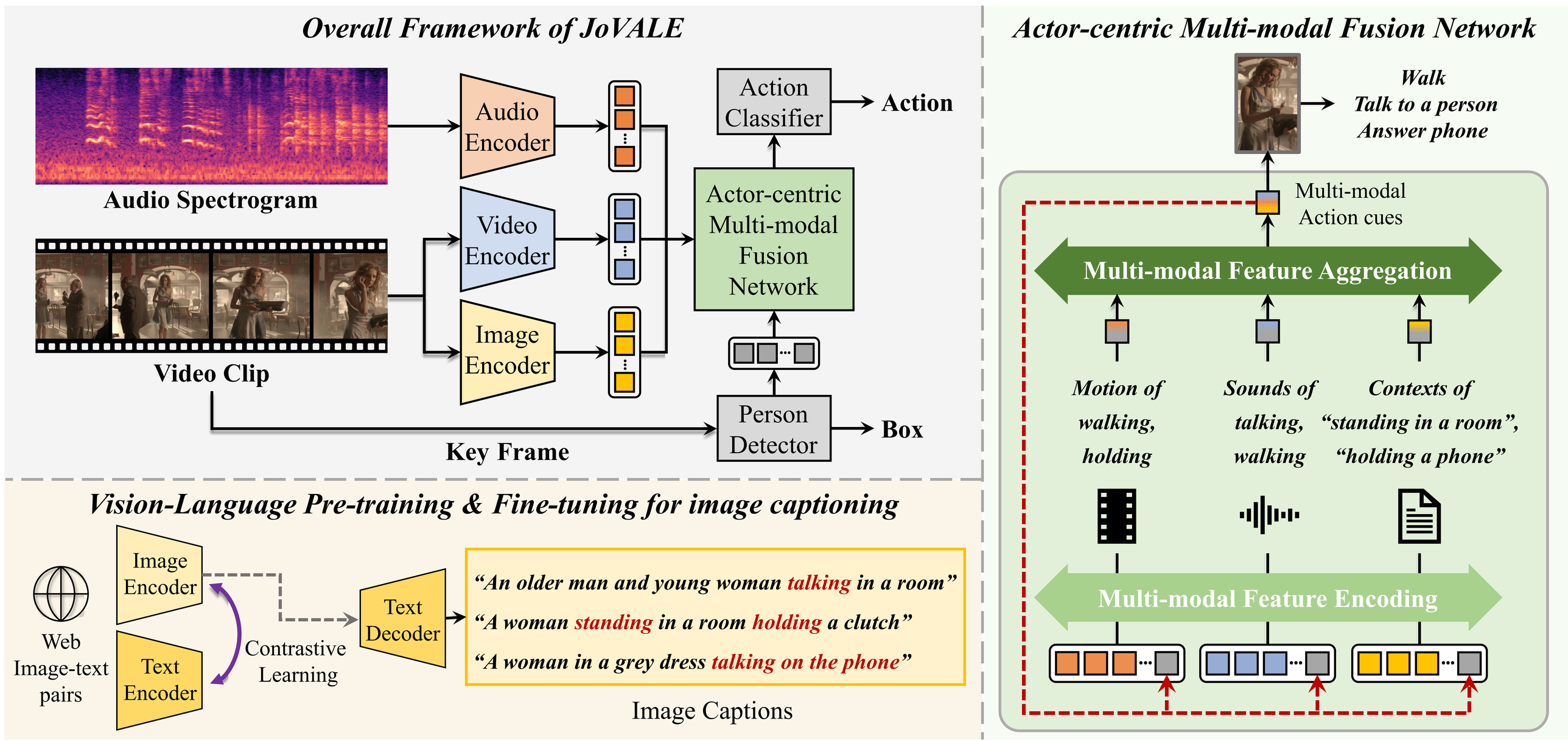
Overall Framework of JoVALE
Audio Spectrogram
Video Clip
Key Frame
Audio Encoder
Video Encoder
Image Encoder
Actor-centric Multi-modal Fusion Network
Action Classifier
Action
Person Detector
Box
Actor-centric Multi-modal Fusion Network
Walk
Talk to a person
Answer phone
Multi-modal Action cues
Multi-modal Feature Aggregation
Motion of walking, holding
Sounds of talking, walking
Contexts of "standing in a room", "holding a phone"
Multi-modal Feature Encoding
Vision-Language Pre-training & Fine-tuning for image captioning
Web Image-text pairs
Image Encoder
Text Encoder
Contrastive Learning
Text Decoder
"An older man and young woman talking in a room"
"A woman standing in a room holding a clutch"
"A woman in a grey dress talking on the phone"
Image Captions