ReAlign: Generalizable Image Forgery Detection via Reasoning-Aligned Representation

Supplementary Material

A. Robustness Study

With the rapid expansion of online platforms and social media, images are often encountered in a degraded form due to transmission artifacts, including JPEG compression and Gaussian-related distortions. Table 1 presents a comprehensive comparison of detection robustness under various image degradations. Across different levels of JPEG compression (QF = 95 / 90 / 75 / 50) and Gaussian blur ($\sigma = 1.0/2.0/3.0/4.0$), most baseline methods exhibit noticeable declines as the perturbations intensify. In contrast, ReAlign maintains consistently strong performance in every setting, showing only a slight reduction from its accuracy on clean images. This stability highlights the effectiveness of our alignment-based design, which enables the model to remain resilient even when structural and frequency cues are weakened by compression or blurring. Even under severe distortions, such as QF=50 JPEG compression or $\sigma = 4.0$ blur, ReAlign delivers robust predictions where other models degrade sharply. These results collectively demonstrate that ReAlign provides the most reliable and degradation-tolerant performance, making it particularly well-suited for deployment in practical environments where image quality can vary widely.

B. More Ablation Study

In this section, we first elaborate on several implementation details related to the ablation experiments. We then present additional ablation results on the AIGCDetectBenchmark.

B.1. Implementation Details

In Tab. 5, all variants perform full parameter fine-tuning for the detection head. The ablation of the fine-tuning method only refers to the training approach for the image encoder.

In cases (b) and (c) of Tab. 5, alignment and classification loss are optimized sequentially. First, we freeze the text encoder and align the outputs of the image encoder and text encoder using the alignment loss, based on image-text pairs. Then, we freeze the image encoder and add the detection head after it, training under the constraint of the classification loss.

B.2. Ablation Study on AIGCDetectBenchmark

To further investigate the importance of our proposed alignment text and training configuration in improving model performance, we conducted additional ablation studies on the AIGCDetectBenchmark [23] based on the experiments in Section 4.3. The results are presented in Tab. 2 and Tab. 3.

Tab. 2 shows the detection accuracy of different text input types across various generative models. Our ReAlign's setting, which uses reasoning text with a class label prefix, achieves the best overall performance with a mean accuracy of 96.14%, demonstrating strong cross-domain generalization. Replacing the reasoning text with image captions in case (a) lowers the accuracy to 92.12%, while case (c), which uses captions alone, drops further to 87.84%. Case (b), using only reasoning text, maintains a high accuracy of 92.11%, indicating that reasoning text alone is highly discriminative. In contrast, case (d), which uses only the class label, performs the worst with 79.95%, a 16.19% decrease compared to ours. These results highlight the essential role of reasoning text in effective alignment and generalizable forgery detection.

Tab. 3 compares the effects of different training strategies on forgery detection performance. The "Ours" setting, which adopts joint optimization and LoRA fine-tuning, achieves the highest mean accuracy of 96.14%, with consistently strong results across all generative models. In case (a), replacing LoRA with full fine-tuning reduces accuracy to 92.12%, proving that the LoRA strategy can effectively preserve the pre-trained semantic structure. Case (b), which uses sequential optimization with full fine-tuning, drops significantly to 62.86%, and case (c), which swaps in LoRA, performs similarly at 62.76%, showing that sequential optimization fails to adapt to the downstream task and learn meaningful forgery representations. Case (d) freezes the pretrained image encoder and trains only the classification head, achieving 75.38%, while case (e), which applies LoRA to the visual encoder, improves slightly to 81.65%. These results highlight the importance of reasoning-text alignment and joint optimization in achieving high detection accuracy and strong cross-domain generalization.

C. Details of Text-image Pairs Construction

In Sec. 3.4, we briefly introduced how we use the GRPO-optimized AIGI-R1 to construct text–image pairs. In this section, we provide a more detailed explanation of several aspects of that process.

