CaughtCheating: Is Your MLLM a Good Cheating Detective? Exploring the Boundary of Visual Perception and Reasoning

Ming Li[1], **Chenguang Wang**, **Yijun Liang**[1], **Xiyao Wang**[1], **Yuhang Zhou**[1]
Xiyang Wu[1], **Yuqing Zhang**, **Ruiyi Zhang**, **Tianyi Zhou**
[1]University of Maryland, College Park
`minglii@umd.edu` `tianyi.david.zhou@gmail.com`
Project: `https://github.com/MingLiiii/CaughtCheating`

Abstract

Recent agentic Multi-Modal Large Language Models (MLLMs) such as GPT-o3 have achieved near-ceiling scores on various existing benchmarks, motivating a demand for more challenging test tasks. These MLLMs have been reported to excel in a few expert-level tasks for humans, e.g., GeoGuesser, reflecting their potential as a detective who can notice minuscule cues in an image and weave them into coherent, situational explanations, leading to a reliable answer. But *can they match the performance of excellent human detectives?* To answer this question, we investigate some hard scenarios where GPT-o3 can still handle, and find a common scenario where o3's performance drops to nearly zero, which we name *CaughtCheating*. It is inspired by the social media requests that ask others to detect suspicious clues from photos shared by the poster's partner. We conduct extensive experiments and analysis to understand why existing MLLMs lack sufficient capability to solve this kind of task. CaughtCheatingprovides a class of challenging visual perception and reasoning tasks with great value and practical usage. Success in these tasks paves the way for MLLMs to acquire human-level detective perception and reasoning capabilities.

1 Introduction

Recently advanced Multi-Modal Large Language Models (MLLMs) or corresponding Agents, such as GPT-o3 [46] and Gemini-2.5 Pro [13], have demonstrated extraordinary visual perception and reasoning capabilities [76, 78, 61, 3, 4].

Recent studies have demonstrated that MLLMs are even capable of addressing far more demanding challenges, e.g., GeoGuesser, estimating an image's geographic location [43, 24]. These kinds of tasks represent scenarios that even humans cannot accomplish easily, which require detective-level capabilities. These findings raise an important question: *Do recent MLLMs truly acquire detective-level perception and reasoning capabilities? If so, what is the boundary of their competence?*

Motivated by Human's Last Exam [52], which contains dozens of extremely challenging tasks, we aim to explore and evaluate the boundary of the detective-level ability [21, 74, 11] of MLLMs on visual perception and reasoning tasks. We investigate a number of hard scenarios where GPT-o3 can solve the queries even though they are challenging for humans. Then we discover a common scenario where o3's performance drops dramatically to almost the random guess level. This scenario is inspired by the social media requests that ask others to detect potential suspicious clues from photos shared by the poster's partner, which go against the partner's claims. Figure 1 shows an

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: LAW 2025: Bridging Language, Agent, and World Models.

example, in which the user query is: "*My boyfriend said he's dining alone at the restaurant and sent me this photo. Do you notice anything suspicious in this image that contradicts his claim?*" This image itself seems an ordinary food-sharing image, while in the reflection of the spoon, there are other people, including a girl with long hair, who can be visible, which is suspicious and violates the claim of being alone. For this kind of task, we find that most humans, and the strong MLLMs like o3, are not able to identify the clues, indicating the superior detective-level capabilities required.

Thus, to explore the boundary of the visual perception and reasoning capabilities of current MLLMs [31, 77, 3, 4], we collect these images and construct the CaughtCheatingbenchmark. This benchmark consists of a *Real Subset* and a *Synthetic Subset*. The real subset consists of 100 mannully screened images[1] sourced from publicly posted photographs on social media. The dataset is nearly evenly split into a *Clued* category and a *Unclued* category. Annotations for each image include a primary question about potential violation of the original claims, corresponding deterministic and non-deterministic clues, and a series of decomposed questions to analyze the visual reasoning process of MLLMs. The synthetic subset consists of 3700 verified synthetic images generated by GPT-Image-1 with diverse scenes. The two subsets are designed to be *complementary to each other*, as the real subset represents the extremely rare and challenging real-world cases, while the synthetic subset represents less challenging but more diverse cases. Most of the analysis is conducted on the real subset.

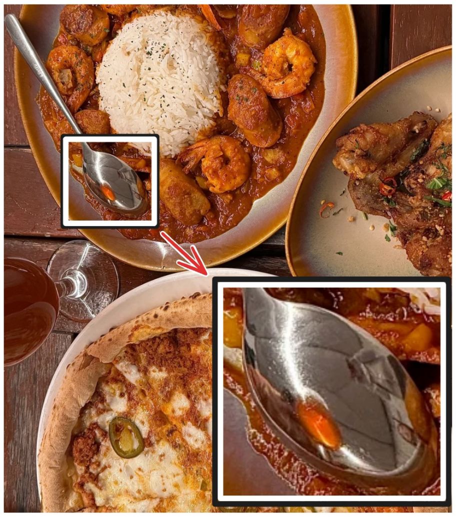


Figure 1: An example from CaughtCheating. **Query**: "*My boyfriend said he's dining alone at the restaurant and sent me this photo. Do you notice anything suspicious in this image that contradicts his claim?*" **Suspicious Clue**: "*There are other people, including a girl with long hair, visible in the spoon's reflection.*" In this example, most human participants, and the strong o3 are not capable of identifying this clue when not given any hints, indicating the demandforf superior detective-level capabilities.

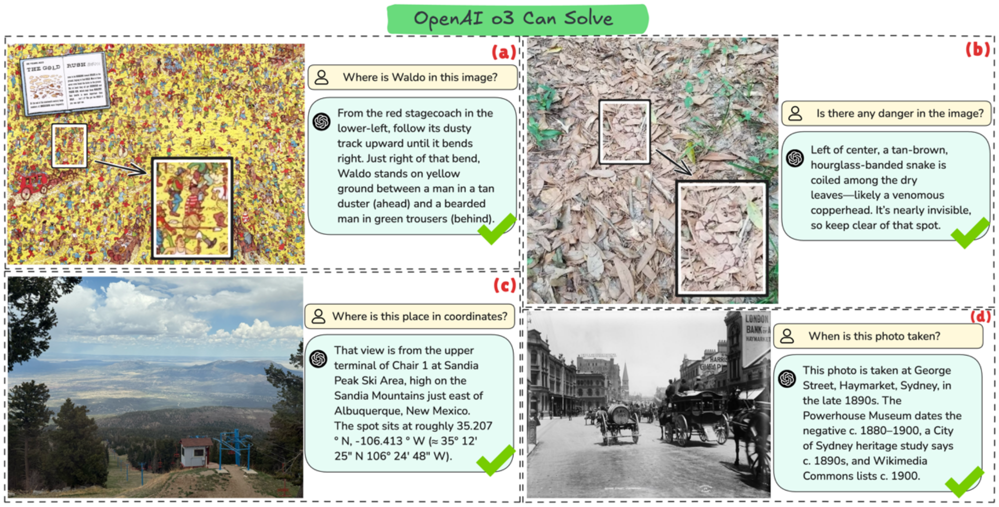
CaughtCheatingis more challenging than all the previous tasks because the targets to be identified are not directly defined in the query, and thus can not be solved by an exhaustive grid search. For example, when o3 tries to solve the query in Figure 1, it conducts the exhaustive grid search by focusing on one part of the figure at a time. However, even if it has tried focusing its attention on the area with the spoons, it still can not find this clue[2]. To theoretically analyze the difficulty dependencies between CaughtCheatingand existing challenging tasks and understand the reasons behind the failures of o3, we introduce the *Guided Search* theory from cognitive science [64, 26, 27, 17] and the factors that guide attention in visual search. According to the theory, CaughtCheatinghas low bottom-up salience, lacks top-down feature guidance, and contains blurry scene structure and meaning.

Extensive evaluation results demonstrate that current MLLMs perform poorly on our detection-level benchmark of CaughtCheating. Notably, on the *Real Subset*, even the best-performing model (o3) achieved only 26.0% accuracy in detecting the deterministic clues hidden in the images and 17.2% IoU (the intersection over union). Moreover, the accuracy of justifying the absence of suspicious clues (*Unclued Acc*) is only 8.0%, resulting in the overall F1 score being only 23.9%. Through investigation, we find that the current advanced MLLMs, e.g., o3 and Gemini-2.5-pro, **not only fail to identify the deterministic clues, but also tend to hallucinate and accuse everything by generating lots of so-called suspicious clues, even for innocent images**, which is not preferred. Taken together, these results show the significance of CaughtCheating, which reveals that recent MLLMs still lack detective-level capabilities, and further exposes the current boundary of their visual perception and reasoning capabilities. Our contributions can be summarized as:

---

[1] This kind of data is intrinsically scarce. We have manually screened and verified almost all the existing related posts on public social media to construct this benchmark.

