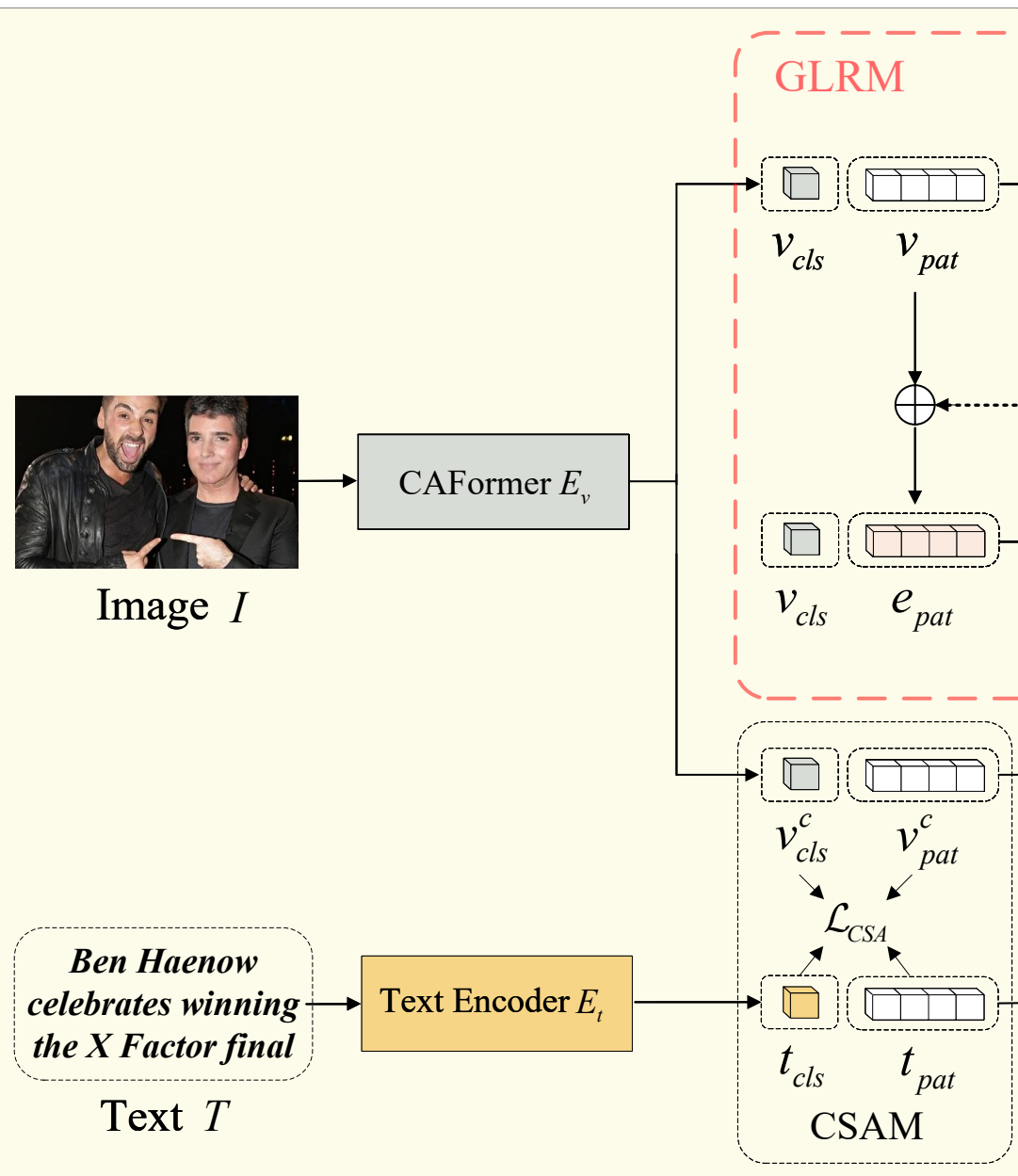
