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

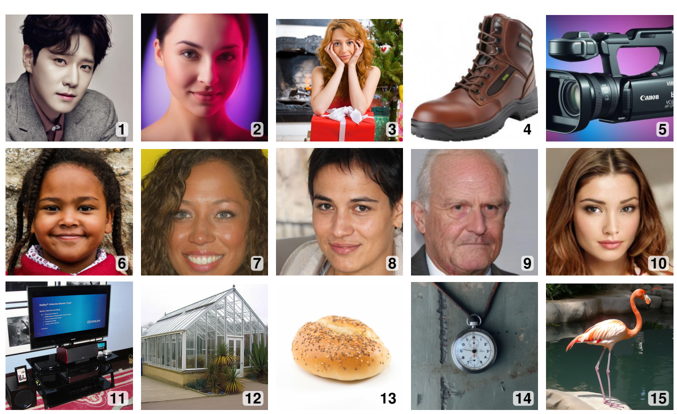
Zoher Kachwala* **Danishjeet Singh** **Danielle Yang** **Filippo Menczer**
Observatory on Social Media
Indiana University, Bloomington, USA

Abstract

As AI models generate increasingly realistic images, growing concerns over potential misuse underscore the need for reliable detection. Traditional supervised detection methods depend on large, curated datasets for training and often fail to generalize to novel, out-of-domain image generators. As an alternative, we explore pre-trained Vision-Language Models (VLMs) for zero-shot detection of AI-generated images. We evaluate VLM performance on three diverse benchmarks encompassing synthetic images of human faces, objects, and animals produced by 16 different state-of-the-art image generators. While off-the-shelf VLMs perform poorly on these datasets, we find that their reasoning can be guided effectively through simple response prefilling — a method we call Prefill-Guided Thinking (**PGT**). In particular, prefilling a VLM response with the task-aligned phrase "*Let's examine the style and the synthesis artifacts*" improves the Macro F1 scores of three widely used open-source VLMs by up to 24%. Our code is publicly available at: `https://github.com/Zoher15/Zero-shot-pgt`

*Corresponding author: `zkachwal@iu.edu`

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: The First Workshop on Generative and Protective AI for Content Creation.



User: [Image] Is this image real or AI-generated?	User: [Image] Is this image real or AI-generated?
Assistant: Let's think step by step	Assistant: Let's examine the style and synthesis artifacts

Figure 1: Top: Sample images from D3 (top row), DF40 (middle row), and GenImage (bottom row) datasets. Can you guess which ones are real? The answer is in the footnote on the next page. Bottom: Guiding model thinking with prefilled responses: chain-of-thought (left) vs task-aligned (right).

1 Introduction

Rapid advancements in image generation have led to a surge in synthetic images (deepfakes) [1, 2, 3]. Improved techniques now enable the easier and cheaper production of high-quality visuals [3, 4, 5]. While beneficial for creative applications, this progress has allowed malicious actors to produce convincing forgeries (e.g., face swaps, synthetic photos) that are nearly indistinguishable from real ones [6, 7, 8]. Such forgeries facilitate impersonation, copyright infringement, and disinformation, highlighting the urgent need for robust detection methods to maintain visual trust [9, 10, 11, 12].

Existing solutions remain limited. Watermarking and metadata approaches are often easily bypassed and require widespread adoption [13]. Supervised methods, particularly feature-based ones, struggle to generalize to new generators [14].

To keep pace with these rapid developments, we investigate the use of pre-trained Vision-Language Models (VLMs), which have demonstrated strong generalization across diverse tasks [15, 16, 4] and whose performance can be further enhanced through prompting [17, 18].

We find that off-the-shelf VLMs perform poorly on the zero-shot detection of AI-generated images. However, prefilling a model response can guide more focused reasoning and significantly improve detection without any fine-tuning. We call this approach Prefill-Guided Thinking (**PGT**). In particular, prefilling the model response with the phrase *"Let's examine the style and the synthesis artifacts"* encourages the VLM to attend more closely to forensic cues relevant for this task.

We evaluate **PGT** on three recent, diverse datasets spanning human faces, objects, and animals, with images generated by 16 state-of-the-art image models (Fig. 1[2]). For three widely used open-source VLMs, our approach boosts Macro F1 scores by up to 24%, demonstrating strong generalization across image categories as well as new generators.

2 Background

Supervised detection methods for AI-generated images typically fall into three categories: artifact-based, frequency-domain, and spatial-domain approaches.

Artifact-based methods use Convolutional Neural Networks (CNNs) or Vision Transformers (ViTs) to detect subtle cues such as unnatural textures or edge inconsistencies [19, 20, 21]. As generative models improve, these cues become less reliable. Models trained on fixed artifacts often overfit to specific generators, leading to poor generalization [22, 23, 24].

Frequency-domain techniques analyze spectral representations using techniques like Fast Fourier Transform or Discrete Cosine Transform [14, 25, 26, 27]. These methods were effective against early Generative Adversarial Networks (GANs), but newer diffusion models exhibit different frequency characteristics, reducing the utility of fixed-frequency detectors [23, 28, 29].

Spatial-domain approaches examine raw pixel patterns to detect structural or textural inconsistencies [29, 30]. While useful for earlier synthetic images, they often fail on photorealistic outputs from diffusion models and are sensitive to post-processing such as compression and resizing [20, 31].

To improve generalization, recent work has explored larger and more diverse training datasets [22, 32, 33, 34, 35, 36, 37], as well as architectural modifications that target universal artifacts (e.g., upsampling patterns) [29, 38]. Approaches like fixed-feature backbones and DiffusionFake have also improved robustness to novel generators [24, 39].

Vision-Language Models (VLMs) are trained on large-scale image-text datasets and demonstrate strong zero-shot generalization across tasks such as classification, captioning, and visual question answering [4, 15, 16]. Prompt-based learning has emerged as a lightweight method to adapt language models using natural language instructions [40]. Chain-of-thought, which encourages step-by-step reasoning, has been effective in multimodal settings [41, 42, 43].