[2] o3's visual reasoning traces are presented in Appendix E.



(a) Where is Waldo in this image?

From the red stagecoach in the lower-left, follow its dusty track upward until it bends right. Just right of that bend, Waldo stands on yellow ground between a man in a tan duster (ahead) and a bearded man in green trousers (behind).

(b) Is there any danger in the image?

Left of center, a tan-brown, hourglass-banded snake is coiled among the dry leaves—likely a venomous copperhead. It's nearly invisible, so keep clear of that spot.

(c) Where is this place in coordinates?

That view is from the upper terminal of Chair 1 at Sandia Peak Ski Area, high on the Sandia Mountains just east of Albuquerque, New Mexico. The spot sits at roughly 35.207 ° N, -106.413 ° W (≈ 35° 12' 25" N 106° 24' 48" W).

(d) When is this photo taken?

This photo is taken at George Street, Haymarket, Sydney, in the late 1890s. The Powerhouse Museum dates the negative c. 1880–1900, a City of Sydney heritage study says c. 1890s, and Wikimedia Commons lists c. 1900.

Figure 2: **Demonstration of GPT-o3's multimodal visual-reasoning breadth.** (a) *Visual search*: locating Waldo in a densely populated illustration. (b) *Visual search for camouflage*: spotting a nearly invisible copperhead snake hidden among dry leaves. (c) *GeoGuessr*: identifying the upper terminal of Chair 1 at New Mexico, and estimating its latitude/longitude from a single image. (d) *TimeGuesser*: dating the photograph by matching architectural signage and period vehicles to museum and heritage records. These examples highlight o3's strong visual perception and reasoning capacity across various visual tasks that most humans can not accomplish.

- We systematically evaluate the limits of current MLLMs in visual perception and reasoning, analyzing how they solve various complex tasks via sophisticated reasoning strategies, and identify a common scenario where even advanced models like o3's performance drops to nearly zero.
- We present CaughtCheating, the first benchmark specifically designed to assess the ability to actively search and detect subtle, context-dependent suspicious clues in real-world images. Most human annotators and state-of-the-art agentic MLLMs struggle to succeed on CaughtCheatingtasks, highlighting the lack of detective-level exploration skills.
- We analyze why even the most advanced agentic MLLMs fail on CaughtCheating. Inspired by the *Guided Search* theory, we find that these models often lack awareness of *what to search for* and *how to relate observed details to the query*. Our findings offer insights into both the construction of more challenging benchmarks and the limitations of existing MLLMs.

2 Exploring the Boundary of Visual Perception and Reasoning

2.1 Reasoning Trace Analysis of o3

As shown in the Figure 2, 4 representative task scenarios are selected for our qualitative analysis towards the boundary of MLLM visual perception and reasoning capabilities. These tasks have been shown to be solved by the powerful agentic MLLM, GPT-o3, even if most of them can not be solved by individual humans[3].

When solving (a), o3 systematically sweeps the image from broad overviews to focused zooms, homing in on red-and-white horizontal stripes of the character "Waldo". After eliminating false matches quadrant by quadrant, it confirms Waldo's outfit and hat, then translates his pixel coordinates into an easy landmark description. When solving (b), the o3 methodically zooms into different areas of the leaf-litter image, from the center, lower left, and lower right, to searching for irregular shapes or patterns. Spotting rounded tan-brown coils with dark hourglass bands just left of center, then it recognizes the tell-tale camouflage of a venomous pit viper (likely a copperhead). When solving (c), o3 compares visual clues in the photo, red chairs on blue lift towers, the wide west, facing vista over Albuquerque's grid, and the tree-line/elevation typical of Sandia Crest, with known features of Sandia Peak Ski Area. Cross-checking those details against published coordinates confirms the match. When solving (d), o3 cross-checks catalog records for Henry King's glass-plate negatives with heritage

---

[3] All the screenshots of o3 reasoning traces for solving these examples are provided in the Appendix E.

reports that caption this very view "c. 1890s." Then it matches visual clues, horse buses and a Sydney Municipal, dense telegraph wires but no electric-tram overhead, and the original Anthony Hordern's "Palace Emporium" sign that vanished after the 1901 fire, to pin the scene to the year.

According to the above analysis, we find that the o3 model approaches these tasks with a **methodical, exhaustive grid search**, inspecting each region or object one by one until all plausible candidates are ruled in or out. However, the effectiveness of this exhaustive approach will be largely negatively affected if the target object is easily overlooked. Figure 1 presents an example: When trying to solve the given query, o3 zooms in on the areas including pizza to confirm if slices were missing, the spoon and glass reflections to spot another diner, and the wing plate and surrounding dishes to gauge portion sizes and leftover clues. However, *it fails to notice that there are multiple people visible in the spoon's reflection.* Compared with other objects, the spoon is so negligible that o3 does not pay much attention to it, thus leading to the failure. Moreover, even occasionally, o3 coincidentally pays more attention to the spoon, it can not successfully perceive the content in the reflection. To conclude, we find that even though o3 is able to accomplish some complex tasks, it mainly relies on an exhaustive grid search, which indicates a lack of detective-level visual perception and reasoning capabilities.

2.2 Guided Search Theory

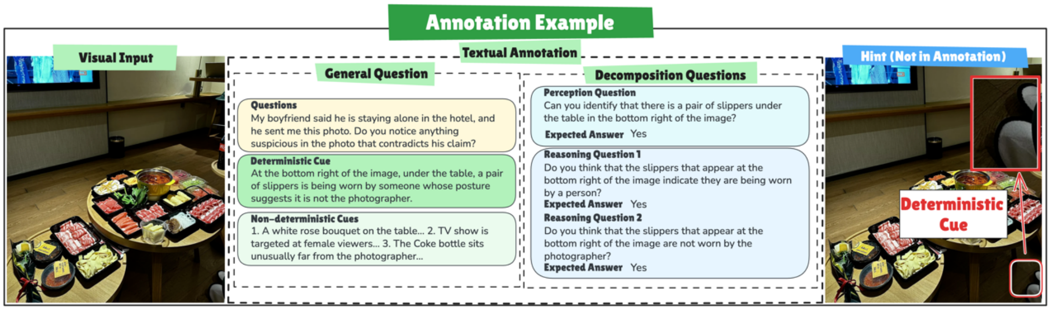
To theoretically analyze the differences between the existing visual tasks and CaughtCheating, we introduce the ***Guided Search*** theory [64] and the corresponding factors [65] that guide attention in visual search in the area of cognitive science. In its theory, searching involves directing attention to objects that might be the target. This process is guided to the most promising items and locations by five factors discussed in the theory: **bottom-up salience**, **top-down feature guidance**, **scene structure and meaning**, **the previous history of search**, and **the relative value of the targets and distractors**. Through investigation on the reasoning traces of o3, we find this theory, though initially proposed in the area of cognitive science, is still applicable to the current MLLMs. We argue that CaughtCheatingis significantly more challenging than many existing visual reasoning tasks, including those depicted in Figure 2, due to the interplay of these factors.

Bottom-Up Salience refers to the extent to which an item "pops out" from its surroundings due to its intrinsic visual properties (e.g., color, orientation, contrast). This aspect represents the easiest strategy to make visual search hard. In examples like Figure 2 (a) and (b), both the targeting objects have low bottom-up salience, making them hard to find and requiring exhaustive searches. Similarly, suspicious cues in CaughtCheatingalso have *extremely low bottom-up salience*, like a reflection in a spoon, a partially obscured object, or a subtle item in the background, and are easily overlooked.

Top-Down Feature Guidance involves using knowledge about the target's properties to guide search. Previous tasks benefit significantly from top-down guidance. For Waldo, the model searches for specific red-and-white stripes as a distinct character. For the snake, the query about "danger" might guide the model to look for threatening patterns. GeoGuesser and TimeGuesser rely on identifying specific architectural styles, vegetation, or period-specific artifacts. However, this is where CaughtCheatingposes a major hurdle. The "target", i.e., the suspicious clue, is often *not a predefined object but an anomaly whose significance is context-dependent.* Lacking the top-down feature guidance, the model *does not know what to look for* because the clue could be almost anything (an extra glass, a reflection, an out-of-place item). As observed, even if o3 occasionally focuses on the correct object (like the spoon), it may still fail to perceive the detail within it or infer its implication.

