IDseq: Decoupled and Sequentially Detecting and Grounding Multi-Modal Media Manipulation

Runxin Liu [1], Tian Xie [2], Jiaming Li [1], Lingyun Yu [1]*, Hongtao Xie [1]

[1] University of Science and Technology of China, Hefei, China
[2] Anhui University, Hefei, China
{lrxsxdl, ljmd}@mail.ustc.edu.cn, xietian@nuaa.edu.cn, {yuly, htxie}@ustc.edu.cn

Abstract

Detecting and grounding multi-modal media manipulation aims to categorize the type and localize the region of manipulation for image-text pairs in both two modalities. Existing methods have not sufficiently explored the intrinsic properties of the manipulated images, which contain both forgery and content features, leading to inefficient utilization. To address this problem, we propose an Image-Driven Decoupled Sequential Framework (**IDseq**), designed to decouple image features and rationally integrate them to accomplish different sub-tasks effectively. Specifically, IDseq employs two specially designed disentangled losses to guide the disentangled learning of forgery and content features. To efficiently leverage these features, we propose a Decoupled Image Manipulation Decoder (DIMD) that processes image tasks within a decoupled schema. We mitigate their exclusive competition by separating the image tasks into forgery-relevant and content-relevant components and training them without gradient interaction. Additionally, we utilize content features enhanced by the proposed Manipulation Indicator Generator (MIG) for the text tasks, which provide the maximal visual information as a reference while eliminating interference from unverified image data. Extensive experiments show the superiority of our IDseq, where it notably outperforms SOTA methods on the fine-grained classification by 3.8% in mAP and the forgery face grounding by 8.7% in IoUmean, even 1.3% in F1 on the most challenging manipulated text grounding.

Introduction

Taking advantage of rapid advances in deep generative models (Goodfellow et al. 2020; Karras, Laine, and Aila 2019; Zhu et al. 2017; Rombach et al. 2022) and large language models (Devlin et al. 2018; Achiam et al. 2023), face (Chen et al. 2020; Gao et al. 2021; Wang et al. 2022; Patashnik et al. 2021) and text manipulation techniques (Sudhakar, Upadhyay, and Maheswaran 2019; Dai et al. 2019) could easily generate high-fidelity multimedia content with minimal human intervention. Although these technologies greatly improve human productivity, their malicious use to produce large-scale multi-modal misinformation causes severe consequences, including damaging reputations and misleading public opinion (Wang 2017; Zellers et al. 2019; Floridi 2021), thereby posing significant threats to both individuals and society.

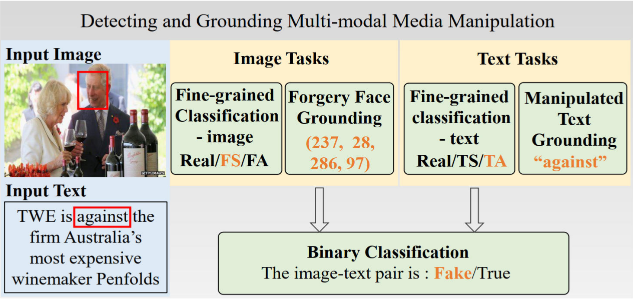

Detecting and Grounding Multi-modal Media Manipulation
Input Image
Input Text
TWE is against the firm Australia's most expensive winemaker Penfolds
Image Tasks
Text Tasks
Fine-grained Classification - image
Real/FS/FA
Forgery Face Grounding
(237, 28, 286, 97)
Fine-grained classification - text
Real/TS/TA
Manipulated Text Grounding
"against"
Binary Classification
The image-text pair is : Fake/True

Figure 1: DGM^4, a multi-modal multi-task problem, conducts a detailed analysis of manipulation within image-text pairs. The red boxes indicate the manipulated regions and words. The predicted results are highlighted in orange.

To alleviate security concerns, early research primarily modelled manipulation detection tasks as binary classification. For example, deepfake detection methods (Li et al. 2020; Haliassos et al. 2021; Jeong et al. 2022; Jia et al. 2021) focus on extracting forensic clues such as blending boundaries and frequency anomalies to distinguish the authenticity of facial images. Similarly, fake news detection methods (Zellers et al. 2019; Potthast et al. 2018; Ahmed, Traore, and Saad 2017; Pérez-Rosas et al. 2017) leveraged writing styles, including psycholinguistic features and readability scores, as key indicators. Despite their considerable success, these binary classification approaches offer only a simple yes-or-no output, failing to provide some explainable evidence to support their predictions. To fill this gap, recent research focuses on detecting and grounding multi-modal media manipulation (**DGM^4**) (Shao, Wu, and Liu 2023; Liu et al. 2023). As shown in Figure 1, DGM^4 aims to perform a detailed analysis of manipulation, encompassing a range of tasks from simple classification to more complex grounding in both visual and textual modalities.

In contemporary research, Hammer (Shao, Wu, and Liu 2023) focuses on learning more aligned visual and textual characteristics for prediction, while UFAFormer (Liu et al. 2023) incorporates frequency characteristics to improve image representation. Although these methods have shown sig-

[*] Corresponding author

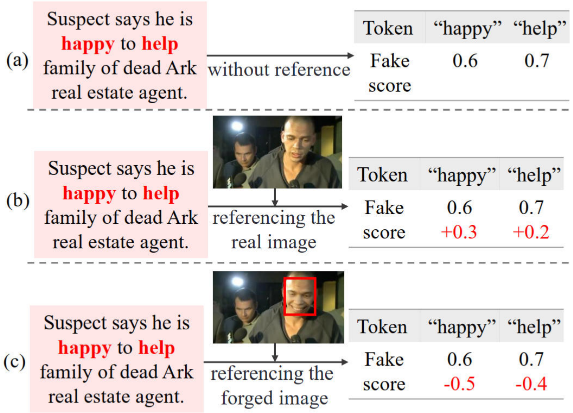

(a)
Suspect says he is happy to help family of dead Ark real estate agent.
without reference
Token "happy" "help"
Fake score 0.6 0.7
(b)
Suspect says he is happy to help family of dead Ark real estate agent.
referencing the real image
Token "happy" "help"
Fake score 0.6 +0.3 0.7 +0.2
(c)
Suspect says he is happy to help family of dead Ark real estate agent.
referencing the forged image
Token "happy" "help"
Fake score 0.6 -0.5 0.7 -0.4

