CorrDetail: Visual Detail Enhanced Self-Correction for Face Forgery Detection

Binjia Zhou[1 †], **Hengrui Lou**[1 †], **Lizhe Chen**[2] and **Haoyuan Li**[3], **Dawei Luo**[4], **Shuai Chen**[4], **Jie Lei**[5], **Zunlei Feng**[1], **Yijun Bei**[1 *]

[1]School of Software Technology, Zhejiang University
[2]Shenzhen International Graduate School, Tsinghua University
[3]School of Computer Science and Technology, Zhejiang University
[4]Ant Group
[5]School of Computer Science and Technology, Zhejiang University of Technology

Abstract

With the swift progression of image generation technology, the widespread emergence of facial deepfakes poses significant challenges to the field of security, thus amplifying the urgent need for effective deepfake detection. Existing techniques for face forgery detection can broadly be categorized into two primary groups: visual-based methods and multimodal approaches. The former often lacks clear explanations for forgery details, while the latter, which merges visual and linguistic modalities, is more prone to the issue of hallucinations. To address these shortcomings, we introduce a visual detail enhanced self-correction framework, designated *CorrDetail*, for interpretable face forgery detection. *CorrDetail* is meticulously designed to rectify authentic forgery details when provided with error-guided questioning, with the aim of fostering the ability to uncover forgery details rather than yielding hallucinated responses. Additionally, to bolster the reliability of its findings, a visual fine-grained detail enhancement module is incorporated, supplying *CorrDetail* with more precise visual forgery details. Ultimately, a fusion decision strategy is devised to further augment the model's discriminative capacity in handling extreme samples, through the integration of visual information compensation and model bias reduction. Experimental results demonstrate that *CorrDetail* not only achieves state-of-the-art performance compared to the latest methodologies but also excels in accurately identifying forged details, all while exhibiting robust generalization capabilities.

1 Introduction

The rapid advancement of generative technologies, exemplified by Artificial Intelligence Generated Content (AIGC) [Cao *et al.*, 2023], has profoundly liberated societal productivity, enabling innovative applications across diverse sectors such as news, literature, entertainment, education, and industry. However, the capability of AIGC to produce high-quality content from simple prompts and guidance has simultaneously lowered the barriers for malicious actors to fabricate deceptive information [Zhao *et al.*, 2023]. This ease of content falsification has introduced significant challenges in verifying identity information, resulting in an increase in financial fraud, misinformation dissemination, and identity theft, thereby posing severe threats to societal security.

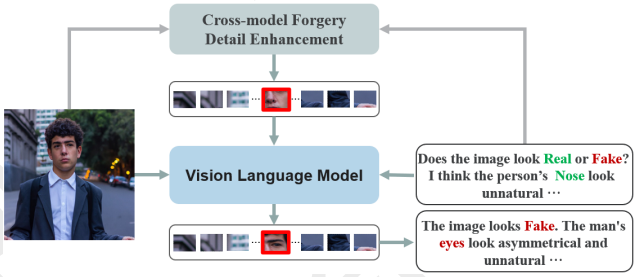

Figure 1: The illustration for *CorrDetail*, which is designed to correct authentic forgery details ("eyes look asymmetrical...") when provided with error-guided questioning ("nose look unnatural...") with the assistance of visual fine-grained detail enhancement.

To address the security threats posed by the facial image forgery technologies, researchers have focused on identifying visual artifacts inherent in forged facial images by analyzing the images themselves. Current approaches [Cao *et al.*, 2022a; Dong *et al.*, 2022; Qian *et al.*, 2020; Wang *et al.*, 2022; Corvi *et al.*, 2023] involve forensic detection from various perspectives, including the network architectures of forgery methods (such as GANs, autoregressive models, and diffusion models) and the domain-specific features of these methods (encompassing both frequency and spatial domains). These techniques have demonstrated commendable performance in facial authentication tasks. However, concentrating solely on forgery detection within a single visual modality encapsulates the forgery process as a black box. Models that rely on a single visual modality typically yield only binary forgery classification results, lacking the provision of critical visual cues necessary for understanding the basis of the discrimination. Consequently, this leads to a significant deficiency in the interpretability of the forgery detection process.

To enhance the interpretability and informational depth of

* Corresponding author. Email: beiyj@zju.edu.cn

facial image forgery detection models, researchers have endeavored to incorporate textual information, thereby developing multimodal deepfake detection frameworks. Leveraging established text-image encoding technologies such as CLIP [Radford *et al.*, 2021] and BLIP [Li *et al.*, 2022], a bridge for feature interaction between the image modality and the text modality has been constructed. Within the field, two primary approaches have emerged: text-image information interaction and the application of large-scale visual language models (VLMs). The first approach involves integrating corresponding textual information with forged facial image data to augment the dimensionality of feature extraction for forgery detection. The second approach builds upon this foundation by replacing traditional neural network models with more extensive and comprehensively pre-trained VLMs. Through a question-and-answer dialogue mechanism, VLMs can provide detailed semantic information during the deepfake detection process, including specifics about the forged regions, methods employed, and the overall forgery framework [Chang *et al.*, 2024; Zhang *et al.*, 2025]. However, within the domain, the training and inference mechanisms of VLMs are susceptible to hallucination phenomena, which not only degrade model performance but also risk supplying erroneous forgery cues. This poses significant challenges to the reliability and accuracy of multimodal deepfake detection.

