last 10 images) were used for evaluation, and the remaining were used for training. For the latter, the last sub-session of session 5 was used for evaluation, while the remaining data was used as ICL training samples. For the three different experiments, different data was used for training and evaluation. For 0-shot strategy, the LVLM was shown images in the test set directly and asked what class out of the 5 it belonged to. This was done to assess its ability to generalize and classify without any prior examples or training on specific data, thereby testing its inherent understanding and adaptability to unseen data. To mitigate overfitting, each sample used for training and evaluation was sufficiently distinct, as different sessions of data were acquired at different points in time, and for both the analyses, the training and testing samples were sufficiently different. The distinction between training and evaluation samples is detailed in the descriptions of the within-session and cross-session analyses.

A. WITHIN-SESSION ANALYSIS

For the 1, 2, and 3-shot strategies, the data-split is described below.

1) 1-Shot

The first image per class from sub-session 1 was shown to the model along with the class label before asking the question. This leads to a total of 5 images and their corresponding class-labels shown. This can be seen in Fig. 3.

2) 2-Shot

The first two images per class from sub-session 1 were shown to the model along with the class labels, leading to a total of 10 images shown.

3) 3-Shot

The first image per class from sub-sessions 1, 2, and 3 were shown to the model along with the class label, leading to a total of 15 images shown.

B. CROSS-SESSION ANALYSIS

For the 1, 2, and 3-shot strategies, the data-split is described below.

1) 1-Shot

The first image per class from sub-session 1 was shown to the model along with the class label before asking the question. This leads to a total of 5 images and their corresponding class-labels shown. The training data shown in similar to the within-session experiment.

2) 2-Shot

The first image per class (sub-session 1) from sessions 1 and 2 were shown to the model along with the class labels, leading to a total of 10 images shown.

3) 3-Shot

The first image from sub-session 1 per class from sessions 1, 2, and 3 were shown to the model along with the class label, leading to a total of 15 images shown.

TABLE 2. Within-session experiment results in percent (with 95% confidence interval in a format of avg$^{\text{upper}}_{\text{lower}}$)

Approach	Accuracy	Precision	Recall	F1 Score
0-shot	19.3 $^{31.4}_{17.9}$	24.7 $^{27.7}_{21.6}$	24.0 $^{26.9}_{21.1}$	24.0 $^{30.8}_{17.2}$
1-shot	68.7 $^{75.5}_{60.9}$	72.0 $^{78.6}_{65.4}$	68.7 $^{73.3}_{64.0}$	68.7 $^{76.1}_{61.2}$
2-shot	75.3 $^{81.5}_{67.9}$	**83.9** $^{88.7}_{79.1}$	75.3 $^{78.0}_{72.6}$	75.3 $^{82.2}_{68.4}$
3-shot	**78.7** $^{84.5}_{71.4}$	**83.9** $^{89.2}_{78.7}$	**78.7** $^{82.9}_{74.4}$	**78.7** $^{85.2}_{72.1}$

TABLE 3. Cross-session experiment results

Approach	Accuracy	Precision	Recall	F1 Score
0-shot	20.7 $^{27.8}_{15.0}$	20.7 $^{23.4}_{17.9}$	20.7 $^{23.4}_{17.9}$	20.7 $^{27.1}_{14.2}$
1-shot	**47.3** $^{55.3}_{39.5}$	—	**47.3** $^{52.1}_{42.6}$	**47.3** $^{55.3}_{39.3}$
2-shot	41.3 $^{49.3}_{33.8}$	—	41.3 $^{42.8}_{39.9}$	41.3 $^{49.2}_{33.5}$
3-shot	35.3 $^{43.3}_{28.1}$	—	35.3 $^{37.2}_{33.4}$	35.4 $^{43.0}_{27.7}$

C. EVALUATION METRICS

To evaluate the performance, the predicted class labels from GPT-4o were compared to the true values. Classification accuracy was used as a metric for evaluating the performance. Confusion matrices were used to visualize the performance for different scenarios. Precision, recall and F1 scores were also calculated for each confusion matrix.

D. RESEARCH QUESTIONS AND EXPERIMENTAL DESIGN

The experiments in this study are designed to address the following key research questions:

- Can GPT-4o classify hand gestures from forearm ultrasound images without fine-tuning?
- How does the classification performance improve with increasing in-context learning (ICL) examples?
- Does retrieval-augmented generation (RAG) enhance the model's classification accuracy?
- How does GPT-4o compare to traditional nearest neighbor classification?

To answer these questions, we conducted within-session and cross-session experiments with 0-shot, 1-shot, 2-shot, and 3-shot ICL strategies. Additionally, we evaluated the effectiveness of RAG and compared GPT-4o's performance to a nearest neighbor classifier. The following sections present the experimental results and analyses.

V. RESULTS

This section provides the results for within-session and cross-session experiments for 0-shot, 1-shot, 2-shot, and 3-shot ICL strategies. The results are described as confusion matrices, classification accuracy, precision, recall and F1 scores. The 95% confidence intervals are also listed along the macro average in the format of $\text{avg}^{\text{upper}}_{\text{lower}}$.

A. WITHIN-SESSION EXPERIMENT

The confusion matrix, summed over the three subjects for the within-session experiment can be seen in Fig. 5 (a)–(d) for 0, 1, 2, and 3-shot strategies respectively.

The classification accuracy, along with the precision, recall, and F1 scores are summarized in Table 2.

