

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

Anonymous ACL submission

Abstract

Traditional supervised methods for detecting AI-generated images depend on large, curated datasets for training and 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 a simple prefiling of responses — a method we call Prefill-Guided Thinking (PGT). In particular, prefiling a VLM response with the 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%. We analyze this improvement by tracking models’ answer confidence at incremental intervals during response generation. For some models, prefills counteract early overconfidence — akin to mitigating the Dunning-Kruger effect — leading to better detection performance.

1 Introduction

Rapid advancements in image generation have led to a surge in synthetic images (deepfakes) (Goodfellow et al., 2014; Ho et al., 2020; Rombach et al., 2022). Improved techniques now enable the easier and cheaper production of high-quality visuals (Rombach et al., 2022; Li et al., 2023; Dhariwal and Nichol, 2021). 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 (Li et al., 2020; Lugmayr et al., 2022; Lu et al., 2023). Such forgeries facilitate impersonation, copyright infringement, and disinformation, highlighting the urgent need for robust detection methods to maintain visual trust (Yang et al., 2024; Somepalli et al., 2023; DiResta and Goldstein, 2024; Chandra, 2024).

Existing solutions remain limited. Watermarking and metadata approaches are often easily bypassed and require widespread adoption (Zhao et al., 2024). Supervised methods, particularly feature-based ones, struggle to generalize to new



User: [Image] Is this image real or AIgenerated?	User: [Image] Is this image real or AIgenerated?
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).

generators (Li et al., 2024).

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 (Radford et al., 2021; Li et al., 2022, 2023) and whose performance can be further enhanced through prompting (Yang et al., 2023; Du et al., 2024).

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 “*Lets 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 diverse datasets spanning human faces, objects, and animals. These images have been generated by 16 state-of-the-art image models that produce photorealistic outputs lacking visible synthesis artifacts (Fig. 1¹). Despite this, our approach boosts Macro F1 scores for three widely used open-source VLMs by up to 24%, demonstrating strong generalization across image categories as well as modern generators.

We further analyze why prefilling is helpful by tracking models’ answer confidence at incremental intervals during their response generation. We find that baseline queries exhibit immediate high confidence that leads to poor performance — a Dunning-Kruger-like effect. For Qwen and LLaVA (on most datasets), prefills counteract this early overconfidence more effectively than prompts. They induce greater uncertainty that leads to gradually improving performance as their responses progress.

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 (Afchar et al., 2018; Rössler et al., 2019; Chollet, 2017). As generative models improve, these cues become less reliable.

Models trained on fixed artifacts often overfit to specific generators, leading to poor generalization (Wang et al., 2020; Cozzolino et al., 2024; Ojha et al., 2023).

Frequency-domain techniques analyze spectral representations using techniques like Fast Fourier Transform or Discrete Cosine Transform (Li et al., 2024; Dzanic et al., 2020; Frank et al., 2020; Tan et al., 2024b). 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 (Cozzolino et al., 2024; Baraldi et al., 2024; Zheng et al., 2024).

Spatial-domain approaches examine raw pixel patterns to detect structural or textural inconsistencies (Zheng et al., 2024; Zhou et al., 2018). 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 (Rössler et al., 2019; Bamme, 2024).

To improve generalization, recent work has explored larger and more diverse training datasets (Wang et al., 2020; Yan et al., 2023; Ye et al., 2024; Zhu et al., 2023; Yan et al., 2024; Zhang et al., 2023; Shao et al., 2023), as well as architectural modifications that target universal artifacts (e.g., upsampling patterns) (Zheng et al., 2024; Tan et al., 2024a). Approaches like fixed-feature backbones and DiffusionFake have also improved robustness to novel generators (Ojha et al., 2023; Sun et al., 2024).

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 (Li et al., 2023; Radford et al., 2021; Li et al., 2022). Prompt-based learning has emerged as a lightweight method to adapt language models using natural language instructions (Brown et al., 2020). Chain-of-thought, which encourages step-by-step reasoning, has been effective in multimodal settings (Wei et al., 2022; Kojima et al., 2022; Gu et al., 2023).

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 training for synthetic image detection, with greater capacity to generalize to increasingly

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

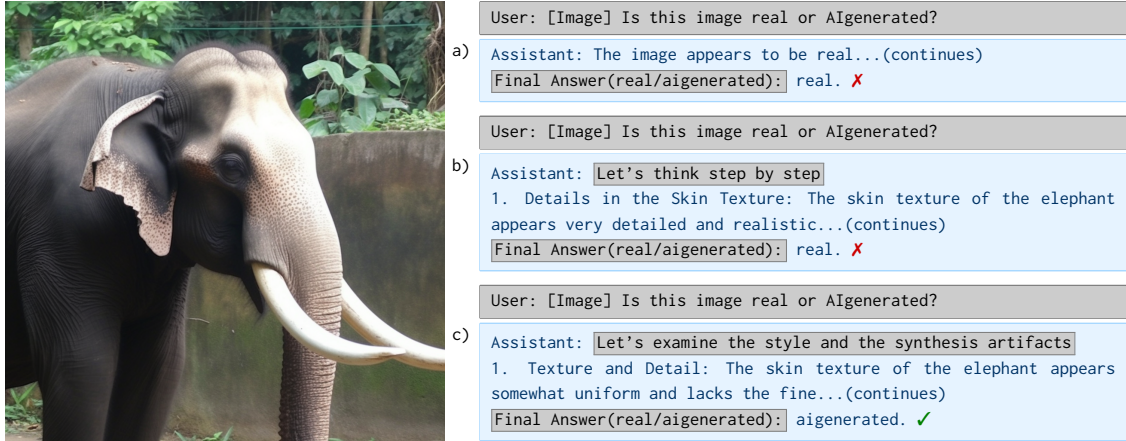


Figure 2: Illustration of **PGT** for the detection of an AI-generated image using a VLM (Qwen2.5-7B). Input text is marked in **grey**, response text in **blue**. (a) A baseline user query results in the incorrect response **real**. (b) Using the chain-of-thought prefill **Let's think step by step** 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.

