Università di Catania

UNIVERSITY OF CATANIA

DEPARTMENT OF MATHEMATICS AND COMPUTER SCIENCE

PHD IN COMPUTER SCIENCE - XXXVIII CYCLE

Claudia Bonanno

Multimodal Conversational Assistance for Industrial Scenarios: From Task-Specific Systems to Large Language Models

---

PhD Thesis

Supervisor: Prof. Giovanni Maria Farinella

Co-supervisor: Prof. Antonino Furnari

Anno Accademico 2024 - 2025

Abstract

This thesis investigates the design and deployment of multimodal conversational assistants for industrial scenarios, bridging the gap between task-specific systems and modern Large Language Models (LLMs). Starting from pre-LLM approaches, the research explores natural language understanding, object recognition, and multimodal fusion techniques to support operators in executing complex procedures. The work introduces the HERO system in successive iterations: from a rule-based assistant leveraging egocentric vision, to a mobile version integrating question answering with LLMs, and finally to a zero-shot, LLM-powered architecture with modular components for contextual reasoning. These systems were developed and validated in the ENIGMA Laboratory, an industrial mock-up laboratory used for dataset creation and user studies. The experimental results, which include qualitative evidence from usability tests with real users and quantitative measures such as accuracy and F_1-score, demonstrate the effectiveness of combining language and vision to improve task guidance, reduce ambiguity, and enhance operator safety. The proposed solutions contribute to advancing multimodal AI assistance in high-stakes, domain-specific environments, offering a flexible architecture adaptable to future industrial and beyond-industrial applications.

Contents

Chapter 1

Introduction

When dealing with a complex task without any prior knowledge or necessary skills, the availability of an assistant is highly beneficial. Human assistants, such as colleagues or experts with greater knowledge of the task or the context, can provide valuable support. However, in many cases, especially in industrial settings, it is unrealistic to expect such assistance to be available when needed. Employing a second person to provide assistance either doubles the cost or, alternatively, reduces available workforce. For this reason, delegating the responsibility of providing assistance to an artificial assistant stands as a good alternative. As a result, we formulated the following research question: can artificial systems replicate the flexibility, the ability to adapt to different contexts, and the capacity to reason over inputs from multiple modalities, that human assistants possess?

To get closer to human behavior, artificial assistants are required to have skills that surpass simple question answering. For example, a human assistant is capable of identifying hardware tools based solely on their appearance

and point them out to the person who asked, when given the name of a specific tool. Conversely, a human assistant is capable of naming a tool when another person just points it out. Thus, an artificial assistant should not only be capable of processing textual questions, but also integrating information presented through different modalities, such as images. Mastering this skill allows artificial assistants to re-enact the way in which humans communicate, interweaving a variety of sensory channels into their interactions to provide context and nuance. Assistance in real-world tasks is often grounded in perception, memory and reasoning, which need to be encoded within artificial assistants in a robust way.

Multimodal conversational assistants combine various forms of input such as speech, text and visual data to improve contextual understanding and enrich user experience. By processing and fusing diverse sources of information, such systems can provide more accurate responses, support more natural interactions and better adapt to a variety of user needs. These capabilities are currently made possible through the use of deep learning systems, especially Language Models (LMs) and Visual Language Models (VLMs). Despite their potential, designing truly effective multimodal assistants remain an open challenge due to the complexities of aligning and grounding different modalities within a unified representation space.

The development of advanced conversational assistants has gained significant momentum in both consumer and industrial domains, driven by rapid advancements in Language Models (LMs) and Vision Language Models (VLMs). Language models have revolutionized Natural Language Processing (NLP), achieving near-human performance in understanding and gen-

erating text. Large Language Models (LLMs), such as GPT-3 [1], BERT [2] and their successors, have proven highly effective in various applications, including virtual assistants, chatbots and customer service, where they contribute to more natural and responsive interactions. Despite their widespread adoption, these models have rarely been integrated into more complex systems targeting domain-specific applications. Their general-purpose nature, while powerful, limits their applicability in specialized scenarios requiring fine-grained knowledge, procedural reasoning and multi-turn task alignment.

In parallel, Visual Language Models have emerged as powerful tools capable of processing multimodal input by combining text and visual information to interpret and respond contextually. This enables their deployment in scenarios requiring comprehension of both language and visual context. VLMs have great potential not only in consumer-oriented applications, but also in industrial settings where they can assist operators in executing complex procedural tasks. In such environments operators often rely on visual cues and step-by-step guidance to complete specific tasks. Embedding these capabilities into an assistant involves not only perception and recognition, but also the ability to follow structured workflows and respond to unforeseen events or user feedback.

More recently, models such as GPT-4 [3] and its successors have made notable progress in handling multimodal inputs. However, they still face limitations when applied to goal-oriented or domain-specific conversations, as they often lack mechanisms for contextual adaptation and tend to rely on their general knowledge. To address these shortcomings, functionalities such as web browsing have been introduced.

Ideally, a conversational assistant should be able to adapt to specific use cases by leveraging offline resources, such as incorporating a semantic memory when the context is predefined. Although this approach may reduce generality, the trade-off is often justified if it results in more accurate, context-aware and useful interactions. Ultimately, the goal is to move towards conversational assistants that can reason contextually, interact multimodally and adapt dynamically, all while maintaining safety, interpretability and usability in real-world environments.

1.1 Objective

Our research objective stems from the desire to study and apply assistive conversational systems to a real-world use case, such as the industrial domain, while demonstrating the ability to process and use modalities beyond text. The goal of this PhD project has been to investigate this area and propose solutions that could be effectively integrated into industrial scenarios, bridging diverse modalities. This involves not only developing prototypes or models, but also understanding the requirements, constraints and practicalities of deploying such systems in real industrial environments.

We believe that, in order for a multimodal conversational assistant to behave appropriately and effectively, it is essential to integrate insights from different research areas, namely egocentric vision, natural language understanding, semantic memory representation and retrieval, image/video understanding and beyond. These components, when harmonized, enable the system to reason in context, adapt to user needs and deliver timely, mean-

ingful assistance.

While some of these topics have been actively explored in state-of-the-art research, such systems still face significant challenges in being employed to their full potential in specific, real-world scenarios. They are often evaluated in benchmark environments, compared against standard tasks and fine-tuned on curated datasets, but rarely tested and validated in end-to-end real cases. As a result, their actual usability and robustness in live, complex environments remains unexplored.

Due to the unique timing of this PhD, starting in October 2022, just before the public release of Large Language Models, our approach to multimodal assistive systems spans two distinct technological phases. In the initial phase, we relied on pre-LLM techniques, which were more constrained in their understanding capabilities. As the project progressed and LLMs became more accessible and powerful, we began integrating them into our work experimenting with their use in assistive multimodal systems, with varying degrees of success. This transition not only reflects a technological shift, but also requires us to re-evaluate our system architecture and evaluation strategies to adapt to a rapidly evolving research landscape.

Formally, the focus of this thesis lies in the study and deployment of a conversational assistant capable of understanding natural language and interleaving multimodal information. Visual inputs such as images or video frames can act as contextual cues that, when combined with text, reduce the need for excessive clarifying questions, ultimately leading to a more fluid, user-friendly and human-like interaction experience. This aligns with the broader goal of making conversational systems truly assistive: capable of

anticipating user intent, grounding responses in context and reducing friction in communication.

By achieving this objective, this thesis aims to contribute to the advancement of multimodal assistive solutions, providing a solid foundation for industrial applications and offering a modular architecture that can be extended with additional functionalities depending on specific needs. Moreover, the findings and lessons learned throughout this research can inform future efforts in the deployment of conversational assistants in other high-stakes or domain-specific contexts, where interpretability, safety and robustness are paramount.

1.2 Contributions

In this thesis, we address the aforementioned challenges by proposing solutions based on deep learning and computer vision techniques. In particular, the task of natural language understanding is approached through a variety of processing strategies, including rule-based pipelines, Transformer-based architectures and modern large language models (LLMs). The integration of visual information, used to extract relevant cues that help resolve ambiguities or enrich contextual understanding, is tackled through the use of object detectors and different forms of multimodal information fusion. By combining these components into coherent pipelines, we are able to demonstrate the feasibility and effectiveness of our approach in realistic industrial scenarios. We propose several end-to-end pipelines that incorporate the aforementioned tools and achieve satisfactory performance in real-world use cases, as vali-

dated through empirical evaluation and feedback collected from actual users.

The contributions presented in this thesis have been published in the following research papers:

- Claudia Bonanno, Francesco Ragusa, Rosario Leonardi, Antonino Furnari, Giovanni Maria Farinella (2022). HERO: An Artificial Conversational Assistant to Support Humans in Industrial Scenarios [4]. In International Conference on Signal Processing and Multimedia Applications (SIGMAP), pp. 86-93.

- Claudia Bonanno, Francesco Ragusa, Antonino Furnari, Giovanni Maria Farinella (2023). HERO: A Multi-Modal Approach on Mobile Devices for Visual-Aware Conversational Assistance in Industrial Domains [5]. In International Conference on Image Analysis and Processing (ICIAP).

- Luca Strano, Claudia Bonanno, Francesco Ragusa, Giovanni Maria Farinella, Antonino Furnari (2024). GPT-Assist: Zero-Shot Conversational Assistance in Industrial Domains Exploiting Large Language Models [6]. In International Conference on Image Processing and Video Engineering (IMPROVE).

- Francesco Ragusa, Rosario Leonardi, Michele Mazzamuto, Claudia Bonanno, Rosario Scavo, Antonino Furnari, Giovanni Maria Farinella (2024). ENIGMA-51: Towards a Fine-Grained Understanding of Human-Object Interactions in Industrial Scenarios [7]. In IEEE Winter Conference on Application of Computer Vision (WACV).

1.3 Outline

This thesis is organized into eight chapters:

- Chapter 1 introduces the context, motivation and objectives of the research, outlining the overall structure of the work;
- Chapter 2 provides a review of the state-of-the-art methodologies relevant to the main research areas addressed in this work;
- Chapter 3 introduces the considered real-world setting, the ENIGMA laboratory;
- Chapter 4 presents our initial investigation into the development of a rule-based conversational assistant;
- Chapter 5 explores the deployment in a real-world context in combination with the gathering and analysis of user feedback;
- Chapter 6 builds upon the previous work by introducing Large Language Models (LLMs) and Retrieval Augmented Generation (RAG) into the conversational pipeline;
- Chapter 7 investigates the effect of LLM-augmented datasets on the system's overall performance;
- Chapter 8 concludes the thesis by summarizing the key findings, contributions and potential future research directions.

Chapter 2

Related Work

This chapter reviews existing literature, examining the key topics relevant to this thesis. The analysis aims to offer a comprehensive overview of the context, methodologies, datasets and findings from related studies, providing the groundwork for the discussions on our research contributions. Specifically, Section 2.1 describes past work related to conversational assistants. Section 2.2 examines the standard for Natural Language Understanding (NLU). Section 2.3 provides an overview on Language Models (LMs). Finally, Section 2.4 delves into the different approaches on text-image multimodality.

2.1 Conversational Assistance

Conversational assistants have proven themselves useful in various consumer scenarios, the most striking and used application being digital assistants for customer service, cutting the need of a real assistant for users for simpler questions. The main objectives of conversational assistants include do-

main classification, intent classification and slot filling. Domain classification makes the assistant capable of recognizing the domain the conversation is actually based on. For instance, buying a train ticket, booking an accommodation, or getting a refund for an order. Intent classification allows the assistant to identify what's the goal of the user inside the specific domain. For example, knowing the timetable for available rides, requesting information on a existing booking, generating the pre-paid return shipping label. Lastly, slot filling aims at obtaining all necessary information in order to assist the user in their intent and generate the appropriate response.

With the availability of LLM-powered assistants to the general public, the interest has shifted to conversational systems that are not goal-oriented, but generally offer the possibility to chat with users, simulating another human being. This kind of conversational systems include OpenAI's ChatGPT [1], Google's Gemini [2] and Anthropic's Claude [3]. The release of new versions also includes a plethora of functionalities, such as text-to-speech and viceversa, image analysis and image generation. The mentioned systems offer more natural experience, making use of previous answers in the conversation to gain information on the context and use it for the generation of latter responses, following a dialogue-state paradigm.

Some work has explored the use of multimodal conversational assistants within procedural fields as cooking [8] and the industrial domain [9], as well as the shopping domain [10]. However, due to the scarcity of suitable public datasets for the task and arising issues with safety and legal responsibility,

[1] https://chatgpt.com
[2] https://gemini.google.com
[3] https://claude.ai

the topic of multimodal conversational assistance has not been explored to its full potential.

2.2 Natural Language Understanding (NLU)

The main objective of Natural Language Understanding (NLU) is to enable machines to comprehend user input expressed in natural language. By leveraging NLU techniques, conversational assistant can interpret queries, extract relevant information and generate responses in a form that is understandable by humans. This capability is central to a wide range of applications, including virtual assistants, custom service chatbots, question answering systems and human-computer interaction interfaces.

Historically, research in NLU has relied heavily on effective text representation techniques. Several word embedding approaches have been proposed in literature to map words into dense vector spaces that capture semantic and syntactic properties, with notable examples being Word2Vec [11], fastText [12] and GloVe [13]. These methods have been widely used in tasks such as text classification, sentiment analysis and named entity recognition, laying the foundation for modern approaches in Natural Language Processing.

For many years, Natural Language Understanding served as a key framework for studying human-machine interaction. However, with the introduction of architectures based on Transformer [14] and, subsequently, (Large) Language Models, NLU approaches have been overcome, rendering them less central in contemporary research.

2.2.1 Intent Classification

Recent research in intent classification has explored both traditional supervised approaches and emerging methods for open-set and low-resource scenarios. For open intent classification, [15] proposes a method based on k-center contrastive learning combined with adjustable decision boundary learning, aiming to better distinguish between known and unknown intents while reducing the reliance on large labeled datasets. Similarly, [16] addresses the challenge of manual annotation by introducing a strategy that uses preudo-labels for intent detection, along with the Pre-trained Intent-aware Encoder (PIE), specifically designed to align utterance representations with their corresponding intent names, thus improving classification accuracy in low-label regimes.

Leveraging the capabilities of large language models, [17] investigates out-of-domain and generalized intent discovery by using generative systems such as ChatGPT in place of traditional discriminative classifiers. This line of work highlights the potential of prompt-based approaches in capturing unseen intent categories without explicit retraining.

In the domain of joint intent detection and slot filling, [18] proposes an architecture that integrates BERT and ELMo embeddings to simultaneously predict the intent of an utterance and extract relevant slots, thereby exploiting shared information between the two tasks to boost performance. CTRAN [19] advances this further with a hybrid CNN–Transformer model that captures both local and global dependencies in user utterances. MISCA [20] adopts a different perspective by introducing intent–slot co-attention,

eliminating the need of graph construction to enhance transfer of correlation information between the two tasks.

Explainability has also become a focus in this area: [21] proposes incorporating slot types into the slot filling process, not only improving accuracy but also enabling interpretable reasoning about how slot predictions are made.

Finally, foundational methods that have shaped the intent classification landscape include early attention-based models [22] and transformer-based baselines leveraging BERT [23], which remain strong references for comparison in contemporary research.

