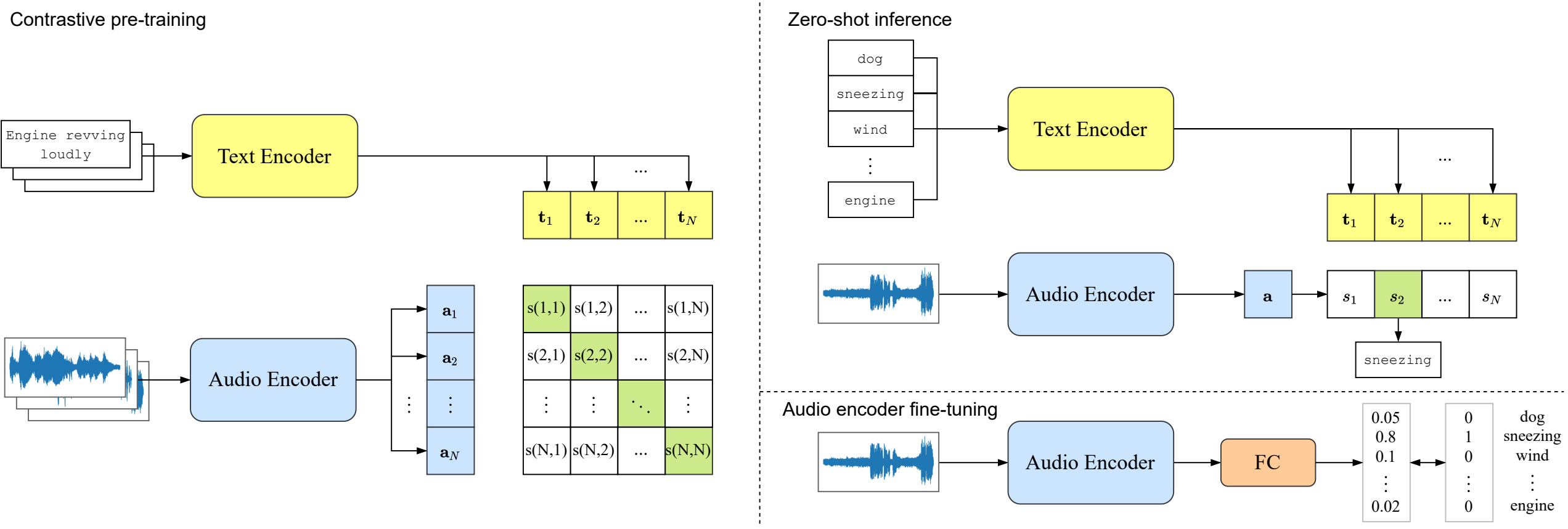
Contrastive pre-training
Engine revving loudly
Text Encoder
$\mathbf{t}_1$ $\mathbf{t}_2$ ... $\mathbf{t}_N$
Audio Encoder
$\mathbf{a}_1$ $\mathbf{a}_2$ $\mathbf{a}_N$
s(1,1) s(1,2) ... s(1,N)
s(2,1) s(2,2) ... s(2,N)
s(N,1) s(N,2) ... s(N,N)
Zero-shot inference
dog
sneezing
wind
engine
Text Encoder
$\mathbf{t}_1$ $\mathbf{t}_2$ ... $\mathbf{t}_N$
Audio Encoder
$\mathbf{a}$
s_1 s_2 ... s_N
sneezing
Audio encoder fine-tuning
Audio Encoder
FC
0.05
0.8
0.1
0.02
0
1
0
0
dog
sneezing
wind
engine