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.

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 constrains the model to select between the two labels presented in the prompt. We extract predictions using keyword matching for *real* and *ai-generated* (or its synonyms: *artificial*, *fake*, *generated*), defaulting to *real* when both appear.

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 (Kojima et al., 2022). 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 (Jia et al., 2024), 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.

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.

Data. We conduct experiments using three diverse benchmarks spanning multiple visual domains and 16 state-of-the-art image generators.

Together, these datasets encompass human faces, natural objects, animals, urban scenes, artwork, and abstract imagery, with synthetic images produced by both GAN and diffusion-based architectures.

D3 is a benchmark dataset introduced as part of the Contrastive Deepfake Embeddings framework (Baraldi et al., 2024). 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 (Rombach et al., 2022), and Stable Diffusion XL (Podell et al., 2023). 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 (Yan et al., 2024). It includes outputs from state-of-the-art models such as Collaborative Diffusion (Huang et al., 2023), Midjourney, StyleCLIP (Patashnik et al., 2021), StarGAN v1 and v2 (Choi et al., 2018, 2020), 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 (Deng et al., 2009), 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 (Dhariwal and Nichol, 2021), BigGAN (Brock et al., 2018), GLIDE (Nichol et al., 2022), Midjourney, VQDM (Gu et al., 2022), Stable Diffusion v1.4 and v1.5 (Rombach et al., 2022), 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).

Collectively, our evaluation encompasses 28,420 images across diverse visual content and generation methods, providing a rigorous test for our method’s generalization capabilities. Importantly, modern diffusion models in these benchmarks (e.g., Stable Diffusion XL, GLIDE, Midjourney) produce highly photorealistic images that largely lack the traditional synthesis artifacts targeted by frequency-domain and spatial-domain detectors (Fig. 1).

Models. We conduct experiments with three Vision-Language Models: *LLaVA-OneVision*², *Qwen2.5-VL*³, and *Llama-3.2-Vision*⁴. We use instruction-tuned variants in evaluation mode and use seed=0 in the vLLM package (Kwon et al., 2023) to ensure reproducibility. Token generation is capped at 512 tokens (max_tokens=512) with temperature=0.0 for deterministic outputs. A single run on a NVIDIA A100 80GB for 10,000 images takes approximately 3 hours.

LLaVA-OneVision has been trained on multi-modal 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 (Bai et al., 2025). 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.

4 Results

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

²<https://huggingface.co/collections/llava-hf/llava-onevision-66bb1e9ce8856e210a7ed1fe>

³<https://huggingface.co/collections/Qwen/qwen25-vl-6795ffac22b334a837c0f9a5>

⁴<https://huggingface.co/collections/meta-llama/llama-32-66f448ffc8c32f949b04c8cf>

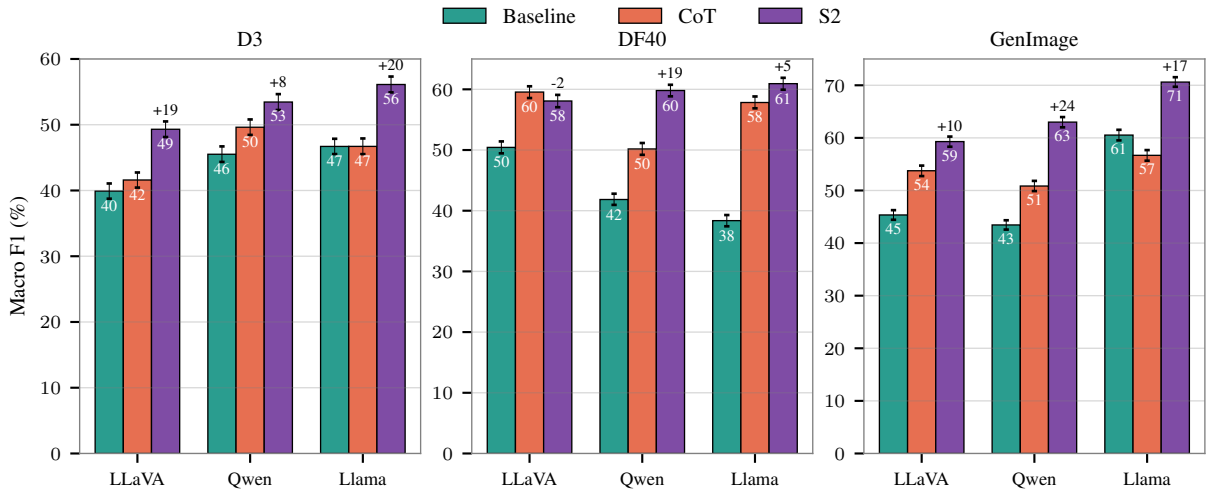


Figure 3: Detection Macro F1 across models, datasets, and **PGT** variations. Bars are annotated with relative improvements of **S2** over the next best method and 95% confidence error bars from 10k bootstrap iterations.

