# ENIGMA-51: Towards a Fine-Grained Understanding of Human Behavior in Industrial Scenarios

Francesco Ragusa [1,2], Rosario Leonardi[1], Michele Mazzamuto[1,2],
Claudia Bonanno[1,2], Rosario Scavo[1], Antonino Furnari[1,2], Giovanni Maria Farinella[1,2]

[1]FPV@IPLab - University of Catania, Italy
[2]Next Vision s.r.l. - Spinoff of the University of Catania, Italy

## Abstract

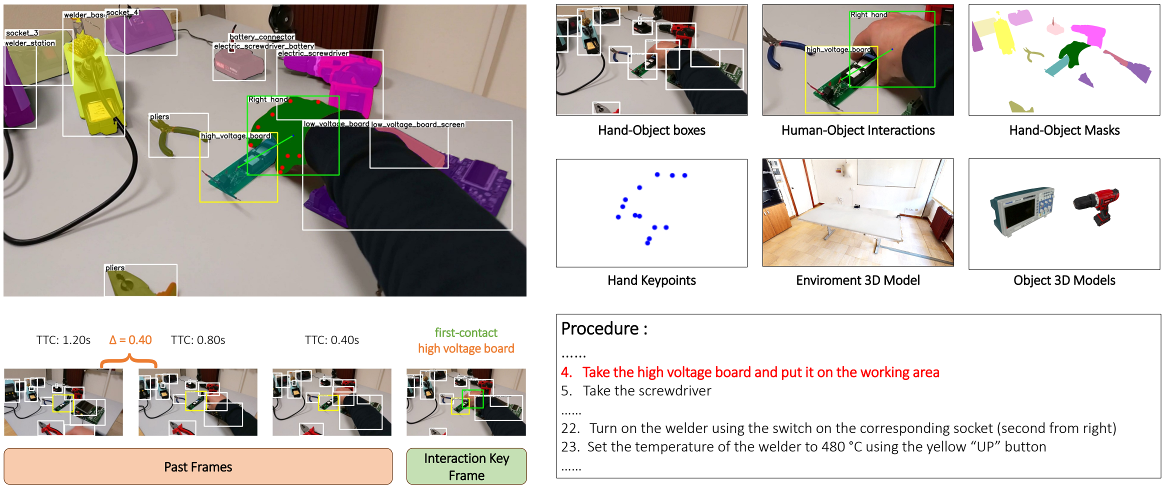
*ENIGMA-51 is a new egocentric dataset acquired in an industrial scenario by 19 subjects who followed instructions to complete the repair of electrical boards using industrial tools (e.g., electric screwdriver) and equipments (e.g., oscilloscope). The 51 egocentric video sequences are densely annotated with a rich set of labels that enable the systematic study of human behavior in the industrial domain. We provide benchmarks on four tasks related to human behavior: 1) untrimmed temporal detection of human-object interactions, 2) egocentric human-object interaction detection, 3) short-term object interaction anticipation and 4) natural language understanding of intents and entities. Baseline results show that the ENIGMA-51 dataset poses a challenging benchmark to study human behavior in industrial scenarios. We publicly release the dataset at* https://iplab.dmi.unict.it/ENIGMA-51.

## 1. Introduction

Every day, humans interact with the surrounding world to achieve their goals. These interactions are often complex and require multiple steps, skills, and involve different objects. For example, in an industrial workplace, when performing maintenance of industrial machinery, a worker interacts with several objects and tools while repairing the machine (e.g., *wear PPEs, take the screwdriver*), testing it (e.g., *press the button on the electric panel*), and writing a report (e.g., *take the pen, write the report*). To properly assist humans, an intelligent system should be able to model human-object interactions (HOIs) from real-world observations captured by users wearing smart cameras (e.g., smart glasses) [9, 13, 33]. It is also plausible that predicting human-object interactions in advance can benefit an intelligent system help workers to avoid mistakes, or to improve their safety. For example, during the execution of a maintenance procedure, an AI assistant should be able to understand when the worker is interacting with the objects, show technical information, provide instructions on how to interact with each object, alert the worker of potential safety risks (e.g., *Before touching the electrical board, turn off the power supply!*), and suggest what the next interaction is. Furthermore, an intelligent system should be able to have a natural language conversation with workers. It should also be able to extract useful information from their speech, and figure out what they are trying to achieve. This way, it can provide assistance for supporting their needs, preferences, and goals.

In general, tasks focused on understanding human behaviour have been extensively studied thanks to the availability of public datasets that consider multiple domains [2, 22, 30, 49] or specific ones, such as kitchens [11, 31, 63], daily life [27, 38], and industrial-like scenarios [41, 47]. However, since data acquisition in a real industrial scenario is challenging due to privacy issues, safety and industrial secret protection, the datasets available to date do not reflect real industrial environments, considering proxy activities such as employing toy models made of textureless parts [41, 47].

Considering what stated above, to enable research in this field, we present ENIGMA-51, a new dataset composed of 51 egocentric videos acquired in an industrial environment which simulates a real industrial laboratory. The dataset was acquired by 19 subjects who wore a Microsoft HoloLens 2 [34] headset and followed audio and AR instructions provided by the device to complete repairing procedures on electrical boards. The subjects interact with industrial tools such as an electric screwdriver and pliers, as well as with electronic instruments such as a power supply and an oscilloscope while executing the steps to complete a specific procedure. Apart the current interactions, we an-



Procedure :

......

4. Take the high voltage board and put it on the working area
5. Take the screwdriver

......

22. Turn on the welder using the switch on the corresponding socket (second from right)
23. Set the temperature of the welder to 480 °C using the yellow "UP" button

......

Figure 1. Frames have been annotated with a rich set of labels (top-left). Sequences have been annotated by determining the interaction key frame (bottom-center), assigning the verb (green) and the active object (orange). For each interaction key frame, we provide objects and hand bounding boxes and the relation between them. In the past frames, we annotated also the next active objects and we derived the time to contact (TTC) (bottom-left). We also generated pseudo-labels for semantic masks and hand keypoints, and we released 3D models for the objects and for the laboratory (top-right). Moreover, a specific instruction belonging to the procedure is associated with each interaction key frame (bottom-right).

notated which objects and hands will be involved in future interactions, as well as the time to contact (TTC) to indicate when the future interaction will start. This allows us to explore the task of predicting the future human-object interactions considering the industrial domain. Textual instructions used for the data acquisition, also allow to study tasks which focus on the knowledge extraction of intents and entities from the text while users are interacting with the objects. In the industrial domain these tasks have not been explored due to the lack of public egocentric datasets explicitly annotated with intents and entities.

