$t^{(1)}$ "a beautiful sceneary is shown and the sight seeing through the train is amazing and place is so lovely to watch"

$t^{(2)}$ "The train is passing through a forest"

$t^{(3)}$ "Outside the train window, there are lakes, mountains, and a small town."

$t^{(4)}$ "the girl shows the boys her medal"

$t^{(5)}$ "a cartoon shows two dogs talking to a bird."

$t^{(6)}$ "Two dogs feel confused in the cartoon."

Video Encoder $F_\theta(\cdot)$

Text Encoder $G_\phi(\cdot)$