2.2.2 Slot Filling

A significant body of research approaches slot filling jointly with intent detection, as the two tasks are inherently related. For instance, [18] proposes an architecture that integrates BERT and ELMo embeddings to simultaneously perform both slot filling and intent classification. By sharing contextual representations, the model can capture dependencies between an utterance's global intent and its local slot annotations, resulting in improved overall performance.

Building on this joint modeling paradigm, CTRAN [19] introduces a hybrid CNN–Transformer architecture. The CNN component captures local sequential features, while the Transformer layers model long-range dependencies, making the system capable of identifying both short, frequent slot patterns and more complex multi-word slot spans. MISCA [20] adopts a co-attention mechanism between intent and slot predictions, replacing graph-

based interaction models. This design allows the slot extraction process to be dynamically guided by the predicted intent, increasing semantic consistency between the two outputs.

In addition to performance improvements, explainability has also emerged as a priority in slot filling. The authors of [21] propose incorporating slot type information into the model's decision-making process. This approach not only enhances classification accuracy but also enables the generation of human-readable explanations detailing how each slot prediction was reached, which is particularly important in domains requiring transparency.

Foundational contributions to the slot filling task include early attention-based sequence labeling methods [22] and transformer-based baselines leveraging BERT [23], which have set strong benchmarks for subsequent research and remain important points of comparison for evaluating newer approaches.

2.2.3 Dialogue State Tracking

Dialogue State Tracking (DST) is a crucial component of task-oriented dialogue systems, responsible for maintaining an up-to-date representation of the conversation state across multiple dialogue turns. This state typically includes the user's goals, requested information and relevant contextual variables, which are used by the dialogue manager to decide the next system action. Accurate DST is essential for ensuring coherent, context-aware interactions, especially in multi-turn conversations where references to previous utterances and implicit information are common.

Recent work by [24] introduces a novel approach that integrates corefer-

ence resolution into the DST pipeline. Their method adopts a text-to-text paradigm based on the T5 architecture, jointly predicting both entity mentions and their links across turns. By explicitly modeling referential relationships in dialogue, the system can more effectively track user goals when information is expressed indirectly, such as through pronouns or ellipsis. This represents a shift towards more unified generative approaches to DST, where state updates are generated directly in natural language form.

Earlier foundational contributions to this field include [25], which explored efficient syntactic and dependency parsing techniques relevant for slot tracking in dialogue, and [26], which proposed an end-to-end neural coreference resolution framework. The latter demonstrated the importance of resolving entity references for coherent state management, influencing subsequent DST architectures that integrate entity-level reasoning.

2.3 Language Models (LMs)

Language Models (LMs) are probabilistic systems designed to estimate the likelihood of a sequence of tokens and to predict the most suitable subsequent token in a sequence given the preceding context. By learning statistical and semantic relationships between words, subwords, or characters, these models capture both local syntactic patterns and long-range semantic dependencies within text.

The most significant advances in recent years have been driven by the introduction of the attention mechanism, and in particular by the Transformer architecture [14], which replaced recurrence with self-attention layers.

This paradigm allows models to process entire sequences in parallel and to dynamically weight the relevance of each token with respect to all others, resulting in improved efficiency and the ability to capture global context more effectively.

Following this breakthrough, several landmark architectures have set new standards in the field. BERT [2] introduced a bidirectional training objective that significantly improved performance on a wide variety of understanding tasks. T5 [27] reframed all NLP tasks into a unified text-to-text format, enabling a single architecture to be adopted to multiple tasks with minimal changes. More recently, models such as LLaMa-4 [28] and GPT-5 [29] represent the current state of large-scale autoregressive language modeling, integrating instruction-tuning and reinforcement learning with human feedback to improve adaptability and alignment with user intent. These models are trained on extensive and diverse corpora, often containing trillions of tokens, and benefit from advanced optimization techniques, large-scale distributed training and architectural refinements. Their scale and training data breadth enable emergent capabilities such as few-shot and zero-shot, where the model can perform novel tasks with little or no explicit fine-tuning.

While the term NLU is still used to refer to the comprehension-focused aspects of NLP systems, the boundary between *understanding* and *generation* has become increasingly blurred. Modern Language Models are not only capable of interpreting user input but also producing coherent, contextually appropriate output, often surpassing traditional systems in both adaptability and performance. This shift reflects a broader trend in artificial intelligence towards more general-purpose, scalable models that learn language repre-

sentations from vast amounts of data, reducing the need for hand-crafted features or domain-specific engineering.

2.4 Text-Image Multimodality

Recent advances in image-text multimodality have been driven by models capable of jointly learning from visual and linguistic signals to produce unified representations. These systems enable a wide range of applications, from image captioning and visual question answering to cross-modal retrieval and grounded language understanding.

One of the most influential contributions in this area is CLIP [30], which learns a shared embedding space for images and text by training on large-scale pair of images and their associated descriptions. This approach allows zero-shot tranfer to a variety of vision-language tasks without task-specific fine-tuning. Similarly, ALIGN [31] extends this paradigm by leveraging an even larger and noisier dataset without the need of expensive filtering or post-processing steps and aligning visual and language representations of the image and text pairs using a contrastive loss, showing that scale can compensate for the lack of a curated dataset, further improving generalization.

Architectures such as ViLBERT [32] adapt the Transformer architecture to multimodal inputs, using separate streams for visual and textual features that interact through co-attention mechanisms. This design enables more fine-grained reasoning between modalities, which is particularly beneficial for tasks requiring complex scene understanding. BLIP [33] and BLIP-2 [34] propose a vision-language pre-training framework which transfers flexibly to

both vision-language understanding and generation tasks, unlike preceding work, and build upon noisy web data by bootstrapping the captions, generating synthetic captions with a captioner and filtering noisy ones.

More recently, Flamingo [35] introduced a novel approach to few-shot learning in the multimodal setting, enabling large language models to process interleaved sequences of text and images without task-specific training. This makes it possibile to perform a wide variety of image-text reasoning tasks simply by providing a few examples at inference time, marking an important step toward general-purpose multimodal conversational systems.

Chapter 3

Setting: the ENIGMA Laboratory

To validate the system, we examined datasets from literature that could meet the following criteria: egocentric acquisition, involvement of procedural activities and specifically a real industrial setting. In general, tasks focused on understanding human behavior have been extensively studied thanks to the availability of public datasets that consider multiple domains [36, 37, 38, 39] or specific ones, such as kitchens [40, 41, 42], daily life [43, 44] and industrial-like scenarios [45, 46]. However, since data acquisition in a real industrial scenario is challenging due to privacy issues, safety and industrial secret protection, the available datasets do not reflect real industrial environments, considering proxy activities such as employing toy models made of textureless parts [46, 45].

To tackle the lack of such datasets and enable research in this field, the ENIGMA-51 dataset was presented. ENIGMA-51 is composed of 51 egocen-

tric videos acquired in an industrial environment which is a mock-up of a real industrial laboratory. The dataset acquisition involved 19 subjects who wore a Microsoft Hololens 2 [47] headset and followed audio and AR instructions provided by the device to complete the repair procedures on electrical boards. The subjects interact with industrial tools such as an electric screwdriver and pliers, as well as electronic instruments such as a power supply and an oscilloscope while executing the steps to complete a specific procedure. The data collection includes annotations with a rich set of labels that enable the systematic study of human behavior in the industrial domain, including:

- Temporal and verb annotations;
- Object interactions;
- Hands annotations;
- Egocentric Human-Object Interaction (EHOI) annotations;
- Next object interaction annotations;
- Utterances.

The dataset and its annotations are designed to support a set of 4 research tasks. Each task focuses on understanding human behavior from first person vision in the considered industrial context, but tackles different perspectives:

- Untrimmed Temporal Detection of Human-Object Interactions;
- Egocentric HOI Detection;
- Short-Term Object Interaction Anticipation;

- Natural Language Understanding of Intents and Entities.

The data collection process, along with the definition of the relevant tasks and the related experiments, led to the publication of the ENIGMA-51 dataset, presented in the following conference paper:

- Francesco Ragusa, Rosario Leonardi, Michele Mazzamuto, Claudia Bonanno, Rosario Scavo, Antonino Furnari, Giovanni Maria Farinella (2024). ENIGMA-51: Towards a Fine-Grained Understanding of Human-Object Interactions in Industrial Scenarios [7]. In IEEE Winter Conference on Application of Computer Vision (WACV).

As the evaluation environment for our research, we chose the ENIGMA laboratory, set in the Department of Mathematics and Computer Science of University of Catania. More specifically, the laboratory replicates a real industrial laboratory, containing 25 different objects that vary in functionality. These objects are required for the completion of step-by-step procedures inside the laboratory. Figure 3.1 shows the laboratory objects. The complete list includes:

- power supply;
- power supply cables;
- oscilloscope;
- oscilloscope probe tip;
- oscilloscope ground tip;
- welder station;
- welder base;
- welder probe tip;
- electric screwdriver;
- electric screwdriver battery;

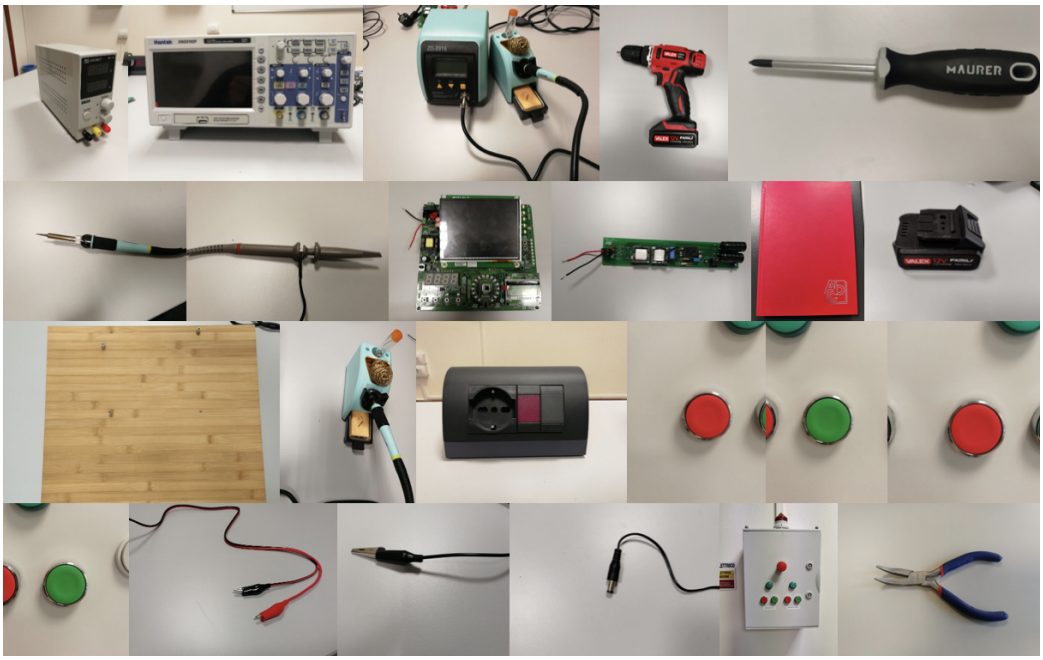

Figure 3.1: Picture of the ENIGMA laboratory objects.

- battery connector;
- screwdriver;
- pliers;
- high voltage board;
- low voltage board;
- low voltage board screen;
- register;
- left red button;
- left green button;
- right red button;
- right green button;
- socket 1;
- socket 2;
- socket 3;
- socket 4.

Inside the laboratory, operators can perform 4 different procedures, each including several steps. The full list of procedures includes:

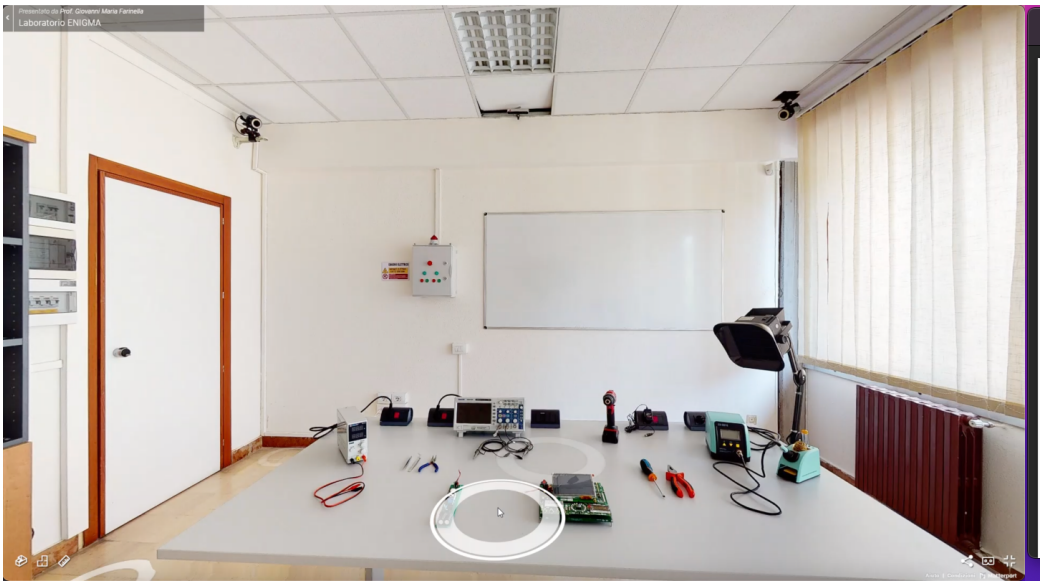

Figure 3.2: Picture of the ENIGMA laboratory.

- repair of high voltage board;
- testing of high voltage board;
- repair of low voltage board;
- testing of low voltage board.

Figure 3.2 shows a picture of the laboratory through the Matterport model acquired inside.

The participants involved in the data acquisition, equipped with a Microsoft Hololens 2 device [47], carried out specific test and repair procedures, representative of real industrial procedures. The first-person perspective provided by the wearable device's integrated camera enabled the acquisition of multimodal data throughout the execution of procedures.

The data collection procedure focused on the synchronous acquisition of video streams (2272 $\times$ 1278 pixel resolution, 30 frames per second) and

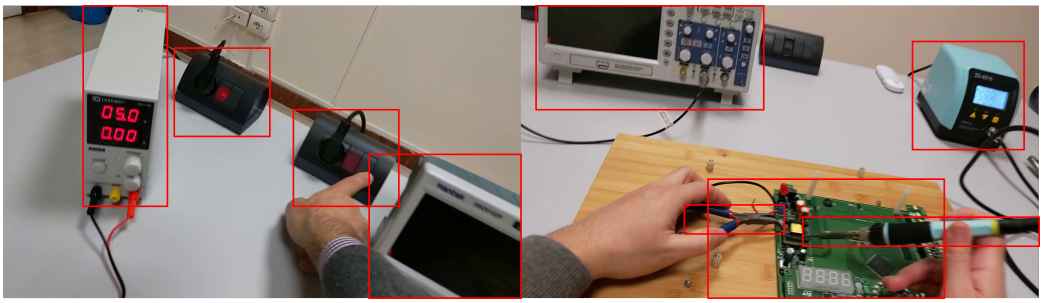

Figure 3.3: Some images with the bounding box annotations for the objects present in the scene.

vocal commands issued by users during the sessions. The egocentric recording modality enabled a natural understanding of the scene, making it possible, for example, to disambiguate vague references such as "This object" or "What should I do now?" by observing what the user was actually looking at.

