ECE 445 Design Document

Project 30: Search and Identify

Name	Email
Shitian Yang	shitian.20@intl.zju.edu.cn
Yitao Cai	yitao.20@intl.zju.edu.cn
Ruidi Zhou	ruidi.20@intl.zju.edu.cn
Yilai Liang	yilai.20@intl.zju.edu.cn

Supervisor: Prof. Howard Yang & Gaoang Wang

ZJU-UIUC Institute
Zhejiang University, Haining, China

03/20/2024

Contents

1 Introduction

1.1 Objective

In contemporary household settings, cleaning and organizing often demands substantial time and effort, making the task of remembering and locating items casually placed in a cluttered indoor environment a common daily challenge. For elderly individuals or those with visual impairments, this task becomes even more challenging, time-consuming, and frustrating. Commonly misplaced everyday essentials include cell phones, wallets, keys, and glasses.

Although there are some existing solutions for locating items, such as using smartwatches or making calls to locate smartphones, or employing tracking devices like AirTags for item retrieval, the existing methods have their own set of limitations. For example, the strategy becomes ineffective when the smartphone is set to silent mode, and the tracking systems require the prior attachment of tracking devices to the items, which is costly and highly inconvenient.

Given these challenges, more direct and effective solutions become an urgent need. With the advancement of technology, numerous artificial intelligence systems, including ChatGPT[1], multimodal models[2–4], have been utilized in managing daily and work-related tasks. Inspired by this, the approach we took integrates image capturing and artificial intelligence, aiming to extend the application of AI in various daily challenges. By utilizing these advanced technologies, we can build more convenient human-computer interaction solutions, facilitating a seamless integration of artificial intelligence into daily life, and opening up new possibilities for smart home solutions that are intuitive and effective.

1.2 Solution

The final product of our project is a voice-activated home-use robot designed for item-seeking and navigating. It operates in response to user-initiated voice commands, providing descriptive information about the desired item for retrieval.

In this academic project, we have integrated a variety of frontier artificial intelligence technologies to enhance the robot's environmental recognition and interactive capabilities. The technologies include speech recognition, natural language processing (NLP), object detection, Visual Question Answering (VQA), and object segmentation. The artificial intelligence models or architectures we use includes:

- **Speech Recognition**: We employed the "Whisper Model"[5], a highly efficient deep learning model developed by OpenAI, designed to process and understand speech data in multiple languages.
- **Object Detection**: We employed the "YOLO" (You Only Look Once) model[6–8], more specifically its more recent version "YOLOv8"[9] and "YOLO-World"[10]. "YOLO" is a popular object detection and image segmentation model. Due to its fast

real-time recognition capability in identifying and locating multiple objects within images, it is widely used for dynamic object detecting, identifying, and tracing tasks.

- **Visual Question Answering (VQA)**: Visual Question Answering refers to the tasks that involve answering questions about an image. We employ the VQAv2 dataset[2] containing image and questioning pairing data for model fine-tuning, testing, and verification. In addition, we have utilized the CLIP[11] architecture to perform image comparisons. CLIP can compare image content based on multiple prompts and deliver the most closely matched results. Compared to other visual models, CLIP can handle more complex natural language inputs, significantly overcoming the limitations of other models that can only recognize single classification keywords from their training datasets.
- **Natural Language Processing**: We implemented the ChatGPT model[1] for the language processing task, which is utilized to refine preliminary prompts given by users into strictly formatted standard prompts suitable for model recognition. Considering the efficiency and cost of commonly used models in the market, employing ChatGPT represents an ideal solution for achieving this functionality.
- **Object Segmentation**: We implemented the "Segment Anything" model (SAM)[12] developed by Meta AI to extract objects. This model is capable of segregating target objects from their backgrounds, enabling us to achieve more precise object recognition and identification.

Upon successful identification, the robot indicates the object's location with a laser pointer. We designed and constructed a dual-axis rotating mechanism suitable for the requirements of the specified indicator, which is controlled by an STM32 microcontroller. This hardware architecture accurately positions the target direction, offering flexible rotation within 360 degrees on the horizontal plane and 120 degrees on the vertical plane.

Our project has great utility potential in helping users find specific objects in broader categories around their houses if later combined with a developed dynamic system or adopted massively in different areas.

1.2.1 Software Components

1. **User Voice Recognition Module:** Utilizing the "Whisper" model, this module is key in capturing and accurately parsing user voice input. It ensures that the system can understand multiple languages and accents, allowing the robot to serve a diverse user base. This module is the initial point of user interaction and its accuracy directly affects the efficiency and correctness of subsequent modules, forming the foundation for the system's usability.
2. **Prompt Processing Module:** Employing the ChatGPT model, this module transforms user natural language instructions into clear, precise query prompts. This step is crucial for ensuring that the robot can accurately understand user requirements and act accordingly. By optimizing prompt processing, this module significantly enhances the relevancy and accuracy of target recognition, acting as a bridge

between user demands and machine operations.

3. **Target Selection Module:** Using the "YOLO-World" model, this module's primary function is to quickly identify objects related to the user's query from complex environments. This preliminary filtering step not only enhances the efficiency of subsequent processing but also reduces the potential for misidentification. It is vital for maintaining system response speed while ensuring operational accuracy, especially important in real-time operating environments.
4. **Target Locating Module:** Integrating the "CLIP" and "Segment Anything" models, this module is responsible for precisely identifying and confirming the specific item requested by the user. Through semantic analysis and detailed image segmentation, this module not only identifies the most matching object but also accurately determines and depicts its specific location and shape. It is the final link in precisely aligning user instructions with the robot's actual actions, crucial for enhancing user satisfaction and the success rate of the robot's operations.

1.2.2 Hardware Components

1. **Image Acquisition Module:**

   Acquires images of the environment through camera with different positions and angles controlled by the steering motor, can also perform appropriate zooming and stitching and transmit the visual data to software for analysis.

2. **Control module:**

   Adopts digital circuit and logic chips, transforming the condition signals into digital signals in controlling the powering module. e.g. This control module will change its output when the software component judges that it suddenly found the desired object.

3. **Drivetrain and Powering Module:**

   This module is the core of the hardware component, where the micro-controller STM32F103c8t6 chip[13] is adopted in controlling and providing the steering engine with impulse signal. We also adopts the software of Keil5 and FlyMCU to help write the hardware codes into the chip board[14].The steering engine will be continuously rotating uniformly, until input signal from the control module shifts. A 12V power source is adopted in charging the micro-controller.

