SHIELD: A Benchmark Study on Zero-Shot Detection of AI-Edited Images with Vision Language Models

Siyuan Cheng[*]
Purdue University
`cheng535@purdue.edu`

Hanxi Guo
Purdue University
`guo778@purdue.edu`

Zhenting Wang
Rutgers University
`zhenting.wang@rutgers.edu`

Xiangyu Zhang
Purdue University
`xyzhang@purdue.edu`

Lingjuan Lyu[†]
Sony AI
`Lingjuan.Lv@sony.com`

Abstract

The rapid progress of generative AI has enabled powerful image editing tools that can convincingly manipulate localized regions of real images. Such *AI-edited images* are increasingly exploited to spread misinformation, yet existing detectors, which primarily designed for whole-image synthesis or DeepFakes, struggle to generalize and often fail against partial manipulations. In this paper, we study AI-edited image detection from a zero-shot perspective, drawing inspiration from how humans approach the task. Humans are generally reliable because they are exposed primarily to authentic images and treat unusual or inconsistent regions as anomalies. Vision language models (VLMs), trained on large and diverse image–text corpora, offer a scalable analogue to this human ability. Leveraging VLMs for zero-shot inference provides a principled framework for anomaly detection while mitigating the overfitting issues that plague training-based detectors. We present SHIELD, the first benchmark study of zero-shot AI-edited image detection using VLMs. Our evaluation covers 24 models under two prompting strategies (direct prompting and Chain-of-Thought prompting) and two inference modes (greedy decoding and sampling). The results show that detection accuracy generally correlates with overall model capability. Notably, direct prompting with greedy decoding achieves the strongest performance, suggesting a "first impression" effect. We also examine detection performance under different datasets, generative models, and editing methods, and discuss potential directions for improving the detection accuracy. The source code used in this study is available at https://github.com/Megum1/SHIELD.

1 Introduction

Generative AI is transforming the way we create and consume information, with breakthroughs spanning art [56, 49, 1], healthcare [15, 53, 64], and autonomous driving [43, 62, 50]. Among the most striking innovations is text-to-image generation [18, 41]: tools like MidJourney [32] can turn a short description into a vivid, photorealistic image within seconds. Despite these advances, the misuse of image generation technologies has become a growing concern. A major risk lies in fabricating fake identities (e.g., DeepFakes [55, 39, 42]) and producing deceptive content such as fake news, which fuels misinformation and public distrust. For example, it is disturbingly straightforward to generate realistic images that fabricate sensational stories, such as falsely depicting public figures in

---

[*] Work partially done during Siyuan Cheng's internship at Sony AI.

[†] Corresponding Author

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

controversial situations[3]. Such misuse undermines the trustworthiness of generative AI and poses significant challenges for its safe deployment. Hence, developing reliable methods to detect synthetic images is of urgent importance.

Existing research [51, 27, 33, 46] has primarily focused on detecting *whole-image synthesis*, i.e., images entirely generated by AI without any image reference. Efforts have largely focused on DeepFake detection [61, 65, 7] or general fake image detection [59, 8, 60], with detectors showing strong performance when test data closely matches the training distribution. However, these methods struggle to generalize. For instance, models trained on GAN-generated images often fail when applied to diffusion-based images [51]. Their robustness is also limited: simple post-processing operations such as JPEG compression can significantly degrade accuracy [8], suggesting that many detectors overfit to pixel-level artifacts in training data. More critically, recent misinformation increasingly exploits *image editing*, where only localized regions of authentic images are manipulated. Such partial edits are both more practical and more convincing, yet detectors trained on whole-image synthesis often struggle to identify them. Prior studies [36] have shown that detectors designed for whole-image fakes perform poorly when faced with partial manipulations.

In this paper, we investigate a more practical and increasingly prevalent misuse scenario: *AI-edited image detection*. Rather than relying on training-based classifiers, which often suffer from poor generalization, we formulate the task as a *zero-shot anomaly detection* problem. The intuition is similar to how humans identify fake images: by leveraging a strong understanding of authentic content and treating deviations as suspicious. Vision language models (VLMs) [29, 28, 67] are particularly well suited for this setting. Pretrained on massive corpora of images and multimodal tasks such as image captioning [63] and object detection [11], VLMs acquire broad semantic knowledge and naturally support zero-shot inference, reducing the risk of overfitting to specific styles or artifacts. While a few recent works have explored using VLMs for detection, they either incorporate VLMs as components of supervised detectors [65, 5] or focus narrowly on DeepFake detection [37]. By contrast, we conduct a comprehensive investigation into leveraging VLMs as *zero-shot detectors for AI-edited images*, aiming to establish their potential as a practical and generalizable solution to synthetic media detection.

We present SHIELD, the first comprehensive *Benchmark Study on Zero-SHot Detection of AI-Edited Images with Vision Language ModDls*. Specifically, we evaluate 24 VLMs under two prompting strategies and two inference modes. The prompting strategies include direct prompting, where the model is asked to judge whether an image is real or fake, and Chain-of-Thought (CoT) [54] prompting, which encourages step-by-step reasoning before producing an answer. For inference, we consider greedy decoding, which selects the most confident token at each step, and sampling-based decoding, which aggregates predictions from multiple sampled responses. Together, these four combinations mirror how humans may approach the task, ranging from instinctive judgments to deliberate reasoning and from confident single responses to reconsideration through multiple trials.

We demonstrate average detection accuracy across the four evaluation settings and compare the results with the Open VLM Leaderboard [35], a well-established benchmark spanning diverse vision tasks (see Figure 1). The overall trend is nearly consistent: models that rank higher on the leaderboard, which reflects general performance across diverse vision tasks, also tend to achieve stronger zero-shot performance on AI-edited image detection. Notably, we also observe some counterintuitive results. Direct prompting with greedy inference achieves the best performance among the four settings. We attribute this to a "first impression" effect, where the initial high-confidence prediction captures decisive cues for abnormality, while extended reasoning or sampling introduces noise and uncertainty. This phenomenon parallels human decision-making, in which immediate judgments can sometimes outperform deliberate overanalysis. Additional observations and detailed analyses are provided in section 4.

2 Background and Related Work

Synthetic Image Generation and Editing. Recent years have witnessed rapid advances in synthetic image generation driven by deep learning. Early approaches were dominated by variational auto-encoders (VAEs) [21, 22] and generative adversarial networks (GANs) [14, 13]. Since 2020, diffusion models (e.g., DDPM [18], LDM [41], Stable Diffusion [41]) have surpassed these earlier paradigms,

---

[3] For example https://www.bbc.com/news/world-us-canada-65069316

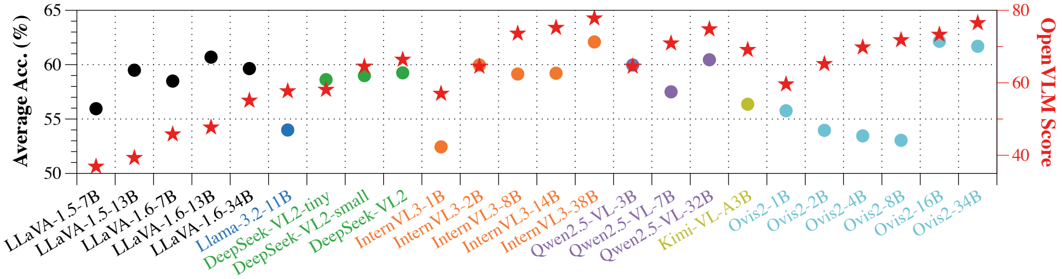


Figure 1: **Comparison between the Open VLM leaderboard and our benchmark SHIELD.** The left y-axis shows the average detection accuracy across four setting, while the right y-axis shows the Open VLM score. The overall trends are largely consistent across the two benchmarks.

achieving state-of-the-art fidelity and, crucially, enabling text-based controllability for both generation and editing. At their core, diffusion models learn to reverse a gradual noising process through a denoising network, thereby generating images from pure noise. Today, they power production-scale applications such as Midjourney [32] and DALL-E [34], supporting diverse styles ranging from photorealistic imagery to animation. In this work, we focus on image editing instead of generation from scratch, where the model receives both an input image and a text description. This can be realized in two ways: (*i*) inpainting [4], where a semantic mask specifies the region to be filled; or (*ii*) prompt-based editing [17], where a textual instruction guides modifications to the original image.

