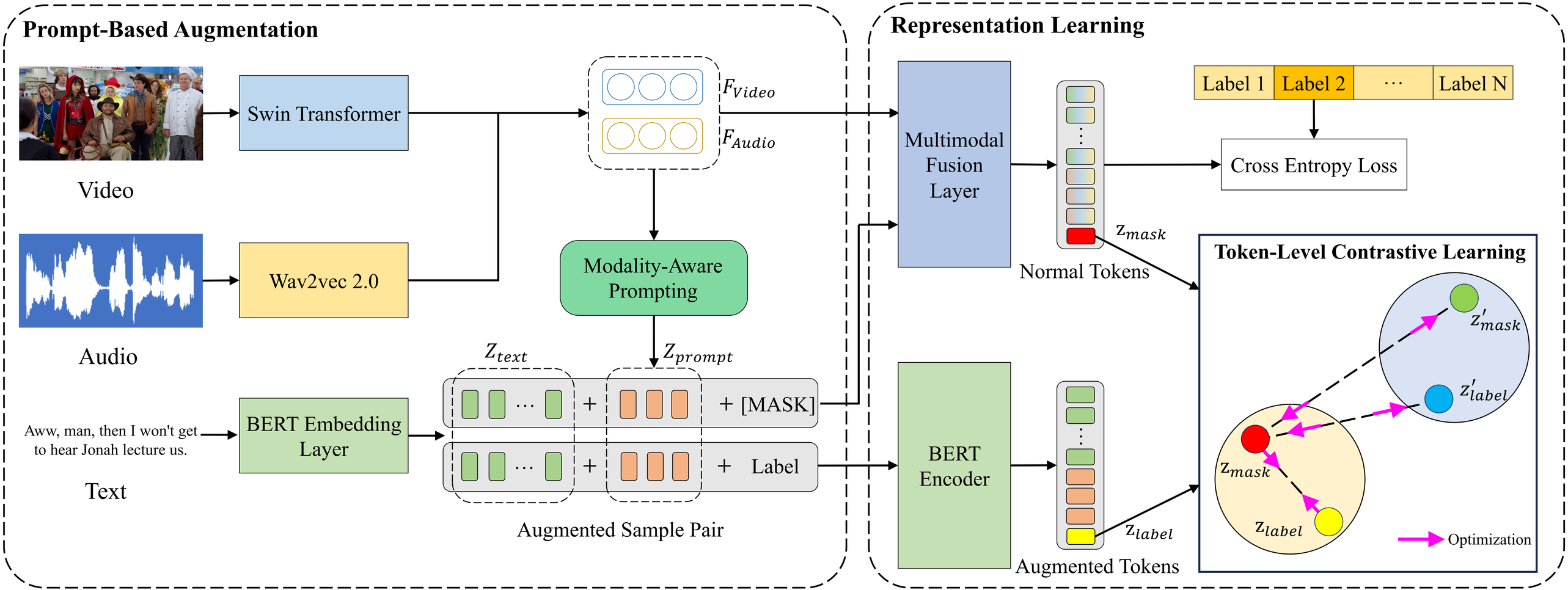
Prompt-Based Augmentation
Video
Swin Transformer
Audio
Wav2vec 2.0
Aww, man, then I won't get to hear Jonah lecture us.
Text
BERT Embedding Layer
F_{Video}
F_{Audio}
Modality-Aware Prompting
Z_{text}
Z_{prompt}
+ [MASK]
+ Label
Augmented Sample Pair
Representation Learning
Multimodal Fusion Layer
BERT Encoder
Normal Tokens
Augmented Tokens
z_{mask}
z_{label}
Label 1
Label 2
⋯
Label N
Cross Entropy Loss
Token-Level Contrastive Learning
z'_{mask}
z'_{label}
Optimization