1.3 Physical Design

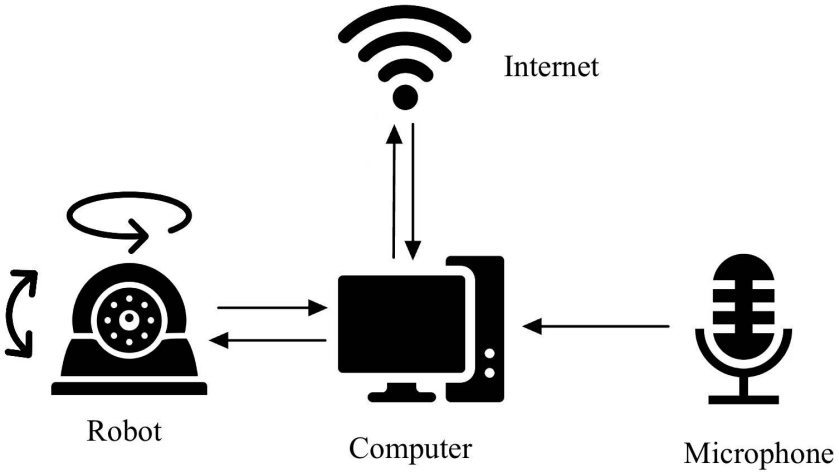

Figure 1: Physical Design System Diagram

1.4 High-level Requirements list

1. **Accurate functionality of steering motor system:**

   Our hardware system mainly centers around the steering system, it should possess the capability to accurately indicate the object based on a specific given coordinate. Once turning on the steering engine without targets, the steering engines will be continuously rotating uniformly. The indicator will only come to a stop when all its surroundings are scanned and a desired target is found. It must also precisely indicate a direction within an angular deviation of no more than 3° from the target's horizontal and vertical rotation. Additionally, it should transmit the data of the captured picture back.

2. **Accuracy of Speech Recognition:**

   The voice recognition system must accurately activate when the user calls it, and correctly extract a clear and concise description of the desired object from the user's commands, achieving a minimum accuracy rate of 95%. This involves discerning the specifics of the desired item from the user's commands under various household noise conditions.

3. **Accuracy of Vision Module:**

   The vision module must accurately detect and pinpoint the location of the target object within the field of view of the camera, achieving at least an 85% accuracy rate. It should be capable of generating an accurate bounding box around the target object

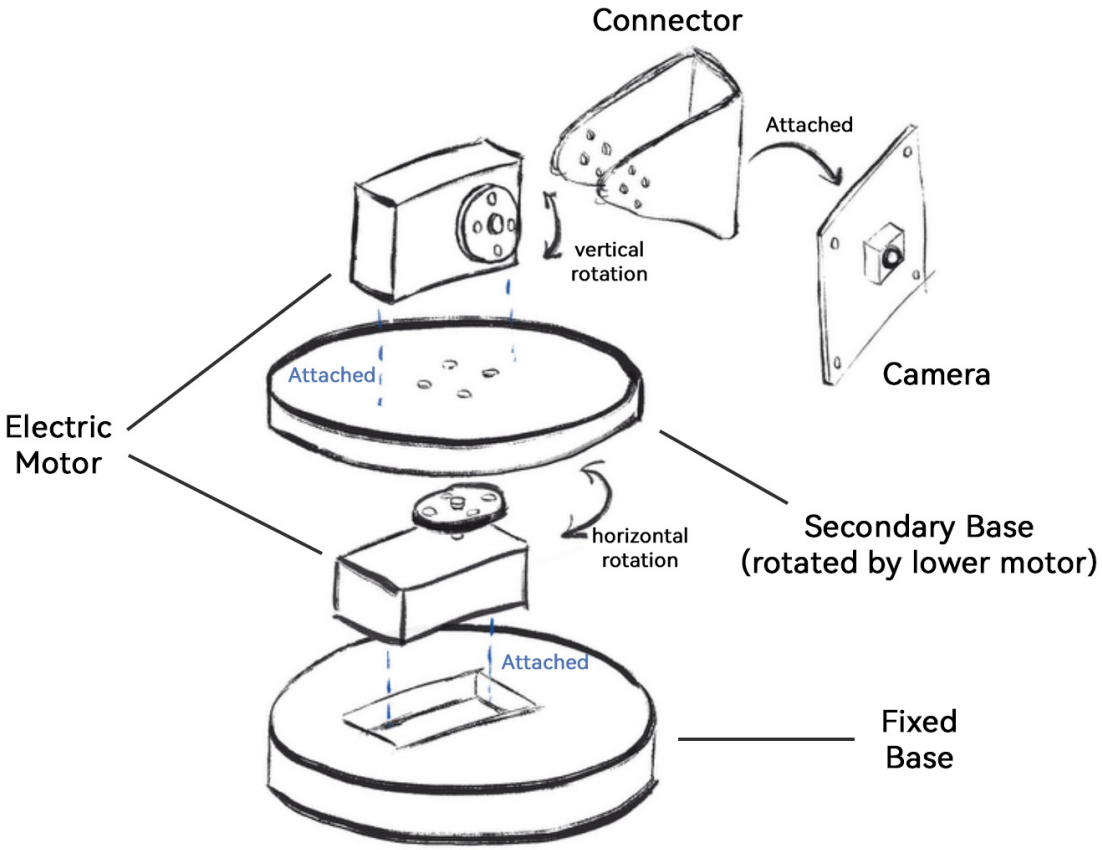

Figure 2: mechanical structure of our robot design

and precisely calculating its central point in relation to the robot's position. Furthermore, it should also be capable of creating a detailed mask that precisely covers the contour of the item, serving as an additional output. This could potentially lay the groundwork for a future manipulator module (even though this project will not develop such a module) or facilitate obstacle avoidance. The module's performance should remain consistent regardless of changes in lighting conditions, the orientation of the object, and background variations.

2 Design and Requirements

2.1 Block Diagram

The block diagram of our project is divided into components of software and hardware.

