Rethinking Vision-Language Model in Face Forensics: Multi-Modal Interpretable Forged Face Detector

Xiao Guo[1], Xiufeng Song[2], Yue Zhang[1], Xiaohong Liu[2], Xiaoming Liu[1]
[1] Michigan State University [2] Shanghai Jiao Tong University
{guoxia11, zhan1624, liuxm}@msu.edu, {akikaze,xiaohongliu}@sjtu.edu.cn

Abstract

Deepfake detection is a long-established research topic vital for mitigating the spread of malicious misinformation. Unlike prior methods that provide either binary classification results or textual explanations separately, we introduce a novel method capable of generating both simultaneously. Our method harnesses the multi-modal learning capability of the pre-trained CLIP and the unprecedented interpretability of large language models (LLMs) to enhance both the generalization and explainability of deepfake detection. Specifically, we introduce a multi-modal face forgery detector (M2F2-Det) that employs tailored face forgery prompt learning, incorporating the pre-trained CLIP to improve generalization to unseen forgeries. Also, M2F2-Det incorporates an LLM to provide detailed textual explanations of its detection decisions, enhancing interpretability by bridging the gap between natural language and subtle cues of facial forgeries. Empirically, we evaluate M2F2-Det on both detection and explanation generation tasks, where it achieves state-of-the-art performance, demonstrating its effectiveness in identifying and explaining diverse forgeries. Source code is available at link.

1. Introduction

Generative Models (GMs) [9, 18, 23, 29, 57] have demonstrated impressive capabilities in synthesizing highly realistic and visually compelling images. However, they also facilitate the proliferation of AI-generated content (AIGC), like *deepfakes*, raising serious concerns over the spread of deceptive or manipulated facial imagery. To counter these threats, substantial efforts have been made to develop deepfake detection techniques, including subtle artifacts indentification [6, 22, 39, 40, 61, 90], frequency analysis [19, 44, 48, 55, 74], disentangling forgery traces via specialized neural networks [15, 20, 26, 43, 46, 78, 79, 89], modeling temporal inconsistencies [24, 62, 75, 91], among others.

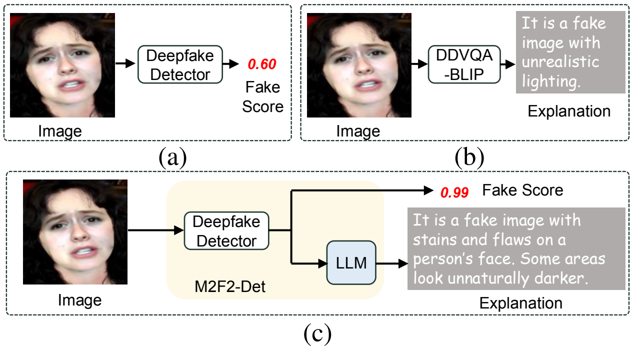
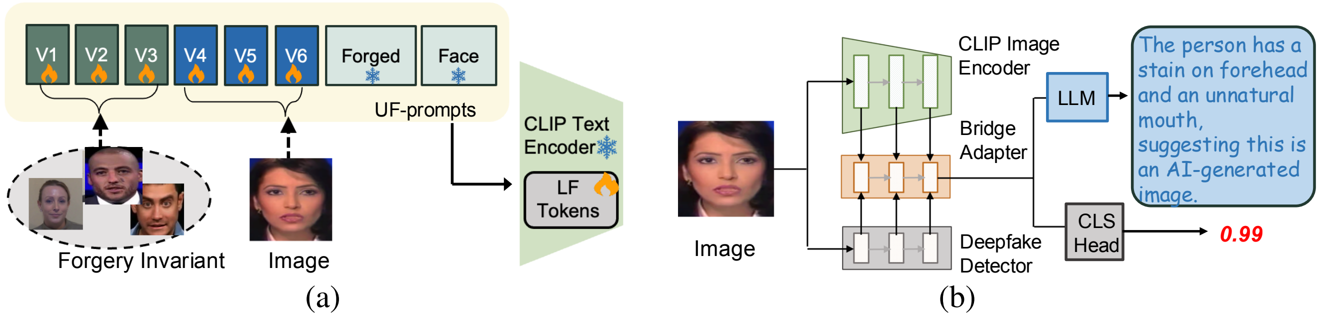


Image
Deepfake Detector
0.60
Fake Score
(a)
Image
DDVQA -BLIP
It is a fake image with unrealistic lighting.
Explanation
(b)
Image
Deepfake Detector
M2F2-Det
LLM
0.99
Fake Score
It is a fake image with stains and flaws on a person's face. Some areas look unnaturally darker.
Explanation
(c)

Figure 1. (a) and (b) represent conventional deepfake detectors and DDVQA-BLIP [86], which take an image as the input and output the fake probability (*e.g.*, score) and textual explanations, respectively. (c) In this work, we propose a multi-modal face forgery detector (M2F2-Det) that produces both fake probability and textual explanations.

Recently, the powerful capability of vision-language models, *e.g.,* CLIP [56], also inspired efforts in detecting deepfakes. For example, DDVQA-BLIP [86] reformulates deepfake detection as an explanation generation task using a vision-language model [36], which enhances interpretability through natural language descriptions (Fig. 1b). In addition, several binary detectors [10, 51, 60] leverage CLIP's robust recognition capabilities to achieve impressive performance. However, three key limitations remain in these works. First, DDVQA-BLIP relies on a general text-generation model without dedicated mechanisms for deepfake detection, resulting in lower detection accuracy compared to conventional binary detectors. Secondly, prior CLIP-based detectors often lack effective input text prompts to describe diverse forgeries, restricting the adaptation of CLIP's multi-modal learning ability in the detection task. Third, while CLIP's open-set recognition capability — enabling it to identify diverse visual semantics — is successfully combined with LLMs in domains like document parsing [25, 47, 81] and medical diagnosis [33, 49, 85], its integration with LLMs for deepfake detection remains largely unexplored.



The person has a stain on forehead and an unnatural mouth, suggesting this is an AI-generated image.

Figure 2. (a) Forgery Prompt Learning (FPL) adapts CLIP to deepfake detection by optimizing *UF-prompts* and *layer-wise forgery tokens* (LF tokens). UF-prompts consist of three segments: trainable general forgery tokens (*i.e.*, $\mathbf{V}_1$, $\mathbf{V}_2$, and $\mathbf{V}_3$), specific forgery tokens (*i.e.*, $\mathbf{V}_4$, $\mathbf{V}_5$, and $\mathbf{V}_6$), and a fixed textual description `` ``Forged Face'' ``. LF tokens are introduced in the CLIP text encoder and detailed in Fig. 3b and Sec. 3.2.1. (b) The Bridge Adapter connects the CLIP image encoder to the deepfake detector. It integrates with an LLM and a classification head, which output textual explanations and a predicted fake score, respectively.

To address these limitations, we propose a multi-modal face forgery detector (M2F2-Det), which contains dedicated forgery detection mechanisms for accurate detection and generating convincing textual explanations (Fig. 1c): explanations enhance detection's trustworthiness, while accurate detection, in turn, supports reliable explanation generation through effective representation learning. Moreover, the M2F2-Det introduces Forgery Prompt Learning, an efficient adaptation strategy that produces discriminative text embeddings for diverse forged face images. We also introduce a Bridge Adapter to leverage the frozen CLIP image encoder, enhancing M2F2-Det's detection performance and facilitating its integration with the LLM for textual explanation generation.

Forgery Prompt Learning (FPL) comprises two key components: *universal forgery prompts* (UF-prompts) and *layer-wise forgery tokens* (LF-tokens) (Fig. 2a). First, UF-prompts include both general and specific forgery tokens: general forgery tokens capture common forgery patterns and invariants shared across various manipulated facial images — critical for generalizing to unseen forgeries; specific forgery tokens, by contrast, encode fine-grained, image-dependent artifacts, such as blurred eyes from attribute manipulation and blending boundaries from face swapping. Secondly, we freeze the CLIP text encoder and introduce trainable layer-wise forgery tokens as inputs to its Transformer [71] layers (Fig. 3b). These task-specific tokens improve CLIP's adaptability to deepfake while largely preserving the recognition ability of its pre-trained weights.

