# Supplementary Materials:
# Error Recognition in Procedural Videos using Generalized Task Graph

Shih-Po Lee
Northeastern University
lee.shih@northeastern.edu

Ehsan Elhamifar
Northeastern University
e.elhamifar@northeastern.edu

| Granularity | Normal | Error | Modification | Slip | Correction | Addition |
|---|---|---|---|---|---|---|
| | Quesadilla (%) | | Quesadilla - Error Only (%) | | | |
| Frame-wise | 86.8 | 13.2 | 25.7 | 21.5 | 20.7 | 32.1 |
| Segment-wise | 90.2 | 9.8 | 33.3 | 22.2 | 22.2 | 22.2 |
| | Oatmeal (%) | | Oatmeal - Error Only (%) | | | |
| Frame-wise | 95.0 | 5.0 | 38.3 | 13.4 | 13.8 | 34.5 |
| Segment-wise | 95.0 | 5.0 | 34.8 | 18.0 | 18.0 | 29.2 |
| | Pinwheel (%) | | Pinwheel - Error Only (%) | | | |
| Frame-wise | 88.1 | 11.9 | 64.0 | 7.1 | 14.2 | 14.7 |
| Segment-wise | 89.9 | 10.1 | 52.2 | 15.8 | 15.8 | 16.3 |
| | Coffee (%) | | Coffee - Error Only (%) | | | |
| Frame-wise | 96.9 | 3.1 | 19.1 | 69.8 | 11.1 | 0.0 |
| Segment-wise | 96.1 | 3.9 | 20.8 | 64.9 | 14.3 | 0.0 |
| | Tea (%) | | Tea - Error Only (%) | | | |
| Frame-wise | 87.7 | 12.3 | 30.3 | 23.6 | 22.6 | 23.5 |
| Segment-wise | 92.4 | 7.6 | 33.3 | 22.2 | 22.2 | 22.2 |

Table S1. Breakdown analysis of the EgoPER dataset

| Granularity | Normal | Error | Prep. | Mea. | Time | Tec. | Temp. | Other |
|---|---|---|---|---|---|---|---|---|
| | Hot Chocolate (%) | | Hot Chocolate - Error Only (%) | | | | | |
| Frame-wise | 66.6 | 33.4 | 2.5 | 25.1 | 40.8 | 20.4 | 11.3 | 0.0 |
| Segment-wise | 80.3 | 19.7 | 5.6 | 36.1 | 22.2 | 30.6 | 5.6 | 0.0 |
| | Sandwich (%) | | Sandwich - Error Only (%) | | | | | |
| Frame-wise | 70.4 | 29.6 | 13.6 | 33.4 | 27.6 | 18.4 | 6.4 | 0.5 |
| Segment-wise | 84.5 | 15.5 | 20.4 | 28.6 | 26.5 | 16.3 | 4.1 | 4.1 |
| | Burritos (%) | | Burritos - Error Only (%) | | | | | |
| Frame-wise | 73.5 | 26.5 | 35.2 | 16.0 | 30.0 | 18.8 | 0.0 | 0.0 |
| Segment-wise | 84.8 | 15.2 | 43.6 | 17.9 | 12.8 | 25.6 | 0.0 | 0.0 |
| | Ramen (%) | | Ramen - Error Only (%) | | | | | |
| Frame-wise | 84.8 | 15.2 | 6.5 | 17.0 | 52.3 | 24.2 | 0.0 | 0.0 |
| Segment-wise | 90.8 | 9.2 | 14.3 | 11.4 | 28.6 | 45.7 | 0.0 | 0.0 |
| | Raita (%) | | Raita - Error Only (%) | | | | | |
| Frame-wise | 83.0 | 17.0 | 41.7 | 28.6 | 0.0 | 29.7 | 0.0 | 0.0 |
| Segment-wise | 87.4 | 12.6 | 38.3 | 25.5 | 0.0 | 36.2 | 0.0 | 0.0 |

Table S2. Breakdown analysis of the CaptainCook4D dataset

## S.1. Prompts for LLMs and Error Descriptions

In this section, we provide the details of prompts for generating different types of error descriptions using LLMs. Each prompt consists of 1) step description, 2) error type definition,and 3) the number of errors to produce. First, we show the prompts for EgoPED [2] as follows:

- **Modification Errors:** Imagine you are doing {task name}. Your next action is {step description}. Describe 10 possible modification mistake you can make. Modification mistake is an error that you use a wrong way or tool to execute the action. The outcome of the action remains the same. Use one sentence to describe each modification mistake.
- **Slip and Slip-Correction Errors:** Imagine you are doing {task name}. Your next action is {step description}. Describe 10 possible accidental mistakes you can make and their corresponding correction actions. Accidental mistake is an error that you make a mistake in the action and need to recover in order to proceed. Use one sentence to describe accidental mistake and use another one to describe correction action. ”Forget” is not type of accidental mistake.
- **Addition Errors:** Imagine you are doing {task name}. Here is the recipe of making {task name}: {recipe}. Describe 20 extra steps you can do. Extra step is a step that is not in the recipe. Extra step is related to the task but will not change the final outcome of the task. Extra step is visually observable. Extra step is visually dissimilar to the steps in the recipe.

where {task name} the is name of the task, {step description} is the step description in the recipe, and {recipe} lists all the steps in the task.