For 0-shot strategy, the average classification accuracy was 19.3% (±1.0%). For 1-shot, 2-shot and 3-shot strategies, we achieved 68.7% (±9.0%), 75.3% (±9.0%), and 78.7% (±8.4%) respectively. It clearly demonstrates that in-context examples can significantly improve the classification accuracy even without fine-tuning the pre-trained LVLM. A slight decline of 2 percentage points is observed when the training examples increase from 2 to 3 per class. It may be within a statistical fluctuation due to the small number of test samples. Using RAG for the within-session analysis demonstrated improved performance with increasing shots.

For 1-shot strategy, using RAG led to a mean accuracy of $99.3 \pm 1.2\%$, significantly higher than the ICL baseline's $68.7 \pm 9.0\%$. For the 2-shot strategy, using RAG led to a perfect accuracy at $100.0 \pm 0.0\%$, while ICL shows $75.3 \pm 9.0\%$. Similar trend was observed for the 3-shot strategy, with the mean accuracy at $100.0 \pm 0.0\%$, whereas ICL obtained $78.7 \pm 8.4\%$. Overall, RAG demonstrates robust performance with low variability, while ICL shows less consistency as shot count increases for the within-session experiment. The results are shown in Fig. 6.

It is to be noted that within-session RAG retrieves the top-k most similar labeled examples from the same session's training pool (excluding the held-out last sub-session). Because the test samples are acquired in the same session with minimal time gap and stable probe placement, retrieved exemplars can be highly similar to the query, which can lead to near-ceiling accuracy; this effect is reduced in cross-session (and randomized-selection) settings.

B. CROSS-SESSION EXPERIMENT

The confusion matrix, summed over the three subjects for the cross-session experiment can be seen in Fig. 5 (e)–(h) for 0, 1, 2, and 3-shot strategies respectively. The classification accuracy, along with the precision, recall, and F1 scores are summarized in Table 3.

For 0-shot case , the classification accuracy is comparable to a random guess because of 5 classes. For 1-shot strategy, it was 52%. For 2-shot, it was 56%, which increased to 70% for 3-shot case. This trend is encouraging since increasing the number of in-context samples can improve the performance of GPT-4o to classify forearm ultrasound images to predict the hand gestures they correspond.

This was repeated for subjects 2 and 3. Fig. 7 shows the classification results averaged over the three subjects. For 0-shot strategy, the average classification accuracy was 20.0% (±0.0%). For 1-shot, 2-shot and 3-shot strategies, it was obtained to be 33.3% (±16.7%), 51.3% (±15.5%), and 61.3% (±22.3%) respectively. These results show a clear improvement in the classifier performance for an increasing number of in-context samples. It was interesting to observe that the standard deviation increases sharply as the number of training examples increases from 2 to 3 per class.

The results for the case where the input samples were picked randomly from the training data are shown in Fig. 7.

