RESEARCH-ARTICLE

DFBench: Benchmarking Deepfake Image Detection Capability of Large Multimodal Models

JIARUI WANG, Shanghai Jiao Tong University, Shanghai, China

HUIYU DUAN, Shanghai Jiao Tong University, Shanghai, China

JUNTONG WANG, Shanghai Jiao Tong University, Shanghai, China

ZIHENG JIA, Shanghai Jiao Tong University, Shanghai, China

WOOYI YANG, Shanghai Jiao Tong University, Shanghai, China

XIAORONG ZHU, Shanghai Jiao Tong University, Shanghai, China

View all

Published: 27 October 2025

Citation in BibTeX format

MM '25: The 33rd ACM International Conference on Multimedia
October 27 - 31, 2025
Dublin, Ireland

Conference Sponsors:
SIGMM

MM '25: Proceedings of the 33rd ACM International Conference on Multimedia (October 2025)
https://doi.org/10.1145/3746027.3758204
ISBN: 9798400720352

DFBench: Benchmarking Deepfake Image Detection Capability of Large Multimodal Models

Jiarui Wang
wangjiarui@sjtu.edu.cn
Shanghai Jiao Tong University
Shanghai, China

Huiyu Duan
huiyuduan@sjtu.edu.cn
Shanghai Jiao Tong University
Shanghai, China

Juntong Wang
wang13029187978@sjtu.edu.cn
Shanghai Jiao Tong University
Shanghai, China

Ziheng Jia
jzhws1@sjtu.edu.cn
Shanghai Jiao Tong University
Shanghai, China

Woo Yi Yang
wooyiyang@sjtu.edu.cn
Shanghai Jiao Tong University
Shanghai, China

Xiaorong Zhu
zhuxiaorong@sjtu.edu.cn
Shanghai Jiao Tong University
Shanghai, China

Yu Zhao
yzhao3@sjtu.edu.cn
Shanghai Jiao Tong University
Shanghai, China

Jiaying Qian
2022qjy@sjtu.edu.cn
Shanghai Jiao Tong University
Shanghai, China

Yuke Xing
xingyuke-v@sjtu.edu.cn
Shanghai Jiao Tong University
Shanghai, China

Guangtao Zhai
zhaiguangtao@sjtu.edu.cn
Shanghai Jiao Tong University
Shanghai, China

Xiongkuo Min*
minxiongkuo@sjtu.edu.cn
Shanghai Jiao Tong University
Shanghai, China

Abstract

With the rapid advancement of generative models, the realism of AI-generated images has significantly improved, posing critical challenges for verifying digital content authenticity. Current deepfake detection methods often depend on datasets with limited generation models and content diversity that fail to keep pace with the evolving complexity and increasing realism of the AI-generated content. Large multimodal models (LMMs), widely adopted in various vision tasks, have demonstrated strong zero-shot capabilities, yet their potential in deepfake detection remains largely unexplored. To bridge this gap, we present **DFBench**, a large-scale DeepFake **Bench**mark featuring **(i) broad diversity**, including 540,000 images across real, AI-edited, and AI-generated content, **(ii) latest model**, the fake images are generated by 12 state-of-the-art generation models, and **(iii) bidirectional benchmarking and evaluating** for both the detection accuracy of deepfake detectors and the evasion capability of generative models. Based on DFBench, we propose **MoA-DF**, Mixture of Agents for DeepFake detection, leveraging a combined probability strategy from multiple LMMs. MoA-DF achieves state-of-the-art performance, further proving the effectiveness of leveraging LMMs for deepfake detection. Database and codes are publicly available at https://github.com/IntMeGroup/DFBench.

CCS Concepts

• **Information systems → Multimedia databases**; *Multimedia streaming*; Multimedia content creation.

Keywords

Deepfake image detection dataset, Large multimodal models (LMM), Mixture of Agents (MoA), AI-generated images

ACM Reference Format:
Jiarui Wang, Huiyu Duan, Juntong Wang, Ziheng Jia, Woo Yi Yang, Xiaorong Zhu, Yu Zhao, Jiaying Qian, Yuke Xing, Guangtao Zhai, and Xiongkuo Min*. 2025. DFBench: Benchmarking Deepfake Image Detection Capability of Large Multimodal Models. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25), October 27–31, 2025, Dublin, Ireland.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3746027.3758204

1 Introduction

The rapid advancement of generative models [2, 8, 14, 28, 31, 49, 57, 63] has significantly improved the ability to generate highly realistic images. However, these advancements raise serious concerns of the generated images regarding misinformation, social manipulation, and erosion of public trust. These concerns have driven the development of deepfake detection models [4, 7, 15, 16, 26, 36, 40, 47]. These models are typically trained on datasets containing real and fake images [5, 17, 44, 46, 60], with the goal of distinguishing between real and fake content. Thus, the generalization ability of these deepfake image detection models remains questionable.

Existing deepfake detection datasets and benchmarks [11, 17, 46, 54, 56] exhibit several critical limitations: **(1) Limited generative models**: most datasets [5, 17, 44, 46, 60] rely on a small number of generative methods. Moreover, many of the generative models [6, 13, 19, 27, 39] used in earlier datasets are now outdated, often

MM '25, Dublin, Ireland

ACM ISBN 979-8-4007-2035-2/2025/10
https://doi.org/10.1145/3746027.3758204

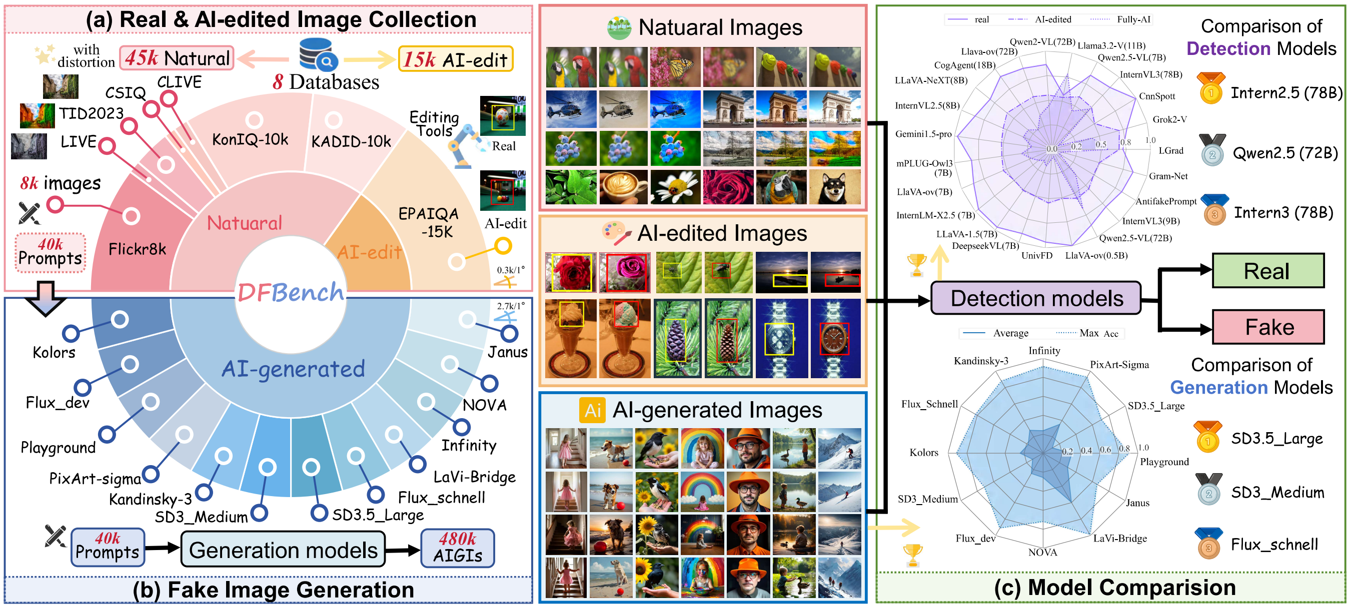

Figure 1: We present the DFBench, a large dataset for benchmarking deepfake image detection capabilities. (a) 45K real and 15K AI-edited images are collected from 8 sources. (b) 480K fake images are generated using 12 state-of-the-art generation models based on 40K prompts from Flickr8k. (c) The database enables evaluation for both detection models and generation models.