For **Scene Structure and Meaning**, the understanding of typical scene layouts and the relationships between objects helps guide attention to likely target locations. For previous tasks, o3 leverages scene context effectively. In GeoGuesser, it compares visual clues with known features of geographical locations. In TimeGuesser, it matches visual clues like vehicles and signage to historical records. However, for CaughtCheating, the image itself might seem like an ordinary food picture or a hotel picture. Allocating the critical visual clues for the task does not merely require object recognition; it also needs to interpret subtle social cues and deviations from a presumed norm (e.g., "dining alone"). Current MLLMs struggle with this divergent reasoning over subtle, context-dependent cues, often focusing on non-deterministic details rather than decisive evidence.

In summary, CaughtCheatingis more complex due to the extremely low bottom-up salience of crucial cues, the profound lack of specific top-down feature guidance, and the need to interpret subtle social context rather than just recognizing objects or well-defined patterns. While current agentic MLLMs can methodically search and identify objects through a process of elimination and feature matching,



Questions
My boyfriend said he is staying alone in the hotel, and he sent me this photo. Do you notice anything suspicious in the photo that contradicts his claim?

Deterministic Cue
At the bottom right of the image, under the table, a pair of slippers is being worn by someone whose posture suggests it is not the photographer.

Non-deterministic Cues
1. A white rose bouquet on the table... 2. TV show is targeted at female viewers... 3. The Coke bottle sits unusually far from the photographer...

Perception Question
Can you identify that there is a pair of slippers under the table in the bottom right of the image?
Expected Answer Yes

Reasoning Question 1
Do you think that the slippers that appear at the bottom right of the image indicate they are being worn by a person?
Expected Answer Yes

Reasoning Question 2
Do you think that the slippers that appear at the bottom right of the image are not worn by the photographer?
Expected Answer Yes

Figure 3: **An example of the annotation for the "Clued" category.** Each image is annotated with a general question assessing overall suspicion and decomposed questions focused on a deterministic clue (here, the feminine bow hair accessory). Decomposed questions include perception-based inquiries (clue identification) and reasoning-based inquiries (social implications and contradictions), all annotated with the expected answer "yes".

CaughtCheatingdemands a more nuanced "detective-level" ability to identify initially inconspicuous details and infer their significance within a specific social claim.

3 Benchmark Construction

3.1 Image Collection for Real Subset

We collect images from publicly posted photographs on social media, focusing on those posts that request others to detect potential suspicious clues that violate their partners' claims from the photos. We only collect images that either clearly contain or lack subtle clues related to potential violation of the claim. Each image is manually reviewed to ensure sufficient resolution quality for identifying such clues. Due to the limited availability of images with naturally occurring subtle clues, we apply minimal cropping to some images that originally show multiple people, transforming them into single-person photos while preserving subtle indicators of another person's presence. This approach allows us to create challenging cases where the clues are interpretable for humans but not immediately obvious. To ensure practical relevance, we exclude any synthetic images generated by image generation models. After careful selection and verification, we construct a dataset of 100 images, split into ***Clued*** and ***Unclued*** categories, with all personal information removed. A detailed version of *Benchmark Construction*, including the image examples, is provided in the Appendix B.

3.2 Annotation for Real Subset

After constructing the image set, we annotate each image with a set of questions and corresponding ground-truth answers. A detailed annotated example is shown in Figure 3. For images in *Clued* category, we annotate each one using a question instantiated from the template: "*My [girlfriend/boyfriend] said [she/he] is [in a certain scenario] and sent me this photo. Do you notice anything suspicious in the image that contradicts [her/his] claim?*" Among the potential clues, the one that deterministically shows the violation of the providing claim (a clearly identifiable, contextually inappropriate element) will be selected as the ***Deterministic Clue***, e.g., a pair of slippers is being worn by someone in Figure 3. The remaining clues are labeled as ***Non-deterministic Clues*** (weaker or more ambiguous signals), e.g., the rose bouquet, the TV shows, and the far-reached drinks. These non-deterministic clues might be suspicious, but apparently not enough to infer the potential claim violation. The reason we provide these clues is to avoid punishing models when they mention these clues.

Furthermore, we construct a series of decomposed questions designed to analyze the visual reasoning process of MLLMs, shown in the right part of Figure 3. This series includes: (1) ***Decomposed Perception Question***, which assesses whether the MLLMs can identify the deterministic clue when we explicitly mention the clue and position. (2) ***Decomposed Reasoning Question***, which assesses whether MLLMs can understand the social implications of the clue, or whether MLLMs can imply the relation between the clue and the potential cheating. The correct answer to each of these decomposed questions is annotated as "yes". We annotate each image in the *Unclued* category using the same initial question template, with "There is no clear evidence." as the ground-truth answer.

3.3 Construction for Synthetic Subset

As a complement to the real subset, we further construct the synthetic subset through a cascading pipeline that balances quality and diversity. First, we define a set of scenes (e.g., bar, office, hotel room, car) and introduce deterministic clues in complete sentences (e.g., There are two glasses on the table) anchored in scene context and gender stereotypes, while ensuring gender balance. Then, for clued samples, the LLM is given both the scene and its associated clue to generate first-person photo prompts that contain elements in contradiction to the claim. For unclued samples, only the scene is provided, and the LLM is instructed to generate first-person photo prompts of a single individual without such contradiction. Candidate prompts are then reviewed by annotators for semantic accuracy and compliance, with the best five prompts per clue retained.

Subsequently, we employ a state-of-the-art image generation model (GPT-Image-1 [**?**]) to synthesize multiple candidates per prompt, from which human evaluators select the best five. Question templates are adapted from the real subset, and answers are directly derived from clue sentences, ensuring consistency with annotation protocols. This procedure yields 3700 images, evenly divided into clued and unclued categories across 10 scenes. Detailed method, including mappings between scenes and clues, prompt design, generation workflow, and distributional statistics, is provided in Appendix B.

3.4 Evaluation Metrics

We employ several evaluation metrics to comprehensively assess MLLMs' performance in detecting potential claim violations from images. ***Clued Accuracy (Clued Acc)*** measures whether MLLMs can successfully identify the key deterministic clues in images from the *Clued* category. ***Intersection over Union (Clued IoU)*** evaluates how well MLLMs identify all relevant non-deterministic clues while avoiding unrelated elements in the *Clued* category. ***Unclued Accuracy (Unclued Acc)*** assesses whether MLLMs can correctly determine the absence of suspicious clues in images from the *Unclued* category. In addition to the above three metrics, we also report the accuracy of MLLMs on the decomposed questions in the analysis, including *Decomposed Perception Accuracy (Dec. P Acc)*, *Decomposed Reasoning Accuracy (Dec. R Acc)*, and *Decomposed Overall Accuracy (Dec. Acc)* for in-depth analysis. These metrics together provide a comprehensive evaluation framework that captures both the accuracy of clue detection and the quality of reasoning in different scenarios.

To compute these metrics, we need to parse the key points from MLLMs' open-ended responses and compare them with the ground-truth answers. Given the complexity of this task and the diversity of the responses, we recommend using human evaluators as the primary judges for the most accurate assessment. However, to enable fair and automated comparison across different models, we also develop an automatic evaluation approach using GPT-4.1 to parse and compare the model response. To validate the reliability of our automatic evaluation method, we calculate the inter-rater agreement between human evaluators and GPT-4.1 using Cohen's Kappa coefficient. The resulting kappa scores of 0.82 for *Clued Acc* and 0.943 for *Unclued Acc* demonstrate strong alignment between human and automatic evaluation, indicating the reliability of our automated assessment approach. A detailed version of *Evaluation Metrics*, including the calculation and transformation between metrics, is provided in the Appendix C.