Building on this, we investigate whether prefilling VLM responses can support generalizable, zero-shot detection of AI-generated images. We hypothesize that this approach could offer an alternative to

[2] Images 3, 10, and 11 are the only real ones in Fig. 1.

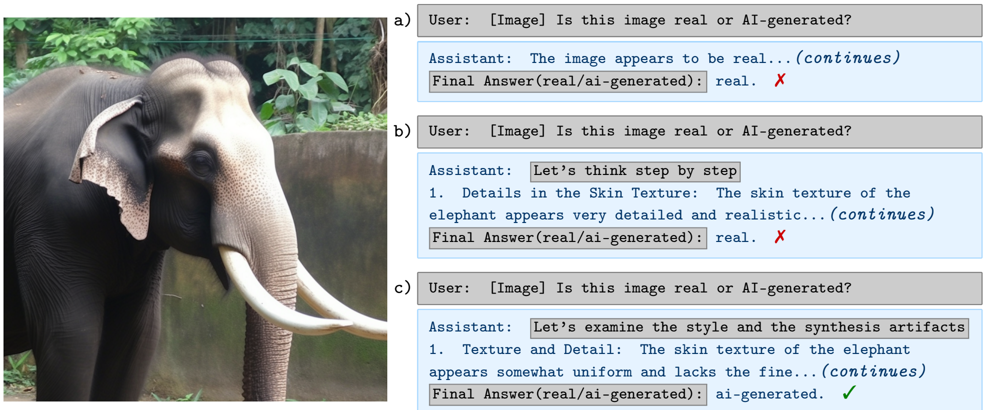


Figure 2: Illustration of **PGT** for the detection of an AI-generated image using a VLM (Qwen2.5-7B). Input text is marked in `grey`, response text in `blue`. (a) A baseline user query results in the incorrect response `real`. (b) Using the chain-of-thought prefill `Let's think step by step` improves reasoning, but the classification remains incorrect. (c) Using our proposed **S2** prefill `Let's examine the style and the synthesis artifacts` leads to the correct classification: `ai-generated`. Full reasoning traces for all three methods can be found in the Appendix.

training for synthetic image detection, with greater capacity to generalize to increasingly sophisticated generative models.

3 Methods

We frame the detection of AI-generated images as a binary classification task: given an image, the goal is to determine whether it is *real* or *AI-generated*. To evaluate overall performance, we use the Macro F1 score, which is robust to class imbalance. To analyze performance across different generators, we additionally report per-generator recall within the *AI-generated* class.

3.1 Prefill-Guided Thinking

Vision-Language Models typically consist of a *system* field for general instructions, a *user* field for inputs, and an *assistant* field for model-generated responses. A *prefill* is a phrase inserted into the *assistant* field, intended to guide the model's response.

In our experiment, an image is presented to a VLM along with a question in the user field: `User: [Image] Is this image real or AI-generated?`. Consistent with instruction-tuned model behavior, the VLM typically generates free-form reasoning in the assistant field (e.g., `Assistant: This image appears to be...`) (Fig. 2a). To obtain a clean label, we query the model a second time by inserting a phrase into the assistant field after its first response: `Final Answer(real/ai-generated):`. This is our *baseline* method.

The chain-of-thought (*CoT*) variant encourages step-by-step reasoning. It uses the same initial question in the user field and inserts the phrase *"Let's think step by step"* into the assistant field [42]. The model's prefilled responses then start with: `Assistant: Let's think step by step` (Fig. 2b). The final clean label is again elicited using the same follow-up answer phrase in the assistant field.

Building on prior work that highlights the importance of synthesis artifacts in detection [44], we introduce **S2**, a task-aligned prefill. We insert the phrase *"Let's examine the style and the synthesis artifacts"* into the assistant's field (Fig. 2c), creating a prefilled response that starts with: `Assistant: Let's examine the style and the synthesis artifacts`. This framing encourages attention to perceptual cues, such as stylistic inconsistencies or generation artifacts, grounding the model's reasoning in visual rather than semantic features. The final classification label is obtained in the same way as in the baseline.

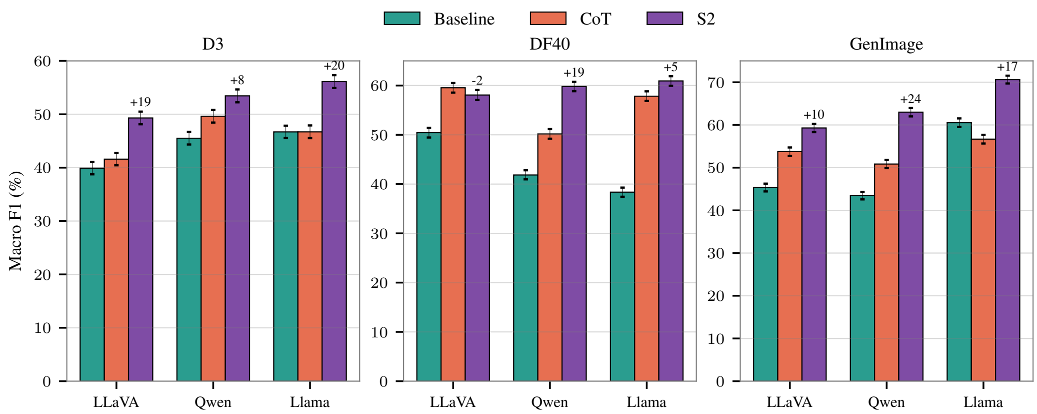


Figure 3: Detection Macro F1 across models, datasets, and **PGT** variations. Bars are annotated with relative improvements of **S2** over the next best method and 95% confidence error bars from 10k bootstrap iterations.

