AIGI-Holmes: Towards Explainable and Generalizable AI-Generated Image Detection via Multimodal Large Language Models

Ziyin Zhou[1*#] Yunpeng Luo[2*] Yuanchen Wu[2] Ke Sun[1] Jiayi Ji[1]
Ke Yan[2†] Shouhong Ding[2] Xiaoshuai Sun[1†] Yunsheng Wu[2] Rongrong Ji[1]
[1] Key Laboratory of Multimedia Trusted Perception and Efficient Computing,
Ministry of Education of China, Xiamen University [2] Tencent YouTu Lab
{ziyinzhou,skjack}@stu.xmu.edu.cn {xssun,rrji}@xmu.edu.cn
jjyxmu@gmail.com {petterluo,kerwinyan}@tencent.com
https://github.com/wyczzy/AIGI-Holmes

Abstract

The rapid development of AI-generated content (AIGC) technology has led to the misuse of highly realistic AI-generated images (AIGI) in spreading misinformation, posing a threat to public information security. Although existing AIGI detection techniques are generally effective, they face two issues: 1) a lack of human-verifiable explanations, and 2) a lack of generalization in the latest generation technology. To address these issues, we introduce a large-scale and comprehensive dataset, Holmes-Set, which includes the Holmes-SFTSet, an instruction-tuning dataset with explanations on whether images are AI-generated, and the Holmes-DPOSet, a human-aligned preference dataset. Our work introduces an efficient data annotation method called the Multi-Expert Jury, enhancing data generation through structured MLLM explanations and quality control via cross-model evaluation, expert defect filtering, and human preference modification. In addition, we propose Holmes Pipeline, a meticulously designed three-stage training framework comprising visual expert pre-training, supervised fine-tuning, and direct preference optimization. Holmes Pipeline adapts multimodal large language models (MLLMs) for AIGI detection while generating human-verifiable and human-aligned explanations, ultimately yielding our model AIGI-Holmes. During the inference stage, we introduce a collaborative decoding strategy that integrates the model perception of the visual expert with the semantic reasoning of MLLMs, further enhancing the generalization capabilities. Extensive experiments on three benchmarks validate the effectiveness of our AIGI-Holmes.

1. Introduction

The rapid evolution of AI technologies like GANs [28, 43] and Diffusion models [21, 73] has made generated images highly realistic. While beneficial for digital art and film, these technologies also pose risks such as misinformation, privacy breaches, and deepfakes. Recent advancements in diffusion models (e.g., FLUX [4], SD3, SD3.5 [22]) and autoregressive techniques (e.g., VAR [87]) have further complicated detection, highlighting the urgent need for effective AI-generated image detection methods.

Recent studies [46, 63, 85, 91, 93, 100] have demonstrated remarkable advancements in AI-generated image detection. While these improvements are noteworthy, two critical issues limit their application and generalization in real-world scenarios: **1. Lack of explanation:** Current detection models are black boxes (Fig. 1(a) , making their detection results difficult for humans to verify. The lack of human-verifiable explanations leads to unreliable detection results. **2. Lack of generalization:** The rapidly evolving AIGC technologies (Fig. 1(c)) persistently challenge their generalization capabilities. Therefore, developing explainable and generalizable AI-generated image detection algorithms is becoming increasingly urgent.

Recent breakthroughs in Multimodal Large Language Models (MLLMs) [13, 33, 34, 54, 69, 83, 99, 101, 106] offer a promising pathway: their exceptional capabilities in commonsense understanding and natural language generation enable semantic-level analysis of visual content. This makes MLLMs a strong candidate for explainable and generalizable AI-generated image detection algorithms. There have been preliminary explorations of using MLLMs for AI-generated images [5, 44]. However, merely employing MLLMs for binary classification predictions does not fully leverage their potential. We aspire for Multimodal Large Language Models (MLLMs) to emulate the capabilities of the renowned de-

This work was done during an internship at Tencent YouTu Lab.
* Equal Contribution.
† Corresponding Author.

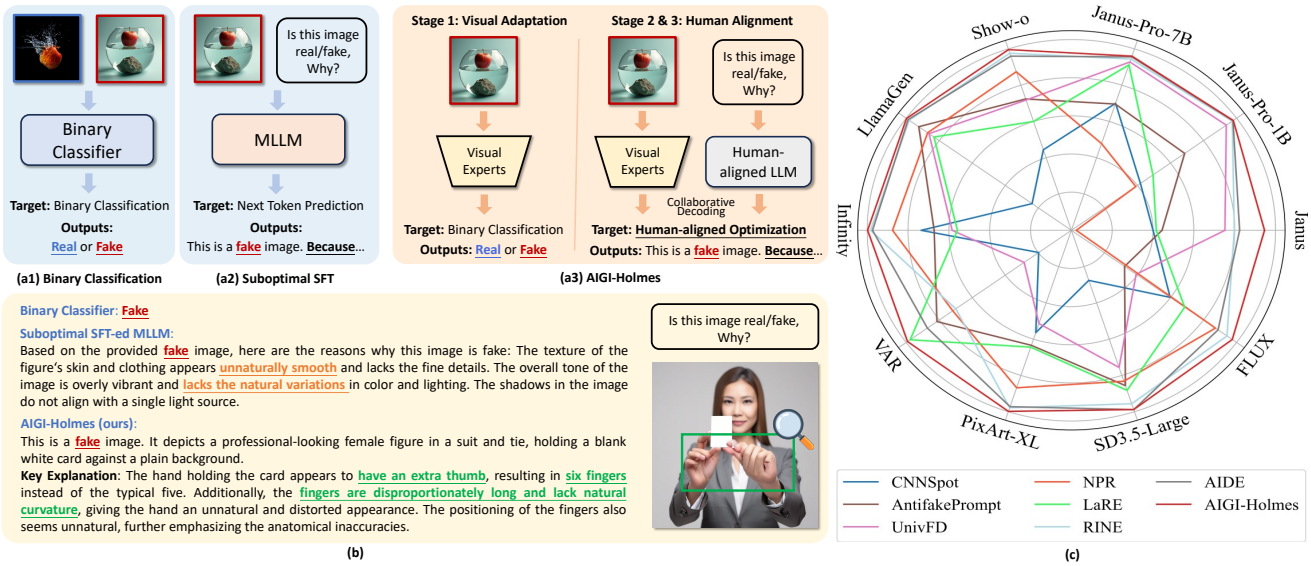

Figure 1. (a): Comparison of AIGI-Holmes with existing methods, (b): A qualitative example to illustrate the effect of AIGI-Holmes. This image is sourced from [24]. (c): AIGI-Holmes outperforms existing baseline methods on state-of-the-art generators under unseen settings.

tective Sherlock Holmes, not only by accurately identifying the culprit but also by providing precise and corresponding evidence, thereby offering human-verifiable explanations.

However, two key challenges hinder MLLM's effective use in AI-generated image detection: 1). **Training Data Scarcity**: As shown in Tab. 1, existing AI-generated image detection datasets, such as CNNDetection [90], GenImage [110], and DRCT [7], consist solely of visual modalities and lack instruction fine-tuning datasets suitable for the SFT(Supervised Fine-tuning) phase of MLLMs. Although FakeBench [53] and LOKI [102] make initial attempts in this field, the use of GPT-4o [1] and the high cost of human labor limit the expansion of these datasets, rendering them unsuitable for SFT of MLLMs. 2). **Suboptimal Supervised Fine-Tuning.** Simply training MLLMs on Supervised Fine-Tuning datasets (even with explanations) often leads to sub-optimal performance (as shown in Tab. 3 and Tab. 4). Possible reasons include: 1. MLLMs exhibit somewhat insufficient capabilities in image classification tasks [105] and low-level perception [96], which are often closely related to the generalization of AI-generated image detection. 2. SFT models may mechanically replicate explanation templates without genuinely understanding the underlying causes of artifacts or semantic errors.

