Put on your detective hat: What's wrong in this video?

Rohith Peddi[1] **Shivvrat Arya**[1] **Bhrath Challa**[1] **Likhitha Pallapothula**[1] **Akshay Vyas**[1] **Qifan Zhang**[1] **Jikai Wang**[1] **Vasundhara Komaragiri**[1] **Eric Ragan**[2] **Nicholas Ruozzi**[1] **Yu Xiang**[1] **Vibhav Gogate**[1]

Abstract

Following step-by-step procedures is an essential component of various activities carried out by individuals in their everyday lives. These procedures serve as a guiding framework that helps achieve goals efficiently, whether assembling furniture or preparing a recipe. However, the complexity and duration of procedural activities inherently increase the likelihood of making errors. Understanding such procedural activities from a sequence of frames is a challenging task that demands an accurate interpretation of visual information and an ability to reason about the structure of the activity. To this end, we collected a new ego-centric 4D dataset comprising 380 recordings (90 hrs) of people performing recipes in kitchen environments. This dataset consists of two distinct activity types: one in which participants adhere to the provided recipe instructions and another where they deviate and induce errors. We provide 5K step annotations and 10K fine-grained action annotations for 20% of the collected data and benchmark it on two tasks: error detection and procedure learning. Our code and data can be accessed from our website

1. Introduction

Remember when you prepared your favorite meal after a long day and missed adding that crucial ingredient and then lost your appetite after a few bites?

Such scenarios are quite common because performing long-horizon step-by-step procedural activities increases the likelihood of making errors. These errors can be harmless, provided they can be rectified with little consequence. Nonetheless, when the procedures in question pertain to the medical field or complex chemical experiments, the cost of errors can be substantial. Therefore, there is a pressing need for building AI systems that can guide users in performing procedural activities (Draper, 2021).

A key problem that we need to solve in order to build such AI systems is *procedural activity understanding*, a challenging and multi-faceted task that demands interpreting what is happening, anticipating what will happen, and planning the course of action to accomplish the goal. For a system to interpret what is happening, the system needs to recognize and segment actions while assessing the current state of the environment. In order to anticipate future events, the system should be able to predict actions at the beginning of an interaction or even beforehand. On the other hand, planning a sequence of actions requires the system to understand the possible outcomes of these interactions.

We introduce a large-scale dataset to aid the learning of AI systems capable of recognizing an error and anticipating it before making an error during the performance of procedural activities. We selected cooking as a domain that is sufficiently complex and encompasses different kinds of errors that are compounding in nature and completely alter the current state of the environment with no point of return. We decided to capture data from an ego-centric view despite ego motions because it helps minimize occlusions more effectively than third-person videos.

This paper makes the following **contributions**:

1. We collected an ego-centric 4D dataset that features individuals following recipes in kitchen settings. This dataset includes two distinct types of activities: one where the participants precisely follow the given recipe guidelines, and another where they deviate, making errors.
2. We provide annotations for (1) Start/End times for each step of the recipe, (2) Start/End times for each action/interaction for 20% of the collected data, (3) Categorize and provide a detailed description of the error performed by a participant (see figure 3, which illustrates the key steps for a recording and the corresponding step and action annotations).

[*] Equal contribution [1] Department of Computer Science, University of Texas at Dallas, Dallas, USA [2] Department of Computer Science, University of Florida, Gainesville, USA. Correspondence to: Rohith Peddi <rohith.peddi@utdallas.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

3. We provide baselines for two tasks, namely, error detection and procedure learning.

2. Preliminaries

We describe the terms used in the subsequent sections. Following the terminology used in scientific disciplines such as neuroscience (Chevignard et al., 2010) and chemistry, we will refer to deviations from procedures as *errors*. Note that the term "errors" used here are equivalent to what is commonly called "mistakes" in the AI community (c.f. (Fadime Sener et al., 2022)). Following (Chevignard et al., 2010; Finnanger et al., 2021; Fogel et al., 2020), we classified common errors performed during a cooking activity into the following categories (1) Preparation Error, (2) Measurement Error, (3) Technique Error, (4) Timing Error, (5) Temperature Error, (6) Missing Steps, and (7) Ordering Errors. Figure 1 displays frames taken from different recipes in our dataset, each corresponding to a distinct type of error as described above. Further, the annotations in our dataset have enabled us to gather a comprehensive overview of different error types and their concise explanations, as depicted in figure 2.

The term **recording** refers to the comprehensive 4D dataset collected while a participant performs a cooking activity. A recording is classified as a **normal recording** when obtained while the participant precisely follows the recipe's procedure. Conversely, a recording is called an **error recording** when it is captured while the individual deviates from the recipe's procedure, thereby inducing errors.

3. Related Work

Temporal Action Segmentation A video understanding task where an untrimmed video sequence is segmented and given a label from a predetermined set of action labels. Procedural activity datasets, encompassing both recorded datasets (such as Breakfast (Kuehne et al., 2014), GTEA(Fathi et al., 2011), and 50Salads(Stein & McKenna, 2013)) and curated datasets (such as CrossTask(Zhukov et al., 2019) and COIN(Tang et al., 2019), YouTubeInstructional()), have been instrumental in driving progress in the field of temporal action segmentation.

Unfortunately, the datasets currently used for this task have various shortcomings. For instance, apart from Assembly-101, the other core datasets for the task, like Breakfast, GTEA, and 50Salads, are relatively smaller and are characterized by limited variations in the temporal order of the steps performed to complete the procedure. Additionally, all the datasets have shorter average lengths for the segments related to each step in the procedure. Our dataset, with its notable scale, combined with a wide range of temporal variations in the steps taken, effectively addresses the limitations observed in existing datasets.

3D activity analysis On the other hand, the NTU RGB+D (Shahroudy et al., 2016) dataset serves as a valuable benchmark for 3D action recognition, but its usage for procedural activity understanding is limited. The Ego4D(Grauman et al., 2021) dataset is a notable exception, offering a large-scale 3D dataset with diverse variations. However, it is constrained by a fixed camera position that captures depth information, resulting in occlusions that hinder fine-grained activity understanding. Multi-view datasets like Assembly-101 (Fadime Sener et al., 2022) offer a compelling choice but lack a depth component, thus limiting their usage here.