We also evaluate the effects of prefill variations on detection performance, as well as approaches where the same phrases are inserted into the user field (traditional user prompting) or in the system field.

3.2 Data

We conduct experiments using three state-of-the-art benchmarks that span a broad spectrum of real and AI-generated images.

D3 is a benchmark dataset introduced as part of the Contrastive Deepfake Embeddings framework [28]. Unlike many generative datasets focused on faces or curated categories, D3 comprises real images collected from the web, covering a wide range of domains, including objects, urban scenes, artwork, animals, abstract visuals, and human figures. Synthetic counterparts were generated using four models: DeepFloyd IF, Stable Diffusion v1.4 and v2.1 [3], and Stable Diffusion XL [45]. We randomly sampled 2,000 sets of five images (one real and four generated). After filtering for copyright restrictions and broken links, the final dataset contains 8,420 images (1,684 real and 6,736 generated). We use 80% of this data (1,344 real and 5,392 generated) for our main evaluation and reserve the remaining 20% (344 real and 1,344 generated) for additional experiments. We refer to the main evaluation set as D3, and the smaller subset as D3 (2k).

DF40 is a facial image dataset containing content generated by 40 deepfake techniques across four categories: face swapping, face reenactment, full-face synthesis, and facial editing [35]. It includes outputs from state-of-the-art models such as Collaborative Diffusion [46], Midjourney, StyleCLIP [47], StarGAN v1 and v2 [48, 49], and WhichFaceIsReal. The dataset spans variations in age, gender, ethnicity, and facial pose. We randomly sample 10,000 images (3,929 real and 6,071 generated) for the main evaluation, and an additional 2,000 images (794 real and 1,206 generated) for extended experiments. We refer to the main evaluation set as DF40, and the smaller subset as DF40 (2k).

GenImage is built using all real images from ImageNet [50], covering diverse object categories such as animals, tools, vehicles, and furniture. Each real image is paired with a synthetic counterpart generated by one of eight models: ADM [5], BigGAN [51], GLIDE [52], Midjourney, VQDM [53], Stable Diffusion v1.4 and v1.5 [3], and Wukong. We use a balanced evaluation sample of 10,000 images (5,000 real and 5,000 generated) for the main experiments and 2,000 images (1,000 real and 1,000 generated) for extended analyses. We refer to the main evaluation set as GenImage, and the smaller subset as GenImage (2k).

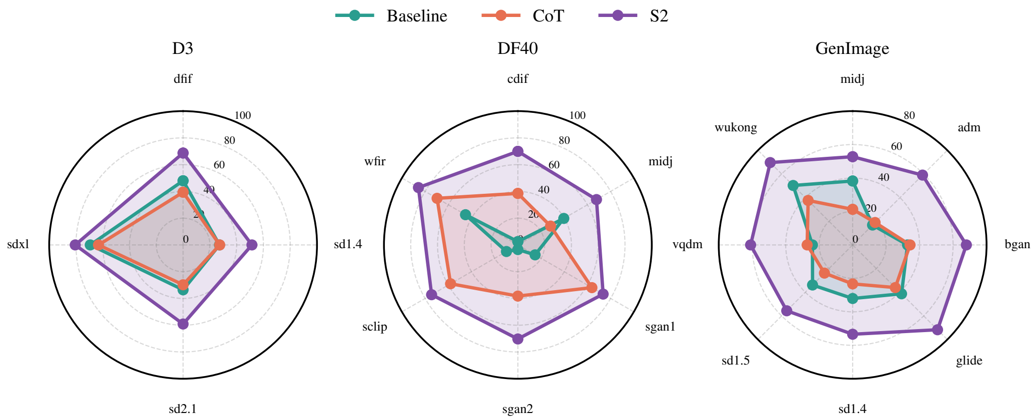


Figure 4: Detection recall (%) for Llama across different datasets and their state-of-the-art synthetic image generators. Similar figures for LLaVA and Qwen are in the Appendix (Figs. 9, 10).

Phrase	**Type**	**D3 (2k)**	**DF40 (2k)**	**GenImage (2k)**
Baseline	—	46.7	42.3	44.5
CoT	Prefill	49.8	48.8	53.3
	Pseudo-Prefill	48.9 (-1.0)	44.1 (-4.7)	48.8 (-4.5)
	Prompt	45.5 (-4.4)	46.7 (-2.1)	50.4 (-2.8)
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 (-14.5)

Table 1: Macro F1 scores (%) across datasets and phrase modes for Qwen. Values in parentheses indicate absolute differences compared to the prefill baselines. Tables for LLaVA and Llama are shown in the Appendix.

3.3 Models

We conduct experiments with three Vision-Language Models: *LLaVA-OneVision*[3], *Qwen2.5-VL*[4], and *Llama-3.2-Vision*[5]. We use instruction-tuned variants in evaluation mode and use `seed=0` in the vLLM package [54] to ensure reproducibility. Token generation is capped at 512 tokens (`max_tokens=512`). A single run on a NVIDIA A100 80GB for 10,000 images takes approximately 3 hours.

LLaVA-OneVision has been trained on multimodal instruction-following data to allow transfer learning across different modalities and scenarios. We use the 7B, chat variant of the model for the main evaluation.

Qwen2.5-VL employs a native dynamic-resolution Vision Transformer trained from scratch using Window Attention [55]. For the main evaluation, we use the 7B-Instruct variant.

Llama-3.2-Vision incorporates a separately trained vision adapter that interfaces with a pre-trained Llama 3.1 language model. We use the 11B-Instruct version for our main evaluation.

