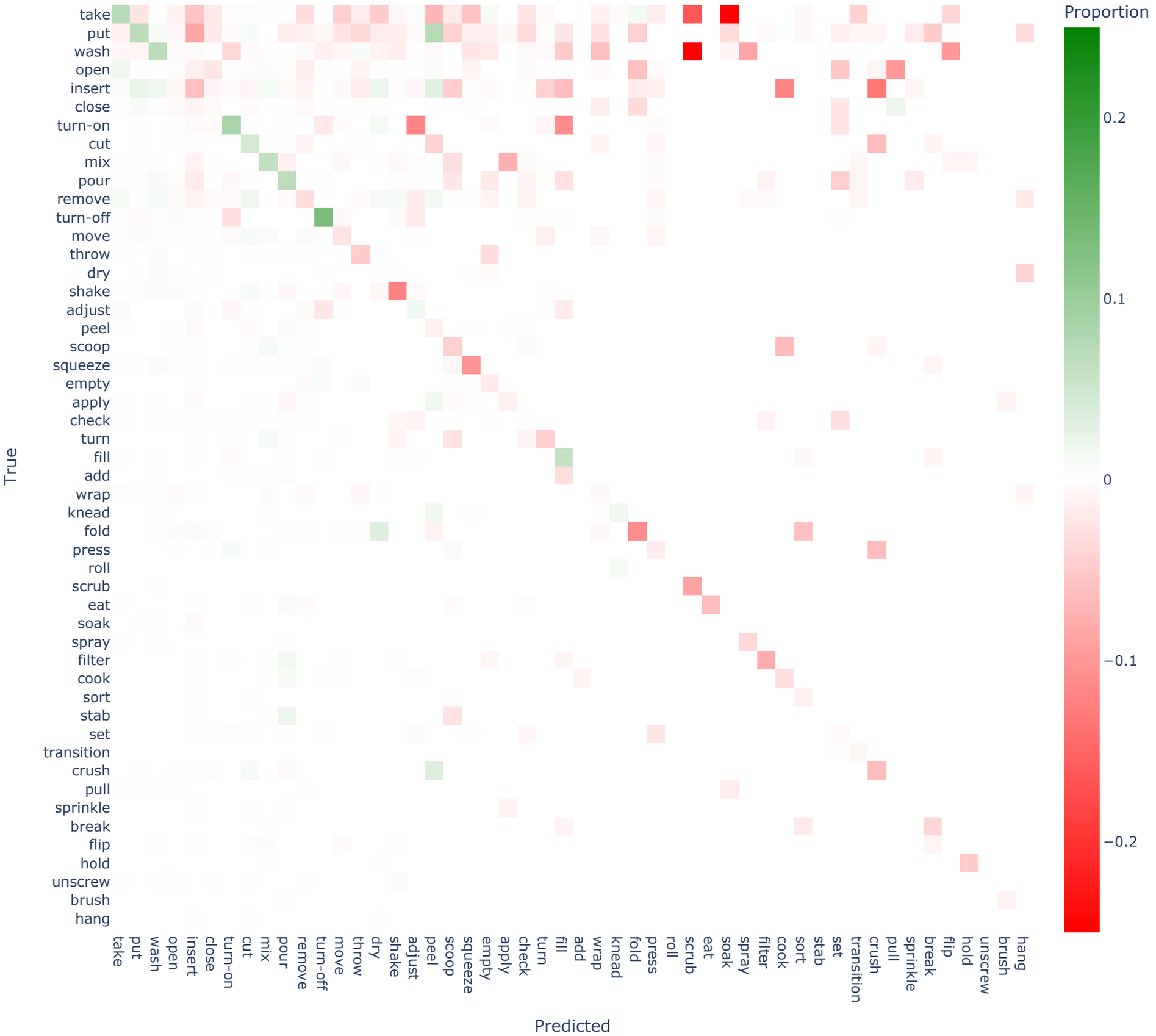Difference in confusion matrix for the top 50 verb classes
between **late-fusion multimodal ensemble (SlowFast + AudioSlowFast)** and SlowFast