For each video, the frames corresponding to the moment when the user directly touches an object and the frame immediately after the hand releases it were selected, marking the start and the end of the action. Representative examples of the annotated frames are shown in Figure 3.3.

Chapters 4, 5 and 6 explore the design and deployment of multimodal conversational assistants that work in our industrial laboratory, while Chapter 7 presents our contribution to the construction of the ENIGMA dataset, which covers the language section and the obtained results.

Chapter 4

HERO: a Vision-and-Language System for Procedural Support in Industrial Environments

Training new operators in industrial contexts is particularly challenging when it comes to learning lengthy step-by-step procedures. It usually requires a preliminary phase of studying the procedures themselves, the hardware manuals and other related documentation. Even experienced operators may find certain steps unclear, yet precision is essential. For this reason, it is common to pause a given step to retrieve the necessary information of clarification, either by consulting various manuals or asking more experienced colleagues. However, without knowing exactly where to look for the relevant information, this process may disrupt the workflow. Additionally, safety concerns may arise, as pausing a procedure at an arbitrary point is not always safe. These issues highlight limitations in industrial environments where a contin-

uous flow of action is desirable, as errors or oversights can lead to serious consequences.

In recent years, growing interest in artificial intelligence systems and their integration into work environments has led to the development of solutions aimed at assisting operators through natural and intuitive interfaces. Among these, conversational assistants, designed to interact with users via natural language, have seen a spike in popularity across several application domains, from customer services to domestic assistance. However, most conversational systems struggle with natural language ambiguity, as they typically lack access to the physical context in which the interaction takes place. This limitation undermines their effectiveness in environments such as industrial workplaces, where situational awareness is crucial for safety and productivity.

To address this challenge, the combination of egocentric vision and conversational systems offers a promising direction for improving interactions in industrial contexts. In such environments, operational complexity and task diversity require systems that are adaptive, fast and context-aware. Using a conversational assistant to ask questions during task execution can reduce friction. This is especially true when the assistant is capable of interpreting images and other visual inputs, making the experience comparable to consulting a highly skilled colleague, without requiring their time and enhancing workflow efficiency. Due to the modular nature of these systems, additional features can be incorporated to suit the typical use cases of a specific procedural task. Ease of use and the elimination of the need for a second person can increase execution speed and help reduce frustration in the workplace.

In industrial scenarios, a conversational assistant should be able to under-

stand requests and respond accurately while ensuring the operator's safety, especially when procedures involve handling hazardous equipment or dangerous steps. At the same time, such systems should be capable of processing and interpreting images or videos provided by the operator in order to better understand the context and deliver accurate, insightful responses. These two requirements lead to a key factor determining the system's success: the ability to fuse information from different modalities (in this case, vision and language).

To meet this need, a first version of HERO (Human Expertise Replication from Observation) was developed. This version of HERO is an early-stage conversational system that combines language and egocentric vision with domain-specific industrial knowledge, aiming to provide helpful, insightful and contextualized responses to operators. It allows the user to request information about the environment or specific objects (e.g. "How can I use this object?") through natural language, accompanied by images captured from their point of view using wearable devices. HERO is therefore capable of interpreting textual commands, identifying objects in the user's visual field and generating coherent responses based on a predefined knowledge of the relevant industrial domain. Unlike conventional chatbots, HERO leverages egocentric vision to perceive the surrounding environment, enabling it to resolve ambiguities, interpret user intent more effectively and provide context-aware and accurate responses.

HERO is designed for use in the industrial domain, where workers frequently perform maintenance, inspection and procedural tasks involving a wide variety of tools and equipment. The system assists users by answering

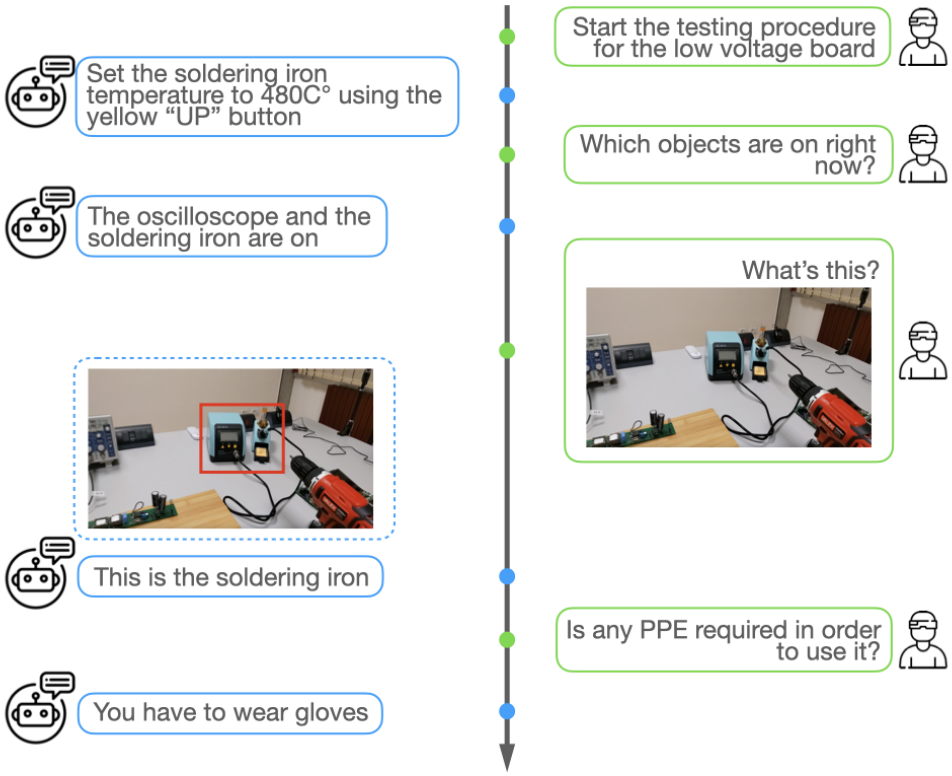

Figure 4.1: Concept of the proposed conversational assistant HERO.

questions such as "What should I do next?" or "How can I use this tool?", using information extracted from both language and visual data captured from a first-person perspective. Figure 4.1 illustrates the concept of the proposed system.

The findings and contributions of this work have been published in the following conference paper:

- Claudia Bonanno, Francesco Ragusa, Rosario Leonardi, Antonino Furnari, Giovanni Maria Farinella (2022). HERO: An Artificial Conversational Assistant to Support Humans in Industrial Scenarios [4]. In International Conference on Signal Processing and Multimedia Applications

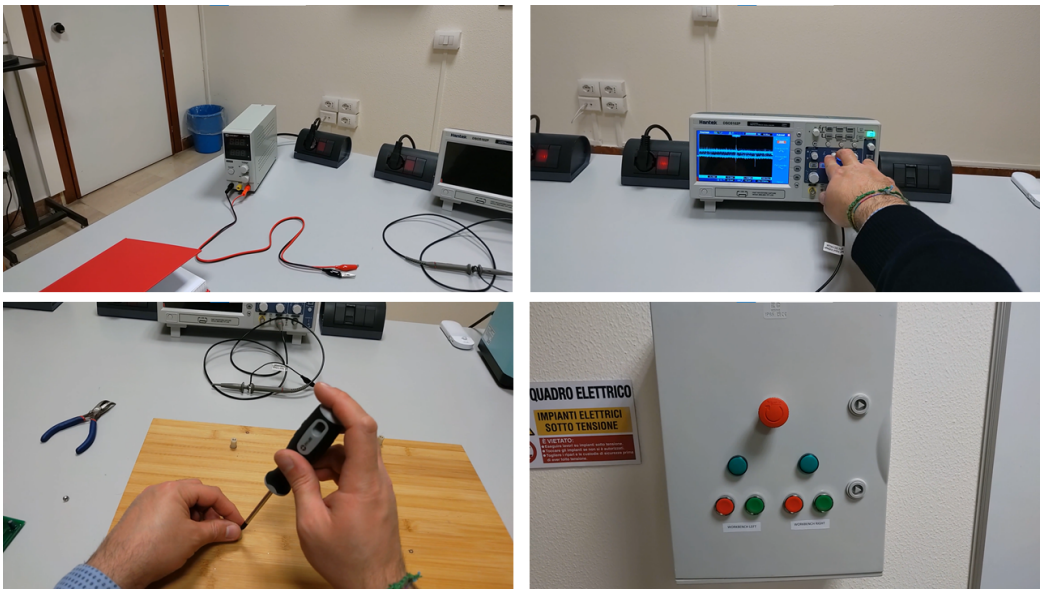

Figure 4.2: Sample images of the industrial laboratory.

(SIGMAP), pp. 86-93 .

4.1 ENIGMA Laboratory: Dataset and Context

We chose laboratory ENIGMA as a real-world setting to evaluate the effectiveness of HERO, inside of which users performed maintenance operations on two different electrical boards. Among the available tools in the laboratory we chose 23 distinct objects, including a power supply, an oscilloscope, power supply cables and other specialized hardware instruments. Figure 4.2 depicts sample images taken at the industrial laboratory.

To train the visual and language understanding modules of the system, the acquired images were manually annotated by identifying objects instances

with bounding boxes and assigning them their respective classes. In total, the dataset includes more than 20,000 annotated objects distributed across 23 classes, as well as over 130 annotated utterances associated with 26 intents and 4 classes of entities: objects, components, electric boards and procedures.

The set of relevant objects in the laboratory include the following:

- power supply;
- working area;
- oscilloscope;
- welder base;
- welder station;
- socket;
- electric screwdriver;
- left red button;
- screwdriver;
- left green button;
- pliers;
- right red button;
- welder probe tip;
- right green button;
- oscilloscope probe tip;
- power supply cables;
- low voltage board;
- ground clip;
- high voltage board;
- battery charger connector;
- register;
- panel A;
- electric screwdriver battery.

Taking the videos depicting the operations performed by users, domain semantics, human actions and objects into account, 26 intents and 4 entity classes were defined.

The entity classes are structured as follows:

- **Object**: refers to the class of the object in the industrial laboratory;
- **Electronic board**: can refer to either high-voltage or low-voltage electronic board;
- **Component**: refers to object's components, as some objects are composed of multiple components (e.g. the high-voltage board includes a display);
- **Procedure**: represents a procedure that the user can perform. Two types of procedure are considered: test and repair.

After a study of the habits and the upcoming questions that can arise while working in an industrial laboratory, the following intents were considered:

- **greet**: greet and start a conversation
- **procedure_tutorial**: ask a specific question about the ongoing procedure;
- **object_warnings**: know if there are alerts for a specific object;
- **turn_object_on**: turn on an object;
- **turn_object_off**: turn off an object;
- **which_PPE_procedure**: know which PPE is required to perform a specific procedure;

- **which_PPE_object**: know which PPE is required to use a specific object;
- **object_instructions**: know how to use a specific object;
- **is_object_on**: find out if an object is turned on or not;
- **object_time**: find out how long an object has been used;
- **where_board**: know where a specific electronic board is located, or identify it on the working area;
- **board_detail**: know the location of a component on an electronic board;
- **where_object**: know where a specific object is located, or identify it on the working area;
- **object_detail**: know the location of a component on an object;
- **start**: start a procedure;
- **next**: hear the next step in the ongoing procedure;
- **previous**: hear the previous step in the ongoing procedure;
- **repeat**: hear the current step in the ongoing procedure;
- **all_objects**: know what objects are present in the laboratory;
- **ok_objects**: know what objects can be used;
- **on_objects**: know what objects are powered;

- **where_PPE**: know where the PPEs are located;
- **inform**: specify an entity;
- **bot_challenge**: knowing if the user is chatting with a bot or a human;
- **goodbye**: end the conversation;
- **out_of_scope**: this category includes all questions that are not relevant to the previous intents.

For each intent multiple example utterances were defined, representing different ways the user might express the same intent. Each example is further annotated to indicate the presence (or absence, as it is not mandatory) of any entities and their corresponding class. In total, 136 annotated examples were created across the 26 defined intents. Figure 4.3 presents a selection of intents along with their associated examples.

4.2 Approach

4.2.1 Object Recognition

For the visual component of the project, the Faster R-CNN object detector [48] was employed. Faster R-CNN is a two-stage object detector model that has proven effective in handling complex environments such as industrial settings. The choice of this specific detector is motivated by the need to reliably distinguish between visually similar objects that differ in function (e.g. battery charger connector, ground clip and power supply cables).

- **intent:** *object_warnings*
  examples:
  - show me the warnings related to the [oscilloscope]**(object)**
  - what are the alerts for the [power supply]**(object)**?
  - warnings for the [welder]**(object)**?
  - are there warnings for this object?
  - what warnings are there for this?
  - has the [electric screwdriver]**(object)** any warnings?

- **intent:** start
  examples:
  - start [low voltage]**(board)** electronic board [repair]**(procedure)** procedure
  - start [high voltage]**(board)** electronic board [test]**(procedure)** procedure
  - start [test]**(procedure)** [low voltage]**(board)** electronic board
  - begin the [repair]**(procedure)** of the [high voltage]**(board)** electronic board

Figure 4.3: Samples of the structured data used to train the NLU module. For each intent (blue) we designed different examples including different entities (red, green and purple).

Given an input image, the detector outputs a quintuple (x, y, w, h, c), where (x, y) represent the coordinates of the upper-left corner of the bounding box, w and h are the width and the height of the bounding box, respectively, and c is the class of the detected object. To enhance the relevance of the interaction, only the object closest to the center was considered, following the principle of visual attention.

4.2.2 Natural Language Understanding

The language model is based on RASA [49], an open-source framework for developing and deploying conversational assistants. RASA allows users to define a custom processing pipeline, enabling them to modify and replace compatible components according to specific use cases. For intent classifica-

tion and entity recognition, the pipeline incorporates DIETClassifier [50], a transformer-based model designed to handle both tasks jointly.

The proposed Natural Language Understanding (NLU) pipeline is composed of the following components:

- **SpacyNLP**: initializes SpaCy [51]'s internal structures, which are required by subsequent components;
- **SpacyTokenizer**: segments each message into tokens (words, puctuation, etc.) based on language-specific rules definedby the SpaCy model in use;
- **CountVectorsFeaturizer (CVF)**: generates bag-of-words representations of user messages, intents and responses. Two instances of CVF are used. The former analyzes messages at word-level, whereas the latter analyzes messages at character-level, considering n-grams raning from lenght 1 to 4;
- **SpacyFeaturizer**: produces dense vector representations of user messages and responses;
- **DIETClassifier (Dual Intent Entity Transformer)**: a multi-task transformer architecture responsible for both intent classification and entity recognition;
- **EntitySynonymMapper**: maps synonymous entity values to a canonical form, ensuring consistency in entity representation when synonyms are defined in the training data;

- **ResponseSelector**: selects the appropriate response from a predefined set of possible responses for those intents that were grouped into a macro-intent.