[3] `https://huggingface.co/collections/llava-hf/llava-onevision-66bb1e9ce8856e210a7ed1fe`
[4] `https://huggingface.co/collections/Qwen/qwen25-vl-6795ffac22b334a837c0f9a5`
[5] `https://huggingface.co/collections/meta-llama/llama-32-66f448ffc8c32f949b04c8cf`

Phrase	Prefill	D3 (2k)	DF40 (2k)	GenImage (2k)
Baseline	—	46.8	42.3	44.5
CoT	`Let's think step by step`	50.0	48.8	53.4
S2	`Let's examine the style and the synthesis artifacts`	53.7	59.5	64.5
Variants	`Let's observe the style and the synthesis artifacts`	55.6 (+2.0)	62.4 (+2.8)	64.6 (+0.1)
	`Let's examine the synthesis artifacts`	54.6 (+0.9)	60.9 (+1.4)	64.0 (-0.6)
	`Let's examine the style`	48.6 (-5.1)	55.8 (-3.8)	54.6 (-9.9)
	`Let's examine the details`	49.4 (-4.3)	50.9 (-8.6)	51.4 (-13.1)
	`Let's examine the flaws`	58.4 (+4.8)	40.0 (-19.6)	54.1 (-10.5)

Table 2: Macro F1 scores (%) of different prefill phrasings for Qwen. Variants are annotated to show absolute differences compared to **S2**.

4 Results

4.1 Detection Performance

With the exception of LLaVA on DF40, **S2** consistently outperforms *CoT* and the baseline across all three benchmarks and models (Fig. 3), achieving up to 24% relative improvement in Macro F1. Across the 16 state-of-the-art generators, the underlying model remains the same, yet **S2** reliably boosts detection recall at inference. The gains in recall range from 15% for DF40's sgan1 to 200% for GenImage's adm compared to the next best method for Llama (Fig. 4). Similar trends hold for the other two models (Figs. 9, 10 in Appendix). These results demonstrate that our task-aligned prefill generalizes effectively, outperforming chain-of-thought for AI-generated image detection across diverse models, datasets, and generators.

4.2 Guidance Comparison

We also compare prefill-guided thinking against simple user prompts that combine `User: [Image] Is this image real or AI-generated?` with `Please think step by step.` (*CoT*) and `Please examine the style and synthesis artifacts.` (**S2**). Additionally, we investigated the effect of using the system field to instruct the model using `System: Please start your responses with` followed by `"Let's think step by step"` and `"Let's examine the style and synthesis artifacts"`. We refer to these variants as Pseudo-Prefills.

As shown in Table 1, user prompts and pseudo-prefills do not work as well as prefill-guided thinking for *CoT* and **S2** using Qwen. Similar trends are observed using LLaVA, whereas user prompts do work better with Llama (Tables 3, 4 in Appendix).

4.3 Prefill Phrasing

Here we examine the model's sensitivity to prefill phrasing variations for guiding the model to detect synthetic images. As shown in Table 2, Qwen's detection performance varies depending on how the prefill is worded. Guiding the model to consider both style and synthesis artifacts using alternative phrasing does not reduce performance relative to our **S2** prefill. In contrast, guiding the model to focus on only style or only synthesis artifacts leads to a drop in detection. Guidance around details and flaws also proves less effective than **S2**. While the specific phrasing of prefills affects the results, note that using any task-aligned prefill consistently improves peformance compared to the baseline.

4.4 Descriptive Analysis of Reasoning

Here, we analyze the top words used by the three methods in correctly identifying images. Responses are represented using a one-hot bag-of-words encoding over the vocabulary, after filtering out stop words and infrequent terms. For each method, we train a logistic regression model on responses from all three benchmarks to predict correctness (`1` for correct detections and `0` for incorrect). The trained regressions are then used to identify the 20 words that are most strongly associated with improved detection performance.

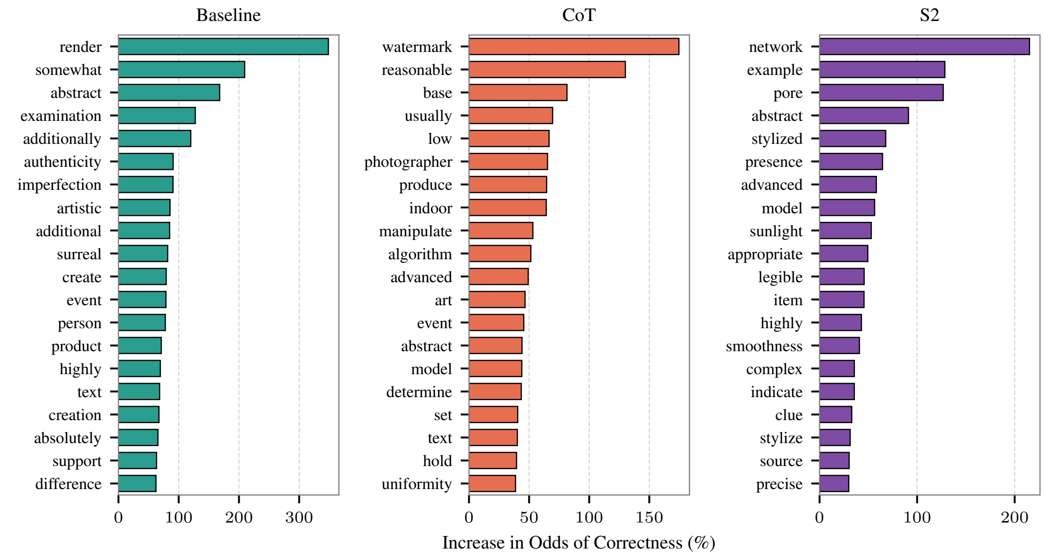


Figure 5: Top 20 words associated with the highest improvement in detection correctness for Qwen. Similar figures for LLaVA and Llama are in the Appendix.

For Qwen, the top words vary widely depending on the method (Fig. 5), and similar differences among methods are observed for Llama and LLaVA (Figs. 11, 12 in Appendix). Even with the same underlying model, the three zero-shot methods induce substantial differences in the vocabulary used for reasoning about the images.

