

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