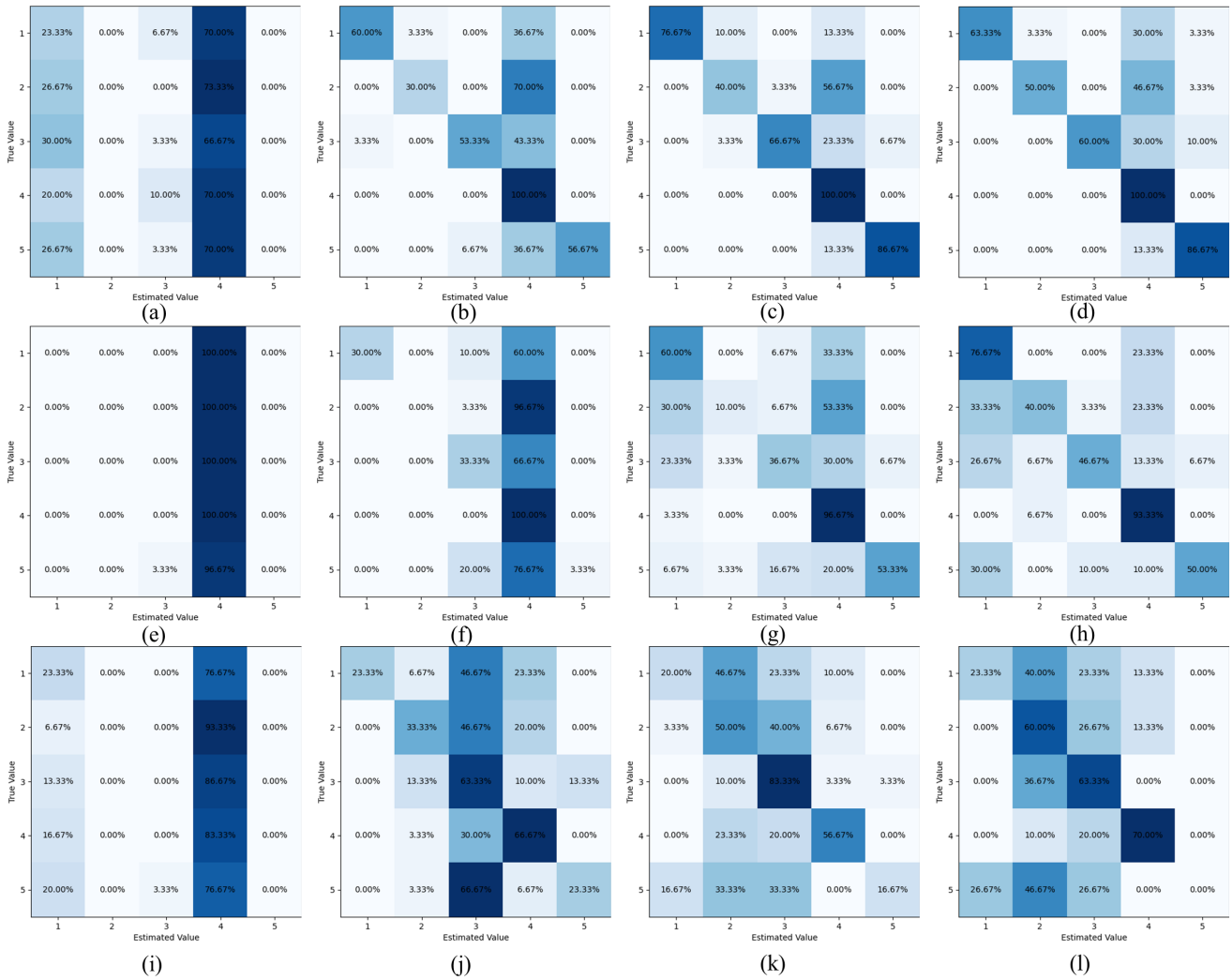


FIGURE 5. Confusion matrices for within-session (a–d), cross-session (e–h), and randomized cross-session (i–l) experiments summed over the three subjects for: 0-shot (a, e, and i), 1-shot (b, f, and j), 2-shot (c, g, and k), and 3-shot (d, h, and l) strategies.

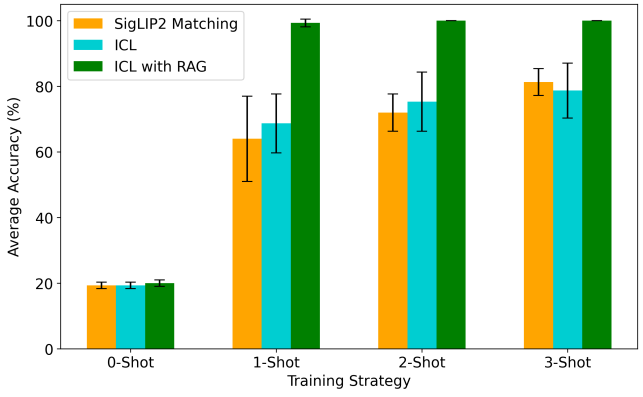


FIGURE 6. Average classification accuracy for within-session experiment. ICL with RAG performed the best across the different number of training images per class provided.

While the performance with in-context learning was better than 0-shot case, it was worse than non-randomized case. Increasing the number of training samples did not clearly improve the average classification across subjects. For 0-shot strategy, the classification accuracy was 21.3% (±3.0%). For 1-shot strategy, the average classification accuracy was 42.0% (±10.6%). For 2-shot strategy, the average classification accuracy was 45.3% (±5.0%). And for 3-shot strategy, the average classification accuracy was 43.3% (±4.2%).

Using RAG for the cross-session analysis demonstrated improved performance, but with a higher standard deviation. For the 1-shot strategy, RAG achieved a mean accuracy of $42.7 \pm 20\%$, higher than the Baseline's $33.3 \pm 16.7\%$. For the 2-shot strategy, RAG led to a mean accuracy of $59.3 \pm 30.7\%$, while the baseline ICL showed $51.3 \pm 15.5\%$. The mean accuracy percentage was the same for the RAG approach for 3-shot strategy, with mean accuracy at $61.3 \pm 22\%$, matching the baseline ICL accuracy of $61.3 \pm 22.4\%$. Overall, using RAG for ICL in this context demonstrates consistent improvement, while the baseline ICL exhibits less variability and performance gains as shot count increases. Unlike traditional

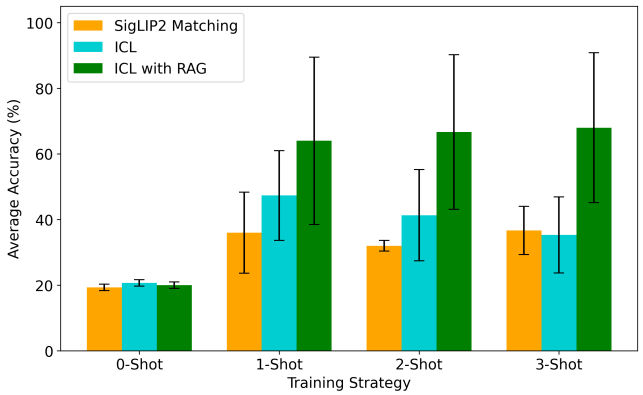


FIGURE 7. Average classification accuracy for cross-session experiment. ICL with RAG outperformed the other methods.

deep learning models that require extensive retraining, our approach demonstrates that GPT-4o can generalize across different sessions of ultrasound image acquisition by leveraging RAG to dynamically construct task-specific prompts. This adaptability allows GPT-4o to incorporate new data without modifying the base model, making it a more flexible and scalable alternative for ultrasound-based biomedical applications.

C. COMPARISON WITH MATCHING NETWORKS

For baseline performance without using LVLM, we consider matching networks [37], which uses metric-based few-shot learning (FSL) classification by finding the closest similarity between few-shot samples and the query sample through matching function. We consider cosine-similarity for two metrics: piece-wise pixel values; and SigLIP2 [38] embedding. SigLIP2 is a state-of-the-art embedding method, recently proposed to enhance the sigmoid contrastive learning. We use `google/siglip2-base-patch16-224` in huggingface transformers library, which is based on vision transformer (ViT) having 86M parameters. The results are shown in Fig. 6 and Fig. 7. For within-session, we confirm that ICL with GPT-4o achieves a competitive performance to the baseline FSL method with the state-of-the-art SigLIP2. Notably, the proposed ICL with RAG achieves far better performance than those FSL baselines. For cross-session, pixel-matching nearest neighbor method achieves its best performance under 2-shot, but still worse than the performance of the proposed ICL with RAG using 2 or 3 shots.