To address the issues with the training dataset, we first introduce two key datasets: Holmes-SFTSet and Holmes-DPOSet. The Holmes-SFTSet provides 65K images annotated with explanations across high-level semantic dimensions (e.g., physical inconsistencies, anatomical errors, and text rendering flaws) and low-level artifacts (e.g., overall hue, texture, edges), rigorously refined through cross-model validation and expert-guided filtering to ensure alignment with human-verifiable evidence. To further the critical need for human-aligned judgment in AI-generated image detection, we introduce the Holmes-DPOSet by constructing contrastive explanation pairs through Positive/Negative prompts on 65K images from the Holmes-SFTSet, along with iterative expert corrections for an additional 4K image explanations, effectively bridging the gap between model perception and human reasoning.

Building upon this training dataset, we propose the Holmes Pipeline, which introduces three key stages to enhance the generalization and interpretability of AIGI detection through systematic MLLM fine-tuning to address the issue of Suboptimal Fine-tuning. The process begins with Visual Expert Pre-training, which leverages the Holmes-SFTset to rapidly adapt the visual encoder through binary classification, establishing domain-specific feature extraction. Following this, Supervised Fine-Tuning enables MLLMs to not only detect synthetic content but also generate human-verifiable explanations. This stage addresses the "black box" limitation of conventional binary classification approaches. Finally, Human-aligned Direct Preference Optimization [71] utilizes the Holmes-DPOset. This stage fundamentally reshapes the reasoning patterns of MLLMs by learning from preference samples, ensuring that the interpretative results align with human judgment standards rather than suboptimal fine-tuning. During inference, our collaborative decoding strategy integrates the model perception of the visual expert with the semantic reasoning of MLLMs, creating a dual-channel verification process that enhances the generalizability of our approach. Our method demonstrates superior performance over state-of-the-art approaches in AI-generated image detection, while producing human-aligned explanations that enhance detection reliability.

In summary, our contributions are threefold:

- **Dataset:** We introduce Holmes-SFTSet and Holmes-DPOSet, the first explanation-rich datasets that include human-verifiable evidence through semantic annotations and contrastive preference pairs, addressing the critical training data scarcity in AI-generated image detection.
- **Methodology:** We propose the Holmes Pipeline, a systematic training pipeline for multimodal large language models (MLLMs) that includes visual expert pre-training, explanation-aware supervised fine-tuning, and human-aligned direct preference optimization. This pipeline synergizes model perception with semantic reasoning through novel collaborative decoding during the inference process.
- **Performance:** Our method achieves state-of-the-art detection accuracy on three benchmark datasets and provides human-verifiable explanations, demonstrating superior generalizability and alignment with human judgment.

2. Related Work

2.1. Detection of AI-Generated Fake Images

With the advancement of AI-based image generation technologies, numerous detection methods have emerged, focusing on training on a single AI-generated image method and generalizing to a wide range of AI-generated images. CNNSpot [91] finds that classifiers trained on ProGAN [42] can generalize to unseen GANs using data augmentation. FreDect [25] detects anomalies in the frequency domain of GAN-generated images. Recent methods have explored new perspectives for better generalization. UnivFD [66] proposes to use pre-trained CLIP-ViT [70] features, generalizing to out-of-distribution (OOD) data through nearest neighbor and linear probing. DIRE [93] introduces Diffusion Reconstruction Error (DIRE), distinguishing real images from Diffusion Model (DM)-generated images by measuring reconstruction errors. DRCT [6] introduces Reconstruction Contrastive Learning (RCL), enhancing generalization by generating challenging samples. PatchCraft [109] detects generated images by segmenting them into small patches, applying SRM filters [26], and examining pixel correlations. NPR [85] introduces neighboring pixel relationships, identifying generated content by analyzing local pixel distribution patterns during upsampling. AIDE [100] develops a two-stream framework using both frequency and semantic information. Despite their acceptable performance, these detection methods cannot explain their underlying principles and struggle to generalize against advanced AI-generated techniques.

2.2. Multimodal Large Language Models

Recent advancements in multimodal language large models (MLLMs) have enhanced Image Forgery Detection and DeepFake Detection, enabling interpretable methods. In DeepFake Detection, DD-VQA [106] and FFAA [34] pioneer the use of MLLMs, leveraging human and GPT-4o [35] annotated datasets to train models based on InstructBLIP [20] and LLaVA [61]. In Image Forgery Detection, FAKESHIELD [99] combines MLLMs with a visual segmentation model (SAM [45]) for explainable detection (e-IFDL), using a GPT-4o-enhanced dataset (MMTD-Set). ForgerySleuth [83] uses MLLMs with a trace encoder to detect tampering and generate detailed analyses, creating the ForgeryAnalysis dataset. For Text Tampering Detection, TextSleuth [69] builds a large dataset (ETTD) using GPT-4o, employing a two-stage analysis paradigm and fusion mask prompts. These works primarily focus on constructing domain-specific SFT datasets for suboptimal fine-tuning, while neglecting alignment with human preferences. Our method introduces a human-aligned preference dataset, Holmes-DPOSet, and employs Direct Preference Optimization [71] to address the suboptimal fine-tuning issue.

Additionally, some works have explored simple multimodal methods for detecting AI-generated images. Liu *et al.* [59] enhanced the generalization of the detection method by considering the text encoding embeddings of CLIP and introducing frequency-related adapters into the image encoder. AntiFakePrompt [5], Bi-Lora [44], and Jia *et al.* [36] redefine the detection task as a visual question-answering task, combining vision-language models to improve performance on unseen data. Our work not only achieves state-of-the-art generalization detection accuracy but also provides further clues as to why an image is or is not AI-generated.

	#Generators	#Image	Explanation	Pref. Data
CNNDetection [90]	11	720K	×	×
GenImage [110]	8	1M+	×	×
DRCT [7]	16	2M	×	×
WildFake [31]	21	3.5M+	×	×
FakeBench [53]	10	6K	✓	×
LOKI [102]	10	3K	✓	×
Holmes-Set	18	65K+4K	✓	✓

Table 1. Comparison between AI-Generated Image Detection Datasets. "Pref. Data" refers to human-aligned preference data.

3. Method

3.1. Data Pipeline

Overview of Data Pipeline. As shown in Fig. 2, our data pipeline consists of four components: Data Source, which includes Data Collection and Image Generation for gathering images to be annotated. These images are then sent to the Automated Annotation section. Subsequently, the annotations are refined in the Preference Modification stage based on expert feedback. Finally, we conduct a Comprehensive Evaluation to assess the model's generalizability and interpretability. In the following sections, we will introduce the detailed methodologies for each component in sequence.

Data Source. To ensure a diverse dataset with various types of forgeries and defects, as shown in Fig. 2, we initially selected 45K images from existing large-scale AI-generated

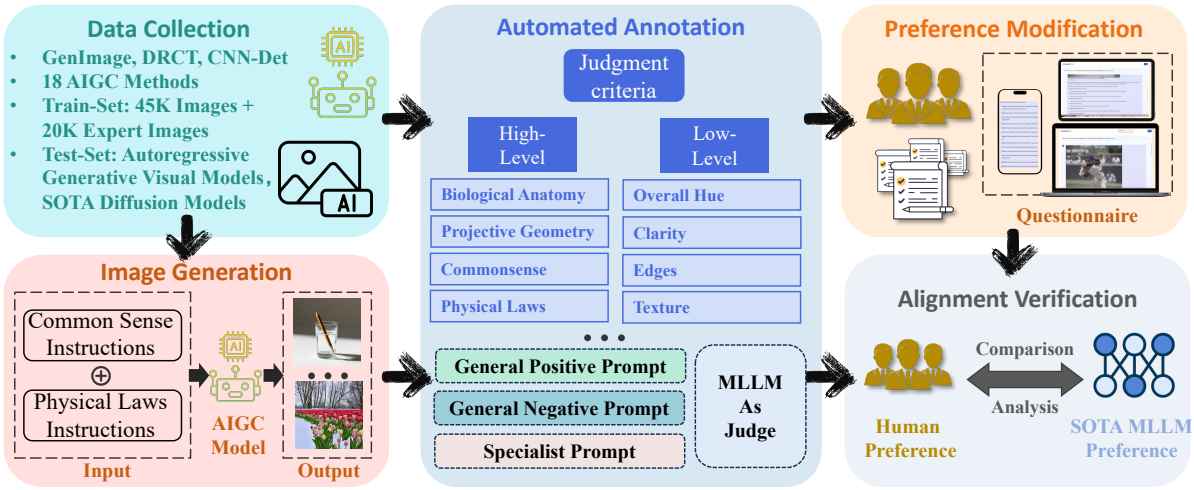

Figure 2. Details of the **Holmes-Set** Construction. The figure illustrates our data pipeline, consisting of four key components: Data Source (including Data Collection and Image Generation), Automated Annotation, Preference Modification (based on human expert feedback), and Comprehensive Evaluation (to assess model generalizability and interpretability).