While several 3D activity analysis datasets are accessible, it's crucial to understand that they all possess constraints when interpreting activities, specifically procedural activities. For instance, the MSR-Daily Activity dataset (Wang et al., 2012) contains 320 samples of 16 daily activities and has a limited sample size and fixed camera viewpoints. Similarly, the RGBD HuDaAct dataset (Ni et al., 2011) includes 1189 videos and 12 daily human actions but lacks variety in its scenarios. As for the G3D (Bloom et al., 2012) and PKUMMD (Liu et al., 2017) datasets provide continuous sequences but are limited to videos captured in a single environment.

Our dataset stands unparalleled in its ability to surmount the challenges presented by other 3D datasets when it comes to procedural tasks. It's worth mentioning that our dataset features recordings that average 32 seconds per segment, and we firmly believe that the 3D insights offered by this dataset will catalyze remarkable progress in 3D activity analysis as well as 3D action recognition.

Procedure Learning Procedure Learning is a two-part process where all video frames are first segregated into K significant steps. Then a logical sequence of the steps necessary to complete the task is identified. Action segmentation differs from procedure learning, which focuses on segmenting actions without considering their relevance to task completion; on the contrary, procedure learning aims to find commonalities among the key steps needed to complete the task captured in multiple videos. Existing procedural activity datasets like CrossTask (Zhukov et al., 2019), COIN (Tang et al., 2019) are predominantly third-person videos; in this light, (Bansal, Siddhant et al., 2022) compiled videos from CMU-MMAC (De la Torre et al., 2008), EGTEA (Fathi et al., 2011), EPIC-Tents (Jang et al., 2019), MECCANO (Ragusa et al., 2020) and curated EgoProceL dataset.

We note that our dataset not only has a higher average step length but also is significantly larger than the currently available datasets for Egocentric Procedure Learning. Through

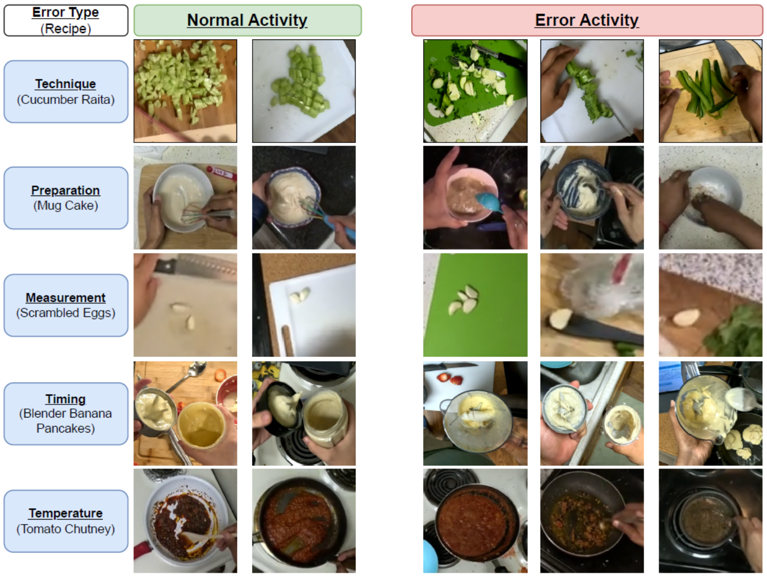

Figure 1. Each row displays frames captured from different recordings of recipes, highlighting both correct and erroneous executions, with a focus on specific types of errors. The first row pertains to the recipe *cucumber raita*. The two frames on the left depict the outcome when the instruction *chop into pieces* is correctly followed, while the three frames on the right show the results when the cucumber is cut improperly, sliced vertically, and sliced horizontally, respectively. The second row corresponds to the recipe *mug cake*, where the two frames on the left demonstrate the proper execution of the instruction *Whisk batter*, while the remaining frames depict incorrect usage of utensils such as a spoon, tablespoon, and hand to perform the same task. The third row corresponds to the recipe *scrambled eggs*. The left two frames exhibit the outcome of correctly following the instruction *Peel 2 garlic cloves*, whereas the subsequent frames display the result when a different number of garlic cloves (4, 1, and 1 respectively) are peeled instead of the intended 2 cloves. The fourth row represents the recipe *blender banana pancakes*, with the two frames on the left showing the result of following the instruction *blitz for 20 seconds*, while the remaining frames showcase the output when the blender is operated for a shorter duration. Finally, the fifth row corresponds to the recipe *tomato chutney*, where the first two frames depict the outcome of the instruction *pan over medium/low heat*, while the subsequent frames demonstrate the result when an incorrect temperature setting is used.

experimental evaluation, we show in 5.1 that average step length indeed significantly hinders the performance of currently proposed approaches.

Error Recognition Given a video clip, error recognition involves identifying errors present in the clip. This task was initially introduced as mistake detection by Assembly-101 (Fadime Sener et al., 2022) and proposed a 3-class classification on the performed procedure to classify the clip as either correct, mistake or correction. We also note that (Soran et al., 2015) is one of the early works in similar lines, where they send a notification every time they miss a step when performing an activity. Anomaly detection, while closely related to error recognition, differentiates itself by utilizing static cameras and backgrounds to identify unusual or abnormal behaviour.

By incorporating a wide range of error types, such as timing, preparation, temperature, technique, and measurement errors (please see figure. 2), our dataset allows researchers to gain insights into error patterns across diverse contexts. This analysis fosters the development of robust error recognition systems and contributes to advancing the field of activity analysis and performance improvement.

Error Anticipation Inspired by the action anticipation proposed by (Damen et al., 2020), we propose a novel error anticipation task that involves proactively identifying and predicting errors before they occur during an activity. The process of error anticipation entails recognizing the indicative patterns that precede errors. Such early recognition provides the opportunity for prompt intervention and correction, ultimately resulting in improved performance and significantly fewer errors.

