WatchThis: A Wearable Point-and-Ask Interface powered by Vision-Language Models for Contextual Queries

Cathy Mengying Fang
catfang@mit.edu
MIT Media Lab
Cambridge, USA

Patrick Chwalek
chwalek@mit.edu
MIT Media Lab
Cambridge, USA

Quincy Kuang
quincyku@media.mit.edu
MIT Media Lab
Cambridge, USA

Pattie Maes
pattie@media.mit.edu
MIT Media Lab
Cambridge, USA

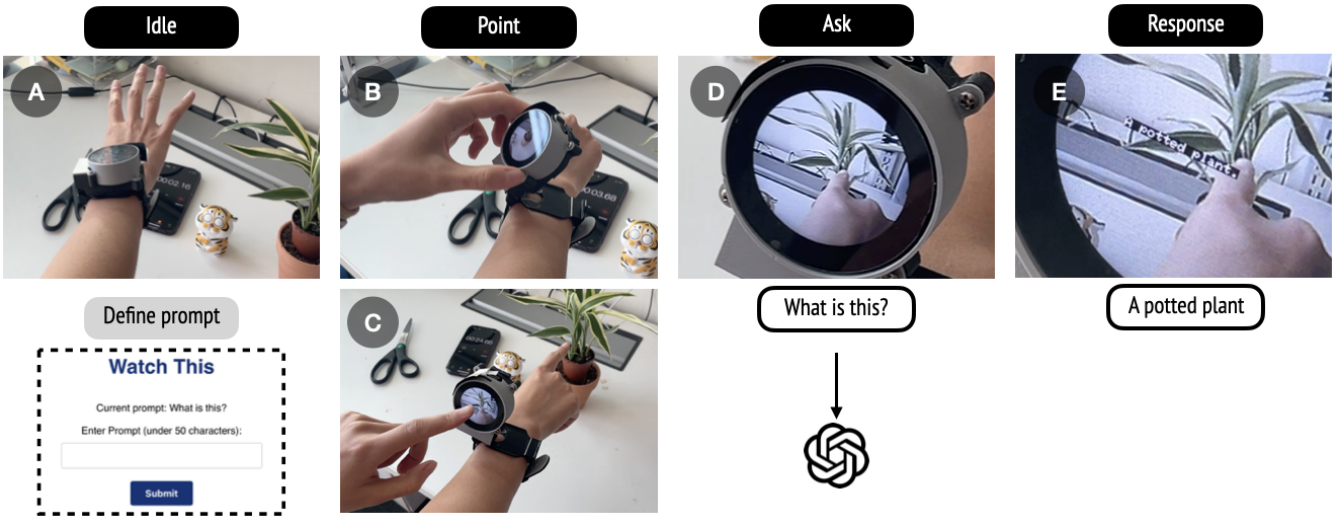

Figure 1: (A) WatchThis is a fully portable device in a watch form factor. (B) The user flips up the device when they encounter an object of interest. (C) The user points at the object with the watch-wearing hand. They see the watch captures along the direction of the finger. (D) The default prompt is can be defined using the accompanying WebApp, and the prompt and the captured image are sent to a VLM such as OpenAI's GPT-4o. E: The user receives a response directly on the watch screen.

ABSTRACT

This paper introduces WatchThis, a novel wearable device that enables natural language interactions with real-world objects and environments through pointing gestures. Building upon previous work in gesture-based computing interfaces, WatchThis leverages recent advancements in Large Language Models (LLM) and Vision Language Models (VLM) to create a hands-free, contextual querying system. The prototype consists of a wearable watch with a rotating, flip-up camera that captures the area of interest when pointing, allowing users to ask questions about their surroundings in natural language. This design addresses limitations of existing systems that require specific commands or occupy the hands, while also maintaining a non-discrete form factor for social awareness. The paper explores various applications of this point-and-ask interaction, including object identification, translation, and instruction queries. By utilizing off-the-shelf components and open-sourcing the design, this work aims to facilitate further research and development in wearable, AI-enabled interaction paradigms.

