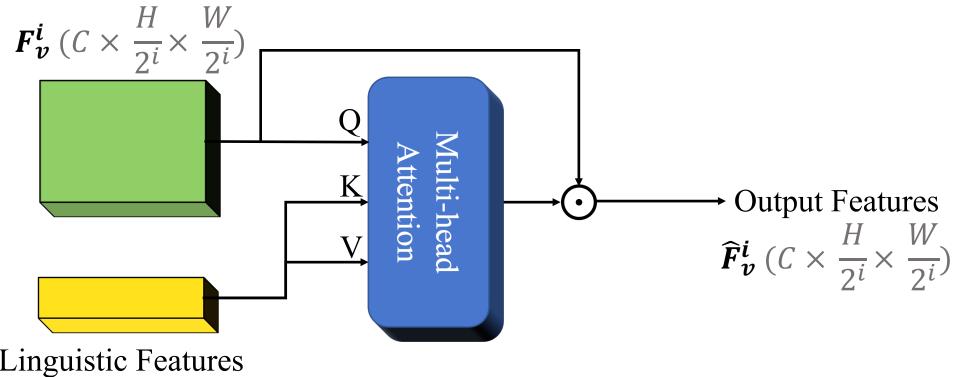Visual Features

$\boldsymbol{F_v^i}\ (C \times \frac{H}{2^i} \times \frac{W}{2^i})$

Q

K

V

Multi-head Attention

Output Features

$\widehat{\boldsymbol{F}}_v^i\ (C \times \frac{H}{2^i} \times \frac{W}{2^i})$

Linguistic Features

$\boldsymbol{F_t}\ (C \times L)$

⊙ Element-wise multiplication