Seed and Temperature. As shown in Fig. 4(b), when constructing the Text component using the trained AIGI-R1, we employ random seeds and temperature values to obtain richer and more diverse textual expressions. Specifically, we randomly select one of the following values as the seed:

Table 1. **Robustness on JPEG compression and Gaussian blur of ReAlign.**

Method	Original	JPEG Compression				Gaussian Blur			
		QF=95	QF=90	QF=75	QF=50	σ = 1.0	σ = 2.0	σ = 3.0	σ = 4.0
CNNSpot	70.78	64.03	62.26	60.65	59.66	68.39	67.26	67.13	65.85
FreDect	64.03	66.95	67.45	66.64	65.33	65.75	66.48	68.58	69.64
Fusing	68.38	62.43	61.39	59.34	57.41	68.09	66.69	66.02	65.58
LNP	83.84	53.58	54.09	53.02	52.85	67.91	66.42	66.20	62.69
LGrad	75.34	51.55	51.39	50.00	50.00	71.73	69.12	68.43	66.22
DIRE-G	68.68	66.49	66.12	65.28	64.34	64.00	63.09	62.21	61.91
UniFD	78.43	74.10	74.02	69.92	68.68	70.31	68.29	64.62	61.18
PatchCraft	89.31	72.48	71.41	69.43	67.78	75.99	74.90	73.53	72.28
AIDE	92.77	75.54	74.21	70.64	69.60	81.88	80.35	80.05	79.86
ReAlign	**96.14**	**82.89**	**81.52**	**74.83**	**70.28**	**84.85**	**82.04**	**81.16**	**80.51**

Table 2. **More Ablation Study across Different Text Types.**

Case	Ours	(a)	(b)	(c)	(d)
Class Label	✓	✓			✓
Image Caption		✓		✓	
Reasoning Text	✓		✓		
ProGAN	100.00	100.00	100.00	100.00	100.00
StyleGAN	97.93	99.67	99.67	97.80	91.16
BigGAN	90.73	88.50	85.07	87.92	80.95
CycleGAN	95.35	92.13	97.77	91.14	88.95
StarGAN	97.47	99.67	99.67	99.75	100.00
GauGAN	86.71	82.64	89.49	77.46	67.29
StyleGAN2	97.62	95.21	95.14	98.23	99.57
WFIR	84.45	77.15	93.20	82.95	60.40
ADM	97.80	95.01	89.17	91.57	74.23
Glide	97.76	92.90	82.86	80.52	65.22
Midjourney	96.63	69.59	90.20	53.51	65.46
SD v1.4	97.85	95.84	92.02	91.41	81.23
SD v1.5	97.79	95.89	93.66	91.37	80.81
VQDM	97.78	96.26	94.53	93.65	71.53
Wukong	97.83	94.83	81.67	91.53	79.34
DALLE2	97.80	91.35	89.70	75.05	62.20
Mean	96.14	92.12	92.11	87.84	79.95

Table 3. **More Ablation Study across Different Training Configuration.**

Case	Ours	(a)	(b)	(c)	(d)	(e)
Joint	✓	✓				
Sequential			✓	✓		
Full		✓	✓			
LoRA	✓			✓		✓
Freeze					✓	
ProGAN	100.00	100.00	88.87	90.10	99.98	100.00
StyleGAN	97.93	97.91	75.18	71.06	87.25	98.73
BigGAN	90.73	84.50	89.65	76.02	98.25	99.90
CycleGAN	95.35	90.73	80.73	74.84	93.34	99.28
StarGAN	97.47	100.00	63.63	70.51	97.57	99.95
GauGAN	86.71	77.91	71.55	74.85	98.71	99.35
StyleGAN2	97.62	98.97	62.43	56.19	74.82	97.21
WFIR	84.45	97.45	64.60	75.53	96.00	98.73
ADM	97.80	90.41	52.28	59.24	89.60	89.48
Glide	97.76	84.89	50.05	52.20	74.38	66.03
Midjourney	96.63	89.06	50.00	51.29	50.98	51.62
SD v1.4	97.85	93.22	49.92	50.32	54.60	67.67
SD v1.5	97.79	93.04	49.91	50.50	54.11	67.95
VQDM	97.78	94.51	50.26	52.04	92.00	94.02
Wukong	97.83	93.08	49.94	50.52	61.59	69.90
DALLE2	97.80	88.30	49.90	48.90	51.50	53.65
Mean	96.14	92.12	62.86	62.76	75.38	81.65

