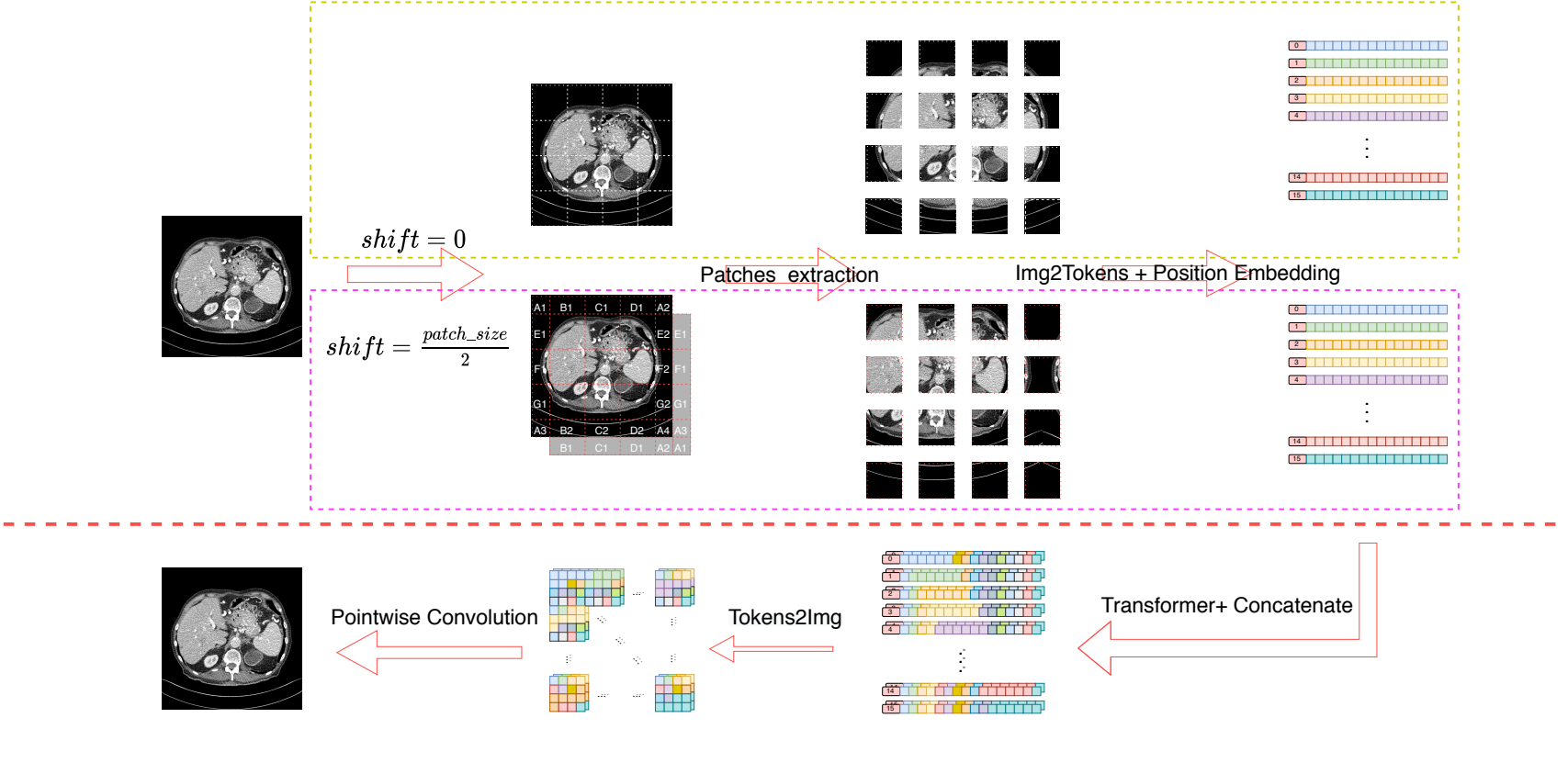
$shift = 0$
$shift = \frac{patch_size}{2}$
Patches_extraction
Img2Tokens + Position Embedding
Transformer+ Concatenate
Tokens2Img
Pointwise Convolution