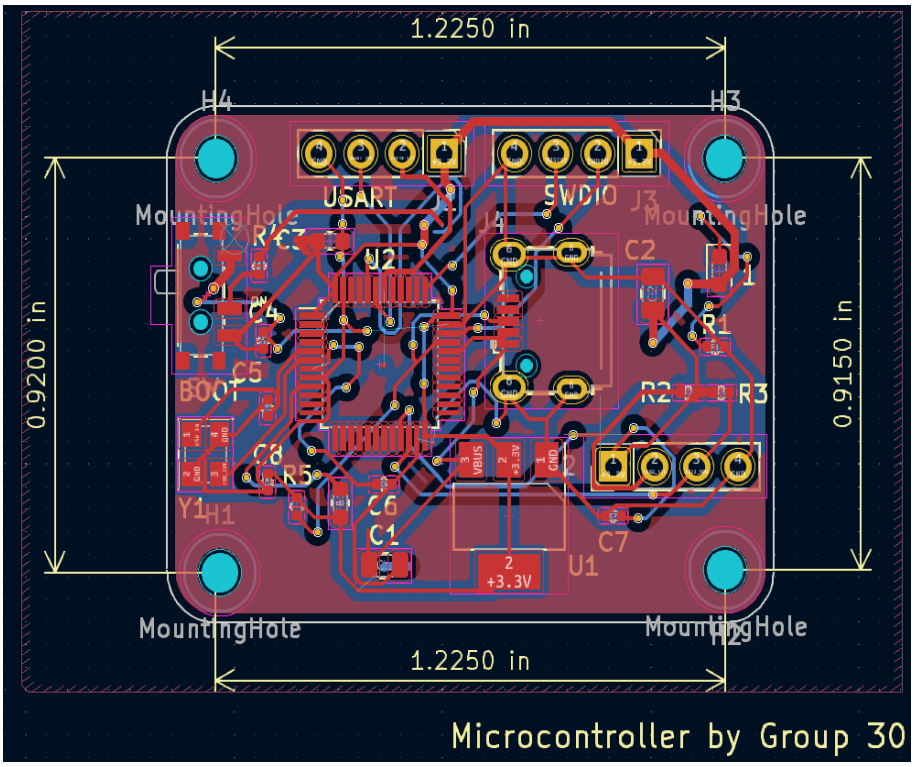

Figure 3: Design of our PCB board of microcontroller

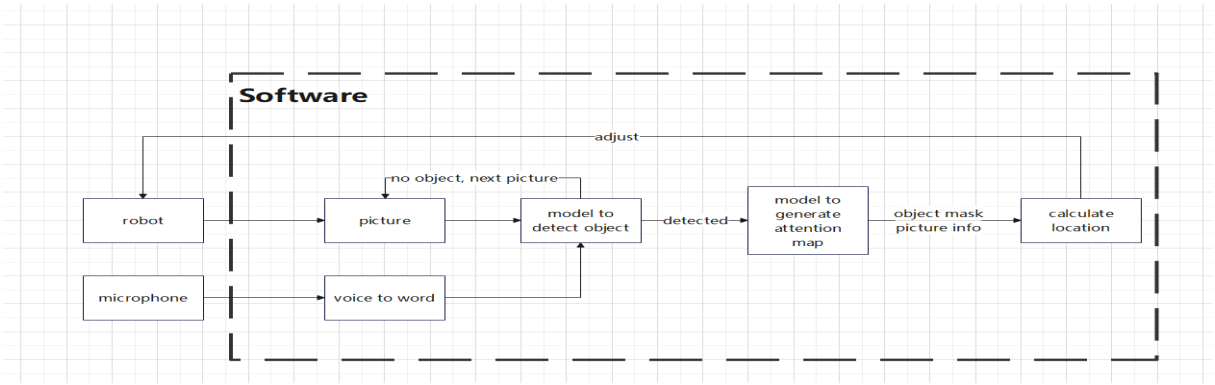

Figure 4: Block Diagram for the Software Component

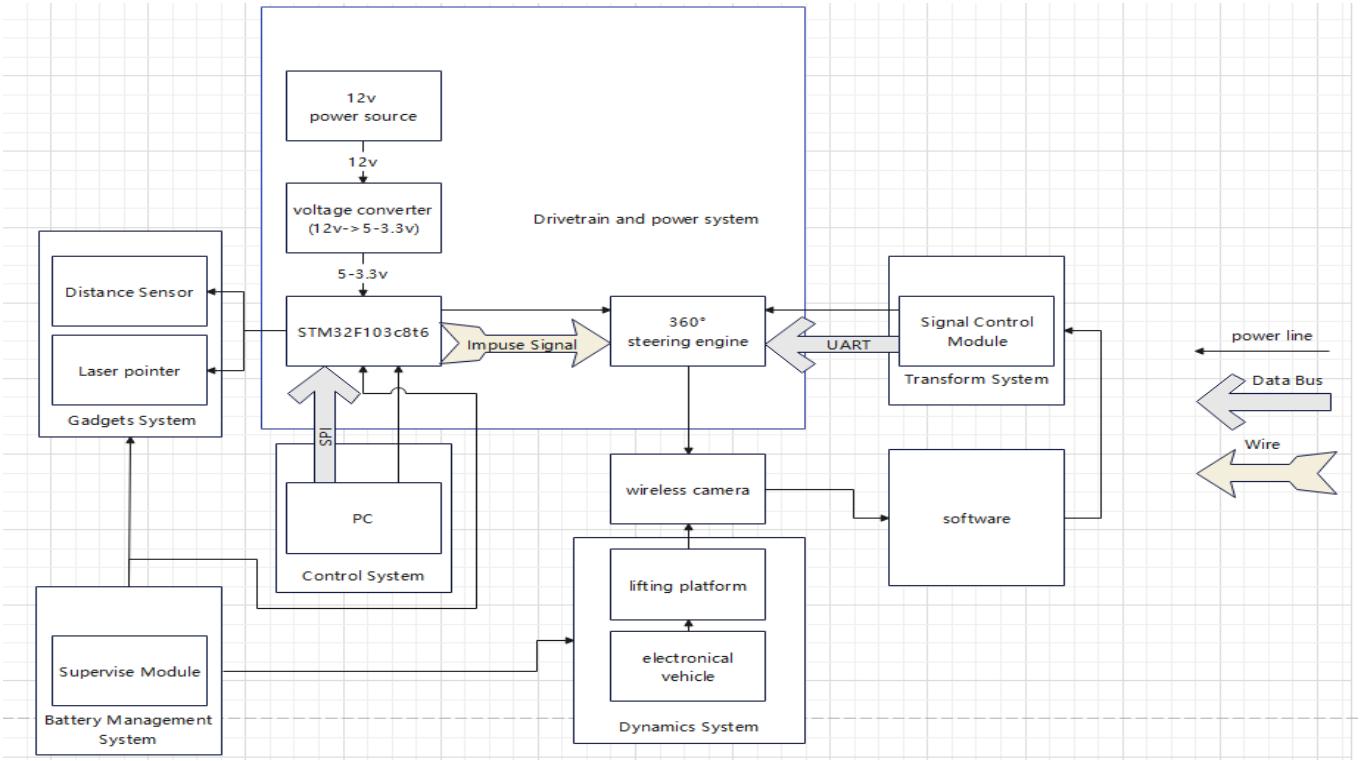

Figure 5: Block Diagram for the Hardware Component

2.2 Schematics for the hardware

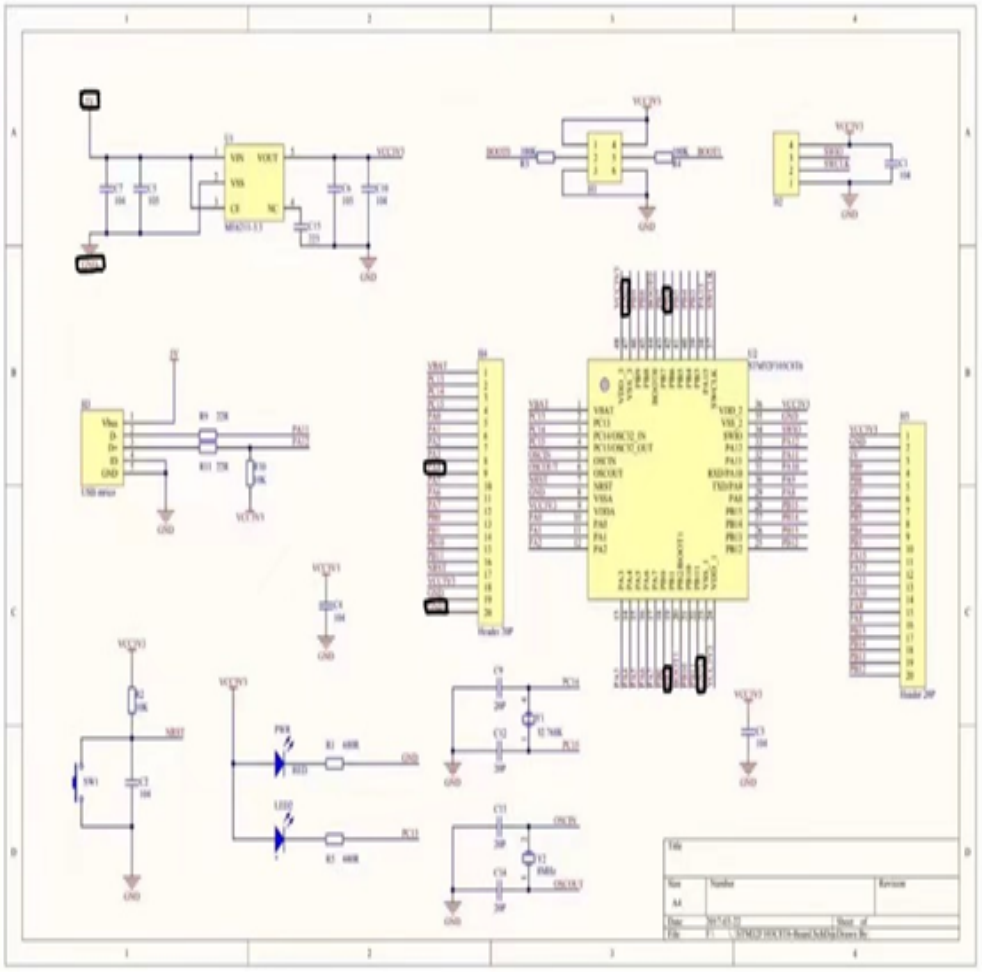

Figure 6: Schematic of STM32 board and ports to be used