5 Discussion

Our results show that prefilling responses for Vision-Language Models provides effective guidance for the detection of AI-generated images — without task-specific fine-tuning. We introduce a simple task-aligned prefill that consistently outperforms chain-of-thought across nearly all models and benchmarks, and generalizes robustly across 16 state-of-the art image generators.

We also find that responses must be prefilled directly, as instructing the model to begin its responses with the prefill does not reliably replicate the effect. For two of the three models we tested, prefilling responses also outperforms standard user prompts. As with prompts, careful phrasing is critical for effectiveness.

Overall, our findings suggest that prefilled guidance is a simple yet powerful approach for steering VLMs toward robust and generalizable synthetic image detection in a zero-shot setting.

6 Limitations

A primary limitation of using VLMs for synthetic image detection is their computational cost. Compared to traditional models like CNNs or ViTs, Vision-Language Models generate longer token sequences, require more memory, and consequently perform slowly. As a result, any method that leverages them inherits these computational demands.

While we found that Prefill-Guided Thinking improves synthetic image detection across three VLMs, we cannot assume it will generalize to all VLMs. Architectures vary widely, as do their training regimes. Moreover, our experiments were limited to mid-sized VLMs (7–11B parameters), so the effectiveness of **PGT** for larger models remains unknown. Finally, although we used state-of-the-art benchmarks from the past two years, the opaque nature of VLM training means we cannot exclude that the models were exposed to some of these images during pre-training.

Finally, while we identified a prefill suited to synthetic image detection, our finding that prefills are sensitive to phrasing means that the ease of discovering effective prefills for any given task remains an open question.

Acknowledgments

This work was supported by the Knight Foundation and the Luddy School of Informatics, Computing, and Engineering at Indiana University. We gratefully acknowledge NVIDIA for the GPU access that made this study possible. This work also used Jetstream2 at Indiana University through allocation CIS240194 from the Advanced Cyberinfrastructure Coordination Ecosystem.

References

[1] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

[2] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. In *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc., 2020.

[3] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685, New Orleans, LA, USA, June 2022. IEEE.

[4] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In *Proceedings of the 40th International Conference on Machine Learning*, pages 19730–19742. PMLR, July 2023. ISSN: 2640-3498.

[5] Prafulla Dhariwal and Alexander Nichol. Diffusion Models Beat GANs on Image Synthesis. In *Advances in Neural Information Processing Systems*, volume 34, pages 8780–8794. Curran Associates, Inc., 2021.

[6] Lingzhi Li, Jianmin Bao, Hao Yang, Dong Chen, and Fang Wen. Advancing High Fidelity Identity Swapping for Forgery Detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5073–5082, Seattle, WA, USA, June 2020. IEEE.

[7] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. RePaint: Inpainting using Denoising Diffusion Probabilistic Models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11451–11461, New Orleans, LA, USA, June 2022. IEEE.

[8] Zeyu Lu, Di Huang, Lei Bai, Jingjing Qu, Chengyue Wu, Xihui Liu, and Wanli Ouyang. Seeing is not always believing: Benchmarking Human and Model Perception of AI-Generated Images. *Advances in Neural Information Processing Systems*, 36:25435–25447, December 2023.

[9] Kaicheng Yang, Danishjeet Singh, and Filippo Menczer. Characteristics and Prevalence of Fake Social Media Profiles with AI-generated Faces. *Journal of Online Trust and Safety*, 2(4), September 2024. Number: 4.

[10] Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Diffusion Art or Digital Forgery? Investigating Data Replication in Diffusion Models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6048–6058, Vancouver, BC, Canada, June 2023. IEEE.

[11] Renée DiResta and Josh A. Goldstein. How spammers and scammers leverage AI-generated images on Facebook for audience growth. *Harvard Kennedy School Misinformation Review*, August 2024.

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

[13] Xuandong Zhao, Kexun Zhang, Zihao Su, Saastha Vasan, Ilya Grishchenko, Christopher Kruegel, Giovanni Vigna, Yu-Xiang Wang, and Lei Li. Invisible Image Watermarks Are Provably Removable Using Generative AI. *Advances in Neural Information Processing Systems*, 37:8643–8672, December 2024.

[14] Hanzhe Li, Jiaran Zhou, Yuezun Li, Baoyuan Wu, Bin Li, and Junyu Dong. FreqBlender: Enhancing DeepFake Detection by Blending Frequency Knowledge. *Advances in Neural Information Processing Systems*, 37:44965–44988, December 2024.

[15] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR, July 2021. ISSN: 2640-3498.

[16] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In *Proceedings of the 39th International Conference on Machine Learning*, pages 12888–12900. PMLR, June 2022. ISSN: 2640-3498.

[17] Lingfeng Yang, Yueze Wang, Xiang Li, Xinlong Wang, and Jian Yang. Fine-Grained Visual Prompting. *Advances in Neural Information Processing Systems*, 36:24993–25006, December 2023.

[18] Yingjun Du, Wenfang Sun, and Cees G. Snoek. IPO: Interpretable Prompt Optimization for Vision-Language Models. *Advances in Neural Information Processing Systems*, 37:126725–126766, December 2024.

[19] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. MesoNet: a Compact Facial Video Forgery Detection Network. In *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–7, December 2018. ISSN: 2157-4774.

[20] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Niessner. FaceForensics++: Learning to Detect Manipulated Facial Images. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1–11, October 2019. ISSN: 2380-7504.

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

[22] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A. Efros. CNN-Generated Images Are Surprisingly Easy to Spot… for Now. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8692–8701. IEEE Computer Society, June 2020.

[23] Davide Cozzolino, Giovanni Poggi, Riccardo Corvi, Matthias Nießner, and Luisa Verdoliva. Raising the Bar of AI-generated Image Detection with CLIP. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 4356–4366, Seattle, WA, USA, June 2024. IEEE.

