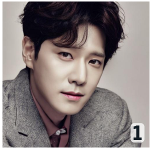
Prefilled responses enhance zero-shot detection of AI-generated images

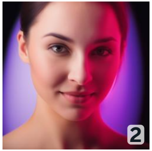
Zoher Kachwala Danishjeet Singh Danielle Yang Filippo Menczer

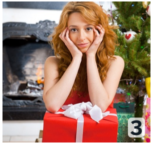
December 5, 2025

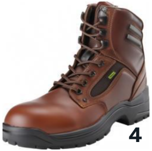
Observatory on Social Media (OSoMe), Indiana University

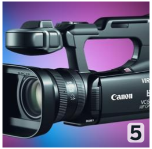
Can You Spot the Synthetic Images?


1

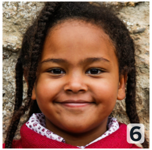

2

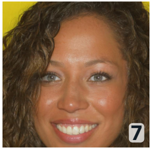

3

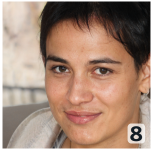

4

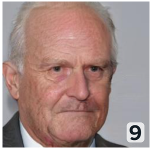


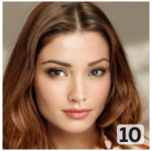
5

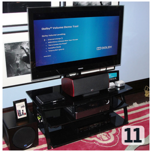

6

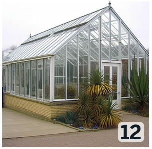

7

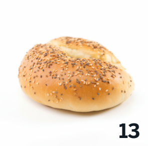

8

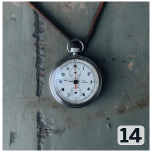

9

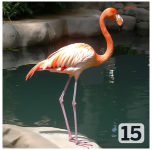

10


11

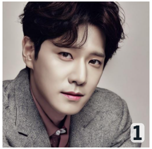

12

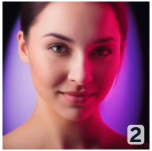

13

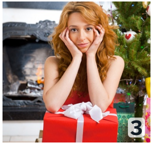

14

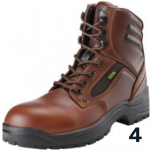

15

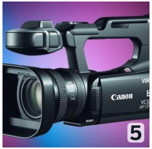
Can You Spot the Synthetic Images?


1

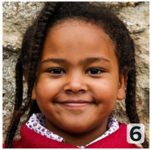

2

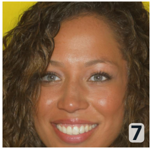

3

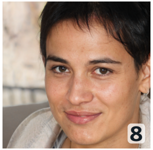

4

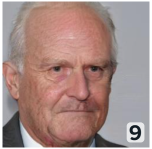


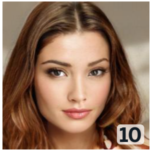
5

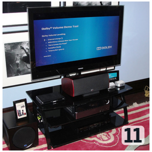

6


7

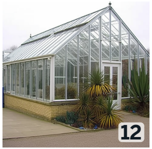

8

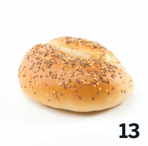

9

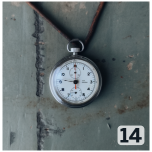

10

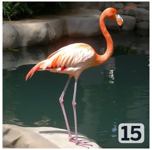


11


12


13


14


15

Images 3, 10, and 11 are real.

Risks of Synthetic Images

- **Image generation quality:** Modern techniques produce synthetic images nearly indistinguishable from real ones [1, 2, 3]
- **Enables malicious use:** Convincing forgeries facilitate impersonation, copyright infringement, and disinformation [4, 5, 6]
- **Threatens societal trust:** Widespread synthetic media undermines visual trust and information integrity [7, 8, 9, 10]

Urgent need for robust detection methods to maintain trust in visual media

Related Work: Supervised Detection

Three main approaches for detecting AI-generated images:

Artifact-based

CNNs/ViTs detect unnatural textures, edges [11, 12]

✗Overfits to specific generators [13, 14]

Frequency-domain

FFT/DCT analyze spectral patterns [15, 16]

✗Fails on diffusion models [17, 18]

Spatial-domain

Examines pixel patterns, structures [19, 20]

✗Fails on photorealistic outputs [21]

✗Sensitive to compression and resizing [22]

Supervised methods struggle with new generators.

