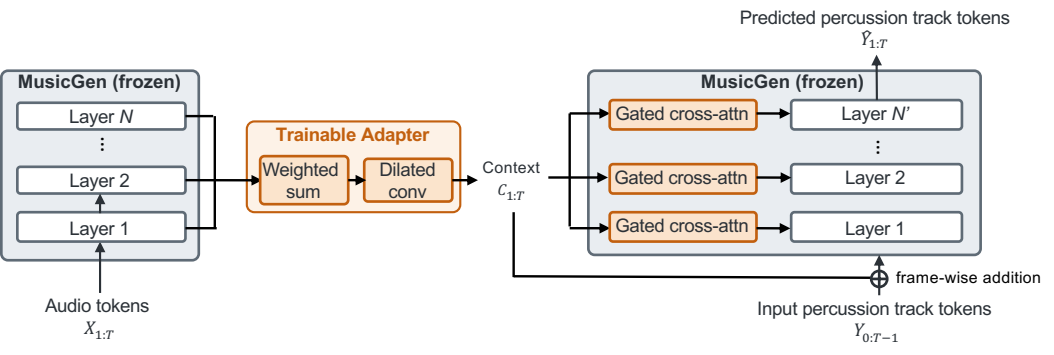

Predicted percussion track tokens
$\hat{Y}_{1:T}$
MusicGen (frozen)
Layer N
Layer 2
Layer 1
Audio tokens
$X_{1:T}$
Trainable Adapter
Weighted sum
Dilated conv
Context
$C_{1:T}$
MusicGen (frozen)
Gated cross-attn
Layer N'
Gated cross-attn
Layer 2
Gated cross-attn
Layer 1
frame-wise addition
Input percussion track tokens
$Y_{0:T-1}$