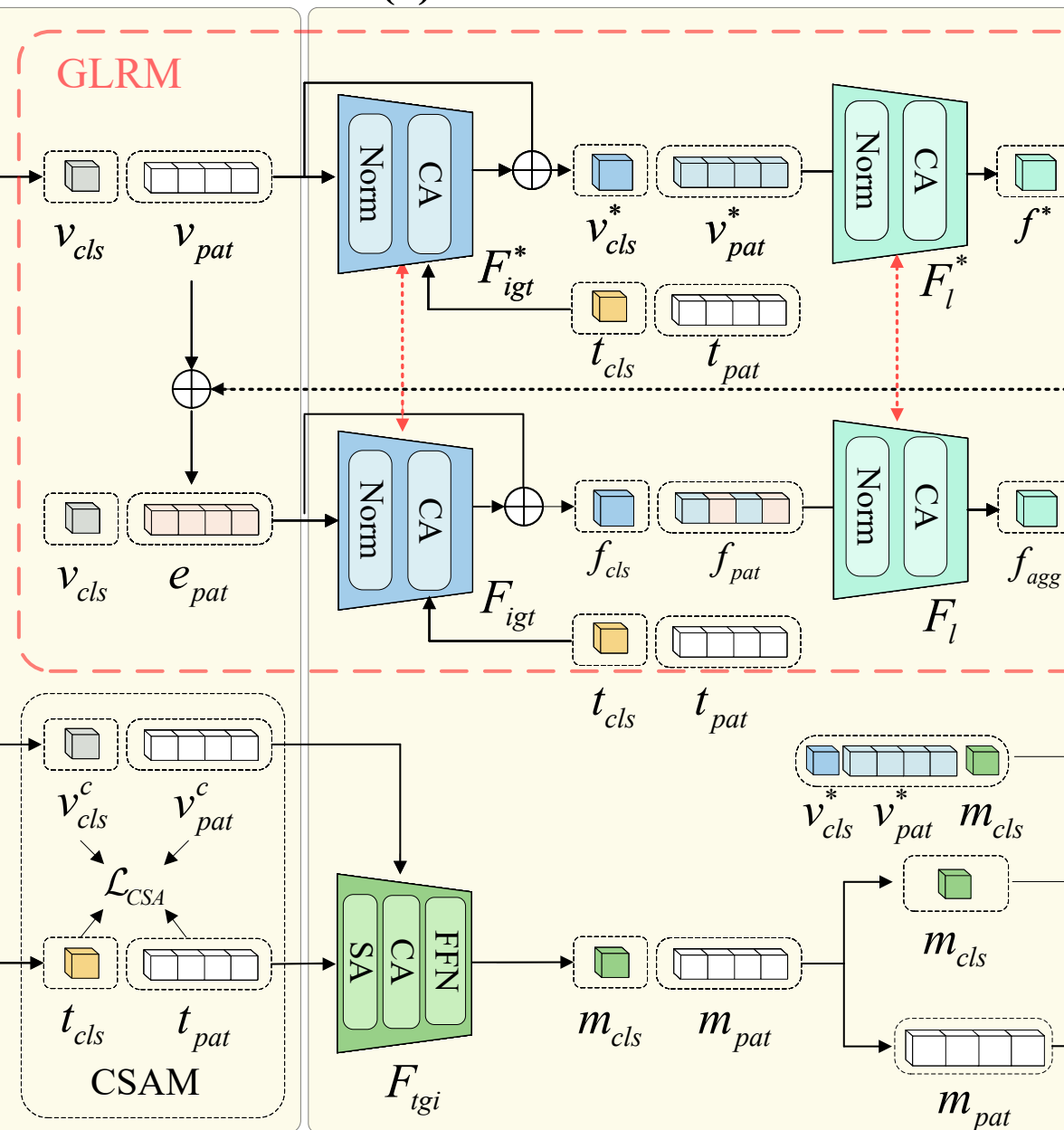


