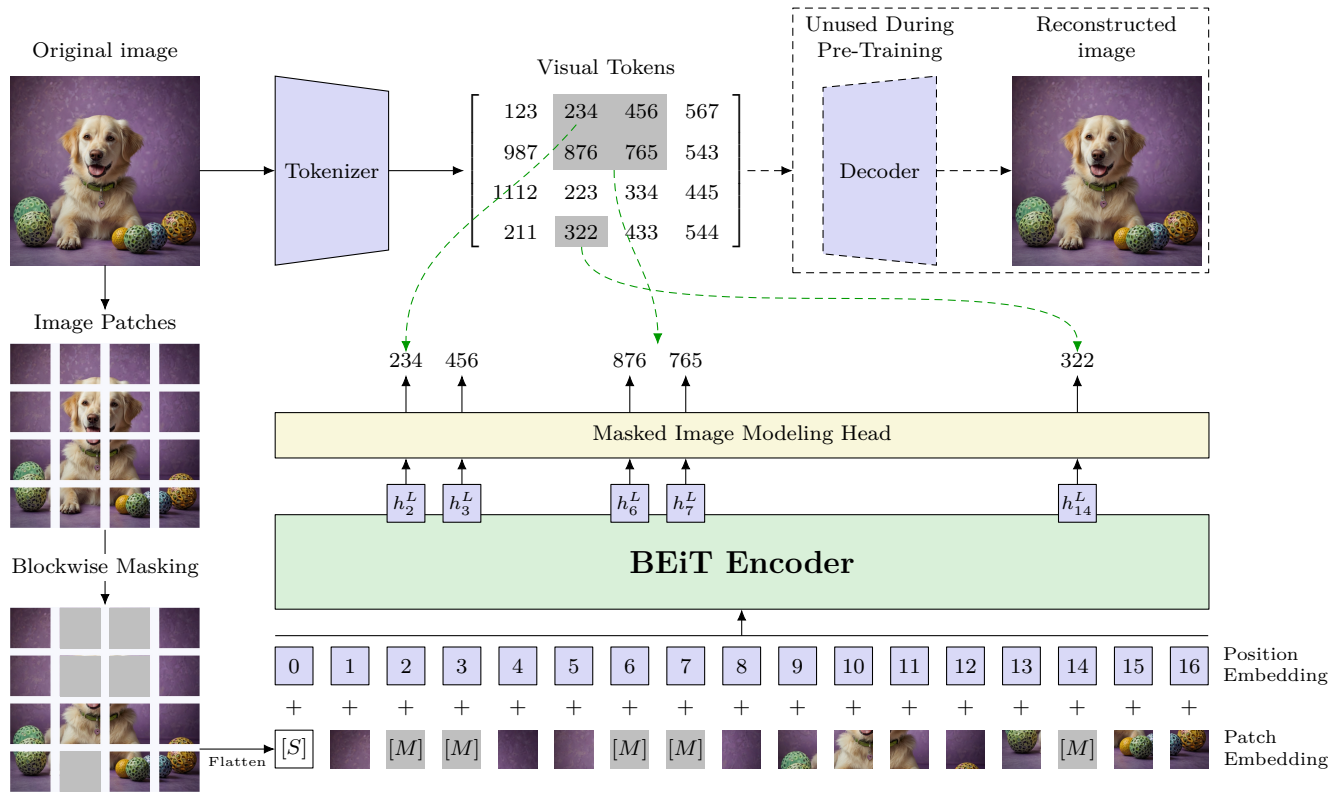
Original image
Tokenizer
Visual Tokens
123 234 456 567
987 876 765 543
1112 223 334 445
211 322 433 544
Unused During Pre-Training
Decoder
Reconstructed image
Image Patches
Blockwise Masking
234 456 876 765 322
Masked Image Modeling Head
h_2^L h_3^L h_6^L h_7^L h_{14}^L
BEiT Encoder
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
Position Embedding
Flatten
[S] [M] [M] [M] [M] [M]
Patch Embedding