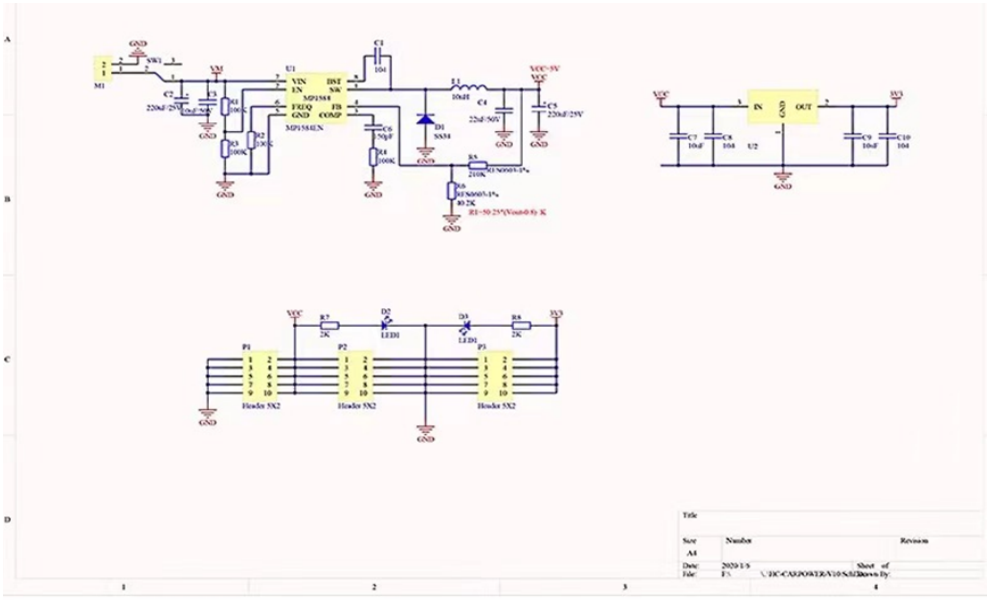

Figure 7: Schematic of interconnections of the battery

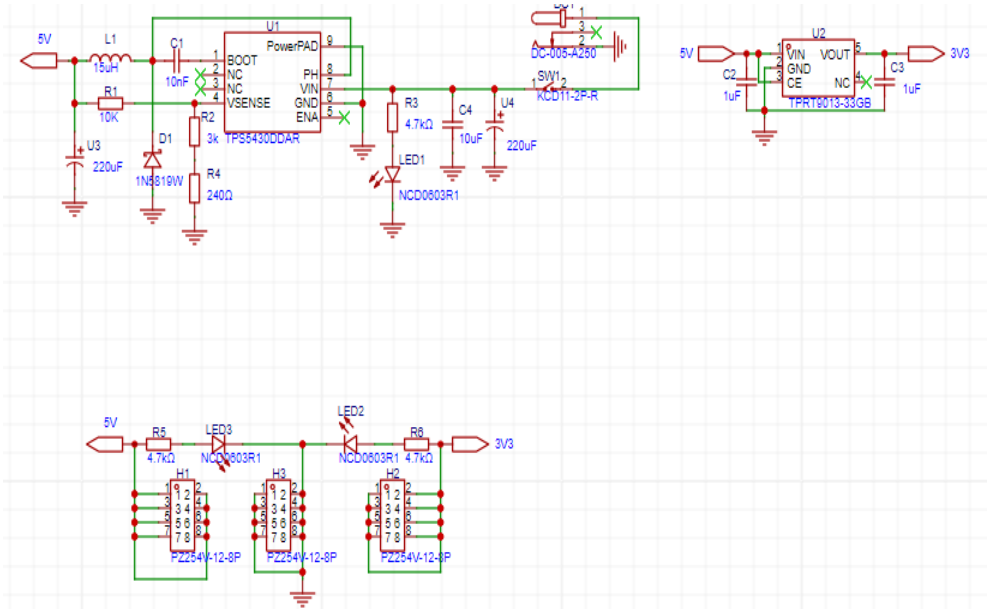

Figure 8: Schematic of the converter

2.3 Block Subsystem Functions & Requirements

2.3.1 Microphone Voice-to-Text Conversion System:

Verification: This mature model is responsible for converting audio captured by the microphone into corresponding text.