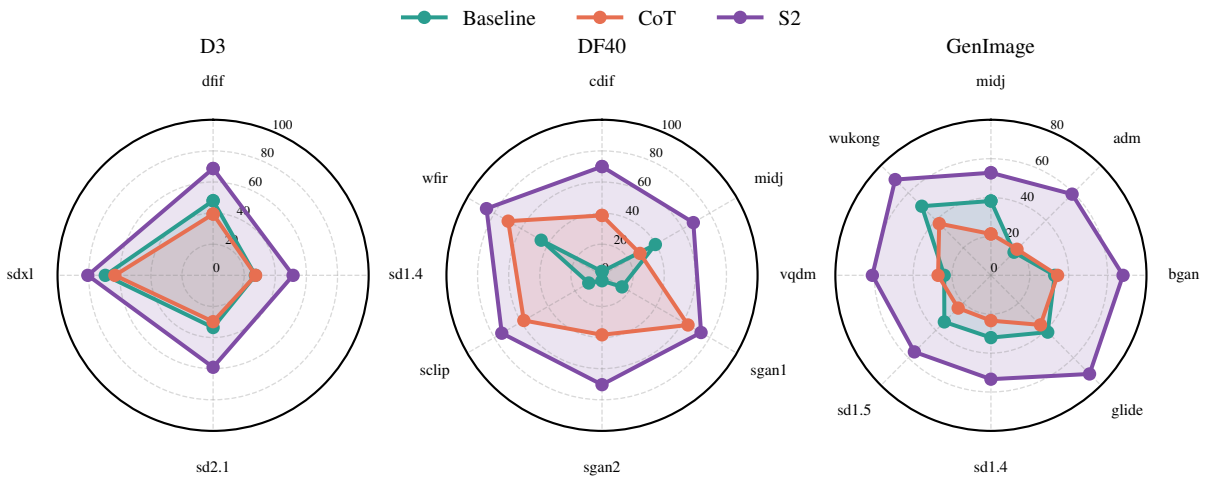


Figure 4: Detection recall (%) for Llama across different datasets and their state-of-the-art synthetic image generators. Similar figures for Qwen and LLaVA are in the Appendix (Figs. 10, 11).

improvement in Macro F1. Across the 16 state-of-the-art generators, **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. 10, 11 in Appendix). Notably, strong recall on modern diffusion models like Stable Diffusion XL, GLIDE, and Midjourney demonstrates that **S2** enables detection of photo-realistic outputs that largely lack traditional synthesis artifacts. These results show that our task-aligned prefill generalizes effectively, outperforming chain-of-thought for AI-generated image detection across diverse models, datasets, and generators.

Guidance Comparison. We also compare prefill-guided thinking against simple user prompts that combine 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).

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 (-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)

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.

Phrase	Prefill	D3 (2k)	DF40 (2k)	GenImage (2k)
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 the synthesis artifacts	53.2	61.2	64.8
Variants	Let’s observe the style and the 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)

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

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 prefill guidance consistently improves performance compared to the baseline.

5 Interpretability

To understand how different guidance modes influences the models, we track the answer confidence and detection performance at incremental intervals during their response generation.

Method. We conduct this analysis on the three smaller evaluation subsets — D3 (2k), DF40 (2k),

GenImage (2k) — across all three models (Qwen, LLaVA, Llama) and three guidance modes (baseline, prefill, prompt). To enable aggregation of answer confidence and detection performance across all responses for a data-model-guidance combination, we normalize each response to five intervals (0%, 25%, 50%, 75%, 100%) based on sentence counts (Fig. 5). Interval boundaries are determined by rounding the product of total sentence count and interval percentage to the nearest integer. For example, a 10-sentence response yields intervals at sentences 0, 2, 5, 8, and 10; for very short responses (< 4 sentences), some intervals may coincide.

At each interval, we construct the input by concatenating the original query, the partial response up to that interval, and the phrase `Final Answer(real/ai-generated): This image is` to elicit an answer. We generate a short completion (temperature=0.0, max_tokens=20) while tracking token-level log probabilities. Answer confidence is the average log probability of the tokens generated in this short response. Answer

User: [Image] Is this image real or AIgenerated?				
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.

Figure 5: Example of Qwen’s incremental response across five intervals for the baseline method. At each interval, we use the phrase `Final Answer(real/ai-generated): This image is` to elicit an answer.

