

(a) Visualization of attention maps of make scrambled egg

b) Visualization of attention maps of make pancake

(c) Cross-modal matching probability statistics