Synthetic Image Detection. The objective of synthetic image detection is to determine whether an image has been generated by AI models. Early research primarily targeted fake face images, known as DeepFake detection [61, 65, 7]. As generative models have advanced and begun producing increasingly high-quality content across diverse domains, detection efforts have expanded to general image categories. Early methods [51] treated the task as typical binary classification, relying on large-scale training sets and carefully designed data augmentations to train deep networks such as ResNet-50 [16]. While these approaches achieved strong performance on in-distribution tests (i.e., images generated by models seen during training), their generalization to out-of-distribution cases remained limited. Subsequent works explored distinctive properties of synthetic images to improve robustness. For example, some methods exploit abnormal patterns in the frequency domain [12, 45], pixel-level artifacts [47, 46, 52], or semantic-level discrepancies compared with real images [33, 30, 6]. More recent approaches integrate multiple types of artifacts to enhance generalizability [59, 8]. In this paper, we focus on *image editing detection*, which differs from whole-image generation detection. Editing detection is often more challenging, as it includes an additional step of localizing modifications within a real image.

Vision-Language Models (VLMs) as Synthetic Image Detectors. VLMs [29, 26, 25] integrate visual information into language model processing, aiming to serve as unified models for diverse vision tasks such as visual question answering (VQA), object detection, and image captioning. Recent works have begun exploring their potential for synthetic image detection. AntifakePrompt [5] was among the first to apply VLMs to deepfake detection by formulating the task as VQA. The authors tuned soft prompts for InstructBLIP [9] using real images and a custom fake dataset containing fully or partially generated images. However, this approach emphasized training for binary classification rather than leveraging zero-shot capabilities. DD-VQA [65] extended deepfake detection into a reasoning-based VQA task. While effective, such training-based approaches remain vulnerable to out-of-distribution data. In contrast, we explore a more reliable, human-inspired zero-shot approach: just as humans detect synthetic content by noticing anomalies against everyday objects, VLMs can leverage their broad visual knowledge to spot abnormalities. A recent work [37] explored zero-shot VLM detection for deepfake faces. In this paper, we study a more challenging problem of detecting image edits across general categories.

3 Design

In this section, we present the experimental design for evaluating the detection performance of VLMs on AI-edited images. An overview is shown in Figure 2, which consists of two main components: (1) Prompting Strategy and (2) Inference Mode. Given an input image, we apply a vision language

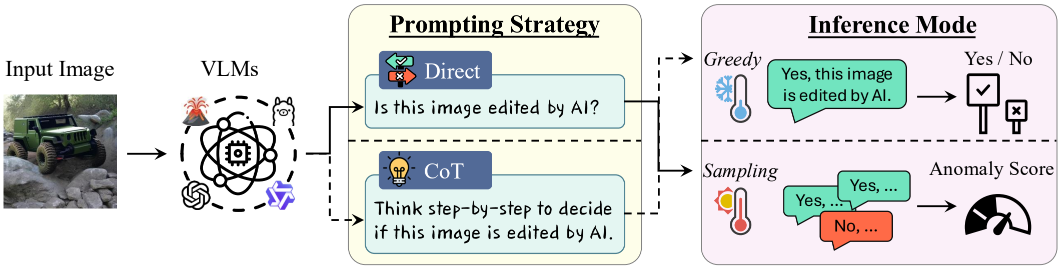

Input Image
VLMs
Prompting Strategy
Direct
Is this image edited by AI?
CoT
Think step-by-step to decide if this image is edited by AI.
Inference Mode
Greedy
Yes, this image is edited by AI.
Yes / No
Sampling
Yes, ...
Yes, ...
No, ...
Anomaly Score

Figure 2: **Overview.** We evaluate the detection performance of VLMs under two prompting strategies and two inference modes to enable a comprehensive assessment.

model (VLM) with a chosen prompt and inference mode to obtain the result. The output can be either a binary decision (0 or 1) or an anomaly score ($[0, 1]$), depending on the inference mode. In the following, we elaborate on these two components in detail.

3.1 Prompting Strategy

As shown in the yellow region of Figure 2, we employ two representative prompting approaches: direct prompting and Chain-of-Thought (CoT) prompting. To facilitate evaluation, we provide light instructions to guide the output format. However, we avoid imposing strict formatting constraints, as these may negatively affect the model's generative ability [44].

Direct Prompting. We directly ask the model to decide whether an image is AI-edited without providing additional reasoning steps. This setting evaluates the model's ability to make a judgment based on its "first impression" rather than a detailed analysis. Surprisingly, this approach yields better results than extended reasoning, as shown in subsection 4.2.

```
Direct Prompt
```

```
Is this image edited or modified by AI? Start your answer with "Yes" or "No".
```

CoT Prompting. Chain-of-thought (CoT) prompting [54] elicits intermediate reasoning before the final decision. By encouraging step-by-step analysis, it can decompose the task into smaller parts and potentially improve accuracy. In our setup, we use a standard CoT instruction that asks the model to reason step by step and then decide whether the input image is AI-edited.

```
CoT Prompt
```

```
Think step-by-step to decide whether this image is edited or modified by AI.
Then conclude your answer with "The answer is" followed by your final answer.
```

3.2 Inference Mode

After selecting a prompt strategy, we select an inference mode before getting the output. The pink part of Figure 2 presents two representative inference modes, namely (1) Greedy and (2) Sampling.

Greedy. This is a greedy way to select subsequent tokens with the highest confidence. Technically, we disable the sampling option and set temperature equal to 0 in the model forwarding function. The greedy inference represent the high confidence output by the model without hesitation and rethinking.

Sampling. In addition to greedy inference, we also evaluate sampling-based inference. Specifically, we generate multiple responses for each input (default: 5) and average the binary outputs to obtain an anomaly score. The sampling temperature is set to 0.7, a commonly used value that balances diversity and stability in generation. Sampling serves as a reference to examine whether introducing stochasticity can improve robustness compared to deterministic greedy decoding.

3.3 Extracting Binary Outputs from Model Responses

With appropriate instructions to constrain the output format, extracting binary predictions from model responses is generally straightforward. For example, under direct prompting, the output typically begins with either "Yes" or "No", which we map to label 1 and 0, respectively. For CoT prompting, we extract the final answer following the phrase "The answer is" and again assign 1 or 0 depending on whether the response begins with "Yes" or "No".

Occasionally, however, models deviate slightly from the prescribed format. In such cases, we employ a secondary judge model to standardize the response and extract a valid binary output. Specifically, we use Llama-3.1-8B as our judge model. We design a prompt into which we insert the `{Raw VLM Response}` and provide three illustrative examples for few-shot guidance. To handle uncertainty, we map the output "Not sure" to label 0, meaning the model considers the input image unedited, as it lacks sufficient evidence of abnormality.

Judge Model Prompt

```
Read the following description and determine the expressed attitude toward
whether the image is edited or modified by AI:
---
{Raw VLM Response}
---

The attitude should be summarized as one of:
- "Yes": the description affirms it is edited or modified by AI
- "No": the description denies it is edited or modified by AI
- "Not sure": the description expresses uncertainty

Examples:
Description: "yes, the image is edited or modified by AI."
Answer: Yes

Description: "the image is not edited by ai"
Answer: No

Description: "can not decide without further information"
Answer: Not sure

Respond only with: "Yes", "No", or "Not sure".
```

4 Evaluation

We conduct extensive experiments to evaluate the ability of different VLMs to detect AI-edited images. Unless otherwise specified, all experiments are run on a server equipped with eight NVIDIA RTX A6000 GPUs.

4.1 Experiment Setup

Dataset. We adopt the recent large-scale dataset *Semi-Truths* [36], which is designed for evaluating AI-augmented image detectors. For a fair yet computationally affordable evaluation, we randomly sample 1,000 real images and 1,000 AI-edited images from three source datasets: ADE20K [66], CelebAHQ [20], and OpenImages [24]. The dataset covers two representative editing types: (1) *image inpainting*, where pixels within a semantic mask are replaced, and (2) *prompt-based editing*, where regions are modified according to textual descriptions. Examples are provided in Appendix B. Five widely used generative models are employed for editing, including Kandinsky 2.2 [40], Stable Diffusion v1.4 [41], Stable Diffusion v1.5 [41], Stable Diffusion XL [2], and OpenJourney [38]. We carefully balance the distribution across editing types and models to ensure diversity and fairness.

Vision-Language Models (VLMs). We benchmark 24 open-source VLMs spanning seven representative families: LLaVA [29, 28], Llama-3.2 [10], DeepSeek-VL2 [57], InternVL3 [67], Qwen2.5-VL [3], Kimi-VL [48], and Ovis2 [31]. A complete list of model sources can be found in Appendix A.