We further propose a Bridge Adapter (Bri-Ada) to harness capabilities of the pre-trained CLIP image encoder for both forgery detection and explanation generation. As depicted in Fig. 2b, the Bri-Ada reuses intermediate features from the CLIP image encoder, preserving its foundational strengths in representation learning, which proves generalizable enough to identify unseen forgeries [10, 51, 60]. To enhance domain-specific discrimination, Bri-Ada incorporates a deepfake encoder that provides forgery-aware knowledge, enabling the construction of more robust and effective visual representations for deepfake detection. In addition, Bri-Ada is employed jointly with FPL in M2F2-Det (Fig. 3a). Such text embeddings generated by FPL are used to produce forgery attention maps, serving as prior knowledge to guide forgery identification. Furthermore, Bri-Ada's output is connected to the LLM, which leverages CLIP's open-set recognition capability to translate visual features into textual explanations. Specifically, Bri-Ada's output is transformed into a frequency-based token. This token then is concatenated with tokens from other modalities to guide the LLM in generating trustworthy explanations for deepfake detection, detailed in Sec. 3.2.3. In summary, our contributions are:

⋄ We propose a multi-modal face forgery detector, M2F2-Det, which innovatively outputs both deepfake detection scores and textual explanations, achieving remarkable detection accuracy and enhanced interpretability.

⋄ M2F2-Det introduces a Forgery Prompt Learning mechanism—automated and effective prompt learning tailored for deepfake detection—that transfers CLIP's powerful multi-modal learning ability into deepfake detection.

⋄ M2F2-Det employs a Bridge Adapter that enhances the integration of LLM, facilitating the generation of trustworthy explanations for detection decisions.

⋄ M2F2-Det achieves state-of-the-art (SoTA) deepfake detection performance, measured by 6 datasets, showing the effectiveness of capturing diverse forgeries. It also obtains SoTA explanation generation performance on the DD-VQA dataset [86], both quantitatively and qualitatively.

2. Related Works

Deepfake Detection. The image forensics community develops various effective deepfake detection techniques, in-

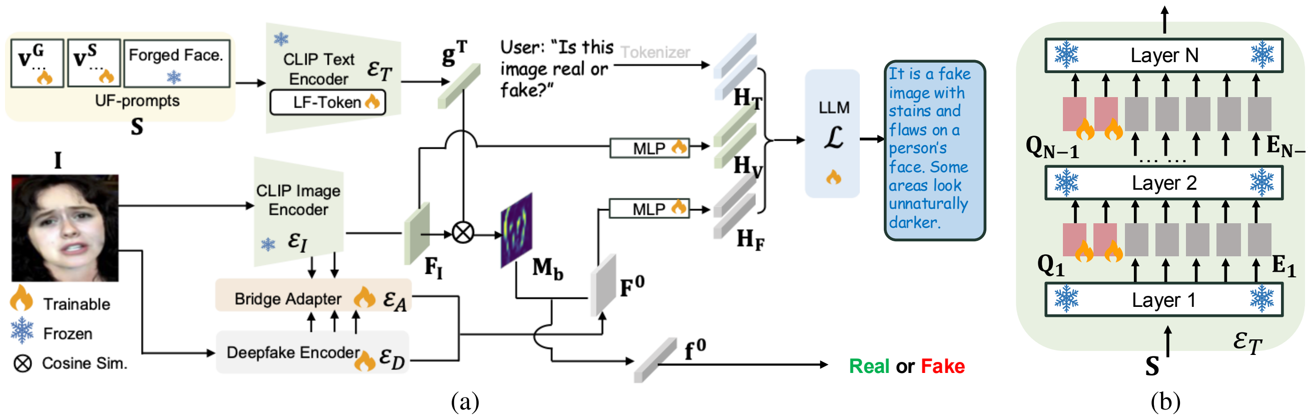

$\mathbf{v}^{\mathrm{G}}$
$\mathbf{v}^{\mathrm{S}}$
Forged Face.
UF-prompts
$\mathbf{S}$
CLIP Text Encoder $\mathcal{E}_T$
LF-Token
$\mathbf{g}^{\mathrm{T}}$
User: "Is this image real or fake?"
Tokenizer
$\mathbf{H}_{\mathrm{T}}$
$\mathbf{H}_{\mathrm{V}}$
$\mathbf{H}_{\mathrm{F}}$
MLP
MLP
LLM $\mathcal{L}$
It is a fake image with stains and flaws on a person's face. Some areas look unnaturally darker.
$\mathbf{I}$
CLIP Image Encoder $\mathcal{E}_I$
$\mathbf{F}_{\mathrm{I}}$
$\mathbf{M}_{\mathrm{b}}$
$\mathbf{F}^0$
Bridge Adapter $\mathcal{E}_A$
Deepfake Encoder $\mathcal{E}_D$
$\mathbf{f}^0$
Real or Fake
Trainable
Frozen
Cosine Sim.
(a)
Layer N
Layer 2
Layer 1
$\mathbf{Q}_{\mathrm{N-1}}$
$\mathbf{E}_{\mathrm{N-1}}$
$\mathbf{Q}_1$
$\mathbf{E}_1$
$\mathbf{S}$
$\mathcal{E}_T$
(b)

Figure 3. (a) The multi-modal face forgery detector (M2F2-Det) comprises pre-trained CLIP image and text encoders (*i.e.*, $\mathcal{E}_I$ and $\mathcal{E}_T$), a deepfake encoder, as well as an LLM. Given the universal forgery prompts (UF-prompts) as input, $\mathcal{E}_T$ generates a global text embedding, *e.g.*, $\mathbf{g}^{\mathrm{T}}$, that guides the generation of a forgery attention mask, *e.g.*, $\mathbf{M}_b$. The deepfake encoder utilizes the bridge adapter, *i.e.*, $\mathcal{E}_A$, for detecting face forgeries (Sec. 3.2.2), while the LLM generates explanations conditioned on a frequency token $\mathbf{H}_{\mathrm{F}}$ transformed from the forgery representation ($\mathbf{F}^0$) (Sec. 3.2.3). (b) In the CLIP text encoder, we introduce trainable layer-wise forgery tokens as inputs to each Transformer [71] encoder layer.

cluding data augmentation [11, 39, 40, 61, 90], frequency clues [44, 48, 74], disentanglement leanring [43, 52, 64, 65, 78, 79], specified networks [21, 28, 68, 73, 80], and biometric information analysis [41, 67, 96]. These works belong to the conventional deepfake detector that outputs binary prediction scores. Recently, DDVQA-BLIP [86] defines a paradigm that generates textual explanations for deepfake detection, enhancing interpretability. In contrast, the M2F2-Det outputs both a prediction score and textual explanations, using the latter to enhance detection interpretability, while the prediction score helps more convincing explanations. Moreover, unlike prior CLIP-based forgery detectors [10, 51, 60] that only rely on detection capabilities of pre-trained CLIP, our M2F2-Det further integrates the open-set visual recognition ability of the CLIP image encoder for enhanced interpretability.

Prompt Learning. Prompt learning offers an efficient strategy adapting the pre-trained CLIP to downstream tasks [27, 31, 62, 66, 88, 92, 94]. For example, CoOp [93] and CoCoOp [92] integrate continuous prompts in the textual space, enhancing the pre-trained CLIP's generalizability. Meanwhile, MaPLE [30] and VPT [27] modify the learning procedure in visual spaces by altering the CLIP image encoder. These works learn global image information for recognition tasks while our FPL conducts the pixel-wise task to localize facial forgeries.

Multimodal Vision-Language Models (MLLMs) [36, 37, 45, 82] use generative capabilities from LLMs [70, 84] to obtain impressive performance across a wide range of tasks. For example, early studies concentrate on generating text-based content grounded on image, video, and audio [3, 13, 38, 45, 76, 83]. Recently, MLLMs broaden applications to more complex downstream domains, including embodied AI [54, 87], document parsing [25, 47, 81], and medical diagnosis [33, 49, 85]. We propose a frequency token that implicitly aligns deepfake domain knowledge with MLLM, bridging the gap between language and subtle facial forgeries.

