

Supplementary Material

Here we show details about the human evaluations illustrated in Section 6 of the main text.

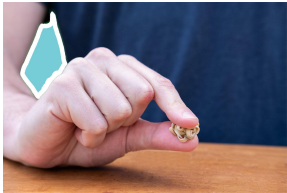



Figure 1	2
Instructions	2
Figure 2	3
Example Stimuli	3
Figure 3	4
Stimuli with Accuracy below Chance Level	4
Figure 4	6
Figure 5	7




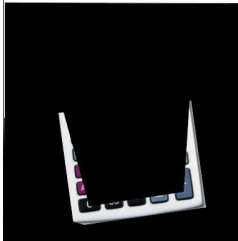


Figure 1

Instructions

The image with the white background is the original image of an item. The image with the black background contains only some regions of the same item. Please select the way you would use your hands to interact with those specific regions of the item from the four possible choices: palm, clench, pinch, poke.

Here are some examples of the four possible hand actions.



Pinch	Palm	Poke	Clench
			

Good Example	Good Example	Bad Example
Poke	Clench	Clench
 	 	 

These were the actual instructions used for the segmentation evaluation task on Mturk.

Figure 2

Example Stimuli


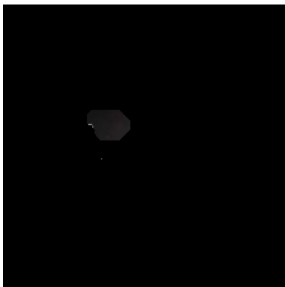
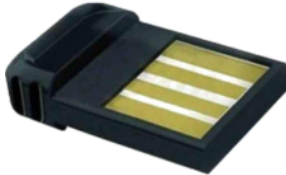

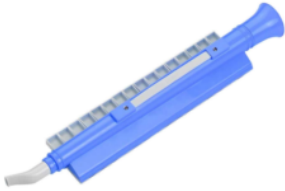
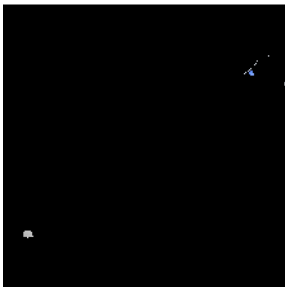





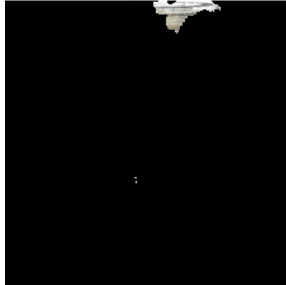




Typical Stimulus	Catch Stimulus
	




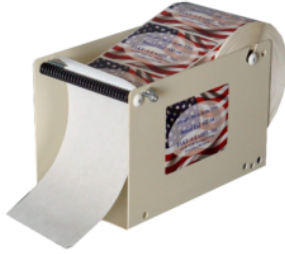

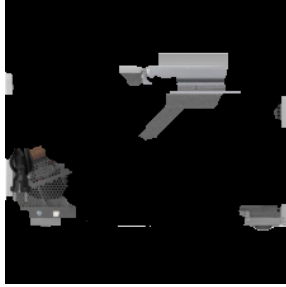






The *typical stimulus* is constructed by having the original image of the object against a white background placed above the segmented region of the highest predicted action against a black background placed below.

The *catch stimulus* is constructed in the same way as a typical stimulus, but there is additional white-colored text on the bottom image that prompts the Mturker to select a particular hand action. This hand action was pre-selected randomly from the four possible hand actions during the creation of the stimulus set, so all Mturkers see the same catch stimuli and prompts.

Figure 3

Stimuli with Accuracy below Chance Level

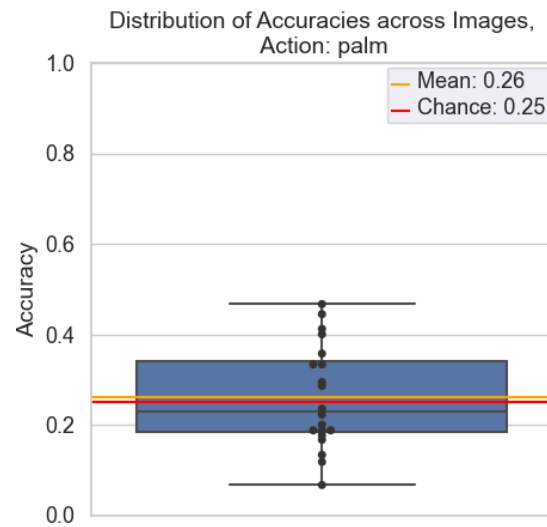
 	 	 	 
Predicted Label: Palm	Predicted Label: Palm	Predicted Label: Clench	Predicted Label: Palm
Accuracy: 0.167	Accuracy: 0.188	Accuracy: 0.133	Accuracy: 0.133
 	 	 	 
