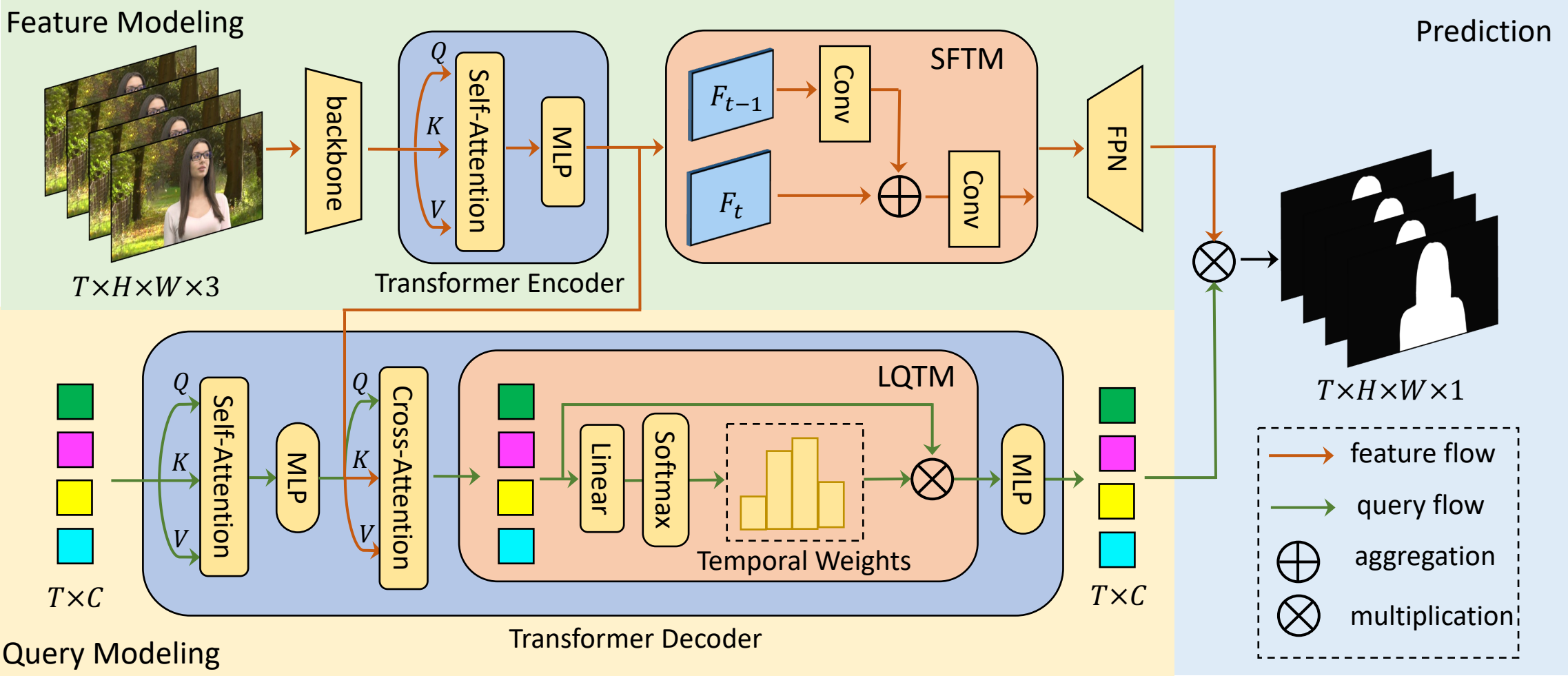
Feature Modeling
Prediction
backbone
Q
K
V
Self-Attention
MLP
Transformer Encoder
SFTM
F_{t-1}
F_t
Conv
Conv
FPN
$T\times H\times W\times 3$
$T\times H\times W\times 1$
LQTM
Linear
Softmax
Temporal Weights
Cross-Attention
$T\times C$
$T\times C$
Transformer Decoder
Query Modeling
feature flow
query flow
aggregation
multiplication