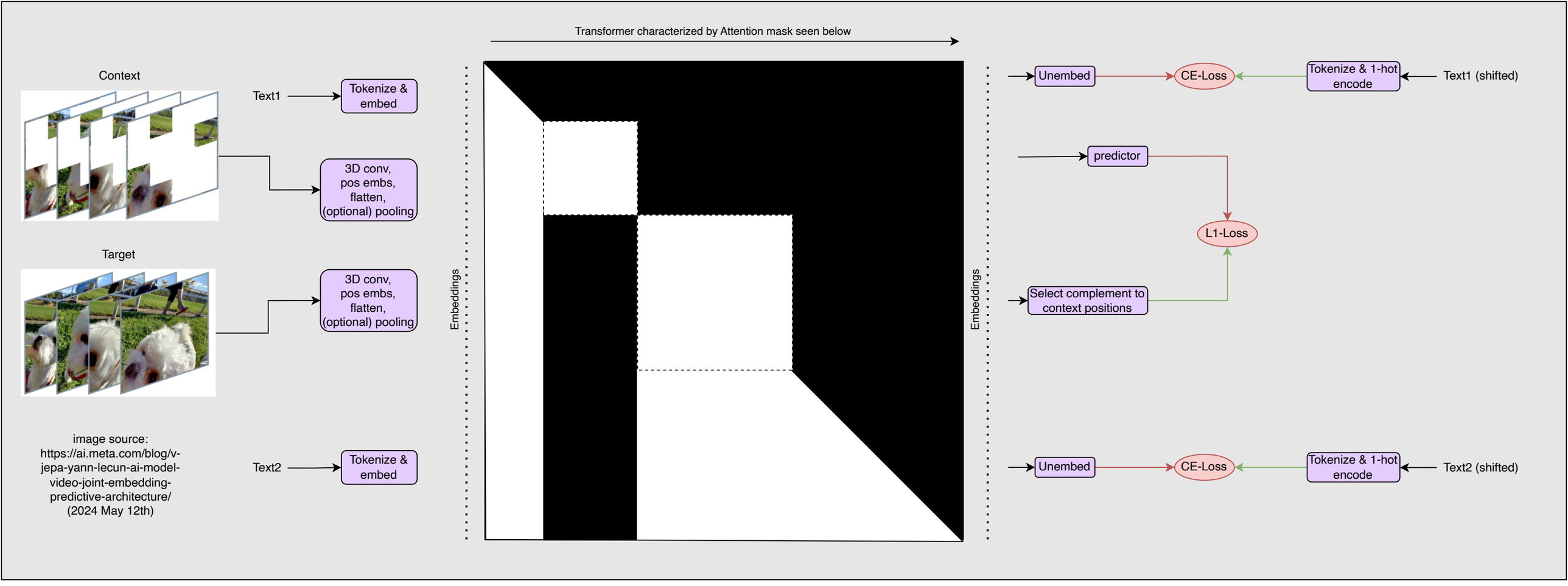
Transformer characterized by Attention mask seen below
Context
Target
Text1
Tokenize & embed
3D conv, pos embs, flatten, (optional) pooling
3D conv, pos embs, flatten, (optional) pooling
Text2
Tokenize & embed
image source: https://ai.meta.com/blog/v-jepa-yann-lecun-ai-model-video-joint-embedding-predictive-architecture/ (2024 May 12th)
Embeddings
Embeddings
Unembed
CE-Loss
Tokenize & 1-hot encode
Text1 (shifted)
predictor
L1-Loss
Select complement to context positions
Unembed
CE-Loss
Tokenize & 1-hot encode
Text2 (shifted)