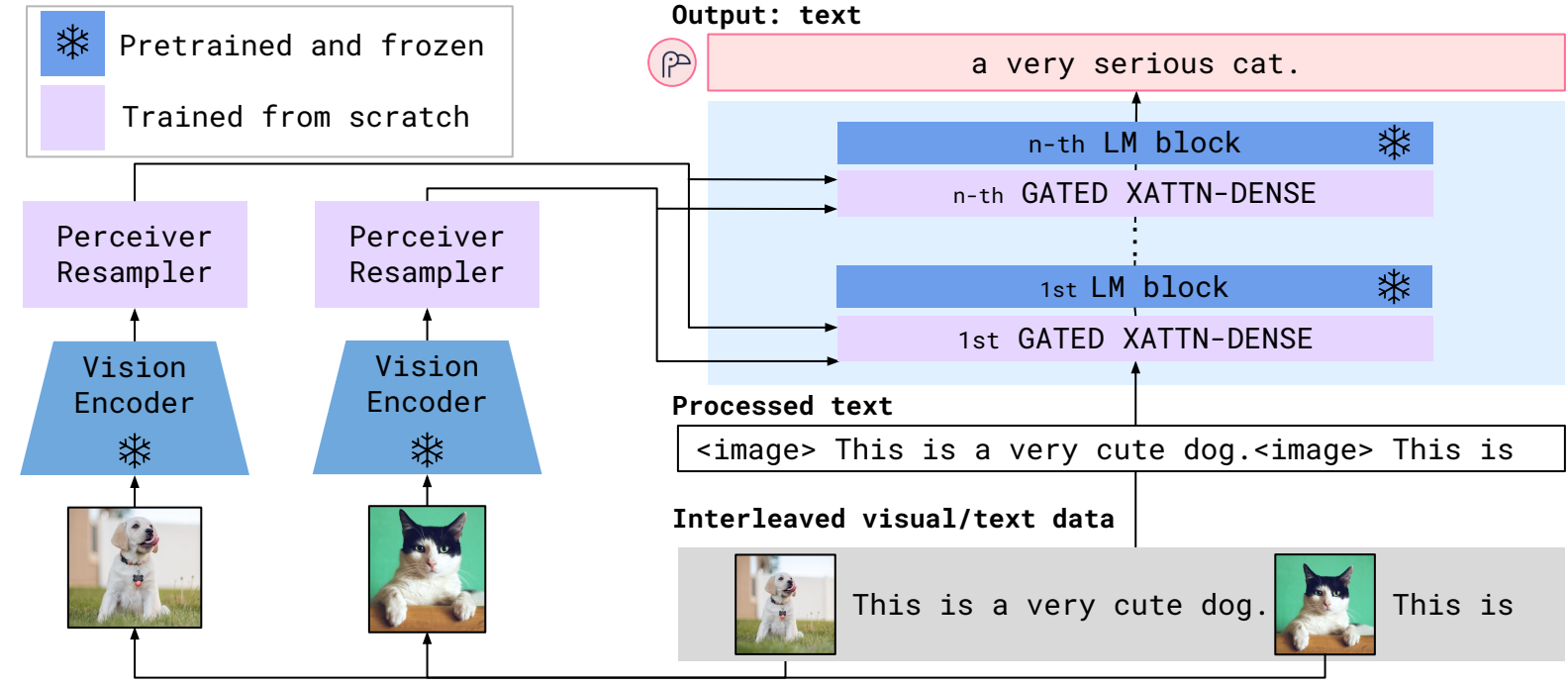
Pretrained and frozen
Trained from scratch
Output: text
a very serious cat.
n-th LM block
n-th GATED XATTN-DENSE
1st LM block
1st GATED XATTN-DENSE
Perceiver Resampler
Perceiver Resampler
Vision Encoder
Vision Encoder
Processed text
<image> This is a very cute dog.<image> This is
Interleaved visual/text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