User utterances were annotated following the intent-entity paradigm. An intent represents the user's objective (e.g. greeting, turning on a device), while entities are key elements extracted from the message (e.g. object names, components). The systems considers 26 distinct intents (e.g. turn_object_on, object_instructions, start_procedure) and 4 entity classes, corresponding to the relevant objects and domain concepts.

At any point during the procedure, the conversational assistant must be able to switch context and respond to unrelated questions pertaining to different intents. To support this flexibility, the RASA's rule-based structure was adopted instead of a story-based approach. Rules define short conversation snippets that should always follow the same path, whereas stories are sequences of user-assistant exchanges that must occur in a specific order, making them less suitable for dynamic, multi-event environments.

4.3 System Architecture

HERO is composed of 4 main modules (Figure 4.4):

- **Input Module**: responsible for receiving utterances from the user and extract raw data (in form of text and images acquired from the first point of view) and forwarding it to the appropriate processing module;

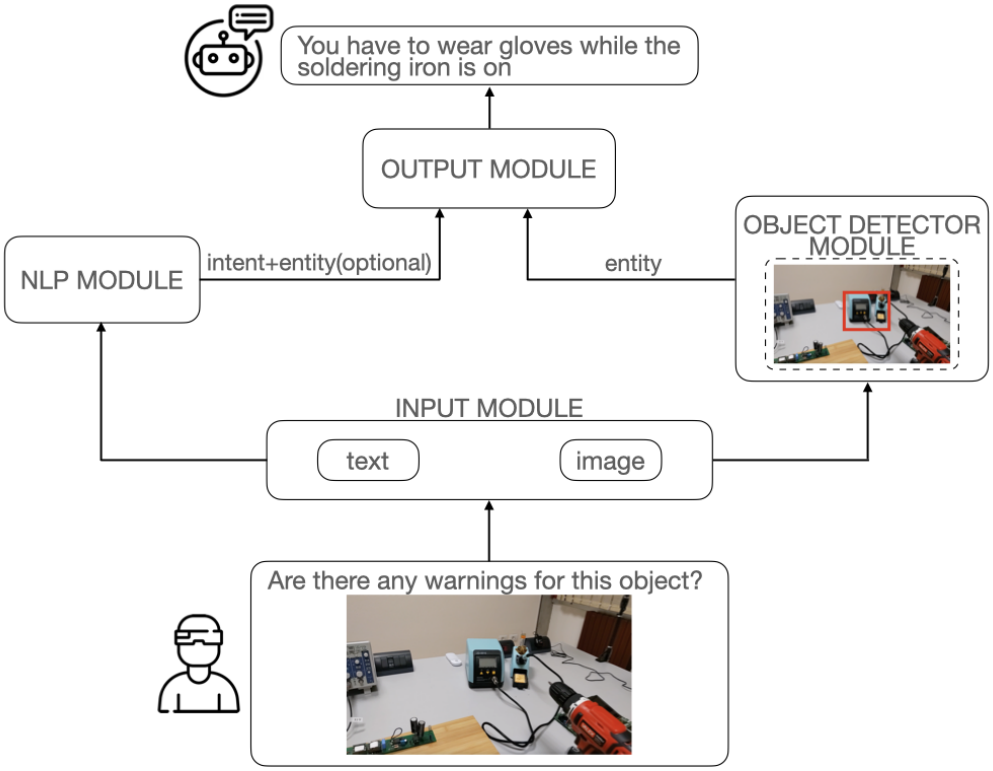

Figure 4.4: The architecture of the HERO system.

- **NLP Module**: processes textual input using the previously described pipeline, extracting intent and entities in the message;
- **Object Detection Module**: analyzes visual input to identify objects within the scene, returning the class of the object closest to the center as an entity;
- **Output Module**: chooses a coherent response for the user by combining the outputs from NLP and Object Detection Modules.

The architecture enables a natural, contextualized and multimodal interaction, with disambiguation capabilities supported by egocentric vision.

4.3.1 Vision-Language Fusion

One of the main innovations introduced by HERO is its ability to integrate visual and linguistic information. When a user formulates a question that lacks explicit entity references, despite such information being necessary to provide an accurate response (e.g. "How do I use this?"), the system analyzes the visual scene to identify the object at the center of the user's attention and links this contextual information with the request expressed through language, providing detailed and object-specific instructions.

4.4 Experiments

4.4.1 Evaluation of NLP module

The performances of the NLP module was evaluated using standard classification metrics: accuracy, precision and F_1-score. The results obtained are reported in Table 4.1.

Table 4.1: Results obtained by the NLP module considering the intent and entity classification of the DIETClassifier (first and second row) and the response selection performed by the ResponseSelector (last row).

	Precision	**Accuracy**	**F_1-score**
Entity Classification	0.944	0.959	0.866
Intent Classification	0.729	0.735	0.700
Response Selection	0.800	0.867	0.822

These results demonstrate that the system is capable of effectively understanding commands expressed through natural language, even in presence of

ambiguities or unstructured terminology.

4.4.2 Evaluation of Object Detection module

The visual module achieved strong performance in terms of object detection and classification accuracy, even under challenging conditions such as partial occlusions and variable lighting. To evaluate this capability, the COCO Mean Average Precision (mAP) metric with an Intersection Over Union (IoU) of 0.5 (mAP@50) was employed. The mAP measures the ability of the model to both localize and classify the objects in the image. The achieved mAP is 73.41%, which indicates that the module is able to correctly localize and recognize the objects present in the images, including small or partially occluded objects. Table 4.2 reports the AP values for each object class, allowing a detailed analysis of strengths and weaknesses across object categories, while Figure 4.5 presents qualitative examples of the model's visual predictions. These examples highlight both successful detections (e.g. oscilloscope, low voltage board, working area) and challenging cases (e.g. battery charger connector), providing a comprehensive view of the module's performances in realistic conditions. Notably, the object detector struggles with the battery charger connector (15.91 AP), given its small size and similarity to other objects such as the power supply cables and the ground clip.

4.5 System Deployment

For testing purposes, HERO was finally tested in a real-world context though a Telegram Bot interface, which enabled users to send images and messages

Table 4.2: Average Precision AP per class.

Category	**AP**	**Category**	**AP**
power supply	80.18	working area	90.18
oscilloscope	90.12	welder base	88.82
welder station	89.87	socket	90.27
electric screwdriver	81.45	left red button	100
screwdriver	58.73	left green button	100
pliers	79.18	right red button	81.82
welder probe tip	50.63	right green button	90.91
oscilloscope probe tip	51.72	power supply cables	41.34
low voltage board	88.53	ground clip	44.84
high voltage board	61.44	battery charger connector	15.91
register	71.07	panel A	89.77
electric screwdriver battery	51.72		

and receive immediate responses. This deployment demonstrated the system's portability and usability, highlighting its potential for procedural scenarios and mobile assistance applications. The experiment highlighted the system's potential to be deployed on mobile platforms for assisting users in the industrial laboratory.

4.6 Summary

In this chapter we presented HERO, a conversational assistant capable of interacting with users through a multimodal approach that combines natural language and egocentric vision. By leveraging the user's point of view,

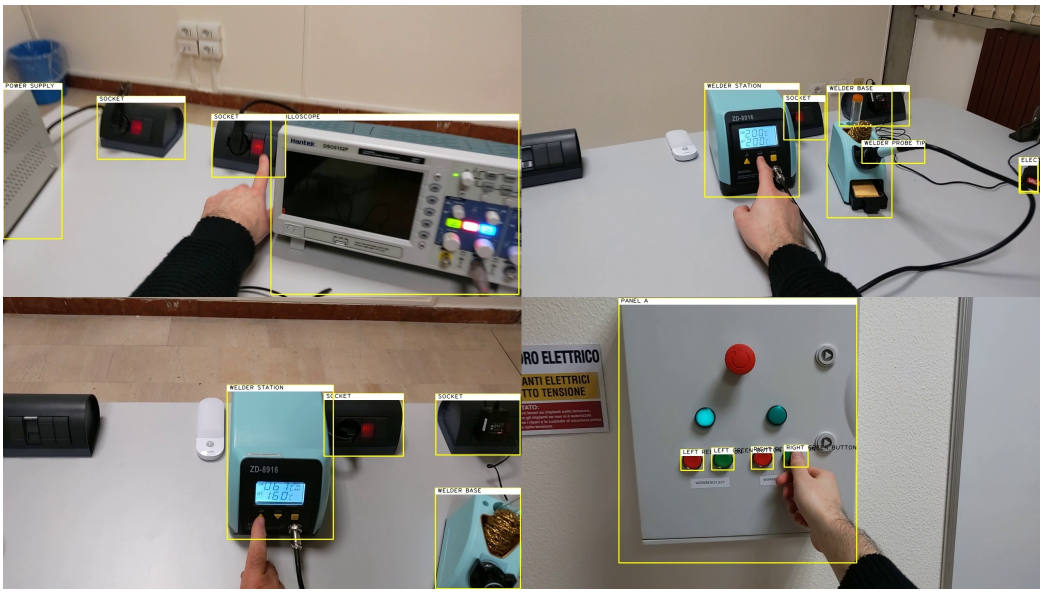

Figure 4.5: Qualitative results of the object detector.

HERO can interpret contextual cues, enabling more accurate and natural interactions. Its application in industrial settings demonstrates the potential of integrating conversational systems with wearable vision to support procedural tasks, reduce ambiguity in user commands, enhance workplace safety and improve overall efficiency.

Chapter 5

HEROv2: Deploying a Multimodal Assistant on Mobile Devices for Real-World Industrial Support

In the previous chapter we introduced HERO, a multimodal conversational assistant capable of combining natural language understanding and egocentric vision to support workers in repair and test activities within industrial environments. The system leveraged data acquired from wearable devices (such as Microsoft Hololens 2[47]), using a first-person point of view to enable natural and contextual interaction between the user and the system though natural language understanding and object detection.

In this chapter we introduce a second version of the system, which is conceived as a direct continuation of the previous version. From both the

architectural and functional perspectives, this version does not represent a radical redesign but rather a targeted evolution of the first. The main modules remain unchanged, while a new Question Answering (QA) module has been introduced. This module constitutes a first attempt at integrating Large Language Models (LLMs) into the system's pipeline, with careful consideration to avoid delegating certain critical intents to LLMs due to their known limitations, such as hallucinations. The processing pipeline remains consistent to the previous version and the interaction logic continues to rely on the same fusion principles based on language and vision.

The primary objective of this new version is to enhance system's accessibility and flexibility by enabling its use on smartphones and tablets, thereby eliminating the need for expensive and cumbersome wearable devices. This evolution responds to specific practical requirements that emerged during the analysis of usage context, where ease of access and portability are key factors for the adoption of this technology. This design choice allowed for deployment to a group of users, who were able to test the system and provide feedback on its usefulness, usability and the naturalness of its interactions. Figure 5.1 shows an example of a possible conversation with the assistant.

The main improvements in this version of the system focus on the new QA module and the expansion of the training dataset, which now includes more complex utterances that emerged during early testing phases and, more significantly, during the system's first operational deployment with real users. While the first version served primarily as a proof of concept, this second iteration marks a decisive step towards real-world applicability: the system was made available through a messaging interface on mobile devices and tested

in a controlled user study, where participants interacted with the assistant while carrying out industrial procedures.

This field validation phase represents the main contribution of the work discussed in this chapter. The aim is not to propose a completely new system from a technological standpoint, but to demonstrate its usefulness, usability and acceptance by real operators in a simulated industrial environment.

The findings and contributions of this work have been published in the following conference paper:

- Claudia Bonanno, Francesco Ragusa, Antonino Furnari, Giovanni Maria Farinella (2023). HERO: A Multi-Modal Approach on Mobile Devices for Visual-Aware Conversational Assistance in Industrial Domains [5]. In International Conference on Image Analysis and Processing (ICIAP) .

5.1 System Architecture

The system architecture broadly replicates that of the first version, with the addition of a new Question Answering (QA) module designed to handle specific user intents.

With the rise of Large Language Models (LLMs), integrating such architectures into our system appears highly promising, given their impressive capabilities in natural language generation. In the case of a general-purpose conversational assistant, this could significantly enhance the naturalness of the interaction, producing responses that closely resemble human communication. However, in an industrial context, the wide range of responses that

LLMs can generate may not be advantageous due to well-known issues such as hallucination.

During an industrial procedure, certain moments or steps require a high degree of precision to ensure tasks are completed accurately and safely. For this reason, we prioritize constrained and verifiable responses over open-ended, generative ones. Consequently, the QA module is designed to handle non-critical intents that do not pose a risk to the user's safety in the industrial domain.

Figure 5.2 illustrates the architecture of this second version of HERO.

In particular, the QA module leverages the OpenAI API to prompt the *text-davinci-003* model. It provides the model with contextual information related to a specific procedure, along with the user's question regarding that procedure, in order to generate an appropriate response. This output is then forwarded to the Output module.

The system was deployed on Facebook Messenger through the use of RASA's channel connector, enabling seamless interaction via mobile devices.

5.2 Dataset

5.2.1 Dataset for the NLP and QA modules

For the training of the language understanding module, the original dataset was extended to include a total of 151 utterance examples, each associated to one of 24 specific intents. Compared to the previous version, the *bot_challenge* and *goodbye* intents were removed due to their limited usefulness in real-world

Table 5.1: Intent and entity classification results obtained by the NLP module.

	Precision	**Accuracy**	F_1**-score**
Entity Classification	0.944	0.959	0.866
Intent Classification	0.729	0.735	0.700

applications.

5.2.2 Dataset for the Object Detection module

The object detection module was trained on a dataset consisting of 16,000 annotated images, extracted from 42 egocentric videos recorded during activities in an industrial environment, following the same methodology as in the previous version. The resulting dataset includes approximately 90,000 object instances. This represents a significant increase compared to the previous version, which relied on fewer videos and ultimately resulted in the extraction of only 20,000 objects.

5.3 Evaluation

5.3.1 NLP Module

The system was evaluated using standard performance metrics, namely precision, accuracy and F_1-score measures. Table 5.1 reports the results obtained for both entity classification (first row) and intent classification (second row).

5.3.2 Object Detector Module

The system was evaluated using the COCO Mean Average Precision (mAP) metric, with an Intersection over Union (IoU) of 0.5 (mAP@50). The model achieved a mAP of 82.57%, indicating a strong ability to accurately localize and recognize the objects present in the images.

5.3.3 User Study

The most significant contribution of the new version lies in its direct experimentation with end users, a crucial step for assessing the system's practical effectiveness and user acceptance in realistic scenarios. To this end, a user study was conducted involving a diverse group of 11 participants. Some participants were already familiar with the laboratory, its procedures and instruments, while the majority had no prior exposure to the environment. Each participant completed two tasks, performed consecutively: one with the support of HERO and the other using a traditional paper-based manual. The manual contained the same information accessible through HERO, covering the laboratory, its objects, and procedures, and was structured into four sections: *Boards*, *Objects*, *Personal Protection Equipment (PPE)* and *Procedures*.

