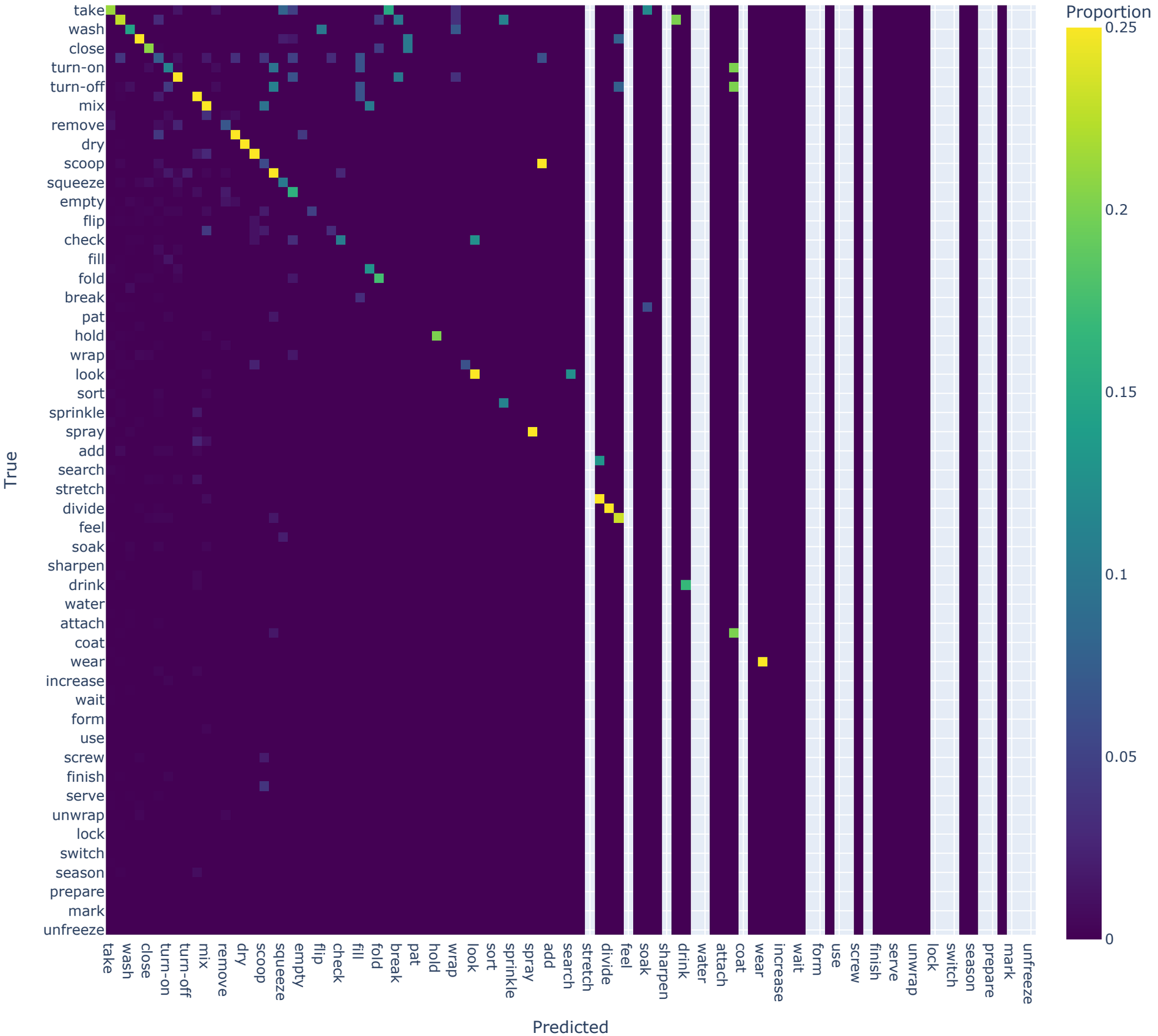Difference in confusion matrix for the top None verb classes between **multi-modal ensemble** and AudioSlowFastGRU