Figure 2: Accomplish text tasks: (a) directly through contextual analysis; (b) referencing the real image; (c) referencing the forged image.

nificant results, they rely on shared image features to address all subtasks. These approaches overlook the fact that manipulated images inherently contain both forgery features, such as forged boundaries (Li et al. 2020) and patch inconsistencies (Huh et al. 2018; Zhao et al. 2021b; Zhou et al. 2023; Ma et al. 2023), and content features, including scenes, characters, and objects. However, different tasks require different features. For image tasks, both content and forgery features play crucial roles: content features are essential for accurately locating all faces within an image, while forgery features are necessary for determining their authenticity. For text tasks, content features provide valuable semantic information, whereas forgery features are often considered noise. Given this distinction, using shared image features across all subtasks leads to exclusive competition, where different tasks compete for the same feature space. This ultimately drives the model toward a suboptimal solution.

Additionally, when handling text tasks, existing methods directly reference visual features without verifying their authenticity, leading to potential interference. In the case shown in Figure 2, given a text input, (a) contextual analysis suggests that the tokens "happy" and "help" are fake; (b) referencing the real image will increase confidence. However, (c) when referencing the forged image, the judgment is completely overturned because the emotion conveyed by the "happy" and "help" aligns closely with the manipulated image. Therefore, directly referencing unverified image information can lead to significant interference.

To tackle these problems, we propose an Image-Driven Decoupled Sequential Framework (**IDseq**) that aims to train forgery-related tasks and content-related tasks in a decoupled schema. Specifically, IDseq employs two Disentangled representation learning losses to capture patch-wise inconsistencies as forgery features and information that is semantically aligned with input text as content features, respectively. Based on the two features, we develop a Decoupled Image Manipulation Decoder (DIMD) that splits the complex image tasks into content-related localization and forgery-related identification and trains them without gradient interaction, thus reducing the exclusive competition for image feature space among tasks. To reduce the interference caused by referencing forged images in text tasks, we implement a sequential task processing order. We introduce a Manipulation Indicator Generator (MIG), which transforms the predicted bounding box of a forged face into an indicator map, effectively highlighting manipulated regions. By integrating pure content features with the indicator map, our approach minimizes interference while preserving as much content information as possible.

Extensive experiments show the superiority of our IDseq. The main contributions are summarized as follows: (1) We propose a novel Image-Driven Decoupled Sequential Framework that deals with DGM^4 by decoupling image information into content and forgery features and leveraging the self-diagnostic ability of images. (2) To avoid exclusive competition between image tasks, we specially design the Decoupled Image Manipulation Decoder to ensure the decoupled training of the complex image tasks without gradient interaction. (3) To mitigate interference for text tasks, we adopt a sequential framework and design the Manipulation Indicator Generator to indicate content features.

Related Work

Previous work on deepfake detection and fake news detection is related to our tasks and discussed in the appendix.

Detection and Grounding of Multi-modal Manipulation. Despite the success of deepfake detection and fake news detection methods, these methods typically model manipulation detection as a binary classification problem. To provide more comprehensive evidence, Shao *et al.* (Shao, Wu, and Liu 2023) introduced the first dataset specifically designed for the DGM^4 problem, which augmented the Human-Centered News Dataset (Liu et al. 2020) by randomly modifying faces in images and words in texts. HAMMER (Shao, Wu, and Liu 2023) serves as the baseline model, which fine-tunes the widely-used visual-language pre-trained model ALBEF (Li et al. 2021). UFAFormer (Liu et al. 2023) aims to detect forgeries by utilizing visual frequency information. It develops a frequency encoder that leverages intra-band and inter-band self-attention layers to capture frequency features from sub-band data following discrete wavelet transformation. Although these models recognize the importance of images, they still use shared image features to address all tasks, which limits the models' performance.

Methodology

Our Image-Driven Decoupled Sequential Framework (**IDseq**) is designed to decouple image features and integrate them rationally to accomplish different sub-tasks in DGM^4. In this section, we first detail the pipeline of the proposed method. Next, we describe the architectures of the Decoupled Image Manipulation Decoder (DIMD) and the Manipulation Indicator Generator (MIG). Finally, we explain the two proposed disentangled losses and other loss functions.