Table 1: **Detection Performance of Open-Source VLMs on AI-Edited Images.** Higher values indicate better performance for Acc., TPR, AUC, and TPR@, while lower values are better for FPR. All metrics are reported in %, and computation time is measured in seconds. The best result in each column is highlighted in **bold underline**, and the average across models is reported in the last row.

VLMs	*Direct Prompting*								*CoT Prompting*							
	Greedy				**Sampling**				**Greedy**				**Sampling**			
	Acc.	TPR	FPR	Time	Acc.	AUC	TPR@	Time	Acc.	TPR	FPR	Time	Acc.	AUC	TPR@	Time
LLaVA-1.5-7B	60.30	40.10	19.50	0.18	58.05	59.95	20.74	0.81	51.40	58.90	56.10	3.26	54.05	53.83	13.96	17.32
LLaVA-1.5-13B	66.05	60.80	28.70	0.30	62.50	65.97	27.33	1.30	55.30	16.10	5.50	8.34	54.10	60.36	19.88	39.73
LLaVA-1.6-7B	**67.15**	69.80	35.50	1.25	61.75	65.87	26.24	5.54	52.50	8.00	**3.00**	14.98	52.55	59.22	21.79	75.57
LLaVA-1.6-13B	66.05	51.70	19.60	1.38	**65.45**	**69.93**	31.36	6.28	56.85	17.50	3.80	16.17	54.40	62.20	23.05	109.04
LLaVA-1.6-34B	65.55	82.30	51.20	158.9	64.00	66.98	21.88	774.6	55.50	16.00	5.00	838.7	53.50	65.50	33.00	4640.6
Llama-3.2-11B	54.90	11.00	12.00	1.90	55.75	61.76	30.30	10.00	50.60	22.20	21.00	13.78	54.70	58.48	16.90	71.11
DeepSeek-VL2-tiny	61.85	47.10	23.40	0.54	61.55	65.03	26.65	2.53	53.35	37.30	30.60	2.56	57.75	59.41	20.47	12.92
DeepSeek-VL2-small	57.50	19.60	4.60	0.48	58.00	61.60	25.11	2.20	58.25	60.40	43.90	10.52	62.20	66.71	25.48	47.38
DeepSeek-VL2	58.75	48.70	31.20	0.79	59.40	62.88	23.10	3.69	57.55	73.80	58.70	15.64	61.30	**66.91**	28.22	73.13
InternVL3-1B	50.50	99.20	98.20	1.65	52.30	58.10	12.65	7.17	53.75	28.10	20.60	17.89	53.20	57.25	16.06	64.16
InternVL3-2B	63.50	44.00	17.00	1.69	61.55	66.60	30.69	7.36	59.55	31.80	12.70	15.65	55.30	59.96	20.60	81.42
InternVL3-8B	61.80	31.50	7.90	3.72	60.85	66.99	31.09	17.70	53.65	48.40	41.10	25.74	60.25	63.63	23.19	119.09
InternVL3-14B	59.65	24.10	4.80	4.84	60.05	64.06	31.42	21.29	59.15	35.70	17.40	51.84	58.00	62.58	24.10	131.52
InternVL3-38B	62.35	38.50	13.80	4.72	61.05	63.60	30.92	20.90	**62.40**	41.20	16.40	26.89	**62.55**	66.41	32.55	132.14
Qwen2.5-VL-3B	64.70	42.00	12.60	0.57	60.00	63.95	29.53	3.31	57.15	60.10	45.80	5.97	58.05	60.92	19.93	33.87
Qwen2.5-VL-7B	57.80	17.70	**2.10**	0.44	57.75	62.72	30.94	2.01	56.35	27.80	15.10	7.91	58.10	63.72	28.45	37.84
Qwen2.5-VL-32B	60.05	23.60	3.50	9.60	59.65	63.29	**35.13**	14.47	60.35	32.30	11.60	48.20	61.75	65.70	33.98	48.47
Kimi-VL-A3B	57.95	22.90	7.00	0.58	58.20	60.95	25.07	2.87	53.20	39.10	32.70	15.64	56.10	59.32	18.16	76.36
Ovis2-1B	50.15	**100.00**	99.70	0.68	50.80	56.47	11.70	4.71	59.85	37.50	17.80	2.60	62.25	66.62	28.86	16.12
Ovis2-2B	50.05	**100.00**	99.90	0.62	50.30	51.64	10.36	4.50	55.90	65.10	53.30	4.66	59.60	62.76	21.25	22.63
Ovis2-4B	52.15	98.10	93.80	2.52	52.10	54.29	11.04	10.08	54.65	**78.40**	69.10	6.26	54.90	57.66	**13.21**	36.70
Ovis2-8B	50.40	74.90	74.10	3.55	50.75	51.43	10.29	13.43	56.95	64.90	51.00	7.85	54.05	57.46	15.73	42.54
Ovis2-16B	63.75	44.50	17.00	6.67	63.40	66.83	32.17	32.27	60.70	29.30	7.90	19.59	60.75	65.52	35.58	90.93
Ovis2-34B	63.60	51.60	24.40	7.55	62.45	65.41	31.32	41.93	60.20	26.20	5.80	24.74	60.50	63.03	31.35	112.75
Average	59.44	51.82	33.40	8.96	58.65	62.35	24.88	42.12	56.46	39.84	26.91	50.22	57.50	61.88	23.57	255.55

Our evaluated model sizes range from 1B to 38B parameters. Due to resource constraints, we exclude models exceeding 40B, though small- and medium-scale models already provide sufficient insights. All models are instruction-tuned, ensuring that they follow the designed prompts and output formats, which facilitates consistent measurement. For comparison, we also include a production-level model, GPT-4o [19]. Each VLM is evaluated under two prompting strategies (i.e., direct and CoT) and two inference modes (i.e., greedy and sampling). Detailed prompts are provided in subsection 3.1.

Evaluation Metrics. We adopt six metrics to assess detection performance. Under greedy inference, where outputs are binary (either true or false), we measure accuracy, true positive rate (TPR), and false positive rate (FPR). Under sampling inference, which yields confidence scores in $[0, 1]$, we compute accuracy (with threshold equals to 0.5), ROC-AUC (AUC), and TPR at 10% FPR (TPR@) which better reflect practical deployment. In addition, we report the average detection time per image for each VLM to quantify computational overhead.

4.2 Detection Performance of Various Open-Source VLMs

In this section, we present the main experimental results on AI-edited image detection using a diverse set of open-source VLMs. Overall, their average detection accuracy closely mirrors their general vision performance, as shown in Figure 1. To gain deeper insights, Table 1 reports results under four settings, involving two prompting strategies and two inference modes, respectively (see Appendix C for illustrative cases). We summarize the key insights from these evaluations as the following findings.

> **Finding 1:** *Direct prompting with greedy inference yields the best overall performance.*

As shown in the last row of Table 1, direct prompting with greedy inference achieves the highest accuracy (59.44%), outperforming the other three settings. This result is somewhat counter-intuitive, since chain-of-thought (CoT) prompting is often reported to enhance reasoning in diverse tasks [54, 58]. We hypothesize that AI-edited image detection fundamentally differs from reasoning-intensive problems: it relies less on multi-step deliberation and more on immediate perceptual judgment. In this context, "overthinking" may actually harm performance, e.g., CoT prompting introduces additional uncertainty, often leading to higher FPR (e.g., DeepSeek family) or overly conservative responses with reduced TPR (e.g., LLaVA family).

This phenomenon suggests that detection is closer to a "*first impression*" task, similar to how humans can often detect anomalies at a glance, whereas prolonged reflection may amplify doubt about borderline cases. From a technical point of view, greedy direct prompting is also less susceptible to drift from verbose text generation: the model's decision (yes/no) appears immediately after the image tokens, while CoT defers the final answer to the end of a long reasoning process, making it more vulnerable to noise or self-contradiction. Likewise, sampling further injects randomness, moving the prediction away from the model's most confident "*first impression*".

> **Finding 2:** *The LLaVA family achieves the best detection performance.*

We observe that models in the LLaVA family outperform other VLM families in detecting AI-edited images. This is also somewhat surprising, as prior large-scale benchmarks [35] often place LLaVA behind families such as InternVL or Qwen on general-purpose vision-language tasks. We suspect there are two possible reasons for this result. First, LLaVA's streamlined architecture directly connects visual features to the language model, enabling rapid "first impression" judgments rather than relying on complex reasoning. Second, LLaVA is primarily optimized for grounded perception and description rather than multi-step reasoning. Since AI-edited image detection resembles a perceptual anomaly recognition task, this design makes LLaVA well-suited for the problem.