4 Experimental Results

4.1 Main Results

The main results are shown in Table 1. *Clued Acc* and *Clued IoU* represent the capability of MLLMs to identify the suspicious clues, which directly reflect the MLLMs' visual perception and reasoning abilities. For previous open-source models, the performance is almost negligible, as no models can reach an accuracy above 5%, indicating their inferior capabilities on visual perception, reasoning, or even instruction following. As for proprietary models before the reasoning era, GPT-4o achieves 4.0% accuracy and 1.0% IoU, and Gemini-2-flash achieves 10.0% accuracy and 0.0% IoU. The performances are slightly better, indicating their better capabilities in instruction understanding and following, but still they can not reach accuracies above 10%. Only for the recent strong large reasoning models, like Gemini-2.5-pro and GPT-o3, the performances can reach above 20% accuracy and 10% IoU, indicating their strong capabilities on visual perception and reasoning. But still, even the best performing model, *GPT-o3*, only achieves 26.0% accuracy and 17.2% IoU, **indicating the**

	Synthetic Subset			Real Subset					
	Overall			Clued		Unclued	Overall		
	Precision ↑	Recall ↑	F1 ↑	Acc ↑	IoU ↑	Acc ↑	Precision ↑	Recall ↑	F1 ↑
InternVL2-1B [8]	15.6	2.3	4.0	0.0	0.0	82.0	0.0	0.0	0.0
LLaVA-OV-1B [34]	33.7	3.2	5.8	0.0	0.0	86.0	0.0	0.0	0.0
InternVL2.5-1B [8]	23.2	3.5	6.1	0.0	0.0	94.0	0.0	0.0	0.0
InternVL2-2B [8]	27.1	5.6	9.3	0.0	0.0	76.0	0.0	0.0	0.0
InternVL2.5-2B [8]	26.3	4.5	7.7	0.0	0.0	68.0	0.0	0.0	0.0
Qwen2.5-VL-3B [1]	52.2	37.6	43.7	2.0	0.0	50.0	3.8	2.0	2.6
LLaVA-v1.6-Mistral-7B [34]	69.7	7.6	13.7	0.0	0.0	82.0	0.0	0.0	0.0
LLaVA-OV-7B [34]	88.7	10.2	18.3	2.0	0.0	52.0	4.0	2.0	2.7
Qwen2.5-VL-7B [1]	62.6	45.6	52.8	2.0	3.9	66.0	5.6	2.0	2.9
InternVL2-8B [8]	44.2	7.6	13.0	0.0	0.0	76.0	0.0	0.0	0.0
InternVL2.5-8B [8]	65.4	34.2	44.9	0.0	0.0	72.0	0.0	0.0	0.0
LLaVA-1.6-Vicuna-13B [34]	20.5	3.3	5.7	0.0	0.0	72.0	0.0	0.0	0.0
InternVL2-26B [8]	61.0	33.7	43.4	2.0	1.8	10.0	2.2	2.0	2.1
InternVL2.5-26B [8]	73.8	38.9	50.9	0.0	0.0	80.0	0.0	0.0	0.0
InternVL2.5-38B [8]	65.5	39.5	49.3	2.0	0.0	76.0	7.7	2.0	3.2
InternVL2-40B [8]	71.8	37.6	49.3	4.0	0.7	12.0	4.4	4.0	4.2
InternVL2-72B [8]	88.4	46.3	60.8	4.0	0.8	16.0	4.5	4.0	4.3
InternVL2.5-72B [8]	75.5	44.7	56.2	2.0	0.8	80.0	9.1	2.0	3.3
LLaVA-OV-72B [34]	77.6	42.3	54.8	0.0	1.3	72.0	0.0	0.0	0.0
GPT-4o [47]	**81.1**	53.2	64.2	4.0	1.0	54.0	8.0	4.0	5.3
Gemini-2-flash [12]	71.4	50.0	58.8	10.0	0.0	6.0	9.6	10.0	9.8
Gemini-2.5-flash [13]	71.1	84.5	77.2	18.0	5.1	22.0	18.8	18.0	18.4
Gemini-2.5-pro [13]	72.6	87.2	79.2	20.0	15.1	22.0	20.4	20.0	20.2
GPT-o3 [46]	68.6	**94.7**	**79.6**	**26.0**	**17.2**	8.0	**22.0**	**26.0**	**23.9**
Human	97.0	95.8	96.4	56.0	/	68.0	63.6	56.0	59.6

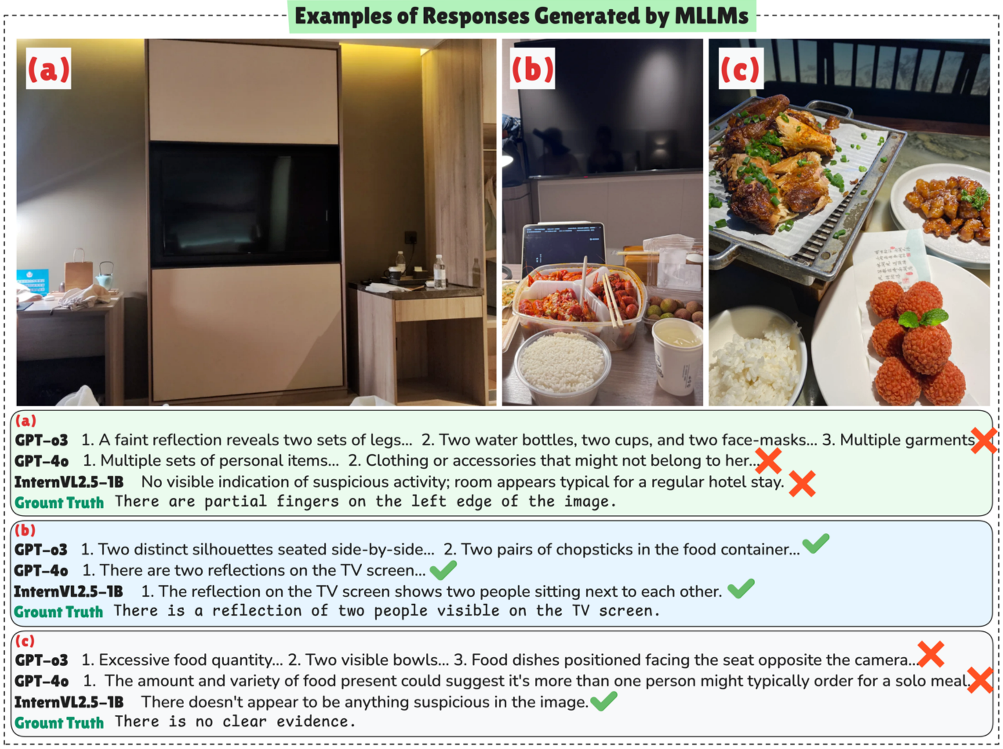
Table 1: The overall precision, recall, and F1 score for the *Synthetic Subset*, and the accuracies, IoU on the *Clued* category, and the accuracy on the *Unclued* category for the *Real Subset*, and the overall precision, recall, and F1 score for the *Real Subset*. Models are grouped by parameter size and type (open-source vs. proprietary). *Clued Acc* and *IoU* represent the capability of MLLMs to identify the suspicious clues, which directly reflects the MLLMs' visual perception and reasoning abilities. *Unclued Acc* represents the capability of MLLMs to not generate any suspicious clues if the image is unclued. On the *Real Subset*, even the best performing model, *GPT-o3*, only achieves 26.0% accuracy and 17.2% IoU, indicating the current boundary of MLLMs' capabilities. *F1* score shows the overall capability of MLLMs on CaughtCheating, where *GPT-o3*, achieves only 23.9%. The highest *F1* score is 23.9%, which is much lower than the human performance, indicating the current boundary of MLLMs' capabilities.

current boundary of MLLMs' capabilities. We believe this benchmark is challenging enough and shows the current boundary of their visual perception and reasoning capabilities.

In the meantime, we also report the *Unclued Acc* to evaluate the capability of MLLMs to not generate any suspicious clues if the image is unclued. This is also important for the real-world application, as **we do not prefer MLLMs to suspect and accuse anything if the image providers are innocent**. As shown in the table, most of the models reach high accuracies on this category; however, this performance is due to their inability to generate any suspicious clues. On the contrary, the advanced agentic models, Gemini-2.5-pro and GPT-o3, achieve low accuracies on this category, indicating their hallucination of nonexistent suspicious clues even on unclued images. These low accuracies reveal **their lack of strong reasoning abilities to identify if something is suspicious or not.**

Finally, the *F1* scores represent the overall performance of the model, which is the harmonic mean of the *Precision* and *Recall*. The highest *F1* score is 23.9%, which is much lower than the human performance, indicating the current boundary of MLLMs' capabilities.

Alternatively, on the *Synthetic Subset* with $3,700$ images, though previous models still perform poorly, existing state-of-the-art MLLMs like GPT-o3 and Gemini-2.5-pro can reach F1 scores of almost to 80%. The main reason is that the current image synthesis model, GPT-Image-1, is not able to generate scenes with subtle and non-obvious suspicious clues, e.g., the vague reflections in the mirror. On the contrary, current SOTA MLLMs can already have the ability to identify the clues generated by image synthesis models. Further discussion on the quality of synthetic data is provided in Appendix B.



