Task-aligned prompting improves zero-shot detection of AI-generated images by Vision-Language Models

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

Abstract

As image generators produce increasingly realistic images, concerns about potential misuse continue to grow. Supervised detection relies on large, curated datasets and struggles to generalize across diverse generators. In this work, we investigate the use of pre-trained Vision-Language Models (VLMs) for zero-shot detection of AI-generated images. While off-the-shelf VLMs exhibit some task-specific reasoning and chain-of-thought prompting offers gains, we show that task-aligned prompting elicits more focused reasoning and significantly improves performance without fine-tuning. Specifically, prefixing the model's response with the phrase "*Let's examine the style and the synthesis artifacts*"—a method we call **zero-shot-s**^{2}—boosts Macro F1 scores by 8%–29%. These gains are consistent for two widely used open-source models and across three recent, diverse datasets spanning human faces, objects, and animals with images generated by 16 different models—demonstrating strong generalization. We further evaluate the approach across three additional model sizes and observe improvements in most dataset–model combinations—suggesting robustness to model scale. Surprisingly, self-consistency, a behavior previously observed in language reasoning, where aggregating answers from diverse reasoning paths improves performance, also holds in this setting. Even here, zero-shot-s^2 scales better than chain-of-thought in most cases—indicating that it elicits more useful diversity. Our findings show that task-aligned prompts elicit more focused reasoning and enhance latent capabilities in VLMs, like the detection of AI-generated images—offering a simple, generalizable, and explainable alternative to supervised methods. Our code is publicly available on github: `https://github.com/Zoher15/Zero-shot-s2`.

Figure 1: A sample of images from the D3 (top), DF40 (middle), and GenImage (bottom) datasets. Can you guess which ones are real vs AI-generated? The answer is in the footnote on the next page.

*Corresponding author: `zkachwal@iu.edu`

Preprint. Under review.

1 Introduction

Rapid advancements in image generation have led to a surge in synthetic images (deepfakes) [1, 2, 3]. Improved techniques now enable easier and cheaper production of high-quality visuals [3, 4, 5]. While beneficial for creative applications, this progress allows malicious actors to create convincing forgeries (e.g., face swaps, synthetic photos) nearly indistinguishable from real ones [6, 7, 8]. Such forgeries facilitate impersonation, copyright infringement, and disinformation, necessitating 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 wide adoption [13]. Supervised methods, particularly feature-based ones, struggle with generalization to new generators [14]. Conversely, pre-trained Vision-Language Models (VLMs) demonstrate strong generalization in various tasks [15, 16, 4], and prompting techniques can further boost their performance [17, 18].

In this work, we investigate the use of pre-trained Vision-Language Models (VLMs) for the zero-shot detection of AI-generated images. While off-the-shelf VLMs exhibit some task-specific reasoning and chain-of-thought prompting offers gains, we show that task-aligned prompting elicits more focused reasoning and significantly improves performance without requiring fine-tuning. Specifically, prefixing the model's response with the phrase *"Let's examine the style and the synthesis artifacts"*—a method we call zero-shot style and synthesis (**zero-shot-s**^{2})—guides the VLM to attend more closely to forensic cues relevant for this task.

We evaluate **zero-shot-s**^{2} on three recent, diverse datasets spanning human faces, objects, and animals, with images generated by 16 different models (see Fig. 1[2]). For two widely used open-source VLMs, our approach boosts Macro F1 scores by 8%–29% compared to chain-of-thought prompting—demonstrating strong generalization. We further evaluate the approach across three additional model sizes and observe improvements in most dataset–model combinations—suggesting robustness to model scale.

Surprisingly, self-consistency, a behavior previously observed in language reasoning where aggregating answers from diverse reasoning paths improves performance, also holds in this visual setting [19]. Even here, **zero-shot-s**^{2} scales better with self-consistency than chain-of-thought in most cases—indicating that it elicits more useful diversity. Our findings show that task-aligned prompts elicit more focused reasoning and enhance latent capabilities in VLMs, like the detection of AI-generated images—offering a simple, generalizable, and explainable alternative to supervised methods.

2 Background

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 [20, 21, 22]. As generative models improve, these cues become less reliable. Models trained on fixed artifacts often overfit to specific generators, leading to poor generalization [23, 24, 25].

Frequency-domain techniques analyze spectral representations using tools like the Fast Fourier Transform (FFT) or Discrete Cosine Transform (DCT) [14, 26, 27, 28]. 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 [24, 29, 30].

Spatial-domain approaches examine raw pixel patterns to detect structural or textural inconsistencies [30, 31]. 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 [21, 32].

To improve generalization, recent work has explored larger and more diverse training datasets [23, 33, 34, 35, 36, 37, 38], as well as architectural modifications that target universal artifacts such as upsampling patterns [30, 39]. Fixed-feature backbones and approaches like DiffusionFake have also improved robustness to novel generators [25, 40].

