Dataset	*Statistics*		*Characteristics*			*Annotations*		
	Hours	Task Sessions	Natural	Interactive	Mistakes	Action Descriptions	Dialog Intents	Mistake Types
ADL (Pirsiavash and Ramanan, 2012)	10	20	✓			✓		
Charades-Ego (Sigurdsson et al., 2018)	69	68.5K	✓			✓		
EGTEA Gaze+ (Li et al., 2018)	28	86	✓			✓		
Epic-Kitchens (Damen et al., 2018)	55	432	✓			✓		
Ego4D (Grauman et al., 2022)	3.7K	931	✓			✓		
Assembly101 (Sener et al., 2022)	513	362	✓		✓	✓		✓
ALFRED (Shridhar et al., 2020)	–	8.1K				✓		
MindCraft (Bara et al., 2021)	12	100		✓		✓		
TEACh (Padmakumar et al., 2022)	–	3.2K		✓	✓	✓		
WTaG (Ours)	10	56	✓	✓	✓	✓	✓	✓

Table 1: Comparison of WTaG to past egocentric task-oriented video datasets. We compare datasets in terms of data statistics, dataset characteristics (whether videos are natural, involve dialog interaction, or have annotated mistakes), and types of annotation (action type or narration, dialog act categorization, or mistake details) available.

in embodied AI have collected similar collaborative task completion datasets generated through virtual environment simulators (Shridhar et al., 2020; Bara et al., 2021; Padmakumar et al., 2022). They also target task-oriented dialog and mistakes but are not as natural nor can be generalized to the open world as can our work.

While smaller than some existing datasets (Table 1), WTaG emphasizes the interactions between the user and the instructor, prioritizes depth over breadth with this uniquely rich dataset for situated task guidance, and is (to our knowledge) the first of its kind with natural, human-to-human interactive videos annotated with conversations texts, recipe steps, dialog intents, and mistakes.

3.1 Data Collection

To collect data, two human subjects, an *instructor* and a *user*, are paired up. The user is tasked with completing 1 of 3 cooking recipes[4] while communicating with an instructor. To encourage natural mistakes and interaction between the instructor and user to occur, the instructor has access to a complete, detailed ground truth recipe, while the user only has a simplified version of it, with high-level directions and minimal details (Figure 9).

As shown in Figure 1, the instructor is separated from the user during task completion, watching and communicating with them through an egocentric camera view interface and external microphone connected to an augmented reality (AR) headset[5] worn by the user. We use Microsoft Azure automatic speech recognition (ASR)[6] to convert audio recordings from videos into conversations transcripts, and manually corrected them as needed. All data from both sides are synchronized and sent to a server for storage and future processing. Although not used in our experiments, we also collect 12 additional types of synchronized data using Microsoft Psi on top of the egocentric RGB video, user and instructor audios for each recording. More details can be found at https://github.com/microsoft/psi.

A total of 56 recordings were collected from 17 user subjects and 3 instructor subjects. All human subjects were over the age of 18, English-speaking college students with different genders and are from different cultural background with normal or corrected-to-normal vision, recruited through messaging platforms. Members of the study team served as the human instructors.

Together, 4,233 English dialog utterances were collected, evenly distributed over the 3 recipes (19 pinwheels, 18 coffees, 19 cakes). The length of the videos range from 5 to 18 minutes, with the median of 10 minutes. As shown in Figure 2, each video contains a median of 31 instructor utterances and 35 user utterances. The instructors talk a bit longer than the users in each utterance, with a median of 6 words per utterance versus 3 words for users. Instructors also speak faster than the users with a median speed of 398 ms/word versus 522 ms/word.

[4] Recipes include peanut butter and jelly pinwheels, pour-over coffee, and a microwaved mug cake.

[5] We use Microsoft HoloLens 2 (https://www.microsoft.com/en-us/hololens/), but such interaction can be enabled by any wearable audiovisual device.

[6] https://azure.microsoft.com/en-us/products/cognitive-services/speech-to-text/