Table 1: An overview of fake image detection datasets.

Dataset	Image Content	AI Generation Category: Fully AI Generation	AI Generation Category: Partial AI Editing	Public Avalibility	Database Real Sources	AI Models	Fake Images	Total Images
UADFV [60]	Face	✓	✗	✗	Real face	1	252	493
FakeSpotter [54]	Face	✓	✓	✗	CelebA, FFHQ	8	5,000	11,000
DFFD [11]	Face	✓	✗	✓	CelebA, FFHQ, FaceForensics++	4	240,336	299,039
DeepFakeFace [46]	Face	✓	✗	✓	IMDB-WIKI	3	90,000	120,000
APFDD [17]	Face	✓	✗	✗	CelebA	1	5,000	10,000
DeepArt [56]	Art	✓	✗	✓	LAION-5B	5	73,411	137,890
IEEE VIP Cup [50]	General	✓	✗	✗	FFHQ, Imagenet, COCO, LSUN	5	7,000	14,000
DE-FAKE [44]	General	✓	✗	✗	MSCOCO, Flickr30k	2	60,000	80,000
CiFAKE [5]	General	✓	✗	✓	CIFAR	1	60,000	120,000
SID-Set [25]	General	✓	✓	✓	COCO, Flickr30k, MagicBrush	1	200,000	300,000
DFBench (Ours)	General	✓	✓	✓	**8 Datasets**	**12**	**495,000**	**540,000**

producing images with visible distortions, unnatural textures, or structural inconsistencies [51, 52, 52, 59]. **(2) Limited content diversity**: existing datasets for deepfake detection focus mainly on facial imagery [11, 17, 46, 54, 60], overlooking the growing threat of non-facial manipulations. In addition, most datasets include either fully real or fully fake images [44, 50, 56], lacking examples of partially AI-edited content where only specific regions are manipulated [1, 10, 43]. Furthermore, the real images in current datasets are often clean and undistorted [34, 41], making detection easier. **(3) Limited evaluation scope**: Large multimodal models (LMMs) have demonstrated strong zero-shot capabilities in vision tasks, yet their potential in deepfake detection remains largely unexplored.

DFBench is specifically designed to overcome the key limitations of existing datasets. (1) To improve generative diversity, fake images are generated using **12 state-of-the-art models**, covering a wide range of generation contents. (2) DFBench enhances content diversity by including **partially manipulated images** where only specific regions are edited and real images with natural distortions (*e.g.*, blur, compression) to better reflect real-world scenarios. (3) DFBench adopts a **bidirectional evaluation protocol** that assesses both the detection ability of conventional detectors and LMMs, and the evasion ability of generative models in fooling these detectors. As shown in Figure 1, DFBench includes highly deceptive examples that challenge deepfake detection models. Table 1 further highlights its advantages in scale, diversity, and evaluation design compared to existing benchmarks. Based on DFBench, we propose **MoA-DF**, **M**ixture **of** **A**gents for **D**eep**F**ake detection, leveraging a combined probability strategy from multiple LMMs and achieves state- of-the-art performance, proving the effectiveness of LMMs in deepfake detection tasks. In summary, our main contributions are:

- We introduce **DFBench**, a large-scale and diverse benchmark, featuring the **largest scale** of fake images generated by 12 state-of-the-art generative models, and **rich content** including AI-edited images and real-world image distortions (*e.g.*, blur, noise, compression, color distortions).
- We present a **bidirectional evaluation protocol** that benchmarks both the ***detection accuracy*** of deepfake detectors and the ***evasion capability*** of generative models.
- We propose **MoA-DF**, a novel mixture of agents method that combines the probabilistic outputs of LMMs to achieve more robust and accurate deepfake detection.

Figure 2: Visualization of images on the DFBench dataset.

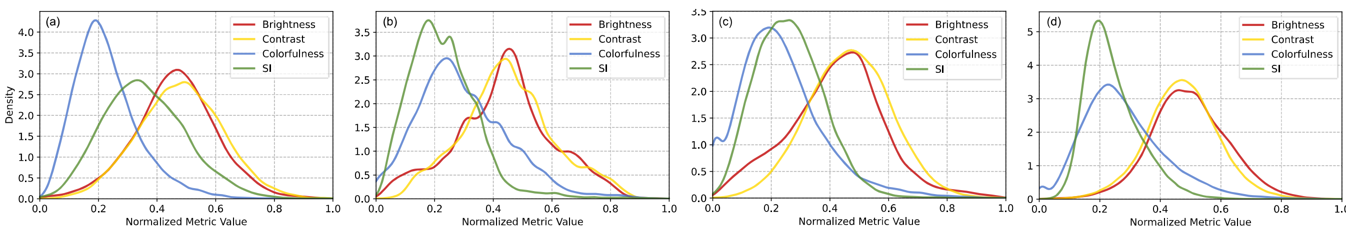

Figure 3: Feature distribution of the DFBench. (a) Feature distribution of real images with no distortion. (b) Feature distribution of real images with distortions. (c) Feature distribution of AI-edited images. (d) Feature distribution of AI-generated images.

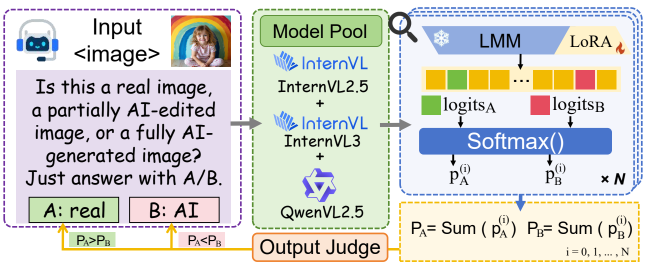

Figure 4: Overview of the MoA-DF architecture. Three LMMs are chosen as core detectors. Each model independently produces log-probabilities of the input image corresponding to A (real) or B (fake). The final decision is made based on the aggregation of these probabilities across all models.