Next, we show the prompts for CaptainCook4D [4] as follows:

- **Preparation Errors:** Imagine you are doing {task name}. Your next action is {step description}. Describe 10 preparation errors of the action. Preparation error means when users use soiled/wrong ingredients or use different tools. Use one sentence to describe each error.
- **Measurement Errors:** Imagine you are doing {task name}. Your next action is {step description}. Describe 2 measurement errors for the action. Measurement error means when users use wrongly measured ingredients. Use one sentence to describe each error.
- **Timing Errors:** Imagine you are doing {task name}. Your next action is {step description}. Describe 2 timing errors for the action. Timing error means when users perform a step in shorter or longer duration than what is prescribed. Use one sentence to describe each error.
- **Technique Errors:** Imagine you are doing {task name}.

| | | | | EgoPER | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Quesadilla | | Oatmeal | | Pinwheel | | Coffee | | Tea | |
| # training Vid. | # test Vid. | # training Vid. | # test Vid. | # training Vid. | # test Vid. | # training Vid. | # test Vid. | # training Vid. | # test Vid. |
| 26 | 35 | 26 | 35 | 26 | 45 | 27 | 38 | 26 | 35 |
| | | | | CaptainCook4D | | | | | |
| Hot Chocolate | | Sandwich | | Burritos | | Ramen | | Raita | |
| # training Vid. | # test Vid. | # training Vid. | # test Vid. | # training Vid. | # test Vid. | # training Vid. | # test Vid. | # training Vid. | # test Vid. |
| 6 | 9 | 5 | 9 | 5 | 7 | 6 | 7 | 10 | 8 |

Table S3. Training and testing split for EgoPER and CaptainCook4D.

| Method | Normal Segments | | | Error Segments | | |
|---|---|---|---|---|---|---|
| | Top-1 | Top-2 | Top-3 | Top-1 | Top-2 | Top-3 |
| | | | | Quesadilla | | |
| Uniform | 12.5 | 25.0 | 37.5 | 25.0 | 50.0 | 75.0 |
| Cos. Sim. | 54.0 | 68.8 | 73.5 | 42.6 | 65.8 | 100.0 |
| Ours | - | - | - | 54.6 | 81.5 | 100.0 |
| | | | | Oatmeal | | |
| Uniform | 6.7 | 13.4 | 20.0 | 25.0 | 50.0 | 75.0 |
| Cos. Sim. | 30.9 | 60.0 | 73.7 | 70.0 | 87.6 | 100.0 |
| Ours | - | - | - | 50.6 | 62.9 | 100.0 |
| | | | | Pinwheel | | |
| Uniform | 7.7 | 15.4 | 23.1 | 25.0 | 50.0 | 75.0 |
| Cos. Sim. | 31.1 | 53.9 | 61.8 | 27.7 | 48.4 | 90.2 |
| Ours | - | - | - | 38.4 | 60.3 | 90.2 |
| | | | | Coffee | | |
| Uniform | 6.7 | 13.4 | 20.0 | 25.0 | 50.0 | 75.0 |
| Cos. Sim. | 26.5 | 49.4 | 61.6 | 48.0 | 72.7 | 100.0 |
| Ours | - | - | - | 51.9 | 90.0 | 100.0 |
| | | | | Tea | | |
| Uniform | 10.0 | 20.0 | 30.0 | 25.0 | 50.0 | 75.0 |
| Cos. Sim. | 41.5 | 59.2 | 76.3 | 27.8 | 45.3 | 85.2 |
| Ours | - | - | - | 34.3 | 53.7 | 85.1 |

Table S4. The performance of different scoring methods on normal and error segments of EgoPER. Notice the our method only works on error segments.

Your next action is {step description}. Describe 10 technique errors of the action. Technique error means when users perform the required action incorrectly, leading to a wrong outcome than expected. Technique error is not related to wrong time, wrong temperature, wrong measurement, or skipping steps. Use one sentence to describe each error.

- **Temperature Errors:** Imagine you are doing {task name}. Your next action is {step description}. Describe 2 temperature errors for each of the step. Temperature error means when users set higher/lower power levels in the microwave or on a stove than what is prescribed. Use one sentence to describe each error.

On the other hand, Table S9 and S10 show the examples of generated error descriptions using LLMs for EgoPER. Notice that the addition errors are irrelevant to the steps. Table S11 and S12 show the examples of generated error descriptions using LLMs for CaptainCook4D.

## S.2. Analysis for VLM features

Table S4 shows the Top-1, Top-2, and Top-3 accuracy with different scoring functions. We compute each accuracy for normal segments based on the average features of frames in normal segments and textual features of steps. On the other hand, we follow the similar way to compute each accuracy for error segments, but use textual features of error descriptions instead. From Table S4, we observe that 1) cosine similarity (Cos. Sims) achieves higher Top-1 accuracy than Uniform on both normal and error segments, indicating the pre-extracted visual and textual features from VLMs are aligned, and 2) the textual features of normal steps are highly correlated to normal segments. Therefore, combining the normal features into our scoring function makes our error features toward error segments.

## S.3. Dataset Analysis

In the section, we highlight the challenges for *ER* in terms of existing datasets.

Table S1 and S2 show the breakdown analysis of EgoPER and CaptainCook4D. For EgoPER, the data imbalance issue is serious as most of the frames and segments (around 90%) are normal. Specifically, for *coffee*, only 3.9% of the segments are errors and around 65% of them are slip errors, making it a hard task for *ER*. For CaptainCook4D, the percentage of error frames is higher than error segments for all tasks. It means that the annotations are not fine-grained. Assume there is a 10 seconds step segment. An error happens in the middle of a step for 2 seconds. The annotator annotates the entire step segment as an error and therefore, the annotations do not provide the precise location of errors, making *ER* becomes even more challenging.

Table S3 shows the training and test split for EgoPER and CaptainCook4D. For EgoPER, every task has at least 26 normal training videos so that the model can have decent *TAS* performance on each action, including background. However, for CaptainCook4D, some tasks only contain 5 normal training videos (e.g., *sandwich* and *burritos*), making *ER* challenging as it is hard to train a good action segmentation model to segment videos into steps, especially background. Still, EgoPER and CaptainCook4D are the only existing datasets that fit our setting with various types of errors and clear definition for them.