> **Finding 3:** *Larger models typically achieve higher detection performance.*

Within the same family, we observe that larger models typically outperform their smaller counterparts, as exemplified by the InternVL3 and Ovis2 families. This trend is expected, since larger models possess greater parameter capacity and broader knowledge, enabling them to capture subtle cues in an image even at a "first impression." The pattern is also similar to human perception. For example, adults (like larger models) generally detect anomalies more reliably than children (like smaller models), reflecting their greater experience and cognitive capacity.

> **Finding 4:** *Sampling slightly reduces performance under direct prompting but provides minor gains under CoT prompting.*

We observe that sampling inference tends to degrade performance for direct prompting, yet offers small improvements for CoT prompting. For instance, within the Qwen2.5 family, sampling lowers accuracy by about 2% under direct prompting but raises it by roughly 1% under CoT prompting. This pattern can also be explained using "first impression" hypothesis: sampling introduces additional uncertainty, which undermines the decisiveness of direct prompting. However, in the more cautious CoT setting, sampling can mitigate hallucinations and provide slight performance improvement.

> **Finding 5:** *Smaller models exhibit a strong bias toward positive predictions.*

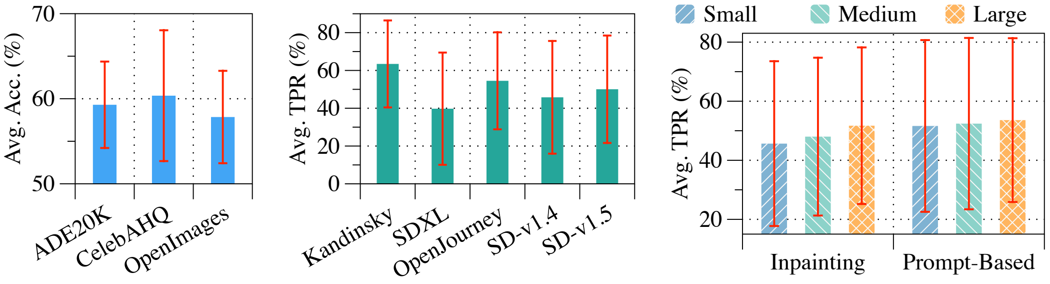
Within a model family, smaller models often predict images as AI-edited regardless of the ground truth. This trend is evident in the Ovis2 and InternVL3 families, where 1B and 2B models yield nearly 100% TPR but also nearly 100% FPR. We attribute this behavior to their limited visual perceptual capacity, which leads them to treat almost any irregularity as evidence of AI-editing.

4.3 Comparison with Production VLMs

In addition to evaluating open-source VLMs, we also include a production-level model, GPT-4o, for comparison. Due to budget constraints, we randomly sample 100 real and 100 AI-edited images for a lightweight test. For baselines, we select the best-performing open-source models under each prompting strategy: LLaVA-1.6-7B with direct prompting and InternVL3-38B with CoT prompting. All models are evaluated using greedy inference without sampling. As shown in Table 2, GPT-4o substantially outperforms the strongest open-source models, achieving 11.5% higher accuracy than

Table 2: **Comparison of the best open-source model with GPT-4o.** Results are reported under greedy inference. The best result in each column is marked in **bold underline**.

VLMs	*Direct Prompting*				*CoT Prompting*			
	Acc. ↑	TPR ↑	FPR ↓	Time (s)	Acc. ↑	TPR ↑	FPR ↓	Time (s)
LLaVA-1.6-7B	67.5%	74.0%	39.0%	1.25	53.5%	8.0%	**1.0%**	14.98
InternVL3-38B	59.5%	31.0%	**12.0%**	4.72	61.5%	35.0%	12.0%	26.89
GPT-4o	**76.0%**	**88.0%**	36.0%	**1.24**	**67.5%**	**53.0%**	18.0%	**5.05**



(a) Across Datasets (b) Across Generative Models (c) Across Editing Methods

Figure 3: **Comparative analysis across different settings.** We evaluate performance across datasets, generative models, and editing methods.

LLaVA-1.6-7B and 6% higher than InternVL3-38B. This result is expected, given GPT-4o's much larger scale and extensive training data. Interestingly, even for GPT-4o, direct prompting outperforms CoT prompting by roughly 10% in accuracy, reinforcing our "first impression" hypothesis that immediate judgments are more reliable than extended reasoning for AI-edited image detection.

4.4 Comparative Analysis Across Different Settings

In this section, we compare detection performance across different experimental settings. All studies use direct prompting under greedy inference, and we report averaged results across all VLMs.

Across Different Datasets. We first evaluate detection performance on different source datasets. Results are shown in Figure 3a, where each bar denotes the averaged accuracy with error bars representing the standard deviation. Detection performance is slightly higher on CelebAHQ compared to ADE20K and OpenImages, likely because CelebAHQ focuses on human faces, a narrower domain that is easier to detect than the broader and more diverse content of the other two datasets.

Across Different Generative Models. We then compare detection performance on AI-edited images generated by different models. Results are presented in Figure 3b, where the bars indicate averaged TPR, since only edited images are considered. Kandinsky is the most vulnerable to detection, potentially due to its relatively poor generation quality. In contrast, SDXL, known for high-quality editing, produces images that are the most difficult to detect.

Across Different Editing Methods. Finally, we examine detection performance across editing methods and editing strengths, as shown in Figure 3c. Bars again indicate averaged TPR. For each editing method, we compare three editing strengths: mask size for inpainting and semantic similarity for prompt-based editing [36]. For example, a larger mask size corresponds to a stronger editing strength (see Appendix B). Observe that prompt-based editing is generally harder to detect than inpainting, as inpainting often leaves artifacts at mask boundaries. Moreover, larger editing strength consistently leads to better detection, as larger modifications are easier for VLMs to capture.

4.5 Insights for Performance Improvement

The detection performance of VLMs remains limited, even for GPT-4o. We therefore explore potential strategies to improve performance. One natural direction is to adopt more fine-grained, object-wise inspection, since some edits only modify small regions of an image. To validate this idea, we conduct

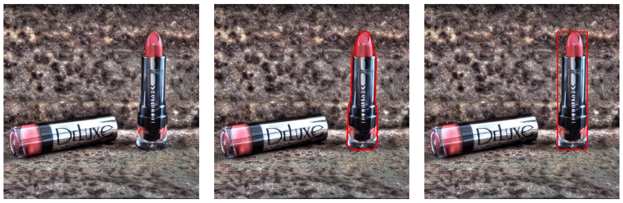

No Hint Given Semantic Mask Given Bounding Box

Figure 4: **Illustration of different hints.** The image is edited with SD-v1.5 (the right lipstick). The hints are highlighted in red.

Table 3: **Performance comparison with different types of hints.** Experiments are conducted under direct prompting using GPT-4o.

GPT-4o	***Direct Prompting***	
	TPR	Time (s)
No hint	85%	0.89
Given mask	86%	6.57
Given bbox	89%	6.26

a lightweight experiment by providing GPT-4o with hints for inpainted images. Specifically, we randomly select 50 inpainted images and compare performance under different types of hints. As illustrated in Figure 4, we highlight the ground-truth inpainting region and ask GPT-4o to focus on that area during detection. We consider two hint types: (1) the semantic outline of the mask, and (2) the bounding box surrounding the mask. The results in Figure 3 show that both hints improve detection accuracy, with the bounding box achieving slightly better gains. We suspect that the semantic outline may sometimes introduce artifacts along the highlighted boundary, which can confuse the model.

In practice, ground-truth hints are unavailable. An alternative is to segment an image into objects and inspect each individually. We conduct a preliminary test using SAM [23] to generate object masks and aggregate GPT-4o's responses for each object. While fine-grained scanning improves sensitivity (TPR), it also increases false positives. For instance, with five objects, a single misprediction can flip the overall decision. Hence, balancing granularity and reliability remains our future work.

5 Conclusion

In this paper, we study zero-shot detection of AI-edited images using vision language models (VLMs). We evaluate 24 VLMs under two prompting strategies and two inference modes, finding that detection accuracy broadly follows overall model capability. Notably, direct prompting with greedy decoding achieves the strongest performance, suggesting a "first impression" effect. However, overall accuracy remains modest (55%–60%), highlighting the need for further improvement.