VI. DISCUSSION

Several additional experiments were carried out for within-session data from subject 1 to understand GPT-4o's performance and reasoning. All these experiments were done for a 1-shot strategy. The baseline confusion matrix is shown in Fig. 9. For this case, the accuracy is 86%, with the macro average precision, recall, and F1 scores being 0.9, 0.86, and 0.85, respectively.

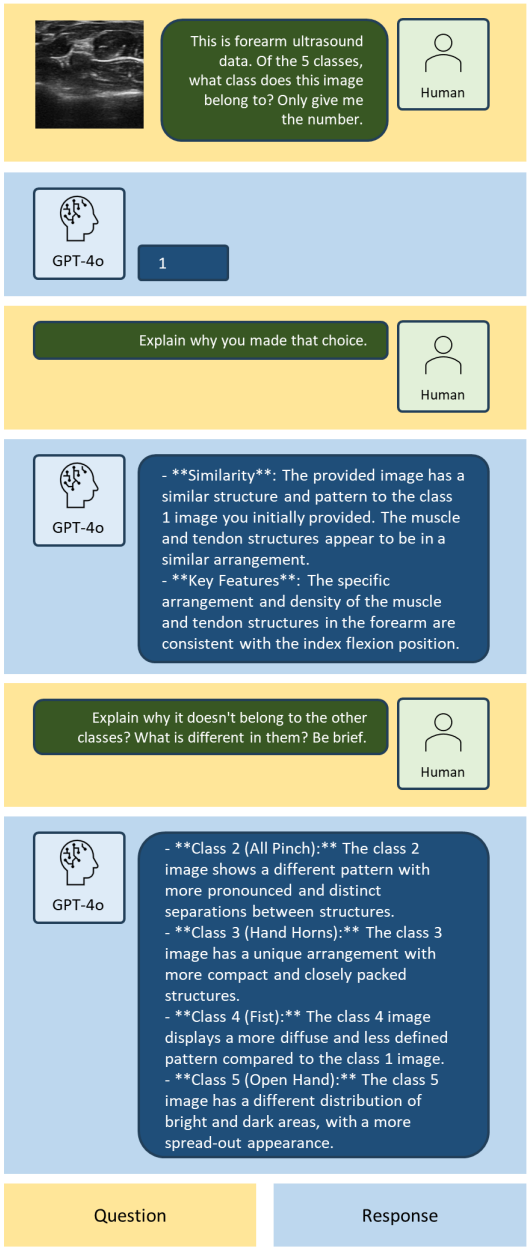


FIGURE 8. Conversation with GPT-4o as a follow up to the 1-shot conversation in Fig. 3 to demonstrate its reasoning capabilities.

A. RESULTS WITH DIFFERENT PROMPTS

We wanted to see how GPT-4o would perform with prompts less and more descriptive than the prompts shown in Fig. 3. These results should be interpreted with caution, as the prompt-ablation analysis was conducted on within-session data from a single subject. While this controlled setting enables focused analysis of prompting effects, the limited sample size may introduce subject-specific bias and does not fully capture inter-subject variability.

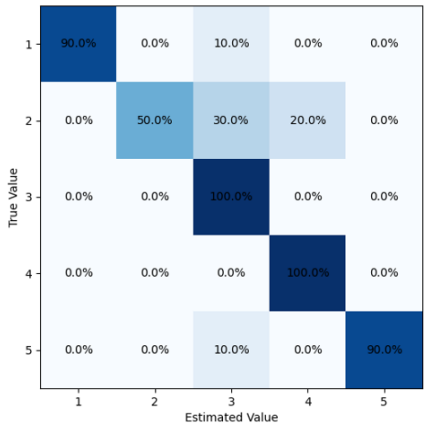


FIGURE 9. Baseline confusion matrix (within-session, subject 1, 1-shot). Accuracy: 86%.

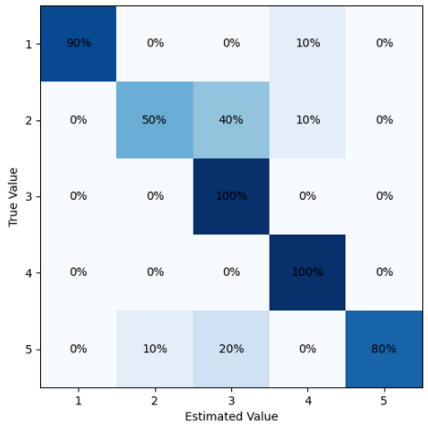


FIGURE 10. Low-descriptive prompt (within-session, subject 1, 1-shot). Accuracy: 82%.

1) Less descriptive information

For this experiment, we did not provide the system message. And for training, we only stated the class label with the image. As the question, we just asked 'What class does the image belong to? Only give the class number.' With this minimal information, the confusion matrix obtained is shown in Fig. 10. For this case, the accuracy is 82%, with the macro average precision, recall, and F1 scores being 0.86, 0.82, and 0.82, respectively. It was interesting to see that there was only a decline of 4% in the classification accuracy from the baseline of Fig. 9, meaning that we can provide it a lot less information without compromising significantly on the accuracy.

2) More descriptive information

For this experiment, we provided a lot more contextual information to GPT-4o both in the system message, as well as in the final question. We mentioned that it should focus on the arrangement of regions with different brightness. We also mentioned that the anatomical and physiological properties visualized in the ultrasound image are distinct for different hand gestures. The confusion matrix is shown in Fig. 11. For this case, the accuracy is 80%, with the macro average precision, recall, and F1 scores being 0.87, 0.8, and 0.8, respectively.