32, 42, 52, 62, or 72, and randomly choose a number from 0.6, 0.7, 0.8 as the temperature.

Expert Proofreading and Refining. After obtaining the reasoning texts, we invited a validation team consisting of 10 human experts to proofread and polish the content to ensure its quality. The specific requirements include: (1) Removing samples where AIGI-R1 made incorrect judgments; (2) Removing clearly low-quality reasoning samples (e.g., refusal to answer, reasoning content contradicting the final answer); (3) Modifying or deleting unreasonable or factually incorrect explanations.

Text-image Pairs Examples. In Fig. 3 and Fig. 4, we show some examples from the text-image pairs dataset we built. AIGI-R1 can efficiently use concise textual content to clearly describe the forged details of the image, which is an effective representation.

D. Details of UltraSynth-10k Construction

In this section, we first introduce the construction details of our UltraSynth-10k, and then provide a detailed description of the five advanced image generation methods used. Finally, we show some examples of generated images.

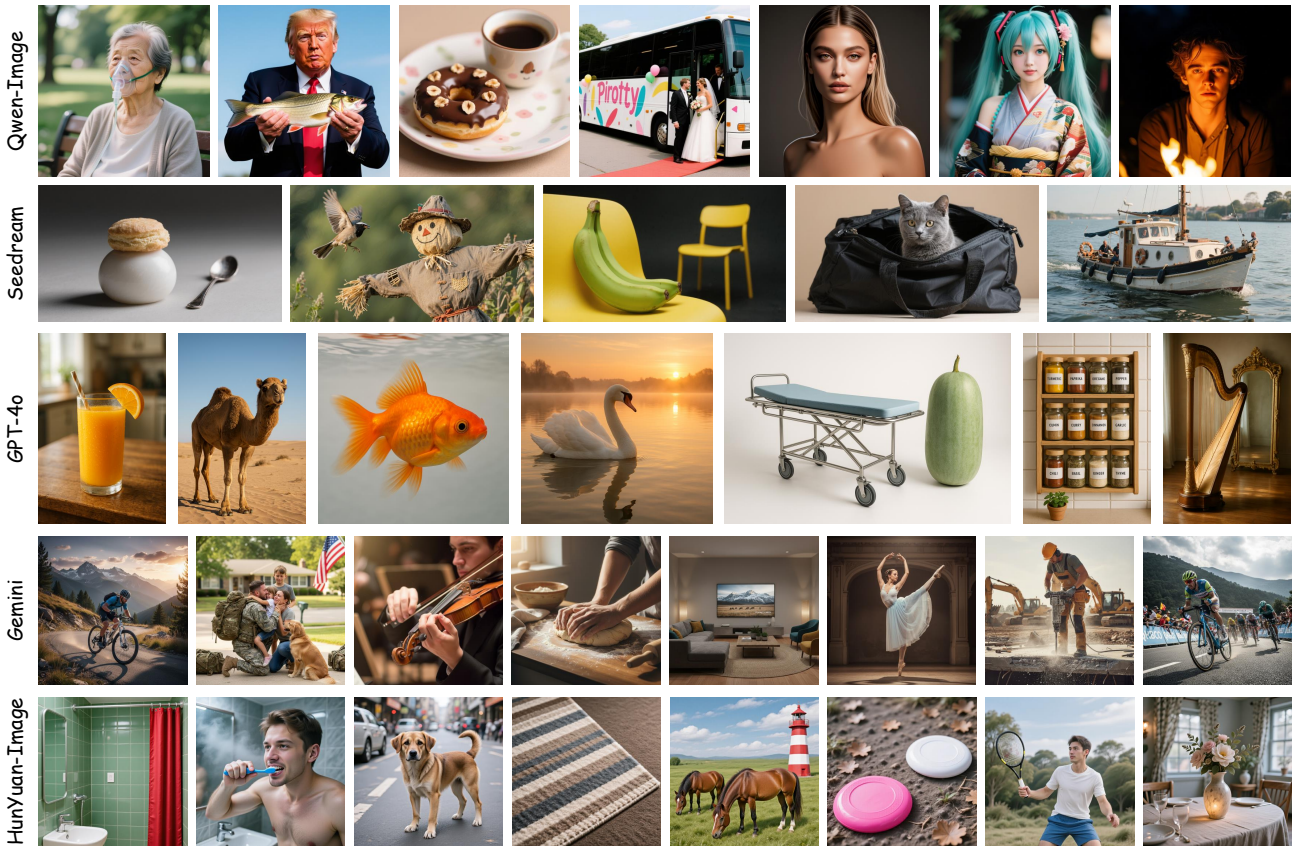


Figure 1. **More Image Examples of UltraSynth-10k**

D.1. Constructing Details

First, we collected a set of high-quality synthetic images that meet our requirements from the internet and publicly available image generation datasets [22]. To further expand the scale of our dataset, we additionally generated fake images using the APIs of these generation methods, based on the image generation prompts provided by T2I-CompBench [7]. In the end, we obtained a dataset consisting of fake images generated by five advanced forgery methods and real images, with a 1:1 ratio between the real and fake images.

D.2. Generation Methods

In this section, we provide a detailed introduction to the five generation methods included in UltraSynth-10k: Qwen-Image [20], Seedream [5], GPT-4o [2], Gemini [3], and HunYuan-Image [16].