2 RELATED WORK

A variety of datasets have been developed to advance deepfake image detection. Early datasets like UADFV [60], FakeSpotter [54], and DFFD [11] focused mainly on facial forgeries, but are limited in both scale and diversity. Later works such as DeepFakeFace [46] and APFDD [17] remained face-centric, while DeepArt [56] and DE-FAKE [44] explored artistic or caption-driven generation. Datasets like CiFAKE [5] and IEEE VIP Cup [50] attempt broader coverage but often rely on low-resolution images or limited models. SID-Set [25] introduces partial edits but uses only a single generation model and is still limited in scale. Our dataset DFBench stands out by its largest scale and broad diversity, including real, AI-edited, and fully generated images constructed from 8 sources.

3 Database Construction

3.1 Image Collection and Generation

To ensure content diversity and realism, DFBench incorporates real images from seven well-known public natural image datasets, including LIVE [45], CSIQ [29], TID2013 [42], KADID-10k [33], CLIVE [18], KonIQ-10k [23], and Flickr8k [21]. Except for Flickr8k, all other datasets include images affected by various forms of degradation that simulate real-world image quality impairments, including compression artifacts, blur, noise, color distortions, and *etc.* The AI-edited images are from EPAIQA-15K [43]. To construct a diverse and challenging set of fake images, we utilize 12 state-of-the-art open-source image generation models, including 10 diffusion-based models: PixArt-sigma [8], Playground [31], Kolors [49], SD3.5-Large [14], SD3-Medium [14], LaVi-Bridge [63], Kandinsky-3 [2], Flux-schnell [28], Flux-dev [28], Janus [57], and two AR-based models: NOVA [12] and Infinity [20]. To maintain fairness, all generative models are employed using their official default weights without further adaptation or tuning. Using 40K prompts from Flickr8k [21], we generated a total of 480K images (12 models × 40,000 images). Each of the 12 models is provided with the same set of prompts from real-world image captions, as shown in Figure 2. Notably, models such as SD3.5-Large [14] and Flux-dev [28] are capable of producing highly detailed outputs that even surpass the real source images, posing substantial challenges to deepfake detection models.

3.2 Database Analysis

As illustrated in Figure 3, we analyze the feature distribution of real distortion-free, real distorted, AI-edited, and AI-generated images in the DFBench across four image quality-related features, including colorfulness, brightness, contrast, and spatial information (SI). It can be observed that distorted real images generally exhibit lower colorfulness and higher SI values compared to distortion-free real images, likely due to the presence of noise, blur, or compression artifacts. AI-generated images exhibit the highest SI, reflecting their rich spatial detail. AI-edited images exhibit feature values between real and synthetic content, due to their mixed authentic and manipulated content. The broad range of feature distributions establishes DFBench as a comprehensive benchmark for evaluating deepfake detection under realistic and challenging conditions.

Table 2: Performance benchmark on real image subsets. ♡Conventional deepfake detection models, ★open-source and △close-source LMMs. ♦* refers to finetuned models. Best and second-best zero-shot results. Best and second-best finetuned results.

Methods / Datasets	LIVE [45]	CSIQ [29]	TID2013 [42]	KADID [33]	CLIVE [18]	KonIQ-10k [23]	Flickr8k [21]	Overall
♡CnnSpott [16]	**99.80%**	96.89%	**99.37%**	**99.83%**	99.74%	**99.72%**	99.65%	**99.29%**
♡AntifakePrompt [7]	81.36%	70.44%	93.23%	93.02%	68.52%	81.82%	89.17%	82.51%
♡Gram-Net [36]	86.97%	75.78%	93.30%	83.40%	86.83%	85.21%	84.29%	85.11%
♡UnivFD [40]	91.24%	86.22%	92.43%	91.79%	99.66%	98.38%	99.83%	94.22%
♡LGrad [47]	82.89%	54.22%	99.53%	70.14%	72.88%	77.14%	45.83%	71.81%
★Llava-one-vision (0.5B) [30]	**99.80%**	**98.44%**	**100.0%**	**98.23%**	**100.0%**	**99.88%**	99.96%	**99.47%**
★DeepSeekVL (7B) [37]	88.90%	87.11%	79.43%	83.95%	96.82%	97.83%	99.68%	90.53%
★LLaVA-1.5 (7B) [35]	93.38%	96.33%	91.30%	95.72%	**100.0%**	99.28%	**100.0%**	96.57%
★Llava-one-vision (7B) [30]	79.53%	76.33%	74.43%	77.47%	99.40%	98.52%	99.91%	86.51%
★mPLUG-Owl3 (7B) [61]	75.76%	76.44%	65.41%	65.32%	97.59%	96.17%	99.73%	82.35%
★Qwen2.5-VL (7B) [3]	77.70%	76.44%	72.00%	74.89%	96.39%	96.20%	99.23%	84.69%
★CogAgent (18B) [22]	81.98%	91.11%	77.83%	84.41%	99.40%	98.67%	99.86%	90.47%
★InternVL2.5 (8B) [9]	78.21%	80.67%	75.67%	70.96%	95.78%	94.14%	99.69%	85.02%
★InternVL3 (9B) [55]	72.40%	73.89%	63.83%	60.77%	94.23%	92.80%	99.62%	79.65%
★InternLM-XComposer2.5 (7B) [62]	90.43%	93.89%	89.60%	92.81%	99.83%	99.33%	100.0%	96.05%
★LLaVA-NeXT (8B) [32]	74.85%	73.22%	68.13%	75.81%	90.71%	84.94%	98.16%	80.83%
★Llama3.2-Vision (11B) [38]	53.56%	49.78%	50.52%	45.65%	66.67%	59.35%	68.39%	56.27%
★Qwen2-VL (72B) [53]	77.60%	76.80%	73.20%	89.60%	95.40%	48.43%	97.20%	86.40%
★Qwen2.5-VL (72B) [3]	79.23%	71.80%	71.60%	90.60%	96.80%	48.35%	98.40%	86.43%
★Llava-one-vision (72B) [30]	77.49%	74.20%	70.60%	92.75%	99.20%	48.32%	99.60%	87.35%
★InternVL2.5 (78B) [9]	67.82%	68.60%	64.40%	84.00%	94.00%	47.19%	95.60%	80.17%
★InternVL3 (78B) [55]	69.25%	96.00%	64.60%	86.00%	96.00%	92.60%	97.80%	86.04%
△Gemini1.5-pro [48]	97.45%	**97.00%**	91.96%	91.70%	**100.0%**	99.60%	**100.0%**	96.82%
△Grok2 Vision [58]	76.48%	74.67%	63.27%	67.22%	76.48%	94.31%	98.28%	78.67%
Model Average (Zero-shot)	81.42%	79.85%	78.57%	81.92%	92.60%	79.14%	94.58%	84.91%
♦LGrad* [47]	56.52%	53.33%	99.00%	97.85%	68.24%	92.53%	81.94%	78.49%
♦InternVL2.5* (8B) [9]	91.30%	99.45%	99.83%	99.85%	95.34%	99.86%	99.20%	97.83%
♦InternVL3* (9B) [55]	83.33%	98.91%	99.02%	99.81%	96.17%	99.38%	99.94%	96.65%
♦Qwen2.5-VL* (7B) [3]	52.17%	98.33%	100.0%	99.17%	90.56%	98.84%	92.16%	90.18%
♦**MoA-DF (Ours)**	90.91%	100.0%	100.0%	99.90%	98.30%	99.91%	99.82%	98.41%

Table 3: Performance benchmark on AI-edit subsets, including real source and four editing types. ♦* refers to finetuned models.

