Supplementary Material
Attribution as Retrieval: Model-Agnostic AI-Generated Image Attribution

Hongsong Wang[1,2], Renxi Cheng[3], Chaolei Han[3], Jie Gui[3,4,5*]
[1]School of Computer Science and Engineering, Southeast University, Nanjing 210096, China
[2]Key Laboratory of New Generation Artificial Intelligence Technology and Its Interdisciplinary Applications (Southeast University), Ministry of Education, China
[3]School of Cyber Science and Engineering, Southeast University, Nanjing 210096, China
[4]Purple Mountain Laboratories, Nanjing 210000, China
[5]Engineering Research Center of Blockchain Application, Supervision And Management (Southeast University), Ministry of Education, China
{hongsongwang, renxi, chaoleihan, guijie}@seu.edu.cn

Shot	Method	Real		Others		DMs		GANs		Avg	
		Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
1-shot	ResNet [7]	1.5	29.7	21.6	53.0	30.2	55.7	**50.8**	**73.6**	26.0	53.0
	DIRE [23]	48.7	74.3	**83.4**	**91.1**	5.0	34.8	3.5	30.7	35.2	57.7
	ESSP [2]	7.0	41.0	76.4	85.9	10.6	35.2	29.1	60.9	30.8	55.7
	Ours	**80.4**	**88.4**	81.4	90.7	**48.7**	**62.0**	18.6	56.2	**57.3**	**74.3**
5-shot	ResNet [7]	23.6	44.6	**69.8**	67.6	9.5	29.3	**50.3**	47.9	38.3	47.4
	DIRE [23]	39.7	41.3	50.3	54.5	25.6	37.0	36.2	**50.2**	37.9	45.8
	ESSP [2]	40.2	44.9	60.3	63.9	23.6	**41.4**	34.2	40.1	39.6	47.5
	Ours	**90.5**	**80.3**	66.3	**86.7**	**35.2**	38.2	47.2	48.8	**59.8**	**63.5**
10-shot	ResNet [7]	53.3	42.1	80.4	59.1	23.6	33.4	27.1	39.0	46.1	43.4
	DIRE [23]	44.2	46.1	58.8	47.6	28.6	36.7	27.6	36.2	39.8	41.7
	ESSP [2]	47.2	39.4	58.3	59.9	19.1	34.9	20.6	31.8	36.3	41.5
	Ours	**83.9**	**81.9**	**86.4**	**87.4**	**83.9**	**54.9**	**35.7**	**44.1**	**72.5**	**67.1**

Table 1. Performance comparison of AI-generated image attribution on the WildFake dataset under the cross-generator setting with different numbers of shots. The best score for each shot setting is highlighted in bold.

Shot	Method	VQV	COM	DD	VQD	Big	Sty	Sta	DF	GA	Gi	Avg
1-shot	ResNet [7]	55.5	39.4	53.8	53.0	56.3	52.3	54.0	36.7	57.0	46.7	50.5
	DIRE [23]	51.3	49.0	48.0	43.9	42.7	50.3	51.5	51.0	51.5	51.0	49.2
	ESSP [2]	52.3	55.3	46.2	51.3	54.8	54.3	48.9	46.2	60.3	38.9	50.8
	Ours	**84.4**	**70.4**	**85.2**	**85.2**	**83.9**	**85.9**	**85.2**	**83.9**	**84.7**	**66.8**	**81.6**
5-shot	ResNet [7]	51.5	52.5	50.5	49.0	49.0	52.0	54.8	51.8	55.3	48.7	51.5
	DIRE [23]	52.5	48.7	53.2	48.7	52.7	51.5	52.0	52.7	53.0	51.0	51.6
	ESSP [2]	53.0	49.7	52.7	53.8	56.8	53.5	56.3	52.5	56.0	50.0	53.4
	Ours	**85.4**	**82.2**	**84.9**	**84.7**	**85.2**	**84.7**	**85.9**	**85.7**	**86.4**	**83.7**	**84.9**
10-shot	ResNet [7]	59.5	43.2	57.3	56.8	58.8	60.8	61.3	53.0	54.5	58.0	56.3
	DIRE [23]	59.3	51.8	54.5	52.5	57.3	58.0	58.5	56.3	59.8	50.8	55.9
	ESSP [2]	54.8	50.5	52.3	58.0	57.9	57.2	56.2	50.8	55.0	51.5	54.4
	Ours	**90.7**	**85.2**	**89.9**	**91.2**	**90.2**	**89.9**	**90.7**	**91.7**	**89.4**	**87.2**	**89.6**

Table 2. Accuracy of AI-generated image detection on the WildFake dataset with different numbers of shots. The best score for each shot setting is highlighted in bold.