To address these challenges, we propose a visually enhanced self-correction framework, denoted as *CorrDetail*. By providing error-guided questions, *CorrDetail* establishes a corrective question-and-answer mechanism that cultivates the ability to identify forged details and rectifies training environments prone to hallucinations. In addition, a visual fine-grained detail enhancement module is integrated to furnish *CorrDetail* with more precise visual forgery cues. Finally, we devise a fused decision strategy that combines visual information compensation with model bias reduction, further strengthening the model's ability to discriminate extreme samples. Our specific contributions are as follows:

- We propose a multimodal forgery detail self-correction framework, alongside the creation of a Self-Correction Visual Question Answering (SCVQA) dataset tailored for facial forgery detection. This framework seamlessly combines a self-correction strategy with a cross-model forgery detail enhancement module, aimed at identifying authentic forgery characteristics and producing intelligible, explanatory text-based forgery clues.
- We introduce a novel inference decision mechanism that merges a dual-round question-answering process with multi-scale visual feature interactions, effectively correcting biases inherent in large models. This mechanism entails the construction of multi-expert reasoning decisions, thereby bolstering the reliability and precision of the forgery detection process.
- Extensive comparative, ablation, and visualization experiments demonstrate that our approach not only achieves state-of-the-art performance but also delivers precise and accurate forgery details, showcasing remarkable generalization capabilities.

2 Related Works

2.1 Face Forgery Technology

GAN-based and autoregressive-based models were pioneers in the field of image generation. Driven by these models, realistic facial modification techniques such as BigGAN [Brock *et al.*, 2018] and StarGAN [Choi *et al.*, 2017] emerged, enabling traditional facial forgery methods to be categorized into face swapping, face editing, and face reenactment. With the rapid advancement of AIGC technologies, prompt-guided facial image generation methods have become widely accessible, allowing for the creation of diverse facial images with minimal input, such as textual descriptions or sketches. Notable products like Midjourney [Midjourney, 2022] and Stable Diffusion [Stability.ai, 2023] have surged in popularity by integrating prompts into the AIGC generation process (e.g., diffusion models), thereby facilitating guided and procedural content generation. This prompt-guided AIGC generation approach starkly contrasts with traditional task-oriented methods [Bei *et al.*, 2024], significantly increasing the difficulty of developing generalized facial forgery detection models.

2.2 Vision-based Face Forgery Detection

Vision-based Face Forgery Detection is the predominant approach in the field, concentrating on the forgery artifacts present in manipulated facial images. This methodology can be specifically categorized into frequency-domain-based and spatial-domain-based techniques [Bei *et al.*, 2024]. The core principle involves amplifying forgery traces through transformations in different domains, thereby capturing clearer and more precise artifacts.

In the frequency domain, forgery features are extracted at a fine-grained level by mapping forged images into high-frequency, mid-frequency, and low-frequency dimensions [Qian *et al.*, 2020]. This allows for the detailed identification of subtle manipulation signs that may not be easily discernible in the spatial domain. Conversely, in the spatial domain, the focus is on identifying inconsistencies introduced during the face modification process, such as artifacts, glare, and blurred regions [McCloskey and Albright, 2018; McCloskey and Albright, 2019; Li *et al.*, 2020a]. To address the novel generation methods introduced by AIGC, new approaches have emerged that rely on reconstruction errors for forgery detection in the spatial domain. These methods analyze discrepancies between the original and reconstructed images to identify potential manipulations.

While the aforementioned methods effectively accomplish facial image forgery detection, single-modality approaches fail to provide specific semantic information regarding the forgeries. The absence of disclosed forgery cues results in insufficient interpretability of these methods, limiting their ability to offer comprehensive insights into the detection.

2.3 Multimodal-Based Face Forgery Detection

Recent advancements in image-text pre-trained models [Liu *et al.*, 2023; Radford *et al.*, 2021; Li *et al.*, 2022] have driven researchers to adopt multimodal approaches to enhance face forgery detection. Currently, these approaches can

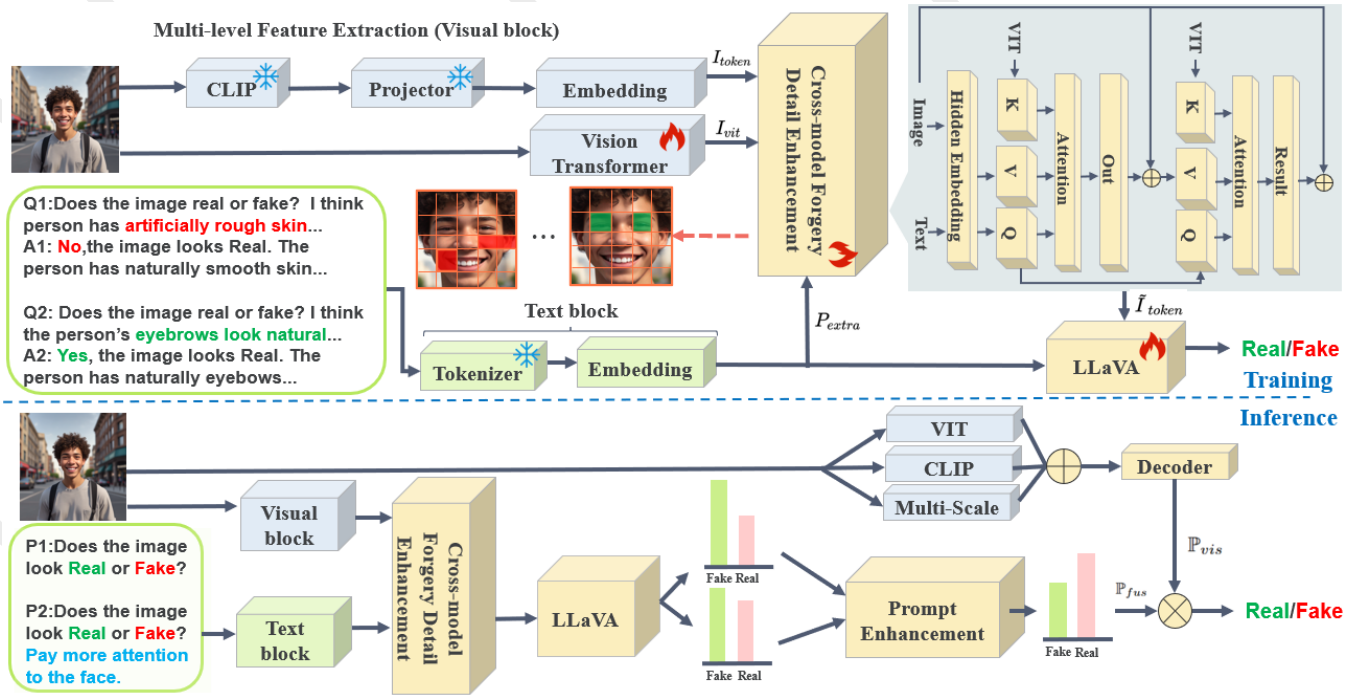

Figure 2: Framework of the proposed *CorrDetail*. During training, the self-correction Q&A strategy is employed to enhance the forgery clue extraction ability of the VLM, while the visual branch integrates a dual-stream architecture (CLIP + ViT) with the Cross-Model Forgery Detail Enhancement (CFDE) module to provide detailed visual features. In the inference phase, the input image and dual-round prompts are initially processed through the CFDE module for feature alignment and concentration. Subsequently, the final detection outcome is generated through a dual-decision strategy.