Related Work: VLMs and Prompting

Large Language Models:

- Trained on diverse reasoning tasks [23]
- Step-by-step reasoning enables adaptation to new tasks [24, 25]

Vision-Language Models:

- Trained on short image-text tasks [26]
- Poorer step-by-step reasoning leads to weaker adaptation [27, 28]

My Approach: Prefill-Guided Thinking

Goal:

- Improve VLM’s step-by-step reasoning, specifically for detection of synthetic images without any training

Key Idea:

- Instead of user-prompting, *prefilled responses* guide VLM’s step-by-step reasoning for detection of synthetic images

Synthetic Image Detection: Task

Ground Truth: AI-generated

Task: Binary classification

- Input: Image
- Output: Real or AI-generated

Evaluation: Macro F1

- Average of F1 scores for both classes
- Balances performance on both classes

Synthetic Image Detection: Datasets & Models

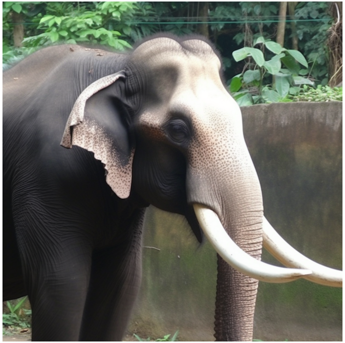
Datasets:

- D3 (first row): 6,736 web images across domains from 4 generators [18]
- DF40 (middle row): 10,000 faces from 6 generators [29]
- GenImage (bottom row): 10,000 ImageNet objects from 8 generators [30]

Models:

- Qwen2.5-VL-7B-Instruct
- LLaVA-OneVision-7B
- Llama-3.2-Vision-11B-Instruct

Method: Baseline



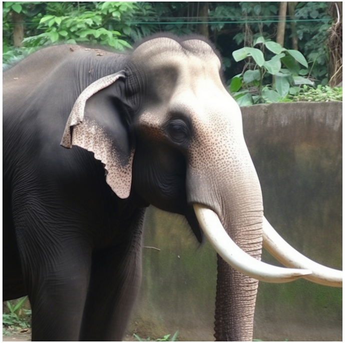
User: [Image] Is this image real or AI-generated?

Assistant: ***The image appears to be real.*** The details in the elephant's skin texture, the natural lighting, and the background vegetation suggest that this is a photograph of a real elephant in a natural or zoo-like setting. There are no visible signs of digital manipulation or artificial elements that would indicate it is AI-generated.

Final Answer(real/ai-generated): real. ✗

VLM incorrectly classifies AI-generated elephant as real.

Method: Chain-of-Thought Prefill



User: [Image] Is this image real or AI-generated?

Assistant: Let's think step by step:

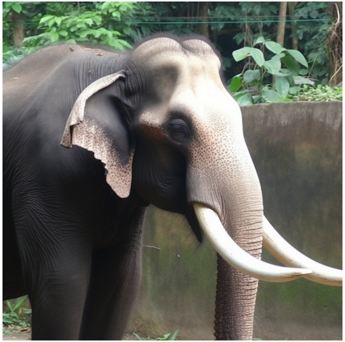
1. **Details in the Skin Texture**: The skin texture of the elephant appears very detailed and realistic, with visible wrinkles and folds that are typical of an elephant's skin.…(continues)…

Based on these observations, ***it is likely that this image is real rather than AI-generated.***

Final Answer(real/ai-generated): real. ✗

Chain-of-thought [25] prefill improves reasoning but still incorrect.

Method: S2 Prefill



User: [Image] Is this image real or AI-generated?