Dimension	Object Enhance		Object Operation		Semantic Change		Style Change		Real Source		Overall	
Methods / Metrics	Acc(%)↑	F1↑	Acc(%)↑	F1↑	Acc(%)↑	F1↑	Acc(%)↑	F1↑	Acc(%)↑	F1↑	Acc(%)↑	F1↑
♡CnnSpott [16]	0.661	0.013	0.861	0.017	1.131	0.022	1.184	0.023	**99.90**	0.629	50.43	0.324
♡AntifakePrompt [7]	43.64	0.479	40.15	0.443	24.14	0.339	25.36	0.340	65.32	0.531	49.32	0.465
♡Gram-Net [36]	10.66	0.176	11.29	0.185	9.641	0.164	9.389	0.161	89.28	0.603	49.76	0.387
♡UnivFD [40]	5.372	0.101	9.103	0.166	12.52	0.220	14.02	0.243	98.54	0.646	54.39	0.414
♡LGrad [47]	**52.07**	**0.577**	**59.15**	**0.627**	**63.36**	**0.729**	**63.98**	**0.708**	76.06	0.679	**67.85**	**0.670**
★Llava-one-vision (0.5B) [30]	0.820	0.016	0.983	0.019	2.439	0.048	1.611	0.032	**99.80**	0.627	50.63	0.328
★DeepSeekVL (7B) [37]	1.983	0.037	3.509	0.065	6.507	0.119	7.560	0.136	95.85	0.621	50.37	0.355
★LLaVA-1.5 (7B) [35]	0.909	0.018	1.225	0.024	2.856	0.055	3.484	0.067	99.45	0.630	50.78	0.335
★Llava-one-vision (7B) [30]	4.711	0.087	6.620	0.121	13.97	0.241	16.77	0.281	96.97	0.640	53.74	0.411
★mPLUG-Owl3 (7B) [61]	5.207	0.092	6.819	0.120	14.96	0.247	17.39	0.280	91.69	0.619	51.39	0.402
★Qwen2.5-VL (7B) [3]	18.10	0.291	21.28	0.334	35.15	0.500	38.04	0.527	92.16	0.651	60.15	0.532
★CogAgent (18B) [22]	6.116	0.112	9.864	0.176	19.62	0.324	21.60	0.349	97.23	0.651	55.77	0.446
★InternVL2.5 (8B) [9]	10.80	0.182	12.62	0.212	20.47	0.324	21.34	0.335	91.49	0.611	53.90	0.437
★InternVL3 (9B) [55]	14.07	0.229	15.87	0.255	23.58	0.361	24.93	0.376	89.43	0.606	54.52	0.456
★InternLM-XComposer2.5 (7B) [62]	0.530	0.010	0.572	0.011	1.047	0.020	1.499	0.029	97.86	0.626	49.39	0.322
★LLaVA-NeXT (8B) [32]	34.05	0.433	36.31	0.469	43.53	0.546	43.80	0.545	77.96	0.624	58.69	0.561
★Llama3.2-Vision (11B) [38]	**43.68**	0.484	43.26	0.481	48.56	0.551	49.88	0.557	61.35	0.544	53.85	0.531
★Qwen2-VL (72B) [53]	8.683	0.143	15.97	0.249	24.63	0.364	27.89	0.402	91.20	0.724	55.25	0.507
★Qwen2.5-VL (72B) [3]	12.28	0.193	18.99	0.286	27.41	0.384	32.04	0.437	88.74	0.719	55.71	0.522
★Llava-one-vision (72B) [30]	8.982	0.150	14.84	0.241	26.95	0.395	31.64	0.454	92.99	**0.736**	56.80	0.523
★InternVL2.5 (78B) [9]	35.76	0.404	40.63	0.460	55.22	0.603	57.63	0.606	74.02	0.693	60.66	0.606
★InternVL3 (78B) [55]	20.06	0.276	28.68	0.386	38.89	0.494	42.04	0.523	84.48	0.719	58.45	0.569
△Gemini1.5-pro [48]	1.321	0.026	3.257	0.063	4.501	0.086	6.374	0.120	99.56	0.634	51.71	0.354
△Grok2 Vision [58]	40.83	**0.533**	**46.48**	**0.597**	**60.19**	**0.714**	**62.22**	**0.727**	88.99	**0.735**	**70.71**	**0.689**
Model Average (Zero-shot)	15.89	0.211	18.68	0.250	24.22	0.327	25.90	34.41	89.18	0.646	55.18	0.464
♦LGrad* [47]	68.24	0.742	65.54	0.711	79.32	0.829	75.48	0.801	72.15	0.771	77.39	0.770
♦InternVL2.5* (8B) [9]	97.27	0.976	92.98	0.952	96.82	0.976	96.34	0.974	95.85	0.970	96.84	0.968
♦InternVL3* (9B) [55]	92.16	0.934	85.92	0.896	96.30	0.965	93.88	0.950	92.06	0.936	93.46	0.934
♦Qwen2.5-VL* (7B) [3]	81.57	0.847	85.12	0.879	93.62	0.931	93.00	0.928	88.33	0.896	89.08	0.891
♦**MoA-DF (Ours)**	96.08	0.972	93.27	0.955	97.85	0.983	96.51	0.976	95.93	0.971	97.07	0.970

4 The MoA-DF Method

To leverage the strong zero-shot capabilities of LMMs for robust deepfake detection, we propose the MoA-DF, mixture of agents for deepfake detection that integrates the knowledge of multiple state-of-the-art LMMs. Specifically, we select Qwen2.5 (7B), InternVL2.5 (8B), and InternVL3 (9B) as the core detection agents. Each model outputs log-probabilities of A (real) or B (fake), denoted as $\log p_i(A)$ and $\log p_i(B)$ for model i, which are then normalized using softmax:

$$p_A^{(i)} = \frac{e^{\log p_A}}{e^{\log p_A} + e^{\log p_B}}, \quad p_B^{(i)} = \frac{e^{\log p_B}}{e^{\log p_A} + e^{\log p_B}} \tag{1}$$

We then aggregate the predictions from all $N = 3$ models:

$$P_A = \sum_{i=1}^{N} p_A^{(i)}, \quad P_B = \sum_{i=1}^{N} p_B^{(i)} \tag{2}$$

Table 4: Performance benchmark on AI-generated subsets. ♡Deepfake detection models, ★open-source and △close-source LMMs. ♦* refers to finetuned models. Best and second-best zero-shot results. Best and second-best finetuned results.