be broadly categorized into contrastive learning-based methods and VLM-based methods.

The contrastive learning-based methods leverage the comparison of encoded image-text features to align and reinforce discriminative features or utilize textual information to amplify forgery traces within the image modality [Wu *et al.*, 2023; Khan and Dang-Nguyen, 2024]. In contrast, VLM-based methods utilize VLMs for forgery detection, leveraging their powerful performance for tasks like data processing, feature extraction, and decision-making.

While both categories have demonstrated performance improvements, they exhibit significant areas requiring enhancement. Firstly, current methodologies primarily involve simple encoding and fusion or alignment of multimodal features. This coarse-grained discrimination approach impedes the construction of comprehensive and deeply nuanced features. Furthermore, VLMs are often treated as black-box tools or auxiliary aids within the domain, lacking an in-depth development of feature extraction and discrimination decision models grounded in the fundamental principles of VLMs. More critically, this coarse-grained approach induces hallucination phenomena within VLMs, hindering their ability to deliver specific and accurate discrimination results and analyses. As a result, the models are unable to provide detailed and reliable insights into the nature and extent of the forgeries, significantly compromising the overall effectiveness and trustworthiness of the detection system.

2.4 Critique and Refine Mechanism

In the course of in-depth research and application of large language models (LLMs), these models have demonstrated remarkable potential in critique and refine. Researchers have leveraged LLMs to assess the results of specific models, facilitating their self-improvement. In the domain of VLMs, critique and refinement mechanisms have also been introduced to perform logical corrections [Saunders *et al.*, 2022; Madaan *et al.*, 2023; Gou *et al.*, 2024]. Specifically, through querying of results and multi-round interactions, LLMs are endowed with self-improvement mechanisms that allow them to critique and refine their reasoning capabilities [Lin *et al.*, 2024b; Yao, 2024; Li *et al.*, 2024b; Wu *et al.*, 2024]. Following these advancements, the critique and refine Mechanism has garnered significant attention, and in the benchmark domain, there has been a growing trend to use critique mechanisms to assess the critique-correction reasoning ability of LLMs [Lightman *et al.*, 2023; Li *et al.*, 2024a; Luo *et al.*, 2024; Huang *et al.*, 2024].

The aforementioned work on the critique and refine mechanism in text or multimodal LLMs primarily focuses on correcting logical reasoning capabilities. Inspired by this, we propose a novel method that employs error-agnostic question-guided mechanisms, enabling the model to learn how to identify visual forgery features and provide accurate responses. This method does not focus on correcting logical reasoning abilities but rather aims to prevent hallucinated responses, ensuring more reliable and precise detection of visual forgeries.

3 Method

Fig. 2 shows the overall framework of *CorrDetail*. In this section, we present the various modules of *CorrDetail*. In Section 3.1, we introduce an innovative self-correction VQA training strategy designed to augment the forgery clue extraction capability of the VLM. Section 3.2 unveils the cross-model forgery detail enhancement module, which provides intricate visual features for facial forgery detection. In Section 3.3, we introduce a multi-expert ensemble decision-making strategy, which strengthens the reliability and precision of the final outcome.

3.1 Self-Correction Visual Question Answering

In the field of face forgery detection, the VQA training of multimodal large models for related detection tasks is essentially a form of knowledge infusion. Although it can continuously improve model performance through large amounts of data, inspired by [Li *et al.*, 2024b] masking strategy and the fact that images dominate the dual-modal interactions with text in the current task, we aim to enable the model to capture the forgery visual features in images to the greatest extent during training. To this end, we propose a novel training method, as shown in Fig. 1.

Based on the VQA format provided by DD-VQA [Zhang *et al.*, 2025] in the FF++ dataset [Rossler *et al.*, 2019a], we enhance the training process by introducing a new prompt in the question. During training, we introduce incorrect and authentic forgery details into the question. There is a 70% probability of incorporating erroneous details into the original prompt (Q1 in Fig. 2 is an example), a 15% probability of adding genuine forgery details through synonym substitution (Q2 in Fig. 2 is an example), while the remaining questions remain unaltered. We deliberately refrain from entirely inserting incorrect details, as this could prompt the model to exploit the prompt directly, deriving answers opportunistically. This Self-Correction Visual Question Answering (SCVQA) dataset equips the VLM to accurately identify authentic forgery details, rather than producing hallucinated responses during the training process.

3.2 Cross-model Forgery Detail Enhancement

In commonly employed VLM, the intricate visual tokens and succinct text tokens are aligned and mapped into a shared latent space. During this alignment and mapping process, the sophisticated visual tokens inevitably lose some of their detailed visual features. However, face forgery detection must prioritize the capture of these forged visual nuances. To address this, we have designed a Cross-model Forgery Detail Enhancement (CFDE) module, which enriches the VLM with additional visual details, enabling it to learn to discern and identify forged visual features.

Specifically, as depicted in Fig. 2, once the image passes through the CLIP and Projector modules, its embedding is devoid of forgery-specific visual details, such as the intrinsic consistency of surrounding elements and the relational information of relevant positions. Therefore, we introduce an additional vision transformer branch to extract the image's intrinsic visual features, providing supplementary information for face forgery detection.