Requirements:
For our project, we will utilize the Whisper model[5] and the SPHINX[15] to implement the Microphone Voice-to-Text Conversion feature. Whisper is a groundbreaking speech recognition model that has been trained on a vast corpus of audio transcripts from the internet, amounting to 680, 000 hours of multilingual and multitask supervised learning. This extensive training enables the Whisper model to deliver high-quality speech recognition capabilities in a zero-shot transfer setting, effectively eliminating the need for dataset-specific fine-tuning. Remarkably, Whisper approaches the accuracy and robustness of human listeners and is designed to handle a wide array of speech processing tasks. These include multilingual speech recognition, speech translation, and spoken language identification, facilitated by its transformer sequence-to-sequence model architecture.

Given that our inputs come through a microphone, we anticipate minimal background noise. In tests conducted on the LibriSpeech.test-clean dataset[16], the Whisper model variants exhibit impressive accuracy: Whisper tiny (approximately 1GB in size and 32x relative speed) shows an error rate of 5.6%, Whisper base (approximately 1GB in size and 16x relative speed) has a 4.2% error rate, and Whisper small (approximately 2GB in size and 6x relative speed) improves further to 3.1%. Given our focus on everyday basic communication, we expect even higher accuracy levels, aligning with our high-level requirements. Moreover, thanks to its extensive training across various languages and datasets, Whisper demonstrates low sensitivity to accents, making it well-suited for a broad user base.

Model	**Test on original dataset**	**Test on recordings**		**Run time**
	Accuracy	**Accuracy**	**False language**	
whisper_tiny	94.4%	94.2%	15%	1.2s
whisper_base	95.8%	95.1%	8%	2s
whisper_s	96.1%	95.2%	8%	3.6s
whisper_m	96.5%	95.1%	6%	5s

Table 1: We tested four models of Whisper. The "False language" refers to the occurrence of non-English words generated by the multi-language model due to our accent or the transformation of short sentences.

2.3.2 Object Detection System:

Verification: The Object Detection Module is tasked with initiating preliminary searches to identify potential objects of interest within captured environmental images. This module employs the YOLO-World and GLIP vision models to conduct object detection tasks across a range of lighting conditions. Its function is to rapidly sift through segmented regions of the environmental images, isolating areas that may correspond to the target objects, thereby facilitating the generation of more precise object masks in subsequent stages.

Requirements:
To fulfill the requirements of rapid and precise object detection, we plan to utilize a dual-model approach for coarse and fine-grained detection. Initially, we will employ ChatGPT with vision capabilities (GPT-4V)[1] for the primary image analysis. GPT-4V introduces the ability to process image inputs alongside textual information, expanding the traditional text-only input framework of language models. This capability allows GPT-4V to understand and answer questions about images, marking a significant step forward in multimodal AI systems.

Following the initial assessment with GPT-4V, we will leverage the YOLO-World model [10] for directed bounding box annotations based on verbal instructions. YOLO-World represents an evolution in the YOLO detector series, incorporating open-vocabulary detection capabilities. It is pre-trained on extensive datasets to enhance its detection and grounding abilities, facilitating efficient user-vocabulary inference through a prompt-then-detect paradigm. This approach allows for real-time, open-vocabulary object detection with remarkable speed and accuracy.

By integrating these two models, we aim to ensure the precision and speed of our Vision Module, meeting our high-level requirement for accuracy in vision. Moreover, leveraging ChatGPT's linguistic intelligence, we intend to design smart feedback mechanisms for scenarios where objects are not detected or only similar items are found. This intelligent interaction with users will clarify the current recognition status and address any issues encountered, thereby enhancing the user experience by providing insightful and constructive feedback.

2.3.3 Object Mask Generation System:

Verification: The Object Mask Generation (OMG) System is designed to produce high-quality masks that enable the precise extraction of target objects from their backgrounds, intended for further verification and display purposes. The system accepts an input image accompanied by a bounding box and a textual description of the desired object, and generates masks for the target object based on these input parameters.

Requirements:

Objects	YOLO-World_m					
	No occlusion		25% occlusion		50% occlusion	
	Recall	False Alarm	Recall	False Alarm	Recall	False Alarm
bottle(10)	1	0.09	0.8	0	0.4	0
book(5)	1	0	0.8	0	0.3	0
scissor(5)	1	0	1	0	0.8	0
pen(10)	1	0	1	0	0.7	0
Objects	YOLO-World_x					
	No occlusion		25% occlusion		50% occlusion	
	Recall	False Alarm	Recall	False Alarm	Recall	False Alarm
bottle(10)	1	0.09	0.7	0	0.2	0
book(5)	1	0	0.8	0	0.3	0
scissor(5)	1	0	0.8	0	0.2	0
pen(10)	1	0	1	0	0.6	0

Table 2: The ability of two YOLO_World models was tested in a messy room in different occlusion situations. We placed 5 or 10 objects for testing. The image of test room you can find below.

In order to generate high-quality masks for the target object, we utilize the Segment-Anything model[12] (SAM) developed by MetaAI for this purpose. And we may use AbsVit[4] for adjustment. SAM adopts a universal segmentation approach that recognizes contextual information and detects object boundaries, enabling it to effectively identify and isolate specified objects from their surrounding environments without having to include them in the training dataset.

However, SAM has limitations in classifying and understanding the objects. During our test, we found out that SAM may fail to find the desired object if only the text description of the object is provided, especially if the object is unique and has not been trained before or if multiple objects of a similar kind confuse the model. Therefore we decided to separate the task into multiple parts, to utilize other models with strength in object detection and let SAM only handle the task of generating object clipping masks.

As explained above, the object detection system will pass the image, text prompt, and the bounding box of the object to the object mask generation system. The object mask generation system will output a clipping mask of the clipped object. We then use it to remove the background and extract the object for display and verification.

Objects	AbsVit		
	No occlusion Accuracy	25% occlusion Accuracy	50% occusion Accuracy
bottle(10)	0.9	0.9	0.6
book(5)	1	1	0.8
scissor(5)	0.8	0.8	0.8
pen(10)	1	1	0.7

Table 3: The ability of AbsVit model was also tested in a messy room in different occlusion situation. Compared with YOLO, AbsVit use a sentence as prompt, so we focus on the accuracy of target object with detail description.