Datasets	Playground		SD3.5_Large		PixArt-Sigma		Infinity		Kandinsky-3		Flux_Schnell		Kolors	
Methods / Metrics	Acc(%)↑	F1↑	Acc(%)↑	F1↑	Acc(%)↑	F1↑	Acc(%)↑	F1↑	Acc(%)↑	F1↑	Acc(%)↑	F1↑	Acc(%)↑	F1↑
♡CnnSpott [16]	0.000	0.000	0.363	0.007	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
♡AntifakePrompt [7]	0.913	0.016	4.775	0.083	1.113	0.020	0.638	0.011	1.625	0.029	7.763	0.131	1.150	0.021
♡Gram-Net [36]	10.98	0.173	4.800	0.080	2.213	0.375	2.425	0.041	0.113	0.001	0.050	0.001	0.025	0.000
♡UnivFD [40]	0.063	0.001	0.100	0.002	0.563	0.011	0.063	0.001	0.113	0.002	0.050	0.001	0.025	0.000
♡LGrad [47]	70.54	0.626	**70.73**	**0.627**	35.89	0.376	**89.94**	0.735	**87.74**	0.724	**67.29**	0.606	5.088	0.064
★Llava-one-vision (0.5B) [30]	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
★DeepSeekVL (7B) [37]	8.513	0.156	0.725	0.014	16.54	0.283	3.488	0.067	14.75	0.256	1.525	0.030	5.763	0.109
★LLaVA-1.5 (7B) [35]	2.038	0.040	0.113	0.002	3.375	0.065	1.875	0.037	13.21	0.233	0.438	0.009	1.413	0.028
★Llava-one-vision (7B) [30]	12.88	0.228	1.063	0.021	24.15	0.390	6.050	0.114	16.64	0.285	4.488	0.086	8.850	0.162
★mPLUG-Owl3 (7B) [61]	19.66	0.328	2.213	0.043	42.15	0.592	16.94	0.289	19.30	0.323	13.03	0.230	20.48	0.339
★Qwen2.5-VL (7B) [3]	54.94	0.706	15.46	0.266	66.60	0.796	27.44	0.428	34.59	0.511	28.60	0.442	45.94	0.626
★CogAgent (18B) [22]	10.86	0.196	1.175	0.023	23.80	0.384	3.763	0.072	16.76	0.287	1.463	0.029	8.100	0.150
★InternVL2.5 (8B) [9]	40.13	0.571	5.013	0.095	44.44	0.614	17.78	0.301	27.70	0.433	13.30	0.234	20.39	0.338
★InternVL3 (9B) [55]	41.28	0.583	6.325	0.119	48.66	0.653	22.63	0.368	30.24	0.463	13.69	0.240	26.88	0.422
★InternLM-XComposer2.5 (7B) [62]	15.69	0.294	1.150	0.023	19.51	0.327	4.025	0.077	16.20	0.279	1.875	0.037	8.438	0.156
★LLaVA-NeXT (8B) [32]	30.16	0.457	4.738	0.089	35.21	0.514	10.00	0.179	23.79	0.379	4.550	0.086	18.78	0.311
★Llama3.2-Vision (11B) [38]	**90.21**	**0.816**	**64.08**	**0.658**	**89.08**	0.812	**91.29**	0.818	**81.50**	0.765	**81.95**	**0.768**	**83.64**	0.796
★Qwen2-VL (72B) [53]	45.60	0.624	14.00	0.244	53.20	0.692	42.80	0.597	43.80	0.607	27.20	0.426	40.00	0.569
★Qwen2.5-VL (72B) [3]	**86.60**	**0.920**	19.60	0.323	**92.40**	**0.953**	72.00	**0.829**	64.60	**0.777**	52.00	**0.677**	**91.00**	**0.945**
★Llava-one-vision (72B) [30]	33.40	0.499	4.200	0.080	48.60	0.652	25.80	0.409	26.80	0.421	18.80	0.315	22.00	0.359
★InternVL2.5 (78B) [9]	69.60	0.800	29.80	0.444	87.40	**0.911**	81.00	**0.874**	74.80	**0.835**	52.20	0.667	78.80	**0.860**
★InternVL3 (78B) [55]	41.28	0.583	6.235	0.119	48.66	0.653	22.63	0.368	30.24	0.463	13.69	0.240	26.88	0.422
△Gemini1.5-pro [48]	9.538	0.175	0.675	0.013	17.39	0.297	4.663	0.089	14.83	0.258	1.250	0.025	6.800	0.128
△Grok2 Vision [58]	32.45	0.484	11.96	0.211	46.23	0.625	23.44	0.375	30.86	0.466	19.96	0.328	58.64	0.451
Model Average (Zero-shot)	30.30	0.387	11.22	0.149	35.30	0.444	23.78	0.295	27.92	0.367	18.10	0.240	24.22	0.304
♦LGrad* [47]	98.69	0.904	98.11	0.901	95.75	0.889	99.88	0.910	100.0	0.911	98.25	0.902	98.19	0.902
♦InternVL2.5* (8B) [9]	100.0	0.996	99.81	0.995	100.0	0.996	100.0	0.996	100.0	0.996	99.94	0.996	100.0	0.996
♦InternVL3* (9B) [55]	99.88	0.999	99.25	0.996	99.81	0.999	99.88	0.999	99.94	0.999	99.44	0.997	100.0	1.000
♦Qwen2.5-VL* (7B) [3]	100.0	0.962	99.94	0.961	100.0	0.962	100.0	0.962	100.0	0.962	100.0	0.962	100.0	0.962
♦MoA-DF (Ours)	100.0	0.998	99.85	0.997	100.0	0.998	100.0	0.998	100.0	0.998	100.0	0.998	100.0	0.998