Predicted Label: Pinch	Predicted Label: Clench	Predicted Label: Palm	Predicted Label: Palm
Accuracy: 0.2	Accuracy: 0.214	Accuracy: 0.188	Accuracy: 0.222

			
			
Predicted Label: Palm	Predicted Label: Palm	Predicted Label: Palm	Predicted Label: Palm
Accuracy: 0.118	Accuracy: 0.2	Accuracy: 0.235	Accuracy: 0.067
			
			
Predicted Label: Palm	Predicted Label: Palm		
Accuracy: 0.176	Accuracy: 0.188		

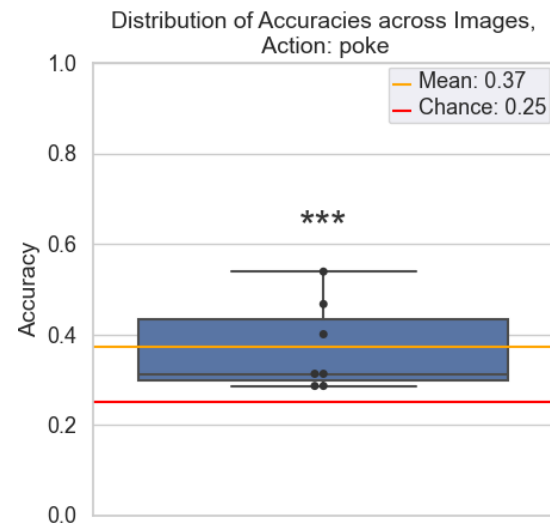
These were all the images that had bad annotations, i.e. the accuracy was below chance level. In the first row, we see that the segmentation regions shown in the bottom images are very dark and sparse. Almost all of these images have *palm* as their predicted label.

Figure 4

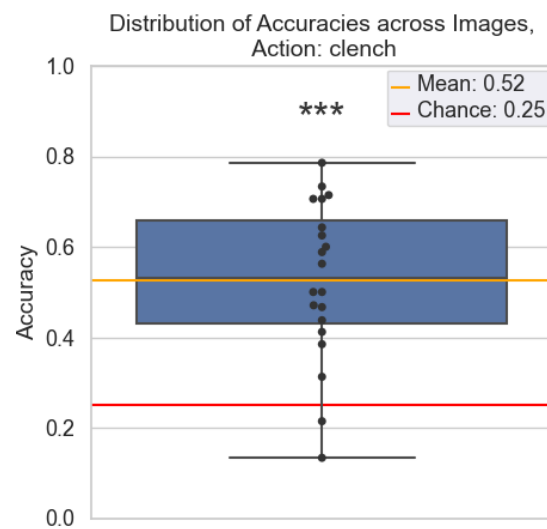
a.



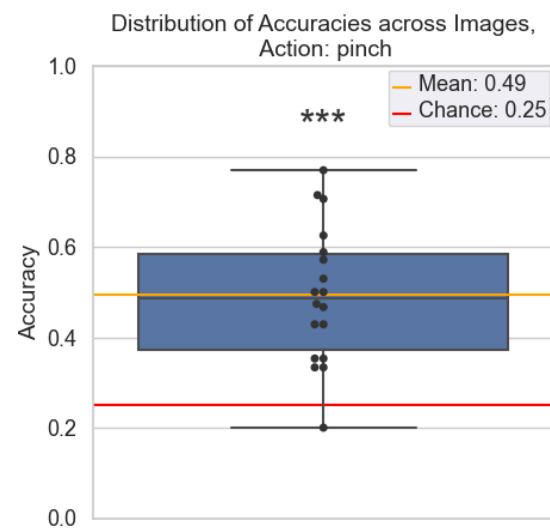
b.



c.



