

# Autoregressive Transformer

features  
 $f_1, \dots, f_L$

K,V

Masked Multi-Head Cross-Attention

Add, Norm

FFN

Add, Norm

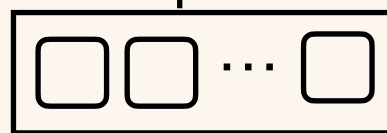
Masked Multi-Head Self-Attention

$\times N_d$

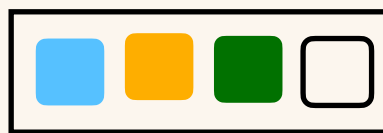
V

K

Q



Begin



$h_1$



$h_2$

...

$h_j$