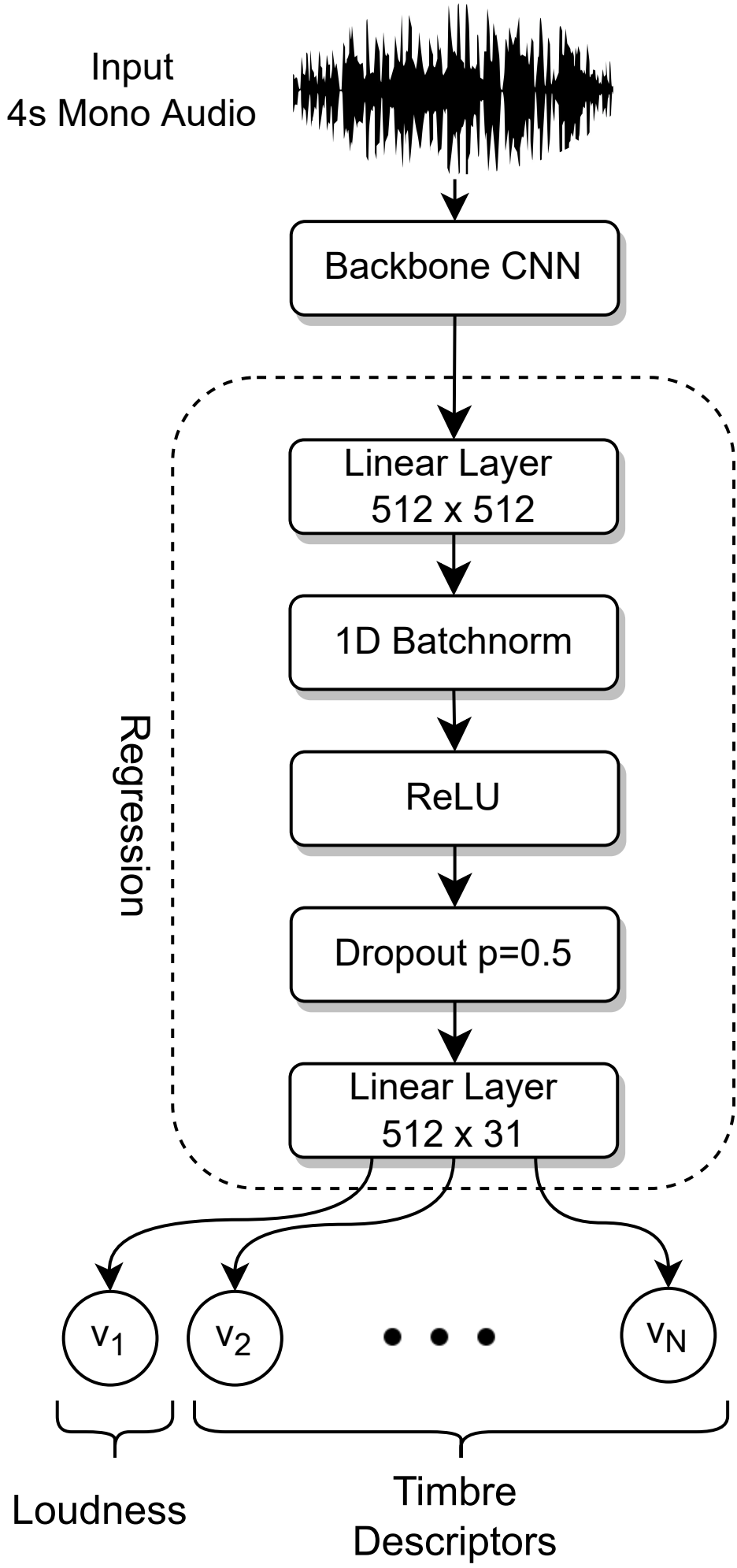
Input
4s Mono Audio
Backbone CNN
Linear Layer
512 x 512
1D Batchnorm
ReLU
Dropout p=0.5
Linear Layer
512 x 31
Regression
v1
v2
vN
Loudness
Timbre
Descriptors