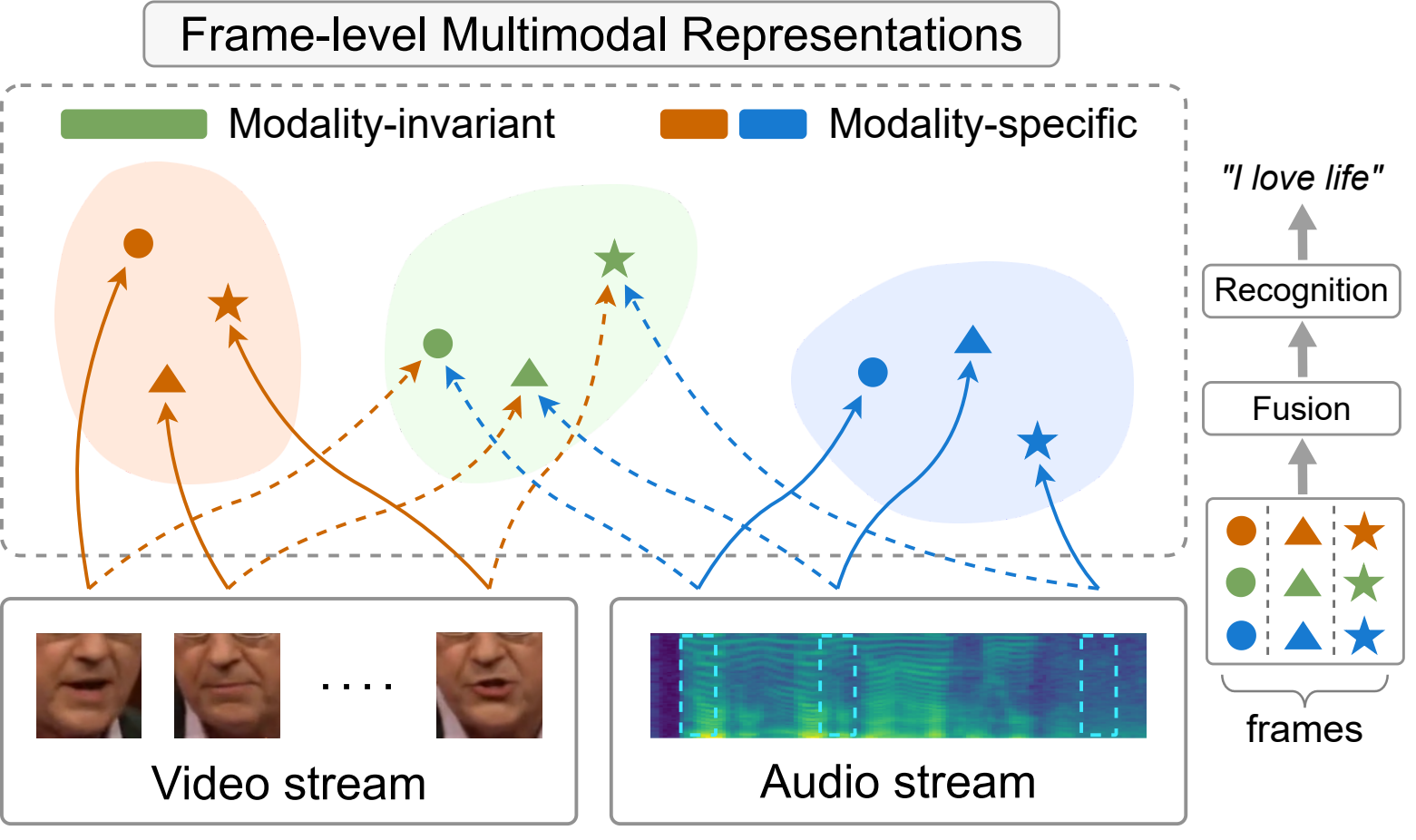
Frame-level Multimodal Representations
Modality-invariant
Modality-specific
"I love life"
Recognition
Fusion
frames
Video stream
Audio stream