Qwen-Image: Qwen-Image [20] is a text-to-image (T2I) model developed by Alibaba, distinguished by its strong ability to generate images with accurate and well-structured embedded text. Powered by the MultiModal Diffusion Transformer (MMDiT) architecture and multimodal spatial encoding, it offers precise control over layout and typography while maintaining high visual fidelity across diverse styles. Its reliable text rendering and multilingual support make it particularly effective for posters, advertising, and other design-oriented generation tasks.

Seedream: Seedream [5] is ByteDance's high-quality T2I model series, with Seedream 3.0 delivering native 2K-resolution generation, improved text rendering, and strong bilingual prompt understanding. Its enhanced training pipeline and efficient sampling enable sharp details, accurate layout control, and fast inference, making it a powerful model for high-resolution creative and commercial visual production.

GPT-4o: GPT-4o [2] is OpenAI's multimodal model that significantly enhances image generation through stronger vision-language reasoning. It produces coherent, detailed, and stylistically consistent images while following fine-grained prompts more accurately than prior GPT models, making it effective for creative illustration, design assistance, and visually grounded content creation.

Gemini: Gemini 2.5 [3] is Google DeepMind's advanced multimodal system with notable improvements in image synthesis quality. It generates detailed and well-structured visuals with strong semantic alignment, benefiting from improved multimodal integration and an efficient expert hybrid architecture, enabling robust performance in complex creative and design tasks.

HunYuan-Image: Hunyuan-Image 3.0 [16] is Tencent's

unified autoregressive T2I model that excels at producing semantically aligned, high-fidelity images from complex prompts. With strong bilingual support, refined visual detail, and reliable multi-object composition, it performs well across photorealistic and artistic styles, making it suitable for high-quality creative content generation.