## S.4. Detailed Analysis for Error Recognition

| Method | Modification | | | Slip | | | Correction | | | Addition | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | P | R | F1 | P | R | F1 | P | R | F1 | P | R | F1 |
| | | | | | | Quesadilla | | | | | | |
| Naive Predictor | 100.0 | 71.4 | 83.3 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| EgPED + MV | 5.8 | 11.5 | 7.7 | 11.9 | 82.6 | 20.8 | 14.3 | 8.3 | 10.5 | 0.0 | 0.0 | 0.0 |
| GTG2Vid + MV | 8.2 | 73.5 | 14.7 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 10.5 | 70.3 | 18.2 |
| GTG2Vid + ERM w/o n.f. | 31.2 | 51.3 | 38.8 | 28.6 | 8.3 | 12.9 | 0.0 | 0.0. | 0.0 | 48.3 | 42.4 | 45.2 |
| GTG2Vid + ERM (Ours) | 19.4 | 17.6 | 18.5 | 23.3 | 28.0 | 25.5 | 25.9 | 29.2 | 27.5 | 35.3 | 61.5 | 44.9 |
| | | | | | | Oatmeal | | | | | | |
| Naive Predictor | 100.0 | 83.3 | 90.9 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| EgPED + MV | 7.3 | 50.0 | 12.8 | 6.1 | 30.0 | 10.2 | 5.9 | 18.8 | 9.0 | 0.0 | 0.0 | 0.0 |
| GTG2Vid + MV | 3.4 | 96.6 | 6.5 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 5.4 | 92.9 | 10.2 |
| GTG2Vid + ERM w/o n.f. | 18.0 | 57.9 | 27.5 | 9.1 | 6.2 | 7.4 | 0.0 | 0.0 | 0.0 | 48.1 | 68.4 | 56.5 |
| GTG2Vid + ERM (Ours) | 22.7 | 51.5 | 31.5 | 3.8 | 11.8 | 5.7 | 15.1 | 68.8 | 24.7 | 29.8 | 90.7 | 44.8 |
| | | | | | | Pinwheel | | | | | | |
| Naive Predictor | 100.0 | 53.6 | 69.8 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| EgPED + MV | 16.7 | 31.7 | 21.9 | 1.5 | 11.8 | 2.6 | 10.6 | 69.7 | 18.5 | 0.0 | 0.0 | 0.0 |
| GTG2Vid + MV | 7.5 | 93.2 | 13.8 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 6.9 | 86.3 | 12.8 |
| GTG2Vid + ERM w/o n.f. | 19.8 | 64.1 | 30.2 | 14.6 | 25.9 | 18.7 | 24.1 | 24.1 | 24.1 | 27.1 | 55.8 | 36.5 |
| GTG2Vid + ERM (Ours) | 12.6 | 25.0 | 16.7 | 7.3 | 44.8 | 12.5 | 1.1 | 3.4 | 1.7 | 18.4 | 73.2 | 29.4 |
| | | | | | | Coffee | | | | | | |
| Naive Predictor | 100.0 | 92.7 | 96.2 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | - | - | - |
| EgPED + MV | 3.2 | 35.3 | 5.8 | 9.5 | 50.0 | 15.9 | 4.5 | 15.4 | 7.0 | - | - | - |
| GTG2Vid + MV | 1.0 | 43.8 | 1.9 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | - | - | - |
| GTG2Vid + ERM w/o n.f. | 4.2 | 28.6 | 7.3 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | - | - | - |
| GTG2Vid + ERM (Ours) | 1.7 | 11.8 | 3.0 | 3.9 | 6.0 | 4.7 | 3.1 | 30.8 | 5.7 | - | - | - |
| | | | | | | Tea | | | | | | |
| Naive Predictor | 100.0 | 74.5 | 85.4 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| EgPED + MV | 6.7 | 17.6 | 9.7 | 6.1 | 37.5 | 10.5 | 9.1 | 4.2 | 5.7 | 0.0 | 0.0 | 0.0 |
| GTG2Vid + MV | 5.7 | 100.0 | 10.9 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 7.5 | 92.9 | 13.9 |
| GTG2Vid + ERM w/o n.f. | 18.5 | 66.7 | 28.9 | 9.1 | 4.2 | 5.7 | 0.0 | 0.0 | 0.0 | 42.9 | 57.1 | 49.0 |
| GTG2Vid + ERM (Ours) | 21.6 | 31.4 | 25.6 | 3.1 | 16.7 | 5.3 | 6.7 | 20.0 | 10.0 | 26.2 | 90.5 | 40.6 |

Table S5. Breakdown analysis of performance on each error type in EgoPER. P, R, and F1 indicate P@0, R@0, and F1@0, respectively. - denotes there is no such error in the test videso.

In this section, we show the breakdown performance and the performance of F1@.1 for *ER*.

Table S5 shows the precision, recall, and F1 for different methods on EgoPER. Our method can recognize every error type across all tasks while EgPED + MV cannot recognize addition errors. This is because the addition errors are too diverse to be generated by LLMs and therefore, the similarities between error frames and generated error descriptions are low. In contrast, our proposed method classifies error frames that are assigned to background (mentioned in the section of implementation) to addition errors. For CaptainCook4D in Table S6, our method also demonstrates a stronger recogntion ability compared to EgPED + MV on *Hot Chocolate*, *Sandwhich*, and *Raita*, as we can recognize all types of errors.

On the other hand, we use a stricter metric, F1@.1, to 1) compare the localization ability of different methods and 2) increase the sensitivity to oversegmentation on error types. In Table S7, our proposed method outperforms both Naive Predictor and EgoPED + MV. We achieves 13.0%, 8.8%, 7.8%, and 9.9% on w-F1@.1 score compared to 2.6%, 4.1%, 3.5%, and 2.3% by EgoPED + MV. Furthermore, our method can still recognize more types of errors than EgoPED + MV.

## S.5. Ablation Studies

In this section, we investigate the performance of GTG2Vid for *ER* with different LLMs and the analysis of runtime and memory usage.

**Effectiveness of Different LLMs.** Table S8 shows the performance of our method with different LLMs. We follow the same generation process to generate error descriptions using each LLM. Our method is robust to error descriptions from different LLMs, specifically achieving 21.5%, 17.6%, and 20.8% on w-F1@0 with GPT4o mini [3], Qwen2.5-14B [5], Llama-3-8B [1].