Together with the manually annotations, we release the pseudo-labels and the pre-extracted features to enable further investigations beyond the current study. In particular, we generated hands and objects segmentation masks [26], and hands keypoints [10]. The provided visual features are extracted with DINOv2 [36] and CLIP [39]. To allow further research in the context of scalable models trained using synthetic data, we share the 3D models of the laboratory and all considered industrial objects. Figure 1 shows examples of images acquired in the industrial environment where the dataset was acquired together with the annotations. To highlight the usefulness of the proposed dataset, we performed baseline experiments related to 4 fundamental tasks focused on understanding human behavior from first person vision in the considered industrial context: 1) Untrimmed Temporal Detection of Human-Object Interactions, 2) Egocentric Human-Object Interaction (EHOI) Detection, 3) Short-Term Object Interaction Anticipation and 4) Natural Language Understanding of Intents and Entities.

In sum the contributions of this work are as follows: 1) we introduce ENIGMA-51, a new dataset composed of 51 egocentric videos acquired in an industrial domain; 2) we manually annotated the dataset with a rich set of annotations aimed at studying human behavior; 3) we propose a benchmark to study human behavior in an industrial environment exploring 4 different tasks, showing that the current state-of-the-art approaches are not sufficient to solve the considered problems in the industrial setting; 4) we provide additional labels and features exploiting foundational models, with the aim to push research on additional tasks on the proposed industrial dataset. The ENIGMA-51 dataset and its annotations are available at the following link: https://iplab.dmi.unict.it/ENIGMA-51.

## 2. Related Work

Our work is related to previous research lines which are revised in the following sections.

### 2.1. Ego Datasets for Human Behavior Studies

Previous works have proposed egocentric datasets focusing on human behavior understanding. The Activity of Daily Living (ADL) [38] dataset is one of the first datasets acquired from the egocentric perspective. It includes 20 egocentric videos where participants were involved in daily activities. It comprises temporal action annotations aimed to study egocentric activities. EGTEA Gaze+ [29] focuses on cooking activities involving 32 subjects who recorded

| Dataset | Year | Video? | EHOI Annotations? | Settings | Hours | Sequences | Subjects |
|---|---|---|---|---|---|---|---|
| ENIGMA-51 (ours) | 2024 | ✓ | ✓ | Industrial | 22 | 51 | 19 |
| MECCANO [41] | 2023 | ✓ | ✓ | Industrial-like | 7 | 20 | 20 |
| Ego4D [22] | 2022 | ✓ | ✓ | Multi-domain | 3670 | 9650 | 923 |
| THU-READ [53] | 2019 | ✓ | ✓ | Daily activities | 224 | 1920 | 8 |
| EPIC-KITCHENS-VISOR [14] | 2022 | ✓ | ✓ | Kitchen activities | 100 | 700 | 45 |
| HOI4D [30] | 2022 | ✓ | ✓ | Objects manipulation | 22 | 4000 | N/A |
| VOST [54] | 2023 | ✓ | ✓ | Daily + Industrial-like | 4 | 713 | N/A |
| ARCTIC [17] | 2023 | ✓ | ✓ | Object manipulation | 2 | 339 | 10 |
| 100 Days of Hands [49] | 2020 | X | ✓ | Daily activities | 3144 | 27000 | 1350+ |
| GUN-71 [44] | 2015 | X | ✓ | Daily activities | N/A | N/A | 8 |
| Assembly101 [47] | 2022 | ✓ | X | Industrial-like | 513 | 362 | 53 |
| EGTEA Gaze+ [29] | 2017 | ✓ | X | Cooking activities | 28 | 86 | 32 |
| ADL [38] | 2012 | ✓ | X | Daily activities | 10 | 20 | 20 |

Table 1. Overview of egocentric datasets with a particular focus on those that allow the study of human-object interactions sorted by the number of hours.

28 hours of videos. It has been annotated with pixel-level hand masks and 10325 action annotations including 19 action verbs and 51 object nouns. The THU-READ [53] dataset is composed of 1920 RGB-D sequences captured by 8 participants who performed 40 different daily-life actions. The EPIC-Kitchens datasets [11, 12] are collections of egocentric videos that capture natural actions in kitchen settings. EPIC-Kitchens-55 [11] consists of 432 videos with annotations for 352 objects and 125 verbs. EPIC-Kitchens-100 [12] is a larger version of EPIC-Kitchens-55 with more videos (700), scenes (45) and hours (100). Assembly101 [47] simulates an industrial scenario and it is composed of 4321 assembly and disassembly videos of toy vehicles made of textureless parts. It offers a multi-view perspective, comprising static and egocentric recordings annotated with 100K coarse and 1M fine-grained action segments and with 18M 3D hand poses.

While these datasets explore actions and activities, other datasets have been proposed to study human-object interactions from the egocentric perspective in a more fine-grained fashion. The Grasp Understanding (GUN-71 [44]) dataset, contains 12,000 images of hands manipulating 28 objects labelled with 71 grasping categories. The 100 Days Of Hands (100DOH) [48] dataset captures hands and objects involved in generic interactions. It consists of 100K frames collected over 131 days with 11 types of interactions. It comprises bounding boxes around the hands and the active objects, the side of the hands and the contact state (which indicates if the hand is touching an object or not). Other works focused on human-object interactions providing egocentric video datasets. EPIC-KITCHENS VISOR [14] contains videos from EPIC-KITCHENS-100 [12] annotated with 272K semantic masks for 257 classes of objects, 9.9M interpolated dense masks, and 67K human-object interactions. The authors of [30] proposed the HOI4D dataset which is composed of 2.4 million RGB-D egocentric frames across 4000 sequences acquired in 610 indoor rooms. The authors of [17] studied hands interacting with articulated objects (e.g., scissors, laptops) releasing the ARCTIC dataset. It comprises 2.1M high-resolution images annotated with 3D hand and object meshes and with contact information. The VOST dataset [54] focuses on objects that dramatically change their appearance. It includes 713 sequences where the objects have been annotated with segmentation masks. Ego4D [22] is a massive-scale dataset composed of 3670 hours of daily-life activity videos acquired in different domains by 923 unique participants. It comes with a rich set of annotations to address tasks concerning the understanding of the past, present, and future.