3. Method

3.1. Preliminaries

We denote the input image and text prompts as $\mathbf{I}$ and $\mathbf{S}$, respectively. The proposed M2F2-Det utilizes CLIP's image and text encoders, $\mathcal{E}_I$ and $\mathcal{E}_T$, together with a deepfake encoder, $\mathcal{E}_D$, for forgery detection. Also, a large language model (*i.e.*, $\mathcal{L}$) is employed to generate textual explanations.

Prompt Learning. Contrastive Language-Image Pre-training, known as CLIP [56], is a large-scale vision-language foundation model that has powerful zero-shot classification capabilities. Given a set of K text prompts $\{\mathbf{S}_1, \mathbf{S}_2, \dots, \mathbf{S}_K\}$, CLIP can estimate the likelihood that $\mathbf{I}$ corresponds to each of these text prompts:

$$p(y|\mathbf{I}) = \frac{\exp\left(\langle \mathcal{E}_I(\mathbf{I}), \mathcal{E}_T(\mathbf{S}_k)\rangle / \tau\right)}{\sum_{k=1}^{K} \exp\left(\langle \mathcal{E}_I(\mathbf{I}), \mathcal{E}_T(\mathbf{S}_k)\rangle / \tau\right)}, \tag{1}$$

where $\langle \cdot, \cdot \rangle$ and τ denote cosine similarity and a temperature hyper-parameter, respectively. To enhance the pre-trained CLIP's performance on downstream tasks, CoOp [93] proposes the prompt learning strategy that uses trainable tokens to automatically learn effective text prompts as follows:

$$\mathbf{S}_k = [\mathbf{v}_1][\mathbf{v}_2]\dots[\mathbf{v}_n][\mathtt{class}_k], \tag{2}$$

where $[\mathbf{v}_1][\mathbf{v}_2]\dots[\mathbf{v}_n]$ ($\mathbf{v}_n \in \mathbb{R}^d$) are trainable tokens, and $[\mathtt{class}_k]$ represents the fixed and non-trainable class name of the k-th class.

Visual Instruction Tuning. MLLMs tackle complicated reasoning tasks by generating responses based on visual and textual inputs. In general, a MLLM consists of three main components: 1) a pre-trained image encoder, *e.g.*, $\mathcal{E}_I$, that transforms $\mathbf{I}$ into a set of visual features. 2) a projector, *e.g.*, `MLP` layers, that converts visual features to visual tokens denoted as $\mathbf{H}_\mathrm{V} \in \mathbb{R}^{N\times D}$. 3) an LLM, *i.e.*, $\mathcal{L}$, that generates free-form responses in an auto-regressive manner when prompted with $\mathbf{H}_\mathrm{V}$ and textual tokens $\mathbf{H}_\mathrm{T} \in \mathbb{R}^{M\times D}$. $\mathbf{H}_\mathrm{T}$ is generated by the tokenizer that takes user-input questions. Let us define target answer tokens as $\mathbf{X}_\mathrm{A} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_\mathrm{z}] \in \mathbb{R}^{Z\times D}$, where Z represents the sequence length, then the probability of generating $\mathbf{X}_\mathrm{A}$ becomes

$$p(\mathbf{X}_\mathrm{A} \mid \mathbf{H}_\mathrm{V}, \mathbf{H}_\mathrm{T}) = \prod_{\mathrm{z}=1}^{Z} p_\theta(\mathbf{x}_\mathrm{z} \mid \mathbf{H}_\mathrm{V}, \mathbf{H}_{\mathrm{T},<\mathrm{z}}, \mathbf{x}_{\mathrm{A},<\mathrm{z}}), \tag{3}$$

where θ are trainable parameters; $\mathbf{H}_{\mathrm{T},<\mathrm{z}}$ and $\mathbf{x}_{\mathrm{A},<\mathrm{z}}$ are instruction and answer tokens in all turns before the current prediction token $\mathbf{x}_\mathrm{z}$, respectively.

3.2. Multi-modal Face Forgery Detector

3.2.1. Forgery Prompt Learning

Forgery Prompt Learning captures forgeries via Universal Forgery Prompts *e.g.*, UF-prompts, that contain two types of learnable tokens, *e.g.*, general-forgery and specific-forgery tokens, *e.g.*, $[\mathbf{v}^\mathrm{G}] \in \mathbb{R}^d$ and $[\mathbf{v}^\mathrm{S}] \in \mathbb{R}^d$, respectively. Formally, we use `MLP` layers to transform the global visual embedding $\mathbf{g}^\mathrm{I} \in \mathbb{R}^d = \mathcal{E}_I(\mathbf{I})$ into $[\mathbf{v}^\mathrm{S}]$, which helps inject image-dependent information into $[\mathbf{v}^\mathrm{S}]$. Consequently, $[\mathbf{v}^\mathrm{G}]$ and $[\mathbf{v}^\mathrm{S}]$ are optimized together to leverage the powerful multi-modal representation capability of the CLIP for capturing both general and image-specific forgery patterns. Next, without loss of generality, we use ``` ``forged face'' ``` as the generic textual description for various input face images, which stabilizes the training empirically. Therefore, we construct UF-prompts as:

$$\mathbf{S} = [\mathbf{v}_1^\mathrm{G}] \ldots [\mathbf{v}_m^\mathrm{G}][\mathbf{v}_1^\mathrm{S}] \ldots [\mathbf{v}_u^\mathrm{S}]\texttt{[forged][face]}, \tag{4}$$

where $m \in \{0, 1...M\}, u \in \{0, 1...U\}$; `forged` and `face` are non-trainable tokens converted by fixed words.

To enhance the conversion of $\mathbf{S}$ into textual embeddings that facilitate CLIP's adaptation, similar to the prior work [27], we introduce trainable layer-wise forgery tokens as inputs to each Transformer encoder layer of $\mathcal{E}_T$ while keeping its pre-trained weights frozen, depicted in Fig. 3b. More formally, for $\mathcal{E}_T$'s $(r+1)$-th layer, *e.g.*, $\mathcal{E}_T^{r+1}$, we denote collections of input O layer-wise forgery tokens and P ordinary tokens as $\mathbf{Q}_r = [\mathbf{q}_1, \mathbf{q}_2, ..., \mathbf{q}_o] \in \mathbb{R}^{O\times d}$ and $\mathbf{E}_r = [\mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_p] \in \mathbb{R}^{P\times d}$, respectively. Then, $\mathcal{E}_T$'s forward propagation of taking $\mathbf{S}$ is:

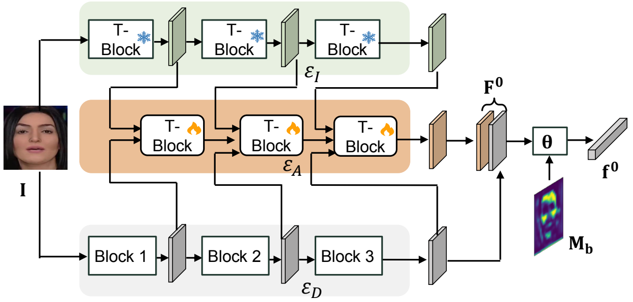


Figure 4. The illustration on the Bridge Adapter, in which Θ represents the transformation conducted by Eq. 7. [Key: T-Block: transformer encoder block; Block: convolution block.]

$$\mathbf{E}_0 = \texttt{Embed}(\mathbf{S}), \tag{5}$$

$$[\_, \mathbf{E}_{r+1}] = \mathcal{E}_T^{r+1}(\mathbf{Q}_i, \mathbf{E}_i), \tag{6}$$

where r is the layer index and `Embed` denotes the procedure that converts $\mathbf{S}$ into the d-dimensional latent space with positional embeddings.

Specifically, taking $\mathbf{S}$ as the input, the $\mathcal{E}_T$ outputs the global textual embedding, denoted as $\mathbf{g}^\mathrm{T} \in \mathbb{R}^d$. Meanwhile, we obtain output features from the last layer of $\mathcal{E}_I$, denoted as $\mathbf{F}_\mathrm{I} \in \mathbb{R}^{N\times d} = \mathbb{R}^{W\times H\times d}$. Then, a forged attention map, *i.e.*, $\mathbf{M}_b \in \mathbb{R}^{W\times H}$, can be obtained by calculating the text-to-image score for its each patch, *i.e.*, $\mathbf{M}_b^{ij} = \langle \mathbf{F}_\mathrm{I}^{ij}, \mathcal{E}_T(\mathbf{S}) \rangle$, where $\mathbf{M}_b^{ij}$ and $\langle\cdot,\cdot\rangle$ represent text-to-image score for (i, j)-th patch of $\mathbf{M}_b$ and a cosine similarity operation, respectively. Next, we use $\mathbf{M}_b$ to help detect global image-level forgeries because it provides prior knowledge of the spatial forgery location, and this technique is similar to prior works [20, 63, 89]. Please note unlike prior works [27, 31, 92, 93] that adopt the prompt learning to capture global information of images, we innovatively propose the FPL to conduct a pixel-wise task of localizing forged regions (*e.g.*, $\mathbf{M}_b$).