Acknowledgements

This work was supported in part by Sony AI, NSF Awards SHF-1901242, SHF-1910300, Proto-OKN 2333736, IIS-2416835, ONR N00014-23-1-2081, and Amazon. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the sponsors.

References

[1] Namhyuk Ahn, Junsoo Lee, Chunggi Lee, Kunhee Kim, Daesik Kim, Seung-Hun Nam, and Kibeom Hong. Dreamstyler: Paint by style inversion with text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 674–681, 2024.

[2] Stability AI. Stable diffusion xl base 1.0, 2023.

[3] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[4] Marcelo Bertalmio, Guillermo Sapiro, Vincent Caselles, and Coloma Ballester. Image inpainting. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 417–424, 2000.

[5] You-Ming Chang, Chen Yeh, Wei-Chen Chiu, and Ning Yu. Antifakeprompt: Prompt-tuned vision-language models are fake image detectors. *arXiv preprint arXiv:2310.17419*, 2023.

[6] Baoying Chen, Jishen Zeng, Jianquan Yang, and Rui Yang. Drct: Diffusion reconstruction contrastive training towards universal detection of diffusion generated images. In *Forty-first International Conference on Machine Learning*, 2024.

[7] Shen Chen, Taiping Yao, Hong Liu, Xiaoshuai Sun, Shouhong Ding, Rongrong Ji, et al. Diffusionfake: Enhancing generalization in deepfake detection via guided stable diffusion. *Advances in Neural Information Processing Systems*, 37:101474–101497, 2024.

[8] Siyuan Cheng, Lingjuan Lyu, Zhenting Wang, Xiangyu Zhang, and Vikash Sehwag. Co-spy: Combining semantic and pixel features to detect synthetic images by ai. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13455–13465, 2025.

[9] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267, 2023.

[10] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407, 2024.

[11] Yongchao Feng, Yajie Liu, Shuai Yang, Wenrui Cai, Jinqing Zhang, Qiqi Zhan, Ziyue Huang, Hongxi Yan, Qiao Wan, Chenguang Liu, et al. Vision-language model for object detection and segmentation: A review and evaluation. *arXiv preprint arXiv:2504.09480*, 2025.

[12] Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. Leveraging frequency analysis for deep fake image recognition. In *International conference on machine learning*, pages 3247–3258. PMLR, 2020.

[13] Hongchang Gao, Jian Pei, and Heng Huang. Progan: Network embedding via proximity generative adversarial network. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1308–1316, 2019.

[14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[15] Sagar Goyal, Eti Rastogi, Sree Prasanna Rajagopal, Dong Yuan, Fen Zhao, Jai Chintagunta, Gautam Naik, and Jeff Ward. Healai: A healthcare llm for effective medical documentation. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pages 1167–1168, 2024.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[17] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022.

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

[19] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

[20] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

[21] Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[22] Diederik P Kingma, Max Welling, et al. An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 12(4):307–392, 2019.

[23] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023.

[24] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International journal of computer vision*, 128(7):1956–1981, 2020.

[25] 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 *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

[26] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022.

[27] Bo Liu, Fan Yang, Xiuli Bi, Bin Xiao, Weisheng Li, and Xinbo Gao. Detecting generated images by real images. In *European Conference on Computer Vision*, pages 95–110. Springer, 2022.

[28] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26296–26306, 2024.

[29] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

[30] Huan Liu, Zichang Tan, Chuangchuang Tan, Yunchao Wei, Jingdong Wang, and Yao Zhao. Forgery-aware adaptive transformer for generalizable synthetic image detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10770–10780, 2024.

[31] Shiyin Lu, Yang Li, Yu Xia, Yuwei Hu, Shanshan Zhao, Yanqing Ma, Zhichao Wei, Yinglun Li, Lunhao Duan, Jianshan Zhao, et al. Ovis2. 5 technical report. *arXiv preprint arXiv:2508.11737*, 2025.

[32] Midjourney. Midjourney., 2023.

[33] Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24480–24489, 2023.

[34] OpenAI. Dall-e 3 understands significantly more nuance and detail than our previous systems, allowing you to easily translate your ideas into exceptionally accurate images., 2023.

[35] OpenCompass. Open vlm leaderboard. Hugging Face Space.

[36] Anisha Pal, Julia Kruk, Mansi Phute, Manognya Bhattaram, Diyi Yang, Duen Horng Chau, and Judy Hoffman. Semi-truths: A large-scale dataset of ai-augmented images for evaluating robustness of ai-generated image detectors. *Advances in Neural Information Processing Systems*, 37:118025–118051, 2024.

[37] Viacheslav Pirogov. Visual language models as zero-shot deepfake detectors. *arXiv preprint arXiv:2507.22469*, 2025.

[38] PromptHero. openjourney. Model hosted on Hugging Face.

[39] Md Shohel Rana, Mohammad Nur Nobi, Beddhu Murali, and Andrew H Sung. Deepfake detection: A systematic literature review. *IEEE access*, 10:25494–25513, 2022.

[40] Anton Razzhigaev, Arseniy Shakhmatov, Anastasia Maltseva, Vladimir Arkhipkin, Igor Pavlov, Ilya Ryabov, Angelina Kuts, Alexander Panchenko, Andrey Kuznetsov, and Denis Dimitrov. Kandinsky: an improved text-to-image synthesis with image prior and latent diffusion. *arXiv preprint arXiv:2310.03502*, 2023.

[41] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[42] Jia Wen Seow, Mei Kuan Lim, Raphaël CW Phan, and Joseph K Liu. A comprehensive overview of deepfake: Generation, detection, datasets, and opportunities. *Neurocomputing*, 513:351–371, 2022.

[43] Hao Shao, Yuxuan Hu, Letian Wang, Guanglu Song, Steven L Waslander, Yu Liu, and Hongsheng Li. Lmdrive: Closed-loop end-to-end driving with large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15120–15130, 2024.

[44] Zhi Rui Tam, Cheng-Kuang Wu, Yi-Lin Tsai, Chieh-Yen Lin, Hung-yi Lee, and Yun-Nung Chen. Let me speak freely? a study on the impact of format restrictions on performance of large language models. *arXiv preprint arXiv:2408.02442*, 2024.

[45] 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 AAAI Conference on Artificial Intelligence*, volume 38, pages 5052–5060, 2024.

[46] Chuangchuang Tan, 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28130–28139, 2024.

[47] Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, and Yunchao Wei. Learning on gradients: Generalized artifacts representation for gan-generated images detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12105–12114, 2023.

[48] Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, et al. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*, 2025.

[49] Dayin Wang, Chong Ma, and Siwen Sun. Novel paintings from the latent diffusion model through transfer learning. *Applied Sciences*, 13(18):10379, 2023.

[50] Lening Wang, Yilong Ren, Han Jiang, Pinlong Cai, Daocheng Fu, Tianqi Wang, Zhiyong Cui, Haiyang Yu, Xuesong Wang, Hanchu Zhou, et al. Accidentgpt: Accident analysis and prevention from v2x environmental perception with multi-modal large model. *arXiv preprint arXiv:2312.13156*, 2023.

[51] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. Cnn-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8695–8704, 2020.

[52] Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. Dire for diffusion-generated image detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22445–22455, 2023.

[53] Ziyu Wang, Hao Li, Di Huang, Hye-Sung Kim, Chae-Won Shin, and Amir M Rahmani. Healthq: Unveiling questioning capabilities of llm chains in healthcare conversations. *Smart Health*, page 100570, 2025.

[54] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[55] Mika Westerlund. The emergence of deepfake technology: A review. *Technology innovation management review*, 9(11), 2019.

[56] Xianchao Wu. Creative painting with latent diffusion models. *arXiv preprint arXiv:2209.14697*, 2022.

[57] Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, et al. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*, 2024.

[58] Guowei Xu, Peng Jin, Ziang Wu, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. Llava-cot: Let vision language models reason step-by-step. *arXiv preprint arXiv:2411.10440*, 2024.

[59] Shilin Yan, Ouxiang Li, Jiayin Cai, Yanbin Hao, Xiaolong Jiang, Yao Hu, and Weidi Xie. A sanity check for ai-generated image detection. *arXiv preprint arXiv:2406.19435*, 2024.

[60] Shilin Yan, Ouxiang Li, Jiayin Cai, Yanbin Hao, Xiaolong Jiang, Yao Hu, and Weidi Xie. A sanity check for ai-generated image detection. *arXiv preprint arXiv:2406.19435*, 2024.