(a)
GPT-o3 1. A faint reflection reveals two sets of legs... 2. Two water bottles, two cups, and two face-masks... 3. Multiple garments ✗
GPT-4o 1. Multiple sets of personal items... 2. Clothing or accessories that might not belong to her... ✗
InternVL2.5-1B No visible indication of suspicious activity; room appears typical for a regular hotel stay. ✗
Grount Truth `There are partial fingers on the left edge of the image.`

(b)
GPT-o3 1. Two distinct silhouettes seated side-by-side... 2. Two pairs of chopsticks in the food container... ✓
GPT-4o 1. There are two reflections on the TV screen... ✓
InternVL2.5-1B 1. The reflection on the TV screen shows two people sitting next to each other. ✓
Grount Truth `There is a reflection of two people visible on the TV screen.`

(c)
GPT-o3 1. Excessive food quantity... 2. Two visible bowls... 3. Food dishes positioned facing the seat opposite the camera... ✗
GPT-4o 1. The amount and variety of food present could suggest it's more than one person might typically order for a solo meal. ✗
InternVL2.5-1B There doesn't appear to be anything suspicious in the image. ✓
Grount Truth `There is no clear evidence.`

Figure 4: **Case studies of the models' performance on the CaughtCheating examples.** 3 representative models are selected, including GPT-o3, GPT-4o and InternVL2.5-1B, and 3 images are selected: (a) A difficult *Clued* image, (b) An easy *Clued* image, and (c) An *Unclued* image. The models' responses are truncated for better visualization.

	Dec. P	**Dec. R**	**Dec.**	**Clued ↑**
GPT-4o	52.0	12.8	2.0	4.0
Gemini-2-flash	74.0	69.6	38.0	10.0
Gemini-2.5-flash	72.0	39.2	20.0	18.0
Gemini-2.5-pro	80.0	52.9	34.0	20.0
GPT-o3	62.0	24.5	2.0	**26.0**
Human	82.0	97.8	80.0	56.0

Table 2: **Performance on decomposed questions.** *Dec. P* and *Dec. R* is the *Decomposed Perception Accuracy* and *Decomposed Reasoning Accuracy* of the model on the decomposed questions. *Dec.* is the *Decomposed Accuracy* which represents the proportion of the model correctly answering all the decomposed questions.

4.2 Decomposition Analysis

To better understand why the current advanced MLLMs can not perform well in this task, we design a set of decomposed questions for each image in the *Clued* category. These questions are divided into two types: perception questions, which test whether the model can accurately identify the key deterministic clue when it is explicitly mentioned, and reasoning questions, which assess whether the model can correctly infer the implications or contradictions associated. This fine-grained analysis helps reveal whether a model's failure is due to not seeing the clue at all, or seeing it but not understanding its significance, thus providing deeper insight into the limitations.

As shown in Table 2, the *Dec. P* is far higher than the *Clued Acc*, indicating that the models can identify the key deterministic clue when it is explicitly mentioned. When the human participants are given the image, it's hard for them to identify the suspicious clues at the first place, e.g. the refelction in Figure 1 and the femine bow hair in Figure 3, but once they are explicitly mentioned or pointed out, they will admit the presense of the items. This human behavior leads to the relatively high *Dec. P* for humans. For the *Dec. R*, the performances are all relatively lower, especially for GPT-4o and o3.

These results together indicate that current advanced MLLMs can identify the key subtle items in the image if they are explicitly mentioned. However, in CaughtCheating, when being asked to identify the suspicious clues without being given any hints, they tend to do an exhaustive search and generate lots of clues without really judging if the clues are suspicious or not, and at the same time, ignore the key but subtle deterministic clues. These behaviors are similar to humans and verify the hypothesis based on *Guided Search* theory.

5 Case Studies

In this section, we provide some examples to show how exactly different models perform on CaughtCheating, shown in Figure 4. In the figure, 3 representative models are selected, including GPT-o3, GPT-4o, and InternVL2.5-1B, and 3 images are selected: (a) A difficult *Clued* image, (b) An easy *Clued* image, and (c) An *Unclued* image.

In (a), there is an elbow, and fingers are visible at the left edge of the photo, clearly indicating the presence of another person. However, all the models fail to identify this subtle but deterministic clue and focus on the reflection of the television, even though there are no visible clues in the reflection, as another person is sitting by the table. What's worse, o3 and 4o keep mentioning the two bottles or cups, which are obviously provided by the hotel and can not be the suspicious clues. On the contrary, InternVL2.5-1B can not provide any clues by saying this is a normal hotel image. In (b), the reflection in the TV clearly shows there are two people on the bed, thus all the selected models can identify this clue. These 2 examples, (a) and (b), show that: (1) models are able to see through reflections, and (2) Reflection does not always contain suspicious clues, which further verifies that **CaughtCheating is challenging since there are no fixed rules for the suspicious clues.**

(c) shows an *Uncled* image, which is merely an ordinary food-sharing image. However, o3 still tries to generate a lot of *so-called suspicious clues*, including the amount of food, the place settings, etc. **This behaviour is not expected since we only want models to generate clues that are really suspicious, rather than accusing everything, which further indicates the values of CaughtCheating.** Similarly to the above examples, InternVL2.5-1B can not provide any clues by saying this is a normal food-sharing image, that's why it reaches the highest on the *Unclued Acc*.

6 Conclusion

In this work, we introduce CaughtCheating, a benchmark that stress-tests detective-level visual perception and social reasoning by asking models to find subtle, context-dependent clues that contradict a subject's claim. Grounded in Guided Search theory, our analysis reveals why these cases are uniquely challenging—targets exhibit extremely low bottom-up salience, lack top-down feature guidance, and hinge on scene meaning, making exhaustive scan strategies ineffective. The benchmark combines a carefully curated real subset and a complementary synthetic subset with a validated automatic evaluator, enabling reproducible, fine-grained assessments. Results on the real subset reveal stark limitations: even strong agentic MLLMs significantly underperform—GPT-o3 attains only 26.0% clued accuracy, 17.2% IoU, 8.0% unclued accuracy, and an overall F1 of 23.9%. As society continuously expects more sophisticated capabilities from LLMs, we anticipate that real-world complex tasks requiring detective-level abilities will become increasingly important for MLLMs to master. In summary, CaughtCheatingnot only provides a challenging new benchmark for evaluating multimodal reasoning, but also exposes critical weaknesses in current MLLMs, highlighting the urgent need for further research into the detective-level capabilities of MLLMs.

7 Ethics Concerns and Limitations

Because our benchmark relies exclusively on publicly available for the *Real Subset*, annotatable social-media photographs, the source pool overwhelmingly features cisgender, heterosexual couples. The scarcity of labeled images depicting LGBTQ+ or non-monogamous relationships, therefore compelled us to centre this demographic. The same constraint restricted the dataset to a handful of commonplace settings, such as hotels, restaurants, cafés, and vacation scenes, leaving contexts such as nightlife, workplaces, or culturally specific environments underrepresented. The benchmark addresses a single, socially charged form of visual reasoning, detecting suspected infidelity, rather than the broader universe of complex social-reasoning inferences people may draw from images. These limitations stem from the paucity of suitable public data. To alleviate these concerns, we further construct the synthetic subset through a cascading pipeline that balances quality and diversity. Although the synthetic subset is not as challenging as the real-world scenarios, it alleviates the potential ethical concerns and expands the dataset to a more diverse set of scenes.

References

[1] 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, 2025.

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[3] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*, 2024.

[4] Qiguang Chen, Libo Qin, Jin Zhang, Zhi Chen, Xiao Xu, and Wanxiang Che. M3 cot: A novel benchmark for multi-domain multi-step multi-modal chain-of-thought. *arXiv preprint arXiv:2405.16473*, 2024.

[5] Ruxiao Chen, Chenguang Wang, Yuran Sun, Xilei Zhao, and Susu Xu. From perceptions to decisions: Wildfire evacuation decision prediction with behavioral theory-informed llms. *arXiv preprint arXiv:2502.17701*, 2025.

[6] Xingyu Chen, Jiahao Xu, Tian Liang, Zhiwei He, Jianhui Pang, Dian Yu, Linfeng Song, Qiuzhi Liu, Mengfei Zhou, Zhuosheng Zhang, et al. Do not think that much for 2+ 3=? on the overthinking of o1-like llms. *arXiv preprint arXiv:2412.21187*, 2024.

[7] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024.

[8] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024.

[9] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023.

[10] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

[11] Edirlei Soares de Lima, Marco A Casanova, Bruno Feijó, and Antonio L Furtado. Characterizing the investigative methods of fictional detectives with large language models. *arXiv preprint arXiv:2505.07601*, 2025.

[12] Google DeepMind. Gemini 2.0 flash, 2025.

[13] Google DeepMind. Gemini 2.5 pro preview model card. Technical report, Google DeepMind, May 2025.

