

# ForenX: Towards Explainable AI-Generated Image Detection with Multimodal Large Language Models

Chuangchuang Tan<sup>1,2\*</sup> Jinglu Wang<sup>2</sup> Xiang Ming<sup>2</sup> Renshuai Tao<sup>1</sup> Yunchao Wei<sup>1</sup>  
Yao Zhao<sup>1</sup> Yan Lu<sup>2</sup>

<sup>1</sup>Beijing Jiaotong University <sup>2</sup>Microsoft Research Asia

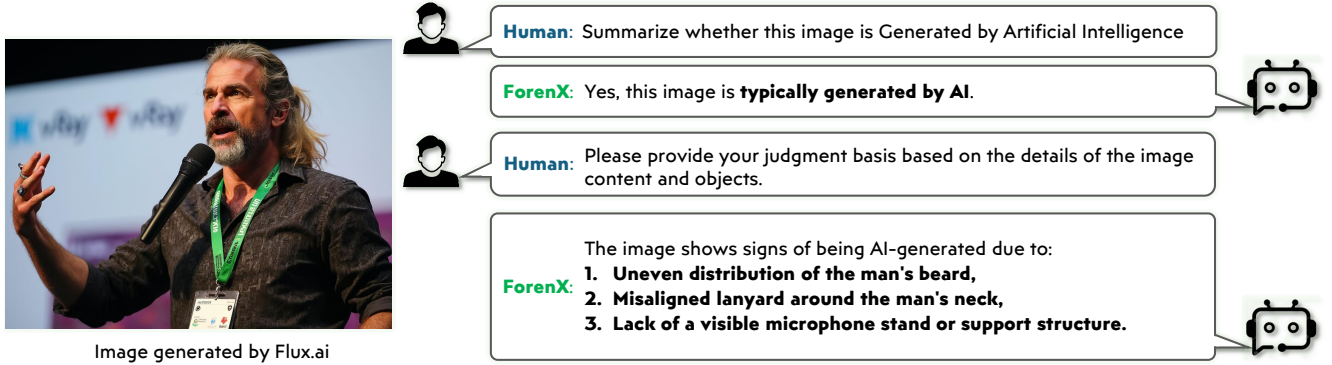


Figure 1. We present a sample result of our explainable AI-generated image detection method, ForenX. Given an image created by Flux.ai [3], ForenX not only delivers the detection result (“Yes”), but also offers a forensic analysis that resonates with human reasoning. For instance, the third evidence, highlighting the lack of a supporting structure for the microphone, is closely aligned with human intuition, which standard MLLMs fail to provide (demonstrated in our experiments).

## Abstract

Advances in generative models have led to AI-generated images visually indistinguishable from authentic ones. Despite numerous studies on detecting AI-generated images with classifiers, a gap persists between such methods and human cognitive forensic analysis. We present ForenX, a novel method that not only identifies the authenticity of images but also provides explanations that resonate with human thoughts. ForenX employs the powerful multimodal large language models (MLLMs) to analyze and interpret forensic cues. Furthermore, we overcome the limitations of standard MLLMs in detecting forgeries by incorporating a specialized **forensic prompt** that directs the MLLMs’ attention to forgery-indicative attributes. This approach not only enhance the generalization of forgery detection but also empowers the MLLMs to provide explanations that are accurate, relevant, and comprehensive. Additionally, we introduce **ForgReason**, a dataset dedicated to descriptions of forgery evidences in AI-generated images. Curated through collaboration between an LLM-based agent and a team of human annotators, this process provides refined data that further enhances our model’s performance. We demon-

strate that even limited manual annotations significantly improve explanation quality. We evaluate the effectiveness of ForenX on two major benchmarks. The model’s explainability is verified by comprehensive subjective evaluations.

## 1. Introduction

Recent advancements in generative models, such as GANs and diffusion models [14, 17, 23, 24, 43], have led to the creation of AI-generated images that are nearly indistinguishable from real-world photographs, raising concerns about potential misuse. As a result, developing effective methods for detecting AI-generated images has become a critical research focus.

Recent studies in AI-generated image detection [7, 12, 30, 31, 50, 51] have predominantly focused on detection-based approaches (Fig. 2 (a)), aiming to differentiate AI-generated and authentic images by detecting specific artifacts or patterns introduced by generative models. These methods rely on artifact-based features to capture subtle traces of forgery. However, they are prone to the classic overfitting issue, restricting their capacity to generalize

\*Work done during C. Tan’s internship at MSRA.

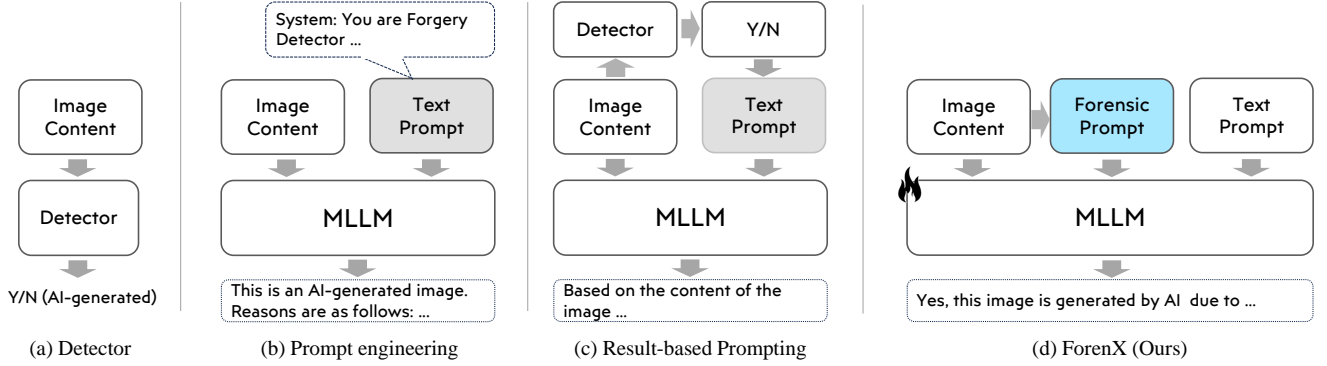


Figure 2. Empower MLLMs to detect AI-generated Images with reliable explanations. Detection model (a) only outputs binary classification without human-understood explanations. Prompt engineering (b), even integrated with detection result (c) may generate explanations without focusing on forensic evidences. Our method (d) enhances the MLLM by incorporating the specialized forensic prompt.

across new generative models.

Moreover, detection models, based on binary classification, typically identify indicative patterns in a black-box manner. Such patterns, often imperceptible or incomprehensible to humans, resulting in difficulties for humans to verify their detection results. In contrast, human observers spot AI-generated images by examining not only artifacts but also the content, such as unnatural human poses, inconsistent lighting conditions, or violations of physical laws. There is a significant discrepancy in the forensic analysis by detection models and humans. To bridge this gap, we introduce an approach that goes beyond binary classification. Our method enriches the forensic analysis of detection results, delivering explanations that highlight both content-related anomalies and generation-based artifacts, thereby aligning the model closer to human perceptual and cognitive strategies. We term our method as the **Forensics eXplainer (ForenX)**. Additionally, model-generated explanations can improve user understanding of the detection results and facilitate a “double-check” process, enabling human verification of the model’s predictions.

Multimodal Large Language Models (MLLMs), such as LLaVA [29] and GPT-4 Vision [4], demonstrate compelling abilities in various cross-modal tasks, including understanding and reasoning. These abilities render them highly advantageous for achieving explainable AI-generated image detection. Adopting prompt engineering in the straightforward manner to guide MLLM outputs (Fig. 2 (b)) tends to generate hallucinations, which may not provide meaningful explanations. Even when prompts are integrated with pre-determined classification results (Fig. 2 (c)), the approach proves unsatisfactory. Likewise, enhancing MLLMs by simply fine-tuning them with textual annotations does not suffice for reliable explanation. All the effectiveness is attributed to the inherent lack of considerations for image

authenticity<sup>1</sup> within the standard MLLMs. To overcome this limitation, we introduce the **forensic prompt**, a specialized input designed to direct MLLMs towards prioritizing authenticity and its related cues, which are beyond conventional text and image inputs. This strategy, integrated with LoRA [18] fine-tuning, enhances MLLMs’ ability to achieve both robust detection and reliable explanation generation.

The forensic prompt is considered to fulfill three essential requirements for effectiveness. (1) Initially, it should enable multi-modal representation with alignment of text and image features. We derive inspiration from the detector [37], which showcases the utility of pretrained CLIP-ViT [11, 42] features. We construct our forensic prompt based on the CLIP-ViT features. However, we observe that the pretrained features, designed for general purpose, may not inherently possess the specific authenticity. (2) The prompt should encapsulate effective cues indicative of authenticity. We integrate the features with authenticity attributes by tuning them specifically for detection with an additional detection loss. (3) Finally, the prompt should be adaptable to MLLMs. To ensure this, we fine-tune these enhanced features along with the forensic prompt during the MLLM training phase. This process is aimed at intensifying MLLMs’ focus on authenticity cues, making them more adept at identifying authentic content. As demonstrated in our experiment, our model with fine-tuned CLIP-ViT features achieves significantly better performance compared to the untuned ones.

Since no existing dataset provides annotations detailing why an image appears AI-generated, we initially train our model using captions generated by the method in Fig. 2(b), as it is impractical to obtain large-scale annotations from human annotators for AI-generated images. However, these

<sup>1</sup>Authenticity in this paper refers to the likelihood an image is captured in the real world rather than generated by AI.

machine-generated captions lack alignment with human reasoning. To address this, we collected human-provided annotations specifying the attributes that make an image seem AI-generated. We then fine-tune the model, initially trained on machine generated low-quality captions, using this human-annotated data. This approach allows the model to attain human-like reasoning capabilities with a minimal set of human annotations.