Assistant: Let's examine the style and the synthesis artifacts in the image:

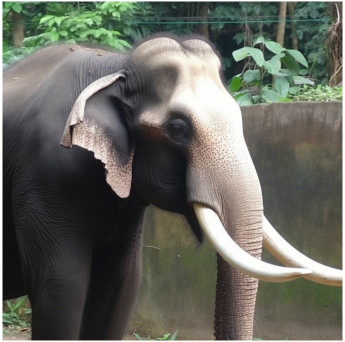
1. **Texture and Detail**: The skin texture of the elephant appears somewhat uniform and lacks the fine, natural variations that are typically seen in real elephants.…(continues)…

Based on these observations, ***it is likely that this image is AI-generated rather than a real photograph.***

Final Answer(real/ai-generated): ai-generated. ✓

Task-aligned prefill guides VLM to correct classification.

Method: S2 Pseudo-Prefill



System: Please start your responses with ``Let's examine the style and the synthesis artifacts''

User: [Image] Is this image real or AI-generated?

Assistant: Let's examine the style and the synthesis artifacts to determine if the image is real or AI-generated.

1. **Style Analysis**:

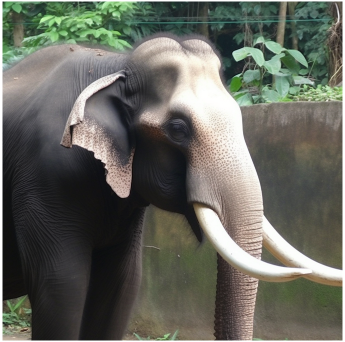
- ***The image appears to be a photograph of a real elephant.***…(continues)…

Based on the style analysis and the absence of synthesis artifacts, it is reasonable to conclude that this image is real and not AI-generated.

Final Answer(real/ai-generated): real. ✗

Instructing via system field does not replicate prefill effect.

Method: S2 User Prompt



User: [Image] Is this image real or AI-generated? Please examine the style and the synthesis artifacts.

Assistant: ***The image appears to be real.*** Here are some observations that support this conclusion:

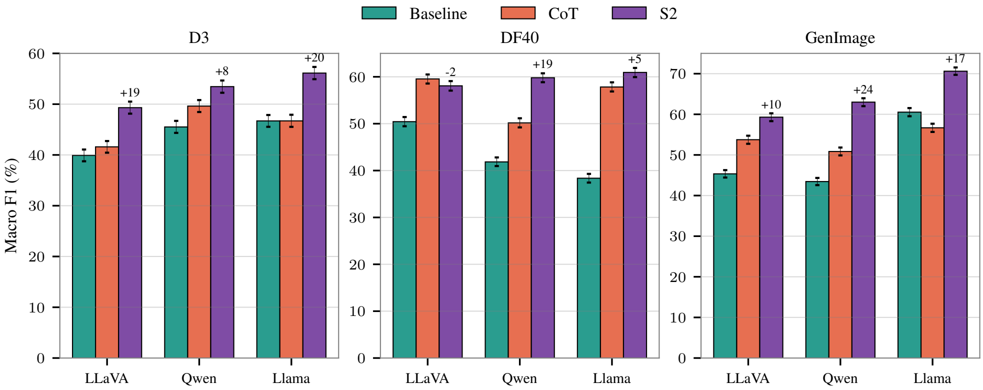
1. **Texture and Detail**: The skin texture of the elephant is highly detailed and realistic.…(continues)…

Overall, based on these observations, the image appears to be real rather than AI-generated.

Final Answer(real/ai-generated): real. ✗

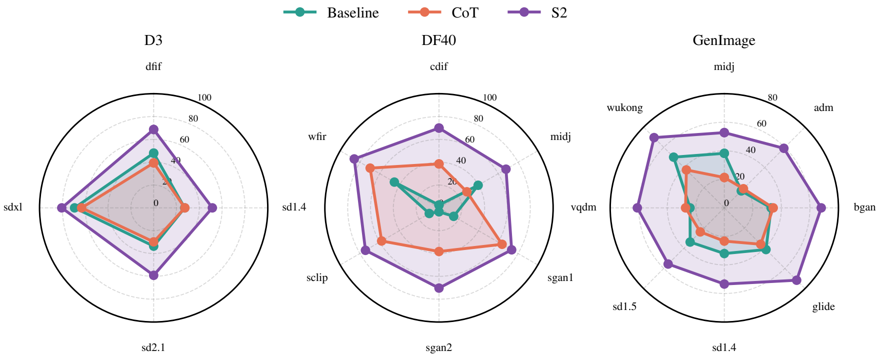
User prompt with task-aligned phrasing also fails.