We first analyze the attribution and detection performance of our Low-bIt-plane-based Deepfake Attribution (LIDA) method on the WildFake dataset [9]. We then vividly visualize the retrieval-based attribution results. Finally, we illustrate how the low-bit generative fingerprint guides the selection of informative image patches.

A. Performance Analyses

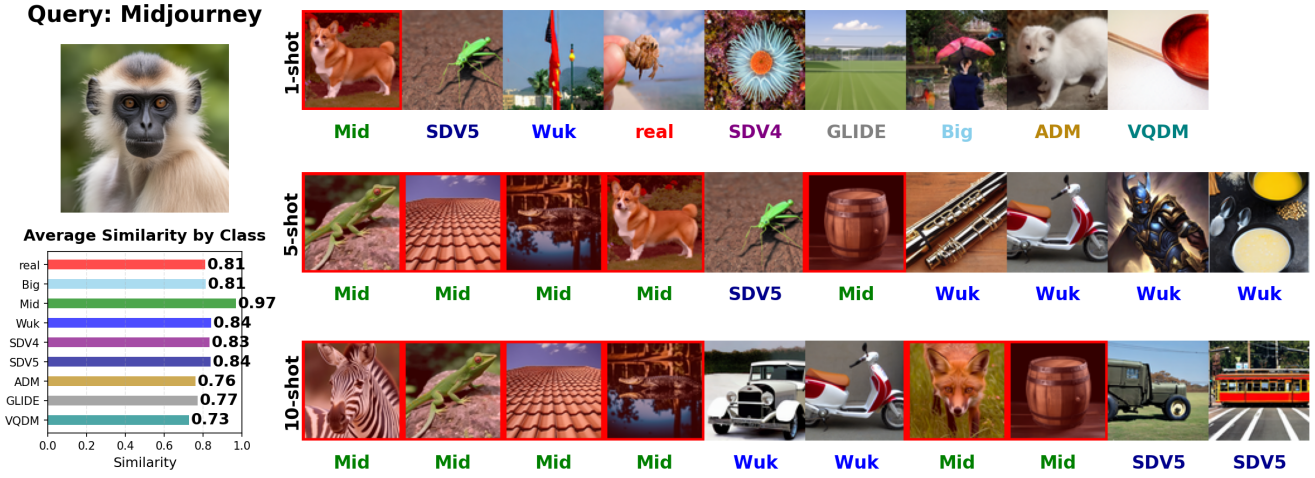
Generator-Level Image Attribution: We further categorize the generative algorithms into three groups based on their model architectures: diffusion-based (DDPM [8], DDIM [19], ADM [5], DALL·E [16], Stable Diffusion [17], Midjourney [14], Imagen [18], VQDM [6]), GAN-based (BigGAN [1], StyleGAN [11–13], GigaGAN [10], StarGAN [3, 4], DF-GAN [20], GALIP [21]), and other (VQ-VAE [22]), to evaluate cross-generator performance. As shown in Table 1, our method achieves the best performance across all evaluation settings. More specifically, the proposed low-bit generative fingerprints outperform the features extracted by DIRE and ESSP by 25.4% and 36.3% in terms of mAP, respectively, indicating their stronger suitability for fake image attribution. As the number of images in the registered database increases, Rank-1 accuracy improves while the task of ranking similarities across all images becomes progressively more challenging.

Deepfake Detection: A similar trend is observed for the Deepfake detection task on the WildFake dataset, where our model consistently outperforms all competitors, as shown in Table 2. Specifically, under the 1-shot, 5-shot, and 10-shot settings, our model exceeds ESSP by 30.8%, 31.5%, and 35.2%, respectively, and outperforms DIRE by 32.4%, 33.3%, and 33.7%, respectively. Notably, all baseline methods remain close to random guessing across generators and shot settings, whereas our method performs substantially better, achieving over 80% accuracy, which demonstrates its superior cross-architecture generalization.