In conclusion, we present ForenX, the first approach to extend MLLMs for Explainable AI-generated Image Detection. Our contributions are three-fold:

- We propose a simple yet effective pipeline, ForenX, to enable MLLMs for explainable AI-generated image detection. ForenX demonstrates strong recognition capabilities, generalization performance, and explainability.
- We introduce the concept of *forensic prompt* to guide MLLMs, enhancing their ability to detect AI-generated images accurately.
- We develop an explainable AI-generated image detection dataset, *ForeReason*, which pairs images with authenticity-related descriptions annotated by human annotators and GPT-4 Vision, facilitating alignment with human reasoning.

## 2. Related Works

**AI-Generated image detection.** To improve the generalizability of models in AI-generated image detection, current research primarily focuses on artifact extraction and detector design within this domain. Various low-level artifact features have been introduced to capture generation traces, including frequency-based attributes [20], gradient information [46], neighboring pixel relationships [50], and random-mapping features [47]. For example, BiHPF [20] enhances artifact magnitudes using dual high-pass filters, while LGrad [46] utilizes gradient data from pre-trained models as representations of artifacts. NPR [50] presents a simple yet effective approach by reconsidering up-sampling operations for artifact representation. In addition to low-level features, large pre-trained models have been leveraged to capture high-level forgery traces in AI-generated content detection tasks. UniFD [37] directly incorporates image features from the CLIP model for linear classification, showcasing robust deepfake detection even with unseen sources. FatFormer [30] combines frequency analysis with a text encoder as an adapter to the frozen CLIP vision model, thereby enhancing detection performance significantly. After investigating the mechanism of CLIP in deepfake detection, C2P-CLIP[48] introduces the category common prompt to enhance detection accuracy.

**Multimodal large language model.** In recent years, the advent of Multimodal Large Language Models (MLLMs)

such as GPT-4V [4] and LLaVA [28] has garnered significant attention due to their unparalleled proficiency in image comprehension and analysis. Several studies [22, 44, 54] have employed prompt engineering to explore the applicability of MLLMs on new tasks, such as face forgery analysis and reasoning. Moreover, research efforts [9, 25, 26] have been directed towards instruction tuning to augment the performance of large language models on specialized tasks. Recent advancements in deepfake detection have introduced methodologies leveraging MLLM. SNIFFER [40] employs InstructBLIP [9] for the enhancement of misinformation detection capabilities, showcasing its significant contributions to the domain of information security. FFAA [19] integrates a fine-tuned Multimodal LLM with a Multi-answer Intelligent Decision System for Open-World Face Forgery Analysis VQA. ForgeryGPT [27] utilizes precise forgery mask data to refine LLM capabilities for explainable image forgery detection. Jia *et al.* [22] meticulously crafted prompts to assess MLLM’s effectiveness in face forgery detection. Despite initial explorations into leveraging MLLM for detecting forgery image, existing studies have mainly focused on identifying facial manipulations and exhibit a gap in AIGC forgery detection.

## 3. Method

In this work, we aim to develop a Multimodal Large Language Model (MLLM) capable of detecting AI-generated images, with a focus on three key attributes: recognition ability, generalizability, and explainability. An overview of the ForenX architecture is shown in Fig. 3. Our primary approach involves incorporating a forensic prompt into the MLLM, such as LLaVA[29]. In the remainder of this section, we describe the technical aspects of ForenX, emphasizing the construction of the forensic prompt and the two-stage training strategy. We also outline the creation of the *ForeReason* dataset.

### 3.1. The Architecture of ForenX

As discussed in Sec. 1, we introduce an forensic prompt to guide the MLLM’s focus on assessing the authenticity of the image. Leveraging the strong performance of CLIP-ViT [11] for this task, we use features extracted from CLIP-ViT to construct the forensic prompt. To maintain feature consistency and reduce training instability, we also use CLIP-ViT features as the representation of image content, ensuring aligned feature distributions across inputs.

Formally, the construction of the forensic prompts is as follows:

$$\begin{aligned} F_v &= g(X_v), \\ F_v^f &= l_f(F_v), \\ H_v^f &= m_f(F_v^f), \end{aligned} \tag{1}$$

in this setup,  $X_v$  denotes the input image,  $g$  represents the

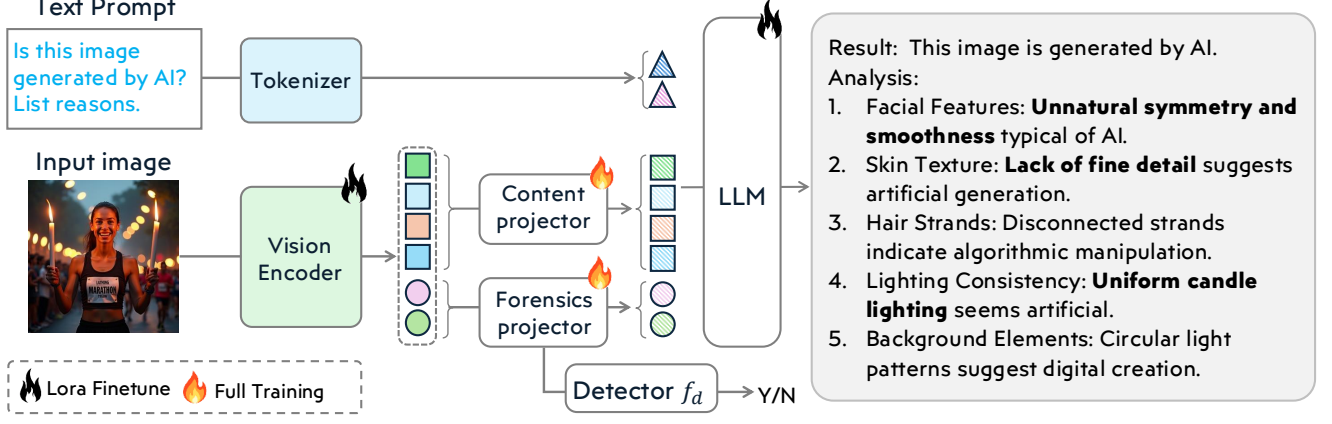


Figure 3. Overview of our ForenX: A Simple Yet Effective Explainable AI-Generated Image Detection Pipeline. To overcome the limitations of standard MLLMs in detecting forgeries, we incorporate a specialized forensic prompt that directs the MLLMs’ attention to forgery-indicative attributes.

feature extractor and  $F_v$  refers to the extracted features. The  $l_f$  and  $m_f$  constitute the forensic projector designed to extract forensic prompt. We apply forensics encoder  $l_f$  to project these features into the forensics embedding space, resulting in  $F_v^f$ , which also serves as the input for an auxiliary forgery detection loss, providing additional constraints on the forensic embedding space. Here  $H_v^f$  serves as the forensic prompt in our model and  $m_f$  is the mapping function from the forensics embedding space to the word embedding space.

We use a trainable forensics embedding  $d$  as the forensics encoder  $l_f$  in our experiments. Specifically, the forensics embedding  $d$  transforms the visual feature  $F_v$  into forensics features  $F_v^f$  as follows:

$$F_v^f = F_v \otimes d. \quad (2)$$

In our experiments,  $\otimes$  denotes the Hadamard product. To further integrate forensics information into  $F_v^f$ , we impose constraints using forgery detection labels  $y_d \in \{0, 1\}$ :

$$\mathcal{L}_{detection} = Loss(f_d(F_v^f), y_d), \quad (3)$$

in which  $f_d$  is a function that maps feature  $F_v^f$  to the classification space. In our experiment, we employ a simple summation function  $sum(\cdot)$  as  $f_d$ . We then connect  $F_v^f$  into the word embedding space using  $m_f$ . This mapping is implemented using two MLP layers.

To achieve comprehensive content understanding, AI-generated image detection and forensics analysis, we feed the text tokens  $H_t$  (from text prompts), visual content tokens  $H_v^c$  (same as Llava), and forensic prompt  $H_v^f$  into the Large Language Model (LLM), as follows:

$$T_p = llm(H_t, H_v^c, H_v^f). \quad (4)$$

We utilize conversations with image captions, detection results, and forensics descriptions as labels for instruction

fine-tuning of LLM:

$$\mathcal{L}_{instruction} = Loss(T_p, T_l). \quad (5)$$

in which  $T_l$  is the conversational data for instruction tuning. Following the foundation MLLM, we adopt instruction-tuning using an auto-regressive objective, enabling model to predict subsequent outputs based on input instructions.

In our pipeline, it is essential that the LLM can independently identify AI-generated images to generate convincing forensic reasoning. To achieve this, we avoid directly converting detection results into prompts (e.g., “this image is real” or “this image is AI-generated”). Instead, we introduce a forensic prompt to guide the LLM’s attention to both the semantic content and authenticity elements within the image, prompting the model to generate more insightful forensic analyses.

### 3.2. Dataset Construction

To our knowledge, no existing dataset directly supports training MLLMs for AI-generated image detection, as we require image captions focused on authenticity. However, since modern AI-generated images closely mimic real images, creating large-scale labeled data through human annotation alone is impractical. A small sample size would also risk quick overfitting. To address this, we first use LLMs to generate a large set of preliminary captions focused on image authenticity. We then manually filter for cases that are difficult for state-of-the-art binary classifiers but readily identifiable by human judgment. The data from LLMs serves as a pretraining set, while the human-labeled examples are used for fine-tuning, enhancing the model’s alignment with human reasoning on image authenticity.

#### 3.2.1. Data Generation with LLM

For fair comparison with other model in the paper, we only use the Llava [29] to generate the image cap-



tions for two widely used forgery detection datasets, *Genimage* and *ForenSynths*. For each image  $X_v$  in the datasets, it is annotated with two-round conversational data  $\{(X_q^1, X_a^1), (X_q^2, X_a^2)\}$ . This includes two types of question-and-answer pairs: content-related  $(X_q^1, X_a^1)$ , and detection results  $(X_q^2, X_a^2)$ .

