(a) Multimodal Feature Extraction

(b) Multimodal Feature Fusion

(c) Detection and grounding

GLRM

Image I

CAFormer E_v

v_{cls} v_{pat}

e_{pat}

Norm CA

F^*_{igt} F_{igt}

v^*_{cls} v^*_{pat} t_{cls} t_{pat}

f_{cls} f_{pat}

Norm CA SA FFN

F^*_l F_l f^* f_{agg}

Text T

Ben Haenow celebrates winning the X Factor final

Text Encoder E_t

v^c_{cls} v^c_{pat} $\mathcal{L}_{CSA}$ t_{cls} t_{pat}

CSAM

F_{tgi}

m_{cls} m_{pat}

Task1 : Manipulated Face Grounding:

BD

$(x_1^*, y_1^*, x_2^*, y_2^*)$

$G(\cdot)$ LIG

f_g

(x_1, y_1, x_2, y_2)

Ground Truth

$\mathcal{L}_{IMG}$

Task2 : Binary Classification:

EMLC

TS TA FS FA

$\mathcal{L}_{MLC}$

Task3 : Multilabel Classification:

BC

Fake Real

$\mathcal{L}_{BIC}$

Task4 : Manipulated Text Grounding:

TD

Ben Haenow celebrates winning the X Factor final

$\mathcal{L}_{TMG}$

CAFormer

Norm MHA MLP ×8

Norm MHA FFN

Norm MHA MLP ×4

Conv Norm FFN

I_{pat}

Image I

v_{cls} v_{pat} v^c_{cls} v^c_{pat}

⊕ : 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

(d) Convolution-Augmented Focus Transformer