[14] Huilin Deng, Ding Zou, Rui Ma, Hongchen Luo, Yang Cao, and Yu Kang. Boosting the generalization and reasoning of vision language models with curriculum reinforcement learning. *arXiv preprint arXiv:2503.07065*, 2025.

[15] Yihe Deng, Hritik Bansal, Fan Yin, Nanyun Peng, Wei Wang, and Kai-Wei Chang. Openvlthinker: An early exploration to complex vision-language reasoning via iterative self-improvement. *arXiv preprint arXiv:2503.17352*, 2025.

[16] Yihe Deng, Pan Lu, Fan Yin, Ziniu Hu, Sheng Shen, Quanquan Gu, James Y Zou, Kai-Wei Chang, and Wei Wang. Enhancing large vision language models with self-training on image comprehension. *Advances in Neural Information Processing Systems*, 37:131369–131397, 2024.

[17] John Duncan and Glyn W Humphreys. Visual search and stimulus similarity. *Psychological review*, 96(3):433, 1989.

[18] Chenrui Fan, Ming Li, Lichao Sun, and Tianyi Zhou. Missing premise exacerbates overthinking: Are reasoning models losing critical thinking skill? *arXiv preprint arXiv:2504.06514*, 2025.

[19] Xingyu Fu, Minqian Liu, Zhengyuan Yang, John Corring, Yijuan Lu, Jianwei Yang, Dan Roth, Dinei Florencio, and Cha Zhang. Refocus: Visual editing as a chain of thought for structured image understanding. *arXiv preprint arXiv:2501.05452*, 2025.

[20] Kanishk Gandhi, Jan-Philipp Fränken, Tobias Gerstenberg, and Noah Goodman. Understanding social reasoning in language models with language models. *Advances in Neural Information Processing Systems*, 36:13518–13529, 2023.

[21] Zhouhong Gu, Lin Zhang, Jiangjie Chen, Haoning Ye, Xiaoxuan Zhu, Zihan Li, Zheyu Ye, Yan Gao, Yao Hu, Yanghua Xiao, et al. Piecing together clues: A benchmark for evaluating the detective skills of large language models. *arXiv preprint arXiv:2307.05113*, 2023.

[22] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[23] Yunzhuo Hao, Jiawei Gu, Huichen Will Wang, Linjie Li, Zhengyuan Yang, Lijuan Wang, and Yu Cheng. Can mllms reason in multimodality? emma: An enhanced multimodal reasoning benchmark. *arXiv preprint arXiv:2501.05444*, 2025.

[24] Jingyuan Huang, Jen-tse Huang, Ziyi Liu, Xiaoyuan Liu, Wenxuan Wang, and Jieyu Zhao. Vlms as geoguessr masters: Exceptional performance, hidden biases, and privacy risks. *arXiv preprint arXiv:2502.11163*, 2025.

[25] Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*, 2025.

[26] Laurent Itti and Christof Koch. Computational modelling of visual attention. *Nature reviews neuroscience*, 2(3):194–203, 2001.

[27] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 20(11):1254–1259, 2002.

[28] Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

[29] Bowen Jin, Chulin Xie, Jiawei Zhang, Kashob Kumar Roy, Yu Zhang, Zheng Li, Ruirui Li, Xianfeng Tang, Suhang Wang, Yu Meng, et al. Graph chain-of-thought: Augmenting large language models by reasoning on graphs. *arXiv preprint arXiv:2404.07103*, 2024.

[30] Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. *arXiv preprint arXiv:2503.09516*, 2025.

[31] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910, 2017.

[32] 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, 2022.

[33] Seungpil Lee, Woochang Sim, Donghyeon Shin, Wongyu Seo, Jiwon Park, Seokki Lee, Sanha Hwang, Sejin Kim, and Sundong Kim. Reasoning abilities of large language models: In-depth analysis on the abstraction and reasoning corpus. *ACM Transactions on Intelligent Systems and Technology*, 2024.

[34] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer, 2024.

[35] Ming Li, Zhengyuan Yang, Xiyao Wang, Dianqi Li, Kevin Lin, Tianyi Zhou, and Lijuan Wang. What makes reasoning models different? follow the reasoning leader for efficient decoding. *arXiv preprint arXiv:2506.06998*, 2025.

[36] Ming Li, Ruiyi Zhang, Jian Chen, Jiuxiang Gu, Yufan Zhou, Franck Dernoncourt, Wanrong Zhu, Tianyi Zhou, and Tong Sun. Towards visual text grounding of multimodal large language model. *arXiv preprint arXiv:2504.04974*, 2025.

[37] Yijun Liang, Ming Li, Chenrui Fan, Ziyue Li, Dang Nguyen, Kwesi Cobbina, Shweta Bhardwaj, Jiuhai Chen, Fuxiao Liu, and Tianyi Zhou. Colorbench: Can vlms see and understand the colorful world? a comprehensive benchmark for color perception, reasoning, and robustness. *arXiv preprint arXiv:2504.10514*, 2025.

[38] Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*, 2023.

[39] 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.

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

[41] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023.

[42] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022.

[43] Weidi Luo, Qiming Zhang, Tianyu Lu, Xiaogeng Liu, Yue Zhao, Zhen Xiang, and Chaowei Xiao. Doxing via the lens: Revealing privacy leakage in image geolocation for agentic multi-modal large reasoning model. *arXiv preprint arXiv:2504.19373*, 2025.

[44] Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Tiancheng Han, Botian Shi, Wenhai Wang, Junjun He, et al. Mm-eureka: Exploring the frontiers of multimodal reasoning with rule-based reinforcement learning. *arXiv preprint arXiv:2503.07365*, 2025.

[45] Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. s1: Simple test-time scaling. *arXiv preprint arXiv:2501.19393*, 2025.

[46] OpenAI. Openai o3 and o4-mini system card. Technical report, OpenAI, April 2025.

[47] OpenAI, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and etc. Gpt-4o system card, 2024.

[48] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[49] Arkil Patel, Satwik Bhattamishra, and Navin Goyal. Are nlp models really able to solve simple math word problems? *arXiv preprint arXiv:2103.07191*, 2021.

[50] Shuai Peng, Di Fu, Liangcai Gao, Xiuqin Zhong, Hongguang Fu, and Zhi Tang. Multimath: Bridging visual and mathematical reasoning for large language models. *arXiv preprint arXiv:2409.00147*, 2024.

[51] Yingzhe Peng, Gongrui Zhang, Miaosen Zhang, Zhiyuan You, Jie Liu, Qipeng Zhu, Kai Yang, Xingzhong Xu, Xin Geng, and Xu Yang. Lmm-r1: Empowering 3b lmms with strong reasoning abilities through two-stage rule-based rl. *arXiv preprint arXiv:2503.07536*, 2025.

[52] Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Chen Bo Calvin Zhang, Mohamed Shaaban, John Ling, Sean Shi, et al. Humanity's last exam. *arXiv preprint arXiv:2501.14249*, 2025.

[53] Leonardo Ranaldi and Andrè Freitas. Self-refine instruction-tuning for aligning reasoning in language models. *arXiv preprint arXiv:2405.00402*, 2024.

[54] Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. Solving math word problems with process-and outcome-based feedback. *arXiv preprint arXiv:2211.14275*, 2022.

[55] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. Git: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100*, 2022.

[56] Xiyao Wang, Jiuhai Chen, Zhaoyang Wang, Yuhang Zhou, Yiyang Zhou, Huaxiu Yao, Tianyi Zhou, Tom Goldstein, Parminder Bhatia, Furong Huang, et al. Enhancing visual-language modality alignment in large vision language models via self-improvement. *arXiv preprint arXiv:2405.15973*, 2024.

[57] Xiyao Wang, Zhengyuan Yang, Chao Feng, Yongyuan Liang, Yuhang Zhou, Xiaoyu Liu, Ziyi Zang, Ming Li, Chung-Ching Lin, Kevin Lin, et al. Vicrit: A verifiable reinforcement learning proxy task for visual perception in vlms. *arXiv preprint arXiv:2506.10128*, 2025.

[58] Xiyao Wang, Zhengyuan Yang, Chao Feng, Hongjin Lu, Linjie Li, Chung-Ching Lin, Kevin Lin, Furong Huang, and Lijuan Wang. Sota with less: Mcts-guided sample selection for data-efficient visual reasoning self-improvement. *arXiv preprint arXiv:2504.07934*, 2025.

[59] Xiyao Wang, Zhengyuan Yang, Linjie Li, Hongjin Lu, Yuancheng Xu, Chung-Ching Lin, Kevin Lin, Furong Huang, and Lijuan Wang. Scaling inference-time search with vision value model for improved visual comprehension. *arXiv preprint arXiv:2412.03704*, 2024.