Each procedure consisted in 10 instructions, such as "Set the soldering iron temperature to 480°C using the yellow UP button" or "Fix the electronic board to the working area using the electric screwdriver". Notably, the two procedures involved non-overlapping sets of objects, ensuring that users engaged with different tools and steps in each case.

Participants interacted with HERO using their smartphones, taking pictures of industrial instruments and submitting questions in natural language. The system responded with contextual information, step-by-step instructions, usage guidelines and safety alerts.

Upon completing each procedure, participants were asked to fill out a questionnaire to evaluate various aspects of the experience. The first questionnaire focused on assessing user satisfaction, perceived usefulness and usability of HERO. The second questionnaire aimed to compare the two experiences, contrasting the use of HERO with the traditional manual.

The order in which participants completed the two procedures was randomized. However, each user was instructed to perform the first task using HERO and the second using the paper manual, ensuring that each procedure was completed with either system approximately 50% of the time. Tables 5.2 and 5.3 list the questions included in the two questionnaires.

The distribution of responses to the questionaires is visualized using boxplots, as shown in Figure 5.3. We include only questions that require to express a score in a range from 1 to 5. As shown by the boxplots, the participants expressed satisfaction with the overall experience and response time (questions 1.1 and 1.10). However, the natural language component was perceived less natural as compared to the vision component (questions 1.2 and 1.4). Some participants reported difficulty in formulating their queries in a way that the system could understand (question 1.7, which has a median score of 3), but found the retrieved information useful and easy to understand (questions 1.8 and 1.9, which have a median score of 4). Even though a few participants expressed a preference for a different type of device (e.g.,

Table 5.2: List of questions included in the questionnaire aimed to evaluate the subjects' degree of satisfaction with our proposed system.

ID	Question
1.1	How satisfied are you overall with the experience in a range from 1 to 5? 1-definitely not satisfied, 5-definitely satisfied
1.2	How natural did you find the interaction with the app in a range from 1 to 5? 1-definitely not natural, 5-definitely natural
1.3	How often did you use the photo sending feature to communicate with the bot? a-never, b-once, c-more than once
1.4	How natural did you find this feature (if you didn't use this feature, you can skip this question) in a range from 1 to 5? 1-definitely not natural, 5-definitely natural
1.5	How helpful do you think the technology demonstrated in this application prototype can be in a range from 1 to 5? 1-definitely not helpful, 5-definitely helpful
1.6	Do you think the technology demonstrated in this prototype can be used in other contexts besides the industrial context? a-yes, b-no
1.7	How often did the system correctly recognize the intent of your questions in a range from 1 to 5? 1-never, 5-each time
1.8	How useful do you think the information received from the application is in a range from 1 to 5? 1-definitely not useful, 5-definitely useful
1.9	How clear do you think the information received from the application is in a range from 1 to 5? 1-definitely not clear, 5-definitely clear
1.10	How satisfied are you with the system response time in a range from 1 to 5? 1-definitely not satisfied, 5-definitely satisfied
1.11	How useful do you think it is for the application to be available on the phone rather than another device (wearable devices, tablets, fixed screens) in a range from 1 to 5? 1-I'd prefer a different device, 5-I prefer a mobile device
1.12	Would you prefer a version with voice dictation? a-yes, b-no

wearable device), most of the participants appear to be satisfied with the mobile application (question 1.11). The participants generally believed that this technology could be useful in various contexts besides the industrial one

Table 5.3: List of questions included in the questionnaire aimed to compare the two experiences.

ID	Question
2.1	Which experience satisfied you the most in a range from 1 to 5? 1-definitely the application, 5-definitely the paper-based manual
2.2	How convenient did you find the use of the paper-based manual in a range from 1 to 5? 1-definitely not convenient, 5-definitely convenient
2.3	How much do you think the technology demonstrated in this application prototype could support you, compared to the use of the paper-based manual in a range from 1 to 5? 1-I found the manual more supportive, 5-I found the application more supportive
2.4	Which tool allowed you to complete the instructions more quickly in a range from 1 to 5? 1-I found the manual as the quickest tool, 5-I found the application as the quickest tool
2.5	How useful do you think the information received from the application is compared to the information obtained through the paper-based manual in a range from 1 to 5? 1-I found the manual instructions more useful, 5-I found the application instructions more useful
2.6	Which tool provided clearer instructions in a range from 1 to 5? 1-I found the manual instructions clearer, 5-I found the application instructions clearer
2.7	Which experience did you prefer overall? a-the use of the application, b-the use of the paper-based manual

(we obtained 100% "yes" preferences in question 1.6). Most users noted that voice dictation can be useful (81.8% "yes" preferences on question 1.12). Despite the participants' pre-existing bias and familiarity with the industrial laboratory, when completing a procedure using a paper-based manual after completing the first procedure using the app, they still preferred our system over the manual. The participants found our system more satisfying, convenient, and quicker to use than the manual (questions 2.1, 2.2, 2.3, and 2.4 in Fig. 5.3 (right)). Although the information contained in the paper-based

manual and in the chatbot's responses was identical, the participants perceived our system's responses as more useful and clearer (questions 2.5 and 2.6). Overall, the participants preferred the chatbot experience (90.9% users replied "the use of the application" to question 2.7).

The results of the questionnaires indicate the following:

- Users appreciated the multimodal nature of the system (combining images and text), considering it an effective solution to disambiguate information;
- The mobile version was perceived as accessible and immediate. However, some participants suggested complementing it with a wearable version, as a hands-free interaction modality was considered advantageous in certain situations;
- Users expressed belief that this technology could be beneficial in contexts beyond the industrial domain;
- Despite some participants' prior familiarity with the industrial laboratory, they consistently preferred using HERO over the paper-based manual, even though the manual was used after the system;
- The system was regarded as more satisfying, convenient and faster to use than the traditional paper-based manual;
- A significant majority (90.9%) of participants stated a preference for interacting with HERO over consulting a printed manual when seeking information about objects, procedures and the laboratory itself, despite the fact that both media contained the same information.

In Fig. 5.4 we report an analysis based on the Spearman's rank correlation coefficient since it's more robust to non-linear correlations that are still monotonic and outliers compared to Pearson's correlation. The figure shows only correlation values which have been found to be statistically relevant (p-value < 0.05). We have identified strong correlations between users' satisfaction with the proposed system (question 1.1), their perception of the technology as helpful (question 1.5), and their evaluation of the clarity and usefulness of responses (questions 1.9, 1.8) - Spearman correlation values of 0.7, 0.73 and 0.55 respectively. Similarly, we found a strong correlation between the system's ability to recognize user intent correctly (question 1.7), the users' preference for it over other devices (question 1.11), as well as their perception on the interaction as natural (question 1.2) - Spearman correlation values of 0.61 and 0.68 respectively. Moreover, the correlations in Fig. 5.4 (right) suggest that the more the users prefer our system over the paper-based manual (question 2.1), the more they perceive it as convenient (question 2.2), supportive (question 2.3), of quicker use (question 2.4), and useful (question 2.5), even though the information provided by both sources is identical - Spearman correlation values of 0.64, -0.7, -0.55 and -0.54 respectively.

This study provides concrete evidence supporting the validity and effectiveness of the new system in assisting during industrial activities. It also offers valuable insights into potential avenues for improvements and scalability.

5.4 Summary

The new version of HERO does not represent a radical evolution over the previous one, however it constitutes a fundamental step towards real-world deployment. While the first version served to demonstrate the technical validity of the multimodal paradigm and the proposed architecture, this second iteration focuses on actual user experience, providing a working, accessible system tested in realistic conditions.

The added value of this work lies not in its technological innovation per se, but in the practical translation of the technology into a real environment: a necessary step to demonstrate the system's maturity, gather user feedback and guide the development of future versions.

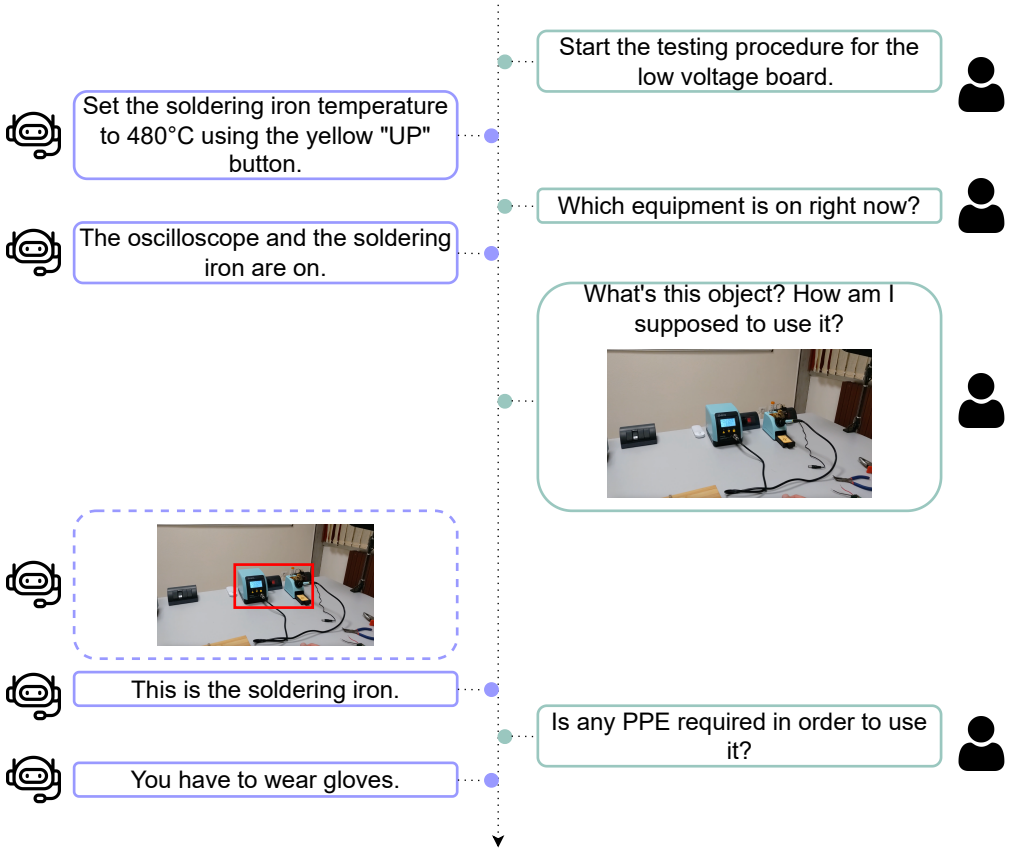

Figure 5.1: Concept of the proposed assistant, HERO. HERO can support users to achieve their goals in the considered industrial setting. The user (right) can ask for instructions on a given procedure, retrieve information on the state of specific objects ("Which equipment is on"), ask visually-grounded questions ("What's this object?") and obtain general information about safety instructions ("Is any PPE required?"). HERO replies using natural language (left), while an object detection module is used to answer visually-grounded questions (e.g., by recognizing the soldering iron).

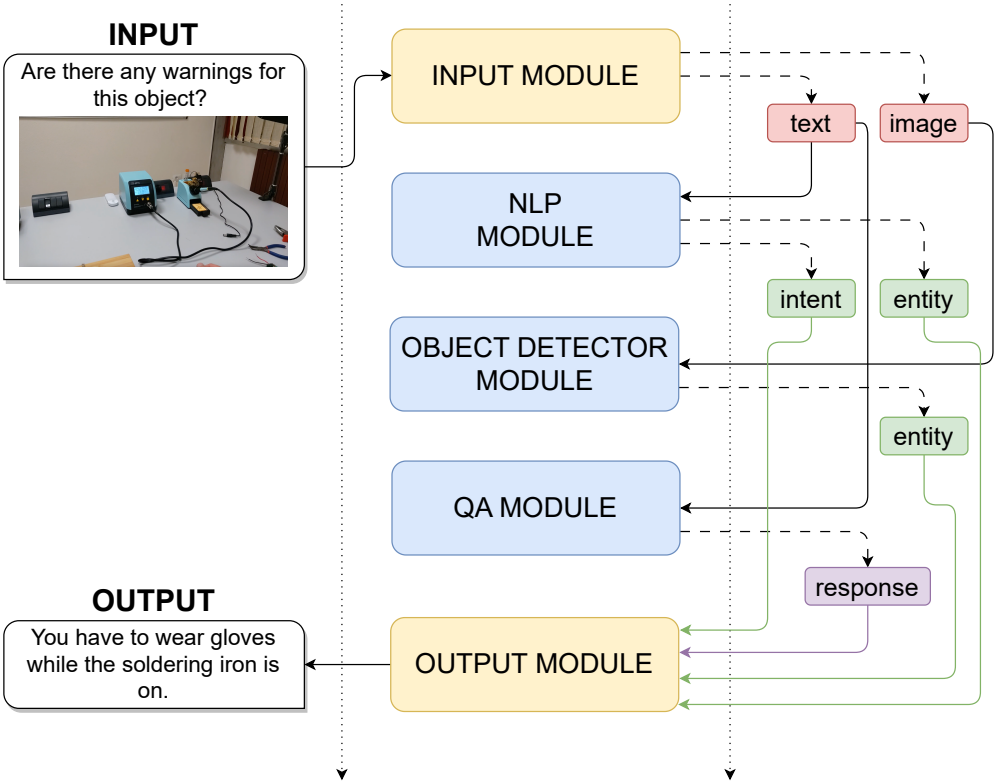

Figure 5.2: The architecture of the proposed HERO system. The left column lists two utterances (input/output) exchanged by the user and the system, while the central column lists the main modules of HERO. Yellow boxes are used for modules that extract/assemble data, while blue boxes represent modules that process raw data to obtain higher level information. The right column shows extracted and processed data, where red boxes denote raw data and green and purple boxes indicate output high level information. Given an utterance comprised of an image and some text, the input module extracts raw data such as text and the images. The NLP module receives text from the input module, predicts an intent and extracts entities if present. The object detector module extracts the class of the object closest to the center of the image sent by the input module as an entity. If the intent of the user is to obtain information on a procedure, the QA module generates a custom response from text sent by the input module. The output module selects the final response based on high level information received by the NLP, object detector and QA modules.

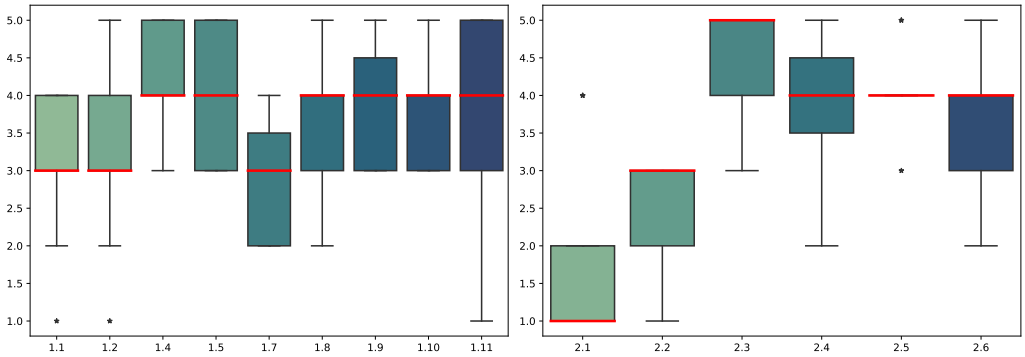

Figure 5.3: Boxplot of the results for each question that required a discrete answer in the first questionnaire (left) and in the second questionnaire (right).