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

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

[26] Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. Leveraging Frequency Analysis for Deep Fake Image Recognition. In *Proceedings of the 37th International Conference on Machine Learning*, pages 3247–3258. PMLR, November 2020. ISSN: 2640-3498.

[27] Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Frequency-aware deepfake detection: improving generalizability through frequency space domain learning. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, volume 38 of *AAAI'24/IAAI'24/EAAI'24*, pages 5052–5060. AAAI Press, February 2024.

[28] Lorenzo Baraldi, Federico Cocchi, Marcella Cornia, Lorenzo Baraldi, Alessandro Nicolosi, and Rita Cucchiara. 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*, pages 199–216, Berlin, Heidelberg, November 2024. Springer-Verlag.

[29] Chende Zheng, Chenhao Lin, Zhengyu Zhao, Hang Wang, Xu Guo, Shuai Liu, and Chao Shen. Breaking Semantic Artifacts for Generalized AI-generated Image Detection. *Advances in Neural Information Processing Systems*, 37:59570–59596, December 2024.

[30] Peng Zhou, Xintong Han, Vlad I. Morariu, and Larry S. Davis. Learning Rich Features for Image Manipulation Detection. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1053–1061, June 2018. ISSN: 2575-7075.

[31] Quentin Bammey. Synthbuster: Towards Detection of Diffusion Model Generated Images. *IEEE Open Journal of Signal Processing*, 5:1–9, 2024.

[32] Zhiyuan Yan, Yong Zhang, Xinhang Yuan, Siwei Lyu, and Baoyuan Wu. DeepfakeBench: A Comprehensive Benchmark of Deepfake Detection, October 2023. arXiv:2307.01426 [cs].

[33] Junyan Ye, Baichuan Zhou, Zilong Huang, Junan Zhang, Tianyi Bai, Hengrui Kang, Jun He, Honglin Lin, Zihao Wang, Tong Wu, Zhizheng Wu, Yiping Chen, Dahua Lin, Conghui He, and Weijia Li. LOKI: A Comprehensive Synthetic Data Detection Benchmark using Large Multimodal Models. In *The Thirteenth International Conference on Learning Representations*, October 2024.

[34] 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, December 2023.

[35] Zhiyuan Yan, Taiping Yao, Shen Chen, Yandan Zhao, Xinghe Fu, Junwei Zhu, Donghao Luo, Chengjie Wang, Shouhong Ding, Yunsheng Wu, and Li Yuan. DF40: Toward Next-Generation Deepfake Detection. *Advances in Neural Information Processing Systems*, 37:29387–29434, December 2024.

[36] Lingzhi Zhang, Zhengjie Xu, Connelly Barnes, Yuqian Zhou, Qing Liu, He Zhang, Sohrab Amirghodsi, Zhe Lin, Eli Shechtman, and Jianbo Shi. Perceptual Artifacts Localization for Image Synthesis Tasks. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7545–7556, Paris, France, October 2023. IEEE.

[37] Rui Shao, Tianxing Wu, and Ziwei Liu. Detecting and Grounding Multi-Modal Media Manipulation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6904–6913, Vancouver, BC, Canada, June 2023. IEEE.

[38] Chuangchuang Tan, Huan Liu, 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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 28130–28139, Seattle, WA, USA, June 2024. IEEE.

[39] Ke Sun, Shen Chen, Taiping Yao, Hong Liu, Xiaoshuai Sun, Shouhong Ding, and Rongrong Ji. DiffusionFake: Enhancing Generalization in Deepfake Detection via Guided Stable Diffusion. *Advances in Neural Information Processing Systems*, 37:101474–101497, December 2024.

[40] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, pages 1877–1901, Red Hook, NY, USA, December 2020. Curran Associates Inc.

[41] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *Advances in Neural Information Processing Systems*, 35:24824–24837, December 2022.

[42] Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large Language Models are Zero-Shot Reasoners. *Advances in Neural Information Processing Systems*, 35:22199–22213, December 2022.

[43] Jindong Gu, Zhen Han, Shuo Chen, Ahmad Beirami, Bailan He, Gengyuan Zhang, Ruotong Liao, Yao Qin, Volker Tresp, and Philip Torr. A Systematic Survey of Prompt Engineering on Vision-Language Foundation Models, July 2023. arXiv:2307.12980 [cs].

[44] Shan Jia, Reilin Lyu, Kangran Zhao, Yize Chen, Zhiyuan Yan, Yan Ju, Chuanbo Hu, Xin Li, Baoyuan Wu, and Siwei Lyu. Can ChatGPT Detect DeepFakes? A Study of Using Multimodal Large Language Models for Media Forensics. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 4324–4333, Seattle, WA, USA, June 2024. IEEE.

[45] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. In *The Twelfth International Conference on Learning Representations*, October 2023.

[46] Ziqi Huang, Kelvin C.K. Chan, Yuming Jiang, and Ziwei Liu. Collaborative Diffusion for Multi-Modal Face Generation and Editing. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6080–6090, Vancouver, BC, Canada, June 2023. IEEE.

[47] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. StyleCLIP: Text-Driven Manipulation of StyleGAN Imagery. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2065–2074, Montreal, QC, Canada, October 2021. IEEE.

[48] 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 *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8789–8797, Salt Lake City, UT, June 2018. IEEE.

[49] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. StarGAN v2: Diverse Image Synthesis for Multiple Domains. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8185–8194, June 2020. ISSN: 2575-7075.

[50] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, June 2009. ISSN: 1063-6919.

[51] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large Scale GAN Training for High Fidelity Natural Image Synthesis. In *International Conference on Learning Representations*, September 2018.

[52] Alexander Quinn 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. In *Proceedings of the 39th International Conference on Machine Learning*, pages 16784–16804. PMLR, June 2022. ISSN: 2640-3498.

[53] 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 *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10686–10696, New Orleans, LA, USA, June 2022. IEEE.

[54] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient Memory Management for Large Language Model Serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, SOSP '23, pages 611–626, New York, NY, USA, October 2023. Association for Computing Machinery.