D.3. More Image Examples of UltraSynth-10k

In Fig. 1, we present additional sample images from UltraSynth-10k generated by advanced methods. Unlike images produced by traditional approaches such as GANs, these results are exceptionally realistic and visually impeccable, with virtually no perceptible artifacts. They preserve fine-grained details remarkably well and exhibit very high resolution.

E. Comparison Method

To thoroughly assess AI-generated image detection, we select a broad range of representative baseline methods, spanning frequency domain analyses, gradient-driven detectors, semantic level approaches, reconstruction-based techniques, and others. In this section, we provide detailed descriptions of these comparative methods.

CNNSpot [18]. CNNSpot uses ResNet-50 [6] and shows that simple augmentations like JPEG compression and Gaussian blur greatly improve generalization, enabling the detector to handle unseen generative models and datasets.

FreDect [4]. FreDect finds that GAN images contain distinctive artifacts in the frequency domain. By detecting these abnormal frequency patterns, the method effectively identifies fake images.

Fusing [8]. Fusing adopts a dual branch model: one branch extracts global features, while the other selects key local patches. A multi-head attention module fuses these features, enabling robust cross-model fake detection.

LNP [10]. LNP observes that real images share consistent noise patterns, whereas generated images do not. By extracting noise via a denoising model and analyzing its frequency characteristics, LNP distinguishes real from fake images.

LGrad [14]. LGrad uses the gradient map of a pretrained classifier as a fingerprint of GAN images. This gradient-based representation allows a simple binary classifier to separate real and generated images.

UniFD [12]. UniFD relies on features from a large pretrained model such as CLIP. The smooth and robust feature space ensures strong generalization across diverse generative models.

DIRE [19]. DIRE detects diffusion-generated images by comparing the input with its reconstruction from a pretrained diffusion model. Fake images reconstruct well, while real images do not, making reconstruction error a useful fingerprint.

PatchCraft [23]. PatchCraft contrasts pixel correlations between rich texture and low texture regions. By reconstructing images from these two types of patches and comparing their correlation differences, it captures a universal signature of generated images.

NPR [15]. NPR shows that CNN-based upsampling introduces distinctive artifacts beyond frequency domain patterns. By measuring neighbor pixel relationships that capture these upsampling traces, NPR provides a simple yet highly generalizable feature for detecting AI-generated images.

AntifakePrompt [1]. AntifakePrompt treats fake image detection as a visual question-answering task using a prompt-tuned vision language model, thereby leveraging strong zero-shot generalization from the VLM to distinguish real from generated images with minimal additional parameters.

LaRE [11]. LaRE builds on DIRE by using latent space reconstruction loss, comparing an image with its latent space reconstruction to achieve more efficient and accurate detection of diffusion-generated images.

RINE [9]. RINE extracts representations from intermediate Transformer blocks of a frozen CLIP image encoder and uses a lightweight network with trainable block-importance weights to map them into a forgery perceptual space, achieving strong generalization for synthetic image detection.

AIDE [21]. AIDE proposes a hybrid model that jointly models low-level and high-level features in order to capture both semantic and frequency differences between real and generated images.

AIGI-Holmes [24]. AIGI-Holmes integrates CLIP and NPR visual expert features into an LLaVA-based multimodal pipeline, using LoRA fine-tuning together with supervised and preference-based optimization to detect AI-generated images while producing explanations that humans can easily verify.

F. Limitation

Although ReAlign shows strong cross-domain generalization and detection performance, it still has several limitations. Firstly, although ReAlign leverages high-quality reasoning text for vision-language alignment during training, it is unable to generate textual explanations during inference, which weakens the model's interpretability. In the future, a lightweight explanation module could be introduced.

Secondly, we only used the classic CLIP-ViT-/14-336 [13], but it has some design flaws, such as not supporting arbitrary resolutions. The input images need to be cropped, which may result in a loss of image details and potentially affect performance. Therefore, in the future, we could explore more advanced CLIP models and their variants, such as SigLIP [17].