**Runtime and Memory Analysis.** Figure 1 shows the analysis of runtime and memory usage in terms of graph sizes and video lengths. The offline processing speed of our method is real-time. Specifically, our method can achieve

|  | Prep. |  |  | Mea. |  |  | Time |  |  | Tec. |  |  | Temp. |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Method | P | R | F1 | P | R | F1 | P | R | F1 | P | R | F1 | P | R | F1 |
|  |  |  |  |  |  |  | Hot Chocolate |  |  |  |  |  |  |  |  |
| Naive Predictor | 100.0 | 100.0 | 100.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| EgPED + MV | 0.0 | 0.0 | 0.0 | 6.2 | 10.0 | 7.7 | 33.3 | 25.0 | 28.6 | 15.4 | 25.0 | 19.0 | 16.7 | 33.3 | 22.2 |
| GTG2Vid + MV | 6.5 | 100.0 | 12.1 | 18.2 | 20.0 | 19.0 | 33.3 | 12.5 | 18.2 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| GTG2Vid + ERM w/o n.f. | 7.3 | 100.0 | 13.6 | 0.0 | 0.0 | 0.0 | 20.0 | 25.0 | 22.2 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| GTG2Vid + ERM (Ours) | 3.7 | 75.0 | 7.1 | 6.0 | 33.3 | 10.1 | 10.9 | 62.5 | 18.5 | 10.7 | 37.5 | 16.7 | 2.9 | 100.0 | 5.6 |
|  |  |  |  |  |  |  | Sandwich |  |  |  |  |  |  |  |  |
| Naive Predictor | 100.0 | 81.8 | 90.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| EgPED + MV | 12.5 | 28.6 | 17.4 | 0.0 | 0.0 | 0.0 | 20.0 | 7.7 | 11.1 | 5.0 | 12.5 | 7.1 | 6.9 | 66.7 | 12.5 |
| GTG2Vid + MV | 12.6 | 100.0 | 22.4 | 25.0 | 23.1 | 24.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| GTG2Vid + ERM w/o n.f. | 14.4 | 86.7 | 24.8 | 30.8 | 30.8 | 30.8 | 7.1 | 7.7 | 7.4 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| GTG2Vid + ERM (Ours) | 6.6 | 66.7 | 11.9 | 5.5 | 50.0 | 9.8 | 4.8 | 28.6 | 8.2 | 3.5 | 33.3 | 6.4 | 2.7 | 100.0 | 5.2 |
|  |  |  |  |  |  |  | Burritos |  |  |  |  |  |  |  |  |
| Naive Predictor | 100.0 | 38.9 | 56.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | - | - | - |
| EgPED + MV | 35.7 | 31.2 | 33.3 | 0.0 | 0.0 | 0.0 | 3.1 | 20.0 | 5.4 | 12.5 | 12.5 | 12.5 | - | - | - |
| GTG2Vid + MV | 20.3 | 60.0 | 30.4 | 6.2 | 14.3 | 8.7 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | - | - | - |
| GTG2Vid + ERM w/o n.f. | 34.8 | 38.1 | 36.4 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | - | - | - |
| GTG2Vid + ERM (Ours) | 21.9 | 35.0 | 26.9 | 13.6 | 42.9 | 20.7 | 5.9 | 16.7 | 8.7 | 0.0 | 0.0 | 0.0 | - | - | - |
|  |  |  |  |  |  |  | Pimiento |  |  |  |  |  |  |  |  |
| Naive Predictor | 100.0 | 56.2 | 72.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| EgPED + MV | 12.5 | 7.1 | 9.1 | 15.4 | 18.2 | 16.7 | 16.7 | 25.0 | 20.0 | 25.0 | 18.8 | 21.4 | 0.0 | 0.0 | 0.0 |
| GTG2Vid + MV | 16.0 | 95.0 | 27.3 | 0.0 | 0.0 | 0.0 | 10.0 | 14.3 | 11.8 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| GTG2Vid + ERM w/o n.f. | 0.0 | 0.0 | 0.0 | 15.7 | 89.5 | 26.8 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| GTG2Vid + ERM (Ours) | 13.6 | 83.3 | 23.4 | 4.2 | 44.4 | 7.6 | 5.6 | 85.7 | 10.5 | 5.3 | 8.3 | 6.5 | 0.0 | 0.0 | 0.0 |
|  |  |  |  |  |  |  | Raita |  |  |  |  |  |  |  |  |
| Naive Predictor | 100.0 | 44.4 | 61.5 | 0.0 | 0.0 | 0.0 | - | - | - | 0.0 | 0.0 | 0.0 | - | - | - |
| EgPED + MV | 0.0 | 0.0 | 0.0 | 21.4 | 27.3 | 24.0 | - | - | - | 50.0 | 6.7 | 11.8 | - | - | - |
| GTG2Vid + MV | 17.8 | 59.1 | 27.4 | 13.3 | 60.0 | 21.8 | - | - | - | 0.0 | 0.0 | 0.0 | - | - | - |
| GTG2Vid + ERM w/o n.f. | 23.4 | 68.2 | 34.9 | 3.4 | 11.1 | 5.3 | - | - | - | 0.0 | 0.0 | 0.0 | - | - | - |
| GTG2Vid + ERM (Ours) | 22.4 | 68.2 | 33.7 | 9.1 | 33.3 | 14.3 | - | - | - | 5.9 | 6.2 | 6.1 | - | - | - |

Table S6. Breakdown analysis of performance on each error type in CaptainCook4D. P, R, and F1 indicate P@0, R@0, and F1@0, respectively. - denotes there is no such error in the test videso.

|  | **EgoPER** |  |  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|---|
|  | Quesadilla |  | Oatmeal |  | Pinwheel |  | Coffee |  | Tea |  |
| Method | w-F1@.1 | EAcc | w-F1@.1 | EAcc | w-F1@.1 | EAcc | w-F1@.1 | EAcc | w-F1@.1 | EAcc |
| Naive Predictor | 3.2 | 25.0 | 0.0 | 0.0 | 6.0 | 25.0 | 0.0 | 0.0 | 2.5 | 25.0 |
| EgoPED + MV | 2.6 | 50.0 | 4.1 | 75.0 | 3.5 | 50.0 | **4.6** | 66.6 | 2.3 | 75.0 |
| GTG2Vid + ERM (Ours) | **13.0** | **100.0** | **8.8** | **75.0** | **7.8** | **100.0** | 2.5 | **100.0** | **9.9** | **100.0** |
|  | **CaptainCook4D** |  |  |  |  |  |  |  |  |  |
|  | Hot Chocolate |  | Sandwich |  | Burritos |  | Ramen |  | Raita |  |
|  | w-F1@.1 | EAcc | w-F1@.1 | EAcc | w-F1@.1 | EAcc | w-F1@.1 | EAcc | w-F1@.1 | EAcc |
| Naive Predictor | 0.9 | 20.0 | 0.0 | 0.0 | 3.3 | 25.0 | 0.0 | 0.0 | 7.3 | 33.3 |
| EgoPED + MV | 1.7 | 40.0 | 0.3 | 20.0 | 2.0 | 25.0 | **3.2** | **75.0** | 3.1 | 66.6 |
| GTG2Vid + ERM (Ours) | **1.8** | **80.0** | **1.5** | **80.0** | **2.3** | **50.0** | 1.2 | 50.0 | **4.3** | **66.6** |

