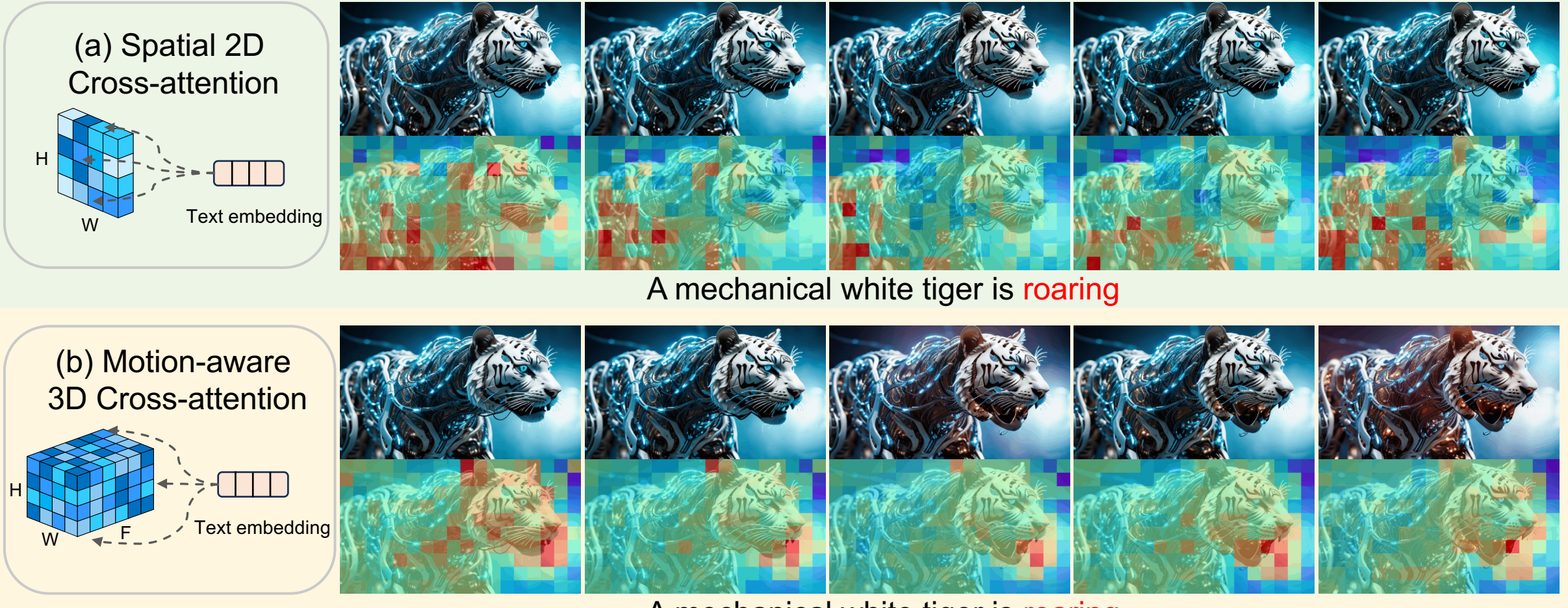
(a) Spatial 2D Cross-attention
H
W
Text embedding
A mechanical white tiger is roaring
(b) Motion-aware 3D Cross-attention
H
W
F
Text embedding
A mechanical white tiger is roaring