3.2.2. Bridge Adapter

As depicted in Fig. 4, the Bridge Adapter (Bri-Ada), *i.e.*, $\mathcal{E}_A$, is composed of Transformer encoder blocks and takes intermediate features from both $\mathcal{E}_I$ and $\mathcal{E}_D$ as inputs. In this way, $\mathcal{E}_A$ fully leverages the detection and open-set recognition capabilities of $\mathcal{E}_I$, further enriched by domain knowledge from $\mathcal{E}_D$ to produce more robust and effective representations for deepfake detection.

Bri-Ada is jointly used with FPL, which generates local forged attention maps. Specifically, we concatenate feature maps output from $\mathcal{E}_D$ and $\mathcal{E}_A$ into a fused feature map $\mathbf{F}^0 \in \mathbb{R}^{w\times h\times c}$, as illustrated in Fig. 4. We then use $\mathbf{M}_b$ and $\mathbf{F}^0$ to obtain refined forgery vector, *e.g.*, $\mathbf{f}^0 \in \mathbb{R}^d$, as the final forgery representation for deepfake detection. Specifically, we have

$$\mathbf{f}^0 = \texttt{AVGPOOL}(\texttt{CONV}(\mathbf{F}^0 \odot \mathbf{M}_b)), \tag{7}$$

where AVGPOOL and CONV represent the average pooling operation and convolution layers, respectively.

The joint use of FPL and $\mathcal{E}_A$ for detection has two key advantages. First, FPL and $\mathcal{E}_A$ mutually benefit each other. For global detection, $\mathcal{E}_A$ reuses intermediate features from $\mathcal{E}_D$ and then employs $\mathbf{M}_b$ as prior knowledge of forgery regions. This encourages FPL to update $\mathbf{S}$ and $\mathbf{Q}_i$, such that $\mathbf{g}^{\mathrm{T}}$ can be used to generate accurate forged attention maps. For generating forgery attention maps, FPL needs feedback from $\mathcal{E}_A$ and $\mathcal{E}_D$, on if its generated $\mathbf{M}_b$ enhances binary detection. Secondly, M2F2-Det is only supervised by the binary ground truth label, indicating that $\mathbf{M}_b$ is learned via an efficient unsupervised manner.

3.2.3. Forgery Explanation Module

The Forgery Explanation Module helps the M2F2-Det generate texts, as depicted in Fig. 3a. These generated texts contain the judgment and explanation, which claims if the image is forged and explicitly describes the rationale behind this decision, respectively. Specifically, we obtain the representation $\mathbf{F}^0$ for the detection task. We convert it into $\mathbf{H}_{\mathrm{F}} \in \mathbb{R}^{N \times D}$. As a result, $\mathbf{H}_{\mathrm{F}}$ informs the LLM (*e.g.*, $\mathcal{L}$) if the input image is fake. Meanwhile, we transform $\mathcal{E}_I$'s output feature into visual tokens $\mathbf{H}_{\mathrm{V}}$, which helps $\mathcal{L}$ describe the facial pattern. Both $\mathbf{H}_{\mathrm{F}}$ and $\mathbf{H}_{\mathrm{V}}$ are fed into $\mathcal{L}$ for explanation generation. Therefore, we update the Eq. 3 into as follows:

$$p(\mathbf{X}_{\mathrm{A}} \mid \mathbf{H}_{\mathrm{V}}, \mathbf{H}_{\mathrm{F}}, \mathbf{H}_{\mathrm{T}}) = \prod_{z=1}^{Z} p_\theta(\mathbf{x}_{\mathrm{z}} \mid \mathbf{H}_{\mathrm{V}}, \mathbf{H}_{\mathrm{F}}, \mathbf{H}_{\mathrm{T},<z}, \mathbf{x}_{\mathrm{A},<z}). \tag{8}$$

3.3. Train and Inference

Training. First, we train the deepfake encoder $\mathcal{E}_D$ as well as $\mathbf{S}$ and $\mathbf{Q}$ in FPL, such that M2F2-Det can perform the binary deepfake classification. We minimize the cross entropy distance between binary classification probability $p(y|\mathbf{I})$ and a ground truth categorical $\hat{y}$.

Secondly, we align $\mathbf{H}_{\mathrm{V}}$ and $\mathbf{H}_{\mathrm{F}}$ with the input space of a frozen LLM, such that outputs from $\mathcal{E}_I$ and $\mathcal{E}_D$ can be interpreted by the LLM. More formally, we maximize the likelihood defined in Eq. 8 via only training MLP layers while freezing other components in this stage.

Thirdly, to better tame the LLM for explanation generation, we again keep the entire model frozen while only updating MLP layers and LLM based on Eq. 8. Trainable parameters, *e.g.*, θ, thus become MLP layers and a subset of LLM's parameters. Please note we use LoRA for efficient LLM fine-tuning.

In second and third-stage training, we use the DD-VQA [86] dataset that contains high-quality image-text pairs annotated by the Amazon mechanical Terk. The DD-VQA dataset consists of $14,782$ question-answer pairs using the train/test IDs from FF++ [58]. This results in $13,559$ question-answer pairs for training and $1,223$ pairs for testing. Note that the second and third training stages are similar to LLaVA [45], and the difference is we align one more representation, *i.e.*, $\mathbf{H}_{\mathrm{F}}$.

Dataset	**Real Samples**	**Fake Samples**
FF++ [58]	$1,000$ V	$4,000$ V
CDF [42]	590 V	$5,639$ V
DFD [1]	363 V	$3,068$ V
WDF [97]	$3,805$ I	$3,509$ I
DFDC [17]	$1,131$ V	$4,113$ V
FFIW [95]	$10,000$ V	$10,000$ V

Table 1. Six datasets used for evaluating detection performance. [Key: V: Video; I: Image].

Inference. Given the input image and user-input questions, we produce the binary result and textual explanations. User-input questions can be flexible, such as ``Determine the authenticity of the image.'' and ``Is this image real or fake?''.

4. Experiment

4.1. Setup

Datasets. We evaluate our method on both detection and explanation generation tasks. For detection, we compare against existing detection approaches using datasets listed in Tab. 1, including FaceForensics++ (FF++) [58], CelebDF [42], WildDeepfake (WDF) [97], DFD [17], DFDC [14], and FFIW [95]. For explanation generation, we evaluate on the publicly available DD-VQA dataset [86].

Metrics. First, we use Area Under the Curve (AUC) and accuracy to measure the detection performance. Second, for explanation generation performance, we evaluate *judgment performance* and *explanation quality*. For judgment performance, we extract keywords, *e.g.*, "Fake" and "Real", from generated texts to compute accuracy and F1-Score. To assess explanation quality, we employ standard natural language generation metrics such as BLUE-4 [53], CIDEr [72], ROUGE_L [59], METEOR [12], and SPICE [2]. These metrics evaluate the similarity between generated and annotated textual answers comprehensively. Additional evaluation details are provided in the supplementary material.

Implmentation Details. We employ EfficientNet-B4 [69] as the deepfake detector, *i.e.*, $\mathcal{E}_D$. Additionally, we use the CLIP/ViT-L-patch14-336 model [16] for the pre-trained CLIP image encoder ($\mathcal{E}_I$) and text encoder ($\mathcal{E}_T$). The LLM is Vicuna-7b [8], and more details are in the supplementary.