Table S7. *ER* results on EgoPER and CaptainCook4D.

| Method | Quesadilla | Oatmeal | Pinwheel | Coffee | Tea | All |
|---|---|---|---|---|---|---|
| Ours (Qwen2.5-14B) | 27.5 | 14.5 | **19.3** | 3.7 | 22.9 | 17.6 |
| Ours (Llama-3-8B) | 26.3 | 29.1 | 17.8 | **4.8** | **26.2** | 20.8 |
| Ours (GPT4o mini) | **31.7** | **31.3** | 17.8 | 4.5 | 22.1 | **21.5** |

Table S8. *ER* results on w-F1@0 for EgoPER with error descriptions from different LLMs.

around 40 FPS with a memory usage of 2600 MB when processing a 335.3-second video on Intel Xeon Gold 5218 64-Core Processor for *Coffee* with a large GTG.

## S.6. Qualitative Results

In this section, we show more qualitative results of EgoPED [2] and our proposed method. Figure 2, 3, 4, and 5 show the qualitative results for *tea*, *quesadilla*, *pinwheels*, and *oatmeal* in EgoPER.

## References

[1] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, et al. The llama 3 herd of models. *arXiv*, 2024. 3

[2] S. Lee, Z. Lu, Z. Zhang, M. Hoai, and E. Elhamifar. Error detection in egocentric procedural task videos. *IEEE Confer-*

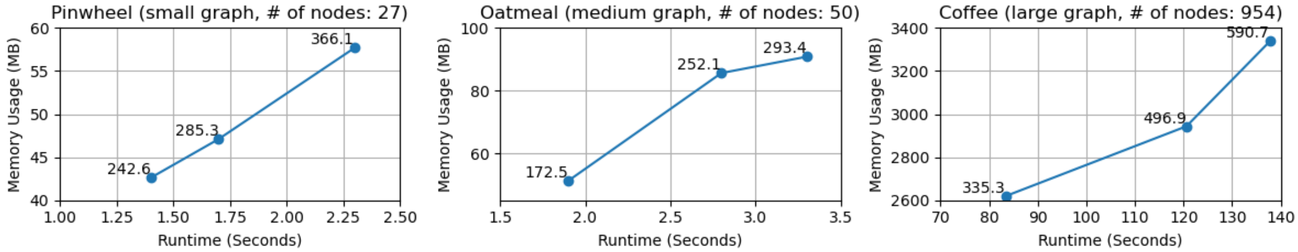

Figure 1. Runtime and memory usage analysis for *Pinwhweel*, *Oatmeal*, and *Coffee* in EgoPER with 27, 50, and 954 nodes. Each point denotes a video with its length in second.

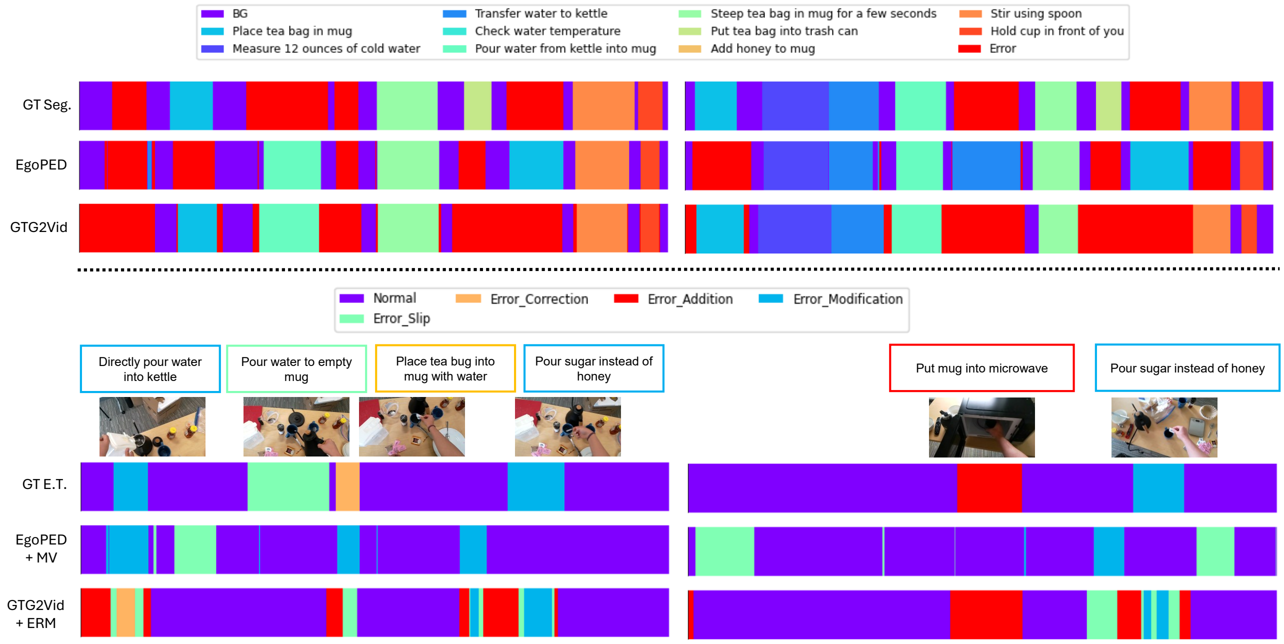

Figure 2. Qualitative error recognition results for *tea* in EgoPER.