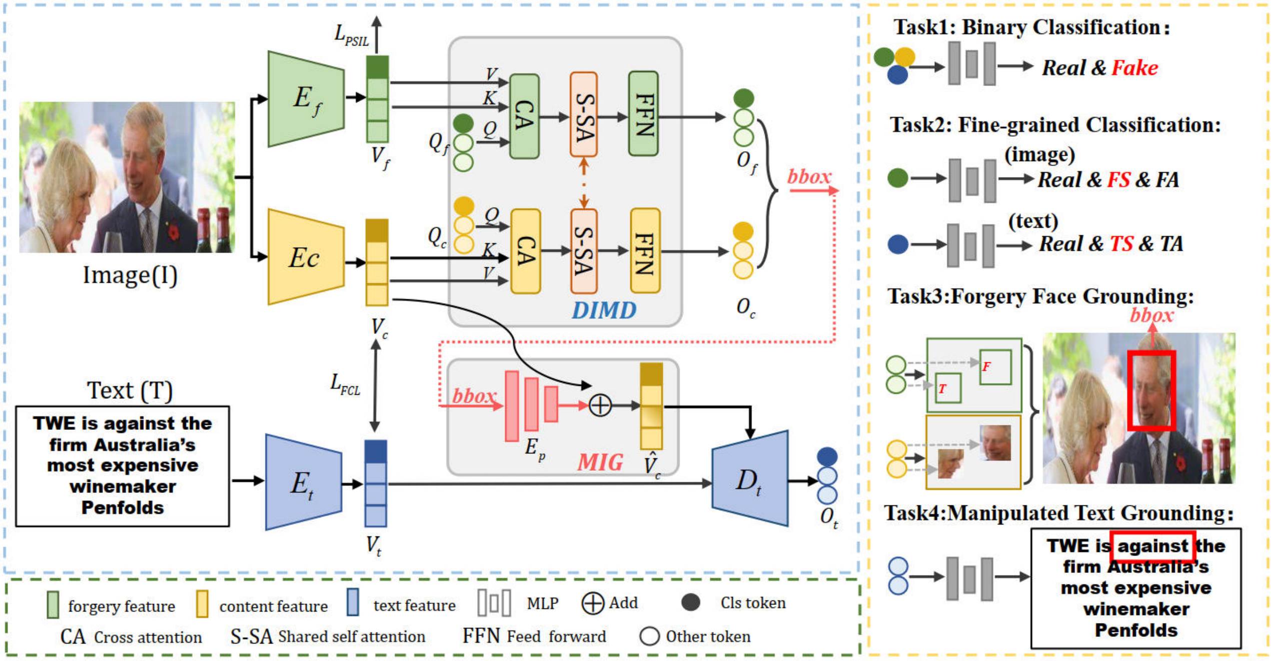


Figure 3: An overview of our IDseq. First, we extract forgery, content, and text features, supervised by two disentangled representation learning losses. We then feed the forgery and content features into the Decoupled Image Manipulation Decoder (DIMD) to handle image tasks. Next, the predicted forgery bounding box is passed to the Manipulation Indicator Generator (MIG) to indicate content features. Finally, we use the enhanced content features along with text features to complete the text tasks and obtain the binary classification.

Pipeline

The pipeline of our **IDseq** is illustrated in Figure 3 . Given an image-text pair (I, T), forgery features V_f, content features V_c, and text features V_t are extracted by a forgery encoder E_f, a content encoder E_c, and a text encoder E_t, respectively. The encoders E_f and E_c share the same architecture but are supervised by two disentangled losses: Patch-wise Self-inconsistency Loss (PSIL) and Fine-grained Contrastive Loss (FCL), respectively. Subsequently, V_f and V_c are fed into DIMD, a specialized DETR-like block designed with two gradient-isolated streams. This process generates the forgery outputs $O_f = \{o_f^i\}_{i=0}^k$ and content outputs $O_c = \{o_c^i\}_{i=0}^k$, where $k = 32$ represents the number of learnable queries in DIMD. By combining O_f and O_c, we can predict the bounding box $\hat{r}$ to address the forgery face grounding task. The first tokens of O_f and O_c, denoted as o_f^0 and o_c^0, are then utilized for fine-grained image classification. After completing image tasks, we first feed the content features and the predicted box $\hat{r}$ into MIG to obtain the enhanced content features $\hat{V}_c$. These enhanced content features are injected into the text features through cross-attention in the text decoder D_t, which then generates text outputs $O_t = \{o_t^i\}_{i=1}^k$. The first token of O_t, denoted as o_t^0, is employed for fine-grained text classification, while the remaining tokens are used for manipulated text grounding. Finally, the concatenation of o_f^0, o_c^0, and o_t^0 is used to perform binary classification of the entire image-text pair.

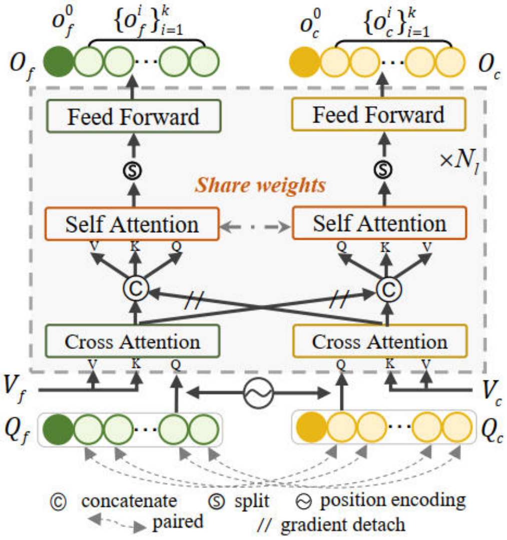


Figure 4: The architecture of Decoupled Image Manipulation Decoder.

Decoupled Image Manipulation Decoder

To train content-related and forgery-related tasks decoupled, it is essential to tackle the challenge of forgery face grounding task, which requires both two kinds of features. Inspired by object detection and grounding methods (He and Todorovic 2022; Zhang et al. 2023; Liu et al. 2024), we design a DETR-like (Carion et al. 2020) Decoupled Image Manipulation decoder (DIMD), which splits the forgery face grounding into content-related localization and forgery-related identification, as shown in Figure 4. By using two sets of queries to interact with forgery and content features independently and employing shared self-attention to propagate information across branches, our DIMD could collect information independently and provide a comprehensive prediction. Specifically, we randomly initial two sets of queries: forgery queries $Q_f = \{q_f^i\}_{i=0}^{k}$ and content queries $Q_c = \{q_c^i\}_{i=0}^{k}$, where $k = 32$ represents the number of queries. At the l-th layer of DIMD, we apply the same positional encoding to both sets of queries to ensure they are paired. These queries, Q_f^l and Q_c^l, interact with the forgery and content features through cross-attention blocks to gather distinct information respectively. Next, we merge the paired queries to generate integrated representations, Q_f^{l+} and Q_c^{l+}, that include both forgery and content information. These representations are subsequently processed by self-attention blocks with shared weights to facilitate information propagation between two branches, thus ensuring that the paired queries focus on the same regions. Notably, we detach the gradients of the queries from the other branch during concatenation to maintain decoupled training. Finally, we resplit the concatenated representations into forgery and content parts and input them into feedforward neural networks. This process is formulated as follows:

$$Q_f^l,\ Q_c^l = \text{CrossAtt}(Q_f^l, V_f),\ \text{CrossAtt}(Q_c^l, V_c), \quad (1)$$

$$Q_f^{l*},\ Q_c^{l*} = \text{Cat}(Q_f^l, Q_c^l.\text{detach}),\ \text{Cat}(Q_f^l.\text{detach}, Q_c^l), \quad (2)$$

$$Q_f^{l*},\ Q_c^{l*} = \text{S-SelfAtt}(Q_f^{l*}),\ \text{S-SelfAtt}(Q_c^{l*}), \quad (3)$$

$$Q_f^{l+1}, Q_c^{l+1} = \text{FFN}(\text{Split}(Q_f^{l*})), \text{FFN}(\text{Split}(Q_c^{l*})). \quad (4)$$

After the decoding process in DIMD, we obtain outputs O_f and O_c, which are then used for identification and localization, respectively. The predicted bounding box with the highest fake score is considered the bounding box of forged face $\hat{r}$.