[61] Zhiyuan Yan, Taiping Yao, Shen Chen, Yandan Zhao, Xinghe Fu, Junwei Zhu, Donghao Luo, Chengjie Wang, Shouhong Ding, Yunsheng Wu, et al. Df40: Toward next-generation deepfake detection. *Advances in Neural Information Processing Systems*, 37:29387–29434, 2024.

[62] Kairui Yang, Zihao Guo, Gengjie Lin, Haotian Dong, Zhao Huang, Yipeng Wu, Die Zuo, Jibin Peng, Ziyuan Zhong, Xin Wang, et al. Trajectory-llm: A language-based data generator for trajectory prediction in autonomous driving. In *The Thirteenth International Conference on Learning Representations*, 2025.

[63] Xu Yang, Yongliang Wu, Mingzhuo Yang, Haokun Chen, and Xin Geng. Exploring diverse in-context configurations for image captioning. *Advances in Neural Information Processing Systems*, 36:40924–40943, 2023.

[64] Ziqi Yang, Xuhai Xu, Bingsheng Yao, Ethan Rogers, Shao Zhang, Stephen Intille, Nawar Shara, Guodong Gordon Gao, and Dakuo Wang. Talk2care: An llm-based voice assistant for communication between healthcare providers and older adults. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(2):1–35, 2024.

[65] Yue Zhang, Ben Colman, Xiao Guo, Ali Shahriyari, and Gaurav Bharaj. Common sense reasoning for deepfake detection. In *European conference on computer vision*, pages 399–415. Springer, 2024.

[66] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127(3):302–321, 2019.

[67] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

A Detailed Description of Evaluated VLMs

For reference, Table 4 provides the evaluated open-source VLMs used in our experiments, along with their source links and release dates.

Table 4: **Sources of Evaluated VLMs.** Mapping of model names used in our experiments to their source links and release dates.

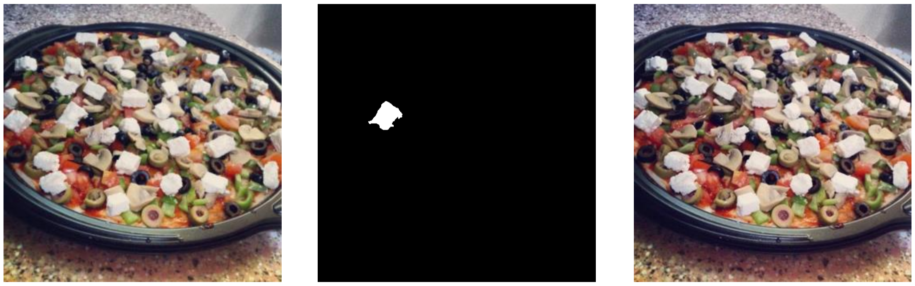
VLM Name	Model ID (HuggingFace Link)	Release Date
LLaVA-1.5-7B	llava-hf/llava-1.5-7b-hf	09.2023
LLaVA-1.5-13B	llava-hf/llava-1.5-13b-hf	09.2023
LLaVA-1.6-7B	llava-hf/llava-v1.6-mistral-7b-hf	07.2024
LLaVA-1.6-13B	llava-hf/llava-v1.6-vicuna-13b-hf	07.2024
LLaVA-1.6-34B	llava-hf/llava-v1.6-34b-hf	07.2024
Llama-3.2-11B	meta-llama/Llama-3.2-11B-Vision	12.2023
DeepSeek-VL2-tiny	deepseek-ai/deepseek-vl2-tiny	12.2024
DeepSeek-VL2-small	deepseek-ai/deepseek-vl2-small	12.2024
DeepSeek-VL2	deepseek-ai/deepseek-vl2	12.2024
InternVL3-1B	OpenGVLab/InternVL3-1B	04.2025
InternVL3-2B	OpenGVLab/InternVL3-2B	04.2025
InternVL3-8B	OpenGVLab/InternVL3-8B	04.2025
InternVL3-14B	OpenGVLab/InternVL3-14B	04.2025
InternVL3-38B	OpenGVLab/InternVL3-38B	04.2025
Qwen2.5-VL-3B	Qwen/Qwen2.5-VL-3B-Instruct	04.2025
Qwen2.5-VL-7B	Qwen/Qwen2.5-VL-7B-Instruct	04.2025
Qwen2.5-VL-32B	Qwen/Qwen2.5-VL-32B-Instruct	04.2025
Kimi-VL-A3B	moonshotai/Kimi-VL-A3B-Instruct	07.2025
Ovis2-1B	AIDC-AI/Ovis2-1B	03.2025
Ovis2-2B	AIDC-AI/Ovis2-2B	03.2025
Ovis2-4B	AIDC-AI/Ovis2-4B	03.2025
Ovis2-8B	AIDC-AI/Ovis2-8B	03.2025
Ovis2-16B	AIDC-AI/Ovis2-16B	03.2025
Ovis2-34B	AIDC-AI/Ovis2-34B	03.2025

B Examples of AI-Edited Images

Examples of AI-edited images are shown in Figure 5 and Figure 6, illustrating both inpainting and prompt-based editing methods across three editing strengths (small, medium, large).

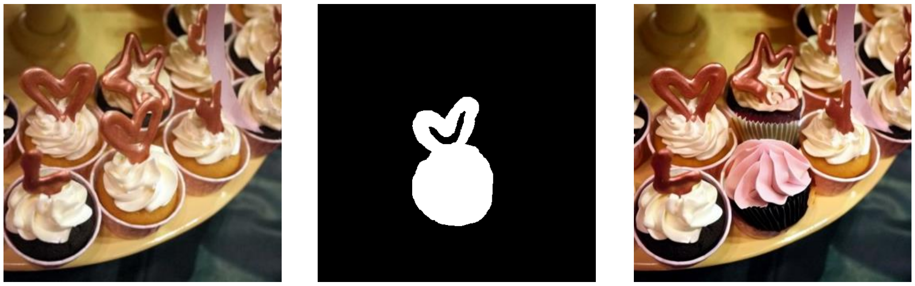
C Examples of VLM Responses

We present example responses from InternVL3-38B [67] under both direct prompting and CoT prompting (subsection 3.1). For each strategy, we illustrate four representative cases: a true positive (Figure 7, Figure 11), a true negative (Figure 8, Figure 12), a false positive (Figure 9, Figure 13), and a false negative (Figure 10, Figure 14). These examples provide a concrete view of how VLMs behave across different prompting methods.



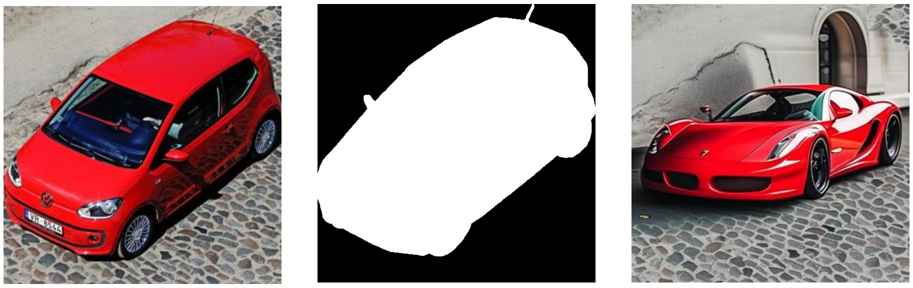
Original Mask Inpainted

(a) Illustration of image inpainting under **small** editing strength.



Original Mask Inpainted

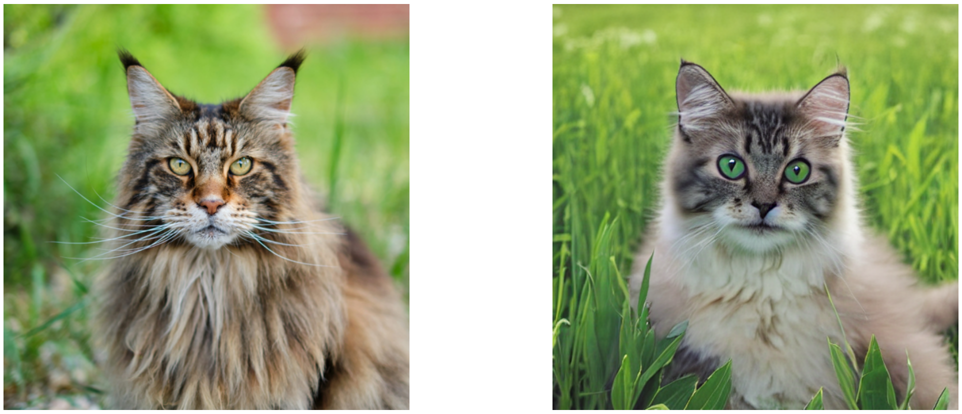
(b) Illustration of image inpainting under **medium** editing strength.