Datasets	SD3_Medium		Flux_dev		NOVA		LaVi-Bridge		Janus		Real Source		Overall	
Methods / Metrics	Acc(%)↑	F1↑	Acc(%)↑	F1↑	Acc(%)↑	F1↑	Acc(%)↑	F1↑	Acc(%)↑	F1↑	Acc(%)↑	F1↑	Acc(%)↑	F1↑
♡CnnSpott [16]	0.025	0.000	0.000	0.000	0.275	0.005	0.013	0.000	0.925	0.018	99.65	0.668	7.789	0.054
♡AntifakePrompt [7]	4.588	0.079	5.713	0.098	0.963	0.017	0.013	0.000	0.863	0.015	89.17	0.625	9.176	0.088
♡Gram-Net [36]	2.688	0.045	0.700	0.012	0.938	0.016	0.450	0.008	7.888	0.127	84.29	0.603	9.918	0.102
♡UnivFD [40]	0.088	0.002	0.000	0.000	2.638	0.051	0.313	0.006	29.63	0.456	99.83	0.675	10.27	0.093
♡LGrad [47]	**47.16**	0.467	**85.34**	0.711	23.60	0.265	81.71	0.693	25.18	0.280	45.83	0.488	56.62	0.512
★Llava-one-vision (0.5B) [30]	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
★DeepSeekVL (7B) [37]	0.688	0.014	6.800	0.127	10.20	0.185	52.78	0.689	12.35	0.219	99.68	0.693	17.98	0.219
★LLaVA-1.5 (7B) [35]	0.125	0.002	4.025	0.077	4.050	0.078	42.88	0.600	5.450	0.103	**100.0**	0.684	13.77	0.151
★Llava-one-vision (7B) [30]	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.225	0.004	99.96	0.669	7.707	0.052
★mPLUG-Owl3 (7B) [61]	2.763	0.054	14.09	0.246	31.16	0.474	84.59	0.915	36.76	0.537	99.73	0.729	30.99	0.392
★Qwen2.5-VL (7B) [3]	19.69	0.327	33.08	0.494	32.35	0.486	86.03	0.921	26.69	0.419	99.23	0.766	43.89	0.553
★CogAgent (18B) [22]	1.100	0.022	7.163	0.134	26.84	0.423	69.61	0.820	47.55	0.644	99.86	0.711	24.47	0.300
★InternVL2.5 (8B) [9]	9.625	0.175	18.83	0.316	18.94	0.318	76.71	0.867	17.98	0.304	99.69	0.730	31.58	0.407
★InternVL3 (9B) [55]	9.288	0.169	17.85	0.304	31.00	0.472	84.34	0.913	23.70	0.382	99.62	0.740	35.04	0.448
★InternLM-XComposer2.5 (7B) [62]	0.925	0.018	7.213	0.135	7.813	0.145	31.83	0.483	2.830	0.055	**100.0**	0.712	17.47	0.216
★LLaVA-NeXT (8B) [32]	5.238	0.098	7.725	0.173	31.95	0.478	87.09	0.922	47.46	0.636	98.16	0.727	31.14	0.388
★Llama3.2-Vision (11B) [38]	**71.14**	**0.702**	**90.80**	**0.816**	**66.60**	0.671	85.60	0.787	51.20	0.560	68.39	0.726	**78.11**	**0.746**
★Qwen2-VL (72B) [53]	14.80	0.256	40.80	0.577	37.20	0.540	85.20	0.917	37.40	0.542	98.96	0.654	44.69	0.557
★Qwen2.5-VL (72B) [3]	30.00	0.456	76.60	**0.860**	71.80	**0.828**	**98.00**	**0.982**	25.60	0.403	98.40	0.843	67.58	0.754
★Llava-one-vision (72B) [30]	6.000	0.113	13.60	0.239	37.60	0.545	89.00	0.940	49.20	0.658	99.60	0.742	36.51	0.459
★InternVL2.5 (78B) [9]	38.40	**0.538**	61.20	0.739	**63.00**	**0.753**	**98.40**	**0.970**	**65.40**	**0.770**	95.60	0.835	**68.89**	**0.769**
★InternVL3 (78B) [55]	25.00	0.393	39.00	0.552	51.80	0.673	93.80	0.957	57.40	0.719	97.80	0.860	55.98	0.666
△Gemini1.5-pro [48]	0.600	0.012	7.475	0.139	10.19	0.185	45.38	0.624	50.19	0.669	**100.0**	0.702	20.69	0.255
△Grok2 Vision [58]	12.65	0.221	46.38	0.374	34.25	0.504	74.11	0.840	**60.06**	**0.742**	98.28	0.709	42.25	0.487
Model Average (Zero-shot)	12.64	0.174	24.68	0.303	25.59	0.351	59.56	0.651	29.79	0.407	94.65	0.708	32.72	0.373
♦LGrad* [47]	96.31	0.892	99.69	0.909	91.38	0.866	99.69	0.909	95.63	0.889	81.94	0.890	96.42	0.898
♦InternVL2.5* (8B) [9]	99.94	0.996	100.0	0.996	99.81	0.995	100.0	0.996	98.76	0.990	99.20	0.995	99.80	0.995
♦InternVL3* (9B) [55]	99.88	0.999	99.88	0.999	99.83	0.999	100.0	1.000	99.31	0.996	99.94	0.999	99.77	0.998
♦Qwen2.5-VL* (7B) [3]	100.0	0.962	100.0	0.962	99.81	0.961	100.0	0.962	100.0	0.962	92.16	0.959	99.38	0.962
♦MoA-DF (Ours)	100.0	0.998	100.0	0.998	99.85	0.998	100.0	0.998	99.69	0.997	99.70	0.998	99.92	0.998

The final decision D is made:

$$D = \begin{cases} \text{A (Real)}, & \text{if } P_A > P_B \\ \text{B (Fake)}, & \text{otherwise} \end{cases} \tag{3}$$

This ensemble strategy effectively leverages the diverse strengths and perspectives of multiple large models by fusing their soft predictions. By combining probabilistic outputs, MoA-DF mitigates individual model biases and uncertainties, resulting in enhanced robustness and improved overall detection accuracy.

5 Benchmark and Evaluation

We benchmark and evaluate the performance of various deepfake detection models across three subsets of DFBench: real, AI-edited, and AI-generated images.

5.1 Experiment Setup

We evaluate the models' ability to correctly classify real and fake images using two standard metrics: accuracy (Acc) and F1-score. Accuracy is defined as the proportion of correctly identified real or

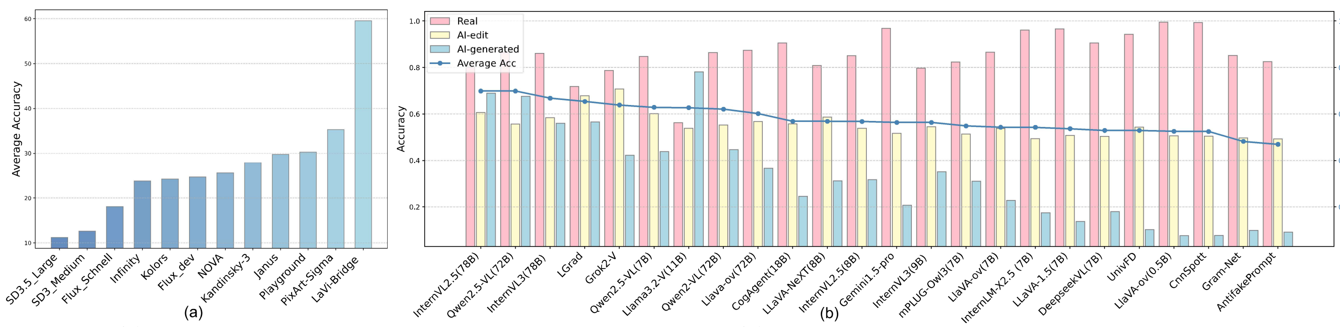

Figure 5: (a) Performance comparison of image generation models (b) Performance comparison of image detection models

fake images out of all relevant samples in the dataset, computed as:

$$\text{Acc} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{4}$$

TP (True Positives) denotes the number of real or fake images correctly identified by the model, while FN (False Negatives) represents the number of images incorrectly classified as the opposite category. To provide a balanced evaluation that considers both precision and recall, we also calculate the F1-score, the harmonic mean of precision and recall, defined as:

$$\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5}$$

For conventional deepfake detection models, we directly utilize publicly available pre-trained weights to conduct inference on the test datasets. For large multimodal models, inference is performed via a prompt-based question-answering approach. We fine-tune three of the LMMs with LoRA [24] (r=8) and LGrad [47] using the same training and testing split (4:1). We set the number of finetuning epoch to 1 for LMMs and 50 for LGrad [47]. The models are implemented with PyTorch and trained on a 40GB NVIDIA RTX A6000 GPU with batch size of 4. The initial learning rate is set to 1e-5 and decreased using the cosine annealing strategy.

5.2 Benchmark on Real Datasets

From the performance results presented in Table 2, it is evident that most models exhibit strong zero-shot identification capabilities on real image datasets. However, detection accuracy generally declines on datasets containing various distortions, such as CSIQ [29] and TID2013 [42], when compared to the distortion-free Flickr8k dataset [21], indicating that image degradations such as noise, blur or compression can impact model reliability and increase the chance of misclassification. CnnSpott [16] and Llava-one-vision (0.5B) [30] perform well on real images mainly because they tend to classify most inputs as real, but may reduce robustness in fake detection.