2.3.4 Calibration and Computation System:

Verification: In order to navigate the user to their desired object, our system must analyze the direction of the target, which can be indicated with coordinates of a polar angle and an azimuth angle. This module receives coordinates of the camera direction when capturing the image, along with the bounding box of the target object. Utilizing this data, the system computes the directional coordinate of the target object's center point relative to the camera. Subsequently, it integrates these calculations with the camera's direction to derive the object's absolute directional coordinates.

Requirements:
In order to actualize this function, a prerequisite is the precision of the input parameters obtained by the module. This encompasses the intrinsic parameters of the camera, such as the angular field of view (AFOV) and focus length, as well as the coordinate parameters of the camera orientation acquired from the mechanical structure, which must be precisely aligned with the camera's actual physical direction under the control of the mechanical structure. This calibration is crucial for subsequent computations.

Then the system employs geometric calculations to transform the object's relative position in the image into real-world directional coordinates. The input data includes the camera AFOV and its directional coordinates (ϕ, θ), the WIDTH and HEIGHT of the image in pixels, and the bounding box of the target with parameters (x, y, w, h).

We apply the following calculations:

1. Target Center Coordinate:

$$x_c = x + \frac{1}{2}w$$
$$y_c = y + \frac{1}{2}h$$

2. Focus Length:

$$F = \frac{\text{WIDTH}/2}{\sin(\text{AFOV}/2)}$$

3. Relative Target Direction:

$$\varphi' = \sin\left(\frac{x_c - \text{WIDTH}/2}{F}\right)$$
$$\theta' = \sin\left(\frac{\text{HEIGHT}/2 - y_c}{F}\right)$$

4. Target Direction:

$$\Phi' = (\varphi + \varphi') \mod 2\pi$$
$$\Theta' = \theta - \theta'$$

Here the AFOV is the camera's angular field of view on its x axis, the coordinate system of the image has the x-axis toward the right and the y-axis downward. The spherical system adopts that polar angle θ is measured between the radical line of the target center and the upward direction, and the azimuthal angle φ is measured between the orthogonal projection of the radial line of the target center onto the horizontal plane and the x-axis. The geometric significance of some aforementioned

calculations is delineated in the Figure 9.

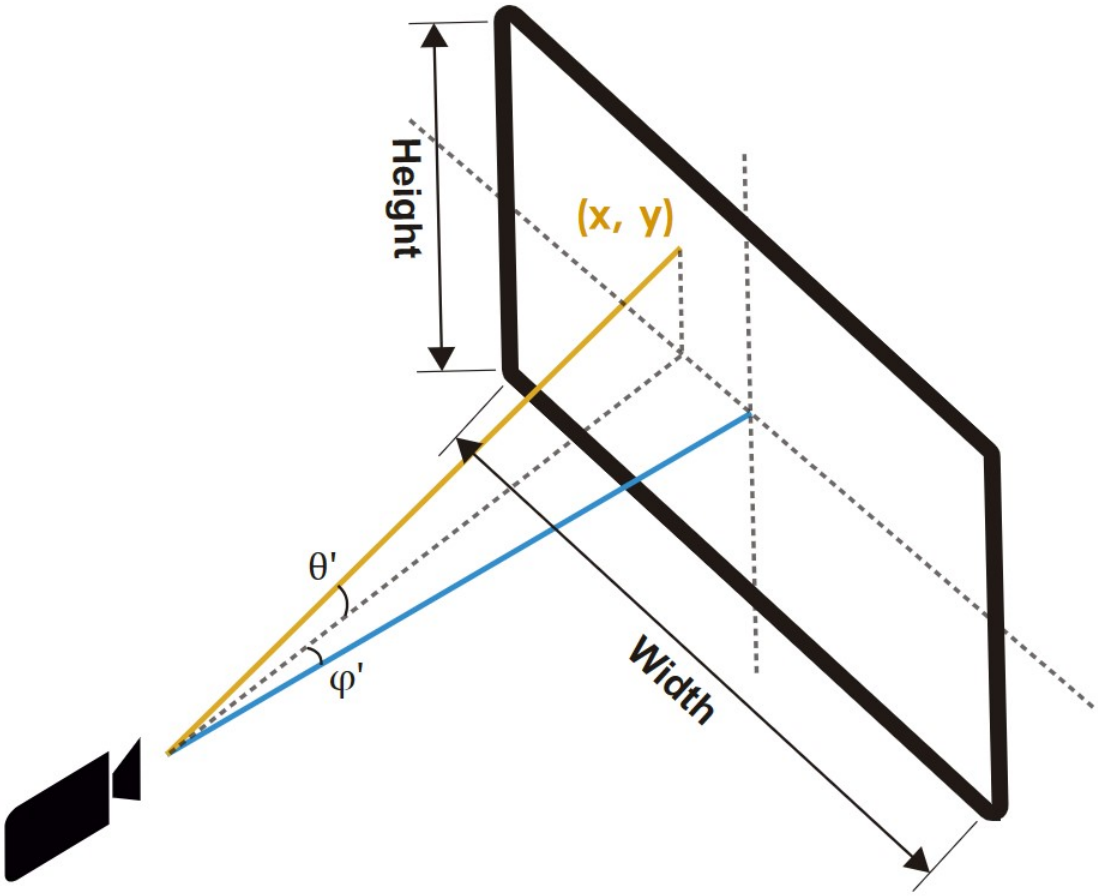

Figure 9: Geometrical explanation of the directional coordinate calculations

2.3.5 3D Picture generation System:

Verification: Pick up some specific pictures and generate the corresponding 3D pictures. Then compare the generated 3D pictures with the exact items, adjust the algorithm, also provide the corresponding data for the model to train so that for the next time about generating similar items, it can generate 3D pictures more accurately.

Requirements: For the final goals of the specific items, it should generate the similar 3D pictures and provide a sample for the users.

2.3.6 Drivetrain and Power System:

Verification: Each time we use serial ports to send the angle to the microcontroller and then we should use the angle measure instruments to measure the angle it rotates. Then we should also use the universal ammeter to measure the voltage of the

power source and the STM32 microcontroller. If the voltage is out of the range, testing the functionality of the power source and examine the connection of the wire lines. If the steering engine's error range is more than 1°, debugging the code of the STM32 microcontroller is the priority.

1. **12V Power Source:** The 12v power supply provides a stable 12v voltage and 10A current, and inputs into the voltage converter, providing power for the entire drivetrain and power system.

   Requirements: The power source must supply stable current and voltage to ensure the normal operation of the system.

2. **Voltage Converter:** The voltage converter connects the power supply with the micro-controller, ensuring the voltage and current do not exceed the limits required by the micro-controller.

   Requirements: Convert 12v to 5-3.3v, the current is 3A for 5v and 1A for 3.3v.

