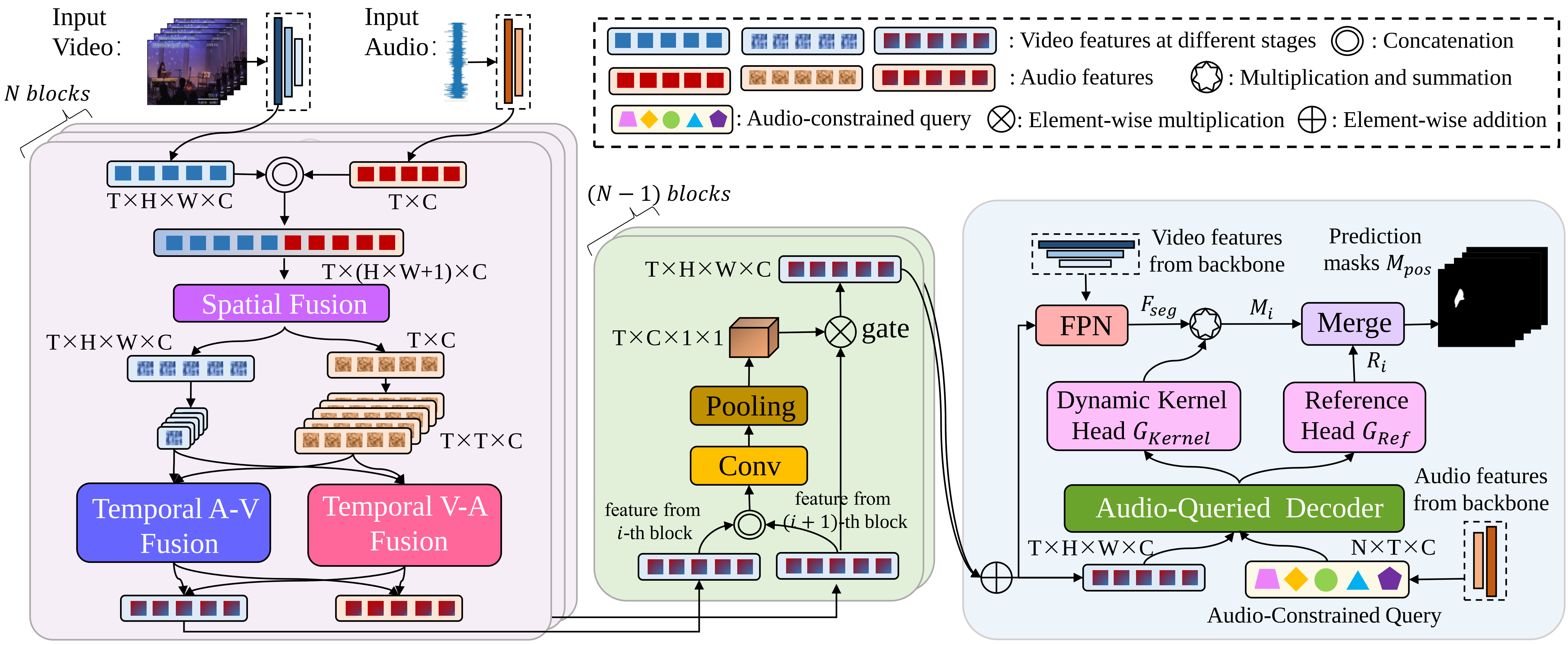

(a) Decoupled Audio-Visual Transformer

(b) Blockwise-Encoded Gate

(c) Audio-Queried Decoding