image detection datasets such as CNNDetection [90], GenImage [110], and DRCT [7] for image description and forgery explanation. To reduce annotation costs, as in previous work [34, 69, 83, 99], we consider using open-source multimodal large language models(MLLMs) for annotation. A key challenge in constructing annotated explanation datasets by MLLMs for AI-generated images lies in two primary aspects. First, unlike deepfake detection and image forgery detection, AI-generated images often lack corresponding real images, complicating annotation efforts through MLLMs. Second, existing MLLMs exhibit limited capability in analyzing fine-grained forgeries within AI-generated images, as evidenced by recent studies [53]. This limitation likely stems from systematic gaps in domain-specific knowledge related to AI-generated image forensics within the training data of MLLMs. Such knowledge deficiencies may prevent MLLMs from reliably identifying common artifacts in AI-generated content, thereby introducing risks of incomplete or inaccurate annotations.

To mitigate the risk of the inaccurate and incomplete annotation results, inspired by [39, 40], we employed expert-guided methods to filter out 20K images with common AI-generated defects, such as text [10, 50], human bodies [23], human faces [34], projective geometry [74], common sense [27], and physical laws [64], from existing datasets. In simple terms, we use expert small models capable of identifying these defects to filter out the images, and then employ MLLMs to determine the presence of these defects. Additionally, due to the lack of attention to commonsense and physical law defects in existing datasets, we supplement the dataset with images generated from [27, 64], as shown in the Image Generation part of Fig. 2. For these images containing common AI-generated defects, we can design targeted prompts to avoid potential annotation hallucinations, thereby reducing the need for extensive human labor in the annotation process. *Further details of the image collection process are elaborated in the Appendix C.*

Automated Annotation. We design the automated annotation system named Multi-Expert Jury, comprising four open-source multimodal large language models (MLLM-Experts): Qwen2VL-72B [89], InternVL2-76B [16], InternVL2.5-78B [15], and Pixtral-124B [65], to ensure high-quality data annotation. The system employs three tailored prompts: 1. **General Positive Prompt** as shown in Fig. 6 and Fig. 7: Annotates 45K randomly selected images using high-level (e.g., anatomy, physical laws) and low-level (e.g., texture, clarity) criteria (see Fig. 1). 2. **General Negative Prompt** as shown in Fig. 6 and Fig. 7: Generates adversarial annotations by asking questions contradicting image authenticity, forming natural positive-negative pairs with the General Positive Prompt annotations to construct the Direct Preference Optimization (DPO) dataset $\mathcal{D}_1$. 3. **Specialist Prompt** as shown in Fig. 9, Fig. 10, Fig. 11, Fig. 12, Fig. 13: Guides MLLM-Experts to annotate 20K expert-filtered defective images, focusing on generation defects (e.g., commonsense). To ensure annotation quality, we adopt an MLLM-as-a-judge approach [8], where MLLM-Experts cross-evaluate each annotation. Only annotations with top consensus scores are retained in the dataset.

Preference Modification. The constructed SFT dataset can be directly used for coarse-grained alignment. However, there may be instances where models mechanically replicate explanation templates without genuinely understanding the underlying causes of artifacts or semantic errors. To further address this issue and narrow the gap between model perception and human reasoning, we introduce the Holmes-DPO dataset. We manually annotate an additional 2K data samples from the same sources as the training set. During this annotation process, we provide the outputs of the SFT model and ask humans to offer modification suggestions, such as

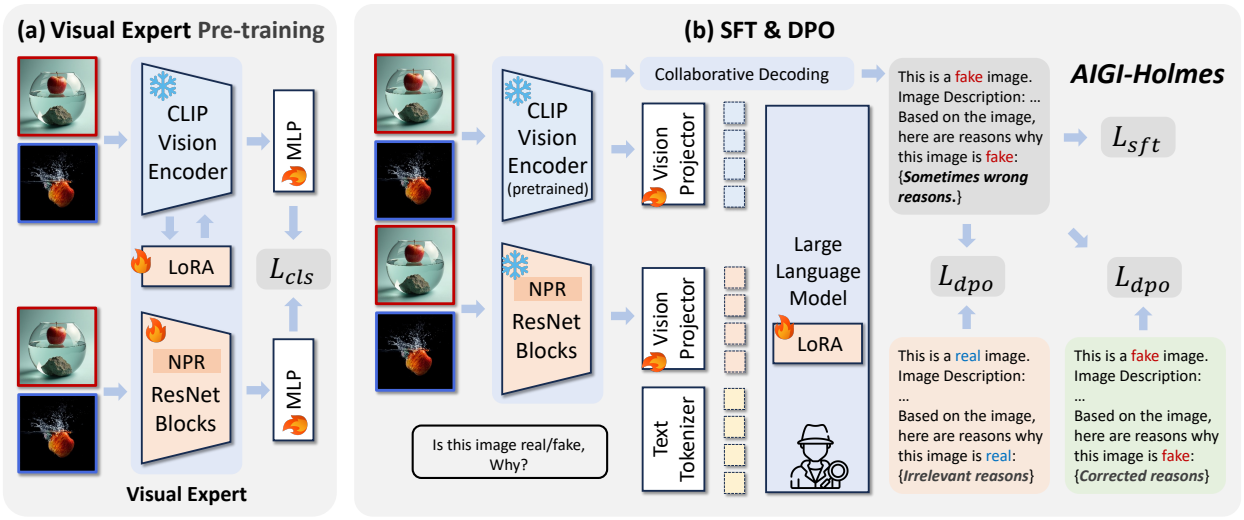

Figure 3. Overview of AIGI-Holmes. We enhance LLaVA [61] with NPR [85] visual expert $\mathcal{R}$ and the Holmes Pipeline, featuring three training stages: Visual Expert Pre-training, SFT, DPO, and a collaborative decoding strategy during inference.

supplementary correct information, removal of incorrect or irrelevant explanations, and other modification opinions. We then use Deepseek-V3 [57], an advanced open-source large language model, to modify the original responses of the SFT model based on the suggestions of human experts, resulting in new, more human-aligned correct explanations. The human experts consist of three authors and two annotators who are strictly guided by us (we reference [39, 40] to create annotation documents for training and guiding the annotators). In addition, we use MLLMs and specialist prompts to modify responses to 2K filtered images with known specific defects. We add the pairs of samples before and after modification to the DPO dataset $\mathcal{D}_2$.

Comprehensive Evaluation. We evaluate the model's capabilities from two perspectives: detection and explanation. **1. For detection,** we output the probabilities of the real/fake tokens to calculate Accuracy (Acc.) and Average Precision (A.P.). **2. For explanation,** we use 1K test samples containing Ground Truth, which have been reviewed by annotators for deficiencies in explanations from a professional perspective and corrected uniformly by the Deepseek-V3 model. We calculate metrics such as BLEU [67], CIDEr [88], METEOR [3], and ROUGE [55] to measure the quality of the explanatory text output by the model. Additionally, we employ multimodal large model scoring and human preference evaluation methods for assessment. For MLLM scoring, we refer to [8], using a prompt that considers relevance, accuracy, comprehensiveness, creativity, and granularity to compare and score the model's responses. For human evaluation scores, we sample 10 images for each type of forgery from $\mathcal{P}_3$, resulting in a total of 100 images. The comparative models generate explanations for these test images, and we use pairwise comparison, as referenced in [17], to calculate the ELO ratings for each model's explanations.