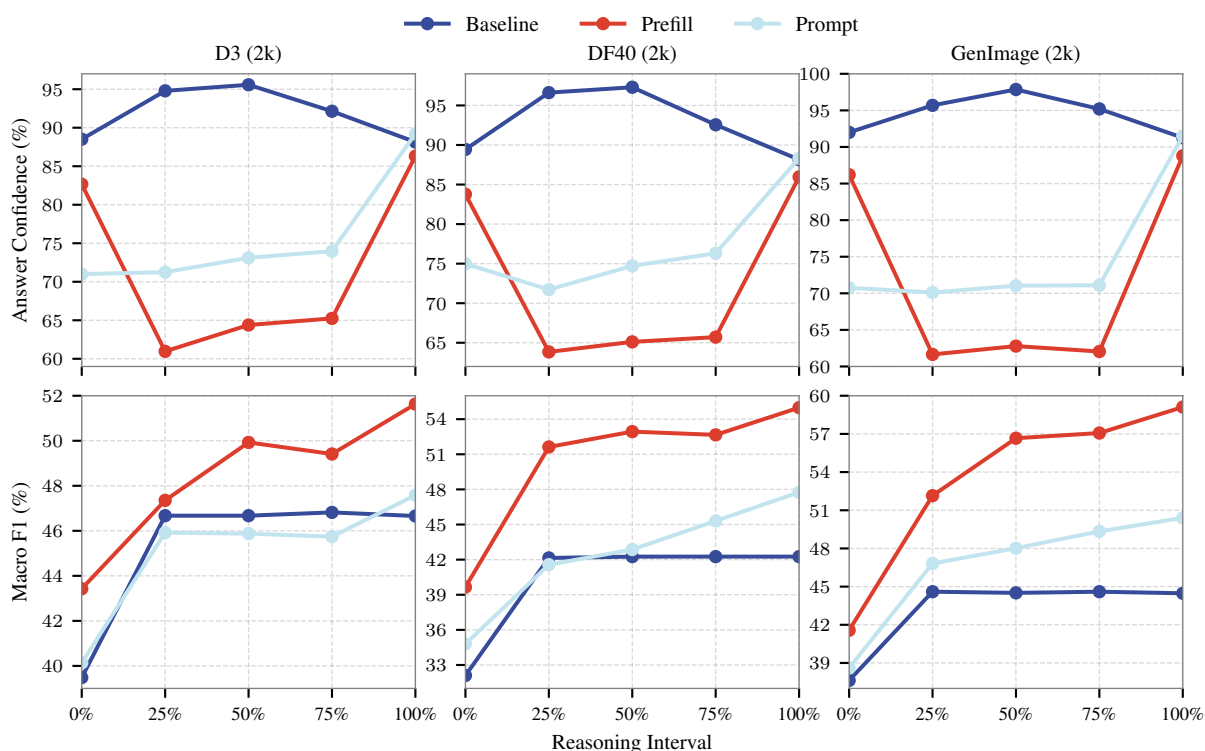


Figure 6: Evolution of answer confidence and Macro F1 across response generation intervals for Qwen on D3 (2k), DF40 (2k), and GenImage (2k) using baseline, prefill, and prompt guidance modes.

390 predictions (real/AI-generated) are extracted using
 391 the same keyword matching approach described in
 392 Section 3.

393 Answer confidence for a data-model-guidance
 394 combination is the average answer confidence
 395 across all responses at each interval. Similarly,
 396 Macro F1 for a data-model-guidance combination
 397 is calculated by using the confusion matrix of all
 398 predictions at each interval. For prompt and pre-

fill, Answer Confidence and Macro F1 are further
 400 averaged across the two phrases (CoT and S2).

401 **Results.** Models guided by prefills show
 402 markedly different reasoning patterns compared
 403 to baseline and prompt methods. For baseline
 404 queries, Qwen exhibits high confidence (>85%)
 405 from the start that remains stable throughout
 406 its response (Fig. 6). Yet this confidence does
 407 not yield strong detection. Macro F1 plateaus

around 42–47% early and stays flat. The model commits to its answer immediately and spends the remainder of its response justifying that decision.

Prefills reverse this pattern. Qwen’s answer confidence drops sharply in early reasoning, remains low through the middle intervals, then recovers only at completion. Crucially, detection performance improves steadily throughout this process. Macro F1 rises from 40–45% at the start to 52–58% at completion — the highest of any method. This suggests that by mitigating early overconfidence, prefilled responses might be forcing the model to reason more carefully before committing to an answer.

User prompts show intermediate behavior. Confidence remains relatively flat with modest performance gains, indicating less effective reasoning than prefills.

LLaVA follows similar patterns for D3 (2k) and GenImage (2k): baseline shows overconfidence, prefills show improved performance with the same confidence dip, and prompts fall in between (Appendix Fig. 12). The exception is DF40 (2k), where prefills show no confidence dip and underperform prompts. Llama shows inconsistent patterns across datasets, with no clear relationship between confidence curves and performance (Appendix Fig. 13).

Overall, these results suggest models may suffer from early overconfidence that harms detection performance — akin to the Dunning-Kruger effect (Kruger and Dunning, 1999). For some models, prefill-guided thinking mitigates this effect by counteracting this early confidence by potentially encouraging more deliberate reasoning before committing to an answer.

6 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, generalizing robustly across 16 state-of-the-art image generators — even those producing photorealistic outputs lacking traditional artifacts.

We also find that responses must be prefilled directly, as instructing the model to begin its responses with the prefill does not reliably repli-

cate 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.

Our interpretability analysis offers a potential explanation for these improvements. For Qwen and LLaVA on most datasets, prefills appear to counteract early overconfidence that otherwise leads to premature answers. While baseline queries trigger immediate high confidence with poor performance, prefills induce uncertainty that persists throughout their response generation. This, in turn, leads to gradually improving detection performance — suggesting that prefills may mitigate a Dunning-Kruger-like effect, forcing more deliberate reasoning.

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.

7 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 require significantly more memory.

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.

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.

Finally, our interpretability analysis revealed patterns of overconfidence mitigation for Qwen and LLaVA on most datasets, but Llama exhibited inconsistent patterns with no clear relationship between answer confidence and detection performance. These findings remain observational and further investigation is needed.

508
509
510
511
512
513

514
515
516
517
518
519
520

521
522
523

524
525
526
527
528
529
530
531

532
533
534
535

536
537
538
539
540
541
542
543
544
545
546

547
548
549

550
551
552
553
554
555
556

557
558
559
560
561
562

References

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

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao 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, and 8 others. 2025. [Qwen2.5-VL Technical Report](#). *arXiv preprint*. ArXiv:2502.13923 [cs].

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

Lorenzo Baraldi, Federico Cocchi, Marcella Cornia, Lorenzo Baraldi, Alessandro Nicolosi, and Rita Cucchiara. 2024. [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. Springer-Verlag.

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

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, and 12 others. 2020. [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. Curran Associates Inc.

Bilva Chandra. 2024. [Analyzing Harms from AI-Generated Images and Safeguarding Online Authenticity](#). RAND Corporation.

Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. 2018. [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. IEEE.

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

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

Davide Cozzolino, Giovanni Poggi, Riccardo Corvi, Matthias Nießner, and Luisa Verdoliva. 2024. [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. IEEE.