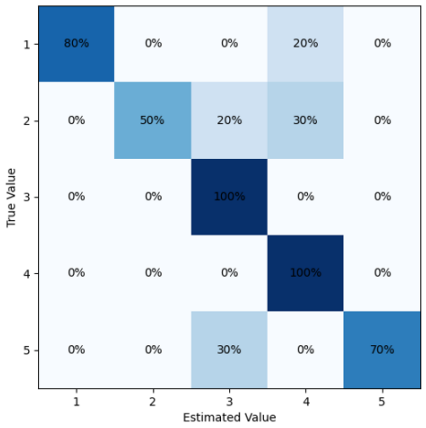


FIGURE 11. High-descriptive prompt (within-session, subject 1, 1-shot). Accuracy: 80%.

It was interesting to see that providing so much extra information did not really help improve the performance. Rather, it decreased the performance compared to the less descriptive information case by 2%.

B. REASONING ABILITY

With the flow shown in Fig. 3, we wanted to understand why GPT-4o made that particular estimation. Fig. 8 shows the user asking questions to GPT-4o, and it answering why it made that particular estimation compared to the other classes. Based on this conversation, we can make the following conclusions.

1) Logical Coherence

GPT-4o demonstrates a structured approach to reasoning, with each successive step logically following the previous one. This indicates an ability to maintain logical consistency.

2) Contextual Understanding

The model incorporates context into its reasoning, ensuring that decisions are relevant to the given scenario. It takes into consideration the information provided during training, as well as in the system message.

3) Decision-Making

GPT-4o was able to express why the image does not belong to the other classes. It provides a clear distinction between the different classes. For example, for class 5 (open hand), the model stated that the image showed a different distribution of bright and dark regions with a more spread-out appearance,

and therefore did not belong to class 5. While the model's reasoning is not fully trustworthy and LVLMs are prone to hallucinations [39], it is encouraging to see that LVLMs like GPT-4o can be used to understand better why it made a particular prediction. More effective conversations with contextual clues may improve its performance.

C. DIFFERENT INPUT FORMATS

Radiologists often look at stacked medical images to understand medical image data. This is done especially with time-varying data to visualize how the physiological features change with time [40]. We wanted to see how GPT-4o would perform for different stacks of ultrasound images. Fig. 12 shows a stacked image sample with 4 ultrasound image frames.

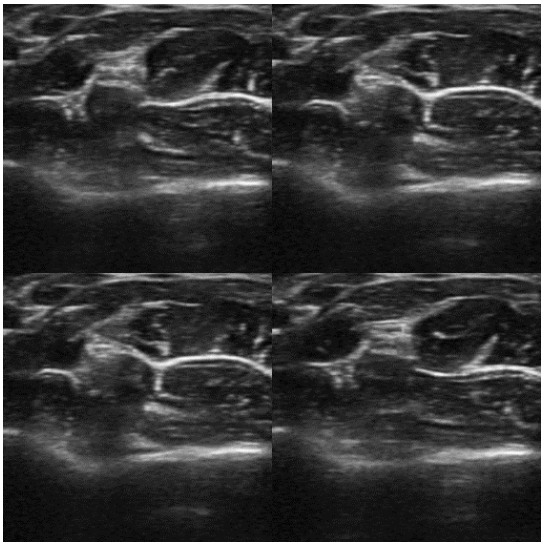


FIGURE 12. Stacked ultrasound images for class 1 with ultrasound image frames taken at different times.

1) Two images as input

Using two stacked ultrasound frames as input for 1-shot strategy, instead of one image per class, 1 image with two ultrasound frames corresponding to the class were shown. This can be visualized in the top row of Fig. 12. The classification results are shown in Fig. 13.

For this case, the accuracy is 78%, with the macro average precision, recall, and F1 scores being 0.83, 0.78, and 0.77, respectively.

2) Four images as input

Using 4 stacked ultrasound frames as input for 1-shot strategy, instead of one image per class, 1 image with 4 ultrasound frames corresponding to the class were shown. This can be visualized in Fig. 12. The classification results are shown in Fig. 13. For this case, the accuracy is 72%, with the macro average precision, recall, and F1 scores being 0.84, 0.72, and 0.68, respectively.

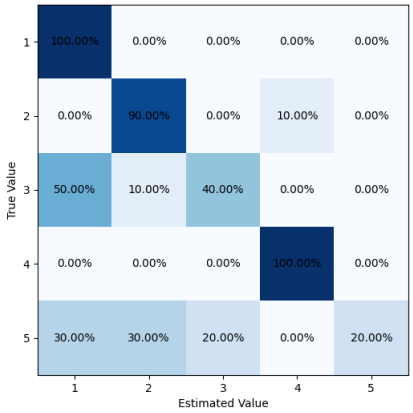


FIGURE 13. Stacked 2-frame confusion matrix. Accuracy: 78%.

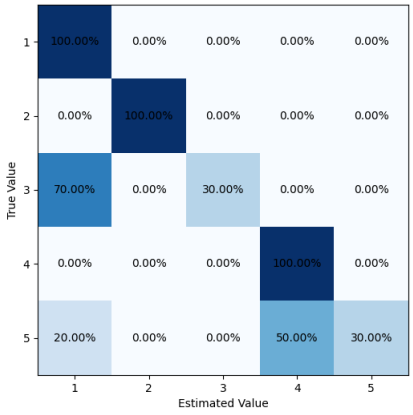


FIGURE 14. Stacked 4-frame confusion matrix. Accuracy: 72%.

Although more training samples are provided by stacking frames, the classification accuracy was degraded. It may be because the image format is different for the testing image and the relative image resolution is lower when stacked. We believe that the performance can be improved by better designing prompts.