*Corresponding author

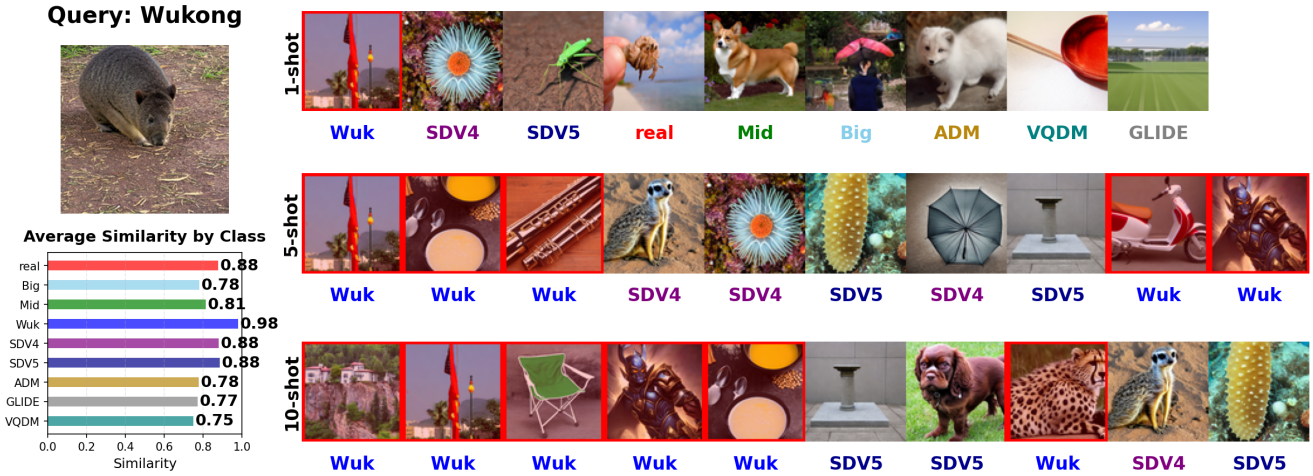
(a) Real query image

(b) Big-GAN generated query image



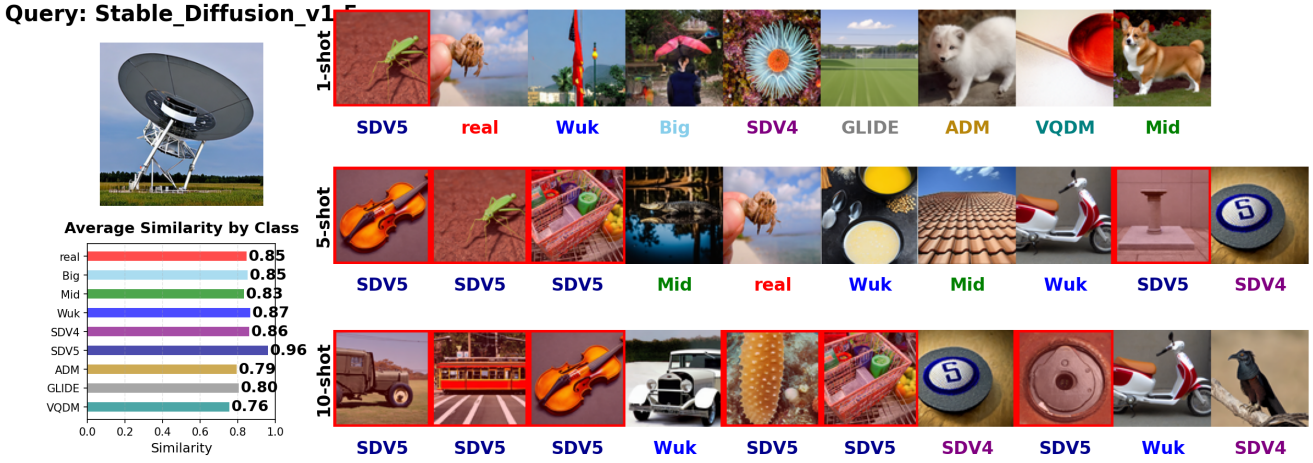
(c) Midjourney generated query image

Figure 1. **Illustration of retrieval-based image attribution.** The top-ranked retrieval results from the database are listed, with the correct result highlighted by a red bounding box. The queries are a real image, a BigGAN-generated [1] image, and a Midjourney-generated [14] image, respectively.



