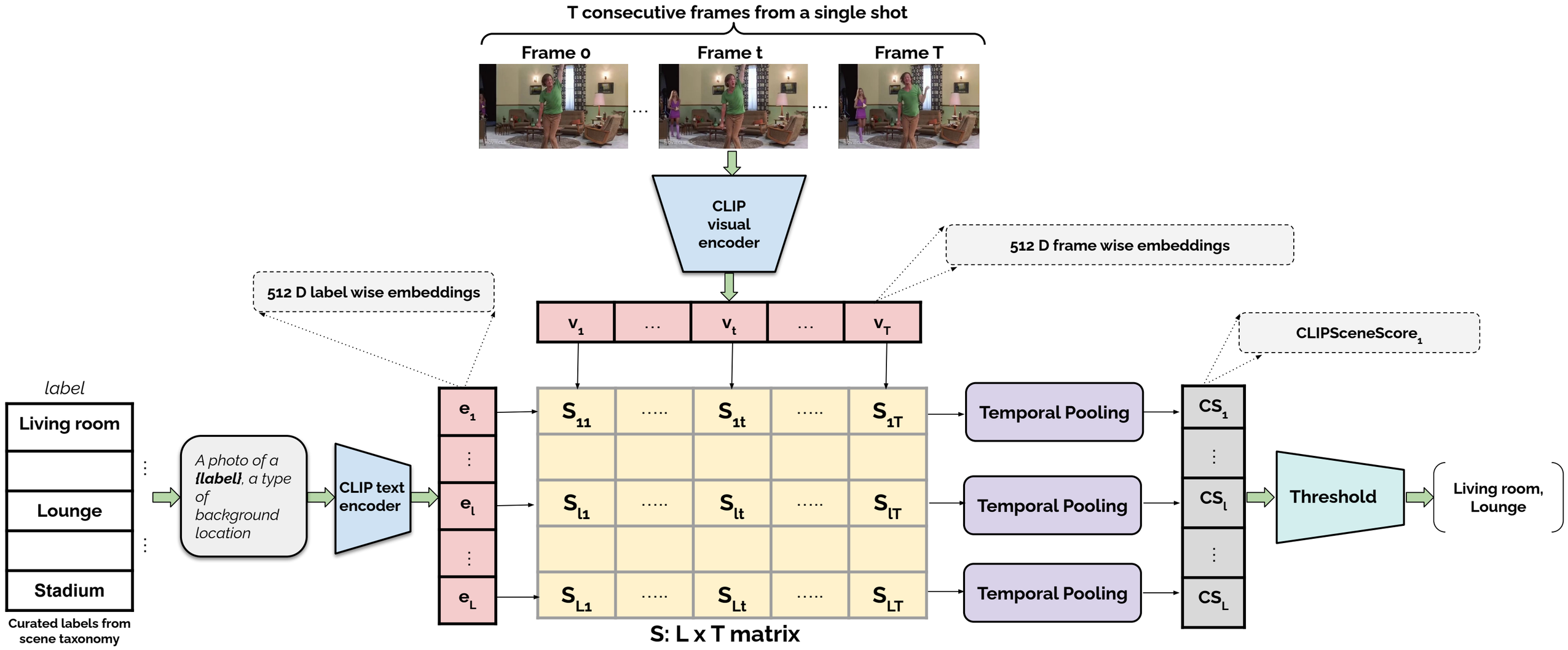
T consecutive frames from a single shot
Frame 0
Frame t
Frame T
CLIP visual encoder
512 D frame wise embeddings
512 D label wise embeddings
v_1 … v_t … v_T
CLIPSceneScore$_1$
label
Living room
Lounge
Stadium
Curated labels from scene taxonomy
A photo of a {label}, a type of background location
CLIP text encoder
e_1 e_l e_L
S_{11} ….. S_{1t} ….. S_{1T}
S_{l1} ….. S_{lt} ….. S_{lT}
S_{L1} ….. S_{Lt} ….. S_{LT}
S: L x T matrix
Temporal Pooling
Temporal Pooling
Temporal Pooling
CS_1 CS_l CS_L
Threshold
Living room, Lounge