[60] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

[61] Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, et al. Charxiv: Charting gaps in realistic chart understanding in multimodal llms. *Advances in Neural Information Processing Systems*, 37:113569–113697, 2024.

[62] 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.

[63] Liang Wen, Yunke Cai, Fenrui Xiao, Xin He, Qi An, Zhenyu Duan, Yimin Du, Junchen Liu, Lifu Tang, Xiaowei Lv, et al. Light-r1: Curriculum sft, dpo and rl for long cot from scratch and beyond. *arXiv preprint arXiv:2503.10460*, 2025.

[64] Jeremy M Wolfe, Kyle R Cave, and Susan L Franzel. Guided search: an alternative to the feature integration model for visual search. *Journal of Experimental Psychology: Human perception and performance*, 15(3):419, 1989.

[65] Jeremy M Wolfe and Todd S Horowitz. Five factors that guide attention in visual search. *Nature human behaviour*, 1(3):0058, 2017.

[66] Junde Wu, Jiayuan Zhu, and Yuyuan Liu. Agentic reasoning: Reasoning llms with tools for the deep research. *arXiv preprint arXiv:2502.04644*, 2025.

[67] Wenshan Wu, Shaoguang Mao, Yadong Zhang, Yan Xia, Li Dong, Lei Cui, and Furu Wei. Minds eye of llms: Visualization-of-thought elicits spatial reasoning in large language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[68] Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. Less: Selecting influential data for targeted instruction tuning. *arXiv preprint arXiv:2402.04333*, 2024.

[69] Yuxi Xie, Anirudh Goyal, Wenyue Zheng, Min-Yen Kan, Timothy P Lillicrap, Kenji Kawaguchi, and Michael Shieh. Monte carlo tree search boosts reasoning via iterative preference learning. *arXiv preprint arXiv:2405.00451*, 2024.

[70] Guowei Xu, Peng Jin, Li Hao, Yibing Song, Lichao Sun, and Li Yuan. Llava-o1: Let vision language models reason step-by-step. *arXiv preprint arXiv:2411.10440*, 2024.

[71] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822, 2023.

[72] Bin Yu, Hang Yuan, Yuliang Wei, Bailing Wang, Weizhen Qi, and Kai Chen. Long-short chain-of-thought mixture supervised fine-tuning eliciting efficient reasoning in large language models. *arXiv preprint arXiv:2505.03469*, 2025.

[73] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023.

[74] Yuan Yuan, Muyu He, Muhammad Adil Shahid, Jiani Huang, Ziyang Li, and Li Zhang. Turnaboutllm: A deductive reasoning benchmark from detective games. *arXiv preprint arXiv:2505.15712*, 2025.

[75] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9556–9567, June 2024.

[76] Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, et al. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. *arXiv preprint arXiv:2409.02813*, 2024.

[77] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6720–6731, 2019.

[78] Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, et al. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *European Conference on Computer Vision*, pages 169–186. Springer, 2024.

[79] Ruohong Zhang, Bowen Zhang, Yanghao Li, Haotian Zhang, Zhiqing Sun, Zhe Gan, Yinfei Yang, Ruoming Pang, and Yiming Yang. Improve vision language model chain-of-thought reasoning. *arXiv preprint arXiv:2410.16198*, 2024.

[80] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*, 2023.

[81] Pei Zhou, Aman Madaan, Srividya Pranavi Potharaju, Aditya Gupta, Kevin R McKee, Ari Holtzman, Jay Pujara, Xiang Ren, Swaroop Mishra, Aida Nematzadeh, et al. How far are large language models from agents with theory-of-mind? *arXiv preprint arXiv:2310.03051*, 2023.

[82] Yiyang Zhou, Zhiyuan Fan, Dongjie Cheng, Sihan Yang, Zhaorun Chen, Chenhang Cui, Xiyao Wang, Yun Li, Linjun Zhang, and Huaxiu Yao. Calibrated self-rewarding vision language models. *arXiv preprint arXiv:2405.14622*, 2024.

Table of Contents for Appendix

A Relate Work **17**

A.1 LLM reasoning 17

A.2 MLLM reasoning 17

B Detailed Benchmark Construction **19**

B.1 Image Collection for Real Subset 19

B.2 Annotation for Real Subset 19

B.3 Construction for Synthetic Subset 20

B.4 Discussion for Synthetic Subset 20

B.5 Data Distribution for Real Subset 21

B.6 Data Distribution for Synthetic Subset 22

B.7 Prompt for Synthetic Subset Construction 24

C Evaluation Metrics **25**

D Evaluation Prompt **27**

E o3 Reasoning Traces for Qualitative Examples **30**

A Relate Work

A.1 LLM reasoning

The chain-of-thought technique [62, 32] represents the early efforts in exploring the reasoning capabilities of large language models (LLMs) [2, 9]. By explicitly generating intermediate reasoning steps, this method notably enhances performance across various reasoning tasks [49, 10]. Moreover, advances in decoding strategies have introduced additional test-time computation to further boost performance. For instance, Self-Consistency sampling [60], which employs voting mechanisms to select from multiple reasoning paths, has notably increased reliability. Expanding beyond linear reasoning processes, structured frameworks such as Tree-of-thought [71] or Graph-of-thought [29]facilitate the exploration of multiple candidate reasoning paths within branched subspaces before reaching a final conclusion. Other research investigates manipulating the reasoning process to generate longer chains of thought than those typically observed, either by explicitly prompting extended reasoning chains [45] or by integrating human-like cognitive theory foundations into the inference process [81, 20, 33, 5]. Furthermore, supervised fine-tuning (SFT) not only improves general instruction-following performance [48, 68] but has also been demonstrated to significantly enhance multi-step reasoning capabilities when trained on structured chain-of-thought (CoT) traces, where models learn to explicitly generate intermediate reasoning steps [53, 63, 45, 72, 35]. Additionally, prior research has employed reward models during training to evaluate each intermediate reasoning step individually, rather than solely assessing final outcomes, further improving reasoning performance [54, 38]. This approach integrates effectively with Monte Carlo Tree Search techniques [69], providing valuable insights into performance gains achieved through fine-grained value estimations. Beyond training, many studies augment the reasoning process with the ability to invoke external tools and knowledge sources, a paradigm known as "agentic reasoning" [66]. In this paradigm, LLMs call tools such as calculators, code interpreters, web search, and other utilities to provide context from the tools' results into the reasoning process to solve complex tasks. For instance, Jin et al. [30] introduce the Search-R1, which lets an LLM query a search engine and condition subsequent reasoning on the retrieved evidence. Recent developments in large-scale reinforcement learning, relying solely on outcome-based rewards, have demonstrated potential for inducing emergent multi-step reasoning capabilities [22, 28]. While the advancements in reasoning also potentially lead to the emergence of overthinking issues [6, 18]. Such advancements underscore the importance of tasks that can be automatically verified (e.g., RL can be effectively scaled up with minimal noise in its reward signals).

A.2 MLLM reasoning

Recent developments in MLLMs [55, 40, 47, 39, 8, 7, 1] have led to the exploration of multimodal chain-of-thought techniques aimed at enhancing performance on visual reasoning tasks [73, 41, 23] with both textual reasoning process [42, 80] and multimodal reasoning path [67, 19]. Methods such as rationale distillation and self-reflection have also been employed to strengthen reasoning capabilities [79, 82, 56, 59, 16]. Besides, LLaVA-o1 [70] proposes a fine-tuning strategy that leverages a dataset enriched with structured reasoning annotations (e.g., summarization, visual analysis, logical deduction, conclusion), achieving substantial performance improvement. Inspired by successes in reinforcement learning of LLMs, recent efforts have similarly applied this method to visual math problems and other visual question-answering tasks [15, 25, 58, 51, 44]. For example, Curr-ReFT [14] introduces a three-stage progression paradigm that blends RL with curriculum design to mimic the student learning process, significantly improving generalization and step-by-step reasoning capability. Although these approaches have improved performance on visual math and STEM-related questions, substantial progress in fine-grained visual perception remains limited. For instance, MMMU [75] shows that current MLLMs, though strong on everyday tasks, stumble on domain-specific reasoning and complex, specialized imagery; many items can be solved from textual cues or memorized facts without genuine visual grounding. Its successor, MMMU-Pro [76], reinforces these findings and demonstrates that prompts encouraging explicit multi-step linguistic reasoning boost performance, provided the model truly incorporates visual evidence at each step. Similarly, MultiMath [50] reveals that many MLLMs are underperforming with purely visual inputs with minimal text, indicating that the understanding of complex spatial reasoning in mathematical or scientific diagrams remains challenging. TRIG [36], proposing the first visual text grounding task, shows the inability of MLLMs to perform visual reasoning and grounding. ColorBench [37] introduces the first comprehensive benchmark for color perception, reasoning, and robustness, showcasing the low capability of MLLMs on color-related perception and reasoning. ViCrit [57], on the other hand, introduces the verifiable reinforcement learning proxy task for visual perception in VLMs.