Results: Detection Performance (Macro F1 %)



With the exception of LLaVA on DF40, S2 (task-aligned prefill) consistently outperforms CoT and baseline across models and datasets, with up to 24% relative improvement.

Results: Per-Generator Recall (Llama)



S2 prefill achieves 15-200% recall improvement across 16 state-of-the-art generators, demonstrating strong generalization.

Results: Guidance Comparison

Phrase	Type	D3	DF40	GenImage
Baseline	—	46.7	42.3	44.5
CoT	Prefill	49.8	48.8	53.3
	Pseudo-Prefill	48.9 (-0.9)	44.1 (-4.7)	48.8 (-4.5)
	Prompt	45.5 (-4.3)	46.7 (-2.1)	50.4 (-2.9)
S2	Prefill	**53.2**	**61.2**	**64.8**
	Pseudo-Prefill	43.4 (-9.8)	45.2 (-16.0)	47.3 (-17.5)
	Prompt	49.7 (-3.5)	48.8 (-12.4)	50.2 (-14.6)

Macro F1 (%) for Qwen across guidance types. Direct prefilling works best.

Results: Prefill Phrasing Sensitivity

Phrase	**Prefill**	**D3**	**DF40**	**GenImage**
Baseline	—	46.7	42.3	44.5
CoT	`Let's think step by step`	49.8	48.8	53.3
S2	`Let's examine the style and synthesis artifacts`	53.2	61.2	**64.8**
Variants	`Let's observe the style and synthesis artifacts`	55.6 (+2.4)	**62.3** (+1.1)	64.6 (-0.2)
	`Let's examine the synthesis artifacts`	54.6 (+1.4)	60.9 (-0.3)	64.0 (-0.8)
	`Let's examine the style`	48.5 (-4.7)	55.7 (-5.5)	54.6 (-10.2)
	`Let's examine the details`	49.2 (-4.0)	50.9 (-10.3)	51.2 (-13.6)
	`Let's examine the flaws`	**58.3** (+5.1)	40.0 (-21.2)	54.1 (-10.7)

Macro F1 (%) for Qwen. Any variation of S2, always performs better than the baseline.

Why Do Prefills Work?

Experiment: Measure class probabilities at intervals during reasoning

- **Probe:** `Final Answer(real/ai-generated): This image is`
- **Intervals:** 0%, 25%, 50%, 75%, 100% of reasoning (by sentence count)
- **Metric:** Maximum class probability (class confidence)

Question: How do prefills affect the reasoning process and confidence, compared to baseline and prompts?

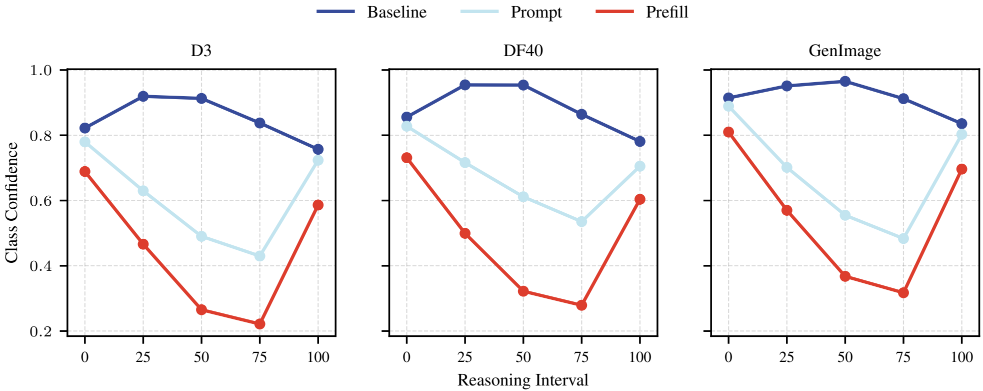
Example: Reasoning Intervals (Baseline)