Methods	Venue	FF++ (c23)		FF++ (c40)		Celeb-DF		WDF	
		Metric: Acc (% ↑) / AUC (% ↑)							
RFM [73]	CVPR21	95.69	98.79	87.06	89.83	97.96	99.94	77.38	83.92
Freq-SCL [35]	CVPR21	96.69	99.28	89.00	92.39	–	–	–	–
Add-Net [97]	ACMMM20	96.78	97.74	87.50	91.01	96.93	99.55	76.25	86.17
F3-Net [55]	ECCV20	97.52	98.10	90.43	93.30	95.95	98.93	80.66	87.53
MultiAtt [89]	CVPR21	97.60	99.29	88.69	90.40	97.92	**99.94**	82.86	90.71
RECCE [5]	CVPR22	97.06	99.32	91.03	95.02	98.59	**99.94**	83.25	92.02
TALL [77]	ICCV23	98.65	**99.87**	92.82	94.57	97.57	98.55	-	-
DDVQA-BLIP [86]	ECCV24	80.69	-	72.73	-	-	-	-	-
M2F2-Det		**98.79**	99.34	**93.83**	**96.58**	**98.98**	99.92	**86.05**	**93.14**

Table 2. Intra-dataset Detection Performance. Results of prior works are mainly cited from [5, 77]. [Key: **Best**, Second Best].

Method	Venue	Training set		Test set AUC (% ↑)			
		Real	Fake	DFDC	FFIW	Celeb-DF	DFD
LocalRL [7]	AAAI21	✓	✓	76.53	-	78.26	89.24
CADDM [78]	CVPR23	✓	✓	-	-	93.88	99.03
UCF [78]	ICCV23	✓	✓	80.50	-	82.40	94.50
SBI [61]	CVPR22	✓		86.15	84.83	93.18	97.56
AUNet [4]	CVPR23	✓		86.16	81.45	92.77	**99.22**
Seeable [32]	ICCV23	✓	-	86.30	-	87.30	-
LAA-Net (w/ SBI) [50]	CVPR24	✓		86.94	-	**95.40**	98.43
FreqBlender [34]	NeurIPS24	✓		87.56	86.14	94.59	-
M2F2-Det (w/SBI)		✓		**87.80**	**88.70**	95.10	97.70

Table 3. Inter-dataset Detection Performance. Results of prior works are mainly cited from [4, 34, 50, 61]. [Key: **Best**, Second Best].

4.2. Detection Performance

Tab. 2 and 3 report detection performance on intra- and inter-dataset setups, respectively.

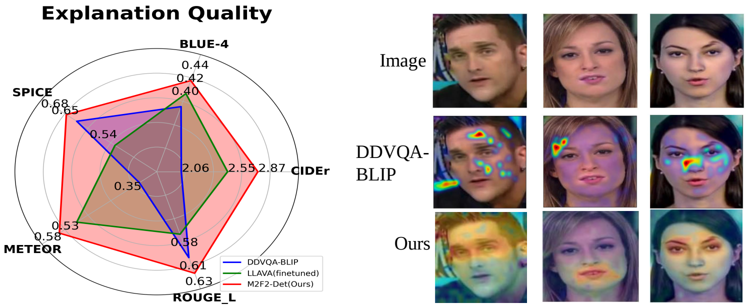
Intra-dataset performance. Tab. 2 shows our M2F2-Det achieves the best overall detection performance in FF++. For example, in FF++ (c40), M2F2-Det has 1.01% higher accuracy and 2.01% higher AUC than the second-best method, *i.e.*, TALL. Please note that TALL takes multiple frames as inputs, which generally contain more forgery information than using one frame as the input. While being a detector that takes single-frame input, our method still outperforms TALL, indicating the effectiveness of the proposed Forgery Prompt Learning and Bridge Adapter. Such effectiveness can be further demonstrated in Celeb-DF, in which our method surpasses TALL by 1.41% accuracy and 1.37% AUC score.

Furthermore, WDF collects diverse real-life forged faces from the web on which our method outperforms the second-best performer, *e.g.*, RECCE, by 2.80% and 1.12% in accuracy and AUC score, respectively. One key difference between RECCE and our method is that we adapt the pre-trained CLIP image encoder, which is also trained on web samples. Such an adaptation improves M2F2-Det's detection ability by leveraging CLIP's robust representation, which proves generalizable enough to detect unseen forgeries [10, 51]. DDVQA-BLIP [86] predicts fake images based on if keywords like "`fake`" in generated sentences, whereas its detection accuracy is 18.1% and 21.1% lower than us on FF++ c23 and c40, respectively. This shows accurate deepfake detection requires a specific mechanism that learns forgeries, like using forged attention masks as local forged contexts, instead of applying the text-generation model [36].

Inter-dataset performance. Tab. 3 reports the performance, in which we follow prior works [50, 61] to train methods on real and pseudo-fake images from FF++. Our M2F2-Det achieves 0.24% and 2.56% higher AUC than FreqBlender [34] on DFDC and FFIW, respectively. We believe our method's superior generalization ability comes from the usage of a pre-trained CLIP image encoder. Specifically, the CLIP is trained on diverse real-world web samples instead of a forgery detection dataset, making its learned features more generalizable and less overfitting on specific forgery patterns [10, 51, 60]. Moreover, M2F2-Det outperforms AUNet with higher AUC scores on three datasets but performs worse on the DFD dataset. AUNet learns forgery clues from relations between different facial action units. Similarly, specific forgery information from action units could also be considered in M2F2-Det's FPL, which uses UF-prompts and layer-wise forgery tokens to adapt CLIP in discerning both general and specific facial forgeries. The learned text embedding is used to generate forged attention maps, as depicted in Fig. 7, that act as local forged contexts to help the global detection task.

Method	Fine-tuned	Detection	
		Acc ↑	F1 ↑
DDVQA-BLIP	✔	87.49	90.07
LLaVA	✗	51.41	37.04
LLaVA	✔	86.41	92.10
Qwen	✗	59.41	51.04
InternVL	✗	42.33	47.35
M2F2-Det *w/o* $\mathbf{H}_{\mathrm{F}}$	✔	85.11	84.21
M2F2-Det	✔	**95.23**	**96.61**

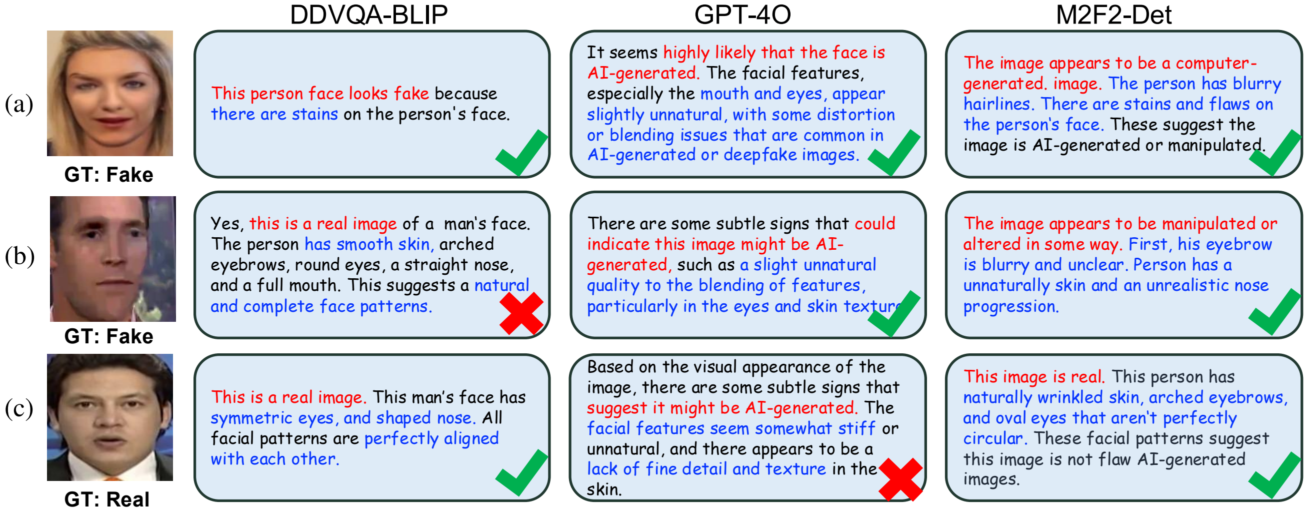
(a)



(b)

(c)

Figure 5. Explanation generation performance on DD-VQA. (a) Judgment performance. [Key: **Best results**, Acc: Accuracy, F1: F1 Score] (b) Explanation quality is measured by 5 metrics. (c) Visualizations of forged attentions.



