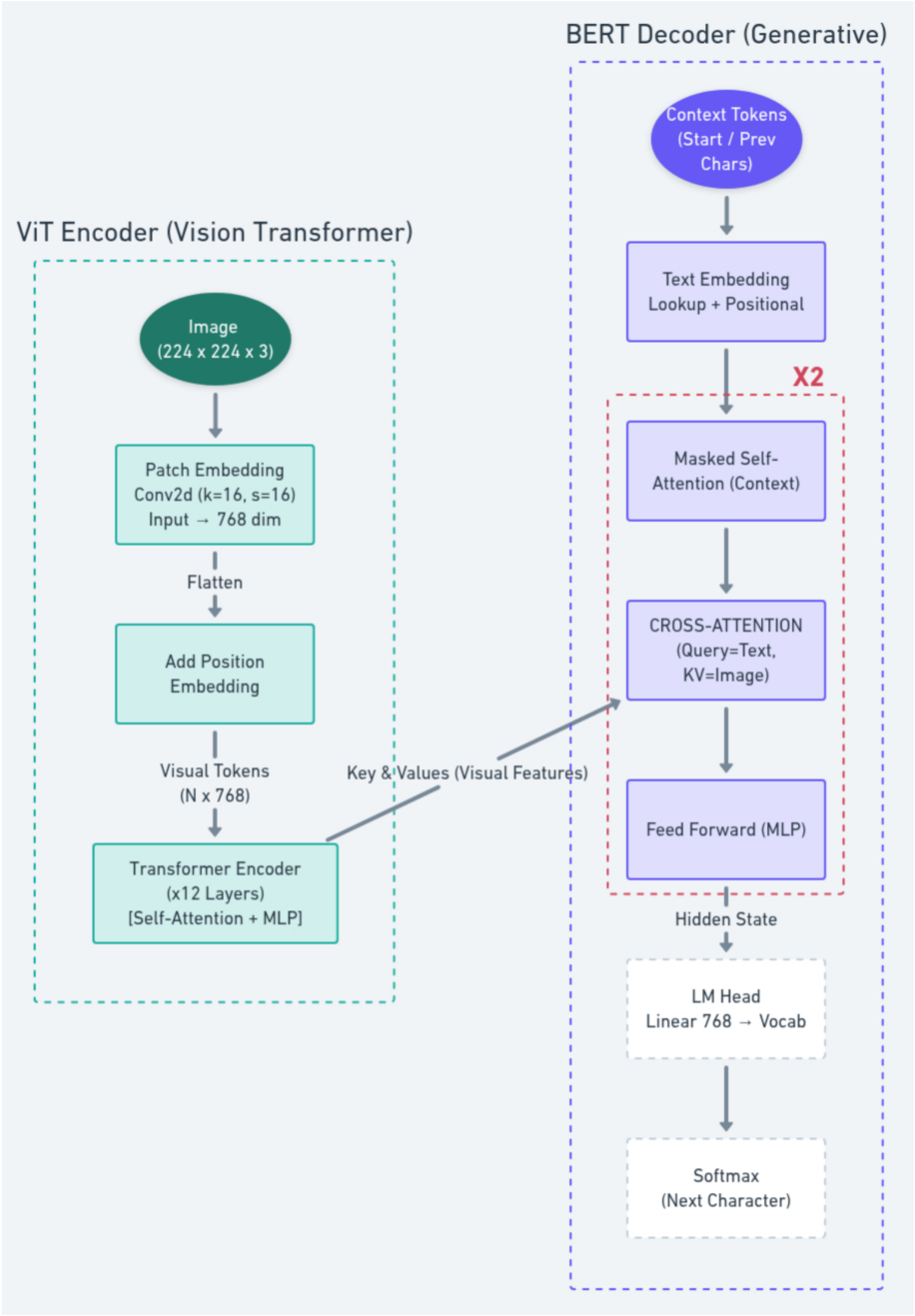
BERT Decoder (Generative)
Context Tokens (Start / Prev Chars)
Text Embedding Lookup + Positional
X2
Masked Self-Attention (Context)
CROSS-ATTENTION (Query=Text, KV=Image)
Feed Forward (MLP)
Hidden State
LM Head Linear 768 → Vocab
Softmax (Next Character)
ViT Encoder (Vision Transformer)
Image (224 x 224 x 3)
Patch Embedding Conv2d (k=16, s=16) Input → 768 dim
Flatten
Add Position Embedding
Visual Tokens (N x 768)
Transformer Encoder (x12 Layers) [Self-Attention + MLP]
Key & Values (Visual Features)