`User: [Image] Is this image real or AI-generated?`

0%	25%	50%	75%	100%
Assistant:	Assistant: The image appears to be real.	Assistant: The image appears to be real. The details in the elephant's skin texture, the natural lighting, and the background vegetation	Assistant: The image appears to be real. The details in the elephant's skin texture, the natural lighting, and the background vegetation suggest that this is a photograph of a real elephant in a natural or zoo-like setting.	Assistant: The image appears to be real. The details in the elephant's skin texture, the natural lighting, and the background vegetation suggest that this is a photograph of a real elephant in a natural or zoo-like setting. There are no visible signs of digital manipulation or artificial elements that would indicate it is AI-generated.

At each interval, we probe with: `Final Answer(real/ai-generated): This image is` and extract the class probabilities.

Results: Confidence Changes During Reasoning (Qwen)



Baseline produces overconfident reasoning, while prompts and prefills induce humility toward class labels—mirroring the Dunning-Kruger effect [31].

Contribution

Prefill-Guided Thinking

- A new way to elicit step-by-step reasoning from vision-language models for zero-shot detection of AI-generated images.
- **Up to 24% MacroF1 improvement** across 16 state-of-the-art generators
- **No training data** required
- **Higher humility** about class labels during reasoning leads to more accurate detection

Under Review: ACL ARR

References i

[1] Robin Rombach et al. “High-Resolution Image Synthesis with Latent Diffusion Models”. en. In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. New Orleans, LA, USA: IEEE, June 2022, pp. 10674–10685.

[2] Junnan Li et al. “BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models”. en. In: *Proceedings of the 40th International Conference on Machine Learning*. ISSN: 2640-3498. PMLR, July 2023, pp. 19730–19742.

[3] Prafulla Dhariwal and Alexander Nichol. “Diffusion Models Beat GANs on Image Synthesis”. In: *Advances in Neural Information Processing Systems*. Vol. 34. Curran Associates, Inc., 2021, pp. 8780–8794.

[4] Lingzhi Li et al. “Advancing High Fidelity Identity Swapping for Forgery Detection”. en. In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Seattle, WA, USA: IEEE, June 2020, pp. 5073–5082.

References ii

[5] Andreas Lugmayr et al. “RePaint: Inpainting using Denoising Diffusion Probabilistic Models”. en. In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. New Orleans, LA, USA: IEEE, June 2022, pp. 11451–11461.

[6] Zeyu Lu et al. “Seeing is not always believing: Benchmarking Human and Model Perception of AI-Generated Images”. en. In: *Advances in Neural Information Processing Systems* 36 (Dec. 2023), pp. 25435–25447.

[7] Kaicheng Yang, Danishjeet Singh, and Filippo Menczer. “Characteristics and Prevalence of Fake Social Media Profiles with AI-generated Faces”. en. In: *Journal of Online Trust and Safety* 2.4 (Sept. 2024). Number: 4.

[8] Gowthami Somepalli et al. “Diffusion Art or Digital Forgery? Investigating Data Replication in Diffusion Models”. en. In: *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Vancouver, BC, Canada: IEEE, June 2023, pp. 6048–6058.

References iii

[9] Renée DiResta and Josh A. Goldstein. “How spammers and scammers leverage AI-generated images on Facebook for audience growth”. en-US. In: *Harvard Kennedy School Misinformation Review* (Aug. 2024).

[10] Bilva Chandra. *Analyzing Harms from AI-Generated Images and Safeguarding Online Authenticity*. en. RAND Corporation, 2024.

[11] Darius Afchar et al. “MesoNet: a Compact Facial Video Forgery Detection Network”. In: *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*. ISSN: 2157-4774. Dec. 2018, pp. 1–7.

[12] François Chollet. “Xception: Deep Learning with Depthwise Separable Convolutions”. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. ISSN: 1063-6919. July 2017, pp. 1800–1807.

References iv

[13] Sheng-Yu Wang et al. "CNN-Generated Images Are Surprisingly Easy to Spot... for Now". English. In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, June 2020, pp. 8692–8701.