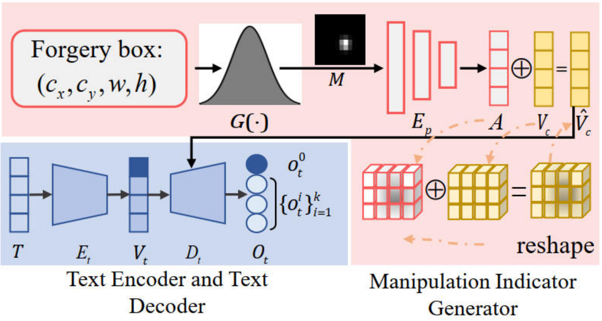


Figure 5: The architecture of text encoder, text decoder and Manipulation Indicator Generator.

Manipulation Indicator Generator

Since all text tasks rely on content features as an additional reference, it is essential to minimize the interference caused by using unverified image information. Thus, we propose the Manipulation Indicator Generator (MIG), a simple yet effective module that transforms the coordinates of a forged face into an embedding A, which indicates whether a specific region has been manipulated. As illustrated in Figure 5, the MIG consists of two key components: a Box-based Adaptive Gaussian Kernel Function $G(\cdot)$ and a three-layer convolutional network E_p. Initially, the coordinates (c_x, c_y, w, h) of the detected forgery face are transformed into a fake map M using the Gaussian kernel function:

$$G(x, y) = \exp\{(-\frac{(x - c_x)^2 + (y - c_y)^2}{\sigma^2}\}), \quad (5)$$

where $\sigma = \sqrt{\left(\frac{h}{2}\right)^2 + \left(\frac{w}{2}\right)^2}$. The values on this map are constrained between 0 and 1, with higher values indicating a greater likelihood that the region has been affected by the forgery. This fake map is then encoded by E_p to obtain an indicative embedding A, which is subsequently added to the content features V_c to produce the indicated content features $\hat{V}_c$. Finally, the indicated content features and text features are fed into the text decoder to complete the text tasks. By training the entire IDseq with MIG, we could mark the manipulated content features, which allows our model to automatically select the content information while guarding against manipulated information, thus alleviating interference caused by unverified image features.

Training Objective

Task-specific Losses. We define three loss functions: binary classification loss L_{bc}, fine-grained classification loss L_{fc} and manipulated text grounding loss L_{tg}, all based on the cross-entropy loss L_{ce}, formulated as follows:

$$L_{bc} = L_{ce}(\hat{y}_{bc}, y_{bc}), \quad (6)$$

$$L_{fc} = L_{ce}(\hat{y}_i, y_i) + L_{ce}(\hat{y}_t, y_t), \quad (7)$$

$$L_{tg} = \frac{1}{k}\sum_{i=1}^{k} L_{ce}(\hat{y}_w^i, y_w^i), \quad (8)$$

where $\hat{y}_{bc}$, $\hat{y}_i$, $\hat{y}_t$ and $\hat{y}_w^i$ represent the predictions for the image-text pair, image manipulation type, text manipulation type and the fake score of i-th token in text, while y_{bc}, y_i, y_t and y_w^i are the corresponding ground truth labels.

For forgery face grounding, we use L1 loss and GIoU loss between the predicted forgery face bounding box $\hat{r}$ and the ground truth r, formulated as:

$$L_{fg} = L_1(\hat{r}, r) + L_{GIoU}(\hat{r}, r). \quad (9)$$

Disentangled Representation Learning Losses. Patch-wise Self-Inconsistency Loss (**PSIL**) is used to maximize the patch-level inconsistencies between real and forged regions, which have been validated as important forgery clues. Following the previous deepfake detection method (Zhao et al. 2021b), we calculate the cosine similarity between the features of any patch pair (v_f^i, v_f^j) from $V_f \in \mathbb{R}^{n_i \times d_i}$ and obtain a similarity map $S \in \mathbb{R}^{n_i \times n_i}$ with values between 0 and 1, where n_i represents the number of patches in the image. Next, we create a binary ground truth map $\hat{S} \in \mathbb{R}^{n_i \times n_i}$, where a value of 1 means both patches are either real or forged, and 0 means one patch is real and the other is forged. The patch-wise self-inconsistency Loss is formulated as follows:

$$L_{PSIL} = \frac{1}{n_i \times n_i} \sum_{(p,q)} L_{bce}(S_{(p,q)}, \hat{S}_{(p,q)}). \tag{10}$$

Fine-grained Contrastive Loss (**FCL**) forces the content encoder to focus on content-relevant information that is semantically aligned with the paired text, which is formulated as follows:

$$\begin{aligned} L_{FCL} &= \frac{1}{2} L_{it} + \frac{1}{2} L_{ti}, \\ L_{it} &= -\frac{1}{b} \log \frac{\exp(S(I, T^+))}{\sum_j \exp(S(I, T^-))}, \\ L_{ti} &= -\frac{1}{b} \log \frac{\exp(S(T, I^+))}{\sum_j \exp(S(T, I^-))}, \end{aligned} \tag{11}$$

where b is the batch size, I denotes the collection of image tokens, T^+ refers to text tokens from the same pair that remain unmanipulated, and T^- includes other tokens within the batch. The definitions of T, I^- and I^+ are similar. Unlike the contrastive loss used in HAMMER (Shao, Wu, and Liu 2023), which relies solely on the class token, our score function $S(\cdot)$ measures the similarity across all tokens, following (Yao et al. 2021), to ensure fine-grained supervision.

