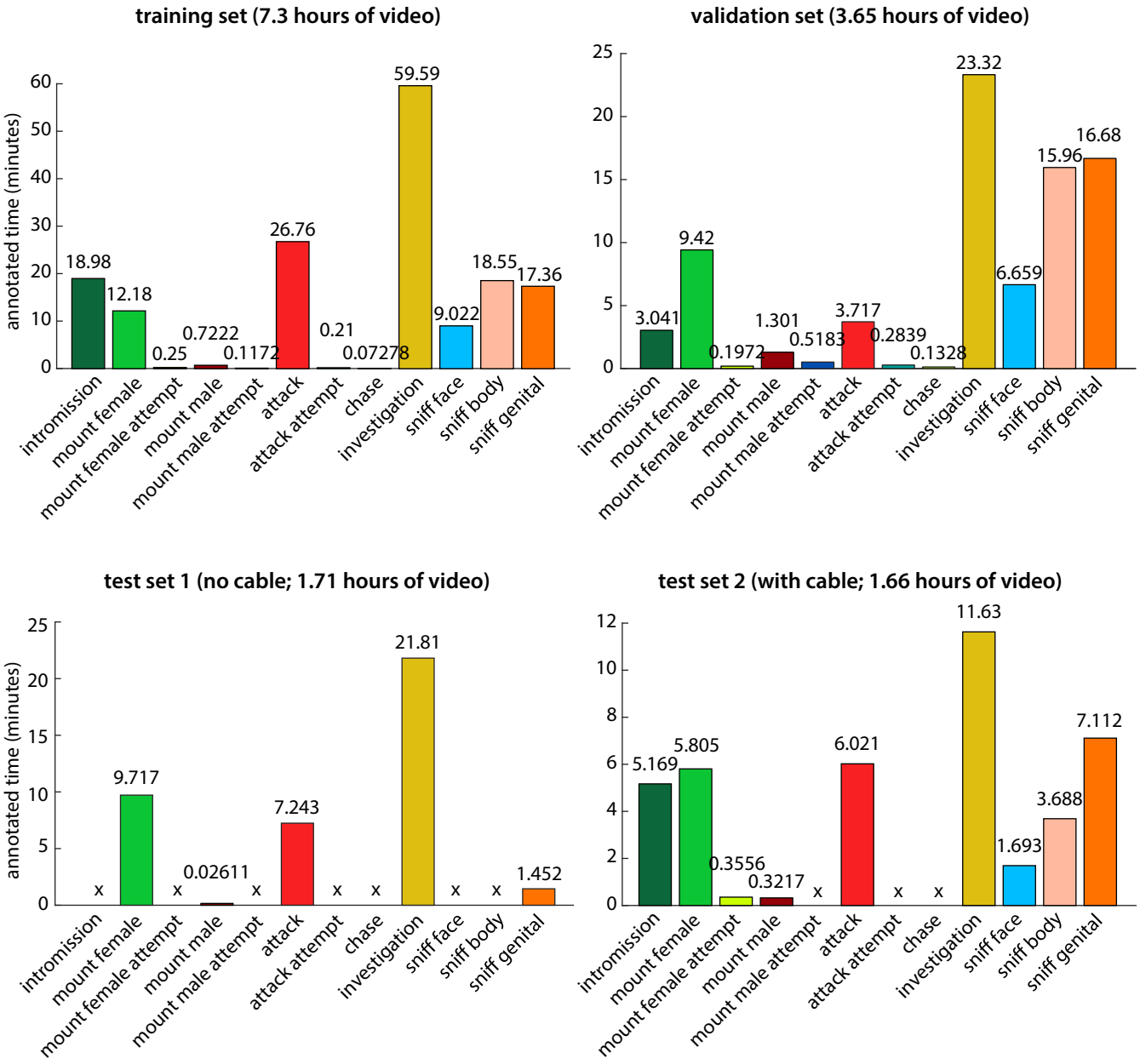
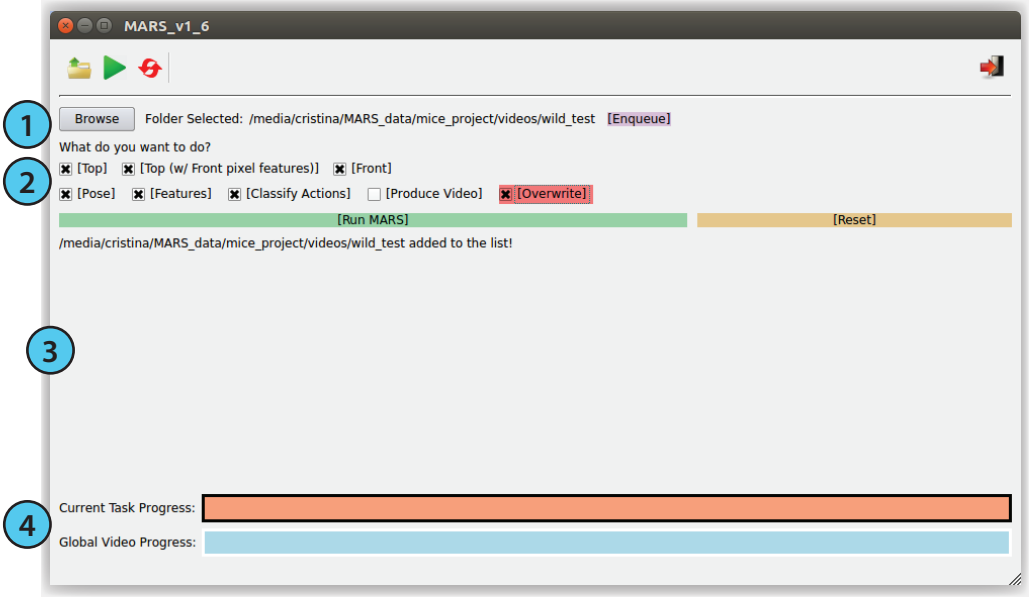
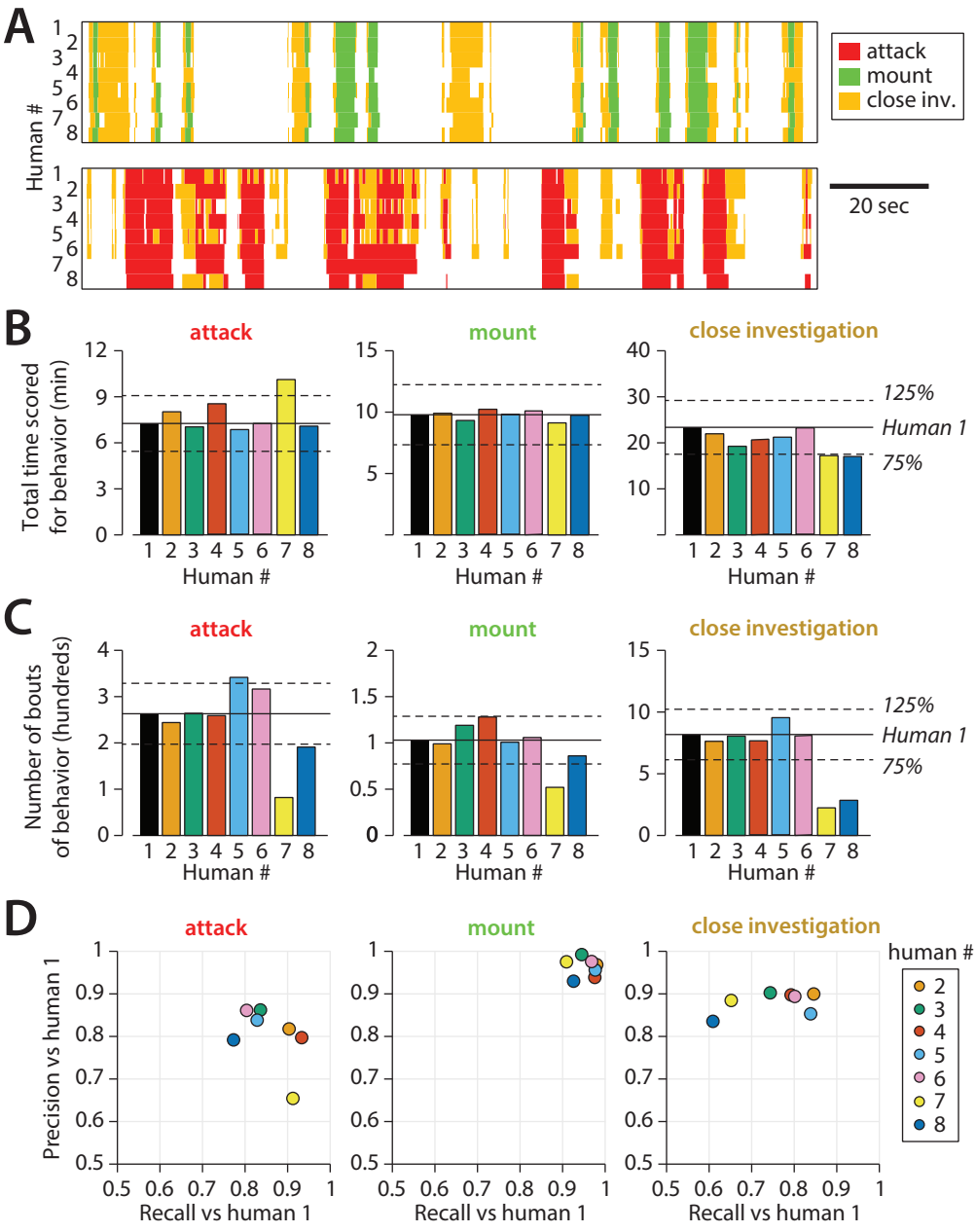


ED Figure 1. The MARS annotation dataset. Number of hours scored for each of 12 behaviors in the 14.3-hour MARS dataset, broken down by training, validation, and test sets. While all videos were scored for attack, mount, and close investigation behaviors, the remaining behaviors were not always scored explicitly, but instead were scored as attack, mounting, or close investigation (see Methods for details). These bar graphs should therefore not be taken as indicative of the relative frequencies of these remaining behaviors in the dataset.

