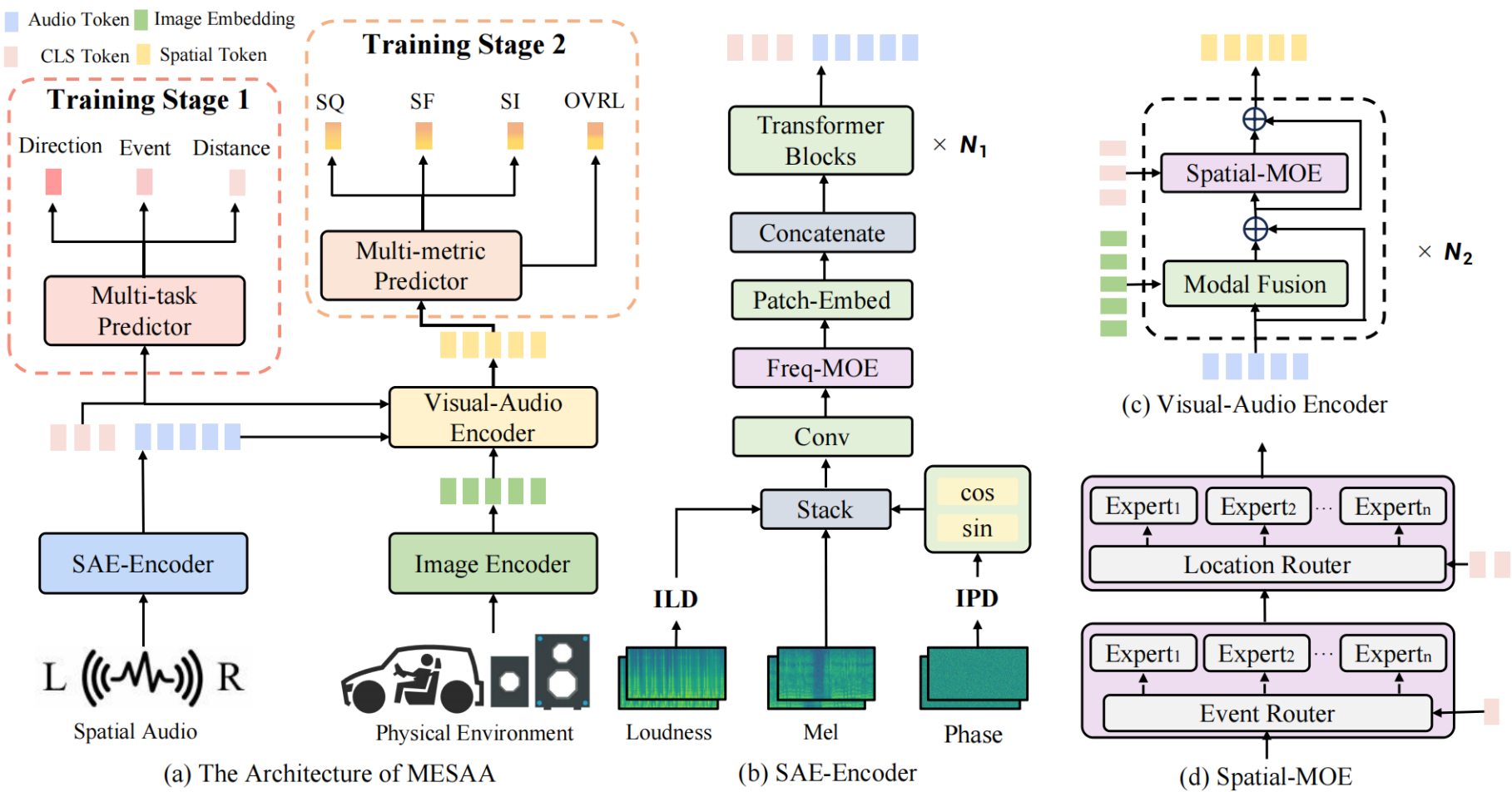
Audio Token
Image Embedding
CLS Token
Spatial Token
Training Stage 1
Direction
Event
Distance
Multi-task Predictor
Training Stage 2
SQ
SF
SI
OVRL
Multi-metric Predictor
Visual-Audio Encoder
SAE-Encoder
Image Encoder
L
R
Spatial Audio
Physical Environment
(a) The Architecture of MESAA
Transformer Blocks
× N_1
Concatenate
Patch-Embed
Freq-MOE
Conv
Stack
cos
sin
ILD
IPD
Loudness
Mel
Phase
(b) SAE-Encoder
Spatial-MOE
Modal Fusion
× N_2
(c) Visual-Audio Encoder
Expert1
Expert2
Expertn
Location Router
Event Router
(d) Spatial-MOE