Specifically, responses related to image content are generated by the pre-trained LLaVa model, using questions that are randomly chosen from a predefined list within LLaVa<sup>2</sup>. For questions about detection results, answers are directly derived from image labels. For example, if an image is labeled as AI-generated, the corresponding question and answer pair might be:

- **Question:** Summarize whether this image is generated by AI, please return beginning with **Yes** or **No**.
- **Answer:** Yes, this image is typically generated by AI.

The test samples will be marked as AI-generated/authentic if the answer starts with **Yes**, and authentic if the answer starts with **Yes/No** in the evaluation process.

### 3.2.2. Data Annotation with Human

To further enhance the forensics interpretability of the LLM, we involve human annotators to label 2,215 images with explanations on why each image is identified as AI-generated. The detailed annotation process is as follows:

1. **Select Images:** Obtain images from Midjourney and choose those that have a realistic style.
2. **Annotate with boxes:** Manually annotate areas in the images that appear unreasonable using box annotations. Provide a description explaining why these areas are deemed unreasonable.
3. **Summarize with GPT-4 Vision:** Use GPT-4 Vision to summarize the images and the manual annotations, generating final explanations for why the images are considered AI-generated.

These images were annotated with captions and detection results following the methods described in Sec. 3.2.1. The detailed flowchart of Human-annotation is provided in the supplementary materials. Before using GPT-4 Vision to generate summaries, we convert bounding box coordinates into textual descriptions that convey relative spatial positions. In this process, some detailed annotations are omitted to adhere to OpenAI’s restrictions.

### 3.3. Optimization Strategy

Our training process follows a two-stage optimization strategy to address the imbalance between annotations obtained from the LLM and those from human annotators. Additionally, we introduce real-world images as negative samples in

the second stage, as the initial 2,215 human-annotated images consist solely of AI-generated samples, which could lead to overfitting. We further sample 5000 real and 1000 fake from Genimage dataset, resulting in a total of 8,215 images. Following is the optimization details for each stage:

#### Stage 1: Initial training for general recognition and explainability.

We begin by using the pre-trained LLaVa model as the foundational MLLM, fine-tuning it on the ForenSynths and GenImage datasets, which provide content captions and detection labels. During this stage, both the CLIP-ViT encoder and the LLM are fine-tuned using LoRA. The overall loss function applied in this stage is:

$$\mathcal{L} = \mathcal{L}_{\text{detection}} + \mathcal{L}_{\text{instruction}} \quad (6)$$

#### Stage 2: Fine-tuning with the dedicated **ForgReason**.

In this stage, we utilize the 8,215 images to further enhance the explainability of the LLM. During this phase, only the LLM is fine-tuned using LoRA with the instruction-based loss  $\mathcal{L}_{\text{instruction}}$ , while all other layers, including the CLIP-ViT, remain frozen. This fine-tuning process focuses on improving the model’s ability to generate detailed and accurate forensics explanations, guiding the model to better interpret and explain the authenticity of AI-generated images.

## 4. Experiments

**Dataset.** We conduct our experiment on two major benchmarks, *Genimage* and *ForenSynths*, for detection accuracy evaluation and one constructed dataset, *ForgReason*, for explainability fine-tuning.

- **Genimage.** This dataset predominantly utilizes various diffusion models for image generation, including Midjourney [1], SDv1.4 [43], SDv1.5 [43], ADM [10], GLIDE [36], Wukong [2], VQDM [15], and BigGAN [5].
- **ForenSynths.** Following Wang et al. [53], our dataset employs ProGAN as part of its training configuration. We utilize a four-class categorization scheme consisting of horse, chair, cat, and car classes as outlined by Tan et al. [50] and Liu et al. [30]. The test set encompasses eight subsets derived from various generative models: ProGAN [23], StyleGAN [24], BigGAN [5], CycleGAN [56], StarGAN [8], GauGAN [38], and Deepfake<sup>3</sup>.
- **ForgReason.** This dataset comprises 2215 realistic fake images from Midjourney, alongside 5000 real and 1000 fake sampled from Genimage dataset. Each image from Midjourney includes manually annotated fake descriptions. All these images collectively for training.

**Implementation detail.** Our implementation uses LLaVa 8B as the foundational Multimodal Large Language Model

<sup>2</sup>Some samples can be found in the supplementary materials.

<sup>3</sup>Deepfake

| Methods          | Venues      | Abilities |         | Test Models |        |        |      |       |        |      |        |      |
|------------------|-------------|-----------|---------|-------------|--------|--------|------|-------|--------|------|--------|------|
|                  |             | Detection | Explain | Midjourney  | SDv1.4 | SDv1.5 | ADM  | GLIDE | Wukong | VQDM | BigGAN | mAcc |
| ResNet-50[16]    | CVPR2016    | ✓         | ×       | 54.9        | 99.9   | 99.7   | 53.5 | 61.9  | 98.2   | 56.6 | 52.0   | 72.1 |
| DeiT-S[52]       | ICML2021    | ✓         | ×       | 55.6        | 99.9   | 99.8   | 49.8 | 58.1  | 98.9   | 56.9 | 53.5   | 71.6 |
| Swin-T[32]       | ICCV2021    | ✓         | ×       | 62.1        | 99.9   | 99.8   | 49.8 | 67.6  | 99.1   | 62.3 | 57.6   | 74.8 |
| CNNSpot[53]      | CVPR2020    | ✓         | ×       | 52.8        | 96.3   | 95.9   | 50.1 | 39.8  | 78.6   | 53.4 | 46.8   | 64.2 |
| Spec[55]         | WIFS2019    | ✓         | ×       | 52.0        | 99.4   | 99.2   | 49.7 | 49.8  | 94.8   | 55.6 | 49.8   | 68.8 |
| F3Net[41]        | ECCV2020    | ✓         | ×       | 50.1        | 99.9   | 99.9   | 49.9 | 50.0  | 99.9   | 49.9 | 49.9   | 68.7 |
| GramNet[33]      | CVPR2020    | ✓         | ×       | 54.2        | 99.2   | 99.1   | 50.3 | 54.6  | 98.9   | 50.8 | 51.7   | 69.9 |
| UnivFD[37]       | CVPR2023    | ✓         | ×       | 93.9        | 96.4   | 96.2   | 71.9 | 85.4  | 94.3   | 81.6 | 90.5   | 88.8 |
| NPR [50]         | CVPR2024    | ✓         | ×       | 81.0        | 98.2   | 97.9   | 76.9 | 89.8  | 96.9   | 84.1 | 84.2   | 88.6 |
| FreqNet [49]     | AAAI2024    | ✓         | ×       | 89.6        | 98.8   | 98.6   | 66.8 | 86.5  | 97.3   | 75.8 | 81.4   | 86.8 |
| FatFormer[30]    | CVPR2024    | ✓         | ×       | 92.7        | 100.0  | 99.9   | 75.9 | 88.0  | 99.9   | 98.8 | 55.8   | 88.9 |
| DRCT[7]          | ICML2024    | ✓         | ×       | 91.5        | 95.0   | 94.4   | 79.4 | 89.2  | 94.7   | 90.0 | 81.7   | 89.5 |
| CLIP(336px)-Lora | -           | ✓         | ×       | 94.2        | 99.0   | 99.1   | 50.7 | 94.2  | 98.9   | 88.2 | 61.2   | 85.7 |
| LLAVA            | NeurIPS2023 | ✓         | ✓       | 51.6        | 50.5   | 50.6   | 50.2 | 51.7  | 52.8   | 50.6 | 54.0   | 51.5 |
| LLAVA-PE         | NeurIPS2023 | ✓         | ✓       | 50.6        | 50.4   | 50.4   | 50.1 | 50.9  | 52.1   | 51.1 | 52.9   | 51.1 |
| LLAVA-FT         | NeurIPS2023 | ✓         | ✓       | 90.8        | 95.2   | 95.1   | 64.5 | 97.5  | 93.9   | 95.6 | 95.0   | 91.0 |
| <b>ForenX-S1</b> | ours        | ✓         | ✓       | 97.9        | 97.8   | 97.7   | 97.4 | 98.0  | 98.0   | 97.7 | 97.8   | 97.8 |
| <b>ForenX-S2</b> | ours        | ✓         | ✓       | 97.9        | 97.8   | 97.7   | 96.9 | 98.0  | 98.0   | 97.8 | 97.0   | 97.6 |

Table 1. **Cross-Diffusion-Sources Evaluation on the GenImage Dataset.** The model is trained using SDv1.4 as described in [57]. LLAVA-PE refers to LLAVA-Prompt-Engineering, while LLAVA-FT denotes LLAVA-FineTuning. We highlight the highest and the second highest numbers in green and blue respectively.