More related to our work are datasets acquired in the industrial-like domain [41, 47]. Unlike Assembly101 [47] and MECCANO [41] we consider an industrial setting which simulates a real industrial laboratory. Unlike Assembly101, we provide fine-grained annotations to study different aspects of human behavior.

Table 1 shows the key attributes of the analyzed datasets. Previous datasets have focused on kitchens, daily activities, and industrial-like scenarios exploring different aspects of the human behavior. In order to perform a systematic study on human behaviour and human-object interactions in an industrial domain, we present the ENIGMA-51 dataset with a rich set of fine-grained egocentric videos together with annotations.

## 2.2. Untrimmed Temporal Detection of Human-Object Interactions

The proposed untrimmed temporal detection of human-object interactions task is related to previous research on untrimmed action detection. Existing approaches focus on one-stage methods, performing both temporal action detection and classification within a single network, aiming to identify actions without using action proposals. Recent works achieved state-of-the-art results using Vision Transformers. The authors of [60] proposed ActionFormer, a transformer network designed for temporal action localization in videos. This method estimates action boundaries through a combination of multiscale feature representation and local self-attention, which effectively models temporal dependencies. TriDet [50] uses a Trident-head to model the action boundary by estimating the relative probability distribution around the boundary. Features are extracted through a feature pyramid and aggregated with the proposed scalable granularity perception layer.

Other methods focused on masked video modeling for pretraining one-stage methods. In particular, InternVideo [58] uses a combination of generative and discriminative self-supervised learning techniques by implementing masked video modeling and video-language contrastive learning. Recently, the authors of [57] proposed VideoMAE V2, which scaled VideoMAE [55] for building video foundation models through a dual masking strategy.

In this work, we assess the performance of state-of-the-art temporal action detection methods on the proposed ENIGMA-51 dataset considering ActionFormer [60].

## 2.3. Egocentric HOIs Detection

Several works have explored different aspects of human-object interactions (HOIs) from the egocentric perspective. The authors of [49] proposed a method based on the Faster-RCNN [43] object detector to detect the hands and the objects present in the image, categorizing objects as either *active* or *passive*, determining the side of the hands (*left* or *right*), and predicting the contact state between the hand and the associated active object. The authors of [41, 42] investigated human-object interactions predicting bounding boxes around the active objects and the verb which describes the interaction exploiting multimodal signals with different instances of SlowFast networks [18]. The authors of [3] presented an architecture for detecting human-object interactions using two YOLOv4 object detectors [4] and an attention-based technique. The authors of [22] explored object transformations introducing the novel task of object state change detection and classification. While most of the analysis of human-object interactions relies on bounding box annotations, some works exploited hand poses and semantic segmentation masks [14, 31], contact boundaries [62], which represents the spatial area where the interaction occurs, or additional modalities, such as depth maps and instance segmentation masks, to learn more robust representations [28].

In this work, we evaluate the HOIs detection method proposed in [28] exploiting the fine-grained human-object interaction annotations of the ENIGMA-51.

## 2.4. Short-Term Object Interaction Anticipation

Past works addressed different variants of the short-term object interaction anticipation task. The authors of [19] focused their study on the prediction of the next-active objects by analyzing their trajectories over time. The authors of [25] proposed a model that exploits a predicted visual attention probability map and the hands' positions to predict next-active objects. The authors of [15] predicted future actions exploiting hand-objects contact representations. In particular, the proposed approach predicts future contact maps and segmentation masks, which are exploited by the Egocentric Object Manipulation Graphs framework [16] for predicting future actions.

The short-term object interaction anticipation task has been more formally defined in [22]. To tackle the task, the authors of [22] released a two-branch baseline composed of an object detector [43] to detect next-active objects and a SlowFast [18] 3D network to predict the verb and the time to contact. The proposed baseline was extended by the authors of [7] who replaced Faster-RCNN with DINO [61], and SlowFast with a VideoMAE-pretrained transformer network [55]. Recently, StillFast an end-to-end approach has been proposed by [40]. The method simultaneously processes a high-resolution still image and a video with a low spatial resolution, but a high temporal resolution. Recent state-of-the-art performances have been achieved by [37] exploiting language. They proposed a multimodal transformer-based architecture able to summarise the action context leveraging pre-trained image captioning and vision-language models.

Due to its end-to-end training ability, in this work, we used StillFast [40] as a baseline for the short-term object interaction anticipation benchmark on ENIGMA-51.

## 2.5. Natural Language Understanding of Intents and Entities

Understanding intents and entities from text to extract knowledge about human-object interactions in the industrial domain is a task that has not been explored due to the lack of public egocentric datasets suitable for this task.

The authors of [56] addressed both intent classification and slot filling as a seq2seq problem, using an architecture that takes text input, generates ELMo embeddings [45], and incorporates one BiLSTM [46] and self-attention layers for each task, outputting task-specific BIO (Beginning Inside and Outside) labels. In [8] the BERT architecture has been explored to tackle the limited generalization capabilities of natural language understanding and propose a joint intent and classification architecture. The authors of [6] incorporate pre-trained word embeddings from language models and combine them with sparse word and character level n-grams features alongside a Transformer architecture.

While some works use speech-to-text models to convert speech input into text, others handle speech directly (Spoken Language Understanding). Earlier approaches proposed RNN-based or LSTM-based contextual SLU [24, 51] which take into account previously predicted intents and slots. The authors of [23] proposed a BiLSTM-based architecture to manage the interrelated connections between intent and slots. In [59] has been introduced the Token-

and-Duration Transducer (TDT) architecture for Automatic Speech Recognition (ASR), able to jointly predict both a token and its duration, enabling the skipping of input frames during inference based on the predicted duration output, resulting in significantly improved efficiency.