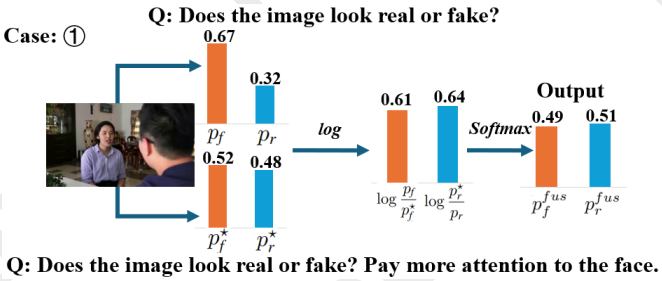

Figure 3: One case for Prompt Enhancement Mechanism handling hard samples (e.g., those with smaller facial proportions).

Formally, the extra text prompt P_{extra} is processed through the same tokenizer and fused with the image tokens I_{token} and the image visual feature I_{vit} (processed through the Vision Transformer) in a cross-modal attention mechanism. The extra text prompt is used as the query Q, the image tokens as the value V, and the image the visual feature as the key K to generate the target attention map. Finally, this attention-enhanced image token $\tilde{I}_{token}$ is obtained by adding the resulting attention map to the original image token in a residual manner. The specific formula is as follows:

$$\tilde{I}_{token} = \mathcal{A}(P_{token}, I_{vit}, \mathcal{A}(P_{token}, I_{vit}, I_{token})),$$
$$P_{token} = Tokenizer(P_{extra}),$$
$$I_{vit} = VIT(I),$$
$$I_{token} = Projector(Clip(I)),$$
$$\mathcal{A}(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V + V,$$

where $CLIP(\cdot)$ represents LLaVA's own visual encoder, and $Projector(\cdot)$ represents the pre-trained image token mapping module, $\mathcal{A}(Q, K, V)$ denotes the self-attention operation in transformer. The final obtained $\tilde{I}_{token}$ will serve as the new image token input to the VLM.

3.3 Decision Fusion

This module is primarily subdivided into two segments. The first segment is the Prompt Enhancement Mechanism, which is chiefly conceived to post-process the VLM, thereby liberating the output from the biases and fairness challenges stemming from the tendentious misjudgment of faces in images with small proportions. The second segment is the Dual-Expert Model Decision, in which an independent visual branch is trained separately to alleviate the information loss induced by image embedding within the VLM. Through collaborative decision-making, the two branches harmonize, resulting in superior performance.

Prompt Enhancement Mechanism. During the testing phase, the same image is provided with a basic question and an additional guiding prompt, such as *"Does the image look real or fake?"* and *"Does the image look real or fake? Pay more attention to the face/background."* Variations in the text input will inevitably lead to changes in the model's binary classification probability output.

Therefore, the Prompt Enhancement Mechanism is designed to address the issue where the detection probability

Methods \ Datasets	Intra-Testing				Cross-Testing		
	FF++ (LQ)		FF++ (HQ)		CelebDF (trained on FF++ (LQ))		
	ACC↑	AUC↑	ACC↑	AUC↑	ACC↑	AUC↑	EER↓
XceptionNet [Rossler *et al.*, 2019a]	86.86	93.50	95.04	96.30	60.11	61.80	41.73
EN-B4 [Tan and Le, 2019]	86.67	88.20	96.63	99.18	66.63	67.22	38.14
Local-relation [Chen *et al.*, 2021]	91.47	95.21	97.59	99.56		–	
RFM [Wang and Deng, 2021]	87.06	89.83	95.69	98.79	64.42	65.64	38.80
RECCE [Cao *et al.*, 2022b]	91.03	95.02	97.06	99.32	67.96	68.71	35.73
CD-Net [Song *et al.*, 2022]	88.12	95.20	98.75	99.90		–	
UIA-VIT [Zhuang *et al.*, 2022]	91.05	94.88	98.62	99.33	<u>69.80</u>	<u>70.15</u>	35.82
GS [Guo *et al.*, 2023b]	92.76	<u>96.85</u>	<u>99.24</u>	**99.95**		–	
HiFi-Net [Guo *et al.*, 2023a]	89.25	92.10	97.41	98.56	67.20	68.80	36.13
PFG-DD [Lin *et al.*, 2024a]	91.03	93.47	97.33	98.68	68.51	69.68	35.94
RECCE (DD-VQA) [Zhang *et al.*, 2025]	<u>92.08</u>	95.36	98.65	99.79	69.46	70.21	<u>35.63</u>
CorrDetail (Ours)	**94.41**	**96.93**	**99.28**	<u>99.92</u>	**72.32**	**72.80**	**33.87**

Table 1: Comparison of *CorrDetail* with SOTA methods across both intra- and cross-datasets. The highest and second-highest performances are highlighted in **best** and <u>second best</u>, respectively. ACC and AUC (↑) denote that higher values are preferred, while EER (↓) signifies that lower values are desired. All scores are expressed in %.

diminishes when the model places excessive focus on the face or background with the added guiding prompt, suggesting that the initial detection probability is dubious. Fig. 3 illustrates the scenario in which the model's predicted probability for "fake" decreases as more attention is directed towards the face, indicating that the initial "fake" prediction is unreliable. A possible explanation is that the model's initial prediction relies on the background features for challenging samples with smaller facial proportions.

For the initial prediction probability $[p_f, p_r]$ and second-round prediction $[p_f^\star, p_r^\star]$ with additional guiding prompt, the final fusion prediction $\mathbb{P}_{fus}$ is adjusted as follows:

$$\mathbb{P}_{fus} = \begin{cases} softmax([\log \frac{p_f}{p_f^\star}, \log \frac{p_r^\star}{p_r}]), \ p_f > p_r \ \& \ p_f > p_f^\star \\ softmax([\log \frac{p_f^\star}{p_f}, \log \frac{p_r}{p_r^\star}]), \ p_r > p_f \ \& \ p_r > p_r^\star \end{cases}$$

when the initial "fake" prediction p_f satisfy the condition $(1-\lambda) \geq p_f \geq \lambda$, where λ is set as a hyperparameter, serving as the exclusion threshold. For $p_f > (1-\lambda)$, $p_f < \lambda$ or any other cases, final fusion prediction $\mathbb{P}_{fus}$ will simply match the initial prediction $[p_f, p_r]$.