5.3 Benchmark on AI-edit Datasets

We further evaluate the performance of different detection models on AI-edit subsets. As shown in Table 3, the high performance of CnnSpott [16] and Llava-one-vision (0.5B) [30] on real source images significantly drops on AI-edited images, resulting in relatively lower F1 scores. The AI-edit datasets consist of four categories: object enhancement, object operation, style change, and semantic change, each posing different challenges for detection models. Among these, models achieve the highest average accuracy on style change and the lowest on the object enhancement category, which involves subtle modifications to the appearance of individual objects. These results suggest that subtle changes at the object level are more difficult for detection models to identify compared to style changes. Models show significant improvements after fine-tuning, especially LMMs trained for 1 epoch outperform conventional best network LGrad [47] trained for 50 epochs, highlighting the effectiveness of LMMs in deepfake detection tasks.

5.4 Benchmark on AI-generation Datasets

From Table 4, we can observe that the detection accuracy on AI-generated datasets is generally lower compared to real image datasets, highlighting the remarkable realism achieved by current generative models and their strong capability to evade detection. Traditional deep learning-based detection models trained on specific deepfake datasets show limited zero-shot generalization, reflecting their insufficient scaling-up capacity to handle more advanced fakes. In contrast, large multimodal models, despite lacking task-specific training for real-fake discrimination, demonstrate relatively robust zero-shot detection performance. Among these, InternVL2.5 (78B) [9] achieves the best results, suggesting that larger parameter scales contribute to better generalization capabilities. On the generation side, detection accuracy also serves as an indirect measure of generative models' evasion effectiveness. As shown in Figure 5, SD3.5-Large [14] attains the lowest detection accuracy, indicating its superior capacity for generating highly realistic images that effectively fool detectors, while LaVi-Bridge [63] exhibits the poorest evasion performance.

6 CONCLUSION

In this paper, we introduce DFBench, a comprehensive benchmark designed to advance deepfake image detection. DFBench features the largest scale of fake images generated by 12 state-of-the-art generative models, and rich content spanning AI-edited images and real-world image distortions. we introduce a bidirectional evaluation protocol that assesses both the detection performance of deepfake models and the evasion strength of generative models. Additionally, we propose MoA-DF, a novel mixture of agents method that integrates LMMs within a unified probabilistic framework, achieving state-of-the-art performance and demonstrating the effectiveness of LMMs for deepfake detection. Through extensive experiments, we demonstrate the increasing realism of generative models and the limited generality of current detection methods. LMMs manifest strong zero-shot generalization ability, highlighting their potential as a promising foundation for developing more robust and generalizable deepfake detection systems.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grants 62271312, 62401365, 62225112, 62132006, U24A20220, and in part by the China Postdoctoral Science Foundation under Grant Number BX20250411, 2025M773473.

References

[1] Donghoon Ahn, Hyoungwon Cho, Jaewon Min, Wooseok Jang, Jungwoo Kim, SeonHwa Kim, Hyun Hee Park, Kyong Hwan Jin, and Seungryong Kim. 2024. Self-rectifying diffusion sampling with perturbed-attention guidance. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 1–17.

[2] Vladimir Arkhipkin, Viacheslav Vasilev, Andrei Filatov, Igor Pavlov, Julia Agafonova, Nikolai Gerasimenko, Anna Averchenkova, Evelina Mironova, Anton Bukashkin, Konstantin Kulikov, et al. 2024. Kandinsky 3: Text-to-image synthesis for multifunctional generative framework. *arXiv preprint arXiv:2410.21061* (2024).

[3] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. 2025. Qwen2.5-VL Technical Report. *arXiv preprint arXiv:2502.13923* (2025).

[4] Xiuli Bi, Bo Liu, Fan Yang, Bin Xiao, Weisheng Li, Gao Huang, and Pamela C. Cosman. 2023. Detecting Generated Images by Real Images Only. *Arxiv* (2023).

[5] Jordan J Bird and Ahmad Lotfi. 2023. CIFAKE: Image Classification and Explainable Identification of AI-Generated Synthetic Images. *arXiv preprint arXiv:2303.14126* (2023).

[6] Andrew Brock, Jeff Donahue, and Karen Simonyan. 2018. Large scale GAN training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096* (2018).

[7] You-Ming Chang, Chen Yeh, Wei-Chen Chiu, and Ning Yu. 2023. AntifakePrompt: Prompt-Tuned Vision-Language Models are Fake Image Detectors. *Arxiv* (2023).

[8] Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. 2024. Pixart-σ: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 74–91.

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

[10] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. 2022. Diffedit: Diffusion-based semantic image editing with mask guidance. *arXiv preprint arXiv:2210.11427* (2022).

[11] Hao Dang, Feng Liu, Joel Stehouwer, Xiaoming Liu, and Anil K Jain. 2020. On the detection of digital face manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern recognition (CVPR)*. 5781–5790.

[12] Haoge Deng, Ting Pan, Haiwen Diao, Zhengxiong Luo, Yufeng Cui, Huchuan Lu, Shiguang Shan, Yonggang Qi, and Xinlong Wang. 2024. Autoregressive Video Generation without Vector Quantization. *arXiv preprint arXiv:2412.14169* (2024).

[13] Prafulla Dhariwal and Alexander Nichol. 2021. Diffusion models beat gans on image synthesis. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 34. 8780–8794.

[14] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. 2024. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the International Conference on Machine Learning (ICML)*.

[15] Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. 2020. Leveraging frequency analysis for deep fake image recognition. In *International Conference on Machine Learning (ICML)*. PMLR, 3247–3258.

[16] Joel Frank and Thorsten Holz. 2021. CNN-generated images are surprisingly easy to spot...for now. *Arxiv* (2021).

[17] Apurva Gandhi and Shomik Jain. 2020. Adversarial perturbations fool deepfake detectors. In *International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–8.

[18] Deepti Ghadiyaram and Alan C Bovik. 2015. Massive online crowdsourced study of subjective and objective picture quality. *IEEE Transactions on Image Processing (TIP)* 25 (2015).

[19] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. 2022. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10696–10706.

[20] Jian Han, Jinlai Liu, Yi Jiang, Bin Yan, Yuqi Zhang, Zehuan Yuan, Bingyue Peng, and Xiaobing Liu. 2024. Infinity: Scaling Bitwise AutoRegressive Modeling for High-Resolution Image Synthesis. *arXiv preprint arXiv:2412.04431* (2024).

[21] Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research (JAIR)* 47 (2013), 853–899.

[22] Wenyi Hong, Weihan Wang, Qingsong Lv, Jiazheng Xu, Wenmeng Yu, Junhui Ji, Yan Wang, Zihan Wang, Yuxuan Zhang, Juanzi Li, Bin Xu, Yuxiao Dong, Ming Ding, and Jie Tang. 2024. CogAgent: A Visual Language Model for GUI Agents. *arXiv preprint arXiv:2312.08914* (2024).

[23] Vlad Hosu, Hanhe Lin, Tamas Sziranyi, and Dietmar Saupe. 2020. KonIQ-10k: An ecologically valid database for deep learning of blind image quality assessment. *IEEE Transactions on Image Processing (TIP)* (2020), 4041–4056.

