FCFS Beam Decoding with Controlled Patience

k: beam size, M: maximum length, $\mathcal{V}$: Vocabulary
score$(\cdot)$: scoring function, p: patience factor.

$B_0 \leftarrow \{\langle 0, \text{BOS} \rangle\}$, $F_0 \leftarrow \varnothing$
for $t \in \{1, \ldots, M-1\}$:
  $H \leftarrow \varnothing$, $F_t \leftarrow F_{t-1}$
  for $\langle s, \mathbf{y} \rangle \in B_{t-1}$: # Expansion.
    for $y \in \mathcal{V}$:
      $s \leftarrow \text{score}(\mathbf{y} \circ y)$, $H.\text{add}(\langle s, \mathbf{y} \circ y \rangle)$
  $B_t \leftarrow \varnothing$
  while $|B_t| < k$: # Find top k w/o EOS from H.
    $\langle s, \mathbf{y} \rangle \leftarrow H.\max()$
    if $\mathbf{y}.\text{last}() = \text{EOS}$:
      $F_t.\text{add}(\langle s, \mathbf{y} \rangle)$ # Finished hypotheses.
    else $B_t.\text{add}(\langle s, \mathbf{y} \rangle)$
    if $|F_t| \geq k \cdot p$: # Originally, $p=1$.
      return $F_t.\max()$
    $H.\text{remove}(\langle s, \mathbf{y} \rangle)$
return $F_t.\max()$