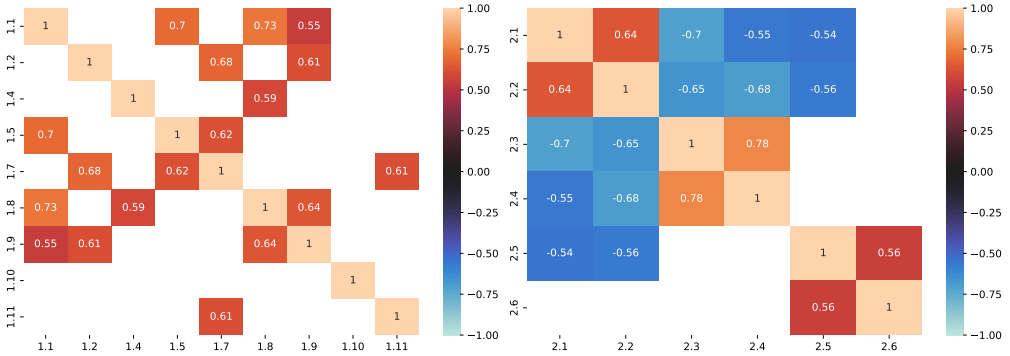

Figure 5.4: Filtered Spearman correlation with p-value<0.05 for the first questionnaire (left) and the second questionnaire (right).

Chapter 6

HERO-GPT: Enabling Zero-Shot Conversational Support in Industrial Domains through Large Language Models

After outlining the evolution of the HERO system and its first two versions in the previous chapters, with particular attention to its multimodal capabilities and its deployment in both simulated and real-world scenarios, this chapter introduces a further advancement of the HERO paradigm, culminating in the development of HERO-GPT.

This new version marks a methodological turning point. While mintaining the original goal of providing multimodal conversational support in indus-

trial contexts, HERO-GPT moves away from the traditional use of intents, entities, and manually trained models. Instead, it adopts a modular architecture powered by Large Language Models (LLMs) and a dynamic Knowledge Base (KB) composed of contextual documents.

The primary objective of HERO-GPT is to enable zero-shot interaction, eliminating the need for domain-specific training. The system is designed to flexibly adapt to new context by simply updating or replacing the documents that serve as the source for the knowledge base, without requiring architectural modifications or additional training sessions.

Moreover, in comparison with earlier versions, HERO-GPT introduces a novel Multi-Agent System architecture. In this configuration, multiple modules (some LLM-powered, others not) collaborate to interpret user requests, access visual context, retrieve information from technical documentation, and generate a coherent responses. Rather than relying on datasets, hand-crafted rules, or example dialogues, the system leverages dynamically interchangeable documents to supply information within a given scenario. As a result, HERO-GPT requires minimal domain-specific data during setup. Figure 6.1 illustrates an example interaction with the system.

This chapter presents the details of HERO-GPT, with a focus on its system architecture, the integration of document-based sources, the role of visual inputs, and the evaluation of its effectiveness through a user study, following a methodology similar to the one adopted for the previous version of HERO.

The findings and contributions of this work have been published in the following conference paper:

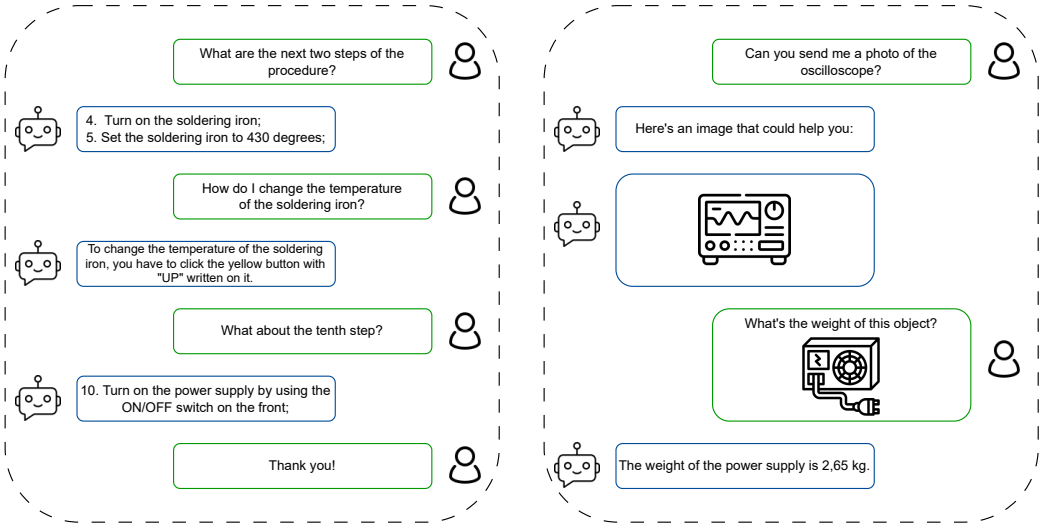

Figure 6.1: Examples of interactions between users and HERO-GPT. Left: users can ask information on procedures and objects through textual interactions. Right: the system also allows for multi-modal interactions, giving information about objects recognized from visual observations and providing images as responses.

- Luca Strano, Claudia Bonanno, Francesco Ragusa, Giovanni Maria Farinella, Antonino Furnari (2024). GPT-Assist: Zero-Shot Conversational Assistance in Industrial Domains Exploiting Large Language Models [6]. In International Conference on Image Processing and Video Engineering (IMPROVE) .

6.1 System Architecture

HERO-GPT's architecture is composed of 5 main modules:

- **Router Module**: manages a set of context-agnostic courtesy intents and, in case of a different intent, routes the user input to the appropriate module;

- **GPT Manager Module**: coordinates the generation of textual responses, based on a Retrieval Augmented Generation (RAG) approach;
- **ObjectDetector Module**: recognizes objects inside images provided by the user;
- **ImageManager Module**: retrieves relevant images from the Knowledge Base;
- **ProcedureManager Module**: manages step-by-step execution and explanation of procedures.

Figure 6.2 illustrates a detailed scheme of each module within the system architecture. Several of the main modules are supported by multiple LLM-based subcomponents, responsible for different natural language understanding tasks. The modules interact via a central Router, which decomposes the multimodal request (text + images) and constructs customized prompts for the language models involved. Each main module may also rely on secondary components, shown as dashed boxes in Figure 6.2.

The system leverages a Knowledge Base composed of documents (e.g. PDF manuals, procedure descriptions, or best practices) and images relevant to the target environment. These documents undergo a pre-processing stage in which they are divided into smaller chunks and indexed using vector representations generated by an embedding model. All LLM-related operations are handled via the LangChain [1] framework.

HERO-GPT retains and expands upon the multimodal capabilities introduced in previous versions. Users can:

[1] https://www.langchain.com

- Submit images of unknown objects to receive explanations or instructions;
- Request contextual information (e.g. "What's the next step?") related to a specific procedure;
- Ask for explicative images (e.g. "Show me a picture of the oscilloscope");
- Use visual inputs to disambiguate linguistic references.

The adopted approach for fusing natural language and visual information is robust: images are analyzed by an object detector (based on Faster R-CNN[48]), which identifies the object closest to the image center, extracts its class, and forwards this information to the language module.

This approach allows the system to:

- Answer specific questions related to ongoing procedures;
- Access heterogeneous technical documentation;
- Operate in entirely new scenarios by simply updating the documents in the Knowledge Base.

6.1.1 Router Module

A set of courtesy intents is defined to help familiarize users with the assistant's functionalities. Courtesy intents, namely *user_greet*, *user_start*, *user_deny*, *user_bot_challenge* and *user_send_image*, are standard and remain consistent across different application contexts. Recognizing these intents requires a

brief training phase using the standard RASA's [49] Natural Language Processing pipeline. User utterances categorized under these intents are managed through RASA's rule-based system (e.g. the assistant will greet the user when the *user_greet* intent is detected).

Any other inquiry that doesn't align with these predefined intents is forwarded to the Router Module. The module leverages the general-purpose language understanding capability of Large Language Models, thereby avoiding the need of context-specific intent classification. Instead, it accurately routes the user's query to the appropriate module based on the inferred intent category.

Intent categories encompass:

- Procedures (e.g., "What's the next step?");
- Images (e.g., "Can you send me an image of the oscilloscope?");
- Questions (e.g, "How do I turn on the soldering iron?");
- Visual Questions (e.g., "What's this object needed for?")

Queries falling under these categories are respectively routed to the ProcedureManager, ImageManager, GPTManager and Dispatcher modules for further processing.

6.1.2 ProcedureManager Module

HERO-GPT is capable of outputting specific steps from a selected procedure contained in the Knowledge Base. A procedure is defined as a sequence of steps required to accomplish a particular objective. This concept is broadly

applicable across various domains. For instance, in a culinary setting, a procedure may correspond to a cooking recipe; in an industrial setting, it may refer to the repair procedure of a high voltage board.

Once a user initiates a procedure though a natural language query, they have the option to request the previous or next steps, as well as specify a particular step. Procedures are sourced from documents inside the Knowledge Base that are marked with the "procedure" keyword.

To interpret the user's request, the LLM instance is tasked with generating a JSON object containing command (next, previous or specific) and steps number. For instance, if the user asks "What are the next four steps?", the Language Model is expected to return a JSON object containing the "next" command and the integer value 4.

This JSON object is subsequently processed by the ProcedureManager complementary module (named P.M. Output Processor in Figure 6.2). This module accesses the procedure loaded in memory, retrieves the appropriated steps, and presents them to the user.

6.1.3 ImageManager Module

HERO-GPT possesses the capability of forwarding images sourced from the Knowledge Base upon user request. This functionality proves to be especially valuable when users are unfamiliar with their environment; indeed, visual information often grants better assistance compared to Natural Language responses.

When a user requests a visual output, the LLM instance is prompted to

select the most relevant image based on the user's query. Image search relies on filenames for retrieval. Lastly, the ImageRetriever Module retrieves the selected image from the Knowledge Base and forwards it to the user.

6.1.4 GPTManager Module

The GPTManager Module coordinates the generation of Natural Language responses to user queries. HERO-GPT's responses are generated through the use of a Retrieval Augmented Generation (RAG) approach, which retrieves the contextual information required to correctly answer the user query from the Knowledge Base.

To minimize the number of calls to the LLM instance, the GPTManager Module forwards the incoming query to the HistoryManager, which caches questions and related previously generated responses. If the current query is sufficiently similar to an already cached question, the associated response is directly returned to the user, bypassing further processing.

If no match is found, the query is forwarded to the EntityExtractor, Retrieval Augmented Generation and Language Model Modules. The EntityExtractor Module uses an appropriate prompt to extract key entities from the user input. The RAG Module computes a similarity measure to retrieve the k most similar documents chunks to the query. Subsequently, a prompt is dynamically constructed by incorporating the retrieved document chunks along with the user's query. Lastly, the formatted prompt is forwarded to the LLM instance to generate contextually relevant responses. The user input, along with the extracted entities, associated response and other relevant

information is forwarded to the HistoryManager Module, which caches the response for future use and outputs it to the user.

6.1.5 ObjectDetector Module

When the Router Module detects a visual question (i.e., a question complemented with an image), the whole bundle is sent to the Dispatcher Module, which forwards the textual part to the GPTManager and makes use of the ObjectDetector Module to extract the appropriate entity (i.e., the object's identity) from the image.

The Object Detector deployed for this module consists of a two-stage Object Detector Faster R-CNN [48]. The ObjectDetector Module extracts the class of the closest object to the center of the input image (the one the user is likely looking at) as an entity and forwards it to the GPTManager Module. Subsequently, the GPTManager Module constructs a prompt that incorporates the received entity along with every other necessary contextual information (see Figure 6.3).

It is noteworthy that, while the object detector may need to be trained on domain-specific images and object classes, given the modular nature of the system, the described module could be implemented with an Open Vocabulary Object Detector or a vision-capable LLM, such as GPT-4V[2].

The architecture's fundamental feature lies on the distinction between traditional and LLM-powered modules (as GPTManager), keeping balance between computational power and operative flexibility.

A key aspect of innovation introduced in HERO-GPT lies in the use of un-

[2] https://openai.com/research/gpt-4v-system-card

structured technical documentation (PDF manuals, object lists, procedures, etc.) indexed through vector embeddings. The system dynamically queries this Knowledge Base using the Retrieval Augmented Generation paradigm. This approach allows the system to reason on context and produces personalized responses also in never-seen-before environments, without requiring architectural modifications or model retraining.

6.2 Deployment and Experimental Evaluation

The performance of the proposed system was evaluated through a user study. HERO-GPT was deployed as a Telegram bot and tested in a mock-up industrial laboratory with a group of 12 volunteers. The participants were asked to carry out given procedures with the system's support and to compare their experience to both a paper-based manual and the previous version of HERO.

To assess system performances, each volunteer was asked to complete 2 procedures, each consisting of 10 steps. These procedures were randomly assigned to volunteers from a set of four procedures involving activities such as repairing a low voltage board and testing a high voltage one.

We performed two sets of tests. The first one aims to assess the usefulness of HERO-GPT when compared to traditional supporting materials, such as paper-based manuals. For these tests, one of the two assigned procedures was performed with the support of HERO-GPT, whereas the other one was performed with the support of a classic paper-based instruction manual.

After testing the system, the participants were asked to fill two questionnaires: the first report to be filled was focused on assessing user's satisfac-

tion degree of the assistant itself, whereas the second one sought feedback on whether the assistant was deemed superior and more user-friendly compared to the classic paper instruction manual. The second set of tests aimed to assess the degree of satisfaction of the user with respect to the previous version following the traditional protocol based on manual definition of intents, entities, and standard answers. We adopt the same protocol for this tests, asking subjects to perform one of the two activities supported by paper manuals and the other one supported by the previous version. The content of the questionnaires used for both evaluation phases is shown in Tables 6.1 and 6.2.

Figure 6.4 presents the distribution of answers to questions requiring to express a satisfaction score, in comparison with those obtained using the previous version of the system. As shown in the boxplots, overall satisfaction is higher with our proposed system (Question 1.1 - compare top - previous version - to bottom - HERO-GPT). The naturalness of the system is also superior to the previous version (Question 1.2), but participants expressed a preference for the answers and the intent recognition mechanism to visual questions implemented in the previous version (Question 1.4). Similarly, HERO-GPT's intent recognition achieved a slightly lower score compared to the previous version (Question 1.7). This result is expected, given that the previous version's intent recognition component is tailored for the considered context. Usefulness and clarity of answers both obtained higher scores with our proposed system (Questions 1.8 and 1.9). This outcome is attributed to the Language Model's capability to enhance responses by providing additional details on some of the questions proposed by our users. The previous

Table 6.1: Questionnaire 1.