You are a professional AIGC image forensics analyst, with deep expertise in forgery detection, semantic coherence analysis, and pixel-level artifact identification. You are trained to detect inconsistencies commonly found in AI-generated or digitally manipulated images.Your task is to critically examine the input image and provide a comprehensive, structured forensic analysis, focusing on anomalies across two core dimensions:

1. Semantic-Level Inconsistencies

Evaluate the image's logical coherence and contextual realism. Look for signs that suggest semantic incongruity or improbable visual narratives, including but not limited to:

- Logical Contradictions
  e.g., inconsistent shadows/light sources, physically implausible reflections, weather or time-of-day mismatches.
- Improbable Object Relationships
  e.g., unrealistic size ratios between objects, impossible overlaps, distorted spatial arrangements.
- Violation of Physical Laws or Human Anatomy
  e.g., gravity-defying elements, floating objects, unnatural body proportions or impossible poses.
- Incoherent Visual Storytelling
  e.g., characters, actions, or props that conflict with the environment or narrative context.

2. Pixel-Level & Visual Artifact Analysis

Analyze the image at a low-level forensic perspective. Examine for signs of manipulation or synthetic generation by inspecting:

- Texture Inconsistencies
  e.g., inconsistent material rendering (e.g., hair, fabric), smeared edges, or irregular texture tiling.
- Color and Lighting Disparities
  e.g., different color temperatures across regions, inconsistent shadow orientation, unnatural edge glows.
- Compression and Noise Artifacts
  e.g., uneven JPEG compression, patchy noise patterns, tell-tale signs of localized generation or splicing.
- Edge and Boundary Abnormalities
  e.g., halos, double edges, blending failures between foreground and background.

Output Requirements

Write your analysis in a precise, objective, and professional tone, formatted in the following sections:

- Semantic-Level Analysis
  Detail any semantic inconsistencies and explain why they are implausible or suggest synthetic origin.
- Pixel-Level / Artifact Analysis
  Describe specific visual anomalies that may indicate image synthesis or manipulation.
- Suspicious Regions (if any)
  Highlight specific areas in natural language (e.g., "the subject’s left hand", "upper-right background") and associate each with the corresponding issue.

Figure 2. **The system prompt of AIGI-R1.**

This is a fake image. The numbers on the alarm clock face are distorted and deviate significantly from real-world typography and layout. This unnatural rendering of digits, such as irregular spacing, warped forms, and inconsistent alignment, strongly suggests a synthetic origin rather than a photograph of a real object.

This is a fake image. The image appears to have a high level of detail and uniformity that is often characteristic of AI-generated images. The colors are vibrant and the objects are perfectly aligned, which can be indicative of digital rendering rather than a photograph taken with a camera. Additionally, the lighting and shadows seem too consistent and idealized for a real photograph.

This is a fake image. The image is likely AI-generated. Semantically, the man's hat mimics Saturn's rings, an unrealistic and impractical object to wear, yet it blends unnaturally well into the scene. Texture-wise, the skin, shirt, and background lack real-world detail and variation, showing smoothness and stylized lighting typical of synthetic images.

This is a fake image. The image appears to have a high level of detail and uniformity in the arrangement of the vegetables, which is often characteristic of AI-generated images. The lighting and shadows are consistent across the entire image, and the colors are vibrant and slightly exaggerated, which can be a sign of digital manipulation. Additionally, the edges of some vegetables seem slightly too sharp and uniform, which is less common in real photographs.

This is a fake image. The image shows a dandelion with a high level of detail and a vibrant yellow center, which is typical of a real photograph. However, the edges of the dandelion's seeds appear slightly too uniform and the overall lighting and color saturation seem a bit too perfect, which can sometimes be indicative of AI generation. The background is blurred in a way that is common in professional photography, but the sharpness of the dandelion itself might suggest digital manipulation.