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

Figure 2: Illustration of three prompting strategies for zero-shot detection of AI-generated images using a VLM (Qwen2.5-7B). Input text is marked in `grey`, response text in `blue`. (a) A standard user query `Is this image real or AI-generated?` results in the incorrect response `real`. (b) Inserting the chain-of-thought phrase `Let's think step by step` as a prefix to the response elicits reasoning, but the classification remains incorrect. (c) Inserting our proposed phrase `Let's examine the style and the synthesis artifacts` leads to the correct classification: `ai-generated`. Full reasoning traces for all three methods in the appendix (Figs. 6, 7, 8).

An emerging alternative lies in Vision-Language Models (VLMs), which 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 these models using natural language instructions [41]. Chain-of-thought prompting, which encourages step-by-step reasoning, has been especially effective in language models and increasingly in multimodal settings [42, 43, 44].

Building on this, we investigate whether prompt-based reasoning in Vision-Language Models can support generalizable, zero-shot detection of AI-generated images—without task-specific training. We hypothesize that this approach offers a scalable, interpretable alternative to supervised methods, and greater resilience to emerging generative techniques.

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. We now describe our zero-shot prompting approach and experimental setup.

3.1 Prompts

We evaluate three zero-shot prompting strategies using Vision-Language Models (VLMs), which typically consist of a *user* field for inputs and an *assistant* field for model-generated responses. Our methods involves programmatic text insertion into the *assistant* field—possible only with open-source models—allowing us to interrupt and steer the model's generation process[3]. Each strategy guides the model's behavior by inserting text into the model's input (`grey`) and output (`blue`) fields. In additional experiments, we also assess the effect of placing prompt text in the *user* field instead of the *assistant* field.

The **zero-shot** baseline directly queries the VLM regarding the authenticity of an image. Each image is presented 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 binary label,

[3]This makes our methods incompatible with API-based VLMs, which restrict access to intermediate outputs.

we insert a follow-up prompt into the assistant field: `Final Answer(real/ai-generated):`. This strategy involves two interactions: the initial user query and the binary-label elicitation.

The **zero-shot-cot** variant encourages explicit reasoning. It uses the same initial question in the user field and inserts the phrase for chain-of-thought prompting *"Let's think step by step"* into the assistant field [43]. The model's output then begins with: `Assistant: Let's think step by step` encouraging a stepwise reasoning process before classification (Fig. 2b). The final label is again elicited using the same follow-up prompt, resulting in three interactions: the user query, the insertion of the reasoning phrase, and the label request.

Building on prior work that highlights the importance of synthesis artifacts in detection [45], we introduce **zero-shot-s**[2], a task-aligned prompt that extends **zero-shot-cot** by directing the model's attention toward forensic visual cues. We insert the phrase *"Let's examine the style and the synthesis artifacts"* as a prefix to the assistant's response (Fig. 2c), producing outputs such as: `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 the baseline. Like **zero-shot-cot**, this method involves three interactions. We also evaluate variations of this prompt to examine its effect on reasoning quality and detection performance.

3.2 Data

We conduct experiments using three diverse and recent datasets that span a broad spectrum of real and AI-generated images.

D3 is a benchmark dataset introduced as part of the Contrastive Deepfake Embeddings (CoDE) framework [29]. 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 [46], Stable Diffusion v1.4 and v2.1 [3], and Stable Diffusion XL [47]. 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 [36]. It includes outputs from state-of-the-art models such as Collaborative Diffusion [48], Midjourney, StyleCLIP [49], StarGAN v1 and v2 [50, 51], and WhichFaceIsReal [52]. The dataset spans a wide demographic range, including variations in age, gender, ethnicity, and facial pose. We 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 [53], 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 [54], GLIDE [55], Midjourney, VQDM [56], 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 analysis. We refer to the main evaluation set as GenImage, and the smaller subset as GenImage(2k).

3.3 Models

We conduct experiments using two families of Vision-Language Models (VLMs): **Qwen2.5-VL**[4] and **Llama-3.2-Vision**[5]. These models were selected for their broad adoption; at the time of our experiments, both ranked among the most downloaded VLMs on Hugging Face. We use instruction-tuned variants in evaluation mode and disable token sampling (`do_sample=False`)

[4] `https://huggingface.co/collections/Qwen/qwen25-vl-6795ffac22b334a837c0f9a5`

[5] `https://huggingface.co/collections/meta-llama/llama-32-66f448ffc8c32f949b04c8cf`

during the main evaluation to ensure reproducibility. Token generation is capped at 300 tokens (`max_new_tokens=300`). A single run of zero-shot-cot or zero-shot-s^2 on a NVIDIA A100 80GB for 10,000 images takes approximately 4 hours with Qwen2.5-7B (batch size 20, FlashAttention-2) and 12 hours with Llama-3.2-11B (batch size 10).

Qwen2.5-VL employs a native dynamic-resolution Vision Transformer trained from scratch using Window Attention [57]. For the main evaluation, we use the 7B-Instruct variant, which sees over 3.3M monthly downloads. To assess robustness to model scale, we additionally evaluate three other instruction-tuned variants: 3B, 32B, and 72B.

Llama-3.2-Vision incorporates a separately trained vision adapter that interfaces with a pre-trained Llama 3.1 language model. The adapter consists of cross-attention layers that project image encoder outputs into the language model's hidden states. We use the 11B-Instruct version for our main evaluation, which receives approximately 520K monthly downloads.

