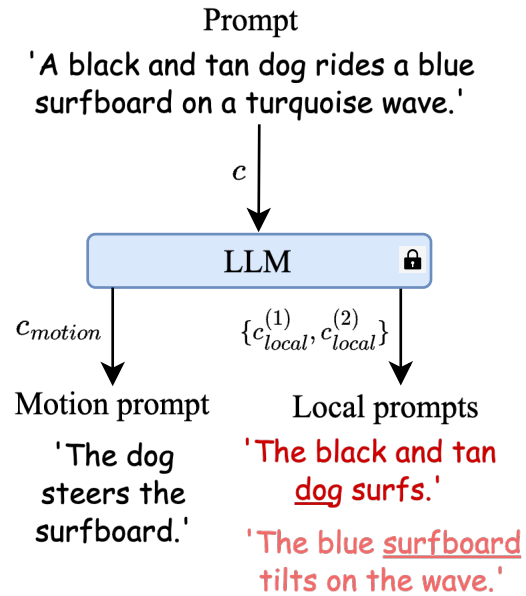


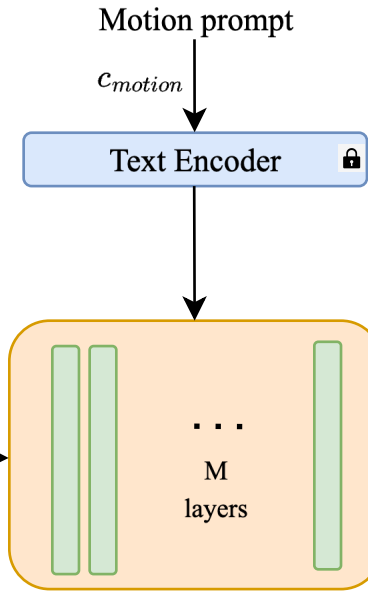
Data Pre-processing



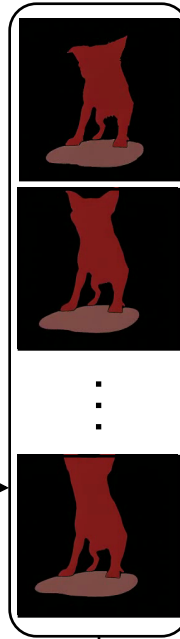
$[\epsilon; s^{(0)}; x^{(0)}]$



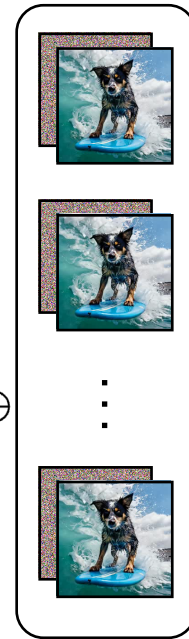
Image-to-Motion



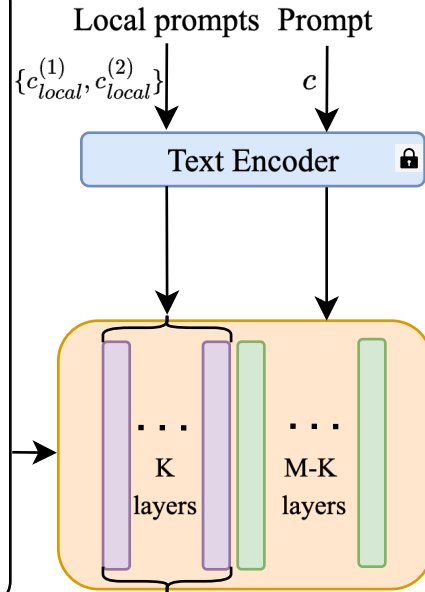
\hat{s}



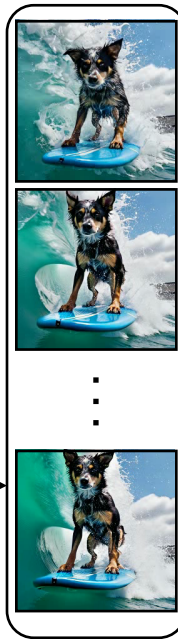
$[\epsilon; x^{(0)}]$



Motion-to-Video



\hat{x}



Attention block



Masked Attention block