Finally, the training objective is outlined as follows:

$$\begin{aligned} L =& \lambda_1 L_{bc} + \lambda_2 L_{fc} + \lambda_3 L_{fg} + \\ & \lambda_4 L_{tg} + \lambda_5 L_{PSIL} + \lambda_6 L_{FCL}, \end{aligned} \tag{12}$$

where $\lambda_1 = 1, \lambda_2 = 1, \lambda_3 = 0.1$ and $\lambda_4 = 1, \lambda_5 = 0.1, \lambda_6 = 0.1$, following the hyperparameter settings of the baseline (Shao, Wu, and Liu 2023).

Experiments

Dataset. We conduct experiments on the DGM$^{\mathbf{4}}$ dataset (Shao, Wu, and Liu 2023), which comprises 230,000 image-text paired samples, including over 77,000 pristine pairs and 152,000 manipulated pairs. The dataset encompasses four types of manipulation: face swap (FS), face attribute (FA), text swap (TS), and text attribute (TA) and presents a challenging scenario where manipulated faces and text are randomly combined. Additionally, half of the dataset is subjected to various perturbations, such as JPEG compression and Gaussian noise, to mimic real-world conditions. We train our IDseq on the training set and evaluate its performance on the test set.

Evaluation Metric. We report our results following the original evaluation protocols and metrics (Shao, Wu, and Liu 2023). Here, we detail the metrics for each DGM$^{\mathbf{4}}$ task. (1) For binary classification, we adopt Accuracy (ACC), Area Under the Receiver Operating Characteristic curve (AUC), and Equal Error Rate (EER). (2) For fine-grained classification, we use the mean Average Precision (mAP), average per-class F1 (CF1), and average overall F1 (OF1). (3) For forgery face grounding, we calculate the mean of intersection over union (IoUmean) and IoU with two thresholds $\{0.5, 0.75\}$ denoted as IoU50 and IoU75 to examine the performance. (4) For manipulated text token grounding, precision (PR), Recall (RE), and F1 score are employed.

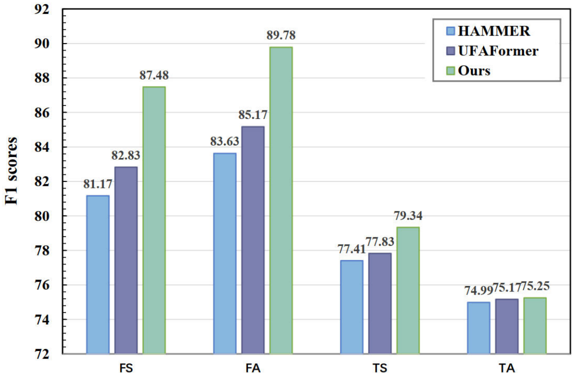


Figure 6: Results of the fine-grained classification of four manipulation types, using the F1-score as the metric.

Implementation Details. For fair comparisons, our training configuration follows the existing methods (Shao, Wu, and Liu 2023; Liu et al. 2023). Two image encoders are implemented by ViT/B with 12 layers (Dosovitskiy et al. 2020). Our text encoder and text decoder are built based on 6-layer transformers, respectively. The above components are initialized by the ALBEF (Li et al. 2021) following the baseline model (Shao, Wu, and Liu 2023). The DIMD is trained from scratch and has 6 layers. The AdamW (Loshchilov and Hutter 2017) optimizer with a weight decay of 0.02 is adopted to optimize the model. The initial learning rates for encoders and the others are set to 1×10^{-5} and 1×10^{-4} under a cosine schedule. The model is trained on four Nvidia A40 GPUs with batch size 128 for 50 epochs. The input images are resized into 224×224, and the text sequence is padded with a max length of 50 for both training and testing. We implement our model on PyTorch (Paszke et al. 2019).

Comparisons with State-of-the-Art

Comparisons with uni-modal methods. We compare our model with some uni-modal methods, as shown in Table 1. In the visual modality, we compare our IDseq with two deepfake detection methods, TS (Luo et al. 2021) and MAT (Zhao et al. 2021a), which can perform the binary classification task and the forgery face grounding task. In the text modality, we compare our model with two sequence tagging models, i.e., BERT (Devlin et al. 2018) and LUKE (Yamada et al. 2020).

Comparisons with multi-modal methods. We present comparisons of our IDseq with state-of-the-art approaches in Table 1 and Figure 6. In the emerging field of DGM^4, HAMMER (Shao, Wu, and Liu 2023) serves as the baseline model, while HAMMER++ (Shao et al. 2024) improves upon it by employing an enhanced contrastive learning loss.

Methods		Binary Cls			Fine-grained Cls			Forgery Face Grounding			Manipulated Text Grounding		
Name	Params(M)	AUC↑	EER↓	ACC↑	mAP↑	CF1↑	OF1↑	IoU_{mean}↑	IoU_{50}↑	IoU_{75}↑	Precision↑	Recall↑	F1↑
TS	-	91.80	17.11	82.89	-	-	-	72.85	79.12	74.06	-	-	-
MAT	-	91.31	17.65	82.36	-	-	-	72.88	78.98	74.70	-	-	-
BERT	-	80.82	28.02	68.98	-	-	-	-	-	-	41.39	63.85	50.23
LUKE	-	81.39	27.88	76.18	-	-	-	-	-	-	50.52	37.93	43.33
CLIP	586.12	83.22	24.61	76.40	66.00	59.52	62.31	49.51	50.03	38.79	58.12	22.11	32.03
ViLT	272.87	85.16	22.88	78.38	72.37	66.14	66.00	59.32	65.18	38.40	66.48	49.88	57.00
HAMMER	389.49	93.19	14.10	86.39	86.22	79.37	80.37	76.45	83.75	76.06	75.01	68.02	71.35
HAMMER++	389.52	93.33	14.06	86.66	86.41	79.37	80.71	76.46	83.77	76.03	73.05	**72.14**	72.59
UFAFormer		93.81	13.60	86.80	87.85	80.31	81.48	78.33	85.39	79.20	73.35	70.30	72.02
Ours	335.64	**94.55**	**11.40**	**88.94**	**90.01**	**83.00**	**84.90**	**83.33**	**89.39**	**86.19**	**75.96**	71.23	**73.52**

Table 1: Comparisons with state-of-the-art methods on the DGM^4 dataset. Bold font indicates optimal results. The up arrow signifies that a larger value corresponds to better model performance, and vice versa. Binary Cls represents the binary classification task, and Fine-grained Cls represents the fine-grained classification task. Params(M) represents the number of parameters, measured in millions.