3.2. Overview of AIGI-Holmes

As shown in Fig. 3, for the architecture, we augment the original multimodal method LLaVA [61] with a low-level information NPR [85] visual expert $\mathcal{R}$. For the training methodology, we introduce the Holmes Pipeline, which includes three key training stages: The visual Expert Pre-training Stage, the SFT Stage, and the DPO Stage. During the final inference, we employ a collaborative decoding strategy to merge the predictions from the visual expert and the language model.

3.3. Architecture

Many existing methods use LLaVA [61], a Multimodal Large Language Model (MLLM) with strong multimodal understanding capabilities, as a baseline. However, this architecture faces several challenges: 1. Limited performance when dealing with classification problems [105]. 2. Inefficiency in handling low-level information [96], which is crucial for AIGI detection. To address these challenges, we augment the CLIP visual encoder $\mathcal{F}$ used in LLaVA with a low-level information NPR [85] visual expert $\mathcal{R}$. Given an image input X_{img}, we extract visual features f_{img} and f_{npr} through $\mathcal{F}$ and $\mathcal{R}$, respectively, as shown in the following equations:

$$X_{npr} = \text{NPR}(X_{img});\ f_{npr} = \mathcal{R}(X_{npr});\ f_{img} = \mathcal{F}(X_{img}). \quad (1)$$

Subsequently, the features are injected into the large language model through a projector and text embedding f_t, producing the final output $\mathbf{H}$. This process is represented by the following equations:

$$\mathbf{H} = \text{LLM}(\text{proj}([f_{img}, f_{npr}]), f_t). \quad (2)$$

3.4. Holmes Pipeline

Visual Expert Pre-training Stage: To transform the MLLM into an expert in the AIGI detection domain, we need to ensure that the visual expert provides a certain level of generalization capability and detection accuracy. Therefore, before

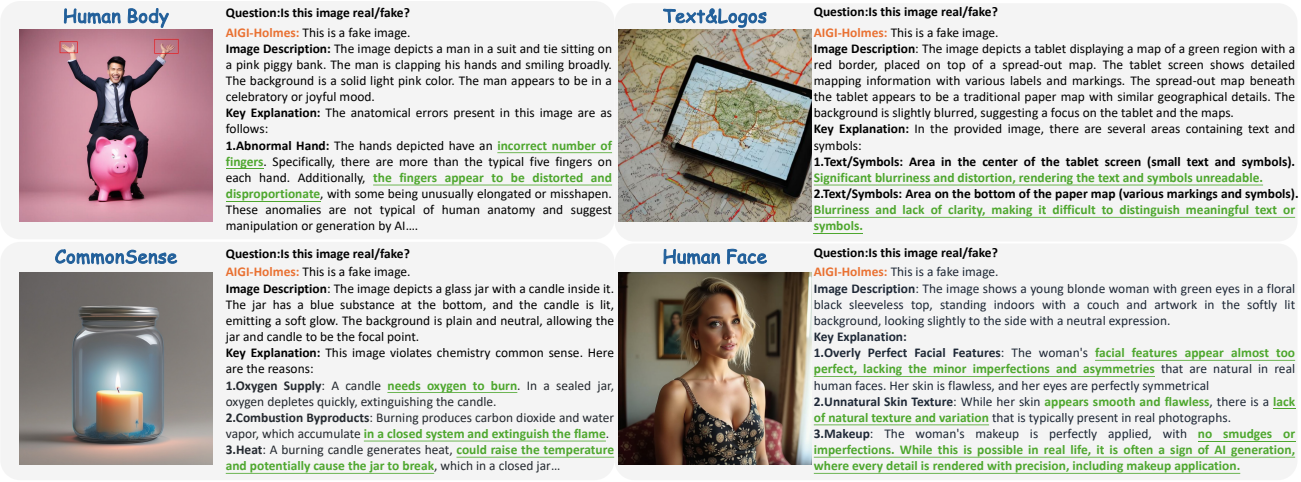

Figure 4. Qualitative results of AIGI-Holmes on AI-Generated images. These images are from [7, 24, 52].

the general LLava training paradigm, we introduce a pre-training stage for the visual expert. We employ LoRA [32] to efficiently fine-tune CLIP-ViT-L/14, setting $r = 4$ and $\alpha = 8$. The CLS features f_{cls} extracted by CLIP are fed into an MLP to obtain classification results. Simultaneously, we independently adjust the NPR-based ResNet. Following [85], we only use the first two layers of ResNet to extract the features f_{npr}. Similarly, an MLP is used to obtain classification results. For ResNet, we perform full parameter fine-tuning. We use binary cross-entropy loss l_{bce} to adjust the visual expert, as shown in the following equation:

$$\begin{aligned} \mathbf{y}_{clip} &= \text{MLP}(f_{cls}), \qquad \mathbf{y}_{npr} = \text{MLP}(f_{npr}), \\ l_{clip} &= l_{bce}(\mathbf{y}_{clip}, \mathbf{y}), \qquad l_{npr} = l_{bce}(\mathbf{y}_{npr}, \mathbf{y}). \end{aligned} \tag{3}$$

SFT Stage: After obtaining the pre-trained visual expert, we integrate it into the large language model. To guide the model to output explanations related to whether an image is AI-generated, we perform SFT on the Holmes-SFT dataset. In this stage, consistent with previous methods, we freeze the visual expert while keeping the linear projector and the LoRA components of the large language model trainable. We optimize these parameters using the autoregressive text loss l_{txt}:

$$l_{txt} = l_{ce}(\mathbf{H}, \mathbf{H}_{txt}). \tag{4}$$

DPO Stage: To further enhance the ability of the SFT model to produce high-quality, human-aligned explanations, we perform human-aligned direct preference optimization on the previously constructed Holmes-DPOSet. Specifically, we sample preference pairs $\{y_w\}, \{y_l\} \sim \mathcal{D}$ from the Holmes-DPOSet, where $\mathcal{D} = \mathcal{D}_1 \cup \mathcal{D}_2$. We then optimize the model using the DPO loss as shown in Eq. 5.

$$\mathcal{L}_{\text{DPO}}(\phi) = -\mathbb{E}_{(x, y_w, y_l)\sim\mathcal{D}} \Bigg[\log \sigma \Bigg(\beta \Big(\log \frac{\pi_\phi(y_w|x)}{\pi_{\text{base}}(y_w|x)} - \log \frac{\pi_\phi(y_l|x)}{\pi_{\text{base}}(y_l|x)} \Big) \Bigg) \Bigg], \tag{5}$$

where σ and β have the same meanings as in the first round. $\pi_\phi(y|x)$ denotes the policy model to be optimized in the second round, with parameters ϕ. $\pi_{\text{base}}(y|x)$ is the model after the first round of preference optimization. During the preference optimization stage, we keep the visual expert frozen, allow the projector to be trainable, and use LoRA to train the large language model part.

Inference Stage: We propose **Collaborative Decoding**, which aims to utilize both our MLLM and its pre-trained expert to jointly decide the authenticity of an image during inference, thereby enhancing the generalization and detection accuracy of AIGI-Holmes. Specifically, we adjust the logit values of the tokens corresponding to "real" and "fake" in the model's output. We denote the logit value for "real" as $\text{logit}(\mathbf{y} = 0)$ and the logit value for "fake" as $\text{logit}(\mathbf{y} = 1)$, for $k \in \{0, 1\}$ as shown in Eq. 6:

$$\begin{aligned} \text{logit}_{new}(\mathbf{y} = k) = \alpha \cdot \text{logit}_{raw}(\mathbf{y} = k) \\ + \beta \cdot \text{logit}(\mathbf{y}_{clip} = k) + \gamma \cdot \text{logit}(\mathbf{y}_{npr} = k), \end{aligned} \tag{6}$$

where $\alpha = 1$, $\beta = 1$, and $\gamma = 0.2$ are the weights assigned to the three prediction results. Through collaborative decoding involving the pre-trained visual expert, we retain the predictions of the MLLM while preventing it from overfitting to existing forgery types, thereby improving the detection accuracy of the MLLM in unseen domains.