Error anticipation in activity analysis involves utilizing contextual cues, temporal patterns, and task-specific knowledge to forecast potential errors accurately. Our dataset, which encompasses both normal and error recordings for each recipe, enables researchers to design and develop algorithms that can understand the current state of the environment and analyze implications for action with common-sense knowledge leading to effective error anticipation techniques.

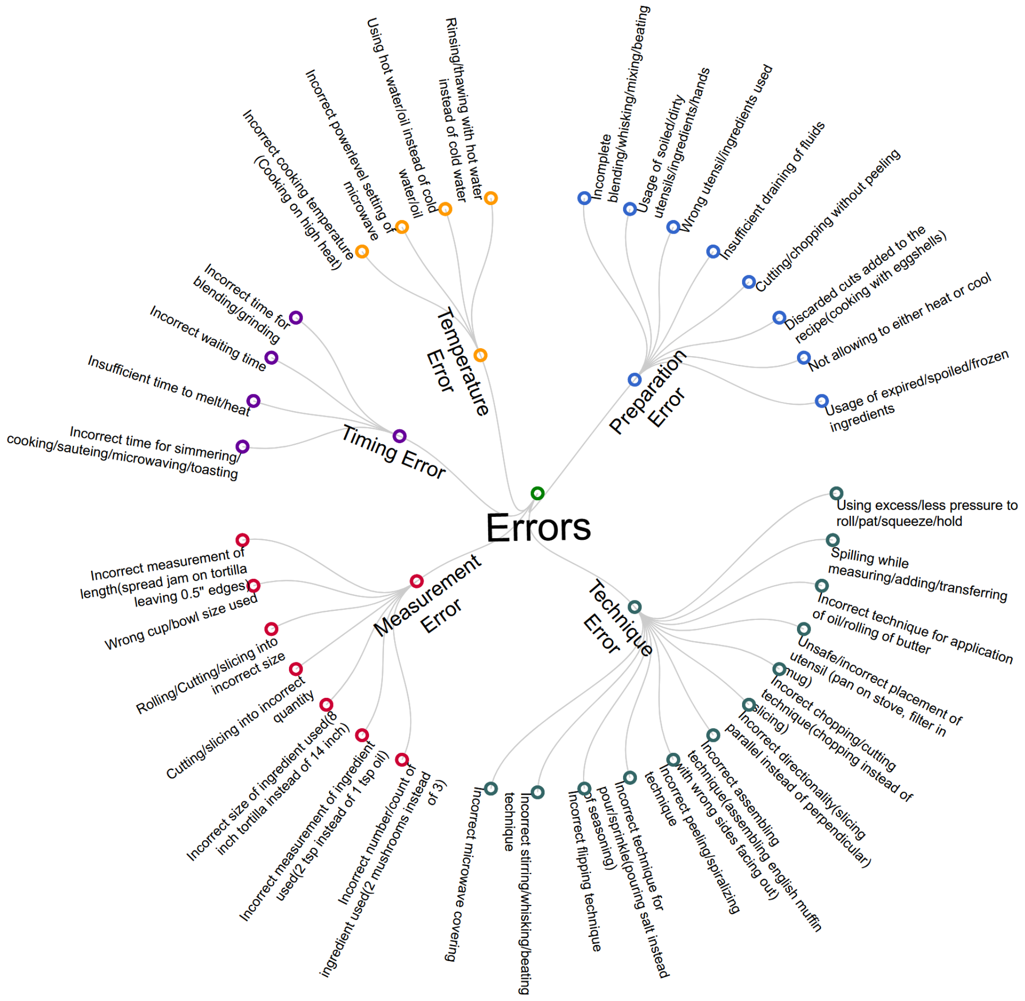
Errors
Temperature Error
Incorrect cooking temperature (Cooking on high heat)
Incorrect powerlevel setting of microwave
Using hot water/oil instead of cold water/oil
Rinsing/thawing with hot water instead of cold water
Preparation Error
Incomplete blending/whisking/mixing/beating
Usage of soiled/dirty utensils/ingredients/hands
Wrong utensil/ingredients used
Insufficient draining of fluids
Cutting/chopping without peeling
Discarded cuts added to the recipe(cooking with eggshells)
Not allowing to either heat or cool
Usage of expired/spoiled/frozen ingredients
Timing Error
Incorrect time for blending/grinding
Incorrect waiting time
Insufficient time to melt/heat
Incorrect time for simmering/ cooking/sauteing/microwaving/toasting
Measurement Error
Incorrect measurement of length(spread jam on tortilla leaving 0.5" edges)
Wrong cup/bowl size used
Rolling/Cutting/slicing into incorrect size
Cutting/slicing into incorrect quantity
Incorrect size of ingredient used(8 inch tortilla instead of 14 inch)
Incorrect measurement of ingredient used(2 tsp instead of 1 tsp oil)
Incorrect number/count of ingredient used(2 mushrooms instead of 3)
Technique Error
Using excess/less pressure to roll/pat/squeeze/hold
Spilling while measuring/adding/transferring
Incorrect technique for application of oil/rolling of butter
Unsafe/incorrect placement of utensil (pan on stove, filter in mug)
Incorrect chopping/cutting technique(chopping instead of slicing)
Incorrect directionality(slicing parallel instead of perpendicular)
Incorrect assembling technique(assembling english muffin with wrong sides facing out)
Incorrect peeling/spiralizing technique
Incorrect technique for pour/sprinkle(pouring salt instead of seasoning)
Incorrect flipping technique
Incorrect stirring/whisking/beating technique
Incorrect microwave covering

Figure 2. A structured synopsis of different types of errors and their short descriptions compiled from the annotations.

4. Data Collection

Sensors In order to gather activity data, we employed a combination of the GoPro Hero 11 camera, which was mounted on the user's head, and the Hololens2 device. To facilitate data collection from the HoloLens2, including its depth sensor, IMU (Inertial Measurement Unit), front RGB camera, and microphone, we utilized a custom tool developed by Dibene, Juan C. & Dunn, Enrique (2022). Furthermore, we captured the processed head and hand tracking information provided by the HoloLens2 device.