As a supervised baseline, we include the recently introduced Contrastive Deepfake Embeddings (**CoDE**), trained on 12 million images from the D3 dataset [29]. We select CoDE for its strong generalization capabilities and ease of reproducibility via Hugging Face[6].

3.4 Lexical Analysis of Elicited Reasoning

One of the key advantages of using Vision-Language Models (VLMs) over traditional Convolutional Neural Networks (CNNs) is their ability to generate interpretable textual explanations. To compare the differences in elicited reasoning across the three prompting strategies—**zero-shot**, **zero-shot-cot**, and **zero-shot-s**^{2}—we perform a lexical analysis using the log-odds ratio, a standard method for identifying words uniquely associated with a given corpus [58]. We begin by collecting all responses from Qwen2.5-7B during the main evaluation across the D3, DF40, and GenImage datasets, yielding approximately 26,736 responses per method. Each response corpus is preprocessed by removing stop words, converting to lowercase, tokenizing, and lemmatizing the text.

After preprocessing, the combined corpus contains roughly 5 million words, with 13,498 unique tokens. The zero-shot subset includes approximately 900,000 words (9,132 unique); zero-shot-cot has around 2.2 million words (10,885 unique); and zero-shot-s^2 consists of 2 million words (7,975 unique).

For each word, we compute the log-odds of its occurrence in the target corpus compared to a combined background corpus formed from the other two subsets. This log-odds ratio is then scaled by its standard deviation to yield a z-score. A high positive z-score indicates that a word is significantly more representative of a specific prompting strategy compared to the others. Finally, we use the top distinctive words identified by this analysis to visualize their frequency distribution across the original responses for each prompting strategy.

3.5 Self-consistency of Elicited Reasoning

This is a decoding strategy in which multiple responses are sampled from a model and their answers are aggregated, typically via majority voting. Originally proposed to improve reasoning in mathematical and multi-step reasoning tasks, this approach leverages response diversity to enhance robustness and accuracy [19]. Unlike standard greedy decoding—which selects the most probable next token and yields deterministic outputs—sampling introduces stochasticity, enabling the model to explore alternative reasoning paths.

We hypothesize that the detection of AI-generated images may similarly benefit from response diversity. Just as there are multiple ways to reason through a math problem, there may be multiple perceptual cues or interpretive strategies when analyzing an image. To test this, we evaluate Llama-3.2-11B on three datasets—D3(2k), DF40(2k), and GenImage(2k)—under three prompting strategies: **zero-shot**, **zero-shot-cot**, and **zero-shot-s**^{2}.

We exclude Qwen2.5-7B from this analysis, as it is not tuned for sampling. Its default configuration employs strongly deterministic (greedy) decoding, which is incompatible with our diversity-based evaluation. In contrast, Llama-3.2-11B supports sampling without restriction in its default settings.

[6] `https://huggingface.co/aimagelab/CoDE`

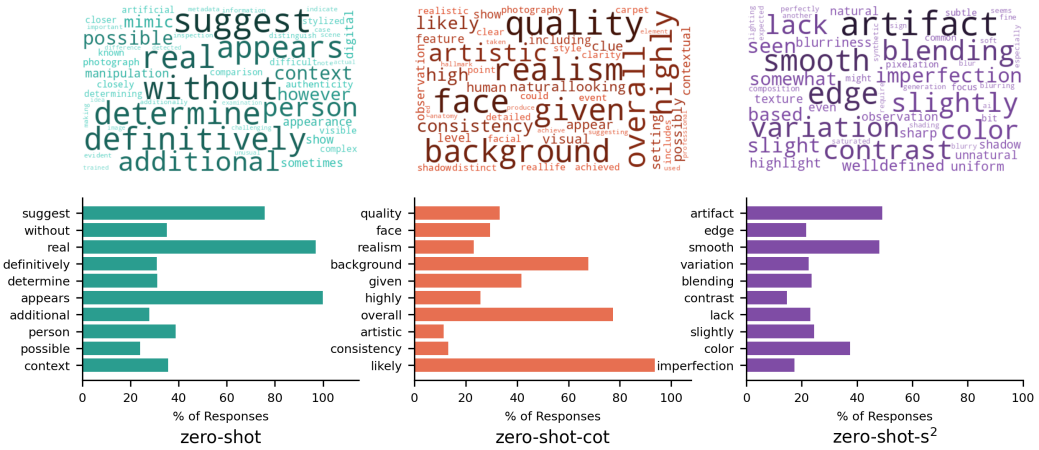

Figure 3: Vocabulary differences in elicited reasoning for Qwen2.5-7B across prompting methods. Word clouds show distinctive words based on z-scores from log-odds ratios relative to the other two methods; word size and color intensity reflect z-score magnitude. Bar charts display the top 10 words per method and their frequency (%) in responses. Full distributions in the appendix (Figs. 9 and 10).

Table 1: Recall scores (%) across models, generators, and prompting methods for DF40. The best-performing prompt per model is shown in bold; absolute improvements of zero-shot-s^2 over zero-shot-cot are shown in parentheses. Scores for D3 and GenImage are in the appendix (Tables 5, 6).