[55] 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. Qwen2.5-VL Technical Report, February 2025. arXiv:2502.13923 [cs].

A Appendix

A.1 Full Reasoning Traces

This section includes the full reasoning traces for all three methods. We can see the full traces produced by Qwen's **PGT** variations for the AI-generated image of an elephant (Figs 6, 7, 8).

```
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. AI-generated images often have a more
artificial or stylized appearance, which is not evident in
this image.
Final Answer(real/ai-generated): real. ✗
```

Figure 6: An example of a baseline reasoning trace for detecting AI-generated images using a VLM (Qwen2.5-7B). Image from Fig. 2.

A.2 Detection Recall

This section shows the detection recall for Qwen and LLaVA (Figs 9, 10).

A.3 Guidance Comparison

This section shows the detection performance for LLaVA and Llama across guidance variations of chain-of-thought and **S2** (Tables 3, 4).

A.4 Descriptive Analysis

This section shows the vocabulary associated with detection correctness for LLaVA and Llama (Figs 12, 11).

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 natural wrinkles and folds that are typical of an elephant's skin. This level of detail is often seen in real photographs.
2. Lighting and Shadows: The lighting and shadows on the elephant's skin are well-rendered, which is another characteristic of real photographs. The way the light falls on the elephant's body and the shadows it casts are consistent with natural lighting.
3. Background and Environment: The background shows a natural setting with greenery and a wall, which looks like a typical environment for an elephant. The details in the background, such as the leaves and the wall, are also quite realistic.
4. Overall Composition: The composition of the image, with the elephant in the foreground and the background elements, is natural and not overly stylized, which is more common in real photographs.
Given these observations, it is likely that this image is a real photograph of an elephant in a natural or semi-natural environment. However, it's important to note that AI-generated images can sometimes be very realistic, so it's always a good idea to cross-reference with other sources or experts if there's any doubt.
Final Answer(real/ai-generated): real. ✗

Figure 7: An example of a chain-of-thought reasoning trace for detecting AI-generated images using a VLM (Qwen2.5-7B). Image from Fig. 2.

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

Assistant Let's examine the style and 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 you would expect in a real photograph. The wrinkles and folds in the skin, while present, do not have the depth and complexity typically seen in real elephants.
2. Lighting and Shadows: The lighting in the image is somewhat flat, and the shadows are not very pronounced. In a real photograph, you would expect more dynamic lighting and more defined shadows, especially around the tusks and the trunk.
3. Background and Environment: The background appears to be a natural setting, but the image quality and the way the elephant is positioned suggest it might be a controlled environment, possibly a zoo or a sanctuary.
4. Overall Composition: The composition of the image is quite good, but the overall quality and the way the elephant is presented suggest it might be an AI-generated image rather than a real photograph.
Based on these observations, it is likely that this image is AI-generated.
Final Answer(real/ai-generated): ai-generated. ✓

Figure 8: An example of a **S2** reasoning trace for detecting AI-generated images using a VLM (Qwen2.5-7B). Image from Fig. 2.

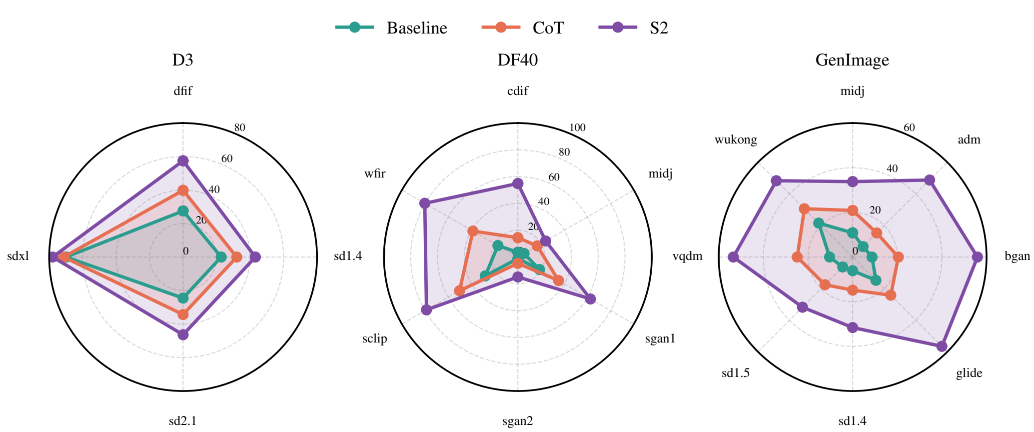


Figure 9: Detection recall (%) for Qwen across different datasets and their state-of-the-art synthetic image generators.

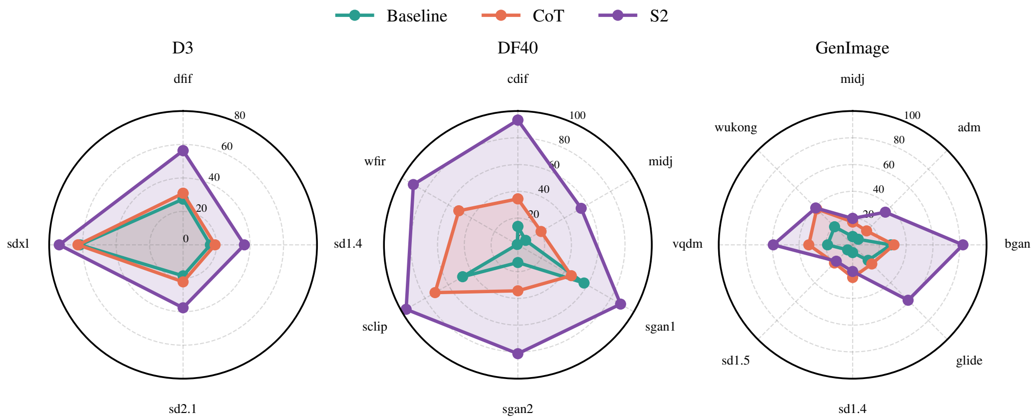


Figure 10: Detection recall (%) for LLaVA across different datasets and their state-of-the-art synthetic image generators.