Dual-Expert Model Decision. As depicted in Fig. 2, an additional purely visual branch is trained independently using cross-entropy loss, enabling it to perform forgery detection autonomously. Its output is subsequently integrated with the VLM branch's output, enhancing the probability for discriminative consistency. The specific formula is as follows:

$$\mathbb{P}_{final} = softmax([p_f^{fus} \times p_f^{vis}, p_r^{fus} \times p_r^{vis}]),$$
$$\mathbb{P}_{fus} = [p_f^{fus}, p_r^{fus}], \mathbb{P}_{vis} = [p_f^{vis}, p_r^{vis}],$$

where $\mathbb{P}_{fus}$ represents the output of the VLM branch, $\mathbb{P}_{vis}$ represents the output of the traditional visual branch, and finally $\mathbb{P}_{final}$ is the final discriminative output.

Figure 4: Qualitative examples of results on certain samples. For failure cases of GPT-4o mini, *CorrDetail* achieves correct classification results along with corresponding explanations.

4 Experiments

In this section, we introduce the experimental setup and environment, present a performance comparison between *CorrDetail* and recent approaches across various datasets, and highlight the advantages of our approach. Finally, we conduct ablation studies to further validate the effectiveness of each module.

Architecture and Implementation Details. The vision branch in 3.3 consists of three components: CLIP, VIT, and Multi-scale. The CLIP module corresponds to the corresponding part in the VLM's architecture, VIT primarily uses vit-L/16, and Multi-scale extracts features at different scales by progressively decomposing through multi-level sampling. More details on the structure of each module can be found in the supplementary materials. We fine-tuned LLaVA-1.5-7B on the SCVQA dataset using a learning rate of $5e^{-5}$, training for 3 epochs with a batch size of 64 on three Nvidia A100 GPUs. The training process employed cosine annealing learning rate decay and the AdamW optimizer. The threshold λ is set as 0.1. More details are given in the *supplementary materials*.

SCVQA Dataset. Based on the original data from the

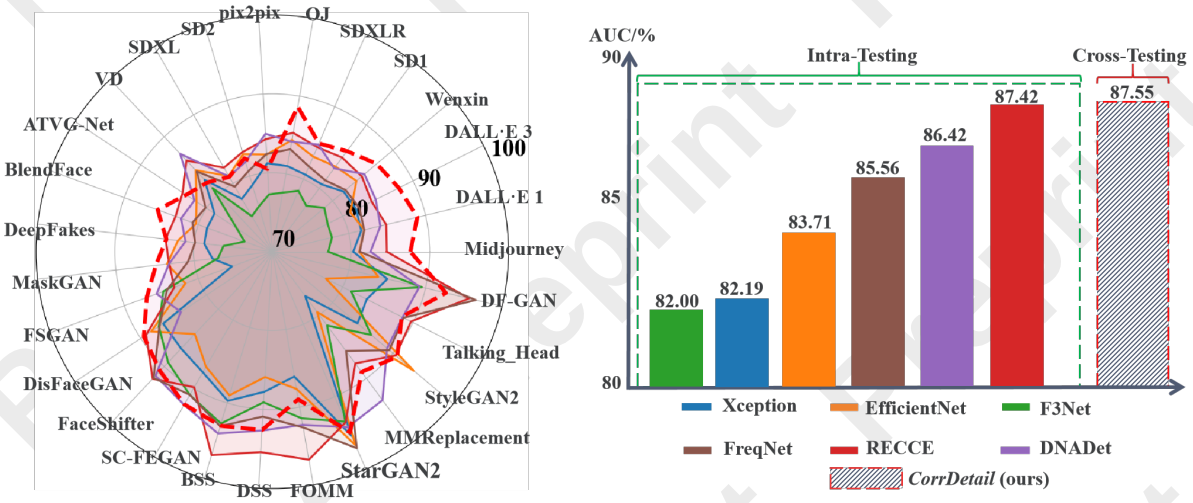

Figure 5: The generalization performance of *CorrDetail* in comparison to the intra-performance of existing methods. The left radar chart illustrates the detection efficacy across various types of forgery samples, with solid lines representing the comparison techniques and a dashed line denoting *CorrDetail*. The bar chart on the right displays the average AUC for each method. The six comparison methods are trained and evaluated on the DeepFaceGen dataset. In contrast, *CorrDetail* is trained on SCVQA and tested on DeepFaceGen.

Methods \ Datasets	DFR		WDF	
	AUC↑	ERR↓	AUC↑	ERR↓
EN-B4 [Tan and Le, 2019]	92.18	15.51	67.89	37.21
Xception [Rossler *et al.*, 2019a]	91.93	15.52	68.90	38.67
VIT-B [Dosovitskiy, 2020]	80.47	26.97	75.29	33.40
LTW [Sun *et al.*, 2021]	–		67.12	39.22
GFF [Luo *et al.*, 2021]	–		66.51	41.52
DCL [Sun *et al.*, 2022]	92.26	14.91	72.95	35.73
RECCE [Cao *et al.*, 2022b]	92.93	14.74	74.38	32.64
CFM [Luo *et al.*, 2023]	95.18	11.87	78.39	30.79
PFG-DD [Lin *et al.*, 2024a]	94.74	12.59	77.41	29.50
MOE-FFD [Kong *et al.*, 2024]	95.39	11.29	80.64	27.11
CorrDetail (Ours)	**96.77**	**10.68**	**81.58**	**26.88**

Table 2: The generalization performance comparison of *CorrDetail* and other methods trained on FF++ (HQ). Missing data is represented by "–".

FF++ dataset, we have constructed the SCVQA dataset, which consists of 19,797 [image, question, answer] triples to support the self-correction mechanism. The dataset includes three types of data formats: (1) 70% of the data is in the form of [image, question, incorrect detail description as the answer], (2) 15% is in the form of [image, question, null answer], and (3) 15% is in the form of [image, question, correct detail description as the answer].

