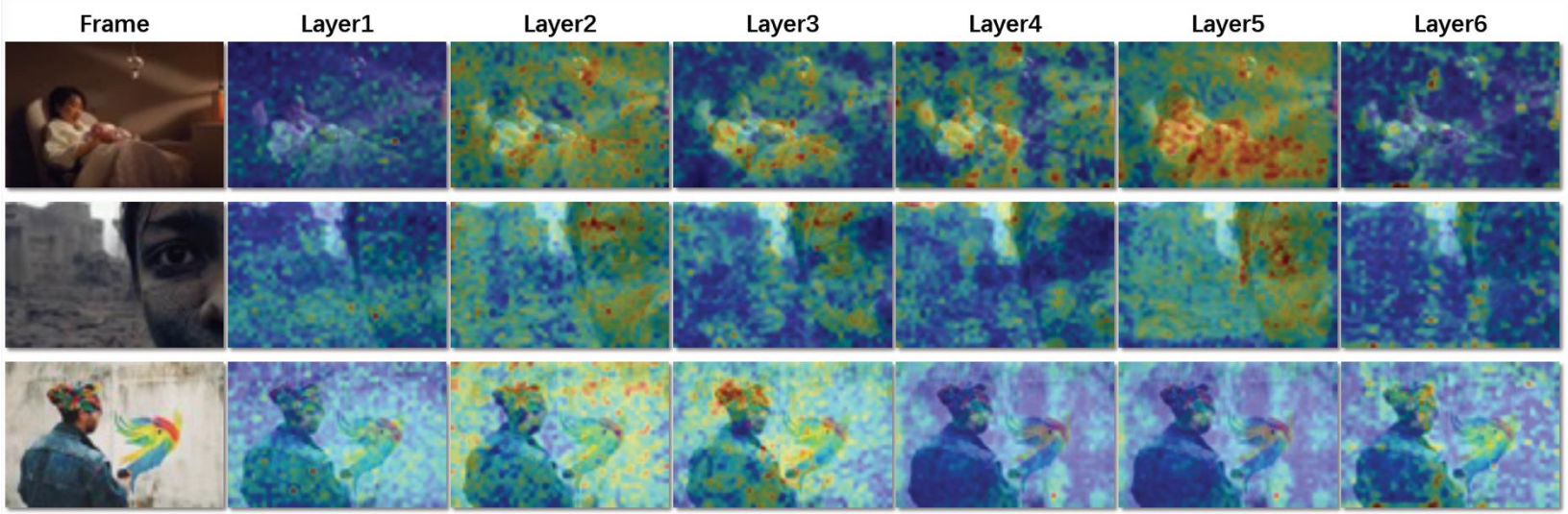
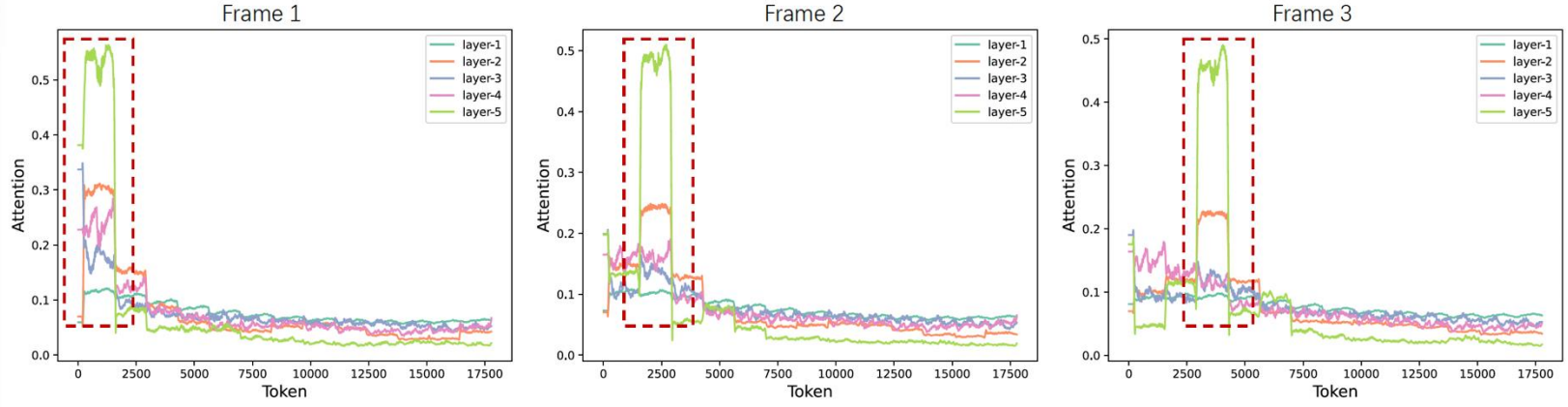
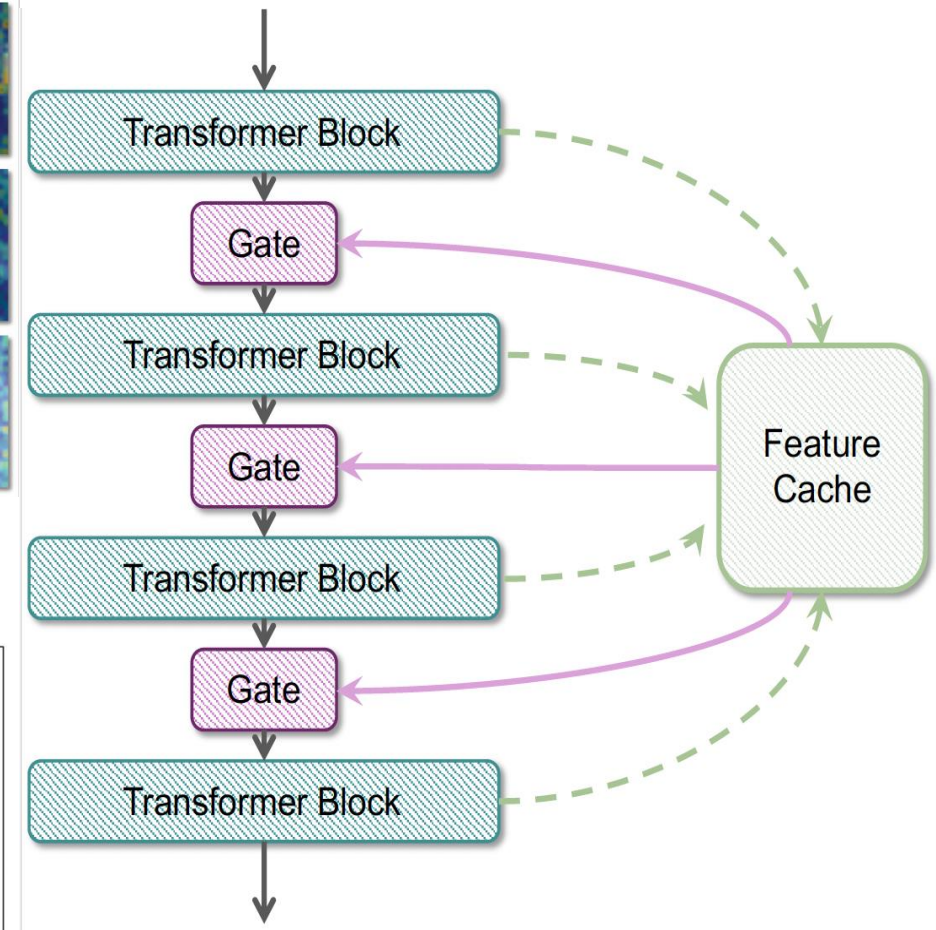


(a) Attention maps show diverse spatial features across layers but lack inter-layer coordination, leading to fragmented representations and weaker spatial semantics per frame.



(b) Attention distributions vary across layers, with deeper layers focusing on same-frame tokens while weakening attention to others.



(c) The architecture of the enhanced cross-layer representation framework.