Original Mask Inpainted

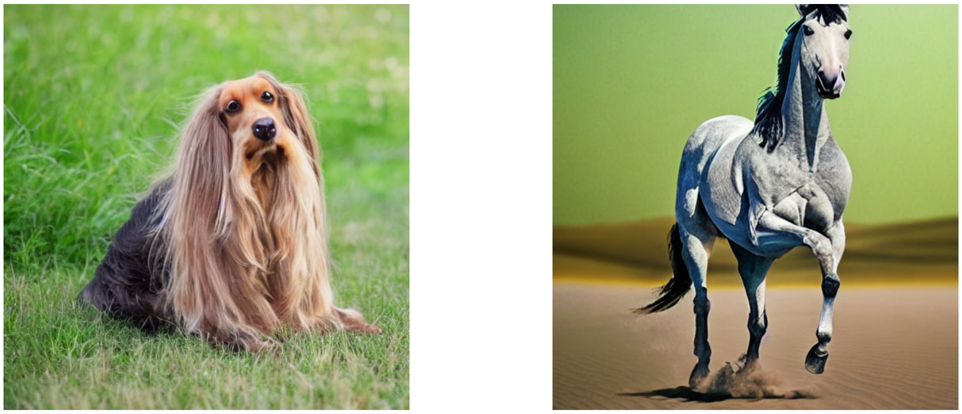
(c) Illustration of image inpainting under **large** editing strength.

Figure 5: **Example of image inpainting as an editing method.** In each row, the first column displays the original image, the second column the semantic mask used for inpainting, and the last column the resulting inpainted image.



(a) **Original.** A long-haired cat sitting in the grass.

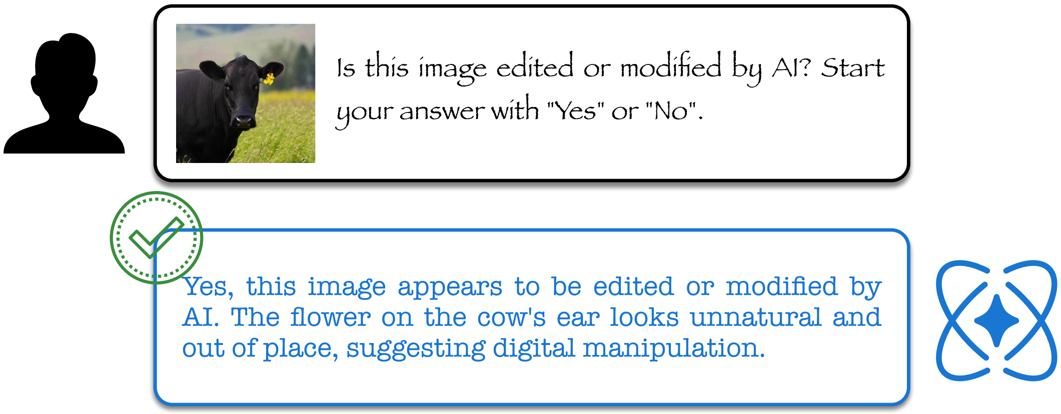
(b) **Small editing strength.** A fluffy cat sitting in the green meadow.



(c) **Medium editing strength.** A long-haired dog sitting in the grass.

(d) **Large editing strength.** A horse galloping across a desert plain.

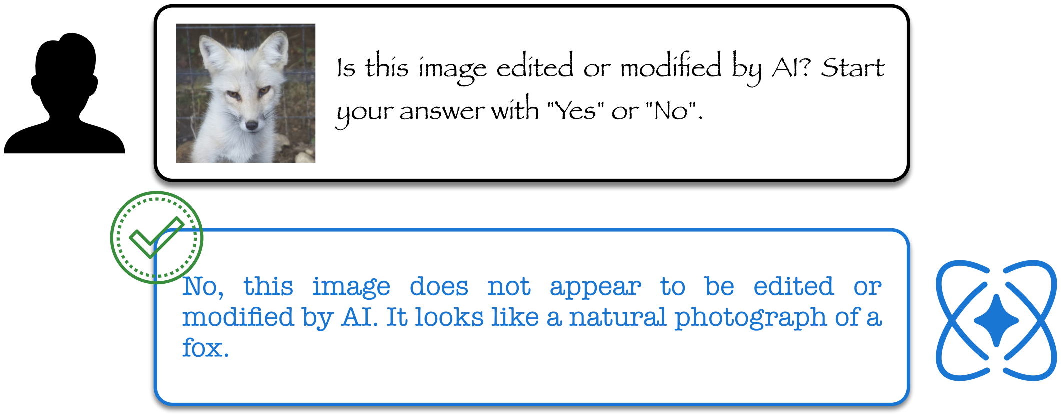
Figure 6: **Example of prompt-based editing.** Subfigure (a) shows the original image with its caption, while (b), (c), and (d) present edited versions generated using the prompts indicated in their respective subcaptions.



Is this image edited or modified by AI? Start your answer with "Yes" or "No".

Yes, this image appears to be edited or modified by AI. The flower on the cow's ear looks unnatural and out of place, suggesting digital manipulation.

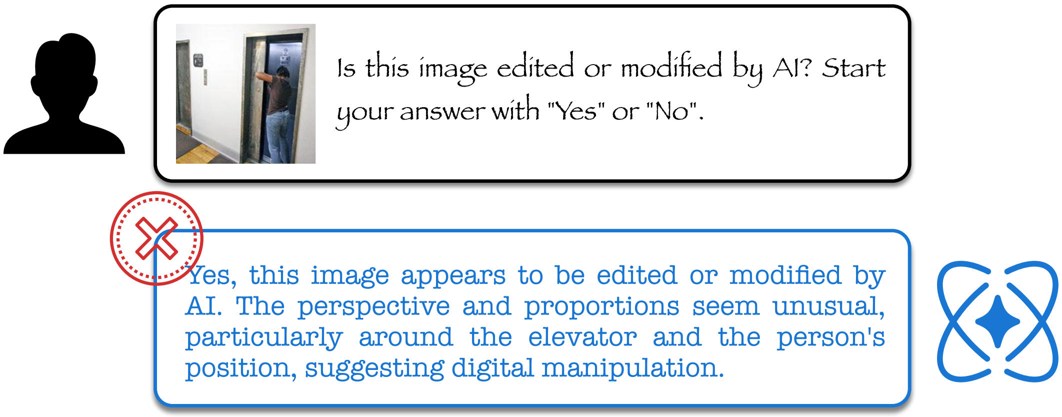
Figure 7: **Direct Prompting.** Illustration of a true positive case.



Is this image edited or modified by AI? Start your answer with "Yes" or "No".

No, this image does not appear to be edited or modified by AI. It looks like a natural photograph of a fox.

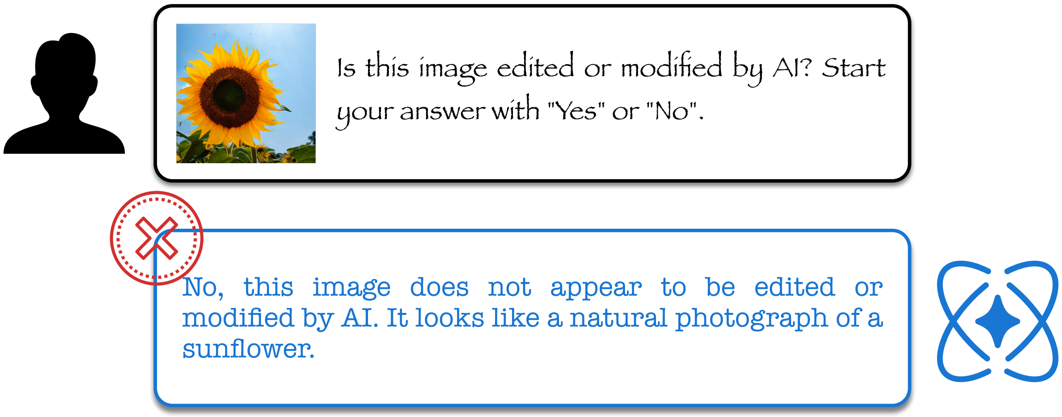
Figure 8: **Direct Prompting.** Illustration of a true negative case.



Is this image edited or modified by AI? Start your answer with "Yes" or "No".