Evaluation Dataset and Metrics. We evaluate the proposed method using the FF++ [Rossler *et al.*, 2019b], CelebDF [Li *et al.*, 2020b], WildDeepfake (WDF) [Zi *et al.*, 2020], DeepForensics-1.0 (DFR) [Jiang *et al.*, 2020], and DeepFaceGen [Bei *et al.*, 2024] datasets. The WDF contains 7,314 facial action sequences extracted from 707 DeepFake videos, and the DFR represents the largest face forgery detection dataset by far, with 60,000 videos constituted by a total of 17.6 million frames. CelebDF employs face swapping as the primary forgery technique, encompassing a total of 5,639 videos. FF++ includes four forgery methods—Deepfake, Face2Face, FaceSwap, and NeuralTextures—comprising a total of 4,000 videos. The videos from both FF++ and CelebDF are cropped into images for facial forgery detection, establishing them as widely adopted evaluation datasets. DeepFaceGen is a facial forgery dataset that includes both video and image modalities. From this dataset, we selected facial forgery images using both traditional task-oriented and novel prompt-guided forgery methods, encompassing a total of 27 approaches. Consistent with the evaluation metrics established in the benchmark papers, we adopt Accuracy (ACC), Equal Error Rate (EER), and Area Under the Curve (AUC) as metrics.

4.1 Comparison with SOTAs

In this section, We compare *CorrDetail* with several mainstream methods, including XceptionNet, HiFi-Net, RECCE, CD-Net, Local-relation, RFM, RECCE (DD-VQA), EN-B4, GS, and UIA-VIT. We fine-tune *CorrDetail* on SCVQA constructed with FF++. As presented in Table 2, *CorrDetail* achieves over 94% in both ACC and AUC in the intra-testing evaluation on FF++, maintains over 72% in ACC and AUC in cross-testing on CelebDF, and reduces the EER to below 34%, attaining the best performance in six out of seven evaluation metrics across intra-testing and cross-testing. The remaining metric also attains the second-best performance. Notably, *CorrDetail* demonstrates a significant advantage over existing SOTA models in the field, specifically RECCE (DD-VQA), which is based on VLM, and GS, which is based on feature decoupling. Fig. 4 further illustrates *CorrDetail*'s ability to accurately detect instances of facial forgery and provide precise forgery cues.

Module			Intra-Testing		Cross-Testing
SCVQA	CFDE	Decision Fusion	LQ	HQ	CelebDF
✗	✗	✗	88.44	91.87	65.58
✗	✗	✓	90.59	93.48	67.15
✗	✓	✗	92.94	95.01	68.99
✓	✗	✗	93.88	96.05	69.37
✗	✓	✓	94.03	97.82	69.88
✓	✗	✓	94.51	98.90	70.44
✓	✓	✗	95.38	99.37	71.54
✓	✓	✓	**96.93**	**99.92**	**72.32**

Table 3: Ablation study of *CorrDetail* with FF++ datasets (LQ and HQ). Best results are highlighted in **bold**. All scores are in %.

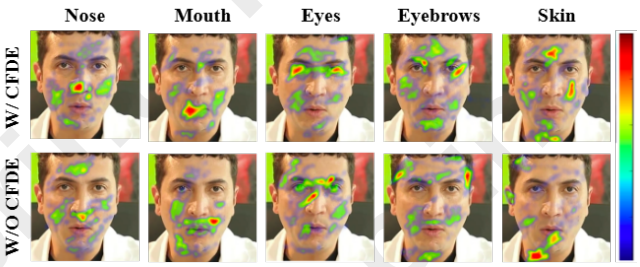

Figure 6: The heatmap visualization of ablation study on CFDE.

4.2 Generalization on Different Datasets

To further validate the effectiveness of *CorrDetail*, we conducted generalization performance experiments. Specifically, on CelebDF, we trained *CorrDetail* on the SCVQA and performed cross-dataset evaluation by testing it on the CelebDF. We compared the performance of *CorrDetail* against other compared methods that were also trained on FF++. On DeepFaceGen, we evaluated *CorrDetail* against six SOTA detection models. It is important to note that the six detection models were subjected to intra-testing, meaning they were both trained and tested on DeepFaceGen. In contrast, *CorrDetail* was trained on SCVQA and tested on DeepFaceGen, representing a cross-testing scenario.

The experimental results for the CelebDF are illustrated in Table 2. In the cross-testing experiments presented in the right column, our method achieved the best performance, surpassing all comparison methods. For the more challenging DeepFaceGen benchmark, the results depicted in Fig. 5 demonstrate that, even under the less favorable experimental setup where *CorrDetail* was trained on SCVQA and tested on DeepFaceGen while the comparison algorithms were trained and tested on DeepFaceGen, *CorrDetail* still outperformed the SOTA model RECCE.

In these unseen data scenarios, encompassing both traditional forged facial data and novel AIGC-generated forged data, *CorrDetail* consistently delivered outstanding performance. These results conclusively demonstrate the robustness and effectiveness of *CorrDetail* in diverse and challenging forgery detection environments.

4.3 Ablation Study

In this section, we discuss practical effectiveness of each module in *CorrDetail*, including the enhancement of learning in SCVQA compared to VQA, the effectiveness of CFDE in accurately enhancing target semantics within images, and, finally, whether the decision enhancement module can effectively address certain challenging samples that VLMs struggle to handle. Detailed results are listed in Table 3.

Illustrating the Effectiveness of SCVQA. To demonstrate the performance of SCVQA, we compared models trained using SCVQA with the GPT-4o mini and visualized the outputs on several face-forged images, as shown in Fig. 4. It is worth noting that during this ablation process, no additional innovative modules were incorporated—only prompt guidance was added. Clearly, for some images that 4o mini fails to classify correctly, our model accurately identifies the reasonable or unrealistic regions and details.

