



Token sequences

Input/output  
( $N$  tokens)

Attention Operators

Latent  
( $M$  tokens)

Self attention  
( $M \rightarrow M$ )

Cross attention  
( $N \rightarrow M$ )

Encoder ( $N \rightarrow M$ )/ decoder ( $M \rightarrow N$ )

Feature projector

Cross attn projector