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

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

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

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

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

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

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

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

Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. 2022. [Vector Quantized Diffusion Model](#)

619	for Text-to-Image Synthesis. In <i>2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)</i> , pages 10686–10696, New Orleans, LA, USA. IEEE.	677
620		678
621		679
622		680
623	Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising Diffusion Probabilistic Models. In <i>Advances in Neural Information Processing Systems</i> , volume 33, pages 6840–6851. Curran Associates, Inc.	681
624		682
625		683
626		684
627		685
628	Ziqi Huang, Kelvin C.K. Chan, Yuming Jiang, and Ziwei Liu. 2023. Collaborative Diffusion for Multimodal Face Generation and Editing. In <i>2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)</i> , pages 6080–6090, Vancouver, BC, Canada. IEEE.	686
629		687
630		688
631		689
632		690
633		691
634	Shan Jia, Reilin Lyu, Kangran Zhao, Yize Chen, Zhiyuan Yan, Yan Ju, Chuanbo Hu, Xin Li, Baoyuan Wu, and Siwei Lyu. 2024. Can ChatGPT Detect DeepFakes? A Study of Using Multimodal Large Language Models for Media Forensics. In <i>2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)</i> , pages 4324–4333, Seattle, WA, USA. IEEE.	692
635		693
636		694
637		695
638		696
639		697
640		698
641		699
642	Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large Language Models are Zero-Shot Reasoners. <i>Advances in Neural Information Processing Systems</i> , 35:22199–22213.	700
643		701
644		702
645		703
646		704
647	J. Kruger and D. Dunning. 1999. Unskilled and unaware of it: how difficulties in recognizing one’s own incompetence lead to inflated self-assessments. <i>Journal of Personality and Social Psychology</i> , 77(6):1121–1134.	705
648		706
649		707
650		708
651		709
652	Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient Memory Management for Large Language Model Serving with PagedAttention. In <i>Proceedings of the 29th Symposium on Operating Systems Principles, SOSP ’23</i> , pages 611–626, New York, NY, USA. Association for Computing Machinery.	710
653		711
654		712
655		713
656		714
657		715
658		716
659		717
660	Hanzhe Li, Jiaran Zhou, Yuezun Li, Baoyuan Wu, Bin Li, and Junyu Dong. 2024. FreqBlender: Enhancing DeepFake Detection by Blending Frequency Knowledge. <i>Advances in Neural Information Processing Systems</i> , 37:44965–44988.	718
661		719
662		720
663		721
664		722
665	Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In <i>Proceedings of the 40th International Conference on Machine Learning</i> , pages 19730–19742. PMLR. ISSN: 2640-3498.	723
666		724
667		725
668		726
669		727
670		728
671	Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In <i>Proceedings of the 39th International Conference on Machine Learning</i> , pages 12888–12900. PMLR. ISSN: 2640-3498.	729
672		730
673		731
674		732
675		733
676		734
	Lingzhi Li, Jianmin Bao, Hao Yang, Dong Chen, and Fang Wen. 2020. Advancing High Fidelity Identity Swapping for Forgery Detection. In <i>2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)</i> , pages 5073–5082, Seattle, WA, USA. IEEE.	735
		736
		737
		738
	Zeyu Lu, Di Huang, Lei Bai, Jingjing Qu, Chengyue Wu, Xihui Liu, and Wanli Ouyang. 2023. Seeing is not always believing: Benchmarking Human and Model Perception of AI-Generated Images. <i>Advances in Neural Information Processing Systems</i> , 36:25435–25447.	739
		740
		741
		742
	Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. 2022. RePaint: Inpainting using Denoising Diffusion Probabilistic Models. In <i>2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)</i> , pages 11451–11461, New Orleans, LA, USA. IEEE.	743
		744
		745
		746
	Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. 2022. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. In <i>Proceedings of the 39th International Conference on Machine Learning</i> , pages 16784–16804. PMLR. ISSN: 2640-3498.	747
		748
		749
		750
	Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. 2023. Towards Universal Fake Image Detectors that Generalize Across Generative Models. In <i>2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)</i> , pages 24480–24489, Vancouver, BC, Canada. IEEE.	751
		752
		753
		754
	Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. 2021. StyleCLIP: Text-Driven Manipulation of StyleGAN Imagery. In <i>2021 IEEE/CVF International Conference on Computer Vision (ICCV)</i> , pages 2065–2074, Montreal, QC, Canada. IEEE.	755
		756
		757
		758
	Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2023. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. In <i>The Twelfth International Conference on Learning Representations</i> .	759
		760
		761
		762
	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. 2021. Learning Transferable Visual Models From Natural Language Supervision. In <i>Proceedings of the 38th International Conference on Machine Learning</i> , pages 8748–8763. PMLR. ISSN: 2640-3498.	763
		764
		765
		766
	Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. 2022. High-Resolution Image Synthesis with Latent Diffusion	767