[14] Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. "Towards Universal Fake Image Detectors that Generalize Across Generative Models". en. In: *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Vancouver, BC, Canada: IEEE, June 2023, pp. 24480–24489.

[15] Hanzhe Li et al. "FreqBlender: Enhancing DeepFake Detection by Blending Frequency Knowledge". en. In: *Advances in Neural Information Processing Systems* 37 (Dec. 2024), pp. 44965–44988.

[16] Tarik Dzanic, Karan Shah, and Freddie Witherden. "Fourier Spectrum Discrepancies in Deep Network Generated Images". In: *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc., 2020, pp. 3022–3032.

References v

[17] Davide Cozzolino et al. “Raising the Bar of AI-generated Image Detection with CLIP”. en. In: *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. Seattle, WA, USA: IEEE, June 2024, pp. 4356–4366.

[18] Lorenzo Baraldi et al. “Contrasting Deepfakes Diffusion via Contrastive Learning and Global-Local Similarities”. In: *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part LXIII*. Berlin, Heidelberg: Springer-Verlag, Nov. 2024, pp. 199–216.

[19] Chende Zheng et al. “Breaking Semantic Artifacts for Generalized AI-generated Image Detection”. en. In: *Advances in Neural Information Processing Systems* 37 (Dec. 2024), pp. 59570–59596.

[20] Peng Zhou et al. “Learning Rich Features for Image Manipulation Detection”. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. ISSN: 2575-7075. June 2018, pp. 1053–1061.

References vi

[21] Andreas Rössler et al. “FaceForensics++: Learning to Detect Manipulated Facial Images”. In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. ISSN: 2380-7504. Oct. 2019, pp. 1–11.

[22] Quentin Bammey. “Synthbuster: Towards Detection of Diffusion Model Generated Images”. In: *IEEE Open Journal of Signal Processing* 5 (2024), pp. 1–9.

[23] Tom B. Brown et al. “Language models are few-shot learners”. In: *Proceedings of the 34th International Conference on Neural Information Processing Systems*. NIPS ’20. Red Hook, NY, USA: Curran Associates Inc., Dec. 2020, pp. 1877–1901.

[24] Jason Wei et al. “Chain-of-Thought Prompting Elicits Reasoning in Large Language Models”. en. In: *Advances in Neural Information Processing Systems* 35 (Dec. 2022), pp. 24824–24837.

References vii

[25] Takeshi Kojima et al. “Large Language Models are Zero-Shot Reasoners”. en. In: *Advances in Neural Information Processing Systems* 35 (Dec. 2022), pp. 22199–22213.

[26] Alec Radford et al. “Learning Transferable Visual Models From Natural Language Supervision”. en. In: *Proceedings of the 38th International Conference on Machine Learning*. ISSN: 2640-3498. PMLR, July 2021, pp. 8748–8763.

[27] Ruohong Zhang et al. “Improve Vision Language Model Chain-of-thought Reasoning”. In: *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Wanxiang Che et al. Vienna, Austria: Association for Computational Linguistics, July 2025, pp. 1631–1662.

References viii

[28] Yangyi Chen et al. “Measuring and Improving Chain-of-Thought Reasoning in Vision-Language Models”. In: *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. Ed. by Kevin Duh, Helena Gomez, and Steven Bethard. Mexico City, Mexico: Association for Computational Linguistics, June 2024, pp. 192–210.

[29] Zhiyuan Yan et al. “DF40: Toward Next-Generation Deepfake Detection”. en. In: *Advances in Neural Information Processing Systems* 37 (Dec. 2024), pp. 29387–29434.

[30] Mingjian Zhu et al. “GenImage: A Million-Scale Benchmark for Detecting AI-Generated Image”. en. In: *Advances in Neural Information Processing Systems* 36 (Dec. 2023), pp. 77771–77782.

[31] J. Kruger and D. Dunning. “Unskilled and unaware of it: how difficulties in recognizing one’s own incompetence lead to inflated self-assessments”. eng. In: *Journal of Personality and Social Psychology* 77.6 (Dec. 1999), pp. 1121–1134.