A batch of video clips with variable length of frames

$\alpha_{S_m,1}, \cdots \alpha_{S_m,j}, \cdots, \alpha_{S_m,L}$

S_m

$\alpha_{S_m,j}$ is the context-aware attention of j-th frame of m-th video clip S_m