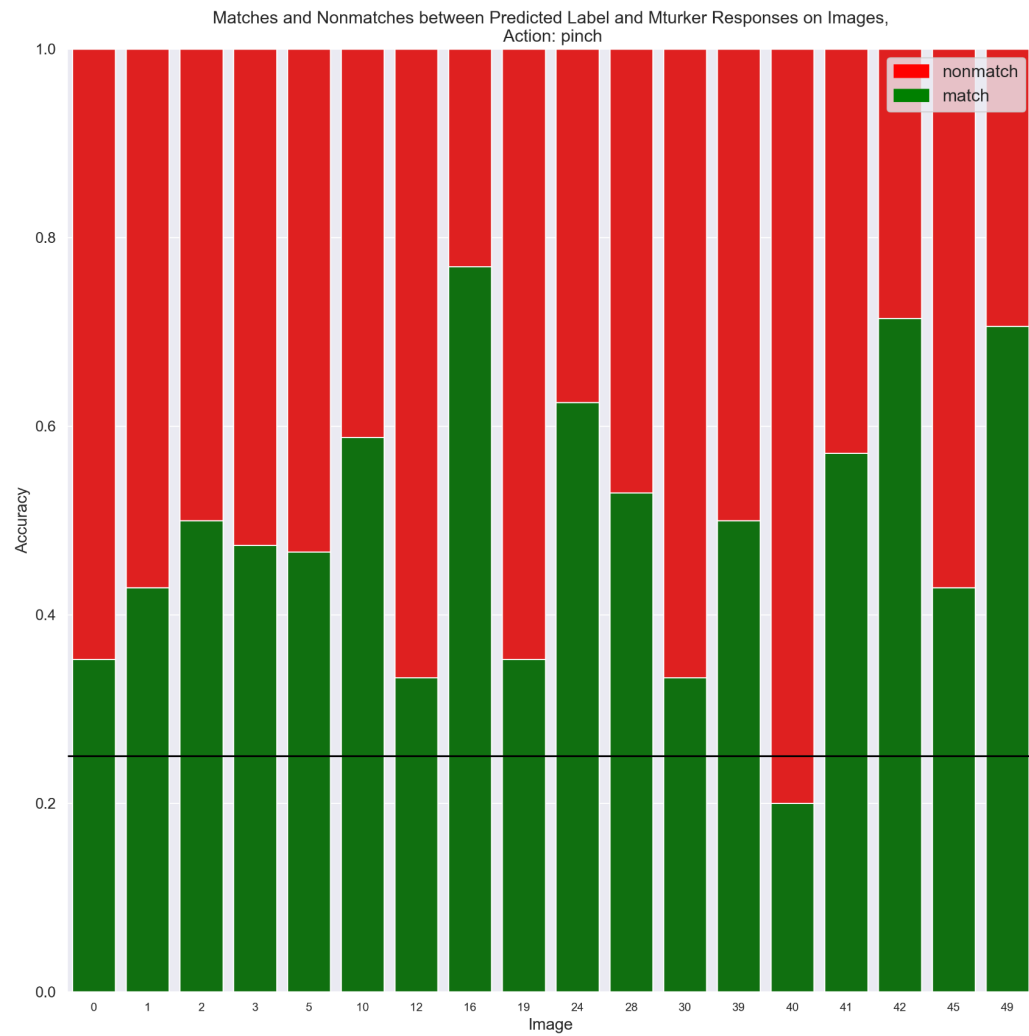
d.