| Methods       | Venues      | Abilities |         | Test Models |          |           |        |          |         |        |          |      |
|---------------|-------------|-----------|---------|-------------|----------|-----------|--------|----------|---------|--------|----------|------|
|               |             | Detection | Explain | ProGAN      | StyleGAN | StyleGAN2 | BigGAN | CycleGAN | StarGAN | GauGAN | Deepfake | mAcc |
| CNNDet[53]    | CVPR2020    | ✓         | ×       | 91.4        | 63.8     | 76.4      | 52.9   | 72.7     | 63.8    | 63.9   | 51.7     | 67.1 |
| Frank[13]     | PMLR2020    | ✓         | ×       | 90.3        | 74.5     | 73.1      | 88.7   | 75.5     | 99.5    | 69.2   | 60.7     | 78.9 |
| Durall[12]    | CVPR2020    | ✓         | ×       | 81.1        | 54.4     | 66.8      | 60.1   | 69.0     | 98.1    | 61.9   | 50.2     | 67.7 |
| Patchfor[6]   | ECCV2020    | ✓         | ×       | 97.8        | 82.6     | 83.6      | 64.7   | 74.5     | 100.    | 57.2   | 85.0     | 80.7 |
| F3Net[41]     | ECCV2020    | ✓         | ×       | 99.4        | 92.6     | 88.0      | 65.3   | 76.4     | 100.    | 58.1   | 63.5     | 80.4 |
| SelfBland[45] | CVPR2022    | ✓         | ×       | 58.8        | 50.1     | 48.6      | 51.1   | 59.2     | 74.5    | 59.2   | 93.8     | 61.9 |
| GANDet[34]    | ICIP2022    | ✓         | ×       | 82.7        | 74.4     | 69.9      | 76.3   | 85.2     | 68.8    | 61.4   | 60.0     | 72.3 |
| BiHPF[20]     | WACV2022    | ✓         | ×       | 90.7        | 76.9     | 76.2      | 84.9   | 81.9     | 94.4    | 69.5   | 54.4     | 78.6 |
| FrePGAN[21]   | AAAI2022    | ✓         | ×       | 99.0        | 80.7     | 84.1      | 69.2   | 71.1     | 99.9    | 60.3   | 70.9     | 79.4 |
| LGrad [51]    | CVPR2023    | ✓         | ×       | 99.9        | 94.8     | 96.0      | 82.9   | 85.3     | 99.6    | 72.4   | 58.0     | 86.1 |
| UniFD [37]    | CVPR2023    | ✓         | ×       | 99.7        | 89.0     | 83.9      | 90.5   | 87.9     | 91.4    | 89.9   | 80.2     | 89.1 |
| FreqNet[49]   | AAAI2024    | ✓         | ×       | 99.6        | 90.2     | 88.0      | 90.5   | 95.8     | 85.7    | 93.4   | 88.9     | 91.5 |
| NPR[50]       | CVPR2024    | ✓         | ×       | 99.8        | 96.3     | 97.3      | 87.5   | 95.0     | 99.7    | 86.6   | 77.4     | 92.5 |
| FatFormer[30] | CVPR2024    | ✓         | ×       | 99.9        | 97.2     | 98.8      | 99.5   | 99.3     | 99.8    | 99.4   | 93.2     | 98.4 |
| LLAVA         | NeurIPS2023 | ✓         | ✓       | 61.9        | 50.2     | 49.9      | 54.8   | 53.6     | 52.8    | 61.3   | 50.5     | 54.4 |
| LLAVA-PE      | NeurIPS2023 | ✓         | ✓       | 63.5        | 50.1     | 50.0      | 54.9   | 59.4     | 53.2    | 59.5   | 51.6     | 55.3 |
| LLAVA-FT      | NeurIPS2023 | ✓         | ✓       | 96.7        | 61.6     | 53.9      | 84.8   | 89.6     | 85.0    | 95.3   | 65.1     | 79.0 |
| ForenX        | ours        | ✓         | ✓       | 99.9        | 94.8     | 89.8      | 98.6   | 96.6     | 93.2    | 99.1   | 83.2     | 94.4 |

Table 2. **Cross-GAN-Sources Evaluation on ForenSynths dataset.** We adopt ProGAN as the training source following [30, 50].

(MLLM), incorporating Llama3 as the large language model component. We optimize the model using an initial learning rate set to  $2 \times 10^{-5}$ . The batch size is configured to 128, and the model is trained for 1 epoch. We apply LoRA on the  $q\_proj$ ,  $k\_proj$ , and  $v\_proj$  layers using the Parameter-Efficient Fine-Tuning (PEFT) [35] library. The hyperparameters for the LoRA layers are set as follows:  $lora.r = 8$ ,  $lora.alpha = 16$ , and  $lora.dropout = 0.9$ . The proposed method is implemented in PyTorch [39] and runs on 4 Nvidia A100 GPUs. Consistent with baseline studies [30, 37, 50], we use mean accuracy (mAcc) as the evaluation metric to assess model performance.

#### 4.1. Recognition and Generalization Evaluation

We evaluate the recognition and generalization capabilities of ForenX using the GenImage and ForenSynths datasets. In the domain of deepfake detection, we believe that the

recognition and generalization abilities of the LLM are crucial. The model must have strong detection capabilities and be able to generalize to unknown sources for its forgery explanations to be considered reliable and acceptable.

**GenImage evaluation.** To assess the recognition and generalization capabilities of our ForenX model, we use Stable Diffusion 1.4 as the training set and evaluate our method on images generated by eight different models: Midjourney, SDv1.4, SDv1.5, ADM, GLIDE, Wukong, VQDM, and BigGAN. The detection results are presented in Tab. 1, showing that ForenX achieves an impressive accuracy of 97.8%. During Stage 2 training, employing ForgReason to enhance the model’s interpretability results in a slight 0.2% impact on its recognition performance. Our model outperforms both non-MLLM and MLLM baselines. Our ForenX is composed of CLIP, LLM, and two projectors, with CLIP primarily used for image feature extraction.

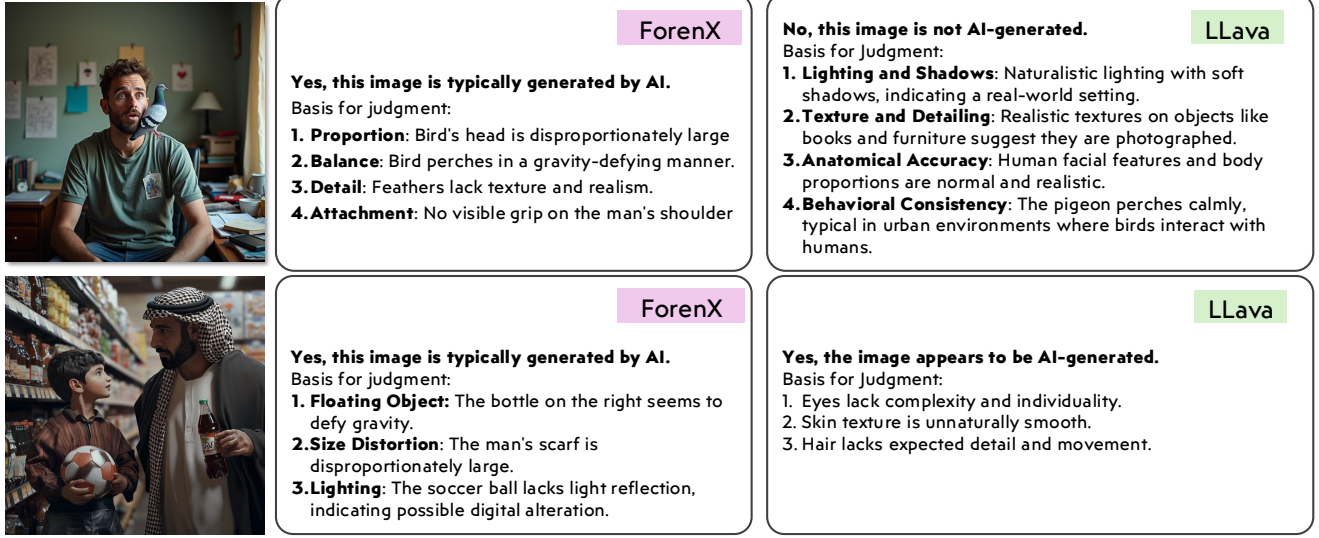


Figure 4. Examples of Explainable AI-Generated Image Detection using LLaVa and our ForenX. ForenX identifies anomalies in the location and status of objects, such as the pigeon and Coca-Cola, within the images. Due to space limitations, the full description is provided in the supplementary materials.

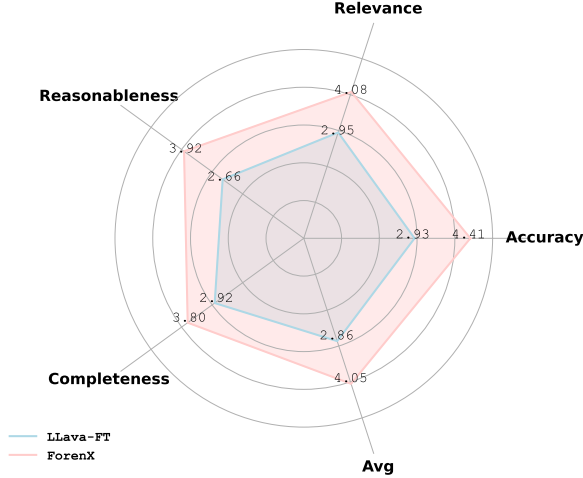


Figure 5. User study for Comparing Explanation Performance Between LLaVa and Our ForenX. We set a full score of 5 points.

Compared to CLIP-based methods such as FatFormer, UniFD, and CLIP336-lora, ForenX demonstrates a performance improvement of 8-9%. This indicates that incorporating an LLM enhances the detection performance of CLIP in AI-generated image detection tasks. Furthermore, DRCT, which uses a pretrained Stable Diffusion model to generate rebuilt images and then performs detection on paired images, achieves lower performance than ForenX. Our method surpasses DRCT by 7.8%, while also offering the ability to explain the forensics reasons.

In comparison to LLM-based baselines, our approach also shows significant improvements. To evaluate the vanilla LLaVa model's deepfake detection capabilities, we used prompts similar to those in ForenX. Additionally, by

altering prompts to implement prompt engineering, named LLaVa-PE, we assess LLaVa's detection ability. Both vanilla LLaVa and LLaVa-PE achieved accuracies of only 51.5% and 51.1%, respectively. This is due to LLaVa's design focus on image content understanding, lacking specific insights needed for AI-generated image detection detection. To address dataset bias, we fine-tuned LLaVa using the same dataset with LoRA. This fine-tuning improved detection performance to 79.7%, still significantly less than ForenX's 97.8% accuracy.

Unlike LLaVa, which exclusively extracts image content for LLM input, our approach extracts both content information and forensics-related details for LLM input, significantly enhancing detection capabilities. ForenX-S1 achieves an accuracy of 97.8%, notably outperforming LLaVa, LLaVa-PE, and LLaVa-finetune by 45.8%, 46.2%, and 17.6%, respectively. This demonstrates that incorporating forensics-relevant information into MLLM can substantially improve detection performance.