Discussing the influence of CFDE. We compared the traditional prompt-based attention enhancement method with CFDE through heatmap visualizations, as shown in Fig. 6. It is evident that feature-level processing yields a more focused overall effect compared to traditional methods. For instance, at the *skin* level, our method produces a broader heat distribution compared to the prompt-based method, while at the *nose* level, it exhibits more concentrated heat around the corresponding parts of the image. This comprehensive comparison demonstrates that enhancing from a cross-modal semantic perspective is more effective in guiding the model's recognition than relying solely on text-based approaches.

Discussing the influence of decision fusion. In this section, we focus on addressing native negative sample issues that can be resolved through integrated decision-making, as shown in Fig. 3. For certain images where the background occupies a large proportion of the frame or where the face is slightly blurred, the VLM's attention to the facial region becomes weaker, leading to incorrect classifications. By visualizing the outputs, we further demonstrate the effectiveness of post-processing. Meanwhile, the enhancements from the visual module are more reflected in the overall improvement of classification performance, as shown in Table 3.

5 Conclusion

In this paper, we present *CorrDetail*, a novel face forgery detection framework that enhances both performance and interpretability through its key modules. The self-correction mechanism utilizes error-guided questions to improve the VLM's efficiency and generalization, achieving SOTA results in face forgery detection tasks. Our visual fine-grained detail enhancement module provides precise forgery cues by strengthening specific semantic features. Additionally, the fused decision strategy employs post-processing to address focus issues the model cannot resolve independently. *CorrDetail* outperforms recent methods across multiple benchmarks, demonstrating superior accuracy and generalization.

Acknowledgments

This work is funded by National Key Research and Development Project (Grant No: 2022YFB2703100), Ant Group

through CCF-Ant Research Fund, and Information Technology Center, ZheJiang University.

Contribution Statement

Authors with † have made the same contribution to this work.

References

[Bei *et al.*, 2024] Yijun Bei, Hengrui Lou, Jinsong Geng, Erteng Liu, Lechao Cheng, Jie Song, Mingli Song, and Zunlei Feng. A large-scale universal evaluation benchmark for face forgery detection, 2024.

[Brock *et al.*, 2018] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *ArXiv*, abs/1809.11096, 2018.

[Cao *et al.*, 2022a] Junyi Cao, Chao Ma, Taiping Yao, Shen Chen, Shouhong Ding, and Xiaokang Yang. End-to-end reconstruction-classification learning for face forgery detection. In *CVPR*, pages 4103–4112, 2022.

[Cao *et al.*, 2022b] Junyi Cao, Chao Ma, Taiping Yao, Shen Chen, Shouhong Ding, and Xiaokang Yang. End-to-end reconstruction-classification learning for face forgery detection. In *CVPR*, pages 4113–4122, 2022.

[Cao *et al.*, 2023] Yihan Cao, Siyu Li, Yixin Liu, Zhiling Yan, Yutong Dai, Philip S. Yu, and Lichao Sun. A comprehensive survey of ai-generated content (aigc): A history of generative ai from gan to chatgpt. *ArXiv*, abs/2303.04226, 2023.

[Chang *et al.*, 2024] You-Ming Chang, Chen Yeh, Wei-Chen Chiu, and Ning Yu. Antifakeprompt: Prompt-tuned vision-language models are fake image detectors, 2024.

[Chen *et al.*, 2021] Shen Chen, Taiping Yao, Yang Chen, Shouhong Ding, Jilin Li, and Rongrong Ji. Local relation learning for face forgery detection. In *AAAI*, volume 35, pages 1081–1088, 2021.

[Choi *et al.*, 2017] Yunjey Choi, Min-Je Choi, Mun Su Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. *CVPR*, pages 8789–8797, 2017.

[Corvi *et al.*, 2023] Riccardo Corvi, Davide Cozzolino, Giada Zingarini, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. On the detection of synthetic images generated by diffusion models. In *ICASSP*, pages 1–5, 2023.

[Dong *et al.*, 2022] S. Dong, Jin Wang, Jiajun Liang, Haoqiang Fan, and Renhe Ji. Explaining deepfake detection by analysing image matching. In *ECCV*, 2022.

[Dosovitskiy, 2020] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv*, 2020.

[Gou *et al.*, 2024] Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Nan Duan, and Weizhu Chen. Critic: Large language models can self-correct with tool-interactive critiquing, 2024.

[Guo *et al.*, 2023a] Xiao Guo, Xiaohong Liu, Zhiyuan Ren, Steven Grosz, Iacopo Masi, and Xiaoming Liu. Hierarchical fine-grained image forgery detection and localization. In *CVPR*, pages 3155–3165, 2023.

[Guo *et al.*, 2023b] Ying Guo, Cheng Zhen, and Pengfei Yan. Controllable guide-space for generalizable face forgery detection. In *ICCV*, pages 20818–20827, 2023.

[Huang *et al.*, 2024] Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. Large language models cannot self-correct reasoning yet. In *ICLR*, 2024.

[Jiang *et al.*, 2020] Liming Jiang, Ren Li, Wayne Wu, Chen Qian, and Chen Change Loy. Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection. In *CVPR*, pages 2889–2898, 2020.

[Khan and Dang-Nguyen, 2024] Sohail Ahmed Khan and Duc-Tien Dang-Nguyen. Clipping the deception: Adapting vision-language models for universal deepfake detection. In *ICMR*, pages 1006–1015, 2024.

[Kong *et al.*, 2024] Chenqi Kong, Anwei Luo, Peijun Bao, Yi Yu, Haoliang Li, Zengwei Zheng, Shiqi Wang, and Alex C Kot. Moe-ffd: Mixture of experts for generalized and parameter-efficient face forgery detection. *arXiv*, 2024.

[Li *et al.*, 2020a] 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 *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5000–5009, 2020.

[Li *et al.*, 2020b] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A large-scale challenging dataset for deepfake forensics. In *CVPR*, June 2020.

[Li *et al.*, 2022] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022.

