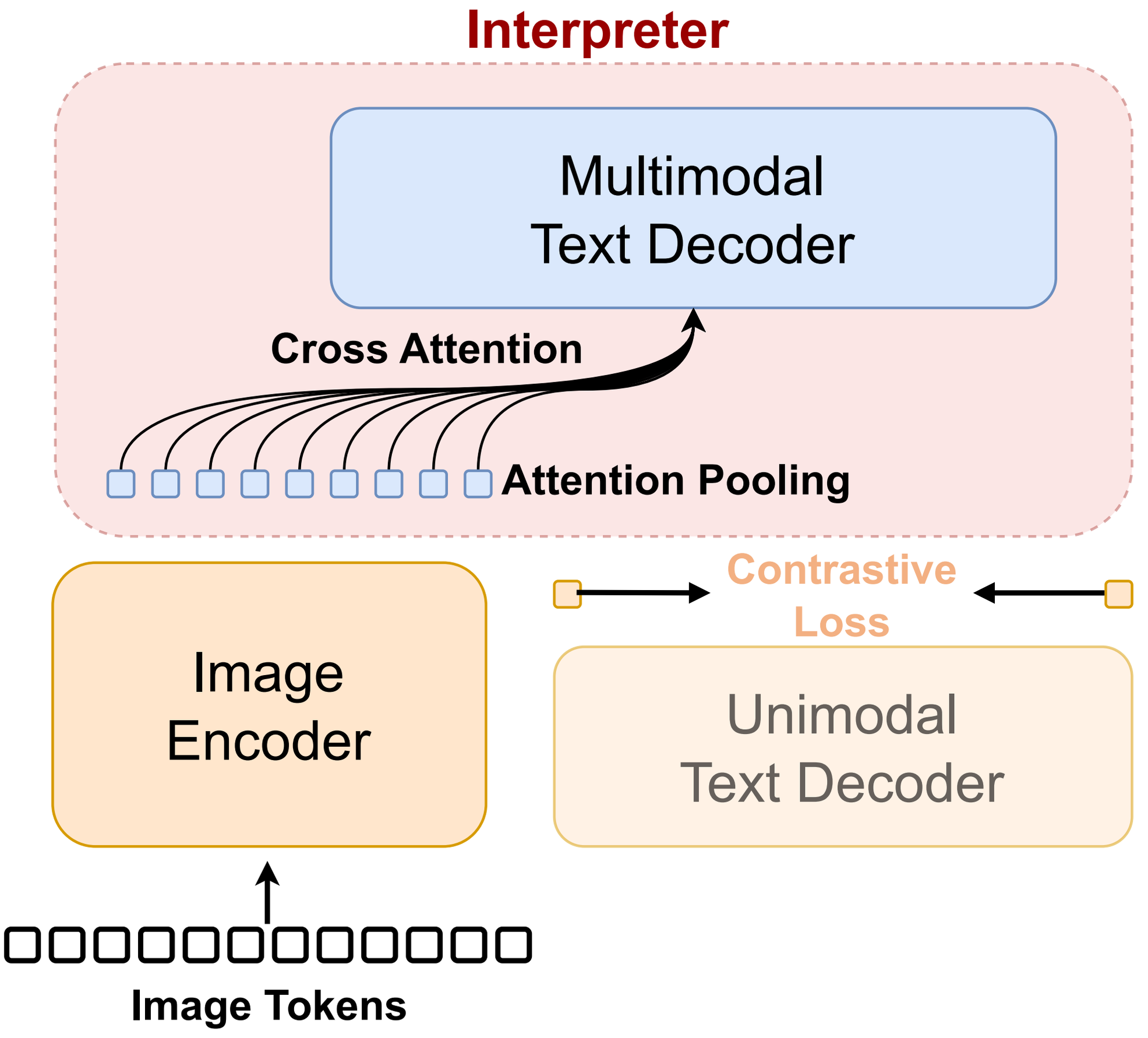
Interpreter
Multimodal
Text Decoder
Cross Attention
Attention Pooling
Contrastive
Loss
Image
Encoder
Unimodal
Text Decoder
Image Tokens