Model	Method	real	cdif	midj	sclip	sgan1	sgan2	wfir
qwen2.5	zero-shot	**98.7**	3.4	5.4	28.6	18.6	1.1	17.9
	zero-shot-cot	87.9	14.1	14.6	49.2	34.1	4.8	36.6
	zero-shot-s^2	73.2 (-14.7)	**53.2** (+39.1)	**23.4** (+8.8)	**78.9** (+29.7)	**62.9** (+28.8)	**11.4** (+6.6)	**81.9** (+45.3)
llama3.2	zero-shot	**98.8**	3.4	37.9	10.8	14.5	4.9	49.9
	zero-shot-cot	79.0	27.6	20.8	49.1	47.5	32.0	65.8
	zero-shot-s^2	54.5 (-24.5)	**67.3** (+39.7)	**65.3** (+44.5)	**67.3** (+18.2)	**70.7** (+23.2)	**70.0** (+38.0)	**76.2** (+10.4)
CoDE	trained on D3	88.2	46.1	38.0	66.5	4.4	25.2	0.7

For all three prompting strategies, we enable sampling during the initial free-form reasoning step (`do_sample=True`, `temperature=1.0`), while disabling other constraints (e.g., `top_p=None`, `top_k=None`). As in the main evaluation, token generation is capped at 300 tokens (`max_new_tokens=300`). Sampling is disabled during the final answer extraction step.

To evaluate how performance scales with response diversity, we generate 5, 10, and 20 reasoning paths per input and aggregate their outputs using majority voting—that is, selecting the most frequently predicted label.

4 Results

Lexical Analysis We begin by analyzing the lexical differences in the reasoning elicited by the three prompting strategies for detecting AI-generated images (Fig. 3). The word cloud presents the top 50 distinctive words used in responses under zero-shot, zero-shot-cot, and zero-shot-s^2 prompting, relative to each other. The bar plot shows the frequency of the top 10 words across responses for each method. Despite using the same underlying model (Qwen2.5-7B), the vocabulary shifts toward more forensic-focused language. This emphasis increases from zero-shot to zero-shot-cot, with terms such as *background*, *realism*, *quality*, and *consistency*, and sharpens further under zero-shot-s^2, which introduces terms like *edge*, *smooth*, *variation*, *blending*, *contrast*, and *imperfection*. We observe a similar trend with Llama-3.2-11B and include the full distributions for both models in the appendix (Figs. 9, 10). Reasoning traces for all three methods can also be found in the appendix (Figs. 6, 7, 8).

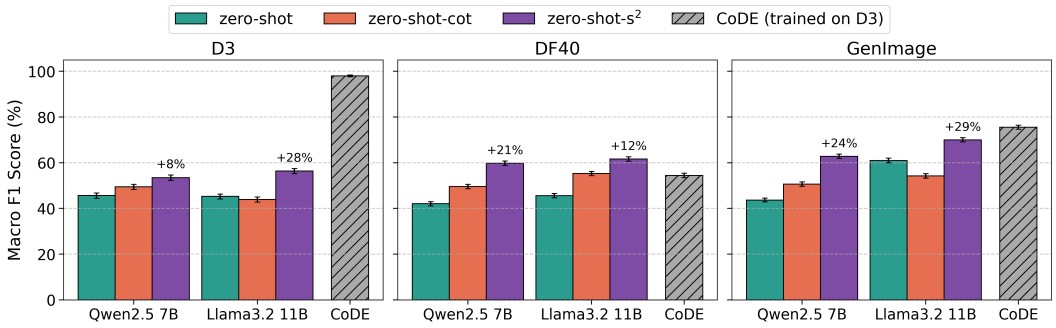

Figure 4: Detection performance (Macro F1, %) across models, datasets, and prompting methods. Results are compared to the supervised CoDE (trained on D3). Bars are annotated to show relative improvements of zero-shot-s^2 over zero-shot-cot and 2-sigma error bars from 1000 bootstrap iterations.

Table 2: Detection performance (Macro F1, %) across model sizes, datasets, and prompting methods for Qwen2.5 family. Best-performing prompt per model is shown in bold; improvements of zero-shot-s ^{2} over zero-shot-cot are indicated in parentheses.

Model	Size	Method	D3 (2k)	DF40 (2k)	GenImage (2k)
qwen2.5	3B	zeroshot	**61.5**	48.8	**74.0**
		zeroshot-cot	52.8	51.8	59.4
		zeroshot-s^2	55.7 (+2.9)	**53.0** (+1.2)	64.2 (+4.8)
	7B	zeroshot	46.9	42.0	44.4
		zeroshot-cot	49.2	49.2	52.9
		zeroshot-s^2	**54.6** (+5.4)	**59.8** (+10.6)	**62.6** (+9.7)
	32B	zeroshot	51.7	60.1	56.5
		zeroshot-cot	45.9	60.6	54.5
		zeroshot-s^2	**55.7** (+9.8)	**62.3** (+1.7)	**69.8** (+15.3)
	72B	zeroshot	48.5	65.2	50.9
		zeroshot-cot	**53.0**	62.7	53.6
		zeroshot-s^2	52.5 (-0.5)	**68.1** (+5.4)	**55.9** (+2.3)