Since the ENIGMA-51 dataset comprises textual instructions about the activities performed by subjects, we exploited this textual information to explore the task of predicting intents and entities to extract knowledge about human-object interactions in the industrial domain.

# 3. The ENIGMA-51 Dataset

In our ENIGMA laboratory there are 25 different objects that can be grouped into fixed objects (such as an *electric panel*) and movable objects (such as a *screwdriver*). Differently than other egocentric datasets [41, 47] that contain industrial-like objects without textures, ENIGMA-51 includes real industrial objects as shown in Figure 1. The complete list of the objects present in the ENIGMA laboratory is reported in the supplementary material.

## 3.1. Data Acquisition

To collect data suitable to study human behavior in industrial domain, we designed two procedures consisting of instructions that involve humans interacting with the objects present in the laboratory to achieve the goal of repairing two electrical boards (see Figure 1 for visual examples). In particular, we designed two repairing procedures, one for each electrical board (*high and low voltage*), with the help of industrial experts. For each procedure, we considered 4 different versions varying the use of a *screwdriver* or *electric screwdriver* and the electrical component to solder (*resistor, capacitor or transformer*). Each procedure is composed of more than 100 steps, referencing objects and actions that were expected to be carried out in the industrial laboratory such as *Turn on the welder using the switch on the corresponding socket (second from right)* and *Set the temperature of the welder to 480 °C using the yellow "UP" button*. Based on these instructions, we developed a custom Microsoft HoloLens 2 [34] application which provided the instructions through audio, images and AR during the acquisition phase[1]. Considering that we designed two different repair procedures, each subject acquired at least one repairing video for each electric board obtaining a total of 51 videos. The 19 participants had different levels of experience in repairing electrical boards and using industrial tools. An example of the captured data is reported in Figure 1. For each participant, we acquired the RGB stream from the Microsoft HoloLens 2 with a resolution of 2272x1278 pixels with a framerate of *30 fps*. The average duration of the captured videos is 26.32min for a total of 22 hours of videos.

| Splits | Train | Val | Test | Total |
|---|---|---|---|---|
| # Videos | 27 | 8 | 16 | 51 |
| # Videos Length | ≃11h | ≃4h | ≃7h | ≃22h |
| # Images | 25,311 | 8,528 | 11,666 | 45,505 |
| # Objects | 152,865 | 53,486 | 68,784 | 275,135 |
| # Active Objects | 4,709 | 1,700 | 2,933 | 9,342 |
| # Hands | 31,249 | 11,322 | 13,902 | 56,473 |
| # Hands in contact | 5,039 | 1,833 | 3,171 | 10,043 |
| # Interactions frames | 6,386 | 2,150 | 4,061 | 12,597 |
| # Interactions | 7,133 | 2,406 | 4,497 | 14,036 |
| # Past frames | 19,090 | 6,437 | 7,683 | 33,210 |
| # Next Object Interactions | 21,535 | 7,280 | 8,499 | 37,314 |

Table 2. Statistics of the ENIGMA-51 dataset considering the Training, Validation and Test splits.

We also synchronized the audio instructions with the captured video by assigning a timestamp when the user moved to the next instruction.

## 3.2. Data Annotation

We labelled the ENIGMA-51 dataset with a rich set of fine-grained annotations that can be used and combined to study different aspects of human behavior. Table 2 summarizes statistics about the collected dataset.

**Temporal and Verb Annotations:** We identified all *interaction key frames* in the 51 videos. For each identified interaction key frame, we assigned a timestamp and a verb describing the interaction. Our verb taxonomy is composed of 4 verbs: *first-contact, de-contact, take, and release*. The 4 considered verbs represent the basic actions that a user performs to interact with objects. Note that the difference between *first-contact* and *take* is that *first-contact* happens when the hand touches an object without taking it (e.g., pressing a button), while *de-contact* is the first frame in which the hand-object contact breaks (e.g., end of pressing a button) and *release* when the object is no longer held in the hand (e.g., put the screwdriver on the table). With this procedure, we annotated 14,036 interactions. Figure 1 reports an example of an interaction key frame with all the provided annotations, while Figure 2-left shows the verbs distribution in the 51 videos.

**Object Annotations:** We considered 25 object classes which include both fixed (e.g., electric panel, oscilloscope) and movable objects (e.g., screwdriver, pliers) to assign a class to the objects present in the interaction key frames and in the past frames[2]. Each object annotation consists in a tuple $(class, x, y, w, h, state)$, where $class$ represents the class of the object, $(x, y, w, h)$ are the 2D coordinates which define the bounding box around the object in the frame, and the $state$ indicates if the object is involved in an interaction or not (*active object vs. passive object*). With this annotation procedure, we annotated 275,135 objects. Figure 2-right reports the distributions of the objects over the 51 videos of the ENIGMA-51 dataset.

**Hands Annotations:** We annotated hands bounding boxes

[1] Additional information about the repairing procedures are available in the supplementary material.

[2] Additional details about our object taxonomy are available in the supplementary material

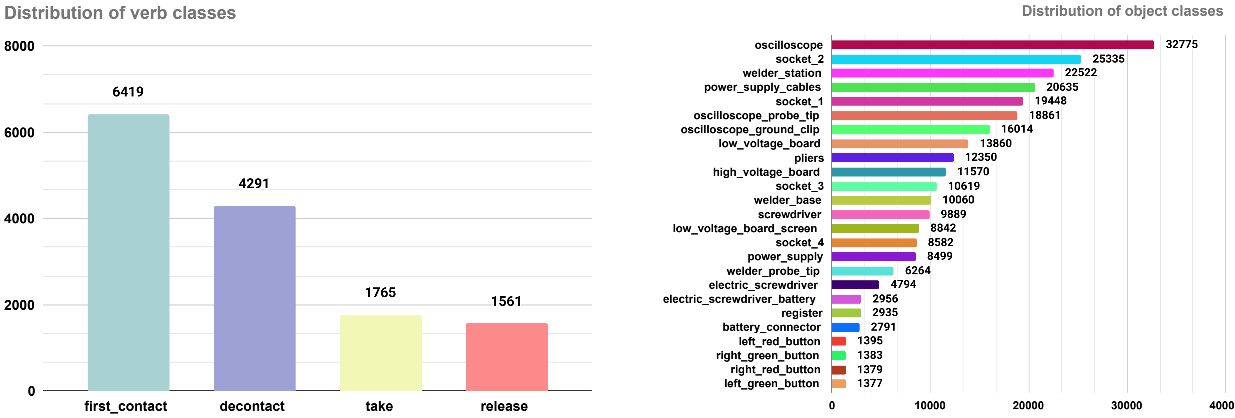


Figure 2. Distribution of verb (left) and object (right) classes over the 51 videos composing the ENIGMA-51 dataset.