ID	Question
1.1	How satisfied are you overall with the experience in a range from 1 to 5? 1-definitely not satisfied, 5-definitely satisfied
1.2	How natural did you find the interaction with the app in a range from 1 to 5? 1-definitely not natural, 5-definitely natural
1.3	How often did you use the photo sending feature to communicate with the bot? a-never, b-once, c-more than once
1.4	How natural did you find this feature (if you didn't use this feature, you can skip this question) in a range from 1 to 5? 1-definitely not natural, 5-definitely natural
1.5	How helpful do you think the technology demonstrated in this application prototype can be in a range from 1 to 5? 1-definitely not helpful, 5-definitely helpful
1.6	Do you think the technology demonstrated in this prototype can be used in other contexts besides the industrial context? a-yes, b-no
1.7	How often did the system correctly recognize the intent of your questions in a range from 1 to 5? 1-never, 5-each time
1.8	How useful do you think the information received from the application is in a range from 1 to 5? 1-definitely not useful, 5-definitely useful
1.9	How clear do you think the information received from the application is in a range from 1 to 5? 1-definitely not clear, 5-definitely clear
1.10	How satisfied are you with the system response time in a range from 1 to 5? 1-definitely not satisfied, 5-definitely satisfied
1.11	How useful do you think it is for the application to be available on the phone rather than another device (wearable devices, tablets, fixed screens) in a range from 1 to 5? 1-I'd prefer a different device, 5-I prefer a mobile device
1.12	Would you prefer a version with voice dictation? a-yes, b-no
1.13	How often did the system correctly recognize the object in a photo you submitted in a range from 1 to 5? 1-never, 5-every time
1.14	Which version did you prefer the most? a-the previous version, b-today's version, c-no preference.

Table 6.2: Questionnaire 2.

ID	Question
2.1	Which experience satisfied you the most in a range from 1 to 5? 1-definitely the paper-based manual, 5-definitely the application
2.2	How convenient did you find the use of the paper-based manual in a range from 1 to 5? 1-definitely not convenient, 5-definitely convenient
2.3	How much do you think the technology demonstrated in this application prototype could support you, compared to the use of the paper-based manual in a range from 1 to 5? 1-I found the manual more supportive, 5-I found the application more supportive
2.4	Which tool allowed you to complete the instructions more quickly in a range from 1 to 5? 1-I found the manual as the quickest tool, 5-I found the application as the quickest tool
2.5	How useful do you think the information received from the application is compared to the information obtained through the paper-based manual in a range from 1 to 5? 1-I found the manual instructions more useful, 5-I found the application instructions more useful
2.6	Which tool provided clearer instructions in a range from 1 to 5? 1-I found the manual instructions clearer, 5-I found the application instructions clearer
2.7	Which experience did you prefer overall? a-the use of the application, b-the use of the paper-based manual

version achieved a faster response time compared to HERO-GPT (Question 1.10) due to real-time response generation in the latter. During the testing phase of HERO-GPT, 83.3% of participants repeatedly used the photo-sending feature to communicate with the assistant (Question 1.3) with an accuracy of about 88% (Question 1.13), demonstrating the essential role of multi-modality in modern AI assistants. The entirety of participants believed that our assistant can be used in other contexts (Question 1.6), while only 30% favored the previous version over our proposed system (Question 1.14, with 50% preferring our assistant, and the remaining 20% expressing no

preference). Lastly, participants exhibited a preference for the proposed assistants over the provided paper instruction manuals (Questions 2.1 through 2.6) in both tests, with 100% of participants demonstrating a preference for one of the assistants.

The results show that:

- A majority of users prefer HERO-GPT over traditional support methods such as paper-based manuals;
- The system performs on-par with previous versions, despite requiring significantly less domain-specific data;
- The system is perceived as clearer, more useful and more natural in its interactions;
- Multimodal interaction is considered a fundamental feature by users, enabling them to combine visual and textual input seamlessly to achieve better task understanding;
- The document-based approach proves to be effective while requiring minimal initial configuration, thus facilitating quick deployment in real-world scenarios;
- The majority of users (80%) preferred the use of conversational systems to get assistance in our real-world application over a paper-based manual, while the remaining 20% expressed no preference.

6.3 Summary

HERO-GPT represents a significant advancement in the evolution of HERO, maintaining its focus on multimodal conversational assistance for industrial environments, while introducing an innovative architecture based on Large Language Models and contextual documents. The system enables highly flexible zero-shot interaction, making it adaptable to a wide range of scenarios, drastically reducing developing and adaptation costs. The results of the user study confirm the validity of this paradigm, paving the way for future developments involving wearable devices and voice-based interfaces.

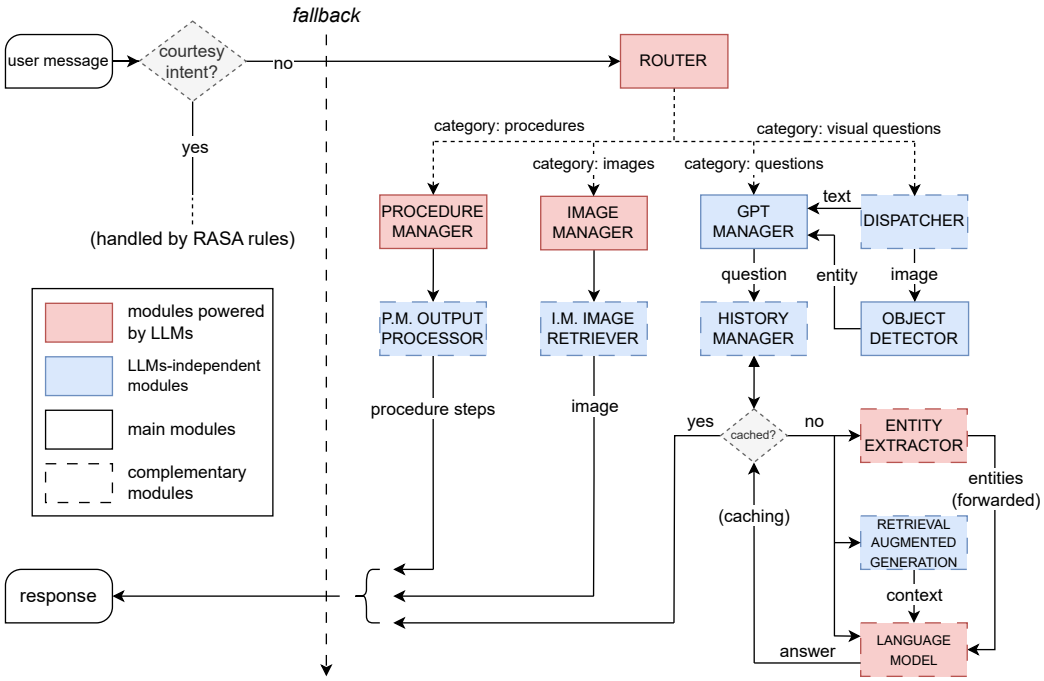

Figure 6.2: High level overview of HERO-GPT's Multi-Agent system. Red boxes represent LLM-powered modules, whereas blue boxes delineate LLM-independent modules. Solid contour represents main modules, while dashed lines depict complementary modules. The diagram illustrates the five principal modules within the system: 1) The Router Module has the role of choosing the appropriate path to fulfill the request; 2) The GPTManager Module is responsible for managing relationships between complementary modules tasked with Natural Language output generation (bottom right modules on the figure); 3) The ObjectDetector Module is designed to identify entities within images submitted by users; 4) The ImageManager Module retrieves images from the Knowledge Base depending on user input; 5) The ProcedureManager Module has the role of retrieving the desired steps of the initiated procedure. See text for additional details.

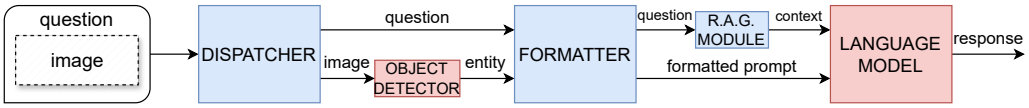

Figure 6.3: Architecture of the Image-to-Prompt system. The Dispatcher Module divides user input into question and image. The former is forwarded through the Formatter Module, which is part of the GPTManager Module, into the Retrieval Augmented Generation (R.A.G.) Module to retrieve context, while the latter is forwarded to the Object Detector Module, which outputs the class of the closest object to the center as an entity. Finally, context and formatted prompt (e.g. "question about (entity): (question)") get integrated and transmitted to a Language Model, which generates the response.

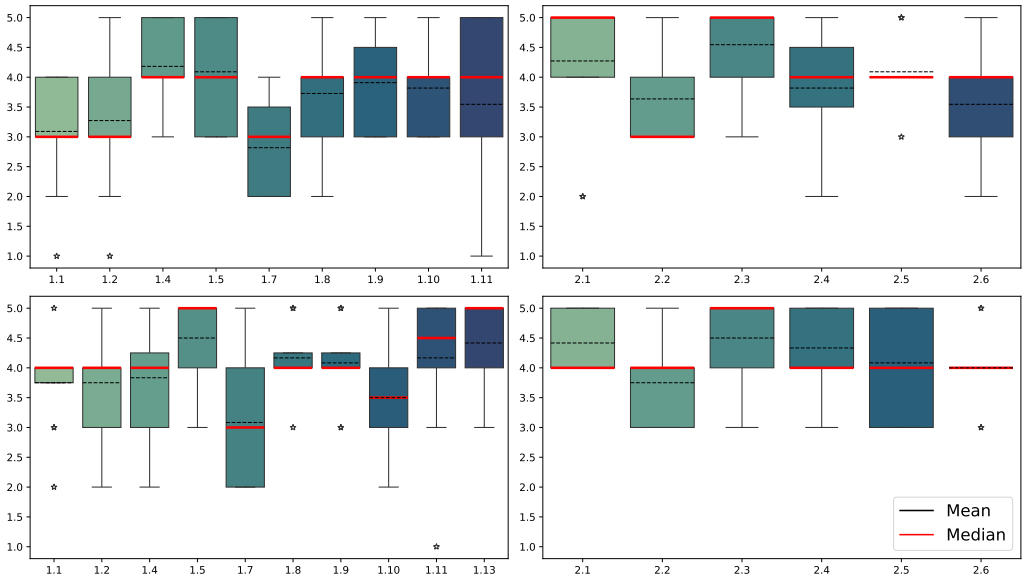

Figure 6.4: Distribution of satisfaction scores. Plots positioned on the left represent responses from the first questionnaire, whereas plots on the right illustrate responses from the second questionnaire. The upper boxplots correspond to the previous version, while the lower boxplots pertain to HERO-GPT. Please refer to the supplementary material for additional discussion and visualizations.

Chapter 7

Expanding Industrial NLU Datasets with Synthetic User Utterances

In the previous chapters various versions of HERO were discussed, with a focus on the different iterations of the systems, aiming at benefitting the naturalness and usefulness of interactions. A key component of each system is the comprehension of natural language, obtained by using typical user utterances in the laboratory as a starting point.

However, during the design and testing of the systems, the dimension of the utterances dataset represented a core problem. Even though language allows a potential user of expressing the same question in different ways, the dataset we came up with was biased on the few people who tested the system and their own way of expressing their intents. For this reason, and since we already extended the dataset after the first version's test, we thought of other

ways of further expanding the dataset.

Due to the recent surge of popularity of Language Models into the general public, we explored the possibility of creating synthetic utterances data by prompting LMs and tested if this addition would benefit our real dataset.

This chapter focuses on the contribution in the language annotations made for the ENIGMA-51 dataset, published in the following conference paper:

- Francesco Ragusa, Rosario Leonardi, Michele Mazzamuto, Claudia Bonanno, Rosario Scavo, Antonino Furnari, Giovanni Maria Farinella (2024). ENIGMA-51: Towards a Fine-Grained Understanding of Human-Object Interactions in Industrial Scenarios [7]. In IEEE Winter Conference on Application of Computer Vision (WACV).

7.1 Data Collection

To enrich our dataset of real utterances, we generated similar utterances through the prompting of ChatGPT [1], obtaining 100 unique utterances for each of the 24 intents. Due to the unique structure of the *inform* utterances, which consists of only an entity and optionally an article, generating a set of 100 utterances was unfeasible: hence, a total of 10 utterances for the *inform* intent were produced. The *inform* intent is defined for conversations in which a worker's question cannot be adequately answered solely by performing slot filling on the initial utterance. This is often the case when some of the required entities for formulating an appropriate response are missing. For

[1] `https://chatgpt.com`

example, in the following conversation:

- Worker: *What's this object? I don't know how to use it.*
- Assistant: *Which object?*
- Worker: *The oscilloscope.*

The worker's first utterance falls under the *object_instructions* intent, whereas the worker's second utterance falls under the *inform* intent.

The used prompt for each intent, except for *inform* and *out_of_scope* intents, was the following: "*Imagine being an operator working inside an industrial laboratory. You can communicate with someone who knows the laboratory perfectly, including all the present objects and possible procedures that can be carried out. There are several intents you could have while operating within this industrial laboratory. This is one: <intent description>. Since you'll have to communicate with the other person through text messages, try to avoid all forms of greeting and politeness. For this intent, imagine 100 unique sentences you would say to your interlocutor to express your intent and achieve the desired result.*" Exceptions were made for the *inform* intent, for which we prompted the model to generate 10 unique sentences, and the *out_of_scope* intent, for which we used the following prompt: "*Imagine being an operator working inside an industrial laboratory. You can communicate with someone who knows the laboratory perfectly, including all the present objects and possible procedures that can be carried out. There are several intents you could have while operating within this industrial laboratory, which I will list below: <full list of intent descriptions>. Since you'll have to communi-*

cate through text messages, try to avoid all forms of greeting and politeness. Knowing these intents, generate 100 unique sentences that are out of scope."

As ChatGPT was not able to generate 100 unique utterances, we carried out additional duplicate filtering and re-prompted the model in order to generate more utterances, until we met the criteria of gathering 100 unique utterances for each intent. We hypothesize that the inability to generate a set of unique utterances is due to the many constraints expressed in our prompt, which, however, was designed to generate utterances that reflected the real ones collected in the same laboratory setting.

7.2 Results

Table 7.1 reports the results obtained for intent and entity classification, considering accuracy and F_1-score. Five different variants of the training set were explored:

- real data;
- real data + G10 (a set containing 10 out of the 100 generated utterances for each intent);
- real data + G50 (a set containing 50 out of the 100 generated utterances for each intent);
- real data + G100 (the full set of generated utterances, consisting of 100 for each intent);
- G100.

	Intent		Entity	
Training	**Accuracy**	**F_1-score**	**Accuracy**	**F_1-score**
real	**0.867**	**0.844**	0.994	0.981
real+G10	0.830	0.815	1.00	1.00
real+G50	0.792	0.773	1.00	1.00
real+G100	0.792	0.784	1.00	1.00
G100	0.584	0.564	**1.00**	**1.00**

Table 7.1: Results for intents and entities classification considering different sets of training data.