(a)

This person face looks fake because there are stains on the person's face. ✓

It seems highly likely that the face is AI-generated. The facial features, especially the mouth and eyes, appear slightly unnatural, with some distortion or blending issues that are common in AI-generated or deepfake images. ✓

The image appears to be a computer-generated. image. The person has blurry hairlines. There are stains and flaws on the person's face. These suggest the image is AI-generated or manipulated. ✓

(b)

Yes, this is a real image of a man's face. The person has smooth skin, arched eyebrows, round eyes, a straight nose, and a full mouth. This suggests a natural and complete face patterns. ✗

There are some subtle signs that could indicate this image might be AI-generated, such as a slight unnatural quality to the blending of features, particularly in the eyes and skin texture ✓

The image appears to be manipulated or altered in some way. First, his eyebrow is blurry and unclear. Person has a unnaturally skin and an unrealistic nose progression. ✓

(c)

This is a real image. This man's face has symmetric eyes, and shaped nose. All facial patterns are perfectly aligned with each other. ✓

Based on the visual appearance of the image, there are some subtle signs that suggest it might be AI-generated. The facial features seem somewhat stiff or unnatural, and there appears to be a lack of fine detail and texture in the skin. ✗

This image is real. This person has naturally wrinkled skin, arched eyebrows, and oval eyes that aren't perfectly circular. These facial patterns suggest this image is not flaw AI-generated images. ✓

Figure 6. Qualitative results of explanation generation from DDVQA-BLIP, GPT-4o, and M2F2-Det.

4.3. Explanation Generation Performance

Quantative Result. Fig. 5a reports explanation generation performance regarding *judgment performance* and *explanation quality*. First, from the judgment perspective, our method achieves the best accuracy and F1 score, which have 7.74% higher accuracy than DDVQA-BLIP and 4.51% higher F1 score than fine-tuned LLaVa. This demonstrates the advantage of the proposed frequency-based token (*i.e.*, $\mathbf{H}_{\mathrm{F}}$), which helps our method generate correct descriptions based on learned deepfake domain knowledge. The usage of $\mathbf{H}_{\mathrm{F}}$ is different from previous MLLMs that only judge if visual artifacts exist in the color domain, which is less effective in capturing the discrepancy between real and fake images at frequency domains. Fig. 5a shows that a removal of $\mathbf{H}_{\mathrm{F}}$ decreases M2F2-Det's judgment performance by 10.12% in accuracy. Secondly, Fig. 5b shows that M2F2-Det achieves the best explanation quality. Specifically, M2F2-Det has the best CIDEr score, which measures semantic consistency between generated sentences and ground truth (GT): Fig. 6 shows other works with erroneous judgments and explanations, different from GT's semantics, hence causing low CIDEr. Also, ROUGE_L measures if generated explanations summarize GT's information with lexical variations, and our results include more diverse phrases, *e.g.*, `naturally wrinkled skin` in Fig. 6's 3rd sample. Fig. 5c shows learned forged maps used in the M2F2-Det can better identify artifacts and forgeries than the cross-attention mechanism of DDVQA-BLIP.

Qualitative Performance. Fig. 6 reports qualitative results, where our M2F2-Det generates explanations with accurate judgments and convincing explanations for both real and fake images. First, for image (a), DDVQA-BLIP cannot offer detailed explanations, whereas our model provides convincing reasonings, *i.e.*, `blurry hairlines`. For (b), M2F2-Det again provides correct judgment and sophisticated explanations, *e.g.*, clearly identifying unnatural patterns of eye regions and unnatural skin textures. We believe that such performance superiority of M2F2-Det comes from the effective deepfake representation (*i.e.*, $\mathbf{H}_{\mathrm{F}}$). Lastly, we also show generated sentences on a real image (c), showing M2F2-Det's effectiveness in discerning genuine faces.

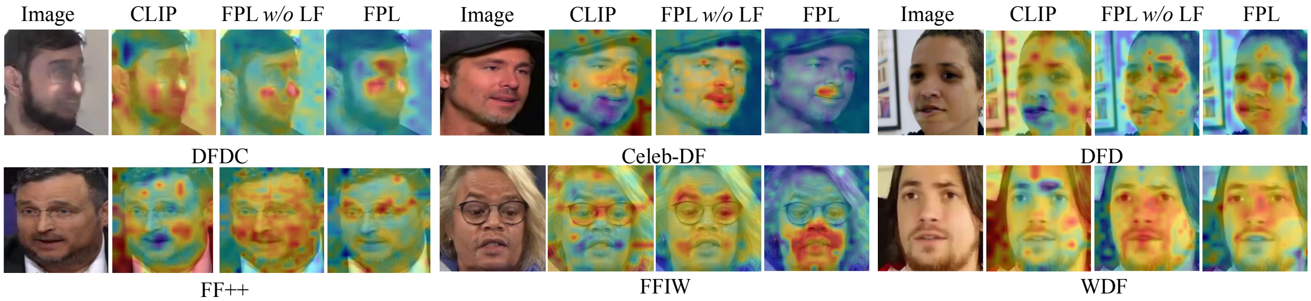


Figure 7. Forged attention maps on samples from 6 datasets introduced in Tab. 1. [Key: LF: layer-wise forgery tokens]

	Forgery PL		Bri-Ada	Test set AUC (%)	
	UF-P	LF		FF++(c40)	Celeb-DF
1				91.03	65.78
2		✓		92.57	67.37
3	✓			92.66	66.08
4	✓	✓		93.65	68.68
5			✓	93.80	70.71
6	✓		✓	94.20	71.08
7	✓	✓	✓	**96.58**	**74.82**

Table 4. Ablation Study. Each model is trained by FF++(c40) and tested on FF++(c40) and Celeb-DF. [Keys: Forgery PL: Forgery Prompt Learning; UF-P: Universal Forgery Prompts; LF: layer-wise forgery tokens; Bri-Ada: Bridge Adapter.]

Method	AUC
Uni.Fake [51]	72.40
DEFAKE [60]	76.84
M2F2-Det	**96.58**

(a)

Method	AUC
CoOp [93]	71.44
CoCoOp [92]	72.62
FPL	**80.75**

(b)

Table 5. Comparisons to (a) existing CLIP-based forgery detection methods and (b) prompt learning methods. The performance is reported on FF++(c40).

4.4. Ablative Study and Analysis

Forged Attention Map. Tab. 4's line #1 represents the deepfake detection baseline performance, *e.g.*, EfficientNet-B4 [69]. Line #2 and #3 show layer-wise forgery tokens (LF tokens) and UF-prompts improve performance — *e.g.*, 1.54% and 1.63% higher AUC scores on FF++, respectively. This shows that LF tokens and UF-prompts are effective designs to learn deepfake domain knowledge. Furthermore, in line #4, we employ both UF-prompts and LF tokens, which is the full version of FPL, further increasing the performance of line #2 by 1.08% AUC on FF++. This is because accurate forged attention maps can assist detection. For example, in Fig. 7's first sample, when using FPL, the person's nose and eye areas are precisely identified as forged regions. Such visualizations further demonstrate the quality of forged attention maps obtained from the FPL.

Bridge Adapter. The comparison between lines #1 and #5 demonstrates the effectiveness of the Bri-Ada — increasing baseline's performance, *e.g.*, line 1, by 2.77% AUC score on FF++. In addition, it enhances the generalizable detection ability and increases performance on Celeb-DF by 4.93%. We believe this is because Bri. Ada. employs the CLIP image encoder that helps become less overfitting on specific manipulation types of FF++ samples. Line #6 indicates the integration between Bri-Ada and FPL further increases detection performance. Lastly, the full M2F2-Det (*i.e.*, line #7) achieves the best overall performance.

Comparison to CLIP-based forgery detectors. Tab. 5a shows that M2F2-Det achieves 24.18% and 19.74% higher AUC scores than previous CLIP-based image forensic methods, Uni-Fake [51] and DEFAKE [60], respectively. These two methods use the CLIP image encoder with simple architecture, *e.g.*, linear layers and ResNet-18, which lack specific facial forgery mechanisms. In contrast, the M2F2-Det not only employs the CLIP but also uses specified forgery detection mechanisms like FPL.