3. **STM32F103c8t6 Micro-controller:** Through hardware programming, we ensure the steering engine can rotate according to the set target. After rotating a specific angle, it stays for a predetermined amount of time, then automatically rotates the same specific angle again, until it completes a full rotation.

   Requirements: STM32F103c8t6 microcontroller's nominal voltage is 3.3v.

4. **360° Steering Engine:** The steering engine can automatically rotate according to the set angle and then stay at the set angle, thus controlling the angle. The steering engine connects with the wireless camera to take pictures of the surrounding environment.

   Requirements: The steering engine can accept the 3.3v signal voltage or 5-8.5v power voltage. The steering engine can carry up to 30kg items.

2.3.7 Transform System:

Verification: Measure the voltage when the software controls are turned on and turned off, and adjust the MOSFET structure when it is not satisfied.

Based on the results of the software part, determine whether to output a 0/1 signal (low-level or high-level signal) to the steering engine. The steering engine will stop running after accepting 0 signal and keep running after accepting 1 signal.

Requirements: The signal Control Module should output the 3.3v signal voltage to the servos.

2.3.8 Control System

Verification: Measure the corresponding time, if it is more than 5s, examine the code and modify the PWM of the signals in the code.

PC writes the program into the microcontroller and gives the angles we want.

Requirements: the steering engines can only rotate in 0-360°, so the input angles should be in the range 0-360°.

2.3.9 Dynamics System:

Verification: Measure the actual position and the goal position, if the distance is more than 10cm, adjust the transmission shaft and minimize the error.

1. **Lifting Platform:** Raise the camera to allow the camera to cover a larger field of view, achieving three divisions rotation.The lifting platform should lift the camera in 0-0.5m and 0.5-1m.

   Requirements: The power source in the lifting platform should supply the platform with enough energy.

2. **Electronical Vehicle:** Carry the lifting platform and the wireless camera to search and identify the specific items in a broader area.

   Requirements: The battery attached to the vehicle should power the vehicle when needed.

2.3.10 Gadgets System:

Verification: Measure the distance between laser pointer's pointing position and the goal's position, and also measure the distance of the camera and the goal's position. If they are out of the range, adjust the position of the laser pointer and the distance sensor on the robot.

1. **Laser pointer:** When detecting the goal items, a laser pointer will be activated and point at the items.

   Requirements: The microcontroller should give the signals to the laser pointer after finding the goal items.

2. **Distance sensor:** When detecting the goal items, the distance sensor should calculate the distance between the goal items and the distance sensor so that the software

can calculate the specific coordinates..

Requirements: The microcontroller should give the signals to the distance sensor after finding the goal items.

2.3.11 Battery Manage System

Verification: The system is capable of alerting users of unexpected circumstances such as overheating or voltage spikes from the safety perspective, it will send out warnings if emergencies such as overheating and overloading occur.
This system consists of the supervising module which mainly handles two tasks. First of all, this system will keep track of the State of Charge (SOC) of the batteries powering the camera, micro-controller, and dynamic cart, and at the same time, monitor their battery health.

Requirements: The supervising module should be able to show real-time data related to the electricity for all of our electrical components. Warnings should be noticeable when physical values detected exceed our set threshold.

2.4 Tolerance (Risk) Analysis

A big difficulty in this project is that the angle control program of the microcontroller cannot receive external data. Thus the rotation angle of the steering motor cannot change according to the results of the software part. Through exact theoretical calculations, we plan to convert the results of the software part into a low- and high-level signal output through the logic gate and to control whether the entire hardware part is running. Also, different hardware items require different nominal voltages and currents. If it exceeds the required range of voltage or current, there are safety hazards, and it will cause damage to the hardware. We ensure that the voltage and current passing through each piece of hardware meet the requirements through accurate theoretical calculations and voltage conversion.

A simple logical equation is:

$$Y = (A \cdot B)' = (A') + (B')$$

A and B represent the results from software and hardware.

In our design, problems may occur when we add adjectives to demonstrate the objects we want. For example, when we want a "red cup", and in the scene, a red cup and a yellow cup are both present. It may be hard for the robot to tell which is the cup we demanded, and it has a great probability of returning the yellow cup, which is the incorrect decision. This indicates that more attention to the heatmap segmentation is demanded in our project. We will tolerate the differences in texture but will try to distinguish between

obvious features such as colors. For the hardware components, we try to adopt the feature of distance sensing, and we try to tolerate this data within 5cm since our objects will only be placed within 1m of the robot. Lastly, lighting conditions will affect the accuracy when conducting 3D reconstruction based on photos taken, so additional measures will be taken to lighten up the photos all to the same standard. The result can be tolerated only when the constructed picture has no or little color difference compared with the actual object.

In addition, Angle deviation is a problem that cannot be ignored. To make the deviation smaller, we need to do a theoretical analysis first. The relationship between the focal length (f) of a lens and the size (h) of the object being imaged can be described by the lens formula:

$$\frac{1}{f} = \frac{1}{d_o} + \frac{1}{d_i}$$

where: - f is the focal length of the lens, - d_o is the object distance (distance between the object and the lens), - d_i is the image distance (distance between the lens and the image).

The magnification (M) of the lens can be calculated using the formula:

$$M = -\frac{d_i}{d_o}$$

where a positive magnification indicates an upright image and a negative magnification indicates an inverted image. This is the foundation of our image acquisition module, which uses a camera with a lens to capture images of the environment and transmit them to the software component for analysis.

3 Cost and Schedule

3.1 Cost Analysis

Our work is to be estimated 10 hours/week for 4 people. One person is about $ 30/hour, we plan to finish ECE 445 design this semester for 16 weeks:

$$\frac{\$30}{hour} \times 4 \times \frac{10hours}{week} \times 16\,\text{weeks} \times 2.5 = \$48000$$

Table 4: Cost analysis

Part	Mft	Desc	For	Price	Qty	Total
Steering Engine	DS	360°, 3.3 V	Camera	188	2	376
STM32F103c8t6	DS	3.3 V	Steering Engine	28	1	28
Power Source	XMS	12 V	Steering Engine	14.5	1	14.5
Converter	SD	12 V $\rightarrow$ 3.3 V	Steering Engine	10	1	10
UVC Camera	HIKVISION	3.3 V	Software	300	1	300
Laser Pointer	HD	12 V	Gadgets	58	1	58
Sonic Sensor	DS	12 V	Gadgets	32	1	32
Total	/	/	/	/	/	818.5

3.2 Schedule

- **Mar 25- Mar 31**
  Shitian Yang:
  Download, install, and locally deploy the Whisper model. Familiarize with the previously installed AbsVit and Segment-anything models and understand their API calls.