Phrase	Type	D3 (2k)	DF40 (2k)	GenImage (2k)
Baseline	—	41.8	49.3	45.1
CoT	Prefill	44.2	59.2	55.1
	Pseudo-Prefill	41.3 (-2.9)	49.1 (-10.1)	45.6 (-9.4)
	Prompt	43.5 (-0.7)	52.2 (-7.0)	55.4 (+0.4)
S2	Prefill	52.1	57.5	59.6
	Pseudo-Prefill	41.6 (-10.5)	49.2 (-8.2)	45.6 (-14.0)
	Prompt	45.0 (-7.2)	60.6 (+3.1)	52.2 (-7.4)

Table 3: Macro F1 scores (%) across datasets and phrase modes for LLaVA. Values in parentheses indicate absolute differences compared to the prefill baselines.

Phrase	Type	D3 (2k)	DF40 (2k)	GenImage (2k)
Baseline	—	47.7	40.1	60.4
CoT	Prefill	47.8	60.3	58.1
	Pseudo-Prefill	40.7 (-7.1)	50.3 (-10.1)	47.4 (-10.7)
	Prompt	50.5 (+2.6)	61.5 (+1.2)	62.9 (+4.8)
S2	Prefill	58.4	62.1	71.8
	Pseudo-Prefill	58.6 (+0.1)	58.7 (-3.4)	64.4 (-7.4)
	Prompt	56.5 (-1.9)	63.1 (+1.0)	76.6 (+4.8)

Table 4: Macro F1 scores (%) across datasets and phrase modes for Llama. Values in parentheses indicate absolute differences compared to the prefill baselines.

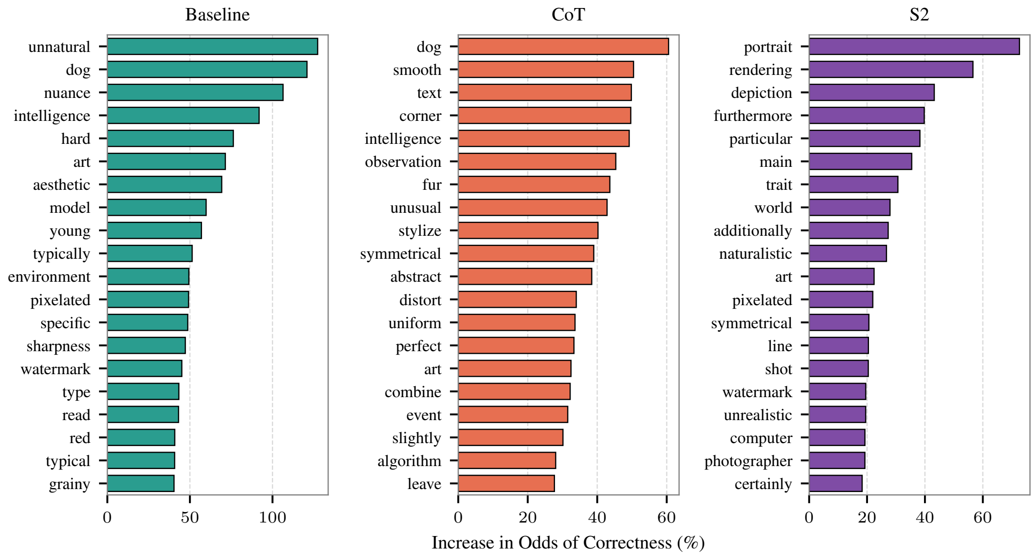


Figure 11: Top 20 words associated with the highest improvement in detection correctness for Llama.

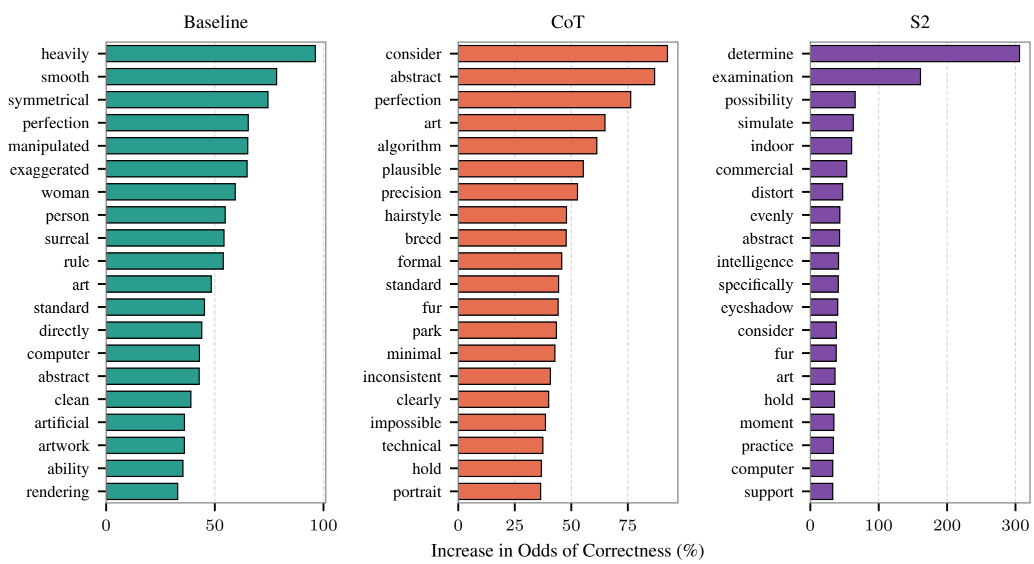


Figure 12: Top 20 words associated with the highest improvement in detection correctness for Llava.