D. GPT MODEL COMPARISONS

Figure 15 reports the average classification accuracies of different GPT variants for 1-shot ICL, with results averaged over the subjects. Several clear trends emerge. First, larger models consistently outperform their smaller counterparts, while nano-scale variants struggle to rise above near-chance performance. Second, variance across runs is also correlated with size, with larger models yielding more stable outcomes.

Examining individual models, GPT-5-chat (∼1.3T parameters) and GPT-4.1 (∼250B) achieved the highest accuracies, reaching 75–80% on average. GPT-4o (∼200B) followed closely, delivering over 70% accuracy while exhibiting rel-

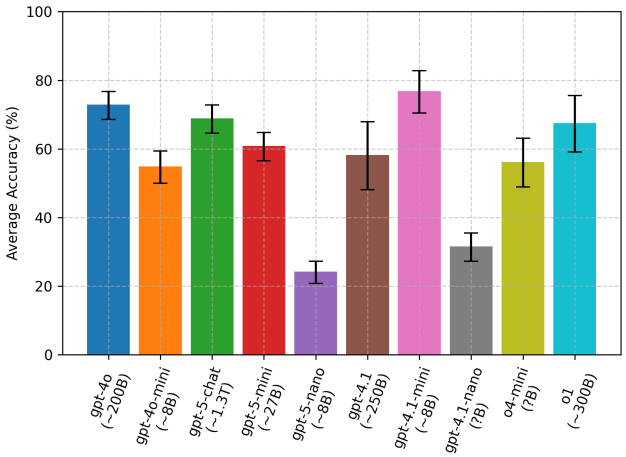


FIGURE 15. Average classification accuracy of GPT variants for ultrasound-based hand gesture recognition for 1-shot ICL averaged over the subjects. Larger models (GPT-5-chat, GPT-4.1, GPT-4o) exceed 70% with lower variance, while mini models (∼8–27B) reach 50–60% and nano models (<10B) remain near chance.

atively low variability. By contrast, mini-scale models such as GPT-5-mini (∼27B) and GPT-4o-mini (∼8B) achieved moderate performance in the 50–60% range, while nano-scale models (e.g., GPT-5-nano, GPT-4.1-nano) plateaued at 20–30%, comparable to random guessing. The error bars in Fig. 15 further highlight that larger models not only achieve higher mean accuracy but also offer more consistent predictions across subjects.

Although GPT-5-chat and GPT-4.1 offer marginally higher accuracy, their extreme size and computational demands limit their practicality. GPT-4o, at ∼200B parameters, provides a favorable trade-off between accuracy, stability, and efficiency, making it a realistic candidate for ultrasound-based human-machine interfacing without the prohibitive cost of trillion-parameter models. A further limitation of this study is its reliance on closed-source LVLMs, which restricts access to internal representations and training details. While this limits transparency and reproducibility at the model level, our evaluation focuses on observable input–output behavior under controlled prompting and data conditions, which remains reproducible across runs and platforms.

E. FUTURE WORK

We conducted experiments to understand capabilities of GPT-4o for hand gesture classification based on forearm ultrasound data. We explored the combination of ICL and RAG for this task. Future work will involve fine-tuning open-source LVLMs such as LLaVA [41] for comparison. Additionally, we plan to conduct extensive cross validation analysis, in addition to acquiring data from more subjects. More rigorous prompt engineering should be considered as well. We are also interested in exploring LVLM's cross-subject generalizability for medical image datasets. In addition, the comparison to parameter efficient fine-tuning (PEFT) [42] methods should follow. Future work will also evaluate cross-subject generalization and robustness to acquisition variations (probe placement/orientation and gain settings), which are known to be challenging in forearm ultrasound gesture recognition even for conventional deep models.

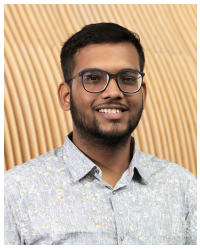
VII. CONCLUSIONS

In this work, we show that we can use a large vision-language model (LVLMs), GPT-4o as a powerful AI assistance tool for understanding and interpreting forearm ultrasound data. We show that by providing some examples of ultrasound images, we can improve its performance for hand gesture classification based on forearm ultrasound data. For within-session performance, we show that the average gesture classification accuracy reached 74.0% for 5 hand gestures with just 2 training samples, and for cross-session performance, it reached 61.3% for just 3 training samples per class. Using retrieval augmented generation (RAG), the within session classification performance reached 100.0% for 2 and 3 training samples per class. Our approach can be used in cases where full fine-tuning of these models is challenging because of enormous compute/memory/dataset requirements. This research opens up exciting avenues for research in utilizing large vision-language models for medical imaging.

