

Interpreter

Multimodal
Text Decoder

Cross Attention

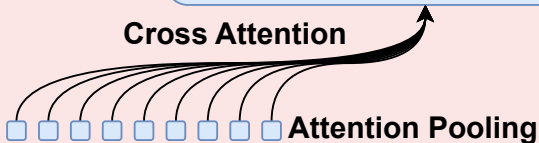


Image
Encoder

Contrastive
Loss

Unimodal
Text Decoder

Image Tokens

The diagram shows a sequence of ten small white squares with black outlines, representing the input image tokens. An upward-pointing arrow indicates that these tokens are fed into the 'Image Encoder'.