Recipes We curated a selection of 24 cooking recipes sourced from WikiHow, specifically focusing on recipes with a preparation time of 30 minutes or less. These recipes encompassed a wide range of culinary traditions (see X-axis of the chart in Fig. 3), showcasing the diversity of cooking styles across various cuisines. Our primary objective was to explore potential errors that may arise when utilizing different cooking tools while preparing various types of cuisine.

Task Graphs A task graph is a visual representation of the sequential steps required to accomplish a given recipe. Each node in the task graph (for a recipe) corresponds to a step in a recipe and a directed edge between a node x and a node y in the graph indicates that x must be performed before y. Thus, a task graph is a directed acyclic graph, and a topological sort over it represents a valid completion of the recipe. In order to construct task graphs for our collection of 24 WikiHow recipes, we meticulously identified all the essential steps involved and established their interdependencies, thereby establishing a topological order of

Table 1. Comparing the datasets relevant for the respective tasks considered above

	Ego	**Multi modal**	Year	#Videos	Hours	Participants
GTEA Gaze+ (Fathi et al., 2011)	✓	RGB	2012	37		26
50 Salads (Stein & McKenna, 2013)	×	RGB-D	2013	50	4.5	25
Breakfast (Kuehne et al., 2014)	×	RGB	2014	1,712	77.0	52
MPII Cooking 2 (Rohrbach et al., 2015)	×	RGB	2015	273	27	30
YouCook2 (Zhou et al., 2017)	×	RGB	2017	2000	176	N.A.
EGTEA Gaze+ (egt, 2018)	✓	RGB	2018	86	29	32
CrossTask (Zhukov et al., 2019)	×	RGB	2019	4700	376	N.A.
COIN (Tang et al., 2019)	×	RGB	2019	11,827	476	N.A.
Assembly101 (Fadime Sener et al., 2022)	✓	RGB	2022	447	53.0	53
Ours	✓	**RGB-D**	2023	380	**90**	8

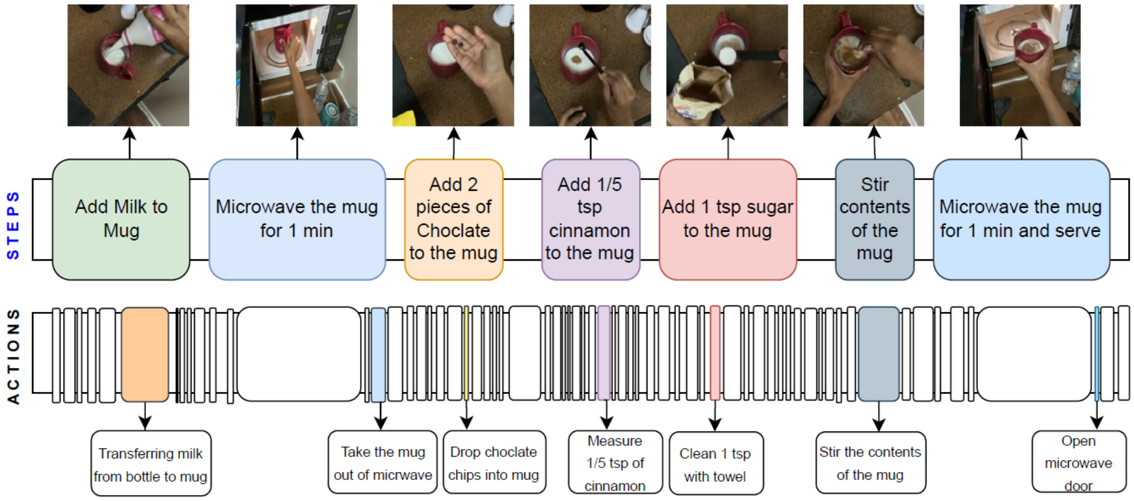
STEPS
Add Milk to Mug
Microwave the mug for 1 min
Add 2 pieces of Choclate to the mug
Add 1/5 tsp cinnamon to the mug
Add 1 tsp sugar to the mug
Stir contents of the mug
Microwave the mug for 1 min and serve
ACTIONS
Transferring milk from bottle to mug
Take the mug out of micrwave
Drop choclate chips into mug
Measure 1/5 tsp of cinnamon
Clean 1 tsp with towel
Stir the contents of the mug
Open microwave door

Figure 3. Figure describes key steps for a recording and the corresponding step and action annotations

tasks.

4.1. Protocol

Our dataset was compiled by eight participants across ten distinct kitchens in the United States. Each participant selected ten recipes and recorded, on average, 48 videos across 5 different kitchens. During filming, all participants were required to ensure that they were alone in the kitchen and remove any items that could potentially identify them, such as personal portraits, mirrors, and smartwatches with portraits. The participants used a GoPro and a HoloLens2 to record and monitor their footage.

Each participant was provided with a tablet-based recording interface accessible through a web browser. This interface facilitated the sorting of their recipe recordings into two categories: standard and error. Furthermore, the recordings were organized based on the specific kitchen setup. To ensure optimal video quality, we asked the participants to configure the GoPro camera such that it captures videos in 4K resolution at 30 frames per second. Additionally, the HoloLens2 device was programmed to stream RGB frames at a resolution of 360p and depth frames using the depth_ahat (Articulated Hand Tracking) mode.

4.1.1. Normal Recordings

Each participant in the study is tasked with selecting a recipe from the available options, which are scheduled within a kitchen setup using the recording interface. Subsequently, they are presented with one of the pre-established topological orders of the recipe, as determined by the previously constructed task graphs. Participants then proceed to follow the provided task graph, commencing from the beginning and progressing through each step in accordance with its dependencies and designated timing.