Comparison to Prompt learning methods. We compute cosine similarities between CLIP image and text encoders' outputs for detection after applying different prompt learning schemes, and the performance is reported in Tab. 5b. Our FPL achieves 9.31% and 8.13% higher AUC scores than CoOp [93] and CoCoOp [92], respectively. This is because prior works are developed to adapt the CLIP to tasks that focus on recognizing semantics, which is different from deepfake detection.

5. Conclusion

In this work, we introduce M2F2-Det for interpretable deepfake detection. Specifically, M2F2-Det uses the Forgery Prompt Learning to adapt CLIP's multi-modal learning ability for deepfake detection. Then, an efficient Bridge Adapter connects the CLIP image encoder with a dedicated deepfake detection network, yielding more robust and effective visual representations for detection and seamlessly integrating with an LLM for enhanced interpretability.

References

[1] Contributing data to deepfake detection research. https://ai.googleblog.com/2019/09/contributing-data-to-deepfake-detection.html. Accessed: 2021-11-13. 5

[2] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pages 382–398. Springer, 2016. 5

[3] Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*, 2023. 3

[4] Weiming Bai, Yufan Liu, Zhipeng Zhang, Bing Li, and Weiming Hu. Aunet: Learning relations between action units for face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24709–24719, 2023. 6

[5] Junyi Cao, Chao Ma, Taiping Yao, Shen Chen, Shouhong Ding, and Xiaokang Yang. End-to-end reconstruction-classification learning for face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4113–4122, 2022. 6

[6] Liang Chen, Yong Zhang, Yibing Song, Lingqiao Liu, and Jue Wang. Self-supervised learning of adversarial example: Towards good generalizations for deepfake detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18710–18719, 2022. 1

[7] Shen Chen, Taiping Yao, Yang Chen, Shouhong Ding, Jilin Li, and Rongrong Ji. Local relation learning for face forgery detection. In *Proceedings of the AAAI conference on artificial intelligence*, pages 1081–1088, 2021. 6

[8] Wei-Lin Chiang, Zhuohan Li, Ziqing Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6, 2023. 5

[9] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation. In *CVPR*, 2018. 1

[10] Davide Cozzolino, Giovanni Poggi, Riccardo Corvi, Matthias Nießner, and Luisa Verdoliva. Raising the bar of ai-generated image detection with clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4356–4366, 2024. 1, 2, 3, 6

[11] Debayan Deb, Xiaoming Liu, and Anil Jain. Unified detection of digital and physical face attacks. In *arXiv preprint arXiv:2104.02156*, 2021. 3

[12] Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, pages 376–380, 2014. 5

[13] Soham Deshmukh, Benjamin Elizalde, Rita Singh, and Huaming Wang. Pengi: An audio language model for audio tasks. *Advances in Neural Information Processing Systems*, 36:18090–18108, 2023. 3

[14] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. The deepfake detection challenge dataset. *arXiv preprint arXiv:2006.07397*, 2020. 5

[15] Shichao Dong, Jin Wang, Renhe Ji, Jiajun Liang, Haoqiang Fan, and Zheng Ge. Implicit identity leakage: The stumbling block to improving deepfake detection generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3994–4004, 2023. 1

[16] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2020. 5

[17] Nicholas Dufour, Andrew Gully, Per Karlsson, Alexey Victor Vorbyov, Thomas Leung, Jeremiah Childs, and Christoph Bregler. Deepfakes detection dataset by Google & Jigsaw, 2019. 5

[18] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014. 1

[19] Qiqi Gu, Shen Chen, Taiping Yao, Yang Chen, Shouhong Ding, and Ran Yi. Exploiting fine-grained face forgery clues via progressive enhancement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 735–743, 2022. 1

[20] Xiao Guo, Xiaohong Liu, Zhiyuan Ren, Steven Grosz, Iacopo Masi, and Xiaoming Liu. Hierarchical fine-grained image forgery detection and localization. In *In Proceeding of IEEE Computer Vision and Pattern Recognition*, 2023. 1, 4

[21] Xiao Guo, Vishal Asnani, Sijia Liu, and Xiaoming Liu. Tracing hyperparameter dependencies for model parsing via learnable graph pooling network. In *Proceeding of Thirty-eighth Conference on Neural Information Processing Systems*, Vancouver, Canada, 2024. 3

[22] Xiao Guo, Xiaohong Liu, Iacopo Masi, and Xiaoming Liu. Language-guided hierarchical fine-grained image forgery detection and localization. *IJCV*, 2024. 1

[23] Xiao Guo, Manh Tran, Jiaxin Cheng, and Xiaoming Liu. Dense-face: Personalized face generation model via dense annotation prediction. *arXiv preprint arXiv:2412.18149*, 2024. 1

[24] Alexandros Haliassos, Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. 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*, pages 5039–5049, 2021. 1

[25] Anwen Hu, Yaya Shi, Haiyang Xu, Jiabo Ye, Qinghao Ye, Ming Yan, Chenliang Li, Qi Qian, Ji Zhang, and Fei Huang.

mplug-paperowl: Scientific diagram analysis with the multimodal large language model. In *ACM Multimedia 2024*, 2024. 1, 3

[26] Baojin Huang, Zhongyuan Wang, Jifan Yang, Jiaxin Ai, Qin Zou, Qian Wang, and Dengpan Ye. Implicit identity driven deepfake face swapping detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4490–4499, 2023. 1

[27] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision*, pages 709–727. Springer, 2022. 3, 4

[28] Shan Jia, Reilin Lyu, Kangran Zhao, Yize Chen, Zhiyuan Yan, Yan Ju, Chuanbo Hu, Xin Li, Baoyuan Wu, and Siwei Lyu. Can chatgpt detect deepfakes? a study of using multimodal large language models for media forensics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4324–4333, 2024. 3

[29] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, pages 4401–4410, 2019. 1

[30] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19113–19122, 2023. 3

[31] Kwanyoung Kim, Yujin Oh, and Jong Chul Ye. Zegot: Zero-shot segmentation through optimal transport of text prompts. *arXiv preprint arXiv:2301.12171*, 2023. 3, 4

[32] Nicolas Larue, Ngoc-Son Vu, Vitomir Struc, Peter Peer, and Vassilis Christophides. Seeable: Soft discrepancies and bounded contrastive learning for exposing deepfakes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21011–21021, 2023. 6

[33] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 3

[34] Hanzhe Li, Yuezun Li, Jiaran Zhou, Bin Li, and Junyu Dong. Freqblender: Enhancing deepfake detection by blending frequency knowledge. *arXiv preprint arXiv:2404.13872*, 2024. 6

[35] Jiaming Li, Hongtao Xie, Jiahong Li, Zhongyuan Wang, and Yongdong Zhang. Frequency-aware discriminative feature learning supervised by single-center loss for face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6458–6467, 2021. 6

[36] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 1, 3, 6

[37] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 3

[38] KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023. 3

[39] Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. Face x-ray for more general face forgery detection. In *CVPR*, 2020. 1, 3

[40] Yuezun Li and Siwei Lyu. Exposing deepfake videos by detecting face warping artifacts. 1, 3

[41] Yuezun Li, Ming-Ching Chang, and Siwei Lyu. In ictu oculi: Exposing ai created fake videos by detecting eye blinking. In *WIFS*, pages 1–7, 2018. 3

[42] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A new dataset for deepfake forensics. *arXiv preprint arXiv:1909.12962*, 2019. 5

[43] Jiahao Liang, Huafeng Shi, and Weihong Deng. Exploring disentangled content information for face forgery detection. In *European Conference on Computer Vision*, pages 128–145. Springer, 2022. 1, 3

[44] Honggu Liu, Xiaodan Li, Wenbo Zhou, Yuefeng Chen, Yuan He, Hui Xue, Weiming Zhang, and Nenghai Yu. Spatial-phase shallow learning: rethinking face forgery detection in frequency domain. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 772–781, 2021. 1, 3

[45] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 3, 5

[46] Xiaohong Liu, Yaojie Liu, Jun Chen, and Xiaoming Liu. Pscc-net: Progressive spatio-channel correlation network for image manipulation detection and localization. *IEEE Transactions on Circuits and Systems for Video Technology*, 2022. 1