  Yitao Cai:
  Prepare the environment for models and install the Segment-Anything model, GLIP model, YOLO model. Investigate UVC camera control protocol with OpenCV.

  Ruidi Zhou:
  Mount the camera onto our motor and achieve basic code-controlled rotation. Basic hardware testing.

  Yilai Liang:
  Read 5+ journals on 3D object reconstruction, make comparison for at least 3 algorithms and comment on efficiency and accuracy.

- **Apr 1 - Apr 7**
  Shitian Yang:
  Perform API calls with the Whisper, AbsVit, and Segment-anything models to ensure they can correctly execute tasks according to our requirements. Begin attempts to streamline the process.

  Yitao Cai:
  Test the functionality of the selected object detection models with images from dataset and real-world collected images. Implement the camera control program and test it.

  Ruidi Zhou:
  Purchase, mount, and code-control the second motor. Achieve the 360-degree rotation ability with testing.

  Yilai Liang:
  Set up the basic environment for construction algorithms to be tested, complete basic testing of a single image.

- **Apr 8 - Apr 14**
  Shitian Yang & Yitao Cai:
  Integrate the entire workflow involving the Whisper and object detection models selected from the models mentioned above based on their quality, and determine the appropriate parameters for each.

  Ruidi Zhou:
  Design and implement the stretchable base of our robot. Explore the possibility of our dynamic system.

  Yilai Liang:
  Explore the data transmission with the camera, perform object reconstruction with real objects using all methods, and choose the most suitable.

- **Apr 15 - Apr 21**
  Shitian Yang & Yitao Cai:
  Continue the tasks from the previous week to establish the software workflow, test its functionality and performance, and modify it if required. If possible, establish a small-scale image dataset aiming at the intended application context and perform fine-tuning training on pre-existing object detection models.

  Ruidi Zhou:
  Implement the application of distance calculating, mainly coding part.

  Yilai Liang:
  Wrap up the 3D construction applications, and help Ruidi make hardware connections.

- **Apr 22 - Apr 28**
  All four members meet together to connect our software and hardware components together. Complete first-time functionality test and debug.

- **Apr 29 - May 5**
  All four members meet together for the second round of testing and sharing information and data about the whole implementation process. Begin planning on the final thesis.

4 Ethics and Safety

There are indeed several concerns on safety and ethics with our project. First of all, a laser pointer is considered to be mounted to the camera for target directing. Though the laser pointer is only designated to be turned on when the desired object is found, it can pose serious risks to human health, including eye injuries and skin burns. To avoid inappropriate pointing, we will fix a baffle and protector to limit the laser pointer in certain angles with a low level and always keep the power off during testing to comply with relevant safety regulations and to minimize the risk of harm to users or bystanders[17]. While our design also adopts electric power sources and motors, special care will be paid on robust testing and validation procedures to ensure the reliability of the system and to prioritize user safety. This complies with the ACM Code of Ethics, Section 2.9, that "Design and Implement Systems That Are Robustly Secure[17]."

As a project involving visual and vocal data utilization, it's crucial that such data is handled securely and with respect for user privacy. With the scope of the course ECE 445, our team members will mainly be the operators and users, and we will take on the responsibility not to use or spread others' data without formal and proper permission[18]. Other users of our project will have total autonomy over whether or to what degree would they like to engage with our robot.

Another concern arises in transparency and explainability. Even if users permit our usage of their vocal data, it's our unshirkable duty to provide clear explanations of its decision-making processes, especially regarding object recognition and task execution, to ensure users understand and trust the system's behavior[17].

Last but not least, to ensure the users' safety and convenience, we adopt a battery management system that will monitor the charge on all our electric components since they are working wirelessly. This design not only provides the users' information about the charging condition of our robot but can also warn ahead if something unexpected or unsafe is about to occur, which also complies with Section 2.9 of the ACM Code of Ethics [6].

All of our group members carefully affirm that we will strictly follow the IEEE and ACM Code of Ethics.

References

[1] OpenAI, *Chatgpt can now see, hear, and speak*, OpenAI Blog, Accessed: 2024-01-10, 2024. [Online]. Available: https://openai.com/blog/chatgpt-can-now-see-hear-and-speak.

[2] S. Antol *et al.*, *Vqa: Visual question answering*, In Proceedings of the IEEE International Conference on Computer Vision, 2015.

[3] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, *End-to-end object detection with transformers*, 2020.

[4] B. Shi, T. Darrell, and X. Wang, *Top-down visual attention from analysis by synthesis*, 2023. [Online]. Available: https://arxiv.org/abs/2303.13043.

[5] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, *Robust speech recognition via large-scale weak supervision*, 2022. arXiv: 2212.04356 [eess.AS].

[6] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, *You only look once: Unified, real-time object detection*, 2016. arXiv: 1506.02640 [cs.CV].

[7] J. Redmon and A. Farhadi, *Yolo9000: Better, faster, stronger*, 2016. arXiv: 1612.08242 [cs.CV].

[8] J. Redmon and A. Farhadi, *Yolov3: An incremental improvement*, 2018. arXiv: 1804.02767 [cs.CV].

[9] Ultralytics. ""You Only Look Once model by Ultralytics"." (2023), [Online]. Available: https://github.com/ultralytics/ultralytics (visited on 04/08/2024).

[10] T. Cheng, L. Song, Y. Ge, W. Liu, X. Wang, and Y. Shan, *Yolo-world: Real-time open-vocabulary object detection*, 2024.

[11] A. Radford *et al.*, *Learning transferable visual models from natural language supervision*, 2021. arXiv: 2103.00020 [cs.CV].

[12] A. Kirillov *et al.*, *Segment anything*, 2023. arXiv: 2304.02643 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2304.02643.

[13] alldatasheet.com, *Stm32f103c8t6 datasheet(pdf)*, en. [Online]. Available: https://www.alldatasheet.com/datasheet-pdf/pdf/201596/STMICROELECTRONICS/STM32f103c8t6.html.

[14] L. Frenzel, "What's the difference between bit rate and baud rate?" *Electronic Design*, Aug. 2022. [Online]. Available: https://www.electronicdesign.com/technologies/communications/article/21802272/whats-the-difference-between-bit-rate-and-baud-rate.

[15] K. F. Lee, H. W. Hon, and R. Reddy, *An overview of the sphinx speech recognition system*, Jan. 1990.

[16] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210. DOI: 10.1109/ICASSP.2015.7178964.

[17] A. Ethics, *Acm code of ethics and professional conduct*, en-US, ACM Ethics - the Official Site of the Association for Computing Machinery's Committee on Professional Ethics, Jan. 2022. [Online]. Available: https://ethics.acm.org/.

[18] *IEEE Code of Ethics*, https://www.ieee.org/about/corporate/governance/p7-8.html.