Detection Performance We next assess how this shift in reasoning vocabulary translates into detection performance. Across all three datasets, zero-shot-s^2 consistently outperforms zero-shot-cot (Fig. 4), with relative improvements ranging from 8% to 29%. Notably, on the DF40 dataset, when paired with either Qwen2.5-7B or Llama-3.2-11B, zero-shot-s^2 surpasses CoDE—a supervised model trained on 12 million D3 images. On GenImage, zero-shot-s^2 performs competitively, approaching CoDE's score. We also examine recall scores across individual generators in the DF40 dataset (Table 1). As we move from zero-shot to zero-shot-cot to zero-shot-s^2, recall on *real* images decreases, while recall on *AI-generated* images increases substantially. Zero-shot-s^2 achieves the highest AI-image recall for nearly all generators, with absolute improvements ranging from 6.6% to 45.3%—for both Qwen2.5-7B and Llama-3.2-11B—and even outperforms CoDE on most of them. We observe a similar trend across the other two datasets (Tables 5, 6, appendix). These results suggest that task-aligned prompting (zero-shot-s^2) generalizes more reliably than chain-of-thought prompting (zero-shot-cot) for detecting AI-generated images, regardless of the generator used.

Robustness Across Model Size We then assess how detection performance varies with model size (Fig. 2). A notable exception emerges with Qwen2.5-3B, where the baseline zero-shot prompt achieves the highest overall performance—surpassing both zero-shot-cot and zero-shot-s^2, as well as larger model variants under all prompting strategies. Interestingly, we also do not observe consistent performance gains from increasing model size—contrary to trends typically seen in natural language tasks. Although official parameter breakdowns are unavailable, Qwen2.5 documentation and model files indicate that the vision encoder remains fixed at approximately 0.5B parameters across all model sizes, with scaling occurring almost entirely in the text decoder. Despite this, zero-shot-s^2

Table 3: Detection performance (Macro F1, %) across datasets and prompt phrasings for Qwen2.5-7B. Best-performing phrase per dataset is shown in bold.

Category	Prompt Phrase	D3 (2k)	DF40 (2k)	GenImage (2k)
Open-ended	`Let's visualize`	48.7	45.5	48.4
	`Let's examine`	47.4	52.3	48.5
	`Let's examine pixel by pixel`	48.2	54.3	53.4
	`Let's zoom in`	51.5	58.8	54.0
	`Let's examine the flaws`	**57.8**	39.5	54.6
	`Let's examine the textures`	49.9	54.6	55.1
	`Let's examine the style`	48.7	53.9	55.2
	`Let's think step by step`	49.2	49.2	52.9
Task-aligned	`Let's examine the synthesis artifacts`	55.9	59.7	63.6
	`Let's examine the synthesis artifacts and the style`	54.7	61.1	63.9
	`Let's observe the style and the synthesis artifacts`	55.8	**62.5**	65.4
	`Let's inspect the style and the synthesis artifacts`	54.2	60.6	63.9
	`Let's survey the style and the synthesis artifacts`	52.2	54.7	60.4
	`Let's scrutinize the style and the synthesis artifacts`	56.3	60.2	**66.0**
	`Let's analyze the style and the synthesis artifacts`	52.3	59.0	61.7
	`Let's examine the details and the textures`	50.5	56.4	54.2
	`Let's examine the style and the synthesis artifacts`	54.6	59.8	62.6

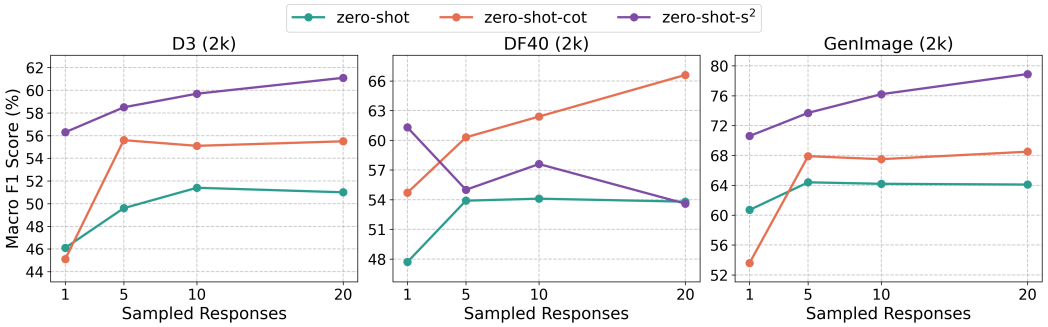

Figure 5: Detection performance (Macro F1,%) across different number of sampled responses for Llama3.2-11B.

outperforms zero-shot-cot across nearly all model–dataset combinations, indicating that task-aligned prompting remains beneficial when additional reasoning capacity is available.

Table 4: Detection performance (Macro F1, %) by phrase location for Qwen2.5-7B. Best-performing prompt per dataset is shown in bold.