**ForenSynths evaluation.** We further validate the performance of the proposed method on the ForenSynths dataset, as shown in Tab. 2. Following the baselines, we adopt four classes setting to train the models, which consist of images of horse, chair, cat, car generated by ProGAN. Then, evaluation on eight generative models containing ProGAN, StyleGAN, BigGAN, CycleGAN, StarGAN, GauGAN and Deepfake. Compared to the Non-LLM methods, our proposed ForenX approach achieved competitive results with an accuracy of 94.4%. Our method achieves an improvement of 1.9% compared to the NPR-based approach and a 5.3% enhancement over the CLIP-only UniFD method. Our approach achieves a competitive accuracy of 94.4%, as opposed to the 98.4% reported by Fatformer. It is important to

note that our method possesses both recognition and interpretation capabilities, whereas Fatformer can only detect. We further compare against LLM-based methods, including LLava, LLava-PE, and LLava-FT. It can be observed that our approach outperforms these three methods.

Additionally, to evaluate the effectiveness of our proposed method on state-of-the-art generative models, we conducted experiments using SDV3 and Flux. We sample 6,000 images from Stable-Diffusion-3 and Flux.1 [dev], respectively. An equal number of real images are sourced from the Gen-Image dataset. Our ForenX achieves an accuracy of **97.7%** on SDv3 and **97.8%** on Flux, respectively.

## 4.2. Explainability Analysis

In this section, we assess the explainability performance of ForenX through qualitative analysis and user-study.

**Quantitative Evaluation** To evaluate the interpretability of the proposed model, we present an assessment framework utilizing LLMs. Specifically, both the generated forgery explanations and their respective manually annotated references are input into the LLM for evaluation across five metrics: comprehensiveness, relevance, similarity, reasonableness, and average performance. The annotation method introduced in Sec. 3.2.2 is employed to annotate images from Midjourney and Flux. ChatGPT-4o is used to compute those metric, with results in Tab. 3 showing that our proposed method, ForenX, significantly outperforms LLAVA-FT. To mitigate the impact of variability introduced by ChatGPT-4’s stochastic nature, we compute the average value across three iterations and use it as the final result.

Table 3. Quantitative results of interpretability (transposed).

| Metric            | LLava-FT | ForenX      |
|-------------------|----------|-------------|
| Comprehensiveness | 80.7     | <b>81.2</b> |
| Relevance         | 71.4     | <b>75.0</b> |
| Similarity        | 60.9     | <b>70.5</b> |
| Reasonableness    | 74.1     | <b>77.2</b> |
| Avg               | 71.8     | <b>76.0</b> |

### 4.2.1. Qualitative analysis

Fig. 4 provides the examples of detection explanations given by LLava and ForenX. The first image depicts a pigeon standing on a man’s shoulder in an indoor scene. As shown in the figure, LLava’s judgment is incorrect, as it relies on superficial reasoning based on general aspects of the image, such as quality, lighting, and contextual clues, without delivering specific insights related to AI-generated image detection. In contrast, ForenX accurately identifies the image’s authenticity and offers detailed, contextually relevant explanations. For instance, ForenX highlights the unusual presence of a pigeon perched on the man’s shoulder,

providing a specific and insightful observation pertinent to the image content in a forensic context.

In addition, in the second image, ForenX also detects anomalies in the cola. Although LLava recognizes the image as fake, it did not notice the anomalies with the cola bottle and soccer ball. All explanations from ForenX are directly tied to the image’s content, showcasing its superior capability to interpret and elucidate the reasoning behind detection decisions. This qualitative analysis underscores the enhanced explainability of ForenX compared to LLava.

### 4.2.2. User-Study

To quantify the explanatory capability of the proposed method, we conducted a user study. Specifically, we select 100 realistic-style images from Midjourney and Flux, then rated the reasons for image forgery generated by ForenX and LLava-FT. We chose 20 users to evaluate five aspects: accuracy, relevance, reasonableness, completeness, and an overall assessment of all aspects. During this process, we randomly shuffle the order of the two reasons so that users could not know which method generated them. The results are shown in Fig. 5. We can see that our ForenX outperforms LLava-FT in all four aspects.

To quantify the explanatory capability of the proposed method, we conduct a user study. Specifically, we select 100 realistic-style images from Midjourney and Flux, then rate the reasons for image forgery generated by ForenX and LLava-FT. We chose 20 users to evaluate five aspects: accuracy, relevance, reasonableness, completeness, and overall evaluation of all aspects. We set a full score of 5 points. During this process, we randomly shuffled the order of the two reasons so that users could not know which method was used to generate them. The results are shown in Fig. 5. It can be seen that our ForenX outperforms LLava-FT in five aspects simultaneously. Our method outperforms LLava-FT on five metrics, indicating that the proposed forensic prompt enhances MLLM’s recognition ability, and the proposed ForgReason dataset significantly improves MLLM’s interpretative ability in low-sample scenarios.

| w/o Lora<br>on CLIP | w/o LLM  | w/o Forensics<br>Projector | w/o $L_{detection}$ | mAcc. |
|---------------------|----------|----------------------------|---------------------|-------|
| <b>X</b>            | <b>X</b> | <b>X</b>                   | <b>X</b>            | 84.7  |
| ✓                   | <b>X</b> | <b>X</b>                   | <b>X</b>            | 85.7  |
| ✓                   | ✓        | <b>X</b>                   | <b>X</b>            | 91.0  |
| ✓                   | ✓        | ✓                          | <b>X</b>            | 93.7  |
| ✓                   | ✓        | ✓                          | ✓                   | 97.8  |

Table 4. Ablation Study regarding the effectiveness of ForenX’s components on Genimage dataset. The results demonstrate an incremental benefit with the inclusion of each module.

## 4.3. Ablation Study

In this section, we analyze the impact of various components of our ForenX on its performance using the GenImage



dataset. Therefore, we focus on the impact of the following factors: 1) the effect of  $\mathcal{L}_{Detection}$ ; 2) the effect of forensics projector; 3) the effect of LLM for CLIP, besides providing interpretability; 4) the effect of Lora on CLIP. When all those module is disable, it is a fixed CLIP(336px) and trainable MLP implementation for detection. The results of are shown in Tab. 4. Compared to non-LLM methods, the addition of LLM effectively improves detection performance. Additionally,  $\mathcal{L}_{Detection}$  and the forensics projector introduce an forensic prompt into the LLM, further enhancing detection performance. The study reveals that each module contributes to overall effectiveness, with improvements observed as additional components are integrated.

## 5. Conclusion

In this paper, we explore methods to enhance Multimodal Large Language Models (MLLMs) with capabilities for detecting AI-generated images, introducing ForenX, a simple yet effective framework for explainable detection. Unlike MLLMs designed solely for content understanding, such as Llava, which focus on extracting image content for LLM input, our approach integrates both content information and a forensic prompt, significantly improving detection performance. Additionally, we believe that further exploration of diverse forensic prompts, such as frequency features, gradients, and NPR[50], presents a promising research direction. We hope this work inspires continued advancements in leveraging MLLMs for explainable AI-generated image detection.

## 6. Ethics statement

The AI-generated images employed in this study are exclusively utilized for training a detection model designed to determine whether an image has been generated by artificial intelligence. These datasets will not be repurposed beyond the scope explicitly defined herein. To prevent any potential misuse or unethical applications, strict adherence to transparency principles, respect for digital content creators and consumers, as well as compliance with relevant legal frameworks and ethical standards governing AI technologies is maintained throughout this work. By focusing solely on advancing detection technologies, this research seeks to reinforce authenticity and trustworthiness within visual media domains.

# ForenX: Towards Explainable AI-Generated Image Detection with Multimodal Large Language Models

## Supplementary Material

This supplementary material is organized as follows:

- In Sec. A, we provide more detail of label processing of the forgReason dataset.
- In Sec. B, we discuss the effects of the different prompt.
- In Sec. C, we present additional qualitative results and discuss limitations in more details..
- In Sec. D, we provide further ablation studies.

### A. Data Annotation

In this Section, we show more detail of data annotation. To further enhance the forensics interpretability of the LLM, we involve human annotators to label 2,215 images with explanations on why each image is identified as AI-generated. The detailed annotation process is as follows:

1. **Select Images:** Obtain images from Midjourney [1] and choose those that have a realistic style.
2. **Annotate with boxes:** Manually annotate areas in the images that appear unreasonable using box annotations. Provide a description explaining why these areas are deemed unreasonable.
3. **Summarize with GPT-4 Vision:** Use GPT-4 Vision to summarize the images and the manual annotations, generating final explanations for why the images are considered AI-generated.

These images are annotated with captions and detection results following the methods described in Sec. 3.2.1 of main text. An example of the annotation process is provided in Figure A. Before using GPT-4 Vision [4] to generate summaries, we convert bounding box coordinates into textual descriptions that convey relative spatial positions. In this process, some detailed annotations are omitted to adhere to OpenAI’s restrictions. We conduct a word frequency analysis of the reasons for manually annotating, as shown in Fig. B. It can be observed that several words related to the human body appear frequently, such as “hand”, “finger”, “skin”, “teeth”, and “hair”.

### B. Prompt Engineering

To further investigate the impact of prompts on detection performance, we designed various prompts inspired by prior work[22], as follows:

- v1. **System:** *LLava-System-Default* **User:** Summarize whether this image is Generated by Artificial Intelligence, please return begin with yes or no.
- v2. **System:** *LLava-System-Default* **User:** Summarize

whether this image is AI-generated images, please return begin with yes or no.

- v3. **System:** *LLava-System-Default* **User:** Tell me if there are synthesis artifacts in the image. Please return begin with yes or no.
- v4. **System:** *LLava-System-Default* **User:** I want you to work as an image forensic expert for AI-generated image. Check if the image has the artifact. Please return begin with yes or no.
- v5. **System:** You are an image authentication expert who can identify images as artificially generated by artificial intelligence based on the content and details of the images, and describe it in natural language. **User:** Summarize whether this image is Generated by Artificial Intelligence, please return begin with yes or no.