The best results for the intent classification have been obtained using only real data, obtaining an accuracy of 0.867 and a F_1-score of 0.844. The results suffer the addition of generated data, which introduces noise and makes performances worse, reaching an accuracy of 0.584 (-0.283) and a F_1-score of 0.564 (-0.280). These results suggest that, in this challenging industrial scenario, generative models, such as GPT [3] are not yet capable of generating appropriate data with regard to understand human's intent in this domain and the use of manually annotated data is still necessary. Instead, considering the ability to predict the entities inside utterances, which represent more simple concepts compared to intents, only generated data is enough. In particular, the model trained with the G100 set obtains better performance than one trained only with real data (1.00 vs 0.994 for accuracy and 1.00 vs 0.981 for F_1-score).

Figure 7.1 presents qualitative results for three distinct utterances for both intent and entity classification tasks. In the top row, we observe an utterance with an incorrect intent prediction (*object_instructions* instead of

object_warnings) with a confidence of 0.32. This utterance did not contain any entities, and the absence of entities was correctly predicted with a confidence of 0.98. These results suggest that our model exhibited uncertainty in determining the appropriate class for the utterance. This uncertainty could be attributed to the fact that both *object_instructions* and *object_warnings* often contain utterances formulated in a very similar manner. However, the model's capabilities concerning entity classification tend to be highly accurate with a significant confidence score. Moving to the middle row, we encounter an utterance with an incorrect intent prediction (*inform* instead of *where_object*) with a confidence score of 0.94. This utterance contained an entity of the *object* type, and the presence and class of this entity were correctly predicted with a confidence of 0.91. These results suggest that our model exhibited a high level of confidence in classifying the utterance, despite its incorrect classification. This observation may indicate that this particular utterance shares significant similarities with those typically found in the *where_object* intent. Similarly, in this instance, the model's capabilities enable accurate entity classification with a high confidence score. Lastly, the bottom row showcases an utterance with a correct intent prediction (*which_PPE_procedure*) with a confidence score of 0.91. This utterance featured entities of both *procedure* and *board* types, and the presence and classes of these entities were correctly predicted with confidence scores of 0.74 and 0.96, respectively. These results suggest that our model exhibited a high level of confidence in classifying the utterance, and its classification was indeed correct. Ultimately, in this latter scenario as well, the model's capabilities enable accurate entity classification, resulting in a confidence score

ranging from moderate to high.

	GROUND TRUTH INTENT	PREDICTED INTENT	CONF.	GROUND TRUTH ENTITIES	PREDICTED ENTITIES	CONF.
should I know more about this object?	object-warnings	object-instructions	0.32	none	none	0.98
point me to the *panel*	where-object	inform	0.94	*panel*	*panel*	0.91
which PPE are required for the *repair* of the *low voltage* board?	which-PPE-procedure	which-PPE-procedure	0.91	*repair*, *low voltage*	*repair*, *low voltage*	0.74, 0.96

Figure 7.1: Qualitative results showing two incorrect intent predictions (first two rows) and a correct prediction (last row), alongside correct entity predictions.

7.3 Summary

The contribution to this work represents a tentative of enriching our real dataset with generated data, with the objective of strengthening the overall performance. However, our analysis showed that the generated data was not capable of being on par with real data for intent classification, as the new data introduced noise and deteriorated the performances. However, entity classification was proved to be stronger with the aid of generated data. We conclude that this duality of results is due to the limitations of LLMs at the time of the experiment, since the prompting phase included a large set of constraints in order to generate utterances.

Chapter 8

Conclusions

This thesis has explored the design, development and deployment of multimodal conversational assistants tailored to industrial scenarios, tracing the evolution from early task-specific systems to advanced architectures integrating Large Language Models (LLMs). The work was grounded in the belief that effective digital assistance in complex operational contexts requires more than textual question answering: it must integrate multiple modalities, interpret context and adapt dynamically to user needs while ensuring safety and reliability. Each chapter addressed different stages into the developing and deploying, evaluating different architecture modifications and their impact on quantitative and qualitative results.

Chapter 4 reflected the first development of HERO, a rule-based assistant capable of combining natural language understanding with egocentric vision for object recognition and procedural support. This initial system demonstrated the feasibility of grounding conversational interactions in visual context, enabling operators to refer to objects implicitly and still receive

accurate, task-relevant guidance. The ENIGMA Laboratory was instrumental in this phase, serving both as a realistic industrial mock-up for prototyping and as the source of a richly annotated multimodal dataset.

Chapter 5 followed the deployment of HERO to mobile platforms, improving accessibility and enabling deployment without specialized wearable devices. A new question answering module, selectively powered by LLMs, was introduced to handle non-critical queries, balancing the generative capabilities of modern models with the safety demands of industrial workflows. A controlled user study confirmed the advantages of the multimodal, mobile-based assistant over traditional paper-based manuals, with participants reporting greater convenience, faster task completion and higher satisfaction.

Chapter 6 presented HERO-GPT, marked a methodological shift away from predefined intents and entities toward zero-shot reasoning and modular LLM integration. This architecture embraced retrieval-augmented generation (RAG) and dynamic routing between specialized modules, allowing the assistant to handle a wider range of user inputs without extensive manual training. The modular design also supports future scalability, enabling the integration of additional sensing modalities or domain-specific knowledge bases, allowing a smoother shift of context.

8.1 Limitations and Future Work

The findings confirm that multimodal conversational assistants can significantly improve task execution, reduce ambiguity in communication and enhance safety in industrial settings. However, some limitations emerge that

are worth considering for future development.

Although the ENIGMA laboratory offers a realistic and controlled industrial mock-up, the experiments have been conducted within a narrower domain compared to real-world industrial settings. For future work, the current evaluation should be expanded toward larger and more realistic industrial environments. Furthermore, the reliance on highly curated datasets both for visual and linguistic components introduce scalability challenges. With the introduction of new appliances and laboratories, the use of open-vocabulary object detectors would strengthen the overall system, allowing generalization across different environments without requiring extra redesign time. Likewise, refining the current "object closest to center" visual-grounding heuristic is necessary to ensure robustness in cluttered or complex scenarios.

Deploying and evaluating the system in actual industrial plants would test its robustness under more challenging conditions. Additionally, the integration of wearable devices could provide hands-free interaction methods, improving ergonomics during manual operations. In addition, refining the integration of LLMs through retrieval-augmented generation and robust consistency checks could enable their safer use even in critical procedures. Incorporating other modalities, for example by considering audio cues, gesture recognition, or IoT sensor data could provide deeper situational awareness, allowing the system to proactively intervene to prevent the user from carrying out an incorrect step or action. Furthermore, prolonged usability studies in such settings would offer valuable insights into the long-term effects on productivity, safety and user satisfaction, helping to guide further refinements of the system. Furthermore, a deeper and more systematic safety analysis

is required, particularly concerning the risks of hallucinations in industrial workflows.

Ultimately, this work demonstrates that the synergy of language, vision and structured procedural knowledge represents a promising path toward conversational assistants that truly replicate the contextual awareness and flexibility of human experts.

Acknowledgements

I would like to acknowledge my advisor, Prof. Giovanni Maria Farinella, for the guidance provided during these years, and my co-tutor, Prof. Antonino Furnari, for his scientific advice and encouragement. I am also grateful to Dott. Francesco Ragusa for his support during the development of this work. I also wish to express my appreciation to the entire research group, and in particular all its members, for the support they have offered me throughout this period.

Bibliography

[1] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

[3] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[4] Claudia Bonanno, Francesco Ragusa, Rosario Leonardi, Antonino Furnari, and Giovanni Maria Farinella. Hero: An artificial conversational assistant to support humans in industrial scenarios. In *Inter-*

national Conference on Signal Processing and Multimedia Applications (SIGMAP), 2022.

[5] Claudia Bonanno, Francesco Ragusa, Antonino Furnari, and Giovanni Maria Farinella. Hero: A multi-modal approach on mobile devices for visual-aware conversational assistance in industrial domains. In *International Conference on Image Analysis and Processing (ICIAP)*, 2023.

[6] Luca Strano, Claudia Bonanno, Francesco Ragusa, Giovanni Maria Farinella, and Antonino Furnari. Gpt-assist: Zero-shot conversational assistance in industrial domains exploiting large language models. In *International Conference on Image Processing and Video Engineering (IMPROVE)*, 2024.

[7] Francesco Ragusa, Rosario Leonardi, Michele Mazzamuto, Claudia Bonanno, Rosario Scavo, Antonino Furnari, and Giovanni Maria Farinella. Enigma-51: Towards a fine-grained understanding of human-object interactions in industrial scenarios. In *IEEE Winter Conference on Application of Computer Vision (WACV)*, 2024.

[8] Elisa Ramil Brick, Vanesa Caballero Alonso, Conor O'Brien, Sheron Tong, Emilie Tavernier, Amit Parekh, Angus Addlesee, and Oliver Lemon. Am i allergic to this? assisting sight impaired people in the kitchen. In *Proceedings of the 2021 International Conference on Multimodal Interaction*, ICMI '21, page 92–102, New York, NY, USA, 2021. Association for Computing Machinery.

[9] Despina Tomkou, George Fatouros, Andreas Andreou, Georgios Makridis, Fotis Liarokapis, Dimitrios Dardanis, Athanasios Kiourtis, John Soldatos, and Dimosthenis Kyriazis. Bridging industrial expertise and xr with llm-powered conversational agents. *arXiv preprint arXiv:2504.05527*, 2025.

[10] Satwik Kottur, Seungwhan Moon, Alborz Geramifard, and Babak Damavandi. Simmc 2.0: A task-oriented dialog dataset for immersive multimodal conversations. *arXiv preprint arXiv:2104.08667*, 2021.

[11] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[12] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*, 2016.

[13] Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics.

[14] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention

is all you need. *Advances in neural information processing systems*, 30, 2017.

[15] Xiaokang Liu, Jianquan Li, Jingjing Mu, Min Yang, Ruifeng Xu, and Benyou Wang. Effective open intent classification with k-center contrastive learning and adjustable decision boundary. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 13291–13299, 2023.

[16] Mujeen Sung, James Gung, Elman Mansimov, Nikolaos Pappas, Raphael Shu, Salvatore Romeo, Yi Zhang, and Vittorio Castelli. Pre-training intent-aware encoders for zero-and few-shot intent classification. *arXiv preprint arXiv:2305.14827*, 2023.

[17] Xiaoshuai Song, Keqing He, Pei Wang, Guanting Dong, Yutao Mou, Jingang Wang, Yunsen Xian, Xunliang Cai, and Weiran Xu. Large language models meet open-world intent discovery and recognition: An evaluation of chatgpt. *arXiv preprint arXiv:2310.10176*, 2023.

[18] Momchil Hardalov, Ivan Koychev, and Preslav Nakov. Enriched pre-trained transformers for joint slot filling and intent detection. *arXiv preprint arXiv:2004.14848*, 2020.

[19] Mehrdad Rafiepour and Javad Salimi Sartakhti. Ctran: Cnn-transformer-based network for natural language understanding. *Engineering Applications of Artificial Intelligence*, 126:107013, 2023.

[20] Thinh Pham, Chi Tran, and Dat Quoc Nguyen. Misca: A joint model for multiple intent detection and slot filling with intent-slot co-attention. *arXiv preprint arXiv:2312.05741*, 2023.

[21] Kalpa Gunaratna, Vijay Srinivasan, Akhila Yerukola, and Hongxia Jin. Explainable slot type attentions to improve joint intent detection and slot filling. *arXiv preprint arXiv:2210.10227*, 2022.

[22] Bing Liu and Ian Lane. Attention-based recurrent neural network models for joint intent detection and slot filling. *arXiv preprint arXiv:1609.01454*, 2016.

[23] Qian Chen, Zhu Zhuo, and Wen Wang. Bert for joint intent classification and slot filling. *arXiv preprint arXiv:1902.10909*, 2019.

[24] Bernd Bohnet, Chris Alberti, and Michael Collins. Coreference resolution through a seq2seq transition-based system. *Transactions of the Association for Computational Linguistics*, 11:212–226, 2023.

[25] Danqi Chen and Christopher Manning. A fast and accurate dependency parser using neural networks. In Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 740–750, Doha, Qatar, October 2014. Association for Computational Linguistics.

[26] Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. End-to-end neural coreference resolution. *arXiv preprint arXiv:1707.07045*, 2017.

[27] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

[28] Meta AI. The llama 4 herd: The beginning of a new era of natively multimodal ai innovation, 2025. Accessed: 2025-08-09.

[29] OpenAI. Introducting gpt-5, 2025. Accessed: 2025-08-09.

[30] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[31] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021.

[32] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019.

[33] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language un-

derstanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022.

[34] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

[35] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

[36] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Q. Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh K. Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Z. Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Christian Fuegen, Abrham Gebreselasie, Cristina González, James M. Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jáchym Kolár, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz Puentes, Merey Ramazanova, Leda Sari, Kiran K. Soma-

sundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Yunyi Zhu, Pablo Arbeláez, David J. Crandall, Dima Damen, Giovanni Maria Farinella, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard A. Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[37] Dandan Shan, Jiaqi Geng, Michelle Shu, and David F. Fouhey. Understanding human hands in contact at internet scale. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9869–9878, 2020.

[38] Siddhant Bansal, Chetan Arora, and C.V. Jawahar. My view is the best view: Procedure learning from egocentric videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022.

[39] Yunze Liu, Yun Liu, Che Jiang, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, Zhoujie Fu, He Wang, and Li Yi. Hoi4d: A 4d egocentric dataset for category-level human-object interaction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21013–21022, June 2022.

[40] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The epic-kitchens dataset. *ArXiv*, abs/1804.02748, 2018.

[41] Yao Lu and Walterio W Mayol-Cuevas. Egocentric hand-object interaction detection and application, 2021.

[42] Dimitri Zhukov, Jean-Baptiste Alayrac, Ramazan Gokberk Cinbis, David Fouhey, Ivan Laptev, and Josef Sivic. Cross-task weakly supervised learning from instructional videos, 2019.

[43] Hamed Pirsiavash and Deva Ramanan. Detecting activities of daily living in first-person camera views. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2847–2854, 2012.

[44] Yong Jae Lee, Joydeep Ghosh, and Kristen Grauman. Discovering important people and objects for egocentric video summarization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1346–1353, 2012.

[45] Fadime Sener, Dibyadip Chatterjee, Daniel Shelepov, Kun He, Dipika Singhania, Robert Wang, and Angela Yao. Assembly101: A large-scale multi-view video dataset for understanding procedural activities. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21064–21074, 2022.

[46] Francesco Ragusa, Antonino Furnari, and Giovanni Maria Farinella. Meccano: A multimodal egocentric dataset for humans behavior understanding in the industrial-like domain. *Computer Vision and Image Understanding (CVIU)*, 2023.

[47] Microsoft hololens 2. `https://www.microsoft.com/en-us/hololens`.

[48] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

[49] Tom Bocklisch, Joey Faulkner, Nick Pawlowski, and Alan Nichol. Rasa: Open source language understanding and dialogue management. *arXiv preprint arXiv:1712.05181*, 2017.

[50] Tanja Bunk, Daksh Varshneya, Vladimir Vlasov, and Alan Nichol. Diet: Lightweight language understanding for dialogue systems. *arXiv preprint arXiv:2004.09936*, 2020.

[51] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. spacy: Industrial-strength natural language processing in python. `https://github.com/explosion/spaCy`, 2020. Accesso: 12 agosto 2025.