	Componets		Binary Cls			Fine-grained Cls			Forgery Face Grounding			Manipulated text Grounding		
	PSIL	FCL	AUC↑	EER↓	ACC↑	mAP↑	CF1↑	OF1↑	IoU_{mean}↑	IoU_{50}↑	IoU_{75}↑	Precision↑	Recall↑	F1↑
(1)	–	–	91.04	16.48	84.11	84.83	77.01	77.39	74.56	80.77	76.78	76.78	64.26	69.96
(2)	–	✓	91.45	15.97	84.56	85.05	77.31	78.02	75.87	81.23	78.76	70.94	**74.18**	72.52
(3)	✓	–	93.59	13.21	87.17	89.21	80.89	82.57	82.88	88.36	85.80	**76.91**	64.92	70.41
(4)	✓	✓	**94.55**	**11.40**	**88.94**	**90.01**	**83.00**	**84.90**	**83.33**	**89.39**	**86.19**	75.96	71.23	**73.52**

Table 2: Ablation study of two disentangled representation learning losses.

UFAFormer (Liu et al. 2023) achieves the best results by incorporating frequency information. We also compare our model with two widely used large-scale pre-trained models, CLIP (Radford et al. 2021) and ViLT (Kim, Son, and Kim 2021), by fully fine-tuning them for our tasks. As the results indicate, our method outperforms all previous state-of-the-art methods, demonstrating the validity and effectiveness of our framework.

Methods	Forgery Face Grounding		
	IoU_{mean}↑	IoU_{50}↑	IoU_{75}↑
(a) $V_f \rightarrow$ SD	81.32	86.73	85.28
(b) $V_c \rightarrow$ SD	74.26	80.57	76.60
(c)V_f+V_c $\rightarrow$ CD	82.62	88.05	**86.64**
(d)V_f $\rightarrow$SD + V_c $\rightarrow$SD	49.22	54.74	49.15
(e) Ours	**83.33**	**89.39**	86.19

Table 3: Ablation study of DIMD with its variants.

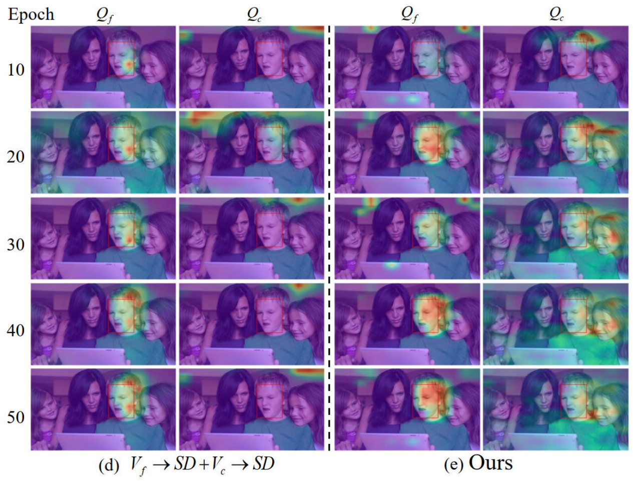
Ablation Study

To examine the effectiveness of our IDseq, we conduct extensive experiments on the DGM^4 dataset.

Effectiveness of PSIL and FCL. As shown in Table 2, we first investigate the impact of patch-wise self-inconsistency loss (PSIL) and fine-grained contrastive loss (FCL) by quantitatively evaluating our IDseq and its variants: (1) IDseq w/o PSIL and FCL; (2) IDseq w/o PSIL; (3) IDseq w/o FCL; (4) our IDseq. Adding PSIL improves the forgery face grounding in IoU_{mean} by 8.32%, indicating that PSIL helps the forgery encoder extract self-inconsistencies in images. Adding FCL improves the forgery face grounding in IoU_{mean} by 1.31% and the manipulated text grounding in F1 score by 2.56%, demonstrating FCL's effectiveness in helping the content encoder focus on semantic visual information. The results show that these losses guide the encoders to extract more discriminative features.

Effectiveness of the DIMD. To validate the effectiveness of our DIMD, we designed several comparative methods for forgery face grounding, as shown in Table 3. First, we compare DIMD with methods that use only forgery features or content features: (a) employing forgery features with a single 6-layer decoder (SD), which includes one set of queries and contains a cross-attention block, a self-attention block, and an FFN block in each layer; and (b) employing content features with a similar single decoder setup. Next, we compare DIMD with methods that use both forgery and content features: (c) utilizing a coupled decoder (CD) with two serial cross-attention blocks in each layer, where queries interact with forgery and content features sequentially; and (d) employing a dual-branch decoder architecture similar to DIMD, but with isolated self-attention layers instead of shared ones, and without the concatenate and split process.

From Table 3, it is evident that forgery face grounding is a hybrid task that requires both forgery and content information, thus supporting our claim. The comparison between (c) and our method highlights the effectiveness of a decoupled training schema over a coupled one. The underperformance of method (d) arises from its inability to ensure that paired queries focus on the same region of interest, due to the lack of shared self-attention layers and the concatenate process. As illustrated in Figure 7, we present the attention



(d) $V_f \rightarrow SD + V_c \rightarrow SD$ (e) Ours

Figure 7: Visualization of the cross-attention maps of DIMD (right) and its variant-(d) (left). Columns signed with Q_f show attention maps of Q_f and columns signed with Q_c show the attention maps of Q_c. The red boxes highlight the fake regions.

maps of paired queries across different training epochs. In method (d), paired queries struggle to converge and focus on the same regions.