REFERENCES

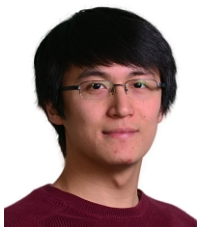
[1] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, *et al.*, "A survey of large language models," *arXiv preprint arXiv:2303.18223*, 2023.
[2] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," OpenAI, San Francisco, CA, USA, 2018.
[3] X. Wang, G. Chen, G. Qian, P. Gao, X.-Y. Wei, Y. Wang, Y. Tian, and W. Gao, "Large-scale multi-modal pre-trained models: A comprehensive survey," *Machine Intelligence Research*, vol. 20, no. 4, pp. 447–482, 2023.
[4] S. Shahriar, B. D. Lund, N. R. Mannuru, M. A. Arshad, K. Hayawi, R. V. K. Bevara, A. Mannuru, and L. Batool, "Putting GPT-4o to the Sword: A Comprehensive Evaluation of Language, Vision, Speech, and Multimodal Proficiency," *Preprints*, 2024.
[5] H. Liu, C. Li, Y. Li, and Y. J. Lee, "Improved baselines with visual instruction tuning," in *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 26296–26306.
[6] N. Zhang, Z. Sun, Y. Xie, H. Wu, and C. Li, "The latest version ChatGPT powered by GPT-4o: what will it bring to the medical field?" *International Journal of Surgery*, 2024.
[7] N. Zhu, N. Zhang, Q. Shao, K. Cheng, and H. Wu, "OpenAI's GPT-4o in surgical oncology: revolutionary advances in generative artificial intelligence," *European Journal of Cancer*, 2024.
[8] Y. Sonoda, R. Kurokawa, Y. Nakamura, J. Kanzawa, M. Kurokawa, Y. Ohizumi, W. Gonoi, and O. Abe, "Diagnostic Performances of GPT-4o, Claude 3 Opus, and Gemini 1.5 Pro in Radiology's Diagnosis Please Cases," *medRxiv*, 2024.
[9] T. Oura, H. Tatekawa, D. Horiuchi, S. Matsushita, H. Takita, N. Atsukawa, Y. Mitsuyama, A. Yoshida, K. Murai, R. Tanaka, *et al.*, "Diagnostic Accuracy of Vision-Language Models on Japanese Diagnostic Radiology, Nuclear Medicine, and Interventional Radiology Specialty Board Examinations," *medRxiv*, 2024.
[10] T. Cesur, Y. C. Gunes, E. Camur, and M. Dagli, "Empowering Radiologists with ChatGPT-4o: Comparative Evaluation of Large Language Models and Radiologists in Cardiac Cases," *medRxiv*, 2024.
[11] F. W. Kremkau, *Sonography: Principles and Instruments*. Elsevier Health Sciences, 2015.
[12] K. Bimbraw, C. J. Nycz, M. Schueler, Z. Zhang, and H. K. Zhang, "Simultaneous estimation of hand configurations and finger joint angles using forearm ultrasound," *IEEE Trans. on Medical Robotics and Bionics*, vol. 5, no. 1, pp. 120–132, 2023.

[13] J. McIntosh, A. Marzo, M. Fraser, and C. Phillips, "Echoflex: Hand gesture recognition using ultrasound imaging," in *Proc. 2017 CHI Conf. on Human Factors in Computing Systems*, 2017, pp. 1923–1934.
[14] Z. Yin, H. Chen, X. Yang, Y. Liu, N. Zhang, J. Meng, and H. Liu, "A wearable ultrasound interface for prosthetic hand control," *IEEE J. of Biomedical and Health Informatics*, vol. 26, no. 11, pp. 5384–5393, 2022.
[15] K. Bimbraw, E. Fox, G. Weinberg, and F. L. Hammond, "Towards sonomyography-based real-time control of powered prosthesis grasp synergies," in *Proc. 42nd Ann. Int. Conf. of the IEEE Eng. in Medicine & Biology Society (EMBC)*, 2020, pp. 4753–4757.
[16] K. Bimbraw, J. Rothenberg, and H. Zhang, "Leveraging Ultrasound Sensing for Virtual Object Manipulation in Immersive Environments," in *Proc. IEEE 19th Int. Conf. on Body Sensor Networks (BSN)*, 2023, pp. 1–4.
[17] K. Bimbraw and H. K. Zhang, "Mirror-based ultrasound system for hand gesture classification through convolutional neural network and vision transformer," in *Medical Imaging 2024: Ultrasonic Imaging and Tomography*, vol. 12932, SPIE, 2024, pp. 218–222.
[18] N. Akhlaghi, C. A. Baker, M. Lahlou, H. Zafar, K. G. Murthy, H. S. Rangwala, J. Kosecka, W. M. Joiner, J. J. Pancrazio, and S. Sikdar, "Real-Time Classification of Hand Motions Using Ultrasound Imaging of Forearm Muscles," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 8, pp. 1687–1698, Aug. 2016, doi: 10.1109/TBME.2015.2498124.
[19] V. Ortenzi, S. Tarantino, C. Castellini, and C. Cipriani, "Ultrasound Imaging for Hand Prosthesis Control: a Comparative Study of Features and Classification Methods," in *Proc. IEEE International Conference on Rehabilitation Robotics (ICORR)*, Singapore, Aug. 2015, pp. 1–6, doi: 10.1109/ICORR.2015.7281166.
[20] Y. Zhai, H. Bai, Z. Lin, J. Pan, S. Tong, Y. Zhou, A. Suhr, S. Xie, Y. LeCun, Y. Ma, *et al.*, "Fine-Tuning Large Vision-Language Models as Decision-Making Agents via Reinforcement Learning," *arXiv preprint arXiv:2405.10292*, 2024.
[21] B. Lester, R. Al-Rfou, and N. Constant, "The power of scale for parameter-efficient prompt tuning," *arXiv preprint arXiv:2104.08691*, 2021.
[22] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, *et al.*, "Language models are few-shot learners," *Advances in Neural Information Processing Systems*, vol. 33, pp. 1877–1901, 2020.
[23] J. Zhang, J. Huang, S. Jin, and S. Lu, "Vision-language models for vision tasks: A survey," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2024.
[24] M. Wang, A. Mahjoubfar, and A. Joshi, "FashionVQA: A Domain-Specific Visual Question Answering System," in *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 3514–3519.
[25] G. Zhang, Y. Zhang, K. Zhang, and V. Tresp, "Can Vision-Language Models be a Good Guesser? Exploring VLMs for Times and Location Reasoning," in *Proc. IEEE/CVF Winter Conf. on Applications of Computer Vision (WACV)*, 2024, pp. 636–645.
[26] A. M. Dollar, "Classifying human hand use and the activities of daily living," in *The Human Hand as an Inspiration for Robot Hand Development*. Springer, 2014, pp. 201–216.
[27] A. Saudabayev, Z. Rysbek, R. Khassenova, and H. A. Varol, "Human grasping database for activities of daily living with depth, color and kinematic data streams," *Scientific Data*, vol. 5, no. 1, pp. 1–13, 2018.
[28] Sonostar, "4L linear palm Doppler ultrasound probe," 2024. Online. Available: http://sonostarmed.com/PalmUS/839.html.
[29] "GitHub - openai/openai-python: The official Python library for the OpenAI API," Online. Available: https://github.com/openai/openai-python. [Accessed 03-07-2024].
[30] "Uploading base-64 encoded images," Online. Available: https://platform.openai.com/docs/guides/vision/uploading-base-64-encoded-images. [Accessed 03-07-2024].
[31] A. Bar, Y. Gandelsman, T. Darrell, A. Globerson, and A. A. Efros, "Visual prompting via image inpainting," *Advances in Neural Information Processing Systems*, vol. 35, pp. 25005–25017, 2022.
[32] X. Wang, W. Wang, Y. Cao, C. Shen, and T. Huang, "Images speak in images: A generalist painter for in-context visual learning," in *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 6830–6839.
[33] J. Zhang, B. Wang, L. Li, Y. Nakashima, and H. Nagahara, "Instruct me more! random prompting for visual in-context learning," in *Proc. IEEE/CVF Winter Conf. on Applications of Computer Vision (WACV)*, 2024, pp. 2597–2606.
[34] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, *et al.*, "Flamingo: a visual language model for few-shot learning," *Advances in Neural Information Processing Systems*, vol. 35, pp. 23716–23736, 2022.
[35] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, and H. Wang, "Retrieval-augmented generation for large language models: A survey," *arXiv preprint arXiv:2312.10997*, 2023.
[36] P. Xia, K. Zhu, H. Li, H. Zhu, Y. Li, G. Li, L. Zhang, and H. Yao, "RULE: Reliable Multimodal RAG for Factuality in Medical Vision Language Models," *arXiv preprint arXiv:2407.05131*, 2024.
[37] O. Vinyals, C. Blundell, T. Lillicrap, and D. Wierstra, "Matching networks for one shot learning," Advances in neural information processing systems, 2016.
[38] M. Tschannen, A. Gritsenko, X. Wang, M.F. Naeem, I. Alabdulmohsin, N. Parthasarathy, T. Evans, L. Beyer, Y. Xia, B. Mustafa, O. Hénaff, "SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features," arXiv preprint arXiv:2502.14786. 2025.
[39] H. Liu, W. Xue, Y. Chen, D. Chen, X. Zhao, K. Wang, L. Hou, R. Li, and W. Peng, "A survey on hallucination in large vision-language models," *arXiv preprint arXiv:2402.00253*, 2024.
[40] F. Gaillard, "Stacks | Radiology Reference Article | Radiopaedia.org," Online. Available: https://radiopaedia.org/articles/stacks?lang=us. [Accessed 05-07-2024].
[41] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *Advances in Neural Information Processing Systems*, vol. 36, 2024.
[42] Z. Han, C. Gao, J. Liu, S. Q. Zhang, *et al.*, "Parameter-efficient fine-tuning for large models: A comprehensive survey," *arXiv preprint arXiv:2403.14608*, 2024.