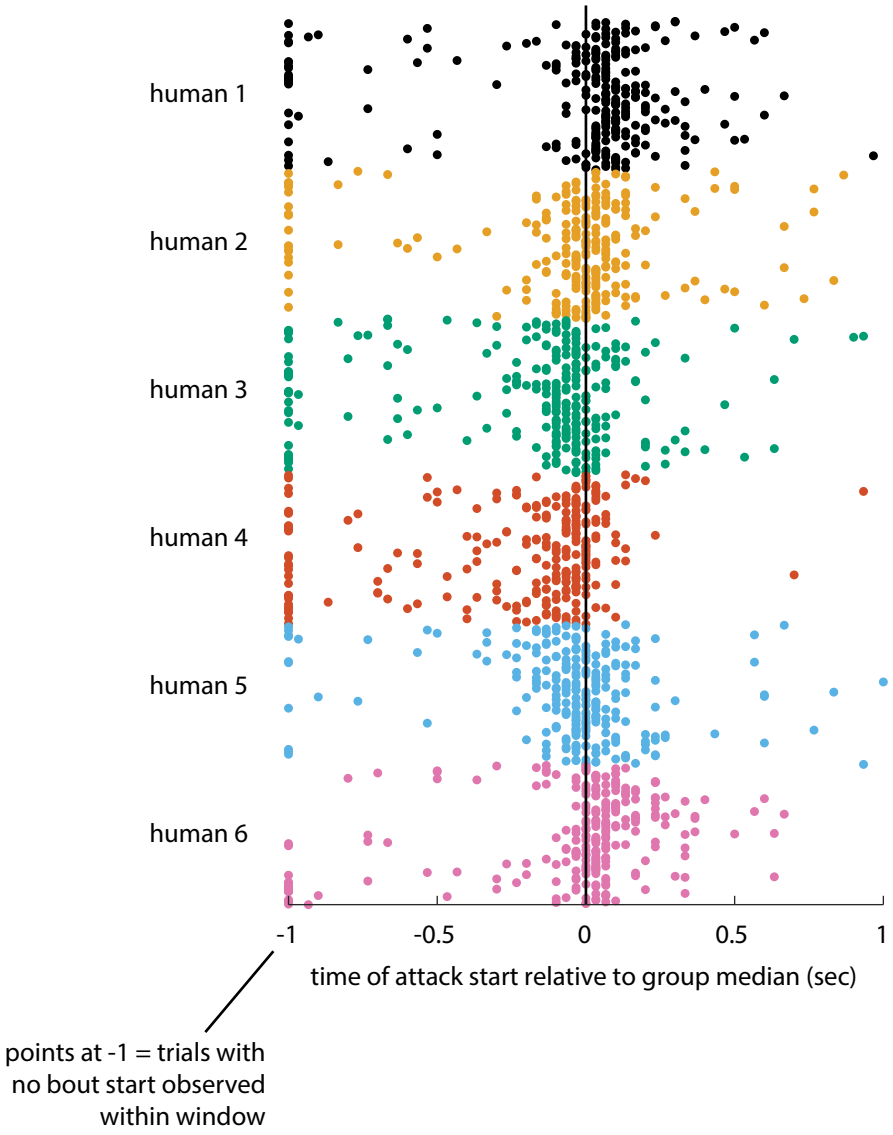


1. File navigator, supporting queueing of multiple jobs while tracking is running.
2. User options: specify video source (top/front view camera), type of features to extract, and analyses to perform (pose estimation, feature extraction, behavior classification, video output.)
3. Display of status updates during analysis.
4. Progress bars for current video and for all jobs in the queue.

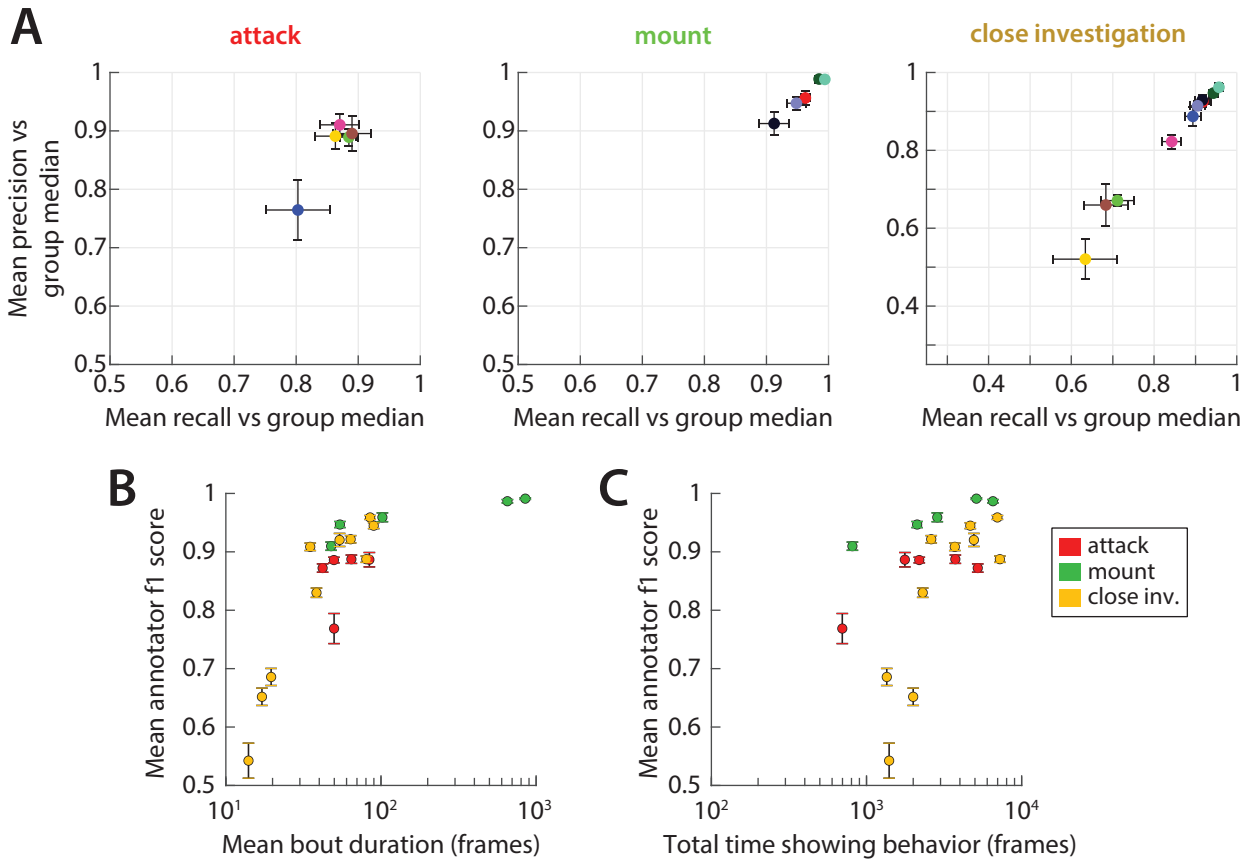
ED Figure 2. MARS graphical user interface. This Python-based GUI allows easy user access to MARS on a desktop computer.



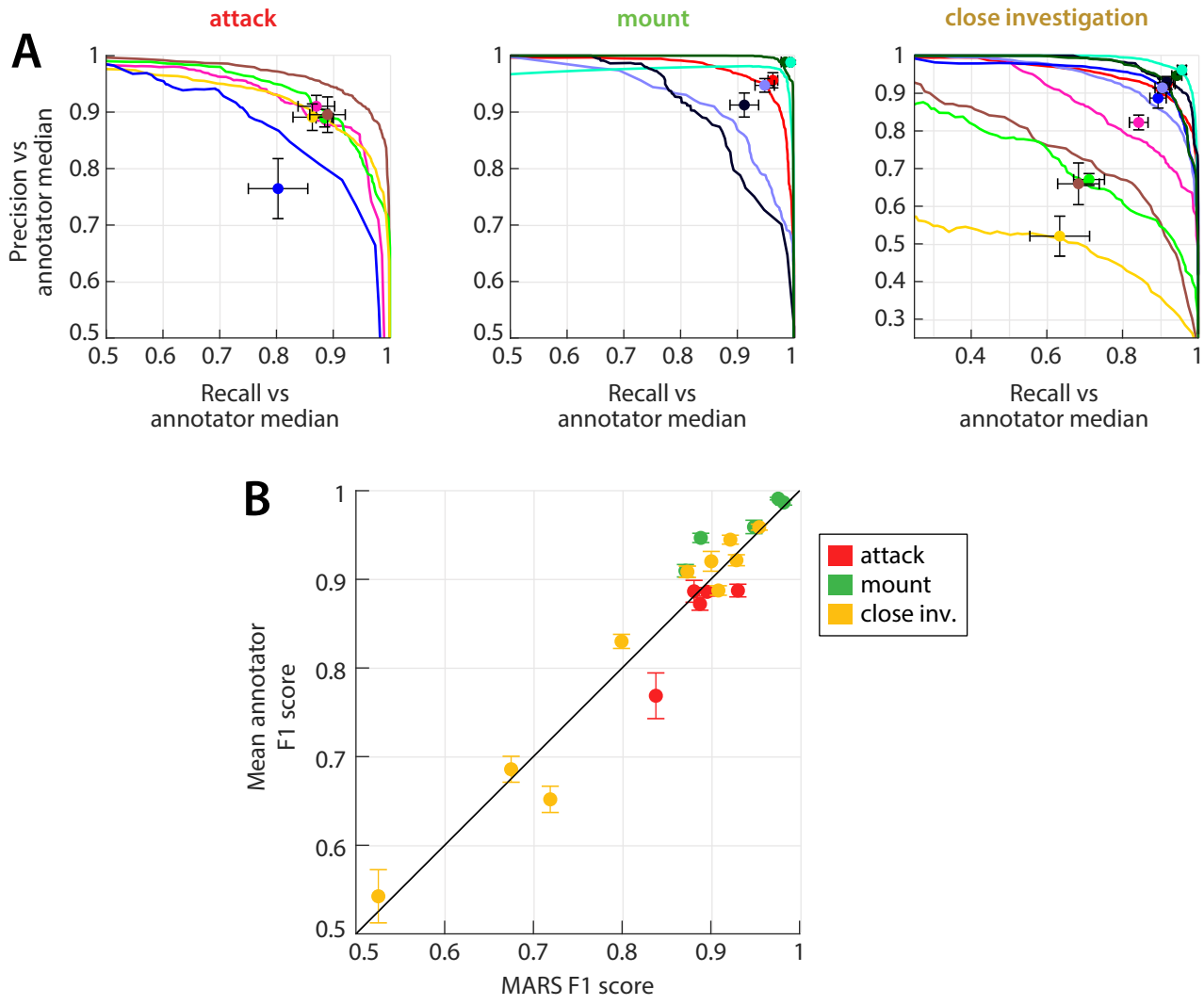