Effectiveness of the MIG. As shown in Table 4, we conduct an ablation study to validate the effect of MIG. To further evaluate the effectiveness of the fake map M generated by the box-based adaptive gaussian kernel function, we also replace it with a binary fake map, where pixels inside the bounding box are assigned a value of 1, and those outside are assigned a value of 0. The improvements across all text tasks demonstrate that the proposed MIG effectively mitigates the interference caused by referencing unverified images.

Methods	TS	TA	Manipulated text grounding		
	F1↑	F1↑	Precision↑	Recall↑	F1↑
w/o MIG	76.28	71.07	71.50	72.79	72.14
w/ MIG(BFM)	77.00	74.59	71.66	**73.11**	72.38
w/ MIG(GFM)	**79.34**	**75.25**	**75.96**	71.23	**73.52**

Table 4: Ablation study of MIG and the box-based adaptive gaussian kernel function."w/o MIG" means not using MIG, "BFM" denotes the binary fake map, and "GFM" denotes the fake map generated by $G(\cdot)$.

Visualization

Results Visualization. In Figure 8, we visualize test samples with different manipulation types. "GT" represents the ground truth annotations and "Pred" indicates the prediction results. The four manipulation types are face swap (FS), face attribute (FA), text swap (TS), and text attribute (TA). In the first row, only one modality is manipulated. In the second row, more challenging scenarios are shown with combinations of FS+TS, FA+TS, FS+TA, and FA+TA, yet our model successfully detects them. These examples demonstrate the effectiveness of our model.

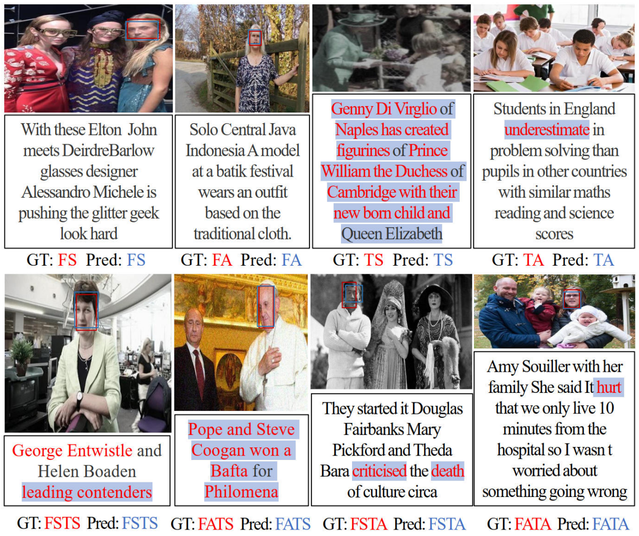

With these Elton John meets DeirdreBarlow glasses designer Alessandro Michele is pushing the glitter geek look hard
Solo Central Java Indonesia A model at a batik festival wears an outfit based on the traditional cloth.
Genny Di Virglio of Naples has created figurines of Prince William the Duchess of Cambridge with their new born child and Queen Elizabeth
Students in England underestimate in problem solving than pupils in other countries with similar maths reading and science scores
GT: FS Pred: FS
GT: FA Pred: FA
GT: TS Pred: TS
GT: TA Pred: TA
George Entwistle and Helen Boaden leading contenders
Pope and Steve Coogan won a Bafta for Philomena
They started it Douglas Fairbanks Mary Pickford and Theda Bara criticised the death of culture circa
Amy Souiller with her family She said It hurt that we only live 10 minutes from the hospital so I wasn t worried about something going wrong
GT: FSTS Pred: FSTS
GT: FATS Pred: FATS
GT: FSTA Pred: FSTA
GT: FATA Pred: FATA

Figure 8: Visualization of detection and grounding results. Ground truth annotations are in red, and prediction results are in blue.

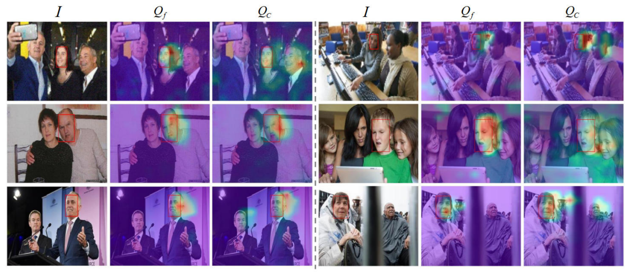


Figure 9: Visualization of the cross-attention maps for forgery queries with forgery features and content queries with content features.

Attention Visualization. As shown in Figure 9, we visualize the cross-attention maps of the paired forgery and content queries with the highest fake scores, which are used for the final prediction for $\hat{r}$. From the maps, we observe that forgery queries focus on the most prominent tampering artifacts, while content queries concentrate on the surroundings of the tampered areas to identify facial boundaries.

Conclusion

We propose a novel Image-Driven Decoupled Sequential Framework (IDseq) to address DGM^4 by providing suitable image information for different tasks and leveraging the self-diagnostic capability of images. To prevent exclusive competition, we use a decoupled training schema supervised by two auxiliary losses and design a DIMD to ensure effective decoupled training for complex forgery face grounding. Additionally, we employ a sequential framework and introduce the MIG, which indicates content features to mitigate interference from unverified image information. We believe this work provides valuable insights into DGM^4 and inspires future research.

Acknowledgments

This work is supported by the National Nature Science Foundation of China (U23B2028, 62121002, 62232006, 62032006, 62472395).

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Ahmed, H.; Traore, I.; and Saad, S. 2017. Detection of online fake news using n-gram analysis and machine learning techniques. In *Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments: First International Conference, ISDDC 2017, Vancouver, BC, Canada, October 26-28, 2017, Proceedings 1*, 127–138. Springer.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers. In *European conference on computer vision*, 213–229. Springer.