(a) Wukong generated query image

(b) Stable Diffusion V1.4 generated query image

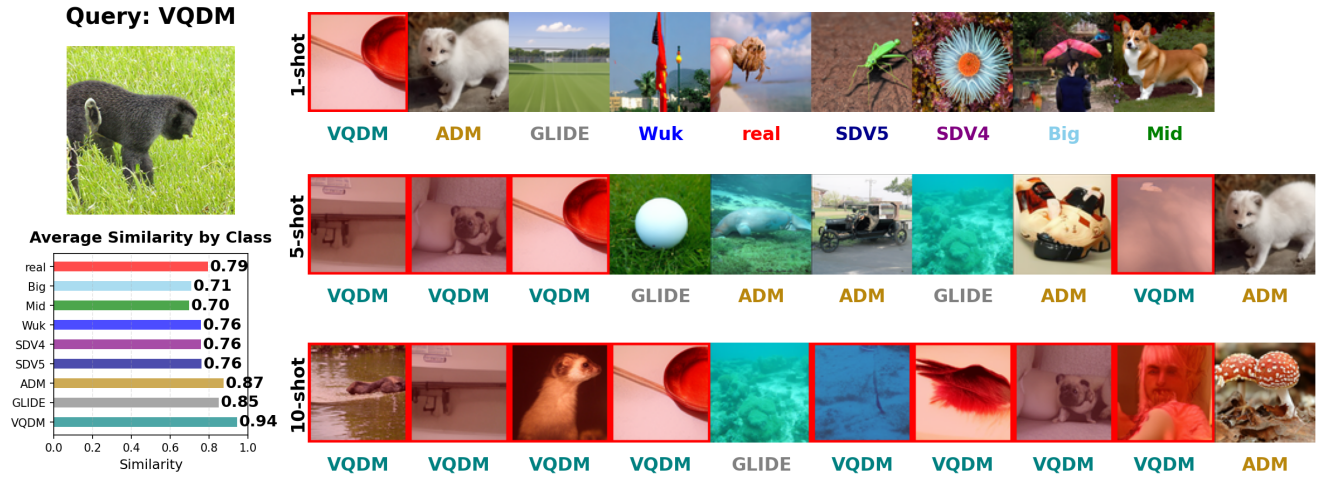


(c) Stable Diffusion V1.5 generated query image

Figure 2. **Illustration of retrieval-based image attribution.** The top-ranked retrieval results from the database are listed, with the correct result highlighted by a red bounding box. The queries are a Wukong-generated [24] image, a SDV4-generated image, and a SDV5-generated [17] image, respectively.

(a) ADM generated query image

(b) GLIDE generated query image



(c) VQDM generated query image

Figure 3. **Illustration of retrieval-based image attribution.** The top-ranked retrieval results from the database are listed, with the correct result highlighted by a red bounding box. The queries are a ADM-generated [5] image, a GLIDE-generated image [15], and a VQDM-generated [6] image, respectively.

B. Visualization

Attribution Visualization: We formulate AI-generated image attribution as an instance-retrieval problem, and representative visualization results are shown in Figure 1, 2, and 3. Specifically, we use a real image and various fake images in GenImage [25] dataset as queries, and rank the most similar images from the different-shot database, with the correctly retrieved result highlighted by a red box. The similarity scores between each query and all images from each generator in the database are visualized using bar charts. Even though Stable Diffusion V1.4 and Stable Diffusion V1.5 [17] share a similar core architecture, our method still achieves strong retrieval results.