4. Experiment

4.1. Experimental Setup

Datasets. To comprehensively evaluate the generalization capabilities of existing methods, we conducted experiments under three settings: **Protocol-I**, **Protocol-II**, and **Protocol-III**. For **Protocol-I**, we trained on the 4-class (car, cat, chair, horse) subset of the CNNDetection dataset, which was widely used in earlier studies, and evaluated the detector on the general and comprehensive benchmark AIGCDetect-

Method	Janus		Janus-Pro-1B		Janus-Pro-7B		Show-o		LlamaGen		Infinity		VAR		PixArt-XL		SD3.5-Large		FLUX		Mean	
	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.
CNNSpot	70.0	86.0	70.9	85.8	85.0	93.6	72.2	86.0	61.9	71.4	86.8	94.6	59.9	75.0	78.2	90.1	63.8	81.1	79.9	92.0	72.9	85.6
AntifakePrompt	72.2	87.4	84.3	94.0	84.8	93.1	86.2	95.5	96.2	99.4	83.6	94.1	90.7	95.6	81.7	92.8	92.8	97.8	66.1	80.8	83.9	93.1
UnivFD	87.6	97.8	96.9	99.5	96.4	99.5	85.9	97.4	93.1	98.6	79.2	96.2	64.3	85.9	75.7	94.4	87.8	97.8	69.6	91.4	83.6	95.9
NPR	51.2	55.9	69.5	75.1	73.9	77.9	93.7	99.6	93.5	99.4	93.8	99.9	85.9	91.2	93.4	99.1	91.6	97.7	93.6	99.5	84.0	89.5
LaRE	70.8	99.3	74.7	97.5	95.6	99.7	80.0	99.0	91.6	99.6	77.9	99.6	98.8	100.0	82.2	99.7	94.1	99.5	84.3	99.0	85.0	99.3
RINE	89.9	98.3	98.7	99.9	97.2	99.6	98.8	99.9	99.1	100.0	99.2	99.9	85.0	97.9	98.9	99.8	97.8	99.7	97.1	99.7	96.2	99.5
AIDE	91.2	99.1	98.9	99.9	97.8	99.8	98.0	99.8	99.4	100.0	98.7	99.9	93.6	99.3	98.6	99.9	99.4	**100.0**	94.4	99.5	97.0	99.7
AIGI-Holmes	**97.3**	**99.9**	**99.0**	**99.9**	**98.0**	**99.9**	**99.8**	**99.9**	**99.9**	**100.0**	**99.9**	**100.0**	**99.6**	**100.0**	**99.9**	**100.0**	**99.4**	99.9	**98.7**	**99.7**	**99.2**	**99.9**

Table 2. Evaluation on the $\mathcal{P}_3$. All baseline results are trained on our training set to ensure a fair comparison.

Benchmark [108] ($\mathcal{P}_1$). For **Protocol-II**, we trained on the training set proposed by AntiFakePrompt [5] and tested on its proposed test set, which includes 18 types of forgeries. For **Protocol-III**, we trained on the dataset containing various Diffusion methods proposed in Sec. 2 and tested on images generated by the latest unseen autoregressive visual generation models Janus [95], Janus-Pro [12], VAR [86], Infinity [30], Show-o [98], LlamaGen [82], and the state-of-the-art diffusion models PixArt-XL [9], FLUX [4], SD3.5 [22] ($\mathcal{P}_3$). Each of the above test sets contains 5K real images from COCO [56] and 5K generated images corresponding to the generation methods.

Implementation Details. During the pre-training phase of the visual expert, we fine-tune CLIP-ViT/L-14 using LoRA (r = 4, α = 8) and fully fine-tune the first two layers of ResNet with NPR as input. The training is conducted for 5 epochs with a batch size of 32. During the SFT phase, we fine-tune LLaVA1.6-mistral-7B [60] using LoRA (rank=128, α=256) while fully training the domain label generator. This model is trained for 3 epochs with a learning rate of $5e-5$, a batch size of 16, and a gradient accumulation step of 1. During the DPO phase, we use LoRA (rank=48, α=96) with a learning rate of $5e-7$, a batch size of 4, a gradient accumulation step of 2, and β=0.1, training for 2 epochs.

Comparison Baselines. We compare various baselines in the main text including CNNSpot [91], NPR [85], AntiFakePrompt [5], LaRE [63], RINE [46], AIDE [100]. *Additional baseline methods will be introduced in the Appendix D.1.*

4.2. Comparisons with SOTA Detection Methods

Protocol-III. The quantitative results in Tab. 2 show the classification accuracy of various methods and generators within the range of $\mathcal{P}_3$. All methods were retrained on our proposed training set to ensure a fair comparison. The test images were generated by unseen state-of-the-art autoregressive visual models and diffusion models. On this challenging benchmark, AIGI-Holmes achieved SOTA results, with accuracy improvements of 15.2%, 3.0%, and 2.2% over the previous best methods NPR, RINE, and AIDE, respectively. For the detection accuracy on the best autoregressive visual generation techniques and diffusion model representatives VAR and FLUX, our method surpassed the best methods by 6.0% and 1.8%, respectively. These three different training settings emphasize the excellent generalization ability of our proposed AIGI-Holmes. *The results of **Protocol-I** and **Protocol-II** can be found in the Appendix D.3.*

4.3. Comparisons with SOTA MLLMs

As demonstrated in Tab. 3, we conduct a quantitative comparison between the textual explanations generated by AIGI-Holmes and those produced by state-of-the-art (SOTA) multimodal large language models (MLLMs). To ensure a fair comparison, the explanations from the baseline models are obtained under the General Positive Prompt query. Our method achieves the highest metrics across nearly all evaluated aspects. For instance, compared to the state-of-the-art closed-source model GPT-4o, our model's output achieves a BLEU-1 score of 0.622, which is an improvement of 0.189 over GPT-4o's score of 0.433. Additionally, for human pairwise scoring, our model achieves an ELO Rating of 11.420, surpassing GPT-4o's score of 10.271 by 1.149 points. *A selection of model outputs is illustrated in the Appendix E.* This result indicates that existing MLLMs have the potential to provide reasonable explanations for AI-generated images. However, due to the lack of targeted downstream task training and the relatively small number of synthetic images included in the dataset, they are unable to perform more precise analyses in the task of explainable AIGI detection. Our method effectively fills this gap.

4.4. Robustness Evaluation

In real-world scenarios, AI-generated images often encounter unpredictable perturbations during dissemination, which can lead to the failure of existing AI detectors. Within the range of $\mathcal{P}_3$, we applied several common perturbations found in real-world scenarios: JPEG compression (QF=75, QF=70), Gaussian blur ($\sigma = 1$, $\sigma = 2$), and downsampling ($\times 0.5$). As shown in Tab. 5, the performance of all methods significantly declines under these distortions. However, AIGI-Holmes achieves higher detection accuracy compared to other baseline methods in these challenging scenarios. A possible reason is our use of the pre-trained CLIP method, which, as indicated in [18], demonstrates good robustness when used as a backbone for AI image detection methods. Additionally, MLLM focuses more on high-level semantic features, reducing the model's reliance on low-level artifacts that are crucial for AI-generated image detection in

MLLM	Automatic Metrics				MLLM-as-Judge Evaluation				Human.
	BLEU-1	ROUGE-L	METEOR	CIDEr	Qwen2VL-72B	InternVL2-76B	InternVL2.5-78B	Pixtral-124B	ELO Ratings
Qwen2VL-72B	0.314	0.227	0.292	0.003	3.874	3.612	4.002	3.163	8.432
InternVL2-76B	0.362	0.224	0.289	0.006	4.042	3.807	4.006	3.463	10.111
InternVL2.5-78B	0.275	0.221	0.293	0.007	4.012	3.531	3.954	3.101	8.623
Pixtral-124B	0.428	0.270	0.302	0.010	3.967	3.990	4.140	4.213	10.472
GPT-4o	0.433	0.308	0.306	0.005	4.102	4.010	4.032	4.010	10.271
AIGI-Holmes(w/o DPO)	0.445	0.315	**0.317**	0.023	4.119	3.918	4.150	4.130	10.670
AIGI-Holmes(w/ DPO)	**0.622**	**0.375**	0.311	**0.107**	**4.196**	**4.011**	**4.189**	**4.227**	**11.420**

Table 3. A comprehensive comparison of the explanations for AI-generated images between pre-trained SOTA MLLMs and AIGI-Holmes. The abbreviations "w/o" stands for "without", and "w/" stands for "with".