[47] Yuliang Liu, Biao Yang, Qiang Liu, Zhang Li, Zhiyin Ma, Shuo Zhang, and Xiang Bai. Textmonkey: An ocr-free large multimodal model for understanding document. *arXiv preprint arXiv:2403.04473*, 2024. 1, 3

[48] Yuchen Luo, Yong Zhang, Junchi Yan, and Wei Liu. Generalizing face forgery detection with high-frequency features. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16317–16326, 2021. 1, 3

[49] Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. Med-flamingo: a multimodal medical few-shot learner. In *Machine Learning for Health (ML4H)*, pages 353–367. PMLR, 2023. 1, 3

[50] Dat Nguyen, Nesryne Mejri, Inder Pal Singh, Polina Kuleshova, Marcella Astrid, Anis Kacem, Enjie Ghorbel, and Djamila Aouada. Laa-net: Localized artifact attention network for quality-agnostic and generalizable deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17395–17405, 2024. 6

[51] Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24480–24489, 2023. 1, 2, 3, 6, 8

[52] Yongyang Pan, Xiaohong Liu, Siqi Luo, Yi Xin, Xiao Guo, Xiaoming Liu, Xiongkuo Min, and Guangtao Zhai. Towards effective user attribution for latent diffusion models via watermark-informed blending. *arXiv preprint arXiv:2409.10958*, 2024. 3

[53] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002. 5

[54] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023. 3

[55] Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *ECCV*, 2020. 1, 6

[56] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 3

[57] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 1

[58] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *CVPR*, 2019. 5

[59] Lin CY Rouge. A package for automatic evaluation of summaries. In *Proceedings of Workshop on Text Summarization of ACL, Spain*, 2004. 5

[60] Zeyang Sha, Zheng Li, Ning Yu, and Yang Zhang. De-fake: Detection and attribution of fake images generated by text-to-image generation models. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, pages 3418–3432, 2023. 1, 2, 3, 6, 8

[61] Kaede Shiohara and Toshihiko Yamasaki. Detecting deepfakes with self-blended images. In *CVPR*, 2022. 1, 3, 6

[62] Xiufeng Song, Xiao Guo, Jiache Zhang, Qirui Li, Lei Bai, Xiaoming Liu, Guangtao Zhai, and Xiaohong Liu. On learning multi-modal forgery representation for diffusion generated video detection. In *NeurIPS*, 2024. 1, 3

[63] Joel Stehouwer, Hao Dang, Feng Liu, Xiaoming Liu, and Anil Jain. On the detection of digital face manipulation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020. 4

[64] Ke Sun, Shen Chen, Taiping Yao, Hong Liu, Xiaoshuai Sun, Shouhong Ding, and Rongrong Ji. Diffusionfake: Enhancing generalization in deepfake detection via guided stable diffusion. *arXiv preprint arXiv:2410.04372*, 2024. 3

[65] Ke Sun, Shen Chen, Taiping Yao, Xiaoshuai Sun, Shouhong Ding, and Rongrong Ji. Continual face forgery detection via historical distribution preserving. *International Journal of Computer Vision*, pages 1–18, 2024. 3

[66] Ximeng Sun, Ping Hu, and Kate Saenko. Dualcoop: Fast adaptation to multi-label recognition with limited annotations. *Advances in Neural Information Processing Systems*, 35:30569–30582, 2022. 3

[67] Zekun Sun, Yujie Han, Zeyu Hua, Na Ruan, and Weijia Jia. Improving the efficiency and robustness of deepfakes detection through precise geometric features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3609–3618, 2021. 3

[68] Chuangchuang Tan, Ping Liu, RenShuai Tao, Huan Liu, Yao Zhao, Baoyuan Wu, and Yunchao Wei. Data-independent operator: A training-free artifact representation extractor for generalizable deepfake detection. *arXiv preprint arXiv:2403.06803*, 2024. 3

[69] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019. 5, 8

[70] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 3

[71] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. 2017. 2, 3

[72] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015. 5

[73] Chengrui Wang and Weihong Deng. Representative forgery mining for fake face detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14923–14932, 2021. 3, 6

[74] Yuan Wang, Kun Yu, Chen Chen, Xiyuan Hu, and Silong Peng. Dynamic graph learning with content-guided spatial-frequency relation reasoning for deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7278–7287, 2023. 1, 3

[75] Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, and Houqiang Li. Altfreezing for more general video face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4129–4138, 2023. 1

[76] Guangyang Wu, Weijie Wu, Xiaohong Liu, Kele Xu, Tianjiao Wan, and Wenyi Wang. Cheap-fake detection with llm using prompt engineering. In *2023 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, pages 105–109. IEEE, 2023. 3

[77] Yuting Xu, Jian Liang, Gengyun Jia, Ziming Yang, Yanhao Zhang, and Ran He. Tall: Thumbnail layout for deepfake video detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 22658–22668, 2023. 6

[78] Zhiyuan Yan, Yong Zhang, Yanbo Fan, and Baoyuan Wu. Ucf: Uncovering common features for generalizable deepfake detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22412–22423, 2023. 1, 3, 6

[79] Xin Yang, Yuezun Li, and Siwei Lyu. Exposing deep fakes using inconsistent head poses. In *ICASSP*, pages 8261–8265. IEEE, 2019. 1, 3

[80] Yuguang Yao, Xiao Guo, Vishal Asnani, Yifan Gong, Jiancheng Liu, Xue Lin, Xiaoming Liu, and Sijia Liu. Reverse engineering of deceptions on machine- and human-centric attacks. *Foundations and Trends in Privacy and Security*, 2024. 3

[81] Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Yuhao Dan, Chenlin Zhao, Guohai Xu, Chenliang Li, Junfeng Tian, et al. mplug-docowl: Modularized multimodal large language model for document understanding. *arXiv preprint arXiv:2307.02499*, 2023. 1, 3

[82] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*, 2023. 3

[83] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023. 3

[84] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022. 3

[85] Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*, 2023. 1, 3

[86] Yue Zhang, Ben Colman, Xiao Guo, Ali Shahriyari, and Gaurav Bharaj. Common sense reasoning for deepfake detection. In *European Conference on Computer Vision*, pages 399–415. Springer, 2024. 1, 2, 3, 5, 6

[87] Yue Zhang, Zhiyang Xu, Ying Shen, Parisa Kordjamshidi, and Lifu Huang. Spartun3d: Situated spatial understanding of 3d world in large language models. In *International Conference on Learning Representations*, 2024. 3

[88] Zhihao Zhang, Shengcao Cao, and Yu-Xiong Wang. Tamm: Triadapter multi-modal learning for 3d shape understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21413–21423, 2024. 3

[89] Hanqing Zhao, Wenbo Zhou, Dongdong Chen, Tianyi Wei, Weiming Zhang, and Nenghai Yu. Multi-attentional deepfake detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2185–2194, 2021. 1, 4, 6

[90] Tianchen Zhao, Xiang Xu, Mingze Xu, Hui Ding, Yuanjun Xiong, and Wei Xia. Learning self-consistency for deepfake detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021. 1, 3

[91] Yinglin Zheng, Jianmin Bao, Dong Chen, Ming Zeng, and Fang Wen. Exploring temporal coherence for more general video face forgery detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15044–15054, 2021. 1

[92] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16816–16825, 2022. 3, 4, 8

[93] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 3, 4, 8

[94] Qihang Zhou, Guansong Pang, Yu Tian, Shibo He, and Jiming Chen. Anomalyclip: Object-agnostic prompt learning for zero-shot anomaly detection. In *International Conference on Learning Representations*, 2023. 3

[95] Tianfei Zhou, Wenguan Wang, Zhiyuan Liang, and Jianbing Shen. Face forensics in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5778–5788, 2021. 5

[96] Xiangyu Zhu, Hao Wang, Hongyan Fei, Zhen Lei, and Stan Z Li. Face forgery detection by 3d decomposition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2929–2939, 2021. 3

[97] Bojia Zi, Minghao Chang, Jingjing Chen, Xingjun Ma, and Yu-Gang Jiang. Wilddeepfake: A challenging real-world dataset for deepfake detection. In *Proceedings of the 28th ACM international conference on multimedia*, pages 2382–2390, 2020. 5, 6