733	Models . In <i>2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)</i> , pages 10674–10685, New Orleans, LA, USA. IEEE.	
734		
735		
736	Andreas Rössler, Davide Cozzolino, Luisa Verdoliva,	
737	Christian Riess, Justus Thies, and Matthias Niessner. 2019. FaceForensics++: Learning to Detect Manipulated Facial Images . In <i>2019 IEEE/CVF International Conference on Computer Vision (ICCV)</i> , pages 1–11. ISSN: 2380-7504.	
738		
739		
740		
741		
742	Rui Shao, Tianxing Wu, and Ziwei Liu. 2023. Detecting and Grounding Multi-Modal Media Manipulation . In <i>2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)</i> , pages 6904–6913, Vancouver, BC, Canada. IEEE.	
743		
744		
745		
746		
747	Gowthami Somepalli, Vasu Singla, Micah Goldblum,	
748	Jonas Geiping, and Tom Goldstein. 2023. Diffusion Art or Digital Forgery? Investigating Data Replication in Diffusion Models . In <i>2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)</i> , pages 6048–6058, Vancouver, BC, Canada. IEEE.	
749		
750		
751		
752		
753		
754	Ke Sun, Shen Chen, Taiping Yao, Hong Liu, Xiaoshuai	
755	Sun, Shouhong Ding, and Rongrong Ji. 2024. DiffusionFake: Enhancing Generalization in Deepfake Detection via Guided Stable Diffusion . <i>Advances in Neural Information Processing Systems</i> , 37:101474–101497.	
756		
757		
758		
759		
760	Chuangchuang Tan, Huan Liu, Yao Zhao, Shikui Wei,	
761	Guanghua Gu, Ping Liu, and Yunchao Wei. 2024a. Rethinking the Up-Sampling Operations in CNN-Based Generative Network for Generalizable Deepfake Detection . In <i>2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)</i> , pages 28130–28139, Seattle, WA, USA. IEEE.	
762		
763		
764		
765		
766		
767	Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua	
768	Gu, Ping Liu, and Yunchao Wei. 2024b. Frequency-aware deepfake detection: improving generalizability through frequency space domain learning . In <i>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</i> , volume 38 of AAAI’24/IAAI’24/EAAI’24, pages 5052–5060. AAAI Press.	
769		
770		
771		
772		
773		
774		
775		
776		
777		
778	Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew	
779	Owens, and Alexei A. Efros. 2020. CNN-Generated Images Are Surprisingly Easy to Spot for Now . In <i>2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)</i> , pages 8692–8701. IEEE Computer Society.	
780		
781		
782		
783		
784	Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten	
785	Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models . <i>Advances in Neural Information Processing Systems</i> , 35:24824–24837.	
786		
787		
788		
789		
	Zhiyuan Yan, Taiping Yao, Shen Chen, Yandan Zhao,	790
	Xinghe Fu, Junwei Zhu, Donghao Luo, Chengjie	791
	Wang, Shouhong Ding, Yunsheng Wu, and Li Yuan.	792
	2024. DF40: Toward Next-Generation Deepfake Detection . <i>Advances in Neural Information Processing Systems</i> , 37:29387–29434.	793
		794
		795
	Zhiyuan Yan, Yong Zhang, Xinhang Yuan, Siwei	796
	Lyu, and Baoyuan Wu. 2023. DeepfakeBench: A Comprehensive Benchmark of Deepfake Detection . <i>arXiv preprint</i> . ArXiv:2307.01426 [cs].	797
		798
		799
	Kaicheng Yang, Danishjeet Singh, and Filippo	800
	Menczer. 2024. Characteristics and Prevalence of Fake Social Media Profiles with AI-generated Faces . <i>Journal of Online Trust and Safety</i> , 2(4). Number: 4.	801
		802
		803
		804
	Lingfeng Yang, Yueze Wang, Xiang Li, Xinlong Wang,	805
	and Jian Yang. 2023. Fine-Grained Visual Prompting . <i>Advances in Neural Information Processing Systems</i> , 36:24993–25006.	806
		807
		808
	Junyan Ye, Baichuan Zhou, Zilong Huang, Junan	809
	Zhang, Tianyi Bai, Hengrui Kang, Jun He, Honglin	810
	Lin, Zihao Wang, Tong Wu, Zhizheng Wu, Yiping	811
	Chen, Dahua Lin, Conghui He, and Weijia Li. 2024. LOKI: A Comprehensive Synthetic Data Detection Benchmark using Large Multimodal Models . In <i>The Thirteenth International Conference on Learning Representations</i> .	812
		813
		814
		815
		816
	Lingzhi Zhang, Zhengjie Xu, Connelly Barnes, Yuqian	817
	Zhou, Qing Liu, He Zhang, Sohrab Amirghodsi, Zhe	818
	Lin, Eli Shechtman, and Jianbo Shi. 2023. Perceptual Artifacts Localization for Image Synthesis Tasks . In <i>2023 IEEE/CVF International Conference on Computer Vision (ICCV)</i> , pages 7545–7556, Paris, France. IEEE.	819
		820
		821
		822
		823
		824
	Xuandong Zhao, Kexun Zhang, Zihao Su, Saastha	825
	Vasan, Ilya Grishchenko, Christopher Kruegel, Giovanni	826
	Vigna, Yu-Xiang Wang, and Lei Li. 2024. Invisible Image Watermarks Are Provably Removable Using Generative AI . <i>Advances in Neural Information Processing Systems</i> , 37:8643–8672.	827
		828
		829
	Chende Zheng, Chenhao Lin, Zhengyu Zhao, Hang	830
	Wang, Xu Guo, Shuai Liu, and Chao Shen. 2024. Breaking Semantic Artifacts for Generalized AI-generated Image Detection . <i>Advances in Neural Information Processing Systems</i> , 37:59570–59596.	831
		832
		833
		834
	Peng Zhou, Xintong Han, Vlad I. Morariu, and Larry S.	835
	Davis. 2018. Learning Rich Features for Image Manipulation Detection . In <i>2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition</i> , pages 1053–1061. ISSN: 2575-7075.	836
		837
		838
		839
	Mingjian Zhu, Hanting Chen, Qiangyu Yan, Xudong	840
	Huang, Guanyu Lin, Wei Li, Zhijun Tu, Hailin	841
	Hu, Jie Hu, and Yunhe Wang. 2023. GenImage: A Million-Scale Benchmark for Detecting AI-Generated Image . <i>Advances in Neural Information Processing Systems</i> , 36:77771–77782.	842
		843
		844
		845

846
847
848
849

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 7, 8, 9).

```
User: [Image] Is this image real or AIgenerated?  
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/aigenerated): real. ✘
```