KESHAV BIMBRAW (Member, IEEE), received the B.E. degree in mechatronics engineering from Thapar University, Patiala, Punjab, India, in 2017, and M.S. degree in computer software and media applications from Georgia Institute of Technology, Atlanta, Georgia, USA.

He is pursuing a PhD in robotics engineering at Worcester Polytechnic Institute, Worcester, Massachusetts, USA. During his PhD, he interned at Nokia Bell Labs, New Providence, New Jersey, USA, and Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA.



YE WANG (Senior Member, IEEE) received the B.S. degree in electrical and computer engineering from Worcester Polytechnic Institute, Worcester, MA, USA, in 2005, and the M.S. and Ph.D. degrees in electrical and computer engineering from Boston University, Boston, MA, USA, in 2009 and 2011, respectively.

In 2012, he joined Mitsubishi Electric Research Laboratories, Cambridge, MA, USA, where he had also previously completed an internship in 2010. His research interests are information theory, machine learning, signal processing, communications, and data privacy/security.

JING LIU (Member, IEEE), received the B.E. degree in electronic engineering from BIT, Beijing, China, in 2010, M.S. degree in electronic engineering from Tsinghua University, Beijing, China, in 2013, and Ph.D. degree in electrical and computer engineering at the University of California, San Diego, CA, USA, in 2019.

He was a Postdoctoral Research Associate at the Coordinated Science Lab of University of Illinois Urbana-Champaign (UIUC) and an Illinois Future Faculty fellow at the Computer Science department of UIUC during 2019 to 2022. He has been a researcher at Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA since 2022.

TOSHIAKI-KOIKE AKINO (Senior Member, IEEE), received the Ph.D. degree from Kyoto University, Kyoto, Japan, in 2005. During 2006-2010, he was a Postdoctoral Researcher at Harvard University, Cambridge, MA, USA, and he is currently Distinguished Research Scientist at MERL, Cambridge, MA, USA.

He was the recipient of the 2008 Ericsson Young Scientist Award, the IEEE GLOBECOM'08 Best Paper Award, the 24th TELECOM System Technology Encouragement Award, and the IEEE GLOBECOM'09 Best Paper Award. He is a Fellow of Optica (formerly OSA).