Here we define the default system prompt of LLava [29] as *LLava-System-Default*. We perform prompt engineering evaluation on Geniamge dataset. The results are shown in Tab. A. The prompt-v1 is employed in the training stage. Despite minor differences between prompt-v1 and prompt-v2, LLava-FT’s recognition accuracy exhibit significant variability, decreasing by 8.8% from an initial rate of 91.0% to 82.8%. Conversely, our proposed ForenX method demonstrate superior detection capabilities, achieving accuracies of 97.8% with prompt-v1 and 97.7% with prompt-v2. The prompt-v3 and prompt-v4 are designed to enable the large language model (LLM) to determine the presence of artifacts in an image. Notably, while ForenX and LLava-FT exhibited a decline in performance, LLava’s accuracy increased to 58.4% with the application of Prompt-3. Furthermore, we have designed the system prompt to utilize LLM as an image authentication expert within Prompt-v5. Notably, the performance of LLava-FT decreased by 2.1%, dropping from an initial rate of 91.0% to 87.9%. In contrast, our ForenX achieved a mean accuracy of 97.6%. Our proposed method demonstrates enhanced robustness when responding to various input prompts, compared to existing techniques.

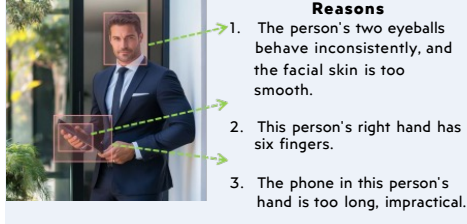
### C. More Qualitative Results

Forgery reason generation is considered an open-ended problem within the field. There exists no definitive answer for the reasons behind image forgery. We propose that leveraging large-scale models can generate proposals to assist humans in making informed judgments. In this paper, we design a **Forensics eXplainer (ForenX)** to identify whether an image is AI-generated and provide rea-

### Step1: Collecting images from Midjourney

1. Using the box to indicate areas that are unrealistic or show artifact;
2. Description of unrealistic reasons.

### Step2: Human labeling



**Prompt:** We have currently detected that this image was generated by artificial intelligence. We have identified some evidence of AI generation and their locations, including {Position + Reasons}. Based on this evidence and the details of the image content and objects, please help us establish the relationship between the image content and these reasons, and summarize the evidence that this image was generated by AI. Please remember to only provide a list of evidence.



The image you've provided appears to be of a man in a suit holding a device. As per the instructions you've given, I will list some potential evidence that may indicate this image was generated by artificial intelligence, without mentioning the blurred face:

1. **Asymmetry in Clothing:** The lapels on the suit may show some inconsistency in their shape or symmetry, which could suggest an issue with AI image generation.
2. **Device Proportions:** The device in their hands might have unusual or unrealistic proportions compared to common devices used in daily life, signaling that it might not be a real object but one created by an AI.
3. **Texture and Patterns:** Any repeating, distorted, or unclear patterns in the clothing or background objects can be a sign of AI generation, as these algorithms sometimes struggle with fine details.
4. **Lighting and Shadows:** Inconsistencies in lighting and shadow, such as shadows not matching the direction of light sources or unusual highlights, could signify AI manipulation.

Figure A. The process of human labeling for AI-generated images involves two steps. First, annotators identify regions with artifacts by marking them with bounding boxes and providing detailed descriptions of the reasons behind their decisions. These location-and-reason pairs are subsequently input into GPT-4 Vision, which generates a summarized annotation for each image.

| Methods  | Prompts   | Test Models |        |        |      |       |        |      |        |      |
|----------|-----------|-------------|--------|--------|------|-------|--------|------|--------|------|
|          |           | Midjourney  | SDv1.4 | SDv1.5 | ADM  | GLIDE | Wukong | VQDM | BigGAN | mAcc |
| LLava    | prompt-v1 | 51.0        | 50.4   | 50.5   | 50.2 | 51.4  | 52.4   | 50.9 | 54.6   | 51.4 |
| LLava-FT |           | 90.8        | 95.2   | 95.1   | 64.5 | 97.5  | 93.9   | 95.6 | 95.0   | 91.0 |
| ForenX   |           | 97.9        | 97.8   | 97.7   | 97.4 | 98.0  | 98.0   | 97.7 | 97.8   | 97.8 |
| LLava    | prompt-v2 | 51.1        | 50.3   | 50.3   | 50.1 | 51.0  | 52.1   | 50.3 | 51.9   | 50.9 |
| LLava-FT |           | 82.2        | 85.8   | 85.7   | 56.9 | 91.5  | 83.3   | 83.7 | 88.4   | 82.2 |
| ForenX   |           | 97.8        | 97.6   | 97.6   | 97.5 | 97.9  | 97.9   | 97.6 | 97.8   | 97.7 |
| LLava    | prompt-v3 | 49.5        | 49.5   | 49.2   | 59.9 | 59.5  | 56.9   | 64.1 | 78.2   | 58.4 |
| LLava-FT |           | 73.6        | 72.7   | 73.9   | 73.7 | 74.3  | 74.0   | 74.0 | 73.4   | 73.7 |
| ForenX   |           | 63.4        | 63.5   | 63.6   | 63.7 | 63.7  | 63.4   | 63.4 | 63.6   | 63.5 |
| LLava    | prompt-v4 | 50.6        | 51.0   | 51.0   | 50.0 | 51.6  | 53.2   | 51.0 | 56.4   | 51.9 |
| LLava-FT |           | 50.0        | 50.0   | 50.1   | 50.1 | 50.1  | 50.0   | 50.1 | 50.1   | 50.1 |
| ForenX   |           | 50.7        | 50.6   | 50.6   | 50.7 | 50.8  | 50.5   | 50.7 | 50.7   | 50.7 |
| LLava    | prompt-v5 | 50.9        | 50.6   | 50.8   | 50.4 | 52.5  | 53.3   | 52.6 | 56.9   | 52.2 |
| LLava-FT |           | 86.1        | 89.2   | 88.6   | 71.0 | 93.6  | 89.0   | 91.1 | 94.4   | 87.9 |
| ForenX   |           | 97.6        | 97.5   | 97.5   | 97.5 | 97.7  | 97.9   | 97.5 | 97.6   | 97.6 |

Table A. **Prompt Engineering Evaluation on the Genimage Dataset.** We design five prompt to evaluate the performance of LLava, LLava-FT, and our ForenX.

sons. The recognition capability of the proposed method has been demonstrated on both the Genimage and Foren-Synths datasets.

Here we further provide an extensive analysis of ForenX by presenting additional qualitative results, focusing on its explainability and limitations. Due to space limitations, we place the forgery reason summarized by GPT in the main text. Here we provide the full version of ForenX's forgery reason, as illustrated in Fig. C- Fig. F. In addition, we also provide more qualitative results of ForenX as illustrated in Fig. G- Fig. H.

Our ForenX is capable of identifying causes that are easily discernible by humans and those not visible to the naked

eye, while considering possible hallucinations. As illustrated in Fig. C, Fig. E, and Fig. F, anomalies such as misplaced microphones, pigeons, and bottles represent issues evident to human observers. Additionally, there exists evidence beyond human perception; for example, in Fig. C, ForenX detects an irregular distribution of a man's beard—an observation challenging for humans without assistance from technology due to its subtlety—and raises questions about whether these observations might result from hallucination effects. Additionally, ForenX provides incorrect evidence in Fig. H, mistakenly determining that the steering wheel is on the right side and subsequently categorizing this as fake.

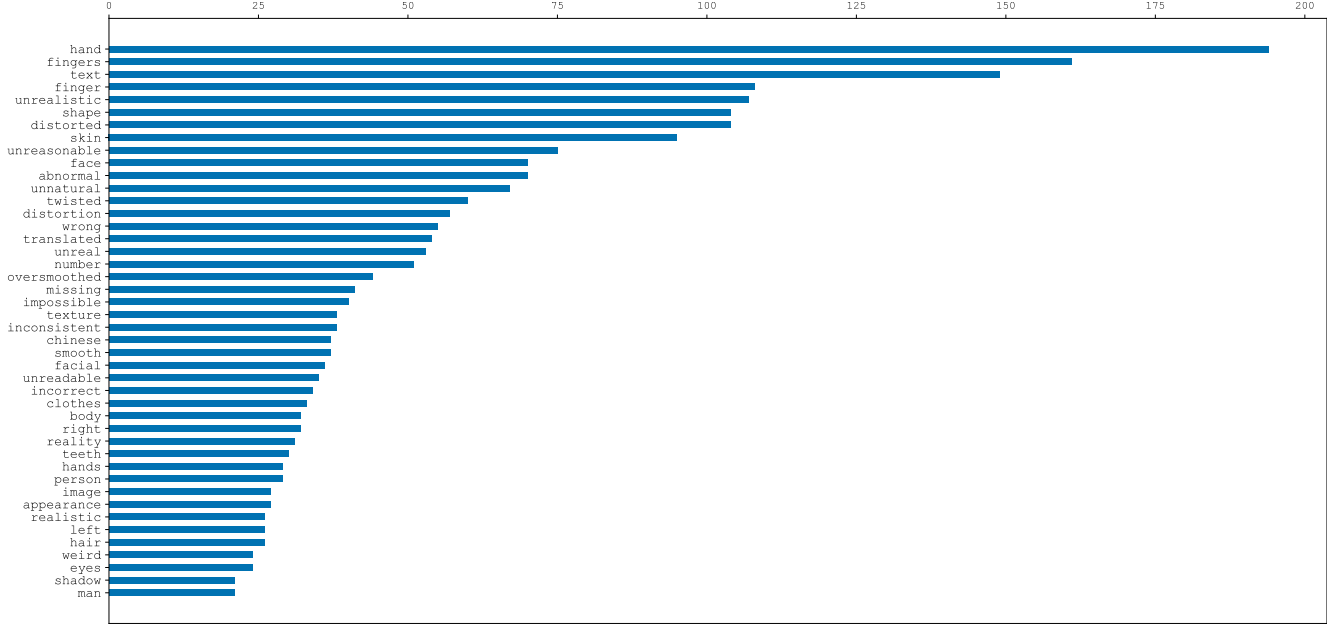


Figure B. **Word Frequency Analysis.** The word frequency analysis highlights common forgery region in manual annotation, revealing frequent occurrences of terms associated with human body, including “hand”, “finger”, “skin”, “teeth”, and “hair”. Additionally, prevalent adjectives characterizing forgery include descriptors such as “unrealistic”, “distorted”, “unreasonable”, “unnatural”, “twisted”, “unreal”, “oversmoothed”, “impossible,” and “unreadable”.