in the interaction key frames and in past frames. To speed up this annotation process, we generated pseudo-labels by processing the interaction key frames with a hand-object detector [49], considering only the information related to the hands. Then, the annotators manually refined the bounding boxes, correcting the side of the hand and associating the hand with the previously labelled active object. Following this procedure, we labelled a total of 56,473 hands.

**EHOI Annotations:** For each of the interaction key frames, we considered: 1) hands and active object bounding boxes, 2) hand side (left and right), 3) hand contact state (contact and no contact), 4) hand-object relationships, and 5) object categories. For each hand, we assigned the *hand contact state* to *contact* if the hand was involved in an interaction of the type *first-contact* or *take*, and *no-contact* for the *release* and *de-contact* categories. Additionally, to make the annotations consistent and uniform, we assigned the *hand contact state* to *contact* even for the hands that were already in physical contact with objects. Following this procedure, we annotated 12,597 interaction frames, 17,363 hands of which 10,043 were in contact, and 9.342 active objects.

**Next Object Interaction Annotations:** Starting from the interaction key frame, we sampled frames every 0.4 seconds going backward up to 1.2 seconds before the beginning of the interaction timestamp. With this sampling strategy, we obtained 33210 past frames. We annotated the past frames with next object interaction annotations which consists of a tuple $(class, x, y, w, h, state, ttc)$ where $class$ represents the class of the object, $(x, y, w, h)$ are the 2D bounding box coordinates, $state$ indicates if the objects will be involved in an interaction and $ttc$ is a real number which indicates the time in seconds between the current timestamp and the beginning of the interaction. Figure 1 - bottom-left shows an example of labelled past frames.

**Utterances:** Based on the instructions used for the acquisition of the dataset, we collected 265 textual utterances, which represent the types of questions that a worker might pose to a supervisor colleague while following a procedure within an industrial setting such as *"How can I use the oscilloscope?"* or *"Which is the next step that I do?"*. We manually annotated user goals as "intents" (e.g. "object-instructions") and key information as "entities" (e.g. "object") considering 24 intent classes and 4 entity types[3]. To enrich this set of utterances, we generated similar synthetic data by interacting with ChatGPT [35]. This study resulted in the creation of 100 unique utterances for each intent[3]. The generated data was divided into three sets, G10, G50, and G100 which contain respectively 10, 50, and 100 generated unique utterances for each intent. Note that, all the utterances in G10 are also in G50 and G100, and all the utterances in G50 are also in G100.

**Additional Resources:** In order to enrich the ENIGMA-51 dataset, we release a set of resources useful to improve the impact of the dataset. We provide segmentation masks for the hands and the objects using SAM-HQ [26] and the 2D pose for the hands with MMPOSE [10]. We also extracted visual representations through DINOv2 [36] and CLIP [39]. The 3D models of ENIGMA Laboratory and of all industrial objects within it have been acquired using the Matterport [32] and ARTEC EVA [1] scanners, to enable the use of synthetic data to train scalable methods[3].

## 4. Benchmark and baselines results

### 4.1. Untrimmed Temporal Detection of Human-Object Interactions

**Task:** We consider the problem of detecting 4 basic human-object interactions (*"take"*, *"release"*, *"first-contact"*, and *"de-contact"*) from the untrimmed egocen-

[3] See the supplementary material for more details.

| Setting | p-mAP (%) temporal offset threshold (s) | | | | | | | | | | mp-mAP (%) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | |
| "take vs. release" | 27.40 | 32.97 | 36.88 | 40.08 | 42.15 | 43.70 | 45.52 | 47.48 | 48.81 | 49.50 | 41.45 |
| "first contact vs. de-contact" | 56.97 | 59.93 | 62.43 | 64.22 | 66.09 | 67.78 | 69.35 | 70.93 | 72.40 | 74.02 | 66.41 |
| "all interactions" | 29.64 | 31.69 | 33.28 | 34.60 | 35.91 | 36.96 | 37.95 | 38.88 | 39.84 | 40.58 | 35.93 |

Table 3. Comparisons of p-mAP under different temporal offset thresholds on 3 different interaction settings.

| AP Hand | AP H.+Side | AP H.+State | mAP H.+Obj | mAP H.+All |
|---|---|---|---|---|
| 90.81 | 90.35 | 73.31 | 46.51 | 46.24 |

Table 4. Results of the baseline for the EHOI detection task.

tric videos of the ENIGMA-51 dataset. Differently from the standard definition of untrimmed action detection, in this task, a prediction is represented as a tuple $(\hat{c}, \hat{t}_k, s)$, where $\hat{c}$ and $\hat{t}_k$ are respectively the predicted class and key timestamp (the timestamp of the interaction key frame) and $s$ is a confidence score.

**Evaluation Measures:** We evaluated our baselines using point-level detection mAP (p-mAP) [52]. We considered predictions as correct when they satisfied two criteria: 1) the interaction class matched the ground truth and 2) the difference between the predicted and ground truth timestamps is within a certain temporal threshold. We considered different temporal offset thresholds ranging from 1 to 10 seconds with an increment of one second [20,21]; we averaged these values obtaining the mp-mAP values.

**Baseline:** Our baseline for this task is based on ActionFormer [60]. It takes the pre-extracted video features as input and gives action boundaries (start and end timestamps) as outputs. Given our focus on predicting the timestamp when the HOI occurs, we considered only the predicted action start[4] as output given by ActionFormer.