Yes, this image appears to be edited or modified by AI. The perspective and proportions seem unusual, particularly around the elevator and the person's position, suggesting digital manipulation.

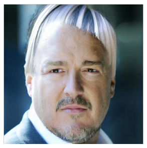
Figure 9: **Direct Prompting.** Illustration of a false positive case.



Is this image edited or modified by AI? Start your answer with "Yes" or "No".

No, this image does not appear to be edited or modified by AI. It looks like a natural photograph of a sunflower.

Figure 10: **Direct Prompting.** Illustration of a false negative case.



Think step-by-step to decide whether this image is edited or modified by AI. Then conclude your answer with "The answer is" followed by your final answer.

To determine if this image is edited or modified by AI, let's analyze it step by step:

1. **Hair Color and Style:** The hair color is an unusual and uniform platinum blonde, which is not common in natural hair. This suggests the possibility of digital alteration.
2. **Lighting and Shadows:** The lighting appears consistent, but the shadows and highlights on the hair and face seem slightly unnatural, which could indicate digital manipulation.
3. **Skin Texture:** The skin texture looks smooth and uniform, which might be a result of AI smoothing or retouching.
4. **Overall Appearance:** The overall appearance of the image, particularly the hair, seems to be more stylized than natural, which is a common trait in AI-generated or heavily edited images.

Based on these observations, it is likely that the image has been edited or modified by AI.

The answer is: Yes, the image is edited or modified by AI.

Figure 11: **CoT Prompting.** Illustration of a true positive case.

Think step-by-step to decide whether this image is edited or modified by AI. Then conclude your answer with "The answer is" followed by your final answer.

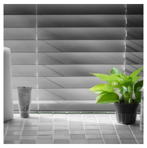
To determine if this image is edited or modified by AI, let's analyze it step by step:

1. **Visual Consistency:** The image shows a plate of flatbread with some dips. The lighting and shadows appear consistent across the image, suggesting a natural setting.
2. **Texture and Detail:** The textures of the flatbread, dips, and the tablecloth look realistic. There are no obvious signs of blurring or unnatural transitions.
3. **Composition:** The composition is typical for a food photograph, with no unusual elements or distortions that might suggest editing.
4. **Color and Lighting:** The colors are natural and the lighting is consistent, without any signs of artificial manipulation.
5. **Background and Context:** The background elements, such as the tablecloth and bowls, are in harmony with the main subject, indicating a coherent scene.

Based on these observations, there are no clear indicators that the image has been edited or modified by AI.

The answer is: The image does not appear to be edited or modified by AI.

Figure 12: **CoT Prompting.** Illustration of a true negative case.



Think step-by-step to decide whether this image is edited or modified by AI. Then conclude your answer with "The answer is" followed by your final answer.

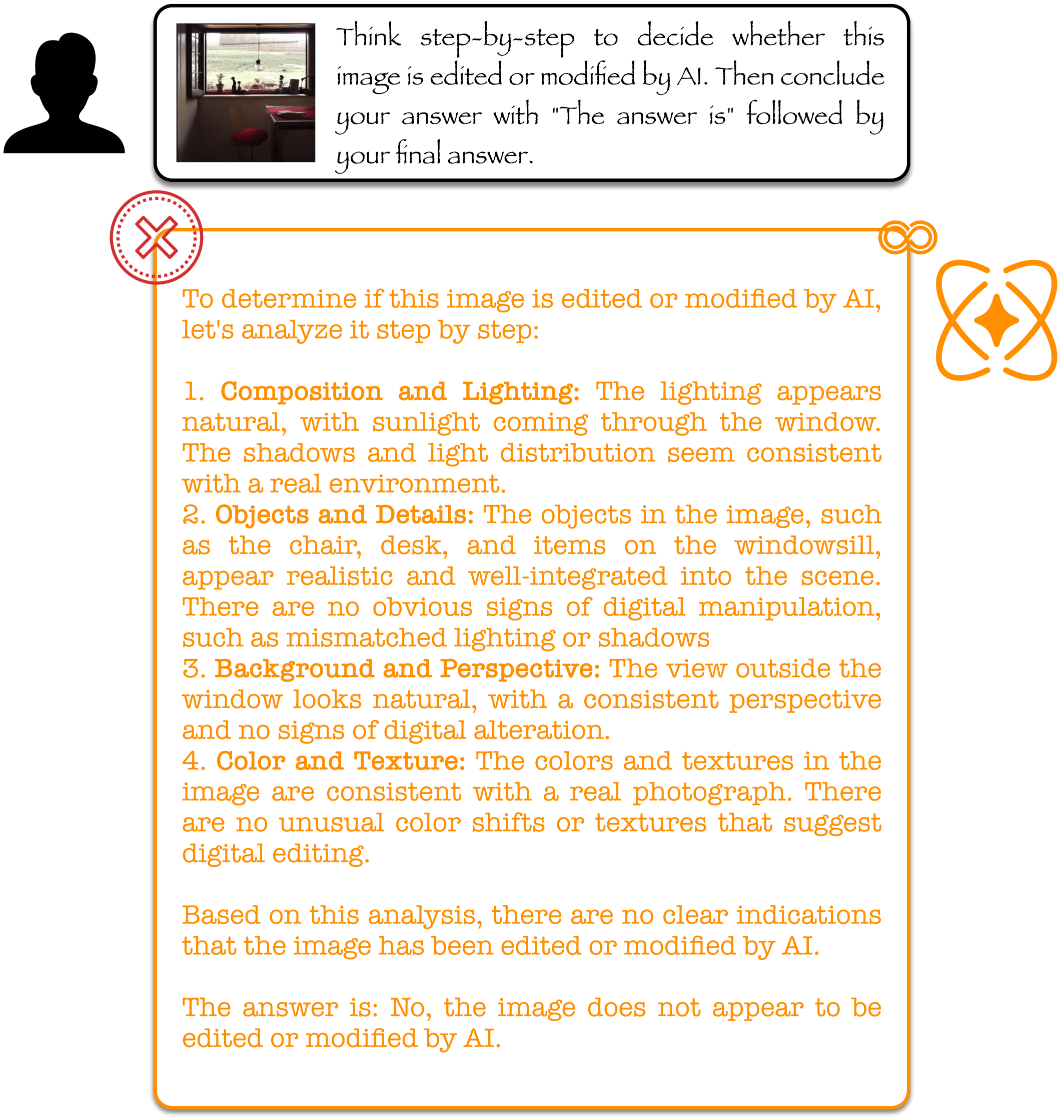
To determine if this image is edited or modified by AI, let's analyze it step by step:

1. **Lighting and Shadows:** The lighting appears consistent across the image. The shadows cast by the blinds are uniform and natural, suggesting that the lighting is not artificially altered.
2. **Objects and Composition:** The objects in the image (the tube, the plant, and the blinds) are all in proportion and perspective. There are no obvious signs of objects being added or removed.
3. **Color and Contrast:** The image is in grayscale except for the green plant. This selective coloring could be a result of editing, as it is unusual for a natural scene to have such a stark color contrast.
4. **Edges and Blending:** The edges of the plant and the tube are sharp and well-defined, which is typical in natural images. There are no visible signs of blending or artifacts that would suggest digital manipulation.
5. **Background and Texture:** The background and textures appear natural. The blinds and tiles have a realistic appearance without any noticeable distortions.

Given these observations, the selective coloring of the plant is the most likely indication of editing. The rest of the image appears natural and unaltered.

The answer is: The image is edited.

Figure 13: **CoT Prompting.** Illustration of a false positive case.



Think step-by-step to decide whether this image is edited or modified by AI. Then conclude your answer with "The answer is" followed by your final answer.

To determine if this image is edited or modified by AI, let's analyze it step by step:

1. **Composition and Lighting:** The lighting appears natural, with sunlight coming through the window. The shadows and light distribution seem consistent with a real environment.
2. **Objects and Details:** The objects in the image, such as the chair, desk, and items on the windowsill, appear realistic and well-integrated into the scene. There are no obvious signs of digital manipulation, such as mismatched lighting or shadows
3. **Background and Perspective:** The view outside the window looks natural, with a consistent perspective and no signs of digital alteration.
4. **Color and Texture:** The colors and textures in the image are consistent with a real photograph. There are no unusual color shifts or textures that suggest digital editing.

Based on this analysis, there are no clear indications that the image has been edited or modified by AI.

The answer is: No, the image does not appear to be edited or modified by AI.

Figure 14: **CoT Prompting.** Illustration of a false negative case.