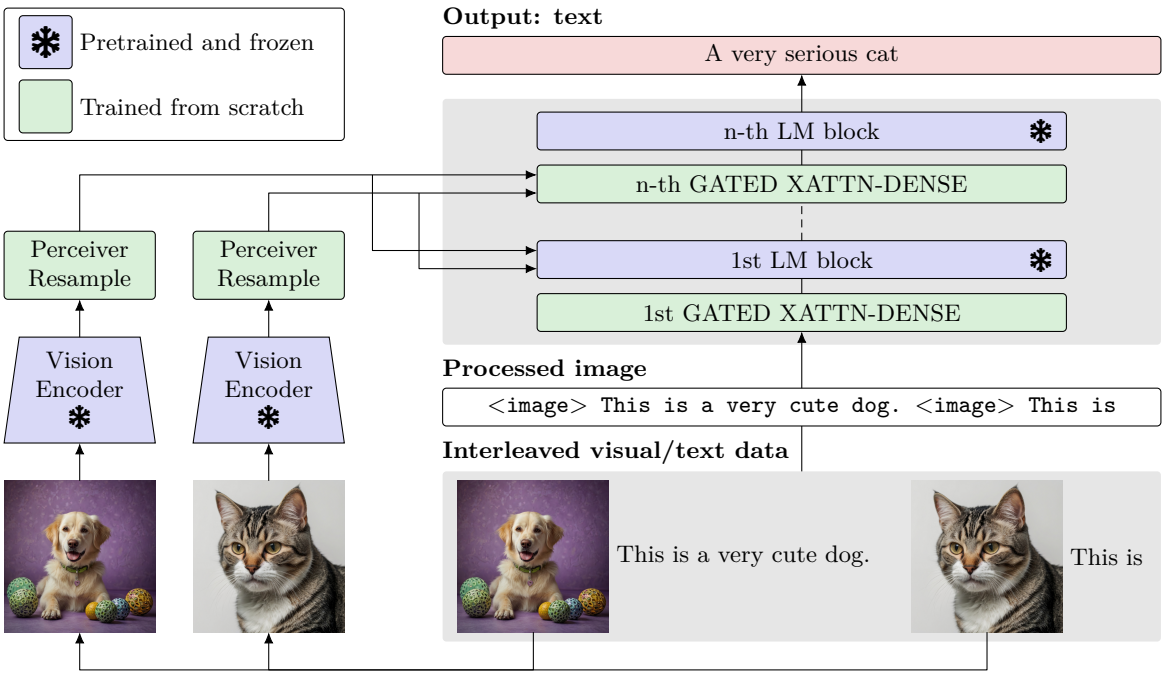
Pretrained and frozen
Trained from scratch
Output: text
A very serious cat
n-th LM block
n-th GATED XATTN-DENSE
1st LM block
1st GATED XATTN-DENSE
Perceiver Resample
Perceiver Resample
Vision Encoder
Vision Encoder
Processed image
<image> This is a very cute dog. <image> This is
Interleaved visual/text data
This is a very cute dog.
This is