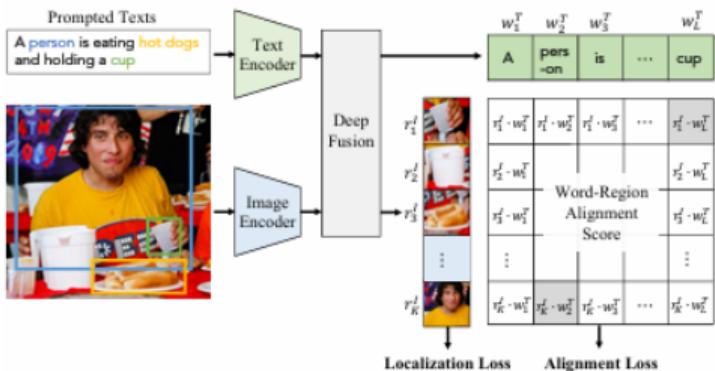Models global correlation between images and texts using a binary classification loss over the alignment score.

Loss Formulation

$$\mathcal{L}_{IT} = p \log S(z^I, z^T) + (1-p) \log(1 - S(z^I, z^T))$$