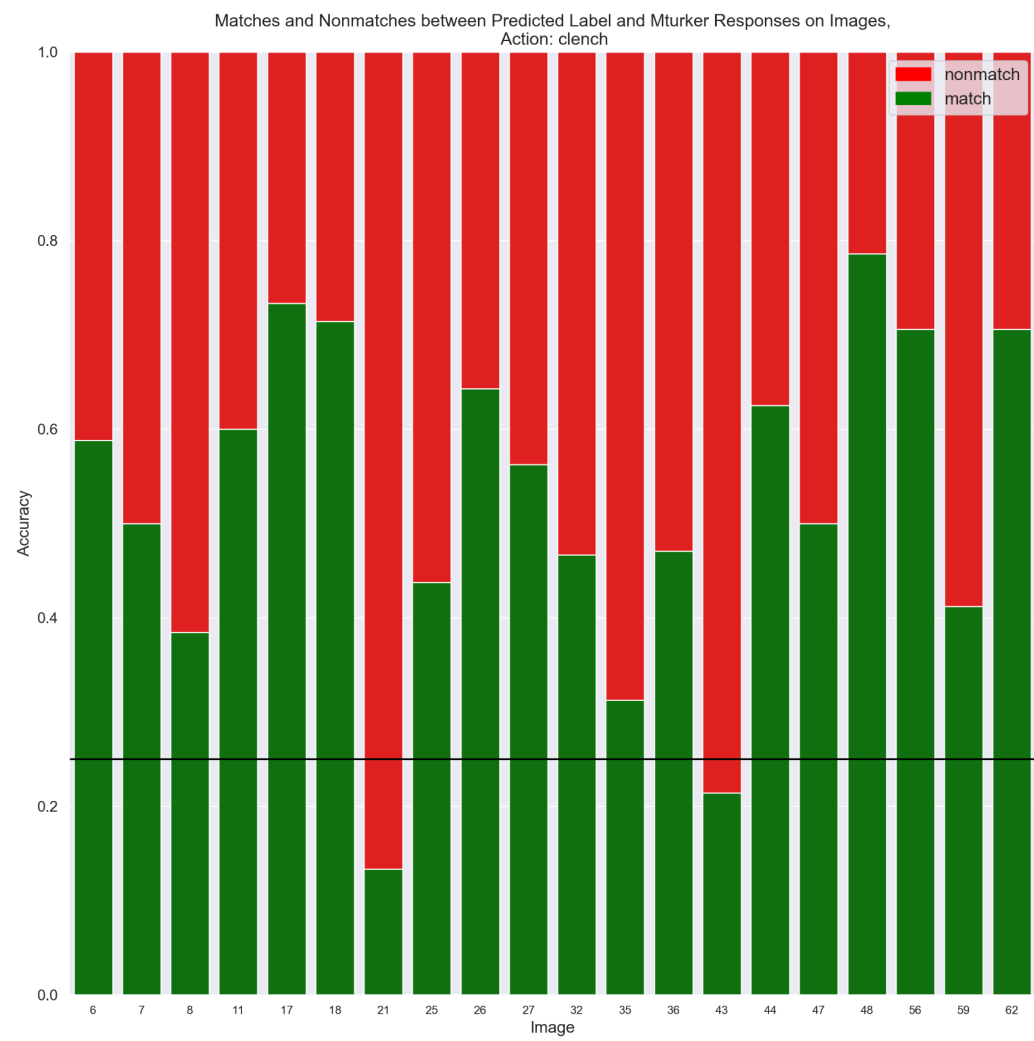
Accuracy distributions for images grouped by the predicted labels from the model. We have performed the same analysis on images separated into different categories according to the predicted label (a, b, c, d)

