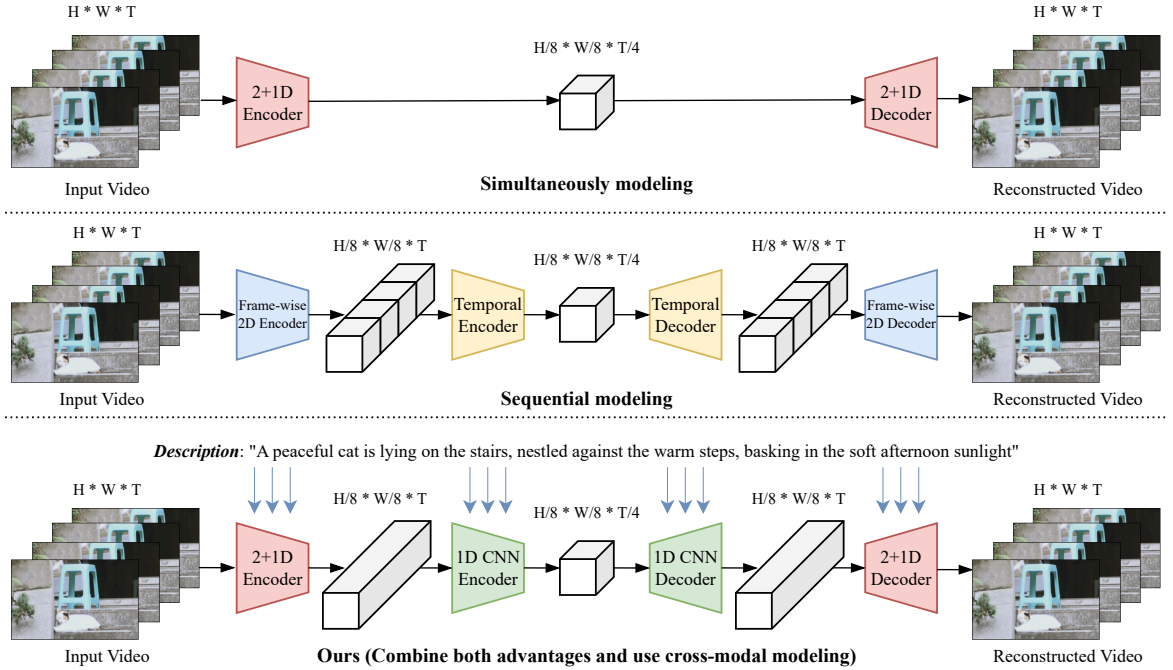
H * W * T
2+1D Encoder
H/8 * W/8 * T/4
2+1D Decoder
H * W * T
Input Video
Simultaneously modeling
Reconstructed Video
H * W * T
Frame-wise 2D Encoder
H/8 * W/8 * T
Temporal Encoder
H/8 * W/8 * T/4
Temporal Decoder
H/8 * W/8 * T
Frame-wise 2D Decoder
H * W * T
Input Video
Sequential modeling
Reconstructed Video
Description: "A peaceful cat is lying on the stairs, nestled against the warm steps, basking in the soft afternoon sunlight"
H * W * T
2+1D Encoder
H/8 * W/8 * T
1D CNN Encoder
H/8 * W/8 * T/4
1D CNN Decoder
H/8 * W/8 * T
2+1D Decoder
H * W * T
Input Video
Ours (Combine both advantages and use cross-modal modeling)
Reconstructed Video