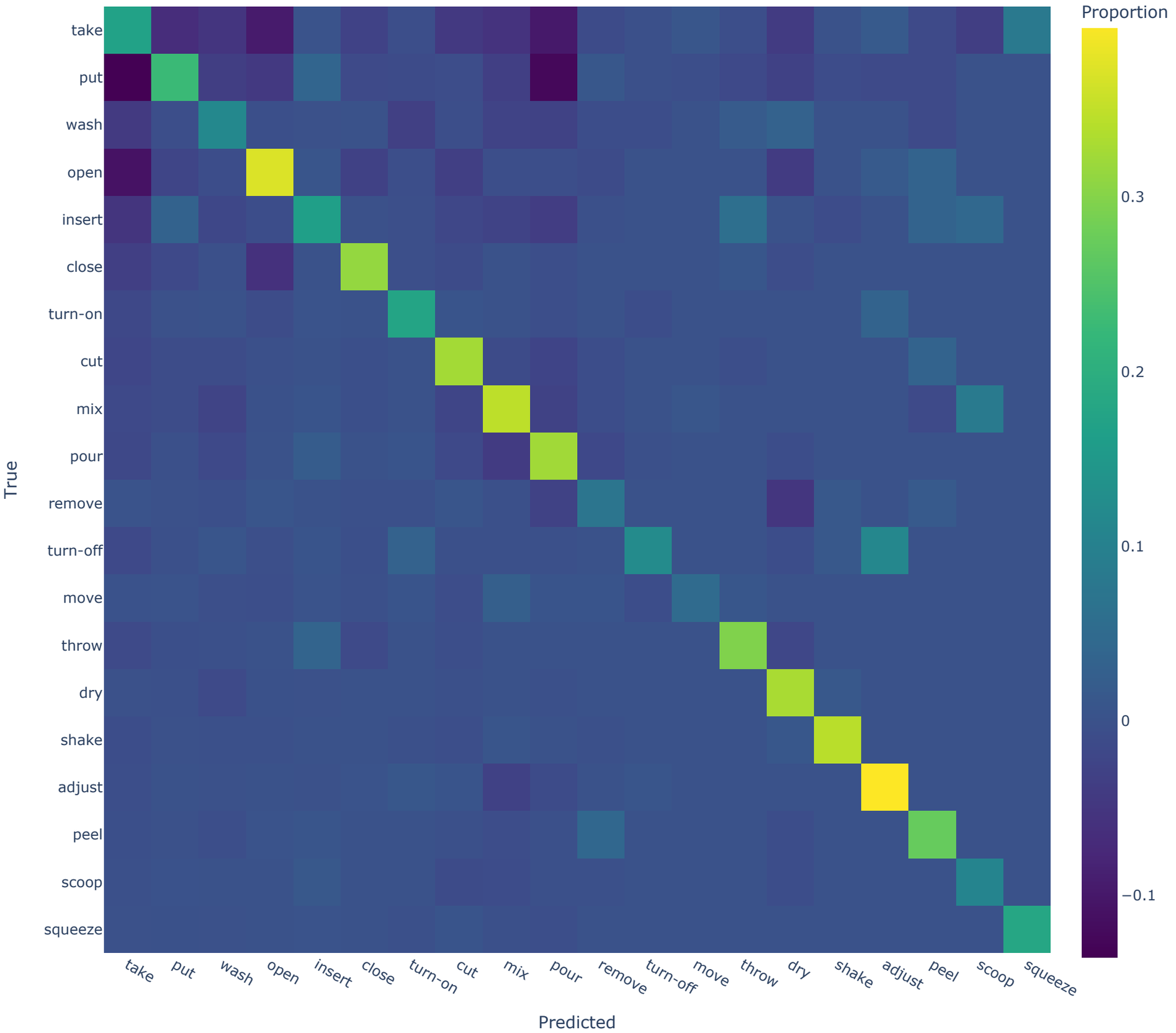Difference in confusion matrix for the top 20 verb classes
between **multi-modal ensemble** and AudioSlowFast