Chen, R.; Chen, X.; Ni, B.; and Ge, Y. 2020. Simswap: An efficient framework for high fidelity face swapping. In *Proceedings of the 28th ACM international conference on multimedia*, 2003–2011.

Dai, N.; Liang, J.; Qiu, X.; and Huang, X. 2019. Style transformer: Unpaired text style transfer without disentangled latent representation. *arXiv preprint arXiv:1905.05621*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Floridi, L. 2021. Artificial intelligence, deepfakes and a future of ectypes. *Ethics, Governance, and Policies in Artificial Intelligence*, 307–312.

Gao, G.; Huang, H.; Fu, C.; Li, Z.; and He, R. 2021. Information bottleneck disentanglement for identity swapping. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3404–3413.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2020. Generative adversarial networks. *Communications of the ACM*, 63(11): 139–144.

Haliassos, A.; Vougioukas, K.; Petridis, S.; and Pantic, M. 2021. Lips don't lie: A generalisable and robust approach to face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5039–5049.

He, L.; and Todorovic, S. 2022. Destr: Object detection with split transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9377–9386.

Huh, M.; Liu, A.; Owens, A.; and Efros, A. A. 2018. Fighting fake news: Image splice detection via learned self-consistency. In *Proceedings of the European conference on computer vision (ECCV)*, 101–117.

Jeong, Y.; Kim, D.; Min, S.; Joe, S.; Gwon, Y.; and Choi, J. 2022. Bihpf: Bilateral high-pass filters for robust deepfake detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 48–57.

Jia, G.; Zheng, M.; Hu, C.; Ma, X.; Xu, Y.; Liu, L.; Deng, Y.; and He, R. 2021. Inconsistency-aware wavelet dual-branch network for face forgery detection. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 3(3): 308–319.

Karras, T.; Laine, S.; and Aila, T. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4401–4410.

Kim, W.; Son, B.; and Kim, I. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *International conference on machine learning*, 5583–5594. PMLR.

Li, J.; Selvaraju, R.; Gotmare, A.; Joty, S.; Xiong, C.; and Hoi, S. C. H. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34: 9694–9705.

Li, L.; Bao, J.; Zhang, T.; Yang, H.; Chen, D.; Wen, F.; and Guo, B. 2020. Face x-ray for more general face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5001–5010.

Liu, F.; Wang, Y.; Wang, T.; and Ordonez, V. 2020. Visual news: Benchmark and challenges in news image captioning. *arXiv preprint arXiv:2010.03743*.

Liu, H.; Tan, Z.; Chen, Q.; Wei, Y.; Zhao, Y.; and Wang, J. 2023. Unified Frequency-Assisted Transformer Framework for Detecting and Grounding Multi-Modal Manipulation. *arXiv preprint arXiv:2309.09667*.

Liu, Z.; Li, J.; Xie, H.; Li, P.; Ge, J.; Liu, S.-A.; and Jin, G. 2024. Towards balanced alignment: Modal-enhanced semantic modeling for video moment retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 3855–3863.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Luo, Y.; Zhang, Y.; Yan, J.; and Liu, W. 2021. Generalizing face forgery detection with high-frequency features. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16317–16326.

Ma, X.; Du, B.; Liu, X.; Hammadi, A. Y. A.; and Zhou, J. 2023. Iml-vit: Image manipulation localization by vision transformer. *arXiv preprint arXiv:2307.14863*.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Patashnik, O.; Wu, Z.; Shechtman, E.; Cohen-Or, D.; and Lischinski, D. 2021. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2085–2094.

Pérez-Rosas, V.; Kleinberg, B.; Lefevre, A.; and Mihalcea, R. 2017. Automatic detection of fake news. *arXiv preprint arXiv:1708.07104*.

Potthast, M.; Kiesel, J.; Reinartz, K.; Bevendorff, J.; and Stein, B. 2018. A Stylometric Inquiry into Hyperpartisan and Fake News. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 231–240.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Shao, R.; Wu, T.; and Liu, Z. 2023. Detecting and grounding multi-modal media manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6904–6913.

Shao, R.; Wu, T.; Wu, J.; Nie, L.; and Liu, Z. 2024. Detecting and grounding multi-modal media manipulation and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Sudhakar, A.; Upadhyay, B.; and Maheswaran, A. 2019. Transforming delete, retrieve, generate approach for controlled text style transfer. *arXiv preprint arXiv:1908.09368*.

Wang, T.; Zhang, Y.; Fan, Y.; Wang, J.; and Chen, Q. 2022. High-fidelity gan inversion for image attribute editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11379–11388.

Wang, W. Y. 2017. " liar, liar pants on fire": A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648*.

Yamada, I.; Asai, A.; Shindo, H.; Takeda, H.; and Matsumoto, Y. 2020. LUKE: Deep contextualized entity representations with entity-aware self-attention. *arXiv preprint arXiv:2010.01057*.

Yao, L.; Huang, R.; Hou, L.; Lu, G.; Niu, M.; Xu, H.; Liang, X.; Li, Z.; Jiang, X.; and Xu, C. 2021. Filip: Fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783*.

Zellers, R.; Holtzman, A.; Rashkin, H.; Bisk, Y.; Farhadi, A.; Roesner, F.; and Choi, Y. 2019. Defending against neural fake news. *Advances in neural information processing systems*, 32.

Zhang, M.; Song, G.; Liu, Y.; and Li, H. 2023. Decoupled detr: Spatially disentangling localization and classification for improved end-to-end object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6601–6610.

Zhao, H.; Zhou, W.; Chen, D.; Wei, T.; Zhang, W.; and Yu, N. 2021a. Multi-attentional deepfake detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2185–2194.

Zhao, T.; Xu, X.; Xu, M.; Ding, H.; Xiong, Y.; and Xia, W. 2021b. Learning self-consistency for deepfake detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, 15023–15033.

Zhou, J.; Ma, X.; Du, X.; Alhammadi, A. Y.; and Feng, W. 2023. Pre-training-free Image Manipulation Localization through Non-Mutually Exclusive Contrastive Learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 22346–22356.

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, 2223–2232.