UIST Adjunct '24, October 13–16, 2024, Pittsburgh, PA, USA

ACM ISBN 979-8-4007-0718-6/24/10.
https://doi.org/10.1145/3672539.3686776

CCS CONCEPTS

• **Human-centered computing** → **Mobile devices**; **Gestural input**.

KEYWORDS

wearable, camera, watch, pointing, vision language model

ACM Reference Format:
Cathy Mengying Fang, Patrick Chwalek, Quincy Kuang, and Pattie Maes. 2024. WatchThis: A Wearable Point-and-Ask Interface powered by Vision-Language Models for Contextual Queries. In *The 37th Annual ACM Symposium on User Interface Software and Technology (UIST Adjunct '24), October 13–16, 2024, Pittsburgh, PA, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3672539.3686776

1 INTRODUCTION

Pointing is a natural human way to communicate intent and highlight an object or area of interest. However, pointing (in particular pointing in mid-air) is not often used as an input gesture to interact with computing devices. The canonical work, Put-That-There [1],

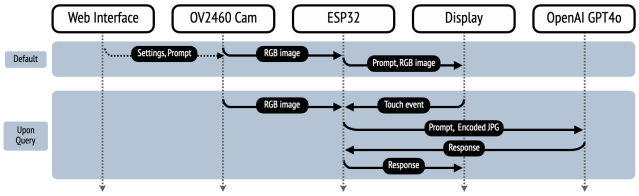

Figure 2: System Diagram

envisions a voice-gesture interface with a "media room" prototype that enables direct pointing of distant objects on a screen and the use of demonstrative pronouns such as "this" and "that" to complete a query. Bringing this vision into daily life, recent works leverage the depth camera and front camera on smartphones to capture the user's pointing gesture of the free hand [2] or their gaze [3] while the other hand holds up the smartphone. Others combined a mobile phone with a separate display worn on the finger that enhances AR interaction [5]. EyeRing [4], a finger-worn camera, assists primarily visually-impaired users with day-to-day tasks such as recognizing currency and navigation.

These implementations either require specific pre-set commands or machine learning algorithms to complete the specific requested tasks. With the recent advancements in Large Language Models (LLM) and Vision Language Models (VLM), individuals can ask questions and receive answers in natural language. This capability frees users from remembering concise, unnatural commands, and VLMs, in particular, enable a broad set of commands that are contextual when given an image input.

We envision an interface that is a natural extension of pointing that does not occupy the hands or require any external tracking setup. We developed a prototype called WatchThis, a wearable VLM-enabled device that allows users to query real-world objects and environments in natural language through pointing in mid-air. The camera on the watch directly captures the area of interest when pointing. The rotate and flip-up design of the watch body is purposefully non-discrete, making it clear to others that a picture is being taken.

Through this protoype, we aim to explore the unique applications enabled by watch-based, VLM-enabled, point-and-ask interactions, such as identifying objects, translation, and asking for instructions. We chose off-the-shelf parts and open-sourced our code and design so that others can build on our work and try it out. We demonstrate a few examples implemented with the minimal setup of our prototype and outline some future concepts with additional add-on functionalities.

2 SYSTEM DESIGN

We set out to design a wearable system to explore interactions enabled by having a camera on a watch with the ability to quickly query a VLM. Specifically, we have the following design goals:

- **Natural Interaction:** Unlike current smartphone-based solutions that occupy one or both hands, WatchThis capitalizes on intuitive pointing gestures for effortless interaction.

