Algorithm 1 Deterministic Hybrid Layout-OCR Fusion

Require: Set of AWS layout blocks B; ordered list of Google OCR word tokens T; tolerance $\tau = 0.02$
Ensure: Structured document object D

for all $b \in B$ **do**
  b.coords $\leftarrow [\text{Left}(b), \text{Top}(b), \text{Right}(b), \text{Bottom}(b)]$
  b.area $\leftarrow (\text{Right}(b) - \text{Left}(b)) \times (\text{Bottom}(b) - \text{Top}(b))$
  b.matched $\leftarrow \emptyset$
end for
Sort B in ascending order of b.area
$U \leftarrow \emptyset$ ▷ Used word indices
for $i \leftarrow 1$ **to** $|T|$ **do**
  $(x, y) \leftarrow \text{Centroid}(T[i])$
  for all $b \in B$ **do**
    if $\text{Left}(b) - \tau \leq x \leq \text{Right}(b) + \tau$ **and** $\text{Top}(b) - \tau \leq y \leq \text{Bottom}(b) + \tau$ **then**
      b.matched $\leftarrow b$.matched $\cup \{i\}$
      $U \leftarrow U \cup \{i\}$
      break
    end if
  end for
end for
$R \leftarrow \emptyset$ ▷ Reconstructed regions
for all $b \in B$ **do**
  $I \leftarrow \text{SortAscending}(b.\text{matched})$
  if $I = \emptyset$ **then**
    $R \leftarrow R \cup \{\text{Region}(b.\text{label}, b.\text{coords}, \epsilon, \infty)\}$
    continue
  end if
  $C \leftarrow \text{SplitIntoContiguousRuns}(I)$
  for all $c \in C$ **do**
    $W \leftarrow \{T[k] \mid k \in c\}$
    $S \leftarrow \text{JoinWordsUsingBreakTypes}(W)$
    $R \leftarrow R \cup \{\text{Region}(b.\text{label}, b.\text{coords}, S, \min(c))\}$
  end for
end for
$O \leftarrow \{i \mid i \in [1, |T|] \wedge i \notin U\}$
if $O \neq \emptyset$ **then**
  $C_o \leftarrow \text{SplitIntoContiguousRuns}(O)$
  for all $c \in C_o$ **do**
    $W \leftarrow \{T[k] \mid k \in c\}$
    $S \leftarrow \text{JoinWordsUsingBreakTypes}(W)$
    $(\bar{x}, \bar{y}) \leftarrow \text{MeanCentroid}(W)$
    $R \leftarrow R \cup \{\text{Region}(\text{“Paragraph”}, [\bar{x}, \bar{y}, \bar{x}, \bar{y}], S, \min(c))\}$
  end for
end if
Sort R in ascending order of sort key
D.annotations $\leftarrow R$
D.wordsFound $\leftarrow |T|$
return D