ED Figure 3. Expanded set of human annotations. All panels as in **Figure 5**, but with the two omitted annotators (human 7 and 8) included. **A)** Example annotation for attack, mounting, and close investigation behaviors by eight trained annotators on segments of male-female (top) and male-male (bottom) interactions. **B)** Inter-annotator variability in the total reported time mice spent engaging in each behavior. **C)** Inter-annotator variability in the number of reported bouts (contiguous sequences of frames) scored for each behavior. **D)** Precision and recall of annotators (humans) 2-8 with respect to annotations by human 1.



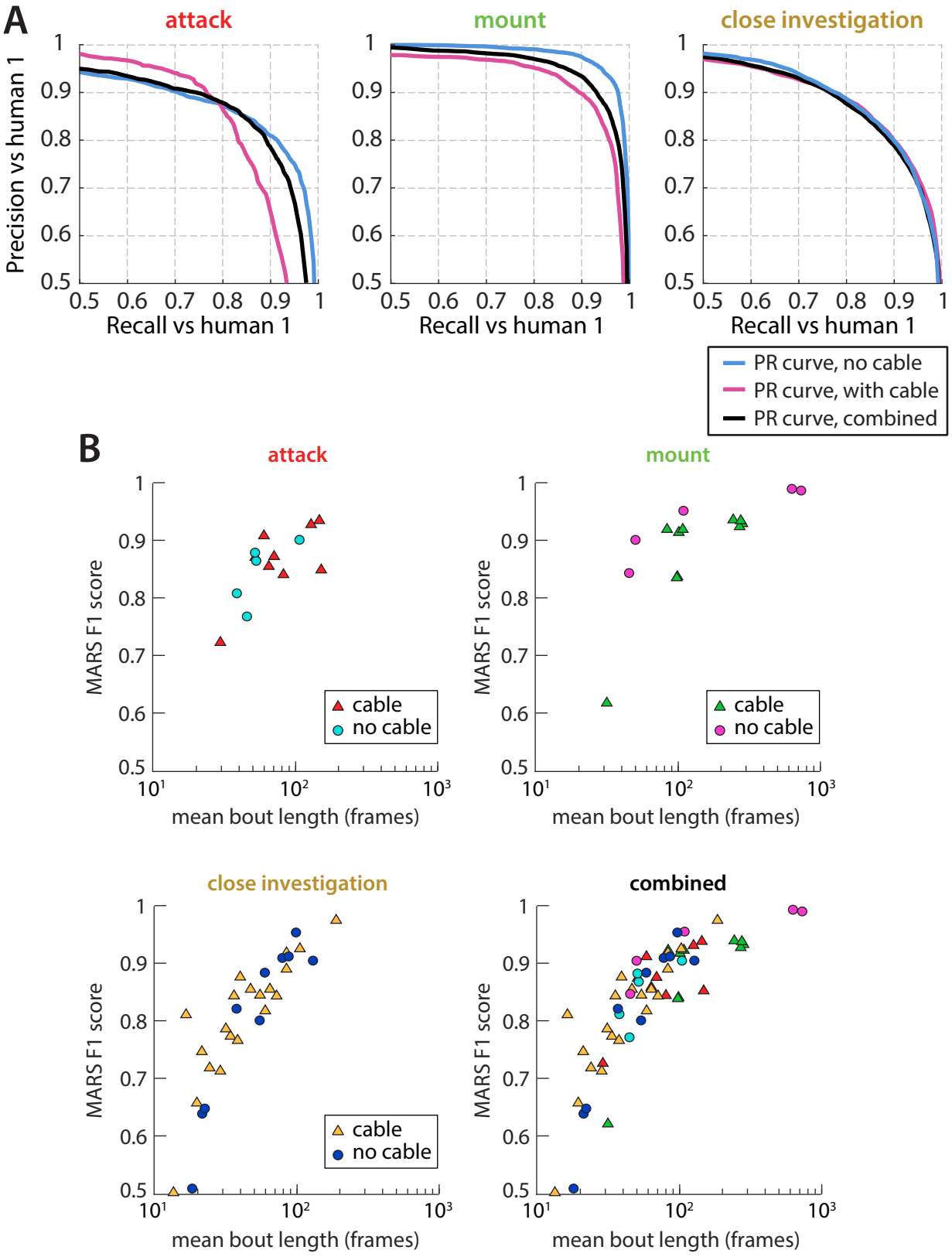