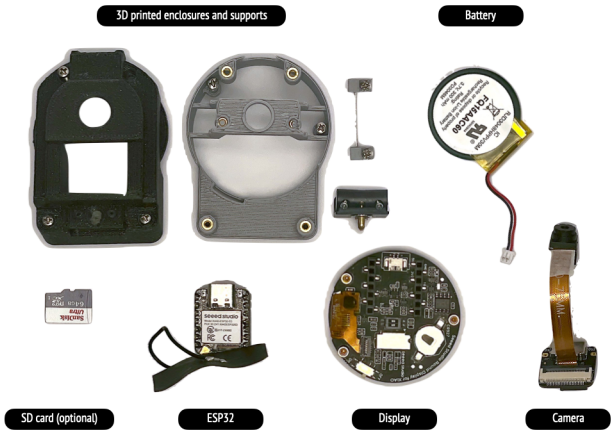

Figure 3: Components of WatchThis.

- **Privacy-Preserving:** Recognizing the potential discomfort caused by always-on cameras in social settings, the mechanical design enforces a visible activation.
- **Seamless and Swift Operation:** The system is designed for quick, almost reflexive use, providing answers at a glance without disrupting the user's flow of activities.
- **Playful:** wearable computing has a history of personalization culture that allows individuals to express their unique selves. The intention is not to replace a phone which is a well-engineered system with reliable functionalities. This device hopes to enable playful uses and spark joy.
- **Open-Source:** We prioritized cost-effectiveness and the use of open-source components to foster further innovation and customization by the wider community.

The final prototype weighs 48g, 30% lighter than an Apple Watch Series 3, and costs around 60 USD. To use the device, the user flips up the watch body from the watch strap, points at the object of interest, and receives the answer on the screen (Figure 1). The system has a default prompt of "What is this?", but the prompt can be easily changed via WebApp (Figure 1 bottom left). Next, we describe the hardware and software components. We provide resources and design for the hardware components and open-source the software component in this Github repository (https://github.com/cathy-mengying-fang/WatchThis) to allow people to recreate and modify the project.

2.1 Hardware

The prototype comprises three main hardware components: a watch display, an ESP32 microprocessor, and a small camera (Figure 3. These components are from Seeed Studio's XiaoESP32S3 and the expansion modules. The prototype can run fully standalone with a LiPo battery and can include an optional SD card for additional storage and functionality. The mechanical design consists of the module housing and a 2-degree-of-freedom swivel joint that attaches to a watch strap.

The flip-up screen allows for ease of aiming at the target. Compared to AR devices, where the results can only be viewed by the

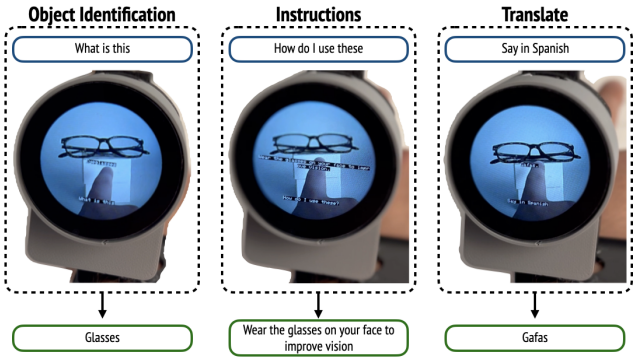

Figure 4: User can easily redefine the quick action prompt via the accompanying WebApp, enabling different contextual queries beyond object identification. These queries can include asking for instructions on how to use the object or translating text into different languages.

user, or projection-based devices that require a surface to receive the projection, the wearable screen is friendly for single-user and multi-user contexts. Applying the WYSIWYG principle, the screen also provides direct confirmation to the user what the system sees.

2.2 Software

The software is written in Arduino-compatible C++ and runs on-device. Once the device is flipped up, the screen displays the camera feed in RGB format. When a touch event is detected, the screen displays the captured image and stops streaming the camera feed. The captured RGB image is compressed to JPG and then decoded to base64 format. An API request is then made to query the image. Once a response is received from the API call, the screen displays the text response overlaid on top of the image. After 3 seconds, when the user has seen the response, the screen returns to streaming the camera feed and waiting for the next command. The end-to-end response time is around 3 seconds.