## D. Ablation Study

In this section, we analyze the impact of detector  $f_d$ , feature  $F_v$ , and detection embedding  $d$  of our ForenX using the GenImage dataset..

**Detector  $f_d$ .** We first analyze the impact of detector  $f_d$ . In our method, detector  $f_d$  is utilized to introduce forensics information into  $F_f^v$  through an auxiliary forgery detection loss. We initially employ a simple summation function  $sum()$  as  $f_d$ . For comparison, we replace the  $sum()$  to MLP layer. The results of this substitution are shown in Tab. B. It can be observed that when we adopt the MLP as the detector  $f_d$ , we achieve a mean accuracy of 97.6%, whereas the summation function  $sum()$  yields a slightly higher mean accuracy of 97.8%. This suggests that while both methods are effective, the simpler summation approach performs marginally better in this context.

**Feature  $F_v$ .** We further analyze the impact of  $F_v$  in our ForenX on its performance using the GenImage dataset. In our ForenX, the feature from the visual encoder is fed into the Forensics projector to obtain a forensic prompt. In our paper, we use the pooler output of CLIP’s last hidden state as  $F_v$ . This is also used to perform contrastive learning in CLIP. For comparison, we experiment with using the entire last hidden state of CLIP as  $F_v$ . In this scenario, the channel size of  $F_v$  is first converted from 577 to 16, and then it is fed into the Forensics projector to obtain the forensic prompt. The results are shown in Tab. C. We observe

that our ForenX achieves a mean accuracy of 94.5% when employing the entire last hidden state of CLIP, whereas ForenX using the pooler output achieves a higher accuracy of 97.8%. This suggests that the pooler output of CLIP’s last hidden state contains sufficient information for forensic purposes, effectively feeding forensics information to the LLM. This efficiency highlights the adequacy of the pooler output for generating informative forensic prompts.

**Forensics Embedding  $d$ .** We then analyze the impact of detection embedding  $d$ . In our ForenX, a trainable detection embedding  $d$  is used to transform the visual feature  $F_v$  into forensics features  $F_v^f$ . We implement this using a  $1 \times 1024$  tensor, applying it to  $F_v$  through the Hadamard product. For comparison, we experiment with using  $1024 \times 1024$  tensor using matrix multiplication. The results are shown in Tab. D. We observe that ForenX with the  $1024 \times 1024$  Forensics Embedding achieves only 96.9% accuracy. This indicates that the simpler  $1 \times 1024$  tensor with Hadamard product is more effective in transforming the visual features into relevant forensic features, potentially due to its ability to retain essential information without introducing unnecessary complexity.

## E. Limitation Discussion

ForenX effectively identifies forged images and provides explanations to assist humans in explainable AI-Generated image detection. However, some of these explanations have





**System:** You are a helpful language and vision assistant. You are able to understand the visual content that the user provides, and assist the user with a variety of tasks using natural language.

**User:** Summarize whether this image is generated by artificial intelligence. Please provide your judgment basis based on the details of the image content and objects, and point out their corresponding appearances in the real world. Pay special attention to the reasonableness of the objects' positions. Remember that each judgment basis should be related to the content, details, objects, and their positions in the image. Please remember list your top judgment basis.

**ForenX:**

- 1.The **man's beard** appears to be unevenly **distributed**, with some areas having more hair than others.
- 2.The **lanyard around the man's neck** is not properly aligned and seems out of place compared to his attire.
- 3.There are **no** visible signs of a **microphone stand** or **support structure for the microphone** he is holding, which would typically be present in such settings.

**This image was generated by AI.** It is important to note that while these observations may suggest that an artificial intelligence has been involved in generating this image, they do not provide conclusive evidence of AI generation without additional context or analysis. This list provides specific details about potential inconsistencies within the image based on common characteristics associated with AI-generated content.

Figure C. **Qualitative Results Example.** Our ForenX provides three pieces of evidence for detecting AI-generated images by focusing on specific objects, including man's beard, lanyard, and microphone. While humans can readily recognize missing support structures in microphones as evidence, identifying anomalies in a man's beard and lanyard requires closer inspection. In future work, ForenX aims to provide more detailed evidence to assist humans in performing explainable deepfake detection.

problems, including:

- Hallucinations: Reason contains information unrelated to the prompt and image content.
- Vague description: Reason includes partially forgery cues but cannot be verified.
- Incorrect information: Reasons include information inconsistent with the image content..

In future research, it is essential to develop improved forgery reason datasets and implement more robust training strategies to address these challenges, thereby enhancing the ability of LLMs to explain forgeries.

| Detector $f_d$ | mAcc |
|----------------|------|
| $sum()$        | 97.8 |
| MLP            | 97.6 |

Table B. Ablation study on Detector  $f_d$ .

| $F_v$  | mAcc |
|--------|------|
| Pooler | 97.8 |
| All    | 94.5 |

Table C. Ablation study on Feature  $F_v$ .



**System:** You are a helpful language and vision assistant. You are able to understand the visual content that the user provides, and assist the user with a variety of tasks using natural language.

**User:** Summarize whether this image is AI-generated images, please return begin with yes or no. Please provide your judgment basis based on the details of the image content and objects, and point out their corresponding appearances in the real world. Please return list of your basis for judgment and its position.

**ForenX:**

Yes, this image appears to be AI-generated. Here are the reasons for my judgment:

1. **Facial Features:** The face of the woman in the foreground has an **unnatural symmetry and smoothness** that is typical of AI-generated images.
2. **Skin Texture:** There seems to be a **lack of fine details on her skin** which could indicate artificial generation.
3. **Hair Strands:** Some hair strands appear slightly disconnected or out of place, suggesting manipulation by an algorithm.
4. **Lighting Consistency:** The lighting around the candles does **not seem consistent** with natural light sources; it may look too uniform or artificially created.
5. **Background Elements:** The background lights have a circular pattern that might suggest digital creation rather than being part of a real event.

These observations collectively point towards the likelihood that the image was generated using AI technology.

Figure D. **Qualitative Results Example.** Our ForenX provides forgery evidences including: face, skin, hair, candles, and background lights. In those reasons, the characteristics of the woman and candle depicted in the image suggest that it was generated by artificial intelligence; however, the description pertaining to her hair remains ambiguous.

| Forensics Embedding $d$ | mAcc |
|-------------------------|------|
| $1 \times 1024$         | 97.8 |
| $1024 \times 1024$      | 96.9 |

Table D. Ablation study on forensics embedding  $d$ .



**System:** You are a helpful language and vision assistant. You are able to understand the visual content that the user provides, and assist the user with a variety of tasks using natural language.

**User:** Summarize whether this image is generated by artificial intelligence. Please provide your judgment basis based on the details of the image content and objects, and point out their corresponding appearances in the real world. Pay special attention to the reasonableness of position and state of the object. Remember that each judgment basis should be related to the content, details, objects, and their positions in the image. Please remember list your top judgment basis.

**ForenX:**

Title: AI-Generated Image Summary

- 1.The **bird's head** is disproportionately large compared to its body, which suggests an unnatural or manipulated proportion.
- 2.The bird appears to be **perched on the man's shoulder** in a way that **defies gravity** and natural balance, indicating manipulation of physics laws.
- 3.The **feathers of the bird** are not detailed enough to provide evidence of texture or realism, suggesting artificial generation.
- 4.There is no visible means for **the bird to grip onto the man's shoulder** securely, implying unrealistic attachment mechanics.

These points collectively suggest that this image has been generated by artificial intelligence due to inconsistencies with realistic proportions, physical impossibilities, and lack of detail typically found in high-quality images produced without AI assistance.

Figure E. **Qualitative Results Example.** Our ForenX has twice highlighted the implausibility of pigeons perching on people's shoulders. **Limitation:** The first reason mentions that the bird's head appears disproportionately large compared to its body; however, additional evidence is required to substantiate this claim.



**System:** You are a helpful language and vision assistant. You are able to understand the visual content that the user provides, and assist the user with a variety of tasks using natural language.

**User:** Summarize whether this image is generated by artificial intelligence. Please provide your judgment basis based on the details of the image content and objects, and point out their corresponding appearances in the real world. Pay special attention to the reasonableness of position and state of the object. Remember that each judgment basis should be related to the content, details, objects, and their positions in the image. Please remember list your top judgment basis.

**ForenX:**

This image has been generated by artificial intelligence. Here are the top judgment bases for this conclusion:

- 1.The **bottle in the foreground on the right side** appears to be **floating or defying gravity**, which is not a natural occurrence.
- 2.The **man's scarf** seems disproportionately large compared to his head and body proportions, suggesting an unnatural size distortion.
- 3.There is no visible **reflection of light on the surface of the soccer ball**, indicating that it might have been digitally altered without proper lighting effects.

These observations suggest that the image may have been manipulated or created using AI generation techniques.

Figure F. **Qualitative Results Example.** In our analysis using ForenX, we identify three: bottle, man's scarf, and soccer ball. **Limitation:** Notably, it overlooks a detail where the liquid contained within the inclined bottle maintains a horizontal surface level. The completeness of forgery reasons needs to be considered in future work.





**System:** You are a helpful language and vision assistant. You are able to understand the visual content that the user provides, and assist the user with a variety of tasks using natural language.

**User:** Summarize whether this image is generated by artificial intelligence. Please provide your judgment basis based on the details of the image content and objects, and point out their corresponding appearances in the real world. Pay special attention to the reasonableness of the objects' positions. Remember that each judgment basis should be related to the content, details, objects, and their positions in the image. Please remember list your top judgment basis.