Location	Full Prompt	D3 (2k)	DF40 (2k)	GenImage (2k)
None	`User: [Image] Is this image real or AI-generated?` `Assistant`	46.9	42.0	44.4
User	`User: [Image] Is this image real or AI-generated? Let's think step by step` `Assistant`	45.5	51.9	49.4
	`User: [Image] Is this image real or AI-generated? Let's examine the style and the synthesis artifacts` `Assistant`	50.1	52.4	54.3
Assistant	`User: [Image] Is this image real or AI-generated?` `Assistant Let's think step by step`	49.2	49.2	52.9
	`User: [Image] Is this image real or AI-generated?` `Assistant Let's examine the style and the synthesis artifacts`	**54.6**	**59.8**	**62.6**

Robustness Across Prompts We also test the robustness of zero-shot-s^2 to changes in prompt phrasing (Table 3). While some sensitivity to wording exists, task-aligned prompts outperform open-ended ones like *"Let's think step by step"* used in zero-shot-cot. Additionally, we evaluate the impact of phrase location (Table 4). Placing the phrase in the assistant's response—as opposed to the user input—leads to significantly better performance for both zero-shot-cot and zero-shot-s^2. Together, these results suggest that task-aligned prompting improves performance across phrase variations, with the greatest gains occurring when the phrase is inserted into the model's response.

Scaling with Self-Consistency Finally, we evaluate whether the prompting strategies benefit from increased response diversity via self-consistency (Fig. 5). On the DF40 dataset, we observe an exception: increasing the number of sampled reasoning paths leads to a decline in performance for zero-shot-s^2. Analysis of the confusion matrix shows that, at higher sample counts, the model tends to over-analyze face-specific features, resulting in reduced recall on real images. However, across most datasets and sample sizes (5, 10, 20), sampling and aggregating responses improves detection performance for all prompting strategies. To the best of our knowledge, this is the first demonstration of self-consistency benefits in the visual domain—extending findings previously observed in natural language and mathematical reasoning tasks [19]. Zero-shot-s^2 scales more effectively than zero-shot-cot in the majority of cases, suggesting that task-aligned prompts elicit more focused and useful diversity under self-consistency.

5 Discussion and Conclusion

Our results show that Vision-Language Models (VLMs), when paired with the right prompts, can effectively detect AI-generated images—without task-specific fine-tuning. We introduce **zero-shot-s^2**, a simple task-aligned prompt that outperforms chain-of-thought prompting across most models, datasets, and sizes, and even surpasses a supervised baseline in some settings.

While chain-of-thought prompting offers a strong baseline by encouraging structured reasoning, incorporating domain-specific cues—such as synthesis artifacts—yields further gains. This highlights prompting not as a fixed technique, but as a flexible interface for aligning general-purpose models with task-specific goals.

Self-consistency, previously explored in language and math reasoning, also improves performance in the visual domain—especially when combined with task-aligned prompts. To our knowledge, this is the first demonstration of self-consistency benefiting visual detection.

Future prompting strategies could embed generator- or dataset-specific cues, offering a lightweight path to adaptation. Prompt-based methods remain highly scalable, interpretable, and adaptable—an asset as generative models continue to evolve.

Overall, our findings suggest that prompt engineering is a powerful tool for steering VLMs toward robust, generalizable, and interpretable image forensics.

6 Limitations

A primary limitation of VLMs is their computational cost. Compared to traditional models like CNNs or ViTs, Vision-Language Models (VLMs) generate longer token sequences, require more memory, and consequently perform slower inference. For example, the ViT-based CoDE baseline has only 6 million parameters and runs significantly faster than the VLMs evaluated in this work.

Model scaling also presents challenges. While zero-shot-s^2 outperforms zero-shot-cot across sizes, larger models do not consistently yield better performance. In some cases, such as Qwen2.5-3B, smaller models with fixed vision encoders outperform larger variants—suggesting diminishing returns from scaling the text decoder for vision-centric tasks.

Task-aligned prompts may also degrade performance under self-consistency. On the DF40 dataset, sampling multiple reasoning paths led to reduced recall on real images, likely due to over-analysis of subtle facial features. Our method additionally requires access to intermediate model responses, which is not supported by API-based VLMs.

Finally, LLMs—and by extension VLMs—remain sensitive to prompt phrasing, with behavior often changing under minor textual variations, posing challenges for stability and generalization [59]. Although we selected recent datasets to reduce the likelihood of overlap with pretraining data, the opaque nature of VLM training corpora means potential leakage cannot be ruled out. Moreover, our analysis focuses on single-image detection; extending these findings to video remains an open direction for future work.

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] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-Consistency Improves Chain of Thought Reasoning in Language Models. September 2022.

[20] 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.

[21] 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.

[22] 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.

[23] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A. Efros. CNN-Generated Images Are Surprisingly Easy to Spot... for Now. pages 8692–8701. IEEE Computer Society, June 2020.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

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

[33] Zhiyuan Yan, Yong Zhang, Xinhang Yuan, Siwei Lyu, and Baoyuan Wu. DeepfakeBench: A Comprehensive Benchmark of Deepfake Detection. *Advances in Neural Information Processing Systems*, 36:4534–4565, December 2023.

[34] 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. October 2024.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] 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.

[41] 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.

[42] 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.

[43] 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.

[44] 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].

[45] 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.

[46] DeepFloyd/IF-I-XL-v1.0 · Hugging Face, December 2024.