Figure 5

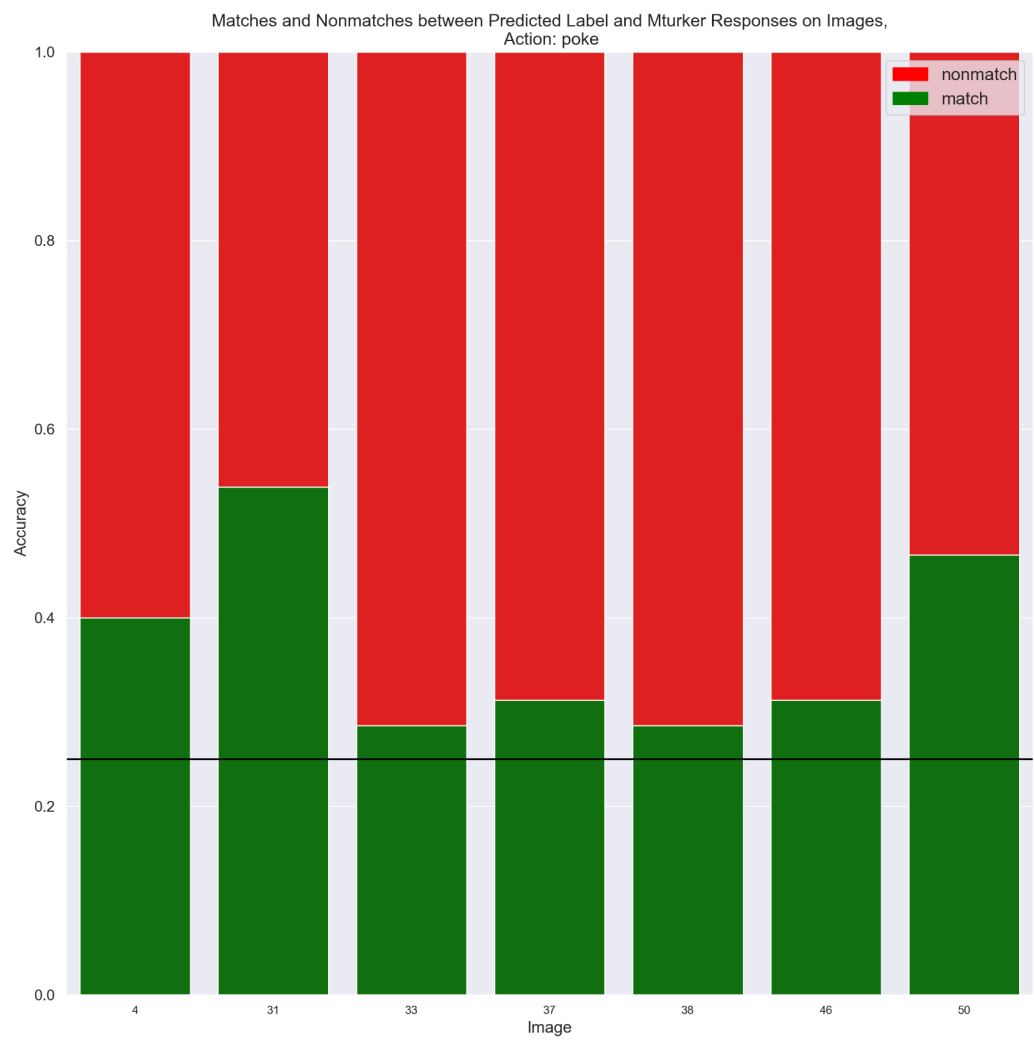
a. Pinch-predicted images



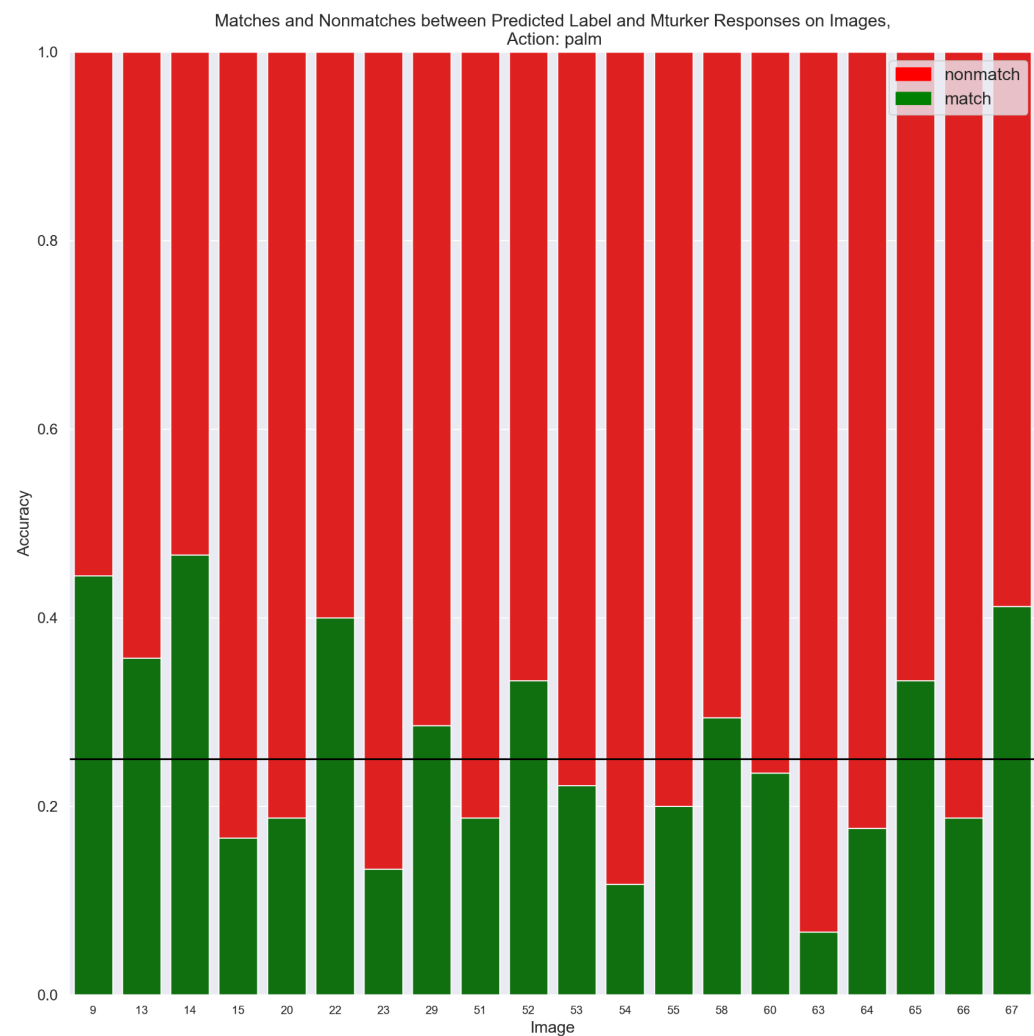
b. Clench-predicted images



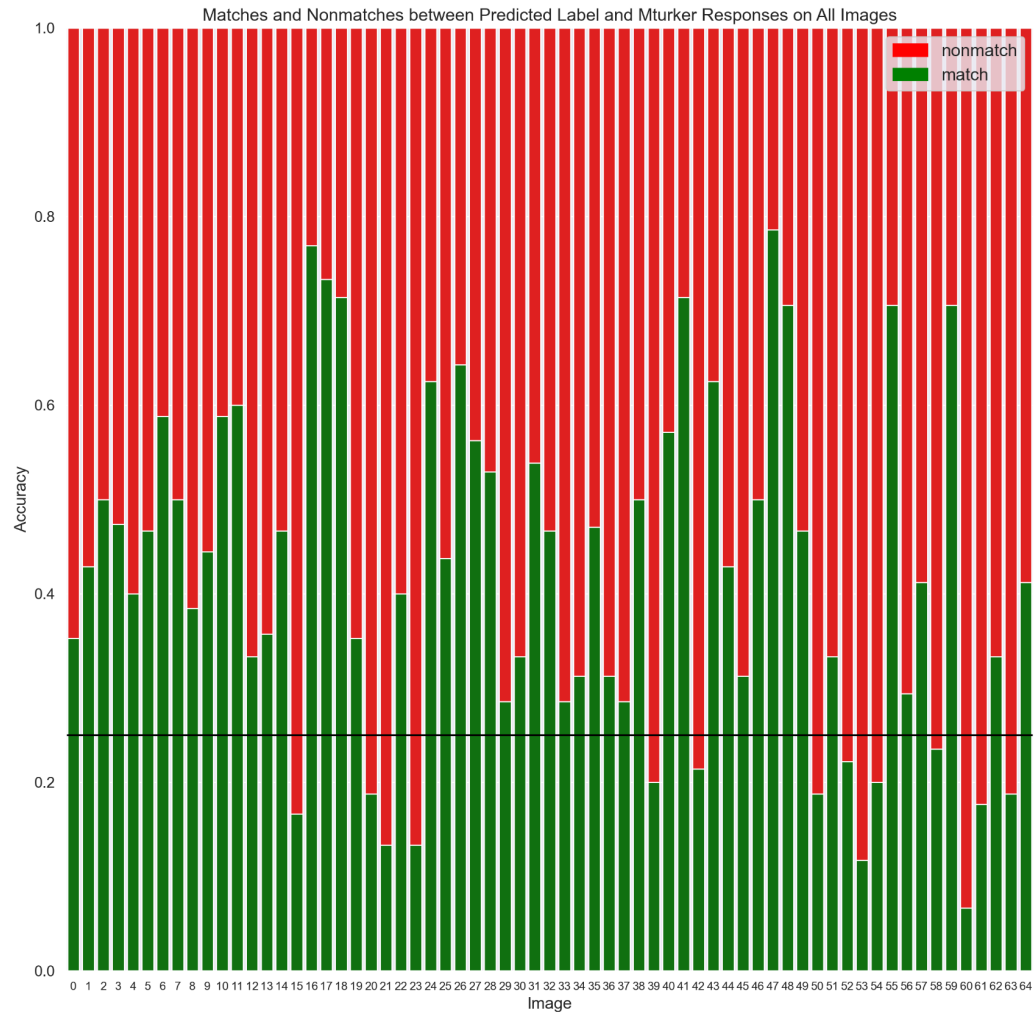
c. Poke-predicted images



d. Palm-predicted images



e. All images



Accuracy of each image by the predicted labels from the model. Each stimulus was assigned a number from 0 to 67, for the 68 stimuli. Of the 68 stimuli, 3 were catch stimuli. We collected an average of 15.23 human annotations for each of the remaining 65 stimuli. Along the x-axis, we see the stimulus index. The height of the green bars shows the percentage of human annotations for each image that matched the predicted label from our prediction model. We have performed the same analysis on images separated into different categories according to the predicted label (a, b, c, d), and for all these images together (e). The black line marks the chance level, 0.25.