VEP-S	DPO	CD	$\mathcal{P}_1$	$\mathcal{P}_3$
			83.3	90.1
	✓		84.8	92.3
✓			86.8	97.2
✓	✓		87.4	97.6
✓		✓	90.8	98.9
✓	✓	✓	**93.2**	**99.2**

Table 4. Ablation Study of core model components.

Method	JPEG Compression		Gaussian Blur		Resize
	QF=75	QF=70	$\sigma = 1.0$	$\sigma = 2.0$	$\times 0.5$
CNNSpot	63.5	62.4	64.5	61.7	59.9
NPR	52.2	51.6	56.8	53.4	74.3
UnivFD	84.7	84.0	81.0	74.9	86.3
LaRE	62.0	63.0	54.3	54.2	51.1
AntifakePrompt	80.1	79.7	78.2	77.6	74.5
AIDE	92.8	92.3	91.9	90.7	89.2
RINE	92.4	91.1	94.2	92.8	92.3
AIGI-Holmes	**99.0**	**98.7**	**98.3**	**97.9**	**95.9**

Table 5. Robustness of Classification Accuracy on JPEG Compression, Gaussian Blur and Resize of AIGI-Holmes. The classification accuracy (%) averaged over 10 test sets in $\mathcal{P}_3$ with specific perturbation.

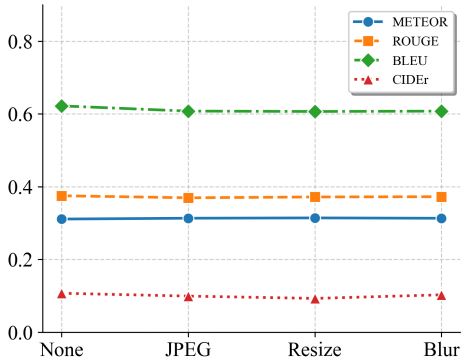

Figure 5. Robustness of the explanation on JPEG Compression (QF=70), Gaussian Blur ($\sigma = 2$), and Resize ($\times 0.5$) of AIGI-Holmes.

other methods. These artifacts are often susceptible to unpredictable perturbations in real-world scenarios. Additionally, as shown in Fig. 5, under these degradation conditions, the evaluation metrics for model explanations such as BLEU-1, ROUGE-L, METEOR, and CIDEr did not exhibit significant declines. This indicates that the explanations generated by the model remain focused on high-level semantic information related to the image content and are not overly affected by these degradation conditions.

4.5. Ablation Study

We conduct ablation experiments on the main innovative methods of AIGI-Holmes: Visual Expert Pre-training Stage (VEP-S), DPO, and Collaborative Decoding (CD), as shown in Tab. 4. The results demonstrate a significant improvement in accuracy when using the Visual Expert Pre-training Stage compared to the original llava paradigm, with detection accuracy increasing by 3.5% and 7.1% on $\mathcal{P}_1$ and $\mathcal{P}_3$, respectively. Adding DPO and Collaborative Decoding on this basis further enhances accuracy. Specifically, DPO helps improve the model's accuracy by 0.6% and 0.4% on $\mathcal{P}_1$ and $\mathcal{P}_3$, respectively, while Collaborative Decoding in conjunction with the Visual Expert Pre-training Stage boosts accuracy by 4.0% and 1.7%. The combined use of these approaches results in improvements over previous combinations, achieving approximately a 10% increase in accuracy compared to the baseline method. These experiments demonstrate the effectiveness of our design. After employing the DPO stage, a comprehensive improvement in the quantitative results of model output explanations is observed, particularly with an increase of 0.75 points in human-scored ELO Ratings compared to the SFT model. This indicates that DPO is an effective post-training method for enhancing the quality of model output explanations to better align with human preferences. *Results of other ablation experiments can be found in the Appendix D.4.*

4.6. Qualitative Results

Fig. 4 showcases the detection and explanation results generated by AIGI-Holmes on various AI-generated images. It can be observed that our method produces precise explanations for defects across different AI-Generation modes. *More visual results can be found in the Appendix E.*

5. Conclusion

In this work, we introduce Holmes-Set, the first explanation-rich dataset with human-verifiable semantic annotations and contrastive preference pairs, addressing the critical data scarcity in AI-generated image detection. Besides, the proposed Holmes Pipeline systematically integrates visual expert pre-training, explanation-aware fine-tuning, and human-aligned preference optimization, synergizing model perception with human reasoning. Extensive experiments demonstrate our AIGI-Holmes' state-of-the-art detection accuracy and human-aligned interpretability. These contributions advance explainable and generalizable AIGI detection for rapidly evolving AIGC scenarios. We hope our dataset and methodology will inspire future research toward building more trustworthy AI-generated image detection systems.

Acknowledgements

This work was supported by the National Science Fund for Distinguished Young Scholars (No.62025603), the National Natural Science Foundation of China (No. U21B2037, No. U22B2051, No. U23A20383, No. U21A20472, No. 62176222, No. 62176223, No. 62176226, No. 62072386, No. 62072387, No. 62072389, No. 62002305 and No. 62272401), and the Natural Science Foundation of Fujian Province of China (No. 2021J06003, No.2022J06001).

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 2

[2] AI@Meta. Llama 3 model card. 2024. 3

[3] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005. 5

[4] Black-Forest-Labs. Flux. https://github.com/black-forest-labs/flux, 2024. 1, 7, 2

[5] You-Ming Chang, Chen Yeh, Wei-Chen Chiu, and Ning Yu. Antifakeprompt: Prompt-tuned vision-language models are fake image detectors. *arXiv preprint arXiv:2310.17419*, 2023. 1, 3, 7, 2, 4

[6] 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*. 3

[7] Baoying Chen, Jishen Zeng, Jianquan Yang, and Rui Yang. DRCT: Diffusion reconstruction contrastive training towards universal detection of diffusion generated images. In *Proceedings of the 41st International Conference on Machine Learning*, pages 7621–7639. PMLR, 2024. 2, 3, 4, 6

[8] Dongping Chen, Ruoxi Chen, Shilin Zhang, Yaochen Wang, Yinuo Liu, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark. In *Forty-first International Conference on Machine Learning*, 2024. 4, 5

[9] Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-σ : Weak-to-strong training of diffusion transformer for 4k text-to-image generation, 2024. 7, 2

[10] Jingye Chen, Yupan Huang, Tengchao Lv, Lei Cui, Qifeng Chen, and Furu Wei. Textdiffuser: Diffusion models as text painters. *Advances in Neural Information Processing Systems*, 36, 2024. 4

[11] Shen Chen, Taiping Yao, Hong Liu, Xiaoshuai Sun, Shouhong Ding, Rongrong Ji, et al. Diffusionfake: Enhancing generalization in deepfake detection via guided stable diffusion. *Advances in Neural Information Processing Systems*, 37:101474–101497, 2024. 1

[12] Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025. 7, 2

[13] Yize Chen, Zhiyuan Yan, Siwei Lyu, and Baoyuan Wu. X^2-dfd: A framework for eXplainable and eXtendable deepfake detection. *arXiv preprint arXiv:2410.06126*, 2024. 1, 5

[14] Zhongxi Chen, Ke Sun, Ziyin Zhou, Xianming Lin, Xiaoshuai Sun, Liujuan Cao, and Rongrong Ji. Diffusionface: Towards a comprehensive dataset for diffusion-based face forgery analysis. *arXiv preprint arXiv:2403.18471*, 2024. 1

[15] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024. 4

[16] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024. 4

[17] Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E Gonzalez, et al. Chatbot arena: An open platform for evaluating llms by human preference. In *Forty-first International Conference on Machine Learning*, 2024. 5, 4

[18] Davide Cozzolino, Giovanni Poggi, Riccardo Corvi, Matthias Nießner, and Luisa Verdoliva. Raising the bar of ai-generated image detection with clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4356–4366, 2024. 7

[19] 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. *arXiv preprint arXiv:2305.06500*, 2023. 2

[20] 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

[21] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 1

[22] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024. 1, 7, 2

[23] Guian Fang, Wenbiao Yan, Yuanfan Guo, Jianhua Han, Zutao Jiang, Hang Xu, Shengcai Liao, and Xiaodan Liang.