[24] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, Vol. 1. 3.

[25] Zhenglin Huang, Jinwei Hu, Xiangtai Li, Yiwei He, Xingyu Zhao, Bei Peng, Baoyuan Wu, Xiaowei Huang, and Guangliang Cheng. 2025. SIDA: Social Media Image Deepfake Detection, Localization and Explanation with Large Multimodal Model. *arXiv preprint arXiv:2412.04292* (2025).

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

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

[28] Black Forest Labs. 2024. FLUX. https://github.com/black-forest-labs/flux.

[29] Eric Cooper Larson and Damon Michael Chandler. 2010. Most apparent distortion: full-reference image quality assessment and the role of strategy. *Journal of Electronic Imaging (JEI)* 19, 1 (2010).

[30] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024. LLaVA-OneVision: Easy Visual Task Transfer. *arXiv preprint arXiv:2408.03326* (2024).

[31] Daiqing Li, Aleks Kamko, Ehsan Akhgari, Ali Sabet, Linmiao Xu, and Suhail Doshi. 2024. Playground v2.5: Three Insights towards Enhancing Aesthetic Quality in Text-to-Image Generation. *arXiv preprint arXiv:2402.17245* (2024).

[32] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. 2024. LLaVA-NeXT-Interleave: Tackling Multi-image, Video, and 3D in Large Multimodal Models. *arXiv preprint arXiv:2407.07895* (2024).

[33] Hanhe Lin, Vlad Hosu, and Dietmar Saupe. 2019. Kadid-10k: A large-scale artificially distorted iqa database. In *Proceedings of the International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE.

[34] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

[35] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 26296–26306.

[36] Zhengzhe Liu, Xiaojuan Qi, and Philip HS Torr. 2020. Global texture enhancement for fake face detection in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*. 8060–8069.

[37] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, Yaofeng Sun, Chengqi Deng, Hanwei Xu, Zhenda Xie, and Chong Ruan. 2024. DeepSeek-VL: Towards Real-World Vision-Language Understanding. *arXiv preprint arXiv:2403.05525* (2024).

[38] AI Meta. 2024. Llama 3.2: Revolutionizing edge AI and vision with open, customizable models. *Meta AI Blog. Retrieved December* 20 (2024), 2024.

[39] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. 2021. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741* (2021).

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

[41] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2641–2649.

[42] Nikolay Ponomarenko, Lina Jin, Oleg Ieremeiev, Vladimir Lukin, Karen Egiazarian, Jaakko Astola, Benoit Vozel, Kacem Chehdi, Marco Carli, Federica Battisti, et al. 2015. Image database TID2013: Peculiarities, results and perspectives. *Signal Processing: Image Communication (SPIC)* 30 (2015).

[43] Jiaying Qian, Ziheng Jia, Zicheng Zhang, Zeyu Zhang, Guangtao Zhai, and Xiongkuo Min. 2025. Towards Explainable Partial-AIGC Image Quality Assessment. *arXiv preprint arXiv:2504.09291* (2025).

[44] Zeyang Sha, Zheng Li, Ning Yu, and Yang Zhang. 2022. DE-FAKE: Detection and Attribution of Fake Images Generated by Text-to-Image Diffusion Models. *arXiv preprint arXiv:2210.06998* (2022).

[45] Hamid R Sheikh, Muhammad F Sabir, and Alan C Bovik. 2006. A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Transactions on Image Processing (TIP)* (2006), 3440–3451.

[46] Haixu Song, Shiyu Huang, Yinpeng Dong, and Wei-Wei Tu. 2023. Robustness and Generalizability of Deepfake Detection: A Study with Diffusion Models. *Arxiv* (2023).

[47] Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, and Yunchao Wei. 2023. Learning on Gradients: Generalized Artifacts Representation for GAN-Generated Images Detection. In *CVPR*.

[48] Google Team. 2024. Gemini1.5-pro. https://gemini.google.com/. Accessed: 2025-03-08.

[49] Kolors Team. 2024. Kolors: Effective Training of Diffusion Model for Photorealistic Text-to-Image Synthesis. *arXiv preprint* (2024).

[50] Luisa Verdoliva, Davide Cozzolino, and Koki Nagano. 2022. 2022 IEEE Image and Video Processing Cup Synthetic Image Detection.

[51] Jiarui Wang, Huiyu Duan, Jing Liu, Shi Chen, Xiongkuo Min, and Guangtao Zhai. 2023. Aigciqa2023: A large-scale image quality assessment database for ai generated images: from the perspectives of quality, authenticity and correspondence. In *Proceedings of the CAAI International Conference on Artificial Intelligence (CICAI)*. 46–57.

[52] Jiarui Wang, Huiyu Duan, Guangtao Zhai, and Xiongkuo Min. 2025. Quality Assessment for AI Generated Images with Instruction Tuning. *arXiv preprint arXiv:2405.07346* (2025).

[53] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191* (2024).

[54] Run Wang, Felix Juefei-Xu, Lei Ma, Xiaofei Xie, Yihao Huang, Jian Wang, and Yang Liu. 2019. Fakespotter: A simple yet robust baseline for spotting ai-synthesized fake faces. *arXiv preprint arXiv:1909.06122* (2019).

[55] Weiyun Wang, Zhe Chen, Wenhai Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Jinguo Zhu, Xizhou Zhu, Lewei Lu, Yu Qiao, and Jifeng Dai. 2024. Enhancing the Reasoning Ability of Multimodal Large Language Models via Mixed Preference Optimization. *arXiv preprint arXiv:2411.10442* (2024).

[56] Yabin Wang, Zhiwu Huang, and Xiaopeng Hong. 2023. Benchmarking Deepart Detection. *arXiv preprint arXiv:2302.14475* (2023).

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

[58] xAI Team. 2024. Grok2 Vision. https://grok.com/. Accessed: 2025-03-08.

[59] Zitong Xu, Huiyu Duan, Guangji Ma, Liu Yang, Jiarui Wang, Qingbo Wu, Xiongkuo Min, Guangtao Zhai, and Patrick Le Callet. 2025. HarmonyIQA: Pioneering Benchmark and Model for Image Harmonization Quality Assessment. *arXiv preprint arXiv:2501.01116* (2025).

[60] Xin Yang, Yuezun Li, and Siwei Lyu. 2019. Exposing deep fakes using inconsistent head poses. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 8261–8265.

[61] Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 2024. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

[62] Pan Zhang, Xiaoyi Dong, Bin Wang, Yuhang Cao, Chao Xu, Linke Ouyang, Zhiyuan Zhao, Shuangrui Ding, Songyang Zhang, Haodong Duan, Wenwei Zhang, Hang Yan, Xinyue Zhang, Wei Li, Jingwen Li, Kai Chen, Conghui He, Xingcheng Zhang, Yu Qiao, Dahua Lin, and Jiaqi Wang. 2023. InternLM-XComposer: A Vision-Language Large Model for Advanced Text-image Comprehension and Composition. *arXiv preprint arXiv:2309.15112* (2023).

[63] Shihao Zhao, Shaozhe Hao, Bojia Zi, Huaizhe Xu, and Kwan-Yee K Wong. 2024. Bridging different language models and generative vision models for text-to-image generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 70–86.