[47] 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. October 2023.

[48] 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.

[49] 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.

[50] 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.

[51] 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.

[52] Which Face Is Real?

[53] 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.

[54] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large Scale GAN Training for High Fidelity Natural Image Synthesis. September 2018.

[55] 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.

[56] 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.

[57] 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].

[58] Burt L. Monroe, Michael P. Colaresi, and Kevin M. Quinn. Fightin' Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict. *Political Analysis*, 16(4):372–403, January 2017.

[59] Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models, October 2024. arXiv:2410.05229.

A Appendix

A.1 Full Reasoning Traces

This section includes the full reasoning traces for all three methods.

```
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 zero-shot 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 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 zero-shot-cot 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 zero-shot-s^2 reasoning trace for detecting AI-generated images using a VLM (Qwen2.5-7B). Image from Fig. 2.

A.2 Recall Scores

This section shows the recall breakdown for the remaining generators across D3 and GenImage.

Table 5: Effect of prompts on recall scores (%) across image generators in D3. The best prompt for each VLM is shown in bold; improvements of zero-shot-s^2 over zero-shot-cot are shown in parentheses.

Model	Method	real	dfif	sd1.4	sd2.1	sdxl
qwen2.5	zero-shot	**87.1**	27.7	23.0	24.5	69.8
	zero-shot-cot	80.5	39.0	31.4	35.3	70.6