[Li *et al.*, 2024a] Junlong Li, Shichao Sun, Weizhe Yuan, Run-Ze Fan, hai zhao, and Pengfei Liu. Generative judge for evaluating alignment. In *ICLR*, 2024.

[Li *et al.*, 2024b] Yian Li, Wentao Tian, Yang Jiao, Jingjing Chen, Na Zhao, and Yu-Gang Jiang. Look before you decide: Prompting active deduction of mllms for assumptive reasoning, 2024.

[Lightman *et al.*, 2023] Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step, 2023.

[Lin *et al.*, 2024a] Li Lin, Xinan He, Yan Ju, Xin Wang, Feng Ding, and Shu Hu. Preserving fairness generalization in deepfake detection. In *CVPR*, pages 16815–16825, 2024.

[Lin *et al.*, 2024b] Zicheng Lin, Zhibin Gou, Tian Liang, Ruilin Luo, Haowei Liu, and Yujiu Yang. CriticBench: Benchmarking LLMs for critique-correct reasoning. In

Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *ACL*, pages 1552–1587, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[Liu *et al.*, 2023] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023.

[Luo *et al.*, 2021] Yuchen Luo, Yong Zhang, Junchi Yan, and Wei Liu. Generalizing face forgery detection with high-frequency features. In *CVPR*, pages 16317–16326, 2021.

[Luo *et al.*, 2023] Anwei Luo, Chenqi Kong, Jiwu Huang, Yongjian Hu, Xiangui Kang, and Alex C Kot. Beyond the prior forgery knowledge: Mining critical clues for general face forgery detection. *IEEE Transactions on Information Forensics and Security*, 19:1168–1182, 2023.

[Luo *et al.*, 2024] Liangchen Luo, Zi Lin, Yinxiao Liu, Lei Shu, Yun Zhu, Jingbo Shang, and Lei Meng. Critique ability of large language models, 2024.

[Madaan *et al.*, 2023] Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback, 2023.

[McCloskey and Albright, 2018] Scott McCloskey and Michael Albright. Detecting gan-generated imagery using color cues. *ArXiv*, abs/1812.08247, 2018.

[McCloskey and Albright, 2019] Scott McCloskey and Michael Albright. Detecting gan-generated imagery using saturation cues. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 4584–4588, 2019.

[Midjourney, 2022] Midjourney. Midjourney. https://www.midjourney.com/home, 2022.

[Qian *et al.*, 2020] Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *ECCV*, page 86–103, Berlin, Heidelberg, 2020. Springer-Verlag.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

[Rossler *et al.*, 2019a] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *ICCV*, pages 1–11, 2019.

[Rossler *et al.*, 2019b] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Niessner. Faceforensics++: Learning to detect manipulated facial images. In *ICCV*, October 2019.

[Saunders *et al.*, 2022] William Saunders, Catherine Yeh, Jeff Wu, Steven Bills, Long Ouyang, Jonathan Ward, and Jan Leike. Self-critiquing models for assisting human evaluators, 2022.

[Song *et al.*, 2022] Luchuan Song, Zheng Fang, Xiaodan Li, Xiaoyi Dong, Zhenchao Jin, Yuefeng Chen, and Siwei Lyu. Adaptive face forgery detection in cross domain. In *ECCV*, pages 467–484. Springer, 2022.

[Stability.ai, 2023] Stability.ai. Stable Diffusion. https://stability.ai/, 2023.

[Sun *et al.*, 2021] Ke Sun, Hong Liu, Qixiang Ye, Yue Gao, Jianzhuang Liu, Ling Shao, and Rongrong Ji. Domain general face forgery detection by learning to weight. In *AAAI*, volume 35, pages 2638–2646, 2021.

[Sun *et al.*, 2022] Ke Sun, Taiping Yao, Shen Chen, Shouhong Ding, Jilin Li, and Rongrong Ji. Dual contrastive learning for general face forgery detection. In *AAAI*, volume 36, pages 2316–2324, 2022.

[Tan and Le, 2019] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *ICML*, pages 6105–6114. PMLR, 2019.

[Wang and Deng, 2021] Chengrui Wang and Weihong Deng. Representative forgery mining for fake face detection. In *CVPR*, pages 14923–14932, 2021.

[Wang *et al.*, 2022] Junke Wang, Zuxuan Wu, Wenhao Ouyang, Xintong Han, Jingjing Chen, Yu-Gang Jiang, and Ser-Nam Li. M2tr: Multi-modal multi-scale transformers for deepfake detection. In *ICMR*, ICMR '22, page 615–623, New York, NY, USA, 2022. Association for Computing Machinery.

[Wu *et al.*, 2023] H. Wu, J. Zhou, and S. Zhang. Generalizable synthetic image detection via language-guided contrastive learning. *arXiv*, 2023.

[Wu *et al.*, 2024] Xueqing Wu, Yuheng Ding, Bingxuan Li, Pan Lu, Da Yin, Kai-Wei Chang, and Nanyun Peng. Visco: Benchmarking fine-grained critique and correction towards self-improvement in visual reasoning, 2024.

[Yao, 2024] Liang Yao. Large language models are contrastive reasoners, 2024.

[Zhang *et al.*, 2025] Yue Zhang, Ben Colman, Xiao Guo, Ali Shahriyari, and Gaurav Bharaj. Common sense reasoning for deepfake detection. In *ECCV*, pages 399–415. Springer, 2025.

[Zhao *et al.*, 2023] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. A survey of large language models, 2023.

[Zhuang *et al.*, 2022] Wanyi Zhuang, Qi Chu, Zhentao Tan, Qiankun Liu, Haojie Yuan, Changtao Miao, Zixiang Luo, and Nenghai Yu. Uia-vit: Unsupervised inconsistency-aware method based on vision transformer for face forgery detection. In *ECCV*, pages 391–407. Springer, 2022.

[Zi *et al.*, 2020] Bojia Zi, Minghao Chang, Jingjing Chen, Xingjun Ma, and Yu-Gang Jiang. Wilddeepfake: A challenging real-world dataset for deepfake detection. In *ACM MM*, pages 2382–2390, 2020.