**Results:** Table 3 reports the results of the baseline. We considered three variants of the task: 1) detecting only *contact* and *de-contact* interactions (first row), 2) considering only *take* and *release* interactions (second row), and 3) considering all the four interactions (third row). We obtained mp-mAP values of 41.45%, 66.41%, and 35.93%, respectively, for *"take vs. release"*, *"first contact vs. de-contact"*, and *"all interactions"*. The results highlight that detecting *"take"* and *"release"* interactions (first row) are more challenging compared to finding *"first contact"* and *"de-contact"* interactions (41.45% vs. 66.41%) due to the different semantic complexity. Moreover, when all the four interactions are considered, the performance decreases, obtaining a mp-mAP of 35.93%[4].

## 4.2. Egocentric HOI Detection

**Task:** We consider the problem of detecting EHOIs from egocentric RGB images following the task definition proposed in [28, 49]. Given an input image, the aim is to predict the triplet *<hand, hand contact state, active object>*. Additional details about the task are reported in [28,49].

**Baselines:** The adopted baseline is based on the method proposed in [49]. We used the implementation proposed in [28] which extends a two-stage object detector with additional modules that exploit hand features to predict the *hand contact state* (in contact or not in contact), the *side of hand* (left and right), and an *offset vector* that indicates which object the hand is interacting with. Since the considered baseline is able to detect at most one contact per hand, we selected a subset of the $12,597$ interaction frames. This subset contains $15,955$ hands of which $8,753$ are in contact with an object, for a total of $7,680$ active objects.

**Evaluation Measures:** We used the following metrics based on standard *Average Precision* [28,49]: 1) *AP Hand*: *AP* of the hand detections, 2) *AP Hand+Side*: *AP* of the hand detections considering the correctness of the hand side, 3) *AP Hand+State*: *AP* of the hand detections considering the correctness of the hand state, 4) *mAP Hand+Obj*: *mAP* of the *<hand, active object>* detected pairs, and 5) *mAP Hand+All*: combinations of *AP Hand+Side*, *AP Hand+State*, and *mAP Hand+Obj* metrics.

**Results:** Table 4 reports the results obtained with the proposed baseline. Results show that the baseline achieved a *AP Hand* of 90.81%, a *AP Hand + Side* of 90.35% ), a *mAP H.+State* of 73.31%, a *mAP H.+Obj* of 46.51% and a *mAP H.+All* of 46.24%, pointing out that the use of domain-specific data in training is needed to exploit the knowledge of the industrial objects to support workers in the industrial domain.

## 4.3. Short-Term Object Interaction Anticipation

**Task:** The short-term object interaction anticipation task [22] aims to detect and localize the next-active objects, to predict the verb that describes the future interaction, and to determine when the interaction will start. Formally, the task consists in predicting future object interactions from a video $V$ and a timestamp $t$. The models can only use the video frames up to time $t$ and have to produce a set of predictions for the object interactions that will occur after a time interval $\delta$. Predictions consist of a bounding box over the next-active objects, a noun label, a verb label describing the future interaction, a real number indicating how soon the next interaction will start, and a confidence score.

**Evaluation Measures:** We evaluated the model's perfor-

[4] Additional information on implementation details, experiments, and results are reported in the supplementary material.

| Noun | N+V | N+TTC | Overall |
|---|---|---|---|
| 78.79 | 62.58 | 35.77 | 27.83 |

Table 5. Results% in Top-5 mean Average Precision for the Short-Term Object Interaction Anticipation task. N stands for noun, N+V stands for Noun+Verb and N+TTC stands for Noun+Time to Contact.

mance with Top 5 mean Average precision measures [22] that capture different aspects of the task: Top-5 mAP Noun, Top-5 mAP Noun+Verb, Top-5 mAP Noun+TTC, and Top-5 mAP Noun+Verb+TTC, which is also referred to as Top-5 mAP Overall.

**Baseline:** We adopted StillFast [40] as the baseline[5]. The model has been designed to extract 2D features from the considered past frame and 3D features from the input video clip. Feature stacks are merged through a combined feature pyramid layer and sent to the prediction head which is based on the Faster-RCNN head [43]. Features are fused and used to predict object (noun), verb probability distributions and time-to-contact (ttc) through linear layers along with the related prediction score $s$.

**Results:** Table 5 reports the results on test set of the ENIGMA-51 dataset considering the Top-5 mAP measures. StillFast obtains a Noun Top-5 mAP of 78.79%, demonstrating the ability to detect and classify the next-active objects processing images and videos simultaneously. When verbs and time to contact are predicted, performance drops according to Noun+Verb Top-5 mAP of 62.58%, Noun+TTC Top-5 mAP of 35.77%, and Overall Top-5 mAP of 27.83%. Qualitative results are reported in the supplementary material.

### 4.4. NL Understanding of Intents and Entities

**Task:** We considered the problem of classifying the intent of a user utterance, falling into one of the considered 24 classes, as well as the problem of entity slot filling, including four different slot types: *"object"*, *"board"*, *"component"* and *"procedure"*. Given an input utterance $U$, the task is to predict the intent class $i$, and to detect any entities $e$, if present, as well as the slot types $t$ associated to them, outputting zero or more $<e, t>$ couples. The complete list of intents/entities is reported in the supplementary material.

**Evaluation Measures:** We evaluate the baseline using the standard accuracy, and F1-score evaluation measures.

**Baseline:** The baseline is based on the DIETClassifier [6]. We performed the tokenization and featurization steps before passing the utterances to the model. Specifically, we used the SpacyNLP, SpacyTokenizer, CountVectorsFeaturizer, SpacyFeaturizer and DIETClassifier components offered by the Rasa framework [5].

**Results:** Table 6 reports the results obtained for intent and entity classification. Five different variants of the training set (see Section 3.2) were explored: real data, real data + G10 data, real data + G50 data, real data + G100 data, and G100. The best results for the intent classification have been obtained using only real data obtaining an accuracy of 0.867 and an F1-score of 0.844. The baseline suffers when generated data are included, which introduces noise and makes performance worse, reaching an accuracy of 0.584 (-0.283) and an F1-score of 0.564 (-0.280). These results suggest that, in this challenging industrial scenario, generative models, such as GPT [35] 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 of human's utterances which represent more simple concepts with respect to the human's intents, only generated data (last row) are 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 F1-score)[6].

| | Intent | | Entity | |
|---|---|---|---|---|
| **Training** | **Accuracy** | **F1-score** | **Accuracy** | **F1-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 6. Results for intents and entities classification considering different sets of training data.

