Network Architecture (55 layers, 2,043,556,210 parameters)

Video Encoder
Audio Encoder
Text Encoder
Bottleneck
Decoder
Generic Layer

video_fc1
Linear, 176,951,296 params
43200 → 4096

video_bn1
BatchNorm1d, 8,192 params
4096 features

video_fc2
Linear, 8,390,656 params
4096 → 2048

video_bn2
BatchNorm1d, 4,096 params
2048 features

video_fc3
Linear, 2,098,176 params
2048 → 1024

video_bn3
BatchNorm1d, 2,048 params
1024 features

audio_fc1
Linear, 1,049,088 params
2048 → 512

audio_bn1
BatchNorm1d, 1,024 params
512 features

audio_fc2
Linear, 131,328 params
512 → 256

audio_bn2
BatchNorm1d, 512 params
256 features

text_fc1
Linear, 524,800 params
1024 → 512

text_bn1
BatchNorm1d, 1,024 params
512 features

text_fc2
Linear, 131,328 params
512 → 256

text_bn2
BatchNorm1d, 512 params
256 features

feature_conv.0
Linear, 2,360,832 params
1536 → 1536

feature_conv.1
BatchNorm1d, 3,072 params
1536 features

layer5.0
Linear, 6,295,552 params
1536 → 4096

layer5.1
BatchNorm1d, 8,192 params
4096 features

layer6.0
Linear, 32,776,000 params
4096 → 8000

layer6.1
BatchNorm1d, 16,000 params
8000 features

layer7_bottleneck.0
Linear, 16,386,048 params
8000 → 2048

layer7_bottleneck.1
BatchNorm1d, 4,096 params
2048 features

bottleneck_to_voxels.0
Linear, 33,570,816 params
2048 → 16384

bottleneck_to_voxels.1
BatchNorm1d, 32,768 params
16384 features

bottleneck_to_voxels.4
Linear, 1,405,996,850 params
16384 → 85810

layer8.0
Linear, 16,392,000 params
2048 → 8000

layer8.1
BatchNorm1d, 16,000 params
8000 features

layer9.0
Linear, 32,772,096 params
8000 → 4096

layer9.1
BatchNorm1d, 8,192 params
4096 features

layer10.0
Linear, 8,390,656 params
4096 → 2048

layer10.1
BatchNorm1d, 4,096 params
2048 features

layer11.0
Linear, 3,147,264 params
2048 → 1536

layer11.1
BatchNorm1d, 3,072 params
1536 features

layer12_video.0
Linear, 3,147,776 params
1536 → 2048

layer12_video.1
BatchNorm1d, 4,096 params
2048 features

layer12_video.4
Linear, 8,392,704 params
2048 → 4096

layer12_video.5
BatchNorm1d, 8,192 params
4096 features

layer12_audio.0
Linear, 3,147,776 params
1536 → 2048

layer12_audio.1
BatchNorm1d, 4,096 params
2048 features

layer12_audio.4
Linear, 4,196,352 params
2048 → 2048

layer12_audio.5
BatchNorm1d, 4,096 params
2048 features

audio_temporal_upsample.0
ConvTranspose1d, 16,779,264 params

audio_temporal_upsample.1
BatchNorm1d, 4,096 params
2048 features

audio_temporal_upsample.3
ConvTranspose1d, 16,779,264 params

audio_temporal_upsample.4
BatchNorm1d, 4,096 params
2048 features

audio_temporal_upsample.6
ConvTranspose1d, 16,779,264 params

audio_temporal_upsample.7
BatchNorm1d, 4,096 params
2048 features

audio_temporal_adjust
Conv1d, 12,584,960 params
2048 → 2048 channels

layer12_text.0
Linear, 1,573,888 params
1536 → 1024

layer12_text.1
BatchNorm1d, 2,048 params
1024 features

layer12_text.4
Linear, 1,049,600 params
1024 → 1024

layer12_text.5
BatchNorm1d, 2,048 params
1024 features

layer13_video.0
Linear, 176,990,400 params
4096 → 43200

layer13_audio
Linear, 33,570,816 params
2048 → 16384

layer13_text
Linear, 1,049,600 params
1024 → 1024

Layer sizes scaled logarithmically by parameter count
Largest layer: 1,405,996,850 parameters