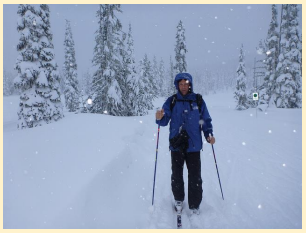
Figure 3. **Examples of AIGI-R1 fake sample.**

This is a real image. The image shows a person in a casual setting, wearing a pink shirt and a red knitted hat, holding a large purple object near their mouth. The lighting and shadows appear natural, and the details in the background, such as the door and wall, look realistic. There are no obvious signs of digital manipulation or distortion that would suggest the image was created by AI. The overall composition and quality align with what one might expect from a photograph taken with a regular camera.

This is a real image. The image shows a hot dog in a bun on a plate, accompanied by a glass of soda. The lighting, shadows, and details in the food and glass suggest a realistic photograph rather than an AI-generated image. There are no visible artifacts or anomalies that would indicate digital manipulation.

This is a real image. The image shows a person in a sports uniform actively engaged in a game, with a ball in motion. The details such as the texture of the skin, the fabric of the jersey, and the natural lighting suggest a real-life scenario rather than an AI-generated image. The depth of field effect, where the background is blurred while the subject is in focus, is a common technique used in photography to emphasize the subject.

This is a real image. The image shows a person skiing in a snowy landscape with trees covered in snow. The details such as the texture of the snow, the person's clothing, and the environment suggest a realistic scene. There are no visible artifacts or anomalies that would typically indicate an AI-generated image. The lighting and shadows also appear natural, supporting the idea that this is a real photograph.



This is a real image. The image shows a group of elephants in a river, with people observing them from a nearby area. The details such as the texture of the skin on the elephants, the water, and the people's clothing appear natural and not overly stylized or unrealistic, which suggests it could be a real photograph rather than an AI-generated image.

Figure 4. **Examples of AIGI-R1 real sample.**

References

[1] 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. 4

[2] Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793*, 2023. 3

[3] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 3

[4] 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. 4

[5] Yu Gao, Lixue Gong, Qiushan Guo, Xiaoxia Hou, Zhichao Lai, Fanshi Li, Liang Li, Xiaochen Lian, Chao Liao, Liyang Liu, et al. Seedream 3.0 technical report. *arXiv preprint arXiv:2504.11346*, 2025. 3

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 4

[7] Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in Neural Information Processing Systems*, 36:78723–78747, 2023. 3

[8] 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. 4

[9] 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. 4

[10] 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. 4

[11] 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. 4

[12] Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In *CVPR*, 2023. 4

[13] 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. 4

[14] 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. 4

[15] 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. 4

[16] Tencent Hunyuan Team. Hunyuanimage 2.1: An efficient diffusion model for high-resolution (2k) text-to-image generation. https://github.com/Tencent-Hunyuan/HunyuanImage-2.1, 2025. 3

[17] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025. 4

[18] 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. 4

[19] 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. 4

[20] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025. 3

[21] 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. 4

[22] Junyan Ye, Dongzhi Jiang, Zihao Wang, Leqi Zhu, Zhenghao Hu, Zilong Huang, Jun He, Zhiyuan Yan, Jinghua Yu, Hongsheng Li, et al. Echo-4o: Harnessing the power of gpt-4o synthetic images for improved image generation. *arXiv preprint arXiv:2508.09987*, 2025. 3

[23] 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*, 2023. 1, 4

[24] Ziyin Zhou, Yunpeng Luo, Yuanchen Wu, Ke Sun, Jiayi Ji, Ke Yan, Shouhong Ding, Xiaoshuai Sun, Yunsheng Wu, and Rongrong Ji. Aigi-holmes: Towards explainable and generalizable ai-generated image detection via multimodal large language models. *arXiv preprint arXiv:2507.02664*, 2025. 4