ED Figure 4. Within-annotator bias and variance in annotation of attack start time. Annotations of all attack bouts in the 10-video dataset by six human annotators. All attack bouts are aligned to the first frame on which at least three human annotators scored attack as occurring. Colored dots then reflect the time when each annotator scored each bout as starting, relative to this aligned time (the group median). Each annotator shows a characteristic bias (a shift in their mean annotation start time before or after the group median) and variance (the spread of annotation start times around this mean) in their annotation style. Some annotators did not score any attack initiated within a +/- 1 second window of the group median for a given bout: these points are plotted at time -1. Note that the average attack bout in the dataset is 1.65 seconds long (using annotations from human 1).



ED Figure 5. Inter-annotator accuracy on individual videos. A) Mean Precision and Recall of annotators 1-6, computed relative to the median of the other five annotators (mean ± SEM.) Each plotted point is one video. **B)** Mean annotator F1 score (harmonic mean of Precision and Recall) plotted against the mean bout duration for each behavior in each video. Plot suggests a close positive correlation between the average duration of behavior bouts in a video (or dataset) and the accuracy of annotators as computed by Precision and Recall. **C)** Mean annotator F1 score plotted against the total number of frames annotated for a given behavior in each video. Correlation is weaker than in **B.**



ED Figure 6. MARS Precision and Recall is closely correlated with that of annotators on individual videos. A) Mean Precision and Recall of annotators 1-6 for each behavior in each of 10 tested videos (plotted points; as in ED Figure 5), and MARS Precision-Recall (PR) curves for those videos. PR curves and points that are the same color correspond to the same video. **B)** Mean annotator F1 score plotted against MARS's F1 score for each behavior in each video. Performance of MARS is well predicted by the inter-human F1 score, which is in turn correlated with mean behavior bout duration (see **ED Fig 5**).



ED Figure 7. Evaluation of MARS on a larger test set. A) Precision-Recall (PR) curves of MARS classifiers for test set 1 ("no cable"), test set 2 ("with cable") and for the two sets combined. **B)** F1 score of MARS classifiers for each behavior in each video, plotted against mean behavior bout duration in that video. Plots show no strong difference in performance between videos in which mice are unoperated ("no cable") and videos in which mice are implanted with a head-attached device ("cable").