Verb-object query: "Slicing vegetables"

Verb-object feature

Video frames $\{I_1, \ldots, I_T\}$

Video frame I_t

Human region features f_t^h

Object region features f_t^o

Contextual frame feature x_t

Region attention

Attended human feature Φ_t^h

Attended object feature Φ_t^o

Object region features from other video frames $\mathcal{F}$

Weakly supervised contrastive loss