4.1.2. Error Recordings

We devised and implemented three strategies for the participants to follow. Each participant was asked to pick his recording strategy for a particular environment and was accordingly guided in preparing for his recording. We

Recipe	**# Steps**	**# Normal Recordings**	**Normal Rec Duration(Hr)**	**# Error Recordings**	**Error Rec Duration(Hr)**
Pinwheels	19	4	0.73	8	1.19
Tomato Mozzarella Salad	9	10	1.21	6	0.53
Butter Corn Cup	12	6	1.63	8	1.51
Tomato Chutney	19	7	3.34	8	2.01
Scrambled Eggs	23	6	1.99	10	3.13
Cucumber Raita	11	13	3.11	8	1.98
Zoodles	13	5	0.74	10	2.19
Microwave Egg Sandwich	12	7	0.88	12	1.67
Sauted Mushrooms	18	6	2.73	8	2.21
Blender Banana Pancakes	14	7	1.78	11	2.42
Herb Omelet with Fried Tomatoes	15	6	1.73	9	1.79
Broccoli Stir Fry	25	11	5.74	6	2.15
Pan Fried Tofu	19	8	3.38	7	2.31
Mug Cake	20	7	2.09	11	2.54
Cheese Pimiento	11	6	1.47	8	1.57
Spicy Tuna Avocado Wraps	17	7	1.70	11	2.66
Caprese Bruschetta	11	7	0.98	11	2.42
Dressed Up Meatballs	16	6	1.63	10	3.10
Microwave Mug Pizza	14	7	1.47	6	1.14
Ramen	15	9	2.10	7	1.45
Coffee	16	7	1.75	7	1.58
Breakfast Burritos	11	6	0.71	10	1.51
Spiced Hot Chocolate	7	6	0.70	10	1.01
Microwave French Toast	11	9	1.94	5	0.66
Ours (Total)	380	173	45.53	207	44.73

Table 2. Statistics for each recipe describe (1) the number of steps present, (2) the number of normal recordings present, (3) the total duration of all the normal recordings, (4) the number of error recordings present, (5) the total duration of all the error recordings present

list the formulated strategies here (1) **Pre-prepared error scripts**: The participants were given pre-prepared error scripts with missing steps and ordering errors. (2) **Prepare error scripts**: Once participants chose this strategy, they were given a web-based interface to create an error script for each error recipe recording and displayed the modified error script on a tablet enabling participants to perform according to their modified error scripts (3) **Impromptu**: During the later stages of the recording process, we implemented a strategy where participants were asked to induce errors intentionally. Following the completion of each recording, participants were given access to a web-based interface to update any errors they made during each step. Figure 4 describes the counts of types of errors in each recipe.

4.2. Data Annotation

We have meticulously annotated the collected data with the following annotations: (1) Annotations for coarse-grained actions or steps, providing the start and end times for each step within the recorded videos. (2) To support learning semi/weakly supervised approaches for action recognition and action anticipation, we have provided fine-grained action annotations for 20% of the recorded data. These annotations include the start and end times for each fine-grained action. (3) We have also categorized and provided error descriptions for the induced errors. These error descriptions are associated with the corresponding step in the provided annotations, allowing for a comprehensive understanding of the errors. Figure 3 describes the granularity of different categories of annotations provided.

To ensure high-quality annotations for our data, we used the following approach. We ensured that each recording was annotated by the person who recorded the video and then reviewed by another. The reviewer was asked to double-check that all errors made by the participant in the recording are included in their corresponding step annotations.

4.2.1. Coarse-Grained Action/Step Annotations

We designed an interface for performing step annotations in Label Studio [1]. Each annotator is presented with this interface to mark the start and end times for each step. Our steps are significantly longer than a single fine-grained action and encompass multiple fine-grained actions necessary for performing the described step. For example, in order to accomplish the step *{Chop a tomato}*, we include the following (1) Pre-conditional actions of *{opening refrigerator, grabbing a polythene bag of tomatoes, taking a tomato, placing the tomato on cutting board, close fridge}* (2) Post-

[1] https://labelstud.io/

Dataset	**Total Hours**	**# Videos**	**Avg Video Length (min)**	**# Segments**	**Avg # Segments**	**Avg # Segments Length (sec)**
50Salads	4.5	50	6.4	899	18	36.8
Breakfast	77	1712	2.3	11,300	6.6	15.1
Assembly 101	513	4321	7.1	104,759	24	16.5
Ours (Total)	90	380	**15**	5000	13.15	**32.00**

Table 3. Comparison of coarse-grained action/step annotations with relevant datasets

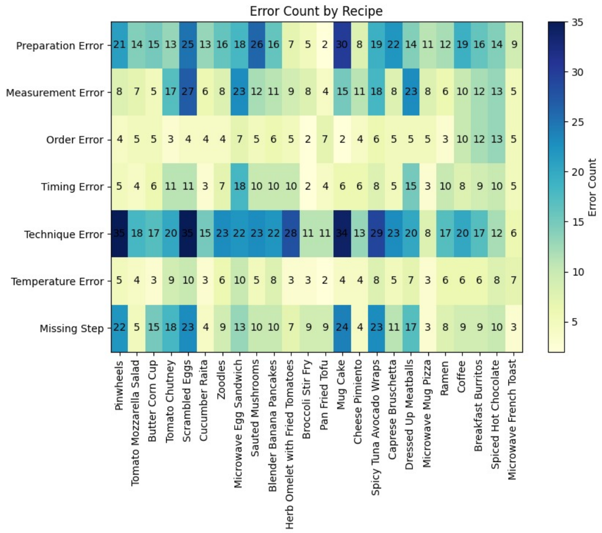

Figure 4. Distribution of error types and counts of different types of errors performed for each recipe

conditional actions of *{placing down the knife, grabbing the polythene bag of tomatoes, open fridge and place the bag in the fridge}*. Table 4.1.2 summarizes and compares coarse-grained action/step annotations for our dataset as well as other popular datasets.

4.2.2. Fine-Grained Action Annotations

Inspired by the pause-and-talk narrator (Damen et al., 2020), we have designed and developed a web-based tool for fine-grained action annotation that utilizes OpenAI's Whisper APIs for speech-to-text translation. Even though we have built this system around the Whisper API, it's flexible enough to accommodate any automatic speech recognition (ASR) system that can serve transcription requests. We will release the developed web-based annotation tool as part of our codebase.