*ence on Computer Vision and Pattern Recognition*, 2024. 1, 4

[3] OpenAI. Gpt-4o mini: Advancing cost-efficient intelligence, 2024. Accessed: February 27, 2025. 3

[4] Rohith Peddi, Shivvrat Arya, Bharath Challa, Likhitha Pallapothula, Akshay Vyas, Bhavya Gouripeddi, Jikai Wang, Qifan Zhang, Vasundhara Komaragiri, Eric Ragan, Nicholas Ruozzi, Yu Xiang, and Vibhav Gogate. CaptainCook4D: A Dataset for Understanding Errors in Procedural Activities, 2024. 1

[5] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv*, 2024. 3

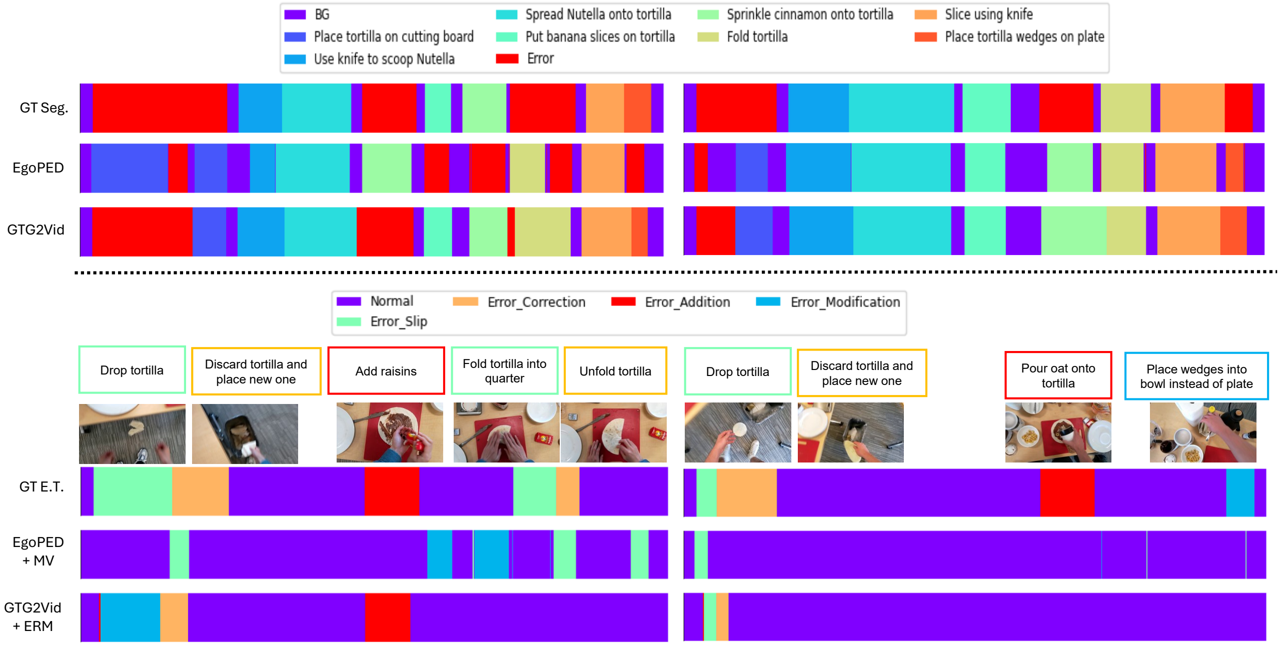

Figure 3. Qualitative error recognition results for *quesadilla* in EgoPER.

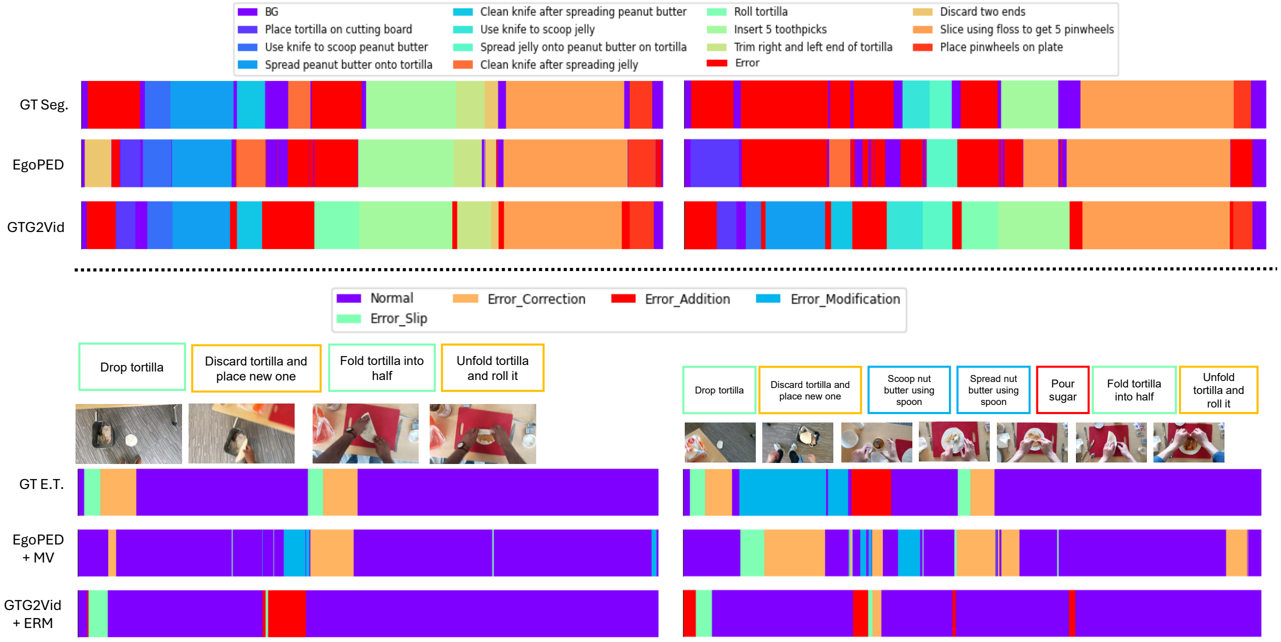

Figure 4. Qualitative error recognition results for *pinwheel* in EgoPER.