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

```
User: [Image] Is this image real or AIgenerated?  
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/aigenerated): real. ✘
```

Figure 8: 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 AIgenerated?  
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/aigenerated): ai-generated. ✔
```

Figure 9: An example of a S2 reasoning trace for detecting AI-generated images using a VLM (Qwen2.5-7B). Image from Fig. 2.

850
851

A.2 Detection Recall

This section shows the detection recall for Qwen and LLaVA (Figs 10, 11).

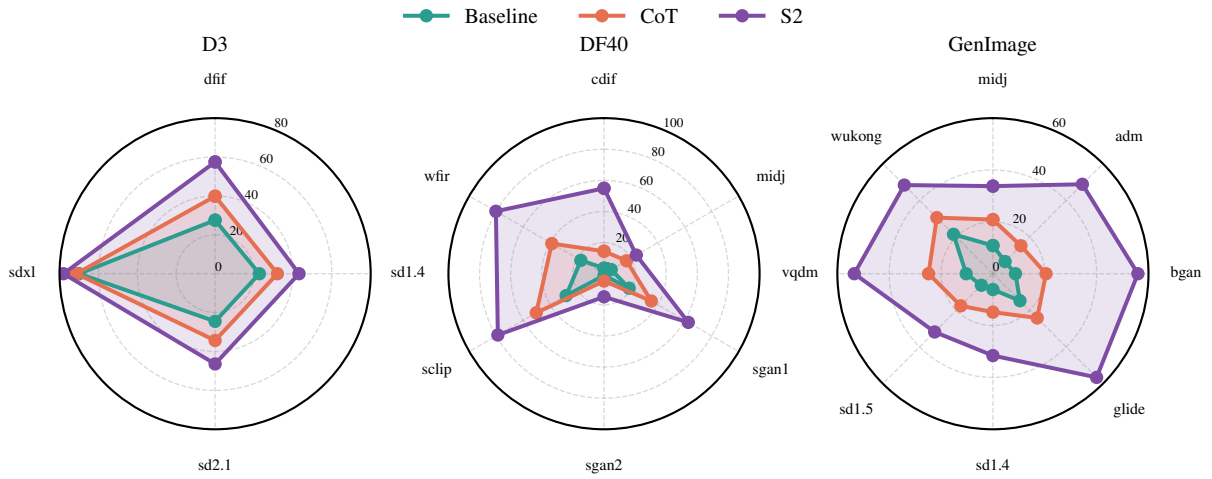


Figure 10: Detection recall (%) for Qwen across different datasets and their state-of-the-art synthetic image generators.

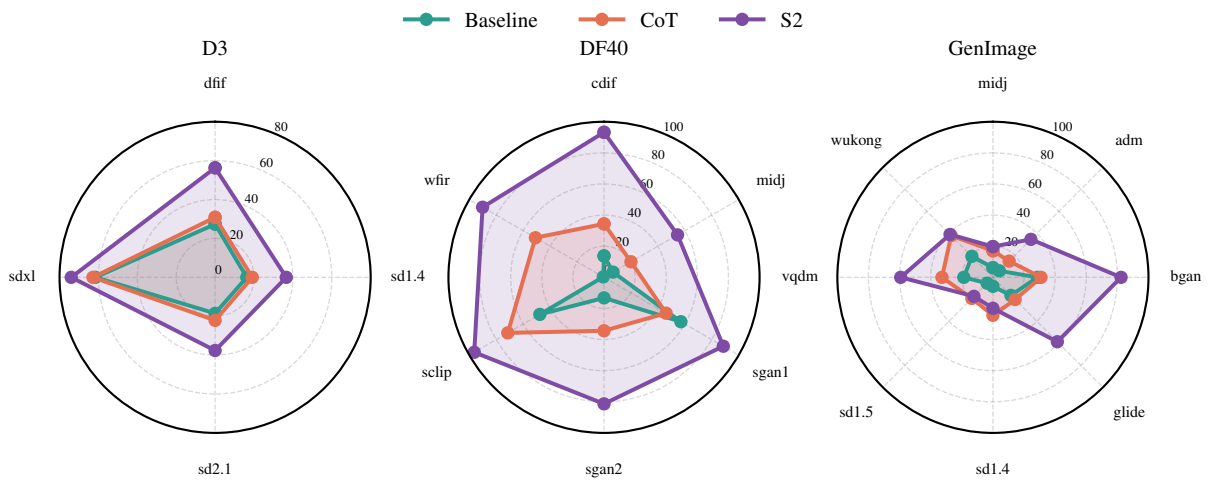


Figure 11: Detection recall (%) for LLaVA across different datasets and their state-of-the-art synthetic image generators.