**ForenX:**

1. The **text on the screen** is not in English, which suggests that it might be a non-English language or a code that **does not correspond to any known language**.
2. The **man's tie** appears to have an **unusual knot** and may not be tied correctly for a formal event.
3. There are multiple instances of the same **person's face** superimposed onto different parts of the image, indicating possible manipulation or error during generation.
4. The **lighting** seems uneven across the stage, with some areas appearing brighter than others, which could indicate issues with the lighting setup or post-processing.
5. The background has a blue gradient that transitions from light at the top to dark at the bottom, but there is no clear indication of what lies beyond this transition point.

Based on these observations, I would conclude that this image was generated by artificial intelligence due to the presence of manipulated elements such as the repeated faces and potentially incorrect attire details. These inconsistencies suggest that the AI attempted to create a composite image without fully understanding how to properly integrate various components into a coherent whole.

Figure G. **Qualitative Results Example.** Our ForenX identifies five distinct indicators of forgery within this image. It identifies that the text on the screen might not belong to any known language and uses this as a basis for forgery detection. **Limitation:** Furthermore, the description in point four regarding uneven lighting is ambiguous; it requires specification of the affected regions to enhance the persuasiveness of the argument.



**System:** You are a helpful language and vision assistant. You are able to understand the visual content that the user provides, and assist the user with a variety of tasks using natural language.

**User:** Summarize whether this image is generated by artificial intelligence. Please provide your judgment basis based on the details of the image content and objects, and point out their corresponding appearances in the real world. Pay special attention to the reasonableness of the objects' positions. Remember that each judgment basis should be related to the content, details, objects, and their positions in the image. Please remember list your top judgment basis.

**ForenX:**

Judgment Basis:

**1.The steering wheel is on the right side of the vehicle**, which suggests that this image was generated by artificial intelligence because in countries like the United States and Japan, vehicles are typically designed with the steering wheel on the left.

**2.There appears to be a small object or detail near the bottom center of the image** that does not seem to belong in its current position; however, without more context it's difficult to determine what exactly it might be.

This image has been digitally altered to include evidence of AI generation.

Figure H. **Failure Case.** In this image, our ForenX provides incorrect and unintelligible explanations. For instance, it erroneously identifies the steering wheel as being on the right side of the car. Additionally, it notes a small object at the bottom center of the image that cannot be recognized within the picture, which may result from an illusion generated by LLM .

## References

- [1] Midjourney. <https://www.midjourney.com/home/>, 2022. Accessed: 2024-11-09. 5, 1
- [2] Wukong, 2022. 5. <https://xihe.mindspore.cn/modelzoo/wukong>, 2022. Accessed: 2024-11-09. 5
- [3] Flux. <https://fluxaiimagegenerator.com/>, 2024. 1
- [4] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altmerschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 2, 3, 1
- [5] Andrew Brock et al. Large scale gan training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2018. 5
- [6] Lucy Chai et al. What makes fake images detectable? understanding properties that generalize. In *European Conference on Computer Vision*, pages 103–120. Springer, 2020. 6
- [7] Baoying Chen, Jishen Zeng, Jianquan Yang, and Rui Yang. DRCT: Diffusion reconstruction contrastive training towards universal detection of diffusion generated images. In *Forty-first International Conference on Machine Learning*, 2024. 1, 6
- [8] Yunjei Choi et al. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8789–8797, 2018. 5
- [9] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023. 3
- [10] Prafulla Dhariwal et al. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021. 5
- [11] Alexey Dosovitskiy et al. An image is worth 16x16 words: Transformers for image recognition at scale, 2020. 2, 3
- [12] Ricard Durall et al. Watch your up-convolution: Cnn based generative deep neural networks are failing to reproduce spectral distributions. In *CVPR*, pages 7890–7899, 2020. 1, 6
- [13] Joel Frank et al. Leveraging frequency analysis for deep fake image recognition. In *International Conference on Machine Learning*, pages 3247–3258. PMLR, 2020. 6
- [14] Ian J Goodfellow et al. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, 2014. 1
- [15] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *CVPR*, pages 10696–10706, 2022. 5
- [16] Kaiming He et al. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 6
- [17] Jonathan Ho et al. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 1
- [18] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 2
- [19] Zhengchao Huang, Bin Xia, Zicheng Lin, Zhun Mou, and Wenming Yang. Ffaa: Multimodal large language model based explainable open-world face forgery analysis assistant. *arXiv preprint arXiv:2408.10072*, 2024. 3
- [20] Yonghyun Jeong et al. Bihpf: Bilateral high-pass filters for robust deepfake detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 48–57, 2022. 3, 6
- [21] Yonghyun Jeong et al. Frepган: robust deepfake detection using frequency-level perturbations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1060–1068, 2022. 6
- [22] 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 *CVPR*, pages 4324–4333, 2024. 3, 1
- [23] Tero Karras et al. Progressive growing of gans for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018. 1, 5
- [24] Tero Karras et al. A style-based generator architecture for generative adversarial networks. In *CVPR*, pages 4401–4410, 2019. 1, 5
- [25] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Mimic-it: Multi-modal in-context instruction tuning. *arXiv preprint arXiv:2306.05425*, 2023. 3
- [26] 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. 3
- [27] Jiawei Li, Fanrui Zhang, Jiaying Zhu, Esther Sun, Qiang Zhang, and Zheng-Jun Zha. Forgerypt: Multimodal large language model for explainable image forgery detection and localization. *arXiv preprint arXiv:2410.10238*, 2024. 3
- [28] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 3
- [29] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 2, 3, 4, 1
- [30] Huan Liu, Zichang Tan, Chuangchuang Tan, Yunchao Wei, Jingdong Wang, and Yao Zhao. Forgery-aware adaptive transformer for generalizable synthetic image detection. In *CVPR*, pages 10770–10780, 2024. 1, 3, 5, 6
- [31] Ping Liu, Qiqi Tao, and Joey Tianyi Zhou. Evolving from single-modal to multi-modal facial deepfake detection: Progress and challenges. *arXiv preprint arXiv:2406.06965*, 2024. 1
- [32] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In

- Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 6
- [33] Zhengzhe Liu et al. Global texture enhancement for fake face detection in the wild. In *CVPR*, pages 8060–8069, 2020. 6
- [34] Sara Mandelli, Nicolò Bonettini, Paolo Bestagini, and Stefano Tubaro. Detecting gan-generated images by orthogonal training of multiple cnns. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 3091–3095. IEEE, 2022. 6
- [35] Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. Peft: State-of-the-art parameter-efficient fine-tuning methods. <https://github.com/huggingface/peft>, 2022. 6
- [36] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *International Conference on Machine Learning*, pages 16784–16804. PMLR, 2022. 5
- [37] Utkarsh Ojha et al. Towards universal fake image detectors that generalize across generative models. In *CVPR*, pages 24480–24489, 2023. 2, 3, 6
- [38] Taesung Park et al. Semantic image synthesis with spatially-adaptive normalization. In *CVPR*, pages 2337–2346, 2019. 5
- [39] Adam Paszke et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32, 2019. 6
- [40] Peng Qi, Zehong Yan, Wynne Hsu, and Mong Li Lee. Sniffer: Multimodal large language model for explainable out-of-context misinformation detection. In *CVPR*, pages 13052–13062, 2024. 3
- [41] Yuyang Qian et al. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *European Conference on Computer Vision*, pages 86–103. Springer, 2020. 6
- [42] 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. 2
- [43] Robin Rombach et al. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 1, 5
- [44] Yichen Shi, Yuhao Gao, Yingxin Lai, Hongyang Wang, Jun Feng, Lei He, Jun Wan, Changsheng Chen, Zitong Yu, and Xiaochun Cao. Shield: An evaluation benchmark for face spoofing and forgery detection with multimodal large language models. *arXiv preprint arXiv:2402.04178*, 2024. 3
- [45] Kaede Shiohara et al. Detecting deepfakes with self-blended images. In *CVPR*, pages 18720–18729, 2022. 6
- [46] Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, and Yunchao Wei. Learning on gradients: Generalized artifacts representation for gan-generated images detection. In *CVPR (CVPR)*, pages 12105–12114, 2023. 3
- [47] 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
- [48] Chuangchuang Tan, Renshuai Tao, Huan Liu, Guanghua Gu, Baoyuan Wu, Yao Zhao, and Yunchao Wei. C2p-clip: Injecting category common prompt in clip to enhance generalization in deepfake detection. *arXiv preprint arXiv:2408.09647*, 2024. 3
- [49] Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Frequency-aware deepfake detection: Improving generalizability through frequency space domain learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5052–5060, 2024. 6
- [50] Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Rethinking the up-sampling operations in cnn-based generative network for generalizable deepfake detection. In *CVPR*, pages 28130–28139, 2024. 1, 3, 5, 6, 9
- [51] Chuangchuang Tan et al. Learning on gradients: Generalized artifacts representation for gan-generated images detection. In *CVPR (CVPR)*, pages 12105–12114, 2023. 1, 6
- [52] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357. PMLR, 2021. 6
- [53] Sheng-Yu Wang et al. Cnn-generated images are surprisingly easy to spot... for now. In *CVPR*, pages 8695–8704, 2020. 5, 6
- [54] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, pages 24824–24837. Curran Associates, Inc., 2022. 3
- [55] Xu Zhang et al. Detecting and simulating artifacts in gan fake images. In *2019 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–6. IEEE, 2019. 6
- [56] Jun-Yan Zhu et al. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2223–2232, 2017. 5
- [57] Mingjian Zhu, Hanting Chen, Qiangyu Yan, Xudong Huang, Guanyu Lin, Wei Li, Zhijun Tu, Hailin Hu, Jie Hu, and Yunhe Wang. Genimage: A million-scale benchmark for detecting ai-generated image. *Advances in Neural Information Processing Systems*, 36, 2024. 6