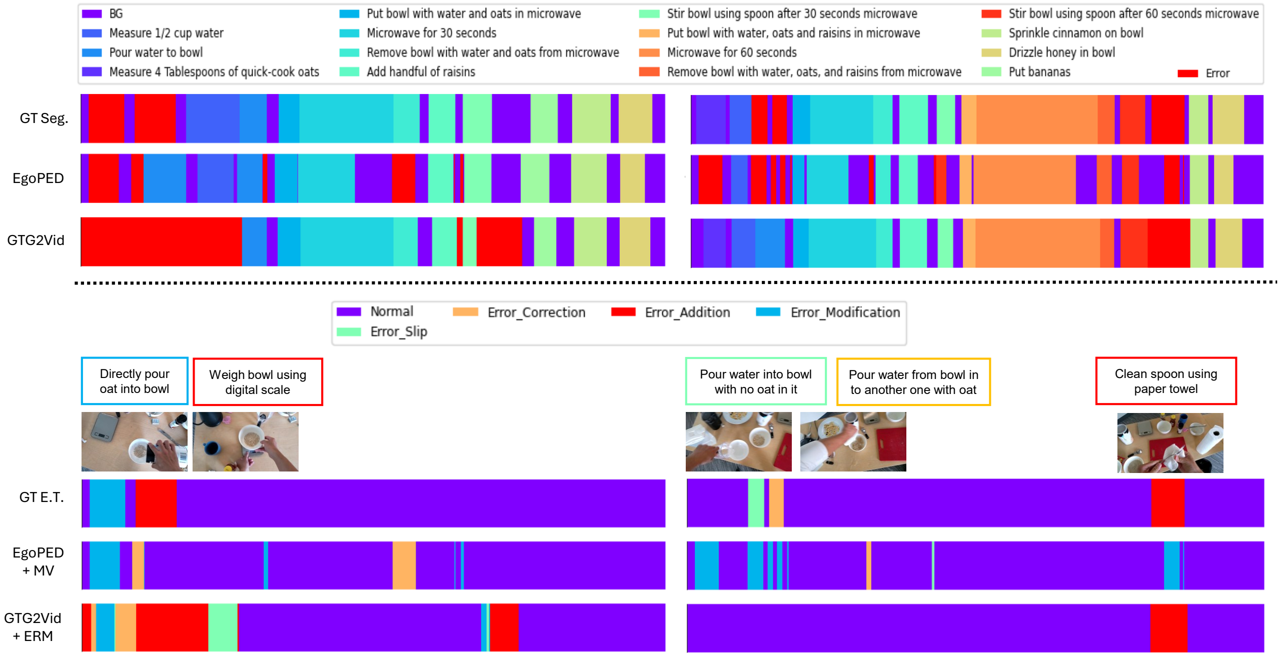

Figure 5. Qualitative error recognition results for *oatmeal* in EgoPER.

| Step | Modification | Slip | Correction | Addition |
|---|---|---|---|---|
| Place tortilla on cutting board | Using tongs instead of hands – You awkwardly struggle to grab the tortilla with tongs, but eventually get it onto the cutting board.<br>Using a spatula – You slide a spatula under the tortilla in the bag, making it harder than necessary but still getting it out. | You rip the tortilla while pulling it from the bag.<br>You accidentally grab two tortillas instead of one. | Carefully unroll the tortilla and patch the tear by gently pressing the edges together before placing it on the cutting board.<br>Separate the tortillas and place one back in the bag before continuing with the action. | Wipe the cutting board before placing the tortilla to ensure a clean surface. |
| Use knife to scoop Nutella | Using the dull side of the knife – You awkwardly try to scoop Nutella with the blunt edge but eventually get some on the knife.<br>Using a serrated knife – You struggle as the Nutella sticks between the serrations, but you still manage to scoop it. | You accidentally dip the knife into the Nutella too deeply and spill some.<br>The knife slips and you scoop more Nutella than needed. | Carefully scrape off the excess Nutella and continue with a smaller amount on the knife.<br>Gently remove the excess Nutella from the knife and place the extra back into the jar. | Use a plate to set aside the banana before slicing, keeping it organized. |
| Spread Nutella onto tortilla | Using the dull side of the knife – You awkwardly push the Nutella around, making spreading harder but still covering the tortilla.<br>Using a serrated knife – The ridges leave uneven streaks, but the Nutella still gets spread. | The Nutella is too thick and doesn't spread easily.<br>You accidentally drop some Nutella off the side of the tortilla. | Warm the Nutella slightly by stirring or microwaving for a few seconds to make it easier to spread.<br>Use the knife to scoop up the spilled Nutella and carefully place it back onto the tortilla. | Position a knife at the edge of the cutting board to be ready for the next steps. |

Table S9. The generated error descriptions for each error type on *quesadilla* of EgoPER. We selectly shows two error descriptions for each error type and show 3 of the steps. Notice that we show 3 error descriptions for addition errors whose descriptions are irrelevant to the steps.

| Step | Modification | Slip | Correction | Addition |
|---|---|---|---|---|
| Measure 4 Tablespoons of quick-cook oats | Using a teaspoon instead of a tablespoon<br>Eyeballing the measurement instead of using a spoon | You accidentally measure 5 tablespoons of oats instead of 4.<br>You spill some oats outside the bowl while pouring them in. | Remove 1 tablespoon of oats from the bowl to correct the measurement.<br>Clean up the spilled oats and add the remaining oats back into the bowl. | Check the microwave settings before starting – Ensure the microwave is set to the correct time and power level before beginning the recipe. |
| Measure 1/2 cup water | Using a liquid measuring cup but reading the measurement from the top instead of eye level<br>Using a dry measuring cup instead of a liquid one | You accidentally overfill the measuring cup, getting more than half a cup of water.<br>You use a different measuring cup size, such as a 1/4 cup instead of a 1/2 cup. | Pour out the excess water until you have exactly half a cup.<br>Measure out two 1/4 cups of water to make a half cup. | Inspect the oats – Take a moment to check the oats to make sure they are not clumped together before starting. |
| Pour water to bowl | Pouring the water too quickly but still getting it all into the bowl<br>Pouring with a shaking hand, causing slight spills outside the bowl | You accidentally spill some water while pouring it into the bowl.<br>You tilt the measuring cup too much, causing the water to pour too quickly. | Clean up the spilled water and pour the remaining water carefully into the bowl.<br>Slow down and tilt the measuring cup more gently to pour the water slowly and steadily. | Smell the oats before adding them to the bowl – Briefly sniff the oats to ensure they have a fresh scent. |