Humanrefiner: Benchmarking abnormal human generation and refining with coarse-to-fine pose-reversible guidance. In *European Conference on Computer Vision*, pages 201–217. Springer, 2024. 4, 1

[24] Guian Fang, Wenbiao Yan, Yuanfan Guo, Jianhua Han, Zutao Jiang, Hang Xu, Shengcai Liao, and Xiaodan Liang. Humanrefiner: Benchmarking abnormal human generation and refining with coarse-to-fine pose-reversible guidance. In *European Conference on Computer Vision*, pages 201–217. Springer, 2025. 2, 6

[25] Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. Leveraging frequency analysis for deep fake image recognition. In *International conference on machine learning*, pages 3247–3258. PMLR, 2020. 3, 2

[26] Jessica Fridrich and Jan Kodovsky. Rich models for steganalysis of digital images. *IEEE Transactions on information Forensics and Security*, 7(3):868–882, 2012. 3

[27] Xingyu Fu, Muyu He, Yujie Lu, William Yang Wang, and Dan Roth. Commonsense-t2i challenge: Can text-to-image generation models understand commonsense? *arXiv preprint arXiv:2406.07546*, 2024. 4, 1, 2

[28] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27, 2014. 1

[29] Xiao Guo, Xiufeng Song, Yue Zhang, Xiaohong Liu, and Xiaoming Liu. Rethinking vision-language model in face forensics: Multi-modal interpretable forged face detector. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 105–116, 2025. 1

[30] Jian Han, Jinlai Liu, Yi Jiang, Bin Yan, Yuqi Zhang, Zehuan Yuan, Bingyue Peng, and Xiaobing Liu. Infinity: Scaling bitwise autoregressive modeling for high-resolution image synthesis, 2024. 7, 2

[31] Yan Hong and Jianfu Zhang. Wildfake: A large-scale challenging dataset for ai-generated images detection. *arXiv preprint arXiv:2402.11843*, 2024. 3

[32] 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. 6

[33] Zhenglin Huang, Jinwei Hu, Xiangtai Li, Yiwei He, Xingyu Zhao, Bei Peng, Baoyuan Wu, Xiaowei Huang, and Guangliang Cheng. Sida: Social media image deepfake detection, localization and explanation with large multimodal model. *arXiv preprint arXiv:2412.04292*, 2024. 1

[34] 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. 1, 3, 4, 5

[35] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 3

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

[37] Albert Qiaochu Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L'elio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b. *ArXiv*, abs/2310.06825, 2023. 3

[38] Yan Ju, Shan Jia, Lipeng Ke, Hongfei Xue, Koki Nagano, and Siwei Lyu. Fusing global and local features for generalized ai-synthesized image detection. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 3465–3469. IEEE, 2022. 2

[39] Negar Kamali, Karyn Nakamura, Angelos Chatzimparmpas, Jessica Hullman, and Matthew Groh. How to distinguish ai-generated images from authentic photographs. *arXiv preprint arXiv:2406.08651*, 2024. 4, 5

[40] Negar Kamali, Karyn Nakamura, Aakriti Kumar, Angelos Chatzimparmpas, Jessica Hullman, and Matthew Groh. Characterizing photorealism and artifacts in diffusion model-generated images. *arXiv preprint arXiv:2502.11989*, 2025. 4, 5

[41] Hengrui Kang, Siwei Wen, Zichen Wen, Junyan Ye, Weijia Li, Peilin Feng, Baichuan Zhou, Bin Wang, Dahua Lin, Linfeng Zhang, et al. Legion: Learning to ground and explain for synthetic image detection. *arXiv preprint arXiv:2503.15264*, 2025. 1

[42] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017. 3

[43] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 1

[44] Mamadou Keita, Wassim Hamidouche, Hessen Bougueffa Eutamene, Abdenour Hadid, and Abdelmalik Taleb-Ahmed. Bi-lora: A vision-language approach for synthetic image detection. *arXiv preprint arXiv:2404.01959*, 2024. 1, 3

[45] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 3

[46] Christos Koutlis and Symeon Papadopoulos. Leveraging representations from intermediate encoder-blocks for synthetic image detection. In *European Conference on Computer Vision*, pages 394–411. Springer, 2024. 1, 7, 2, 4

[47] Rohit Kundu, Athula Balachandran, and Amit K Roy-Chowdhury. Truthlens: Explainable deepfake detection for face manipulated and fully synthetic data. *arXiv preprint arXiv:2503.15867*, 2025. 1

[48] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Pro-*

ceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles, 2023. 2

[49] Binh M Le and Simon S Woo. Quality-agnostic deepfake detection with intra-model collaborative learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22378–22389, 2023. 2

[50] Caoshuo Li, Zengmao Ding, Xiaobin Hu, Bang Li, Donghao Luo, AndyPian Wu, Chaoyang Wang, Chengjie Wang, Taisong Jin, Yunsheng Wu, et al. Oraclefusion: Assisting the decipherment of oracle bone script with structurally constrained semantic typography. *arXiv preprint arXiv:2506.21101*, 2025. 4

[51] Jiawei Li, Fanrui Zhang, Jiaying Zhu, Esther Sun, Qiang Zhang, and Zheng-Jun Zha. Forgerygpt: Multimodal large language model for explainable image forgery detection and localization. *arXiv preprint arXiv:2410.10238*, 2024. 1

[52] Ouxiang Li, Jiayin Cai, Yanbin Hao, Xiaolong Jiang, Yao Hu, and Fuli Feng. Improving synthetic image detection towards generalization: An image transformation perspective. *arXiv preprint arXiv:2408.06741*, 2024. 6, 2

[53] Yixuan Li, Xuelin Liu, Xiaoyang Wang, Shiqi Wang, and Weisi Lin. Fakebench: Uncover the achilles' heels of fake images with large multimodal models. *arXiv preprint arXiv:2404.13306*, 2024. 2, 3, 4

[54] Jingchun Lian, Lingyu Liu, Yaxiong Wang, Yujiao Wu, Li Zhu, and Zhedong Zheng. A large-scale interpretable multimodality benchmark for facial image forgery localization. *arXiv preprint arXiv:2412.19685*, 2024. 1

[55] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004. 5

[56] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 7

[57] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024. 5, 1, 2, 10, 15

[58] Bo Liu, Fan Yang, Xiuli Bi, Bin Xiao, Weisheng Li, and Xinbo Gao. Detecting generated images by real images. In *European Conference on Computer Vision*, pages 95–110. Springer, 2022. 2

[59] Huan Liu, Zichang Tan, Chuangchuang Tan, Yunchao Wei, Yao Zhao, and Jingdong Wang. Forgery-aware adaptive transformer for generalizable synthetic image detection. *arXiv preprint arXiv:2312.16649*, 2023. 3

[60] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 7

[61] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 3, 5, 1

[62] Huan Liu, Zichang Tan, Chuangchuang Tan, Yunchao Wei, Jingdong Wang, and Yao Zhao. Forgery-aware adaptive transformer for generalizable synthetic image detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[63] Yunpeng Luo, Junlong Du, Ke Yan, and Shouhong Ding. Lareˆ 2: Latent reconstruction error based method for diffusion-generated image detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17006–17015, 2024. 1, 7, 2