## 5. Conclusion

We proposed ENIGMA-51, a new egocentric dataset acquired in an industrial environment and densely annotated to study human behavior. In addition, we performed baseline experiments aimed to study different aspects of human behavior in the industrial domain addressing four tasks. Existing methods show promising results but are still far from reaching reasonable performance to build an intelligent assistant able to support workers in the industrial domain. This opens up opportunities for future in-depth investigations.

## ACKNOWLEDGEMENTS

This research is supported by Next Vision[7] s.r.l., by MISE - PON I&C 2014-2020 - Progetto ENIGMA - CUP: B61B19000520008, and MIUR AIM Linea 1 - AIM1893589 - CUP: E64118002540007, and by the project FAIR – PNRR MUR Cod. PE0000013 - CUP: E63C22001940006.

[5] https://github.com/fpv-iplab/stillfast

[6] Additional details about the used prompting, the implementation details, and the results are reported in the supplementary material.

[7] Next Vision: https://www.nextvisionlab.it/

## References

[1] Artec eva. https://www.artec3d.com/portable-3d-scanners/artec-eva. 6

[2] 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. 1

[3] Manuel Benavent-Lledo, Sergiu Oprea, John Alejandro Castro-Vargas, David Mulero-Perez, and Jose Garcia-Rodriguez. Predicting human-object interactions in egocentric videos. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, pages 1–7, 2022. 4

[4] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection, 2020. 4

[5] Tom Bocklisch, Joey Faulkner, Nick Pawlowski, and Alan Nichol. Rasa: Open source language understanding and dialogue management. *arXiv preprint arXiv:1712.05181*, 2017. 8

[6] Tanja Bunk, Daksh Varshneya, Vladimir Vlasov, and Alan Nichol. Diet: Lightweight language understanding for dialogue systems. *arXiv preprint arXiv:2004.09936*, 2020. 4, 8

[7] Guo Chen, Sen Xing, Zhe Chen, Yi Wang, Kunchang Li, Yizhuo Li, Yi Liu, Jiahao Wang, Yin-Dong Zheng, Bingkun Huang, Zhiyu Zhao, Junting Pan, Yifei Huang, Zun Wang, Jiashuo Yu, Yinan He, Hongjie Zhang, Tong Lu, Yali Wang, Liming Wang, and Yu Qiao. Internvideo-ego4d: A pack of champion solutions to ego4d challenges. *ArXiv*, abs/2211.09529, 2022. 4

[8] Qian Chen, Zhu Zhuo, and Wen Wang. Bert for joint intent classification and slot filling. *arXiv preprint arXiv:1902.10909*, 2019. 4

[9] Sara Colombo, Yihyun Lim, and Federico Casalegno. Deep vision shield: Assessing the use of hmd and wearable sensors in a smart safety device. In *Proceedings of the 12th ACM International Conference on PErvasive Technologies Related to Assistive Environments*, PETRA '19, page 402–410, New York, NY, USA, 2019. Association for Computing Machinery. 1

[10] MMPose Contributors. Openmmlab pose estimation toolbox and benchmark. https://github.com/open-mmlab/mmpose, 2020. 2, 6

[11] 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. 1, 3

[12] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Jian Ma, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision (IJCV)*, 130:33–55, 2022. 3

[13] Oscar Danielsson, Magnus Holm, and Anna Syberfeldt. Augmented reality smart glasses for operators in production: Survey of relevant categories for supporting operators. *Procedia CIRP*, 93:1298–1303, 2020. 53rd CIRP Conference on Manufacturing Systems 2020. 1

[14] Ahmad Darkhalil, Dandan Shan, Bin Zhu, Jian Ma, Amlan Kar, Richard Higgins, Sanja Fidler, David Fouhey, and Dima Damen. Epic-kitchens visor benchmark: Video segmentations and object relations. In *Neural Information Processing Systems (NIPS)*, 2022. 3, 4

[15] E. Dessalene, C. Devaraj, M. Maynord, C. Fermuller, and Y. Aloimonos. Forecasting action through contact representations from first person video. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, pages 1–1, 2021. 4

[16] Eadom Dessalene, Michael Maynord, Chinmaya Devaraj, Cornelia Fermüller, and Yiannis Aloimonos. Egocentric object manipulation graphs. *ArXiv*, abs/2006.03201, 2020. 4

[17] Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J. Black, and Otmar Hilliges. ARCTIC: A dataset for dexterous bimanual hand-object manipulation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[18] C. Feichtenhofer, H. Fan, J. Malik, and K. He. Slowfast networks for video recognition. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 6202–6211, 2018. 4

[19] Antonino Furnari, Sebastiano Battiato, Kristen Grauman, and Giovanni Maria Farinella. Next-active-object prediction from egocentric videos. *J. Vis. Commun. Image Represent.*, 49:401–411, 2017. 4

[20] Mingfei Gao, Mingze Xu, Larry S Davis, Richard Socher, and Caiming Xiong. Startnet: Online detection of action start in untrimmed videos. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 5542–5551, 2019. 7

[21] Mingfei Gao, Yingbo Zhou, Ran Xu, Richard Socher, and Caiming Xiong. Woad: Weakly supervised online action detection in untrimmed videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1915–1923, 2021. 7

[22] 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. Somasundaram, 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. 1, 3, 4, 7, 8

[23] E Haihong, Peiqing Niu, Zhongfu Chen, and Meina Song. A novel bi-directional interrelated model for joint intent detection and slot filling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5467–5471, 2019. 4

[24] Dilek Z. Hakkani-Tür, Gökhan Tür, Asli Celikyilmaz, Yun-Nung (Vivian) Chen, Jianfeng Gao, Li Deng, and Ye-Yi Wang. Multi-domain joint semantic frame parsing using bi-directional rnn-lstm. In *Interspeech*, 2016. 4

[25] Jingjing Jiang, Zhixiong Nan, Hui Chen, Shitao Chen, and Nanning Zheng. Predicting short-term next-active-object through visual attention and hand position. *Neurocomputing*, 433:212–222, 2021. 4