Table S10. The generated error descriptions for each error type on *oatmeal* of EgoPER. We selectly shows two error descriptions for each error type and show 3 of the steps. Notice that we show 3 error descriptions for addition errors whose descriptions are irrelevant to the steps.

| Step | Prep. | Mea. | Time | Tec. | Temp. |
|---|---|---|---|---|---|
| Heat-Heat the contents of the mug for 1 minute and serve | Using expired cocoa powder, which makes the hot chocolate taste stale and bitter.<br>Adding too much sugar, resulting in an overly sweet, cloying drink. | The contents of the mug are heated for 2 minutes instead of 1 minute.<br>The mug is not heated for long enough and remains cool. | Heating the contents for only 30 seconds instead of 1 minute might result in a lukewarm beverage.<br>Heating for 2 minutes might cause the drink to become too hot and potentially scald the milk. | Overheating the hot chocolate by microwaving it for too long, causing it to boil and lose flavor.<br>Stirring the hot chocolate too vigorously, leading to the milk separating or forming bubbles. | Using a higher power setting could overheat the contents, making it too hot to drink immediately.<br>Using a lower power setting could result in a lukewarm drink, requiring extra time to heat up. |
| Add-Add 1/5 teaspoon cinnamon to the mug | Adding too much cinnamon, overwhelming the flavor and making the drink bitter.<br>Using ground nutmeg instead of cinnamon, changing the intended spice profile. | 1/2 teaspoon of cinnamon is added instead of 1/5 teaspoon.<br>No cinnamon is added, leaving the drink without any spiced flavor. | Adding the cinnamon before heating could result in a less aromatic flavor.<br>Adding too much cinnamon (more than 1/5 teaspoon) might overwhelm the flavor of the drink. | Adding cinnamon directly to the hot chocolate without properly mixing it with other dry ingredients, leading to clumps.<br>Pouring the cinnamon too quickly, causing uneven distribution in the mug. | Adding cinnamon to a drink that is too hot could cause the spice to burn and release an unpleasant taste.<br>If the drink is too cold when the cinnamon is added, it might not dissolve properly, leading to clumps. |
| Mix-Mix the contents of the mug | Using a dirty spoon to mix, contaminating the hot chocolate with leftover flavors from previous use.<br>Stirring with a plastic spoon, which could melt or release chemicals when in contact with hot liquid. | The contents are not mixed thoroughly, leaving some ingredients unmixed.<br>The contents are over-mixed, causing the texture to become too frothy. | Mixing too briefly (less than 10 seconds) could leave some cinnamon floating on top instead of fully blending.<br>Overmixing for more than 30 seconds could cause the texture to become too frothy, affecting the overall taste. | Stirring the contents of the mug too quickly, causing hot chocolate to splatter and making a mess.<br>Using a spoon that's too large or too small, making it difficult to properly mix the ingredients. | Mixing when the drink is too hot could cause splattering, burning the skin.<br>Mixing when the drink is too cold could make it harder to properly blend the ingredients, leading to an uneven flavor. |

Table S11. The generated error descriptions for each error type on *hot chocolate* of CaptainCook4D. We selectly shows two error descriptions for each error type and show 3 of the steps.

| Step | Prep. | Mea. | Time | Tec. | Temp. |
|---|---|---|---|---|---|
| Pour-Pour 1 egg into the ramekin cup | You accidentally crack two eggs into the ramekin instead of just one.<br>The ramekin cup you use is not microwave-safe, potentially causing it to crack or overheat. | Using two eggs instead of one can cause the ramekin to overflow.<br>Using only half an egg may result in an insufficient amount for the sandwich. | Pouring the egg too quickly may cause spills and uneven distribution.<br>Pouring too slowly might result in egg sticking to the container before it's fully transferred. | Cracking the Egg Directly into the Ramekin – The user cracks the egg directly into the ramekin without checking for shell fragments, leading to unwanted shell pieces in the mixture.<br>Breaking the Yolk Prematurely – The user accidentally punctures the yolk while pouring, preventing it from cooking with an intact center. | The user microwaves the egg at too high a power, causing the egg to cook too quickly and possibly overcook or spill over.<br>The user sets the microwave to too low a power, resulting in undercooking and a runny egg. |
| Place -Place the egg from the cup over the lettuce | You place the egg directly on wilted or soggy lettuce, making it unappetizing and difficult to eat.<br>You mistakenly use spinach or another leafy green instead of lettuce, changing the flavor and texture. | Using too much lettuce can make the sandwich difficult to close.<br>Using too little lettuce may not provide enough texture and freshness. | Placing the egg too early before it's cooked can make it messy and difficult to handle.<br>Placing the egg too late might cause it to cool down and not blend well with the sandwich. | Dropping the Egg Too Forcefully – The user places the egg too roughly, causing it to break apart or slide off the lettuce.<br>Placing the Egg Off-Center – The user positions the egg unevenly, causing it to hang over the edge and making the sandwich unstable. | The user microwaves the sandwich at too high of a temperature, wilting the lettuce and affecting the texture.<br>The user microwaves on too low of a power setting, leaving the egg cold while the lettuce remains unaffected. |
| Coat -Coat a 6-oz. ramekin cup with cooking spray | You forget to shake the cooking spray can, leading to an uneven coating in the ramekin.<br>You accidentally use olive oil instead of cooking spray, making the egg too greasy. | Spraying too much cooking spray can make the egg greasy.<br>Not using enough spray may cause the egg to stick to the ramekin. | Spraying the ramekin for too short a time may result in the egg sticking to the cup.<br>Spraying too much can create an oily texture in the egg. | Holding the Spray Can Too Close – The user sprays too close to the ramekin, causing an uneven, thick layer that may pool at the bottom.<br>Holding the Spray Can Too Far – The user sprays from too far away, resulting in an uneven coating with gaps where the egg may stick. | The user heats the ramekin before spraying, causing the oil to evaporate and making the coating less effective.<br>The user applies cooking spray while the ramekin is too cold, causing the spray to clump and coat unevenly. |

Table S12. The generated error descriptions for each error type on *sandwich* of CaptainCook4D. We selectly shows two error descriptions for each error type and show 3 of the steps.