	zero-shot-s^2	65.7 (-14.8)	**57.3** (+18.3)	**43.1** (+11.7)	**47.1** (+11.8)	**78.6** (+8.0)
llama3.2	zero-shot	69.9	44.9	26.1	30.3	65.4
	zero-shot-cot	**85.1**	35.3	22.4	25.9	53.8
	zero-shot-s^2	66.1 (-19.0)	**62.7** (+27.4)	**48.6** (+26.2)	**54.8** (+28.9)	**76.8** (+23.0)
CoDE	trained on D3	97.5	98.7	99.0	99.0	99.3

Table 6: Effect of prompts on recall scores (%) across image generators in GenImage. The best prompt for each VLM is shown in bold; improvements of zero-shot-s^2 over zero-shot-cot are shown in parentheses.

Model	Method	real	adm	bgan	glide	midj	sd1.4	sd1.5	vqdm	wk
qwen2.5	zero-shot	**98.4**	06.0	8.8	14.9	10.9	5.9	6.8	11.3	21.8
	zero-shot-cot	94.0	14.9	19.3	23.3	20.7	16.6	17.3	21.2	32.5
	zero-shot-s^2	84.4 (-9.6)	**47.4** (+32.5)	**55.0** (+35.7)	**56.6** (+33.3)	**35.2** (+14.5)	**29.5** (+12.9)	**31.5** (+14.2)	**51.6** (+30.4)	**49.4** (+16.9)
llama3.2	zero-shot	**97.6**	16.5	25.3	33.4	40.9	34.1	36.1	22.2	51.5
	zero-shot-cot	97.2	14.6	25.0	26.8	18.5	21.1	22.8	23.4	37.7
	zero-shot-s^2	84.7 (-12.5)	**50.8** (+36.2)	**61.9** (+36.9)	**60.9** (+34.1)	**51.3** (+32.8)	**50.8** (+29.7)	**53.3** (+30.5)	**56.5** (+33.1)	**65.8** (+28.1)
CoDE	trained on D3	97.6	14.6	3.9	46.0	55.8	97.8	97.7	24.7	97.6

A.3 Lexical Analysis

This shows the full distribution of top 50 distinct words for Qwen2.5-7B and Llama3.2-11B.

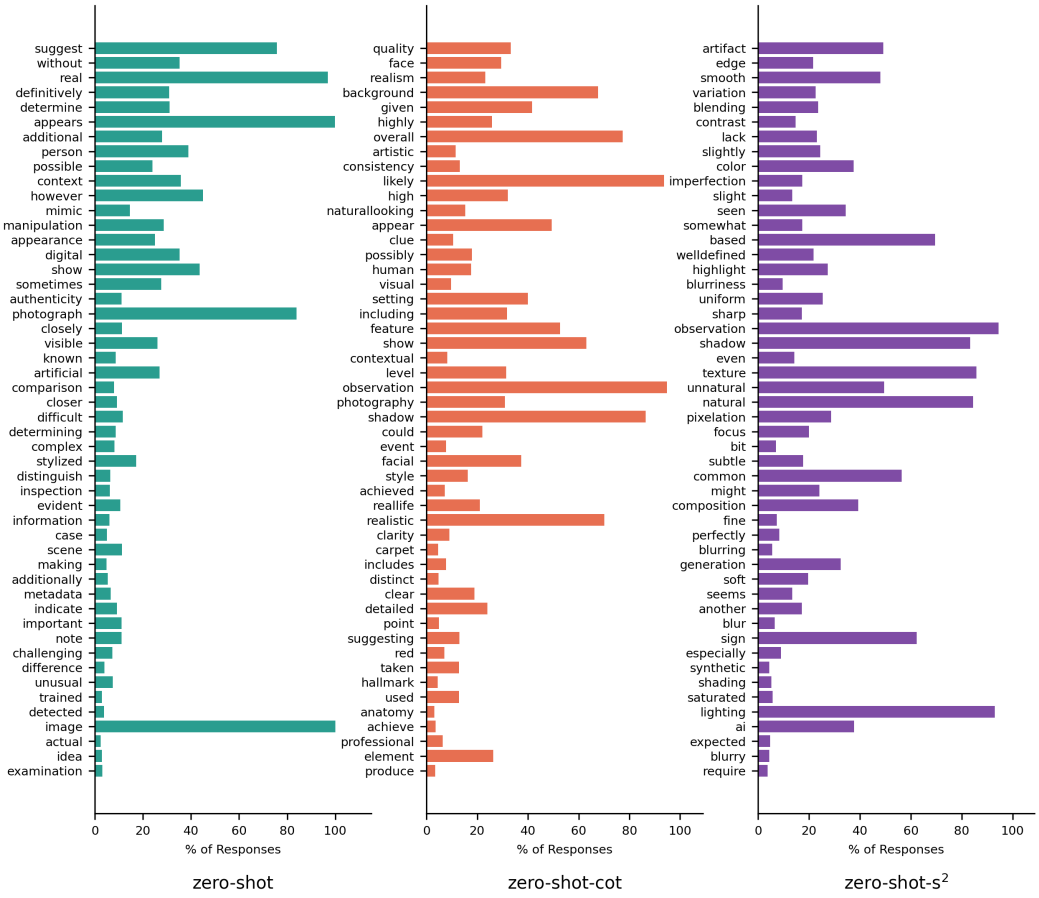

Figure 9: Effect of prompts on elicited reasoning for Qwen2.5-7B. Fifty most distinctive words from the word cloud and their frequency (%) in responses. The distinctive vocabulary is based on the z-scores of the log-odds ratio.

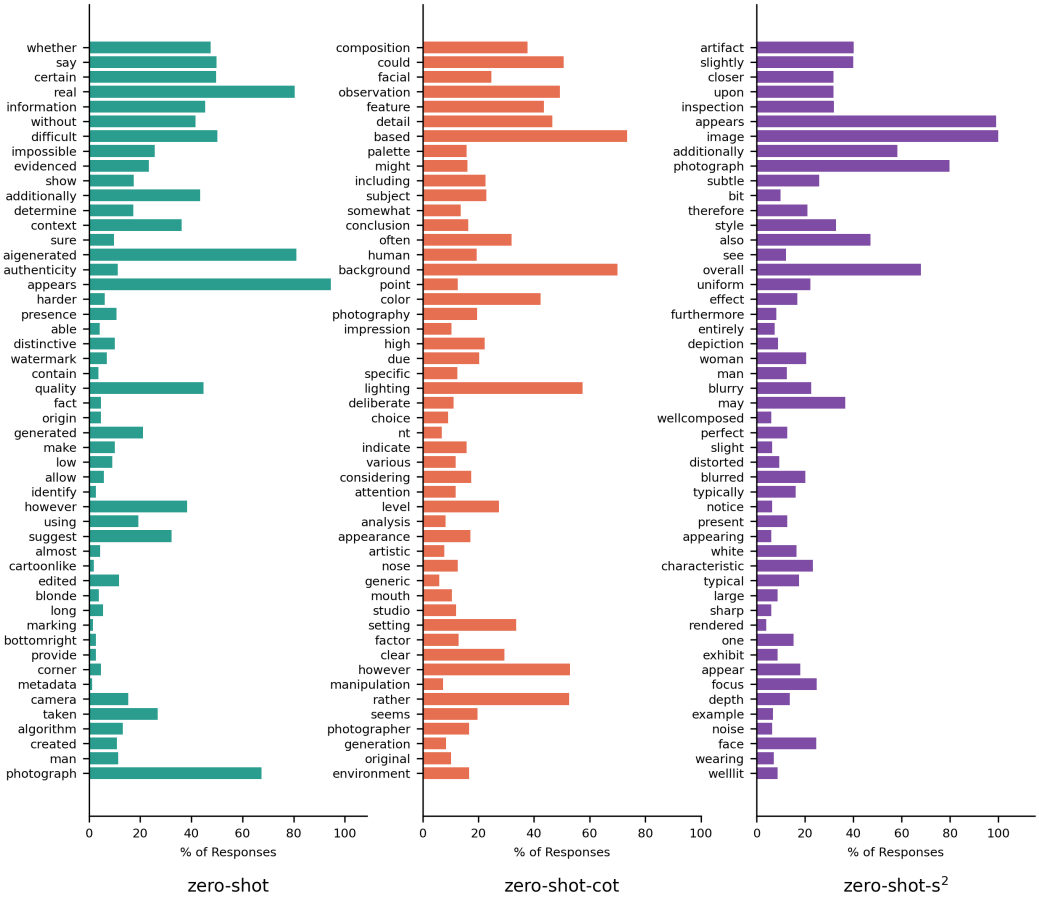

Figure 10: Effect of prompts on elicited reasoning for Llama3.2-11B. Fifty most distinctive words from the word cloud and their frequency (%) in responses. The distinctive vocabulary is based on the z-scores of the log-odds ratio.