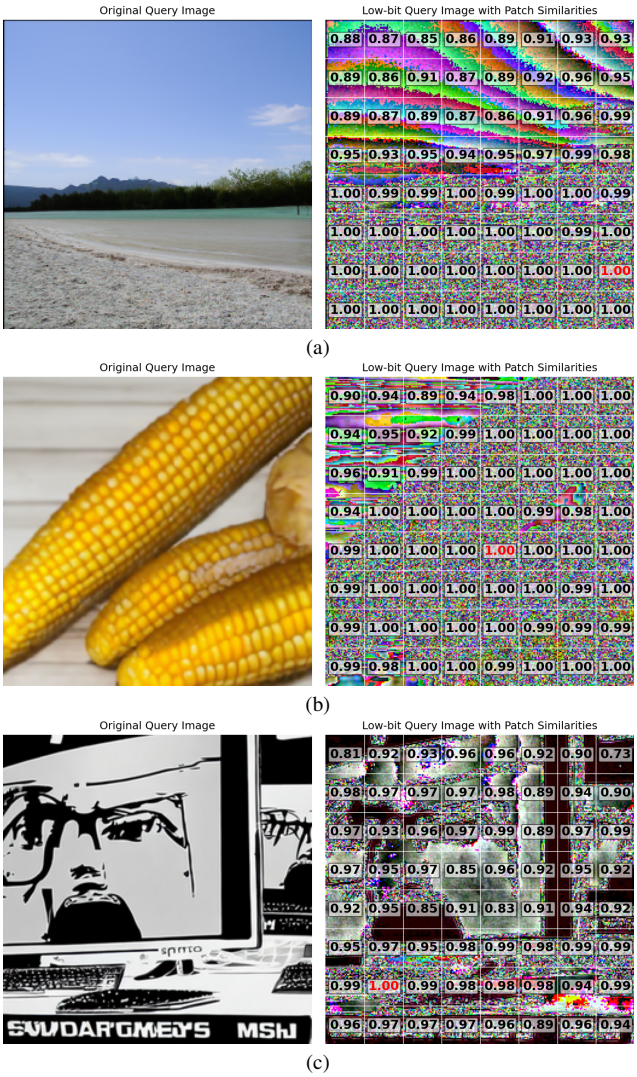


Figure 4. **Illustration of low-bit generative fingerprints.** Each image is divided into multiple patches, and the patch with the highest similarity score is selected for subsequent processing.

Fingerprints Visualization: We visualize the low-bit generative fingerprints of query, which generated by ADM [5], GLIDE [15] and Wukong [24] respectively, as illustrated in Figure 4. In practical implementation, each low-bit generative fingerprint is partitioned into 32×32 patches. For clarity of visualization, we present an 8×8 patch configuration. For each query patch, its similarity is computed against every patch of all images in the database, and the maximum similarity is assigned as the score for that patch. The patch achieving the highest score is subsequently forwarded to the encoder for feature extraction. This selection process ensures that the features carrying the most significant generator artifacts are used for attribution.

References

[1] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018. 1, 2

[2] Jiaxuan Chen, Jieteng Yao, and Li Niu. A single simple patch is all you need for ai-generated image detection. *arXiv preprint arXiv:2402.01123*, 2024. 1

[3] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8789–8797, 2018. 1

[4] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8188–8197, 2020. 1

[5] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 1, 4, 5

[6] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10696–10706, 2022. 1, 4

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 1

[8] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1

[9] Yan Hong, Jianming Feng, Haoxing Chen, Jun Lan, Huijia Zhu, Weiqiang Wang, and Jianfu Zhang. Wildfake: A large-scale and hierarchical dataset for ai-generated images detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3500–3508, 2025. 1

[10] Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. Scaling up gans for text-to-image synthesis. In *Proceedings of*

the IEEE/CVF conference on computer vision and pattern recognition, pages 10124–10134, 2023. 1

[11] 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

[12] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020.

[13] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *Advances in neural information processing systems*, 34:852–863, 2021. 1

[14] Midjourney. https://www.midjourney.com/home/, 2022.5. 1, 2

[15] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 4, 5

[16] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 1

[17] 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, 3, 5

[18] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022. 1

[19] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 1

[20] Ming Tao, Hao Tang, Fei Wu, Xiao-Yuan Jing, Bing-Kun Bao, and Changsheng Xu. Df-gan: A simple and effective baseline for text-to-image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16515–16525, 2022. 1

[21] Ming Tao, Bing-Kun Bao, Hao Tang, and Changsheng Xu. Galip: Generative adversarial clips for text-to-image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14214–14223, 2023. 1

[22] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 1

[23] 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

[24] Wukong. https://xihe.mindspore.cn/, 2022.5. 3, 5

[25] Mingjian Zhu, Hanting Chen, Qiangyu Yan, Xudong Huang, Guanyu Lin, Wei Li, Zhijun Tu, Hailin Hu, Jie Hu, and Yunhe Wang. Genimage: A million-scale benchmark for detecting ai-generated image. *Advances in Neural Information Processing Systems*, 36:77771–77782, 2023. 5