4.3. Sources of bias

While this represents our first attempt at building a comprehensive 4D dataset to study mistakes in procedural tasks, we acknowledge the dataset's inherent biases. The number of participants contributing to this dataset is noticeably smaller than conventional, large-scale action or activity understanding datasets. Yet, it's important to mention that each participant is asked to perform and record the same recipe four times, and each time, the recording script changes, thus making each recording unique. Finally, note that because the participants followed a script, many errors were intentional. However, they also made unintentional errors in the process which they annotated later.

5. Baselines

5.1. Error Detection

As a baseline, we formulate the task of frame-level error detection as anomaly detection. More specifically, we use anomaly detection methods to classify each frame in each video as either normal or abnormal, where the latter is defined as an instance that deviates from the expected behavior (frame where participants made errors). We used two self-supervised anomaly detection methods from literature, self-supervised masked convolutional transformer block (SSMCTB) (Madan et al., 2022) and self-supervised predictive convolutional attentive block (SSPCAB) (Ristea et al., 2022), and trained them on top of ResNet-50 (He et al., 2015), where the latter serves as a neural, image-based feature extractor. Both models were trained using reconstruction loss (Madan et al., 2022). We evaluated the benchmark models using frame-level area under the curve (AUC) and Equal Error Rate (EER) scores. Table 5.1 shows the results. We observe that SSMCTB is slightly better than SSPCAB.

Method	**AUC**	**EER**
SSMCTB(Madan et al., 2022)	50.86 %	49.43 %
SSPCAB(Ristea et al., 2022)	47.94 %	51.12 %

Table 4. Zero-Shot Error Detection using Anomaly Detection Methods.

5.2. Procedure Learning

We chose recently proposed (Bansal, Siddhant et al., 2022) for evaluating on our dataset. This approach uses a self-

RECIPE	CRL			$CRL + CIDM_{0.1}$			$CRL + CIDM_{0.3}$			$CRL + CIDM_{0.5}$		
	P	R	IOU	P	R	IOU	P	R	IOU	P	R	IOU
BlenderBananaPancakes	12.65	9.50	5.16	15.54	9.96	5.72	18.51	10.51	6.38	15.13	10.37	5.75
BreakfastBurritos	18.72	11.46	6.77	16.58	10.77	5.87	16.23	11.21	6.57	20.46	13.16	7.60
BroccoliStirFry	9.92	9.11	3.93	8.20	8.10	3.85	6.98	7.06	3.34	10.03	8.29	3.93
ButterCornCup	13.82	11.85	5.79	15.07	12.30	5.82	16.13	12.52	5.99	14.47	11.11	5.36
CapreseBruschetta	25.55	12.89	7.52	20.53	9.09	5.59	17.46	10.33	5.94	13.07	9.00	4.93
CheesePimiento	19.74	10.48	6.44	17.49	10.32	6.26	14.69	8.88	5.42	17.41	10.43	6.01
Coffee	13.68	9.91	5.49	15.76	10.25	5.63	16.80	11.12	5.92	16.33	11.05	6.03
CucumberRaita	13.58	7.92	5.14	16.15	9.97	6.09	10.44	7.17	4.17	15.85	8.82	5.53
DressedUpMeatballs	15.20	10.80	6.05	17.59	10.27	5.81	16.77	9.71	5.64	17.42	11.62	6.80
HerbOmelet	14.66	14.98	5.50	14.64	11.34	6.29	14.40	10.52	5.21	16.01	15.80	6.27
MicrowaveEggSandwich	16.25	10.44	6.16	19.16	11.29	6.99	15.93	11.00	6.34	21.43	12.31	7.37
MicrowaveFrenchToast	16.82	7.90	5.07	17.31	8.82	5.66	15.91	8.84	5.65	14.06	8.49	5.24
MicrowaveMugPizza	12.82	9.78	5.27	12.69	9.18	5.18	15.65	9.43	5.55	14.50	9.94	5.54
MugCake	16.12	12.95	6.87	10.32	8.85	4.40	10.86	9.63	4.83	13.69	10.80	5.70
PanFriedTofu	8.86	10.39	3.75	9.34	12.44	3.87	9.59	10.20	3.59	8.98	10.58	3.68
Pinwheels	13.58	11.96	5.92	16.08	13.06	7.05	11.72	10.22	4.76	17.07	11.89	6.32
Ramen	11.09	9.97	4.48	12.90	10.92	5.07	12.59	10.77	4.66	12.23	9.74	4.38
SautedMushrooms	15.06	12.22	6.16	19.54	13.83	7.42	19.08	11.61	6.25	19.54	13.82	7.42
ScrambledEggs	11.11	11.08	5.27	11.70	10.96	5.27	14.91	14.17	7.07	15.61	13.82	6.43
SpicedHotChocolate	29.82	10.58	8.49	29.79	11.04	8.74	28.57	13.61	9.95	29.56	9.79	7.87
SpicyTunaAvocadoWraps	15.62	10.52	5.67	12.47	9.61	5.25	18.95	10.18	5.55	14.81	10.38	5.58
TomatoChutney	12.25	10.68	5.42	12.25	10.68	5.42	14.69	11.44	5.71	13.64	10.17	5.23
TomatoMozzarellaSalad	19.77	10.21	6.01	19.20	10.48	5.96	16.01	9.22	5.26	20.80	9.01	5.43
Zoodles	18.32	12.80	6.37	18.32	12.80	6.37	19.60	15.70	7.34	15.54	13.48	5.41
Average	**15.62**	**10.85**	**5.78**	**15.78**	**10.68**	**5.82**	**15.52**	**10.63**	**5.71**	**16.15**	**10.99**	**5.83**

Table 5. Procedure Learning using CNC framework by (Bansal, Siddhant et al., 2022)

supervised Correspond and Cut (CnC) framework for procedure learning. We trained an embedder network using an A-40 GPU and it took us 3 hours to complete the training process. In Table 5.1, We present results on data corresponding to three recipes sampled from our dataset. From the obtained evaluation metrics of precision, recall and IoU scores we did observe a significant drop in performance compared to the results observed for all the other datasets considered for evaluation in the paper.

