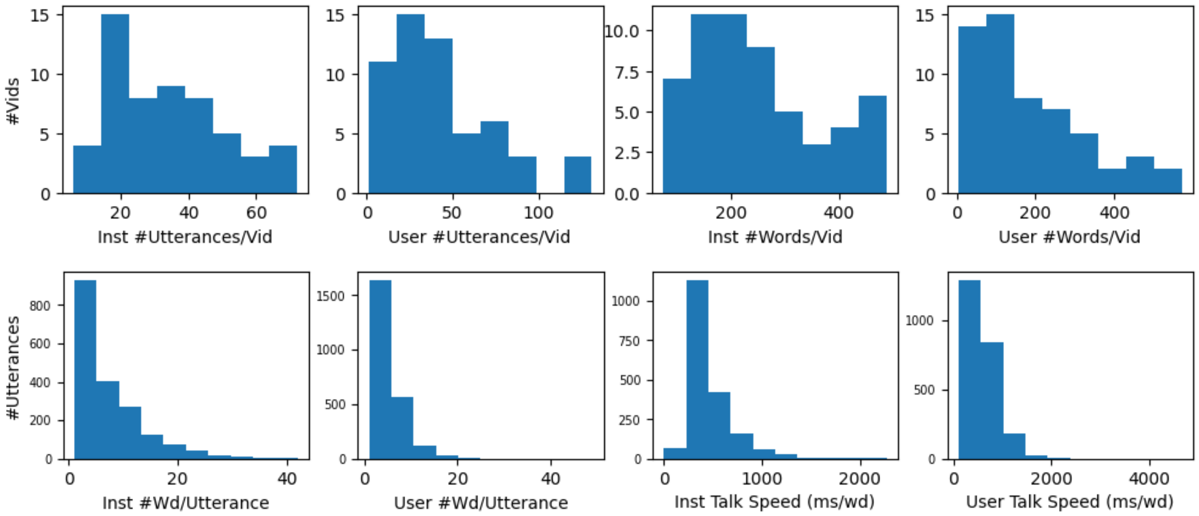

Figure 2: WTaG dataset statistics histogram for videos and dialog interactions.

3.2 Data Annotation

To have an in depth understanding on how human instructors navigate this complex task and nuanced situations, and an easier way to evaluate models' performance on WTaG, we provide rich annotations on recordings. First, we annotate the time span of each recipe step the user performs in each video, facilitating user state detection. Secondly, we manually go through all the ASR results, correct the transcripts and voice activity time span as needed, and filter out any potential harmful speech. Lastly, we categorize user and instructor utterances into a set of dialog intentions, together with mistake types if any. Details are as follows:

User utterances are categorized into 6 intents: *Question*, *Answer*, *Confirmation*, *Self Description*, *Hesitation*, and *Other*.

User mistakes are categorized into 3 classes: *Wrong Action*, *Wrong Object*, and *Wrong State* (including measurement and intensity).

Instructor utterances are categorized into the following 5 coarse-grained intents: *Instruction*, *Question*, *Answer*, *Confirmation*, and *Other*.

If the instructor decided to issue an "instruction", the **instructions** are further classified into 4 types based on what they inform users about: *Mistake Correction*, *Current Step*, *Next Step*, and *Details*.

4 Task Definitions

For benchmark evaluation, we extract query points from the WTaG dataset whenever one of the following conditions is met:

(a) GT user said something

(b) GT instructor said something

(c) GT no one said anything for 10 seconds

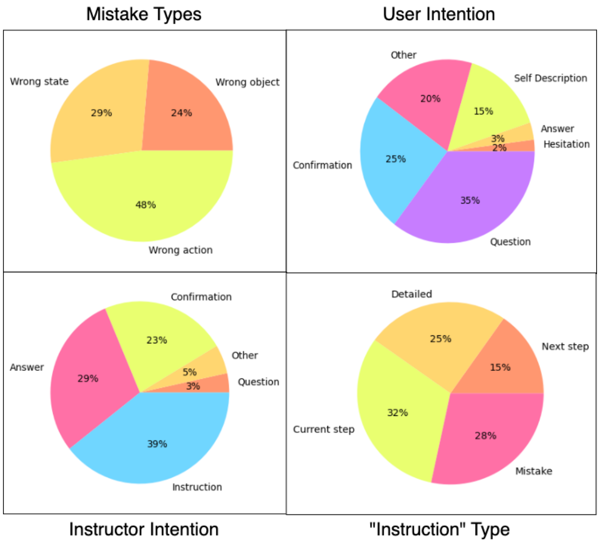

Figure 3: User and Instructor Dialog Intention Type Distributions and Mistake Type Distributions of WTaG

Each query point provides systems with the latest frame image, dialog history, the task recipe, and current elapsed time into the task. This creates a variety of situations, where the instructor may or may not need to intervene and provide guidance.

More specifically, for each query time point t, given the user's egocentric video frame, and the chat history, we formulate the following two tasks for the models to predict.

User and Environment Understanding

1. User intent prediction: Dialog intent of user's last utterance, if any (options).
2. Step detection: Current step (options).
3. Mistake Existence and Mistake Type: Did the user make a mistake at time t (yes/no). If so, what type of mistake (options).