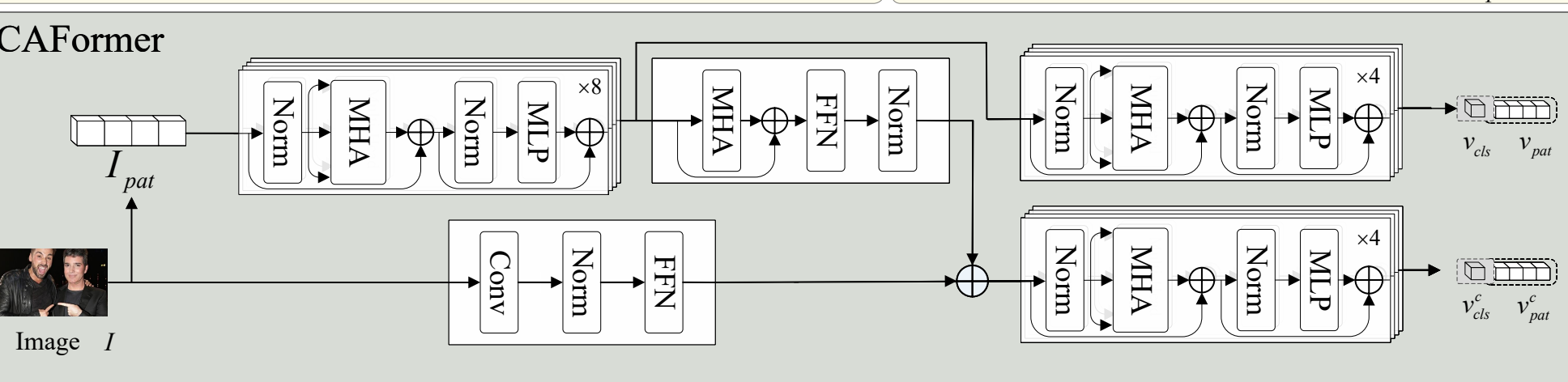
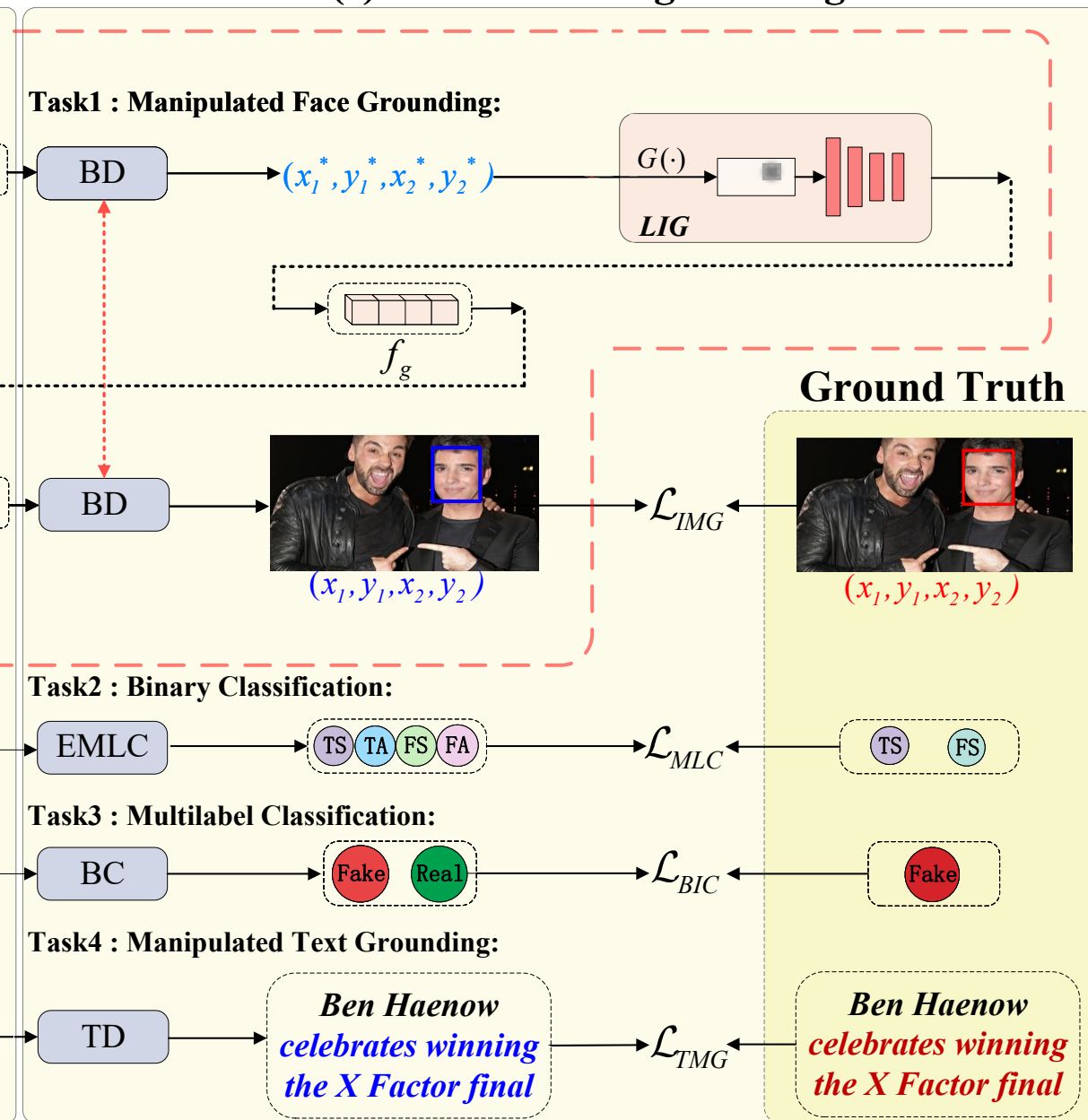
(a) Multimodal Feature Extraction



(b) Multimodal Feature Fusion



(c) Detection and grounding



(d) Convolution-Augmented Focus Transformer

- \oplus : Element-wise Sum
 CA : Cross-Attention
 SA : Self-Attention
 FFN : Feedforward Network
 MHA : Multi-Head Attention
 GLRM : Guided Localization Refinement Module
 CSAM : Cross-modal Semantic-aware Alignment Mechanism
 BD : Bbox Detection
 BC : Binary Classifier
 TD : Token Detection
 EMLC : Enhanced Multi-Label Classifier