6. Discussion

6.1. Summary

In this paper, we have introduced a large ego-centric dataset for procedural activities. Our dataset consists of synchronized egocentric views, audio and depth information specifically designed for tasks such as Temporal Action Segmentation, 3D activity analysis, Procedure Learning, Error Recognition, Error Anticipation, and more. Additionally, we have provided benchmarks for Error Detection and Procedure Learning. Although existing methods have shown encouraging results, they still fail to effectively address these challenges with high precision, as evident from the oracle experiments. This indicates the need for further exploration and future research in this domain.

6.2. Limitations & Future Work

This research opens up several promising avenues for further exploration and expansion in the field of error recognition and activity analysis. First, an exciting direction for future work is the extension of the dataset to include activities from other domains. By incorporating tasks such as performing chemical experiments or executing hardware-related activities (e.g., working with cars or computer parts), the dataset can encompass a wider range of activities and provide insights into error patterns in diverse real-world scenarios. Second, the dataset can be used to compare and develop methods for solving various tasks such as transfer learning (e.g., training on Ego4D, Kinetics 101 and EPIC-Kitchens and testing on our dataset), semantic role labeling, video question answering, long video understanding, procedure planning, improving task performance, reducing errors, etc.

References

In the Eye of Beholder Joint Learning of Gaze and Actions in First Person Video - ECCV-2018_12260412, August 2018. URL https://openaccess.thecvf.com/content_ECCV_2018/papers/Yin_Li_In_the_Eye_ECCV_2018_paper.pdf. [Online; accessed 7. Jun. 2023].

Bansal, Siddhant, Arora, Chetan, and Jawahar, C. V. My View is the Best View: Procedure Learning from Egocentric Videos. *European Conference on Computer Vision*, July 2022. doi: 10.48550/arxiv.2207.10883. ARXIV_ID: 2207.10883 MAG ID: 4288043312 S2ID: f2053238548ceb5d2a1835341594349798506 8de.

Bloom, V., Makris, D., and Argyriou, V. G3D: A gaming action dataset and real time action recognition evaluation framework. In *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pp. 7–12. IEEE, June 2012. doi: 10.1109/CVPRW.2012.6239175.

Chevignard, M. P., Catroppa, C., Galvin, J., and Anderson, V. Development and evaluation of an ecological task to assess executive functioning post childhood tbi: The children's cooking task. *Brain Impairment*, 11(2):125–143, 2010. doi: 10.1375/brim.11.2.125.

Damen, D., Doughty, H., Farinella, G. M., Fidler, S., Furnari, A., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., Will Price, Price, W., and Wray, M. The EPIC-KITCHENS Dataset: Collection, Challenges and Baselines. *arXiv: Computer Vision and Pattern Recognition*, April 2020. doi: 10.1109/tpami.2020.2991965. ARXIV_ID: 2005.00343 MAG ID: 3022491006 S2ID: 1badccbe4a3cbf8662b924a97bbeea14fe2f1ac7.

De la Torre, F., Hodgins, J. K., Bargteil, A. W., Martin, X., Macey, J. R., Collado, A. T., and Beltran, P. Guide to the carnegie mellon university multimodal activity (cmu-mmac) database. In *Tech. report CMU-RI-TR-08-22, Robotics Institute, Carnegie Mellon University*, April 2008.

Dibene, Juan C. and Dunn, Enrique. HoloLens 2 Sensor Streaming. *Cornell University - arXiv*, November 2022. doi: 10.48550/arxiv.2211.02648. ARXIV_ID: 2211.02648 MAG ID: 4308505718 S2ID: b19229b4f8667dae5017cae4df5c37086332da17.

Draper, B. DARPA's Perceptually-enabled Task Guidance (PTG) program, 2021. URL https://www.darpa.mil/program/perceptually-enabled-task-guidance.

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. *Computer Vision and Pattern Recognition*, 2022. doi: 10.1109/cvpr52688.2022.02042. ARXIV_ID: 2203.14712 S2ID: 8699794561b74e461fa86e1a9dcd5de74d6d7f6d.

Fathi, A., Ren, X., and Rehg, J. M. Learning to recognize objects in egocentric activities. In *CVPR 2011*, pp. 3281–3288. IEEE, June 2011. doi: 10.1109/CVPR.2011.5995444.

Finnanger, T. G., Andersson, S., Chevignard, M., Johansen, G. O., Brandt, A. E., Hypher, R. E., Risnes, K., Rø, T. B., and Stubberud, J. Assessment of executive function in everyday life-psychometric properties of the norwegian adaptation of the children's cooking task. *Frontiers in human neuroscience*, 15:761755, 2021. ISSN 1662-5161. doi: 10.3389/fnhum.2021.761755. URL https://doi.org/10.3389/fnhum.2021.761755.

Fogel, Y., Rosenblum, S., Hirsh, R., Chevignard, M., and Josman, N. Daily performance of adolescents with executive function deficits: An empirical study using a complex-cooking task. *Occupational therapy international*, 2020:3051809, 2020. ISSN 0966-7903. doi: 10.1155/2020/3051809. URL https://doi.org/10.1155/2020/3051809.

Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., Hamburger, J., Jiang, H., Liu, M., Liu, X., Martin, M., Nagarajan, T., Radosavovic, I., Ramakrishnan, S. K., Ryan, F., Sharma, J., Wray, M., Xu, M., Xu, E. Z., Zhao, C., Bansal, S., Batra, D., Cartillier, V., Crane, S., Do, T., Doulaty, M., Erapalli, A., Feichtenhofer, C., Fragomeni, A., Fu, Q., Gebreselasie, A., Gonzalez, C., Hillis, J., Huang, X., Huang, Y., Jia, W., Khoo, W., Kolar, J., Kottur, S., Kumar, A., Landini, F., Li, C., Li, Y., Li, Z., Mangalam, K., Modhugu, R., Munro, J., Murrell, T., Nishiyasu, T., Price, W., Puentes, P. R., Ramazanova, M., Sari, L., Somasundaram, K., Southerland, A., Sugano, Y., Tao, R., Vo, M., Wang, Y., Wu, X., Yagi, T., Zhao, Z., Zhu, Y., Arbelaez, P., Crandall, D., Damen, D., Farinella, G. M., Fuegen, C., Ghanem, B., Ithapu, V. K., Jawahar, C. V., Joo, H., Kitani, K., Li, H., Newcombe, R., Oliva, A., Park, H. S., Rehg, J. M., Sato, Y., Shi, J., Shou, M. Z., Torralba, A., Torresani, L., Yan, M., and Malik, J. Ego4D: Around the World in 3,000 Hours of Egocentric Video. *arXiv*, October 2021. doi: 10.48550/arXiv.2110.07058.

He, K., Zhang, X., Ren, S., and Sun, J. Deep Residual Learning for Image Recognition. *arXiv*, December 2015. doi: 10.48550/arXiv.1512.03385.

Jang, Y., Sullivan, B., Ludwig, C., Gilchrist, I., Damen, D., and Mayol-Cuevas, W. Epic-tent: An egocentric video dataset for camping tent assembly. In *Proceedings of the*

IEEE/CVF International Conference on Computer Vision (ICCV) Workshops, Oct 2019.

Kuehne, H., Arslan, A., and Serre, T. The language of actions: Recovering the syntax and semantics of goal-directed human activities. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 780–787, 2014.

Liu, C., Hu, Y., Li, Y., Song, S., and Liu, J. Pku-mmd: A large scale benchmark for continuous multi-modal human action understanding. *arXiv preprint arXiv: 1703.07475*, 2017.

Madan, N., Ristea, N.-C., Ionescu, R. T., Nasrollahi, K., Khan, F. S., Moeslund, T. B., and Shah, M. Self-Supervised Masked Convolutional Transformer Block for Anomaly Detection, September 2022. URL http://arxiv.org/abs/2209.12148. arXiv:2209.12148 [cs].

Ni, B., Wang, G., and Moulin, P. *RGBD-HuDaAct: A Color-Depth Video Database for Human Daily Activity Recognition*, volume 47. November 2011. ISBN 978-1-4471-4639-1. doi: 10.1109/ICCVW.2011.6130379.

Ragusa, F., Furnari, A., Salvatore Livatino, Livatino, S., and Farinella, G. M. The MECCANO Dataset: Understanding Human-Object Interactions from Egocentric Videos in an Industrial-like Domain. *arXiv: Computer Vision and Pattern Recognition*, 2020. doi: 10.1109/wacv48630.2021.00161. ARXIV_ID: 2010.05654 MAG ID: 3092157723 S2ID: a58a0732664b97b471b795df5812f98f24840490.

Ristea, N.-C., Madan, N., Ionescu, R. T., Nasrollahi, K., Khan, F. S., Moeslund, T. B., and Shah, M. Self-Supervised Predictive Convolutional Attentive Block for Anomaly Detection, March 2022. URL http://arxiv.org/abs/2111.09099. arXiv:2111.09099 [cs].

Rohrbach, M., Rohrbach, A., Regneri, M., Amin, S., Andriluka, M., Pinkal, M., and Schiele, B. Recognizing fine-grained and composite activities using hand-centric features and script data. *International Journal of Computer Vision*, pp. 1–28, 2015. ISSN 0920-5691. doi: 10.1007/s11263-015-0851-8. URL http://dx.doi.org/10.1007/s11263-015-0851-8.

Shahroudy, A., Liu, J., Ng, T.-T., and Wang, G. NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis. *arXiv:1604.02808 [cs]*, April 2016. URL http://arxiv.org/abs/1604.02808. arXiv: 1604.02808.

Soran, B., Farhadi, A., and Shapiro, L. Generating notifications for missing actions: Don't forget to turn the lights off! In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ICCV '15, pp. 4669–4677, USA, 2015. IEEE Computer Society. ISBN 9781467383912. doi: 10.1109/ICCV.2015.530. URL https://doi.org/10.1109/ICCV.2015.530.

Stein, S. and McKenna, S. J. Combining embedded accelerometers with computer vision for recognizing food preparation activities. In *UbiComp '13: Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*, pp. 729–738. Association for Computing Machinery, New York, NY, USA, September 2013. ISBN 978-1-45031770-2. doi: 10.1145/2493432.2493482.

Tang, Y., Dajun Ding, Dajun Ding, Ding, D., Rao, Y., Zheng, Y., Zhang, D., Zhao, L., Lu, J., and Zhou, J. COIN: A Large-Scale Dataset for Comprehensive Instructional Video Analysis. *Computer Vision and Pattern Recognition*, pp. 1207–1216, June 2019. doi: 10.1109/cvpr.2019.00130. ARXIV_ID: 1903.02874 MAG ID: 2964094654 S2ID: e27e78c33288728f66f7dab2fe2696ddbc5c1026.

Wang, J., Liu, Z., Wu, Y., and Yuan, J. Mining actionlet ensemble for action recognition with depth cameras. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1290–1297. IEEE, June 2012. doi: 10.1109/CVPR.2012.6247813.

Zhou, L., Xu, C., and Corso, J. J. Towards Automatic Learning of Procedures from Web Instructional Videos. *arXiv*, March 2017. doi: 10.48550/arXiv.1703.09788.

Zhukov, D., Alayrac, J.-B., Cinbis, R. G., Fouhey, D. F., Laptev, I., and Sivic, J. Cross-task weakly supervised learning from instructional videos. *arXiv: Computer Vision and Pattern Recognition*, March 2019. doi: 10.1109/cvpr.2019.00365. ARXIV_ID: 1903.08225 MAG ID: 2923016229 S2ID: 3605e41ce77dfd259eaebed906804bb60f634f75.