We made API calls to OpenAI's GPT-4o model[1] as it accepts both text and image input. Instead of voice recognition and text-to-speech, which can be error-prone and costly in terms of compute and speed, we opt for direct text input for the query. The system has a default prompt that is assumed to be the "Quick Action" for each query. The system has an accompanying WebApp served on the device to quickly change the default prompt. We demonstrate multiple example use cases for the "Quick Action".

3 APPLICATION

Here we show some example use cases enabled by the interaction modality and capabilities of the system (Figure 4). Upon discovering an object, a building, or a plant, one can *point* at it and ask the device to **identify the object**. One can also ask questions that are based on the identification of the objects such as **asking for instructions** or the functionalities of components. For example, the user can point at a word or image on the menu and **translate** it to a different language. For language learners, this device can enable them to practice on the go in real-world contexts.

[1]https://platform.openai.com/docs/models/gpt-4o

Figure 5: Future version of this prototype can incorporate GPS information that allows people to direction point at a building in the distance and receive directions.

Some other ideas but not implemented fully, include:

- **"Remember this"**: Many of us have experienced the scenario of seeing something that reminds us of a task, only to forget it once the object is out of sight. The device can serve as a quick reminder, for example, for people who need to take medicine.
- **"How do I get there"**: If the device is given GPS information, one can point at a distant building and ask how to get there, using the user's current direction (Figure 5. This allows for spatially contextual wayfinding rather than mapping user's physical location onto a 2D map.
- **"Zoom in on that"**: Upon detecting a pointing gesture, the camera can zoom in at the fingertip to help people capture information from afar. disrupting the user's flow of activities.
- **"Turn that off"**: When provided with permission to discover local devices on the same network, IoT devices can expose the API. The VLM can detect the object of interest and understand which API to call based on the user's natural language command. For example, the user can say "turn it down" while pointing at an AC.

4 DISCUSSION

We opt for a simple setup with gesture and image as the main modalities of input. While speech input and output can be nice additions, the focus of this prototype is to investigate unique use cases enabled by "camera-on-the-wrist" and natural gestures in daily interactions. The image quality is influenced by unintended hand movements and the resolution of the camera module. Additionally, the camera's limited field-of-view (FOV) necessitates accurate and steady aiming to capture the desired area effectively. Enhancing the system with increased image stabilization and a variable FOV could provide users with greater flexibility when targeting objects both nearby and afar. Currently, the confirmation of the command is done via touching the watch's screen. Future work can incorporate intelligent vision algorithms that detect the presence of a finger and the action of pointing to enable single-handed interaction. We also plan to evaluate the usability through a user study and explore other use cases.

REFERENCES

[1] Richard A. Bolt. 1980. "Put-that-there": Voice and gesture at the graphics interface. In *Proceedings of the 7th Annual Conference on Computer Graphics and Interactive Techniques* (Seattle, Washington, USA) *(SIGGRAPH '80)*. Association for Computing Machinery, New York, NY, USA, 262–270. https://doi.org/10.1145/800250.807503

[2] Daehwa Kim, Vimal Mollyn, and Chris Harrison. 2023. WorldPoint: Finger Pointing as a Rapid and Natural Trigger for In-the-Wild Mobile Interactions. *Proceedings of the ACM on Human-Computer Interaction* 7, ISS (2023), 357–375.

[3] Sven Mayer, Gierad Laput, and Chris Harrison. 2020. Enhancing mobile voice assistants with worldgaze. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–10.

[4] Suranga Nanayakkara, Roy Shilkrot, Kian Peen Yeo, and Pattie Maes. 2013. EyeRing: a finger-worn input device for seamless interactions with our surroundings. In *Proceedings of the 4th Augmented Human International Conference*. 13–20.

[5] Jing Qian, Meredith Young-Ng, Xiangyu Li, Angel Cheung, Fumeng Yang, and Jeff Huang. 2020. Portalware: A smartphone-wearable dual-display system for expanding the free-hand interaction region in augmented reality. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–8.