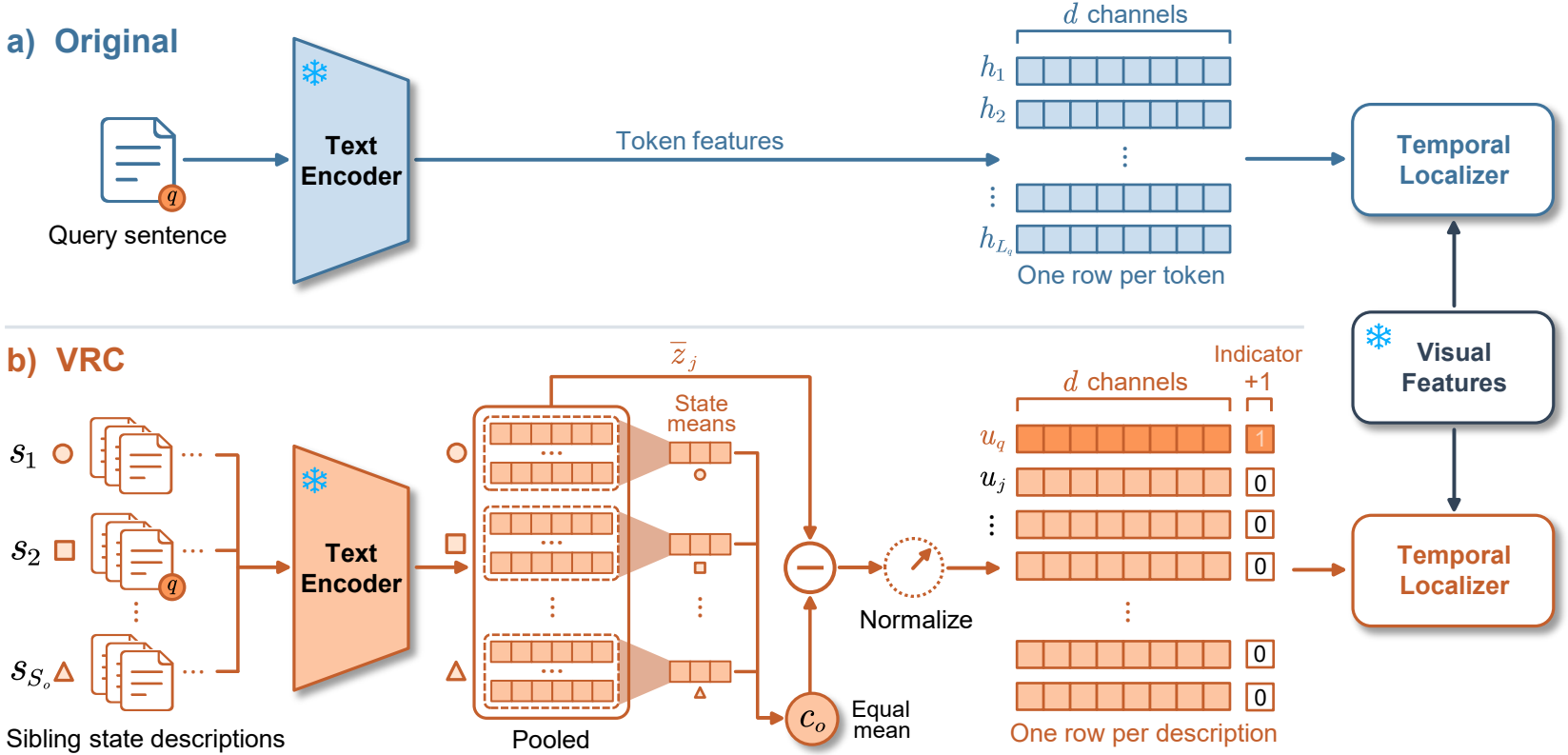

a) Original
Query sentence
Text Encoder
Token features
d channels
h_1
h_2
h_{L_q}
One row per token
Temporal Localizer
Visual Features
b) VRC
s_1
s_2
s_{S_o}
Sibling state descriptions
Text Encoder
$\overline{z}_j$
State means
Pooled
c_o
Equal mean
Normalize
Indicator
+1
d channels
u_q
u_j
One row per description
Temporal Localizer