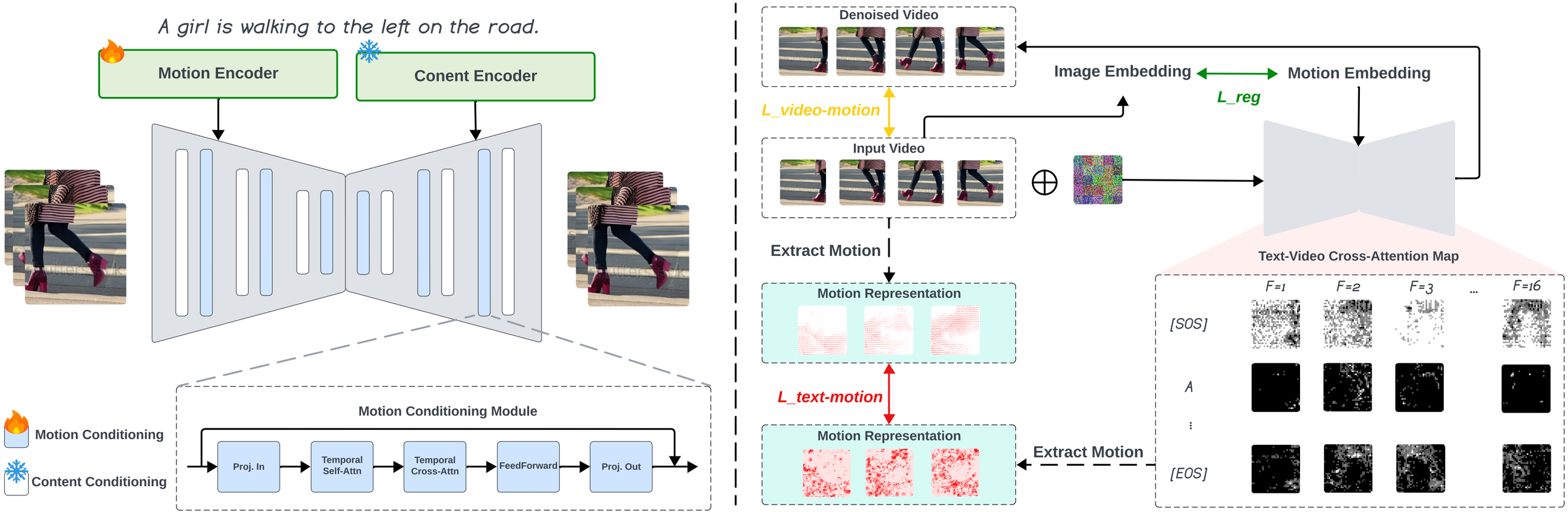
A girl is walking to the left on the road.
Motion Encoder
Conent Encoder
Motion Conditioning Module
Proj. In
Temporal Self-Attn
Temporal Cross-Attn
FeedForward
Proj. Out
Motion Conditioning
Content Conditioning
Denoised Video
L_video-motion
Input Video
Extract Motion
Motion Representation
L_text-motion
Motion Representation
Image Embedding
Motion Embedding
L_reg
Text-Video Cross-Attention Map
F=1
F=2
F=3
...
F=16
[SOS]
A
[EOS]
Extract Motion