1. Experiments

1.1. Implementation Details.

All of our experiments are performed on 8 NVIDIA V100 GPUs with PyTorch framework [7]. Image Encoder is implemented by ViT-B/16 [3] with 12 layers. Both Text Encoder and Multi-Modal Aggregator are built based on a 6-layer transformer initialized by the first 6 layers and the last 6 layers of BERT_base [2], respectively. Binary Classifier, Multi-Label Classifier, BBox Detector, and Token Detector are set up to two Multi-Layer Perception (MLP) layers with output dimensions as 2, 4, 4, and 2. We set the queue size \( K = 65,536 \). AdamW [5] optimizer is adopted with a weight decay of 0.02. The learning rate is warmed-up to \( 10^{-4} \) in the first 1000 steps, and decayed to \( 10^{-5} \) following a cosine schedule.

1.2. Evaluation Metrics.

To evaluate the proposed new research problem DGM comprehensively, we set up rigorous evaluation protocols and metrics for all the manipulation detection and grounding tasks.

- **Binary classification**: Following current deepfake methods [6,8], we adopt Accuracy (ACC), Area Under the Receiver Operating Characteristic Curve (AUC), and Equal Error Rate (EER) for evaluation of binary classification.

- **Multi-label classification**: Like existing multi-label classification methods [1,4], we use mean Average Precision (MAP), average per-class F1 (CF1), and average overall F1 (OF1) for evaluating the detection of fine-grained manipulation types.

- **Manipulated image bounding box grounding**: To examine the performance of predicted manipulated bbox, we calculate the mean of Intersection over Union (IoUmean) between ground-truth and predicted coordinates of all testing samples. Moreover, we set two thresholds (0.5, 0.75) of IoU and calculate the average accuracy (correct grounding if IoU is above the threshold and versa), which are denoted as IoU50 and IoU75.

- **Manipulated text token grounding**: Considering the class imbalance scenario that manipulated tokens are much fewer than original tokens, we adopt Precision, Recall, F1 Score as metrics. This contributes to a more fair and reasonable evaluation for manipulated text token grounding.

References