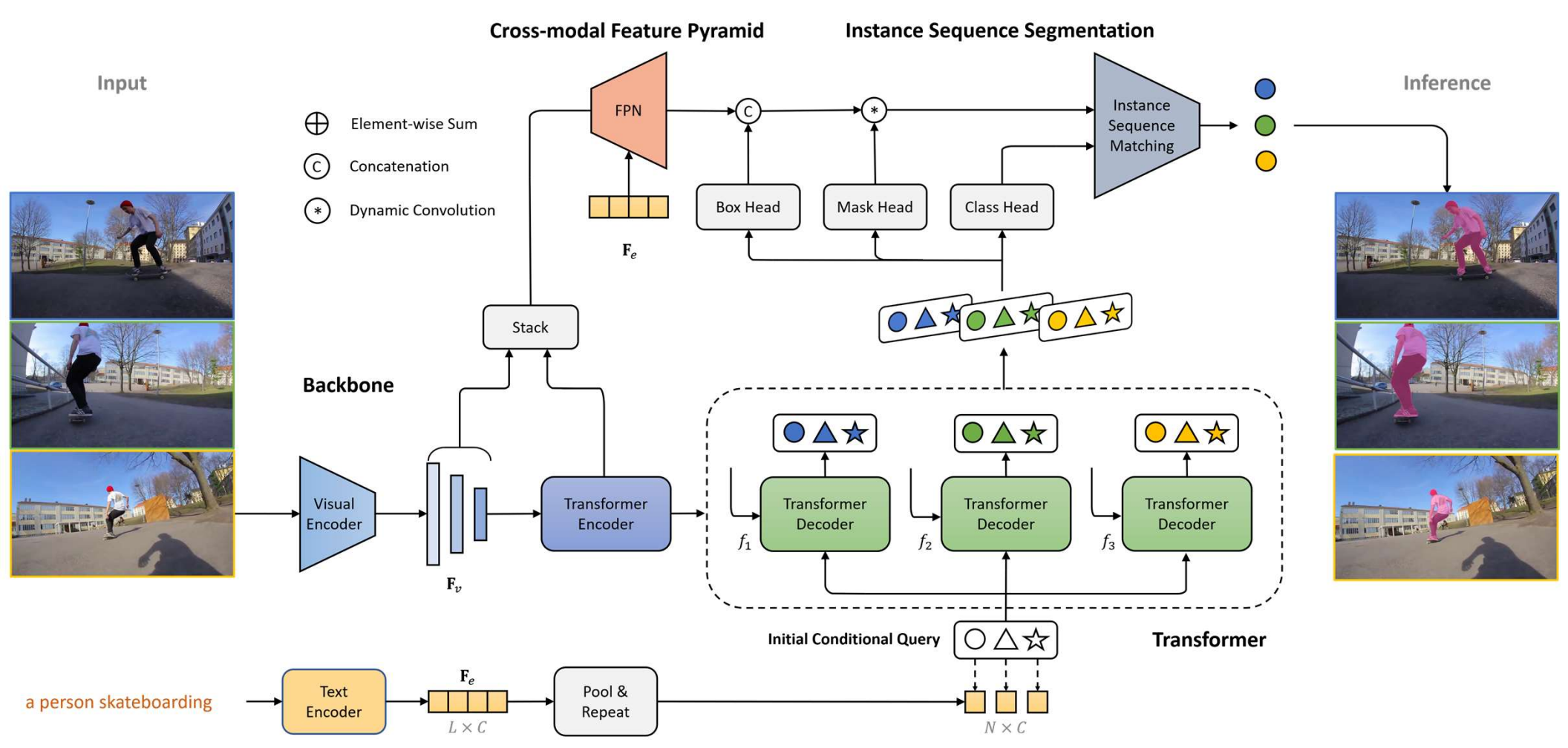
Cross-modal Feature Pyramid
Instance Sequence Segmentation
Input
Inference
Element-wise Sum
Concatenation
Dynamic Convolution
FPN
Box Head
Mask Head
Class Head
Instance Sequence Matching
$\mathbf{F}_e$
Stack
Backbone
Visual Encoder
$\mathbf{F}_v$
Transformer Encoder
Transformer Decoder
Transformer Decoder
Transformer Decoder
f_1
f_2
f_3
Initial Conditional Query
Transformer
a person skateboarding
Text Encoder
$\mathbf{F}_e$
$L \times C$
Pool & Repeat
$N \times C$