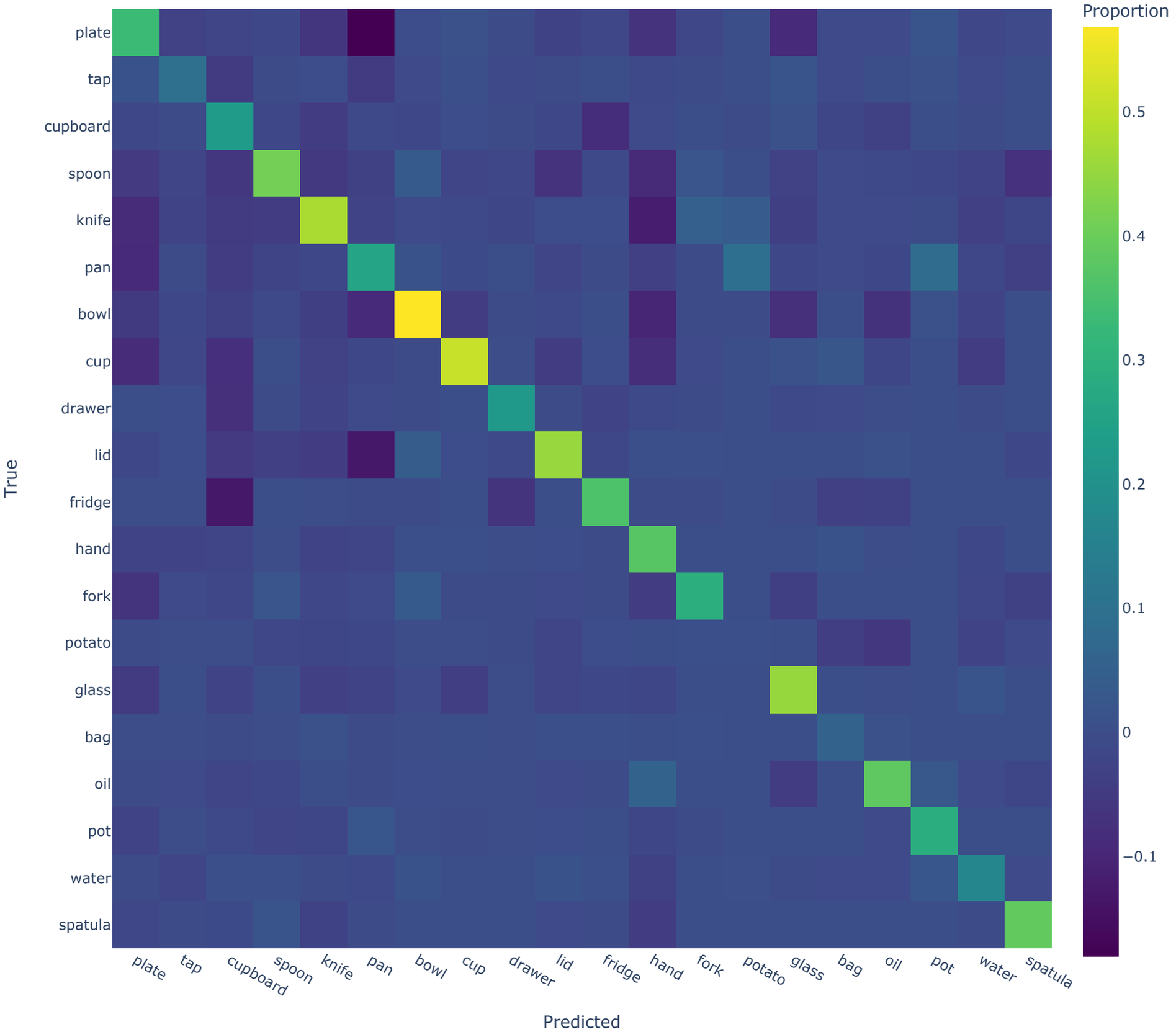Difference in confusion matrix for the top 20 noun classes
between **multi-modal ensemble** and AudioSlowFast