[64] Fanqing Meng, Wenqi Shao, Lixin Luo, Yahong Wang, Yiran Chen, Quanfeng Lu, Yue Yang, Tianshuo Yang, Kaipeng Zhang, Yu Qiao, et al. Phybench: A physical commonsense benchmark for evaluating text-to-image models. *arXiv preprint arXiv:2406.11802*, 2024. 4, 1, 2

[65] Mistral-AI. Pixtral large. https://mistral.ai/news/pixtral-large/, 2024. 4

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

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

[68] Siran Peng, Zipei Wang, Li Gao, Xiangyu Zhu, Tianshuo Zhang, Ajian Liu, Haoyuan Zhang, and Zhen Lei. Mllm-enhanced face forgery detection: A vision-language fusion solution. *arXiv preprint arXiv:2505.02013*, 2025. 1

[69] Chenfan Qu, Jian Liu, Haoxing Chen, Baihan Yu, Jingjing Liu, Weiqiang Wang, and Lianwen Jin. Explainable tampered text detection via multimodal large models. *arXiv preprint arXiv:2412.14816*, 2024. 1, 3, 4

[70] 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. 3

[71] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 2024. 2, 3

[72] Jonas Ricker, Simon Damm, Thorsten Holz, and Asja Fischer. Towards the detection of diffusion model deepfakes. *arXiv preprint arXiv:2210.14571*, 2022. 2

[73] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 1

[74] Ayush Sarkar, Hanlin Mai, Amitabh Mahapatra, Svetlana Lazebnik, D.A. Forsyth, and Anand Bhattad. Shadows don't lie and lines can't bend! generative models don't know projective geometry...for now. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 28140–28149, 2024. 4, 1

[75] Zeyang Sha, Zheng Li, Ning Yu, and Yang Zhang. De-fake: Detection and attribution of fake images generated by text-to-image diffusion models. *arXiv preprint arXiv:2210.06998*, 2022. 2

[76] Xiufeng Song, Xiao Guo, Jiache Zhang, Qirui Li, Lei Bai, Xiaoming Liu, Guangtao Zhai, and Xiaohong Liu. On learning multi-modal forgery representation for diffusion generated video detection. In *Proceeding of Thirty-eighth Conference on Neural Information Processing Systems*, Vancouver, Canada, 2024. 1

[77] 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 *Proceedings of the AAAI conference on artificial intelligence*, pages 2638–2646, 2021. 1

[78] Ke Sun, Hong Liu, Taiping Yao, Xiaoshuai Sun, Shen Chen, Shouhong Ding, and Rongrong Ji. An information theoretic approach for attention-driven face forgery detection. In *European conference on computer vision*, pages 111–127. Springer, 2022.

[79] Ke Sun, Taiping Yao, Shen Chen, Shouhong Ding, Jilin Li, and Rongrong Ji. Dual contrastive learning for general face forgery detection. In *Proceedings of the AAAI conference on artificial intelligence*, pages 2316–2324, 2022.

[80] Ke Sun, Shen Chen, Taiping Yao, Xiaoshuai Sun, Shouhong Ding, and Rongrong Ji. Continual face forgery detection via historical distribution preserving. *International Journal of Computer Vision*, 133(3):1067–1084, 2025. 1

[81] Ke Sun, Shen Chen, Taiping Yao, Ziyin Zhou, Jiayi Ji, Xiaoshuai Sun, Chia-Wen Lin, and Rongrong Ji. Towards general visual-linguistic face forgery detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19576–19586, 2025. 1

[82] Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525*, 2024. 7, 2

[83] Zhihao Sun, Haoran Jiang, Haoran Chen, Yixin Cao, Xipeng Qiu, Zuxuan Wu, and Yu-Gang Jiang. Forgerysleuth: Empowering multimodal large language models for image manipulation detection. *arXiv preprint arXiv:2411.19466*, 2024. 1, 3, 4, 5

[84] Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, and Yunchao Wei. Learning on gradients: Generalized artifacts representation for gan-generated images detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12105–12114, 2023. 2

[85] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28130–28139, 2024. 1, 3, 5, 6, 7, 2

[86] Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. 2024. 7, 2

[87] Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *arXiv preprint arXiv:2404.02905*, 2024. 1

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

[89] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 4

[90] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. Cnn-generated images are surprisingly easy to spot...for now. In *CVPR*, 2020. 2, 3, 4

[91] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. Cnn-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8695–8704, 2020. 1, 3, 7, 2

[92] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*, 2023. 2

[93] Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. Dire for diffusion-generated image detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22445–22455, 2023. 1, 3, 2

[94] Siwei Wen, Junyan Ye, Peilin Feng, Hengrui Kang, Zichen Wen, Yize Chen, Jiang Wu, Wenjun Wu, Conghui He, and Weijia Li. Spot the fake: Large multimodal model-based synthetic image detection with artifact explanation. *arXiv preprint arXiv:2503.14905*, 2025. 1

[95] Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, et al. Janus: Decoupling visual encoding for unified multimodal understanding and generation. *arXiv preprint arXiv:2410.13848*, 2024. 7, 2

[96] Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Chunyi Li, Wenxiu Sun, Qiong Yan, Guangtao Zhai, et al. Q-bench: A benchmark for general-purpose foundation models on low-level vision. *arXiv preprint arXiv:2309.14181*, 2023. 2, 5

[97] H. Wu, J. Zhou, and S. Zhang. Generalizable synthetic image detection via language-guided contrastive learning. *arXiv preprint:2305.13800*, 2023. 2

[98] Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*, 2024. 7, 2

[99] Zhipei Xu, Xuanyu Zhang, Runyi Li, Zecheng Tang, Qing Huang, and Jian Zhang. Fakeshield: Explainable image

forgery detection and localization via multi-modal large language models. *arXiv preprint arXiv:2410.02761*, 2024. 1, 3, 4, 5

[100] Shilin Yan, Ouxiang Li, Jiayin Cai, Yanbin Hao, Xiaolong Jiang, Yao Hu, and Weidi Xie. A sanity check for ai-generated image detection. *arXiv preprint arXiv:2406.19435*, 2024. 1, 3, 7, 2, 4

[101] Fan Yang, Ru Zhen, Jianing Wang, Yanhao Zhang, Haoxiang Chen, Haonan Lu, Sicheng Zhao, and Guiguang Ding. Heie: Mllm-based hierarchical explainable aigc image implausibility evaluator. *arXiv preprint arXiv:2411.17261*, 2024. 1

[102] Junyan Ye, Baichuan Zhou, Zilong Huang, Junan Zhang, Tianyi Bai, Hengrui Kang, Jun He, Honglin Lin, Zihao Wang, Tong Wu, et al. Loki: A comprehensive synthetic data detection benchmark using large multimodal models. *arXiv preprint arXiv:2410.09732*, 2024. 2, 3, 4

[103] Peipeng Yu, Jianwei Fei, Hui Gao, Xuan Feng, Zhihua Xia, and Chip Hong Chang. Unlocking the capabilities of large vision-language models for generalizable and explainable deepfake detection. *arXiv preprint arXiv:2503.14853*, 2025. 1

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

[105] Yuhui Zhang, Alyssa Unell, Xiaohan Wang, Dhruba Ghosh, Yuchang Su, Ludwig Schmidt, and Serena Yeung-Levy. Why are visually-grounded language models bad at image classification? *arXiv preprint arXiv:2405.18415*, 2024. 2, 5

[106] Yue Zhang, Ben Colman, Xiao Guo, Ali Shahriyari, and Gaurav Bharaj. Common sense reasoning for deepfake detection. In *European Conference on Computer Vision*, pages 399–415. Springer, 2025. 1, 3

[107] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623, 2023. 3

[108] Nan Zhong, Yiran Xu, Zhenxing Qian, and Xinpeng Zhang. Rich and poor texture contrast: A simple yet effective approach for ai-generated image detection. *arXiv preprint arXiv:2311.12397*, 2023. 7, 2

[109] Nan Zhong, Yiran Xu, Sheng Li, Zhenxing Qian, and Xinpeng Zhang. Patchcraft: Exploring texture patch for efficient ai-generated image detection. *arXiv preprint arXiv:2311.12397*, pages 1–18, 2024. 3

[110] 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, 2023. 2, 3, 4