A.3 Guidance Comparison

852

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

853

854

A.4 Confidence Analysis

855

This section shows the interval progression for LLaVA and Llama (Figs 12, 13).

856

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 (-9.5)
	Prompt	43.5 (-0.7)	52.2 (-7.0)	55.4 (+0.3)
S2	Prefill	52.1	57.5	59.6
	Pseudo-Prefill	41.6 (-10.5)	49.2 (-8.3)	45.6 (-14.0)
	Prompt	45.0 (-7.1)	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 (-10.0)	47.4 (-10.7)
	Prompt	50.5 (+2.7)	61.5 (+1.2)	62.9 (+4.8)
S2	Prefill	58.4	62.1	71.8
	Pseudo-Prefill	58.6 (+0.2)	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.

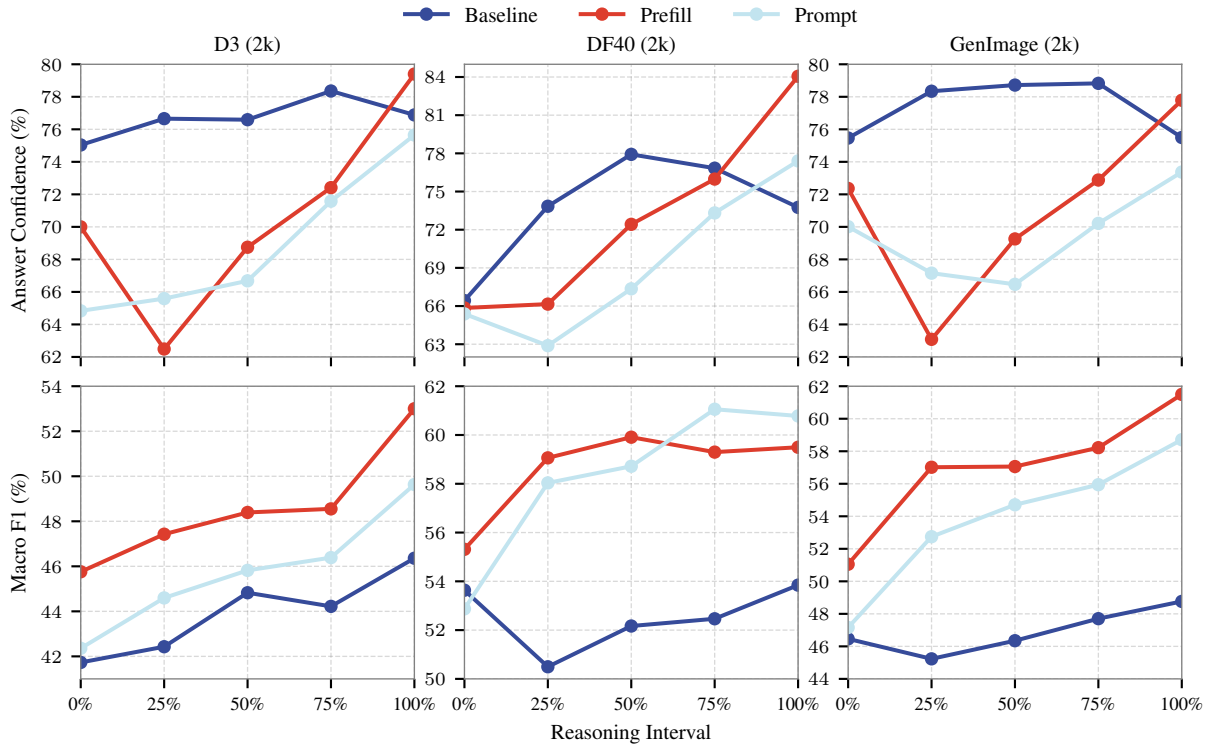


Figure 12: Interval progression for LLaVA.

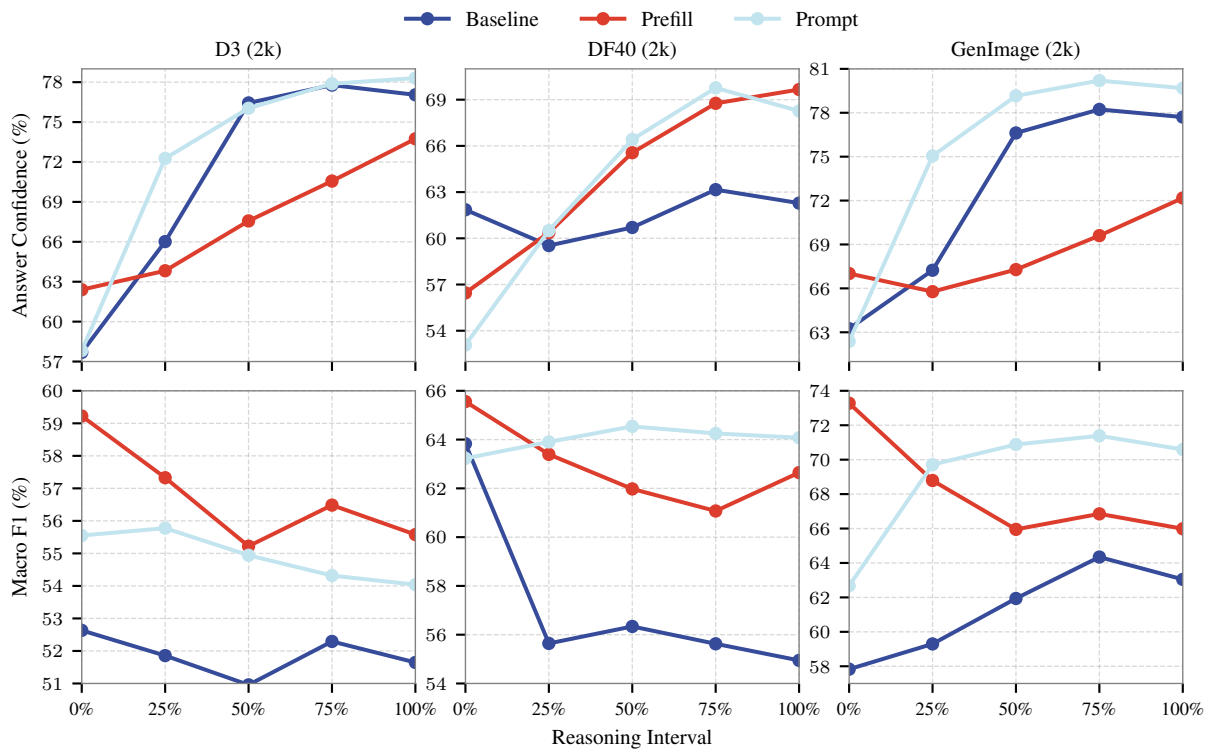


Figure 13: Interval progression for Llama.