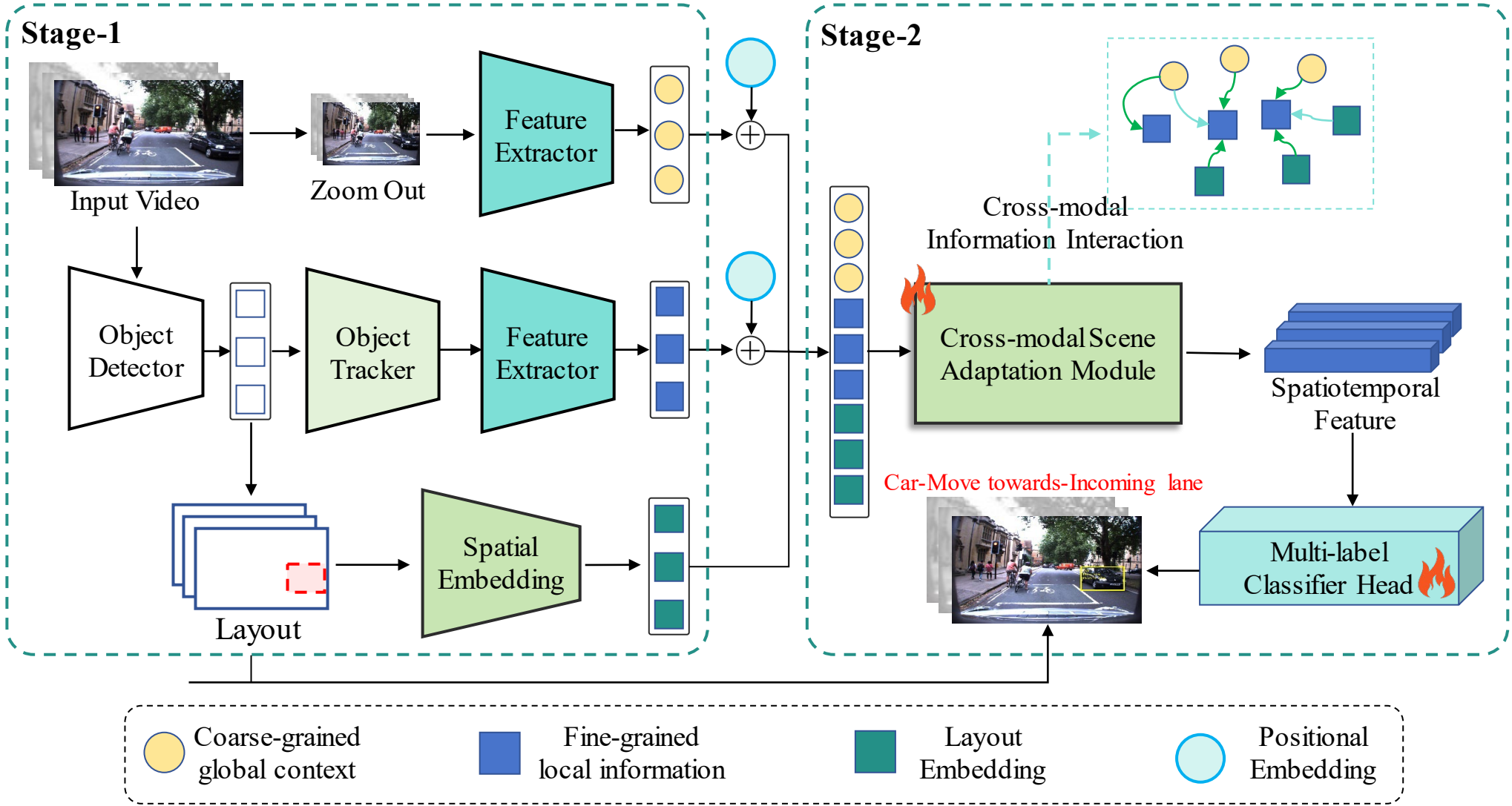
Stage-1
Input Video
Zoom Out
Feature Extractor
Object Detector
Object Tracker
Feature Extractor
Layout
Spatial Embedding
Stage-2
Cross-modal Information Interaction
Cross-modal Scene Adaptation Module
Spatiotemporal Feature
Car-Move towards-Incoming lane
Multi-label Classifier Head
Coarse-grained global context
Fine-grained local information
Layout Embedding
Positional Embedding