[26] Lei Ke, Mingqiao Ye, Martin Danelljan, Yifan Liu, Yu-Wing Tai, Chi-Keung Tang, and Fisher Yu. Segment anything in high quality. *arXiv:2306.01567*, 2023. 2, 6

[27] 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. 1

[28] Rosario Leonardi, Francesco Ragusa, Antonino Furnari, and Giovanni Maria Farinella. Exploiting multimodal synthetic data for egocentric human-object interaction detection in an industrial scenario. *arXiv preprint arXiv:2306.12152*, 2023. 4, 7

[29] Yin Li, Miao Liu, and James M. Rehg. In the eye of the beholder: Gaze and actions in first person video, 2020. 2, 3

[30] 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. 1, 3

[31] Yao Lu and Walterio W Mayol-Cuevas. Egocentric hand-object interaction detection and application, 2021. 1, 4

[32] Matterport. https://matterport.com/. 6

[33] Michele Mazzamuto, Francesco Ragusa, Alessandro Resta, Giovanni Maria Farinella, and Antonino Furnari. A wearable device application for human-object interactions detection. In *International Conference on Computer Vision Theory and Applications (VISAPP)*, 2023. 1

[34] Microsoft hololens 2. https://www.microsoft.com/en-us/hololens. 1, 5

[35] Openai chatgpt. https://openai.com/blog/chatgpt. 6, 8

[36] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023. 2, 6

[37] Razvan-George Pasca, Alexey Gavryushin, Yen-Ling Kuo, Otmar Hilliges, and Xi Wang. Summarize the past to predict the future: Natural language descriptions of context boost multimodal object interaction. *arXiv preprint arXiv:2301.09209*, 2023. 4

[38] 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. 1, 2, 3

[39] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2021. 2, 6

[40] Francesco Ragusa, Giovanni Maria Farinella, and Antonino Furnari. Stillfast: An end-to-end approach for short-term object interaction anticipation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2023. 4, 8

[41] 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. 1, 3, 4, 5

[42] Francesco Ragusa, Antonino Furnari, Salvatore Livatino, and Giovanni Maria Farinella. The meccano dataset: Understanding human-object interactions from egocentric videos in an industrial-like domain. In *Proceedings of the IEEE Workshop on Applications of Computer Vision (WACV)*, 2021. 4

[43] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Neural Information Processing Systems (NIPS)*, 28, 2015. 4, 8

[44] Grégory Rogez, James S. Supancic, and Deva Ramanan. Understanding everyday hands in action from rgb-d images. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 3889–3897, 2015. 3

[45] Justyna Sarzynska-Wawer, Aleksander Wawer, Aleksandra Pawlak, Julia Szymanowska, Izabela Stefaniak, Michal Jarkiewicz, and Lukasz Okruszek. Detecting formal thought disorder by deep contextualized word representations. *Psychiatry Research*, 304:114135, 2021. 4

[46] Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681, 1997. 4

[47] Fadime Sener, Dibyadip Chatterjee, Daniel Shelepov, Kun He, Dipika Singhania, Robert Wang, and Angela Yao. As-

sembly101: 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. 1, 3, 5

[48] Dandan Shan, Jiaqi Geng, Michelle Shu, and David Fouhey. Understanding human hands in contact at internet scale. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[49] 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. 1, 3, 4, 6, 7

[50] Dingfeng Shi, Yujie Zhong, Qiong Cao, Lin Ma, Jia Li, and Dacheng Tao. Tridet: Temporal action detection with relative boundary modeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18857–18866, 2023. 4

[51] Yangyang Shi, Kaisheng Yao, Hu Chen, Yi-Cheng Pan, Mei-Yuh Hwang, and Baolin Peng. Contextual spoken language understanding using recurrent neural networks. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5271–5275, 2015. 4

[52] Zheng Shou, Junting Pan, Jonathan Chan, Kazuyuki Miyazawa, Hassan Mansour, Anthony Vetro, Xavier Giro-i Nieto, and Shih-Fu Chang. Online detection of action start in untrimmed, streaming videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 534–551, 2018. 7

[53] Yansong Tang, Yi Tian, Jiwen Lu, Jianjiang Feng, and Jie Zhou. Action recognition in rgb-d egocentric videos. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, pages 3410–3414, 2017. 3

[54] Pavel Tokmakov, Jie Li, and Adrien Gaidon. Breaking the "object" in video object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[55] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Neural Information Processing Systems (NIPS)*, 35:10078–10093, 2022. 4

[56] Andrea Vanzo, Emanuele Bastianelli, and Oliver Lemon. Hierarchical multi-task natural language understanding for cross-domain conversational ai: Hermit nlu. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 254–263, 2019. 4

[57] Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. Videomae v2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14549–14560, 2023. 4

[58] Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, et al. Internvideo: General video foundation models via generative and discriminative learning. *arXiv preprint arXiv:2212.03191*, 2022. 4

[59] Hainan Xu, Fei Jia, Somshubra Majumdar, He Huang, Shinji Watanabe, and Boris Ginsburg. Efficient sequence transduction by jointly predicting tokens and durations. *arXiv preprint arXiv:2304.06795*, 2023. 4

[60] Chen-Lin Zhang, Jianxin Wu, and Yin Li. Actionformer: Localizing moments of actions with transformers. In *Proceedings of the European Conference on Computer Vision (ECCV)*, volume 13664 of *LNCS*, pages 492–510, 2022. 3, 4, 7

[61] Hao Zhang, Feng Li, Siyi Liu, Lei Zhang, Hang Su, Jun-Juan Zhu, Lionel Ming shuan Ni, and Heung yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *ArXiv*, abs/2203.03605, 2022. 4

[62] Lingzhi Zhang, Shenghao Zhou, Simon Stent, and Jianbo Shi. Fine-grained egocentric hand-object segmentation: Dataset, model, and applications. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 127–145, 2022. 4

[63] Dimitri Zhukov, Jean-Baptiste Alayrac, Ramazan Gokberk Cinbis, David Fouhey, Ivan Laptev, and Josef Sivic. Cross-task weakly supervised learning from instructional videos, 2019. 1