B Detailed Benchmark Construction

B.1 Image Collection for Real Subset

For CaughtCheatingimages, we use publicly posted photographs from social media. We manually search and review all the comments for each image to assess their suitability. Selected images must either contain or lack subtle, suspicious clues related to potential claim violation. The judgment of image candidates is based not only on the comments but also on human evaluation. Additionally, each image must have sufficient resolution quality to allow us to directly identify such clues, rather than rely on implications from blurred or indistinct objects.

Due to the limited availability of images with clear, subtle clues from public sources, we also include minimally modified versions of images containing direct clues (e.g., a clearly visible person or untypical belongings suggesting the presence of another individual). We apply simple cropping to these images to obscure the direct clues. As shown in Figure 5, the original photo shows a person sitting on the couch. After cropping, only their back remains visible, making the clue still interpretable for humans, yet challenging for MLLMs.

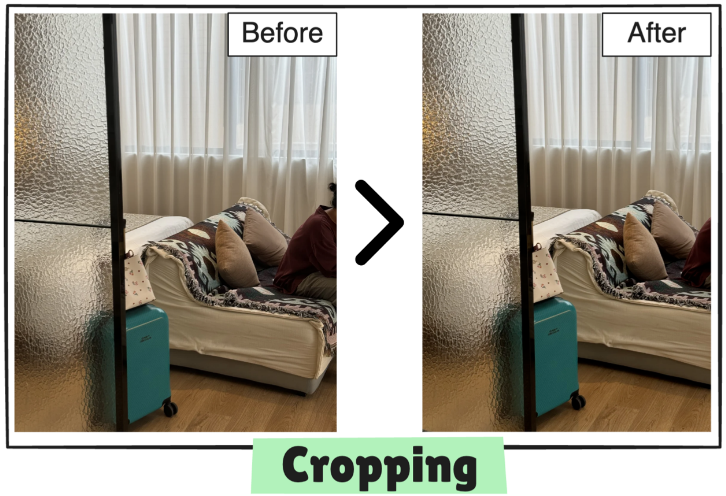


Figure 5: **Example of cropping an image for *with-clue* category**. The original photo shows part of the person sitting on the sofa (Before). By cropping (After), we can still infer there is a person, but identifying the clue is more subtle and challenging for MLLM.

After collecting a sufficient number of candidate images, we meticulously selected 100 images, split into *Clued* and *Unclued* categories, to construct the image set for CaughtCheatingbenchmark. These images are collected from the *Real Subset*. All the images are verified manually to make sure the clues are solid and no personal information exists on the image.

B.2 Annotation for Real Subset

After constructing the image set, we annotate each image with a set of questions and corresponding ground-truth answers. A detailed annotated example is shown in Figure 3. For images in *Clued* category, we annotate each one using a question instantiated from the template: "*My [girlfriend/boyfriend] said [she/he] is [in a certain scenario] and sent me this photo. Do you notice anything suspicious in the image that contradicts [her/his] claim?*" Among the potential clues, the one that deterministically shows the violation of the providing claim (a clearly identifiable, contextually inappropriate element) will be selected as the ***Deterministic Clue***, e.g., a pair of slippers is being worn by someone in Figure 3. The remaining clues are labeled as ***Non-deterministic Clues*** (weaker or more ambiguous signals), e.g., the rose bouquet, the TV shows, and the far-reached drinks. These non-deterministic clues might be suspicious, but apparently not enough to infer the potential claim violation. The reason we provide these clues is to avoid punishing models when they mention these clues.

Furthermore, we construct a series of decomposed questions designed to analyze the visual reasoning process of MLLMs, shown in the right part of Figure 3. This series includes: (1) ***Decomposed Perception Question***, which assesses whether the MLLMs can identify the deterministic clue when we explicitly mention the clue and position. (2) ***Decomposed Reasoning Question***, which assesses whether MLLMs can understand the social implications of the clue, or whether MLLMs can imply the relation between the clue and the potential lie. The correct answer to each of these decomposed questions is annotated as “yes”. These decomposed questions can be utilized for in-depth analysis on why MLLMs can not solve the question.

These decomposed questions can be utilized for in-depth analysis on why MLLMs can not solve the question. (1) If the MLLMs have low accuracy on perception-related decomposed questions, it means the low performance is caused by their poor visual perception ability. (2) If the MLLMs have low accuracy on reasoning-related decomposed questions, it means the low performance is caused by their poor visual reasoning ability. (3) If the MLLMs have relatively high accuracy on both types of decomposed questions, it means they have the necessary capabilities to solve the task, but they do not know where to start. For images in the *Unclued* category, we annotate each using the same initial question template, with the ground-truth answer labeled as "There is no clear evidence."

B.3 Construction for Synthetic Subset

We construct the synthetic subset through a cascading pipeline that balances image quality and diversity, starting from predefined scenes and deterministic clues.

First, we define a set of scenes potentially involving suspicious behaviors (e.g., bar, office, hotel room, car). For each scene, deterministic clues are provided solely for the clued category, and their design is anchored to the scene’s context. These clues are drafted as complete sentences (e.g., there are two glasses with wine on the table) rather than keywords in practice, and they are designed according to gender stereotypes while ensuring a balanced distribution across genders. For instance, wine glasses may appear as a clue in the bar scene but not in other contexts, such as the scene inside the car. While clues are restricted to the clued category, scene definitions apply to both the clued and unclued categories. Notably, while detailed mappings between scenes, clues, and their statistics of our experiment are given in the Table 4 and Table 5, they can still be further scaled as needed.

Subsequently, we employ a LLM to generate multiple candidate prompts for both categories. For the clued category, the LLM is provided with both the scene and its associated clue; for the unclued category, only the scene is provided, with an explicit instruction to depict a single individual. To mitigate potential conflicts arising from gender-sensitive stereotypes, information about the photographer’s gender is also incorporated where relevant. Moreover, to ensure a balanced number of generated images in both categories, the number of prompts corresponding to each gender within the same scene is kept identical. After candidate prompts are generated, annotators review them for semantic accuracy and compliance with the predefined constraints, and then select the five best prompts per clue within each scene. Moreover, to ensure a balanced number of generated images in both categories, the number of prompts corresponding to each gender within the same scene is kept identical. The prompts we used to produce the image generation prompt are illustrated in Figure 7 and Figure 8

Finally, we use a state-of-the-art image generation model (e.g., GPT-image-1 [**?**])to synthesize images from these prompts. To scale up the dataset and improve diversity, multiple candidates are generated for each prompt, and human evaluators select the top five. Meanwhile, we adapt the question templates from the real subset and directly adopt the answers from the clue-defining sentences, so that the constructed question–answer pairs are fully consistent with the annotation protocol and data organization of the real subset.

As a result, the procedure produces 3700 images in total, evenly split between 1850 clued and 1850 unclued samples spanning 10 distinct scenes. Comprehensive statistics and distributions are reported in the Table 6.

B.4 Discussion for Synthetic Subset

To enable efficient large-scale data construction, we generate prompts only from the photographer’s gender, the scene, and a deterministic clue, rather than drafting case-specific, highly detailed descriptions by human experts. This design ensures scalability and consistency, but it also introduces trade-offs: clues are generally more prominent, non-deterministic clues are ignored, and incidental distractions such as irrelevant personal belongings are largely absent. Consequently, compared with the image in the real subset, synthetic images are “cleaner,” with less background variation and fewer competing elements, making them easier for MLLMs to perceive, as demonstrated in Figure 6.

However, although the perceptual difficulty is lower, the deterministic clues still embed non-trivial social semantics and contextual implications. These clues, while visually more salient, continue to convey cultural, gender, and situational meanings that require careful interpretation. For example, the lipstick and handbag on the table in Figure 6 are easy to recognize, yet they still imply the presence of a woman in the room. Furthermore, as shown in Table 1, many MLLMs continue to struggle in detecting the contradiction to the claim under these conditions, reinforcing that the social and semantic challenge remains essential even if perceptual noise has been reduced for scalability. Thus, although the synthetic subset cannot fully replicate the subtlety and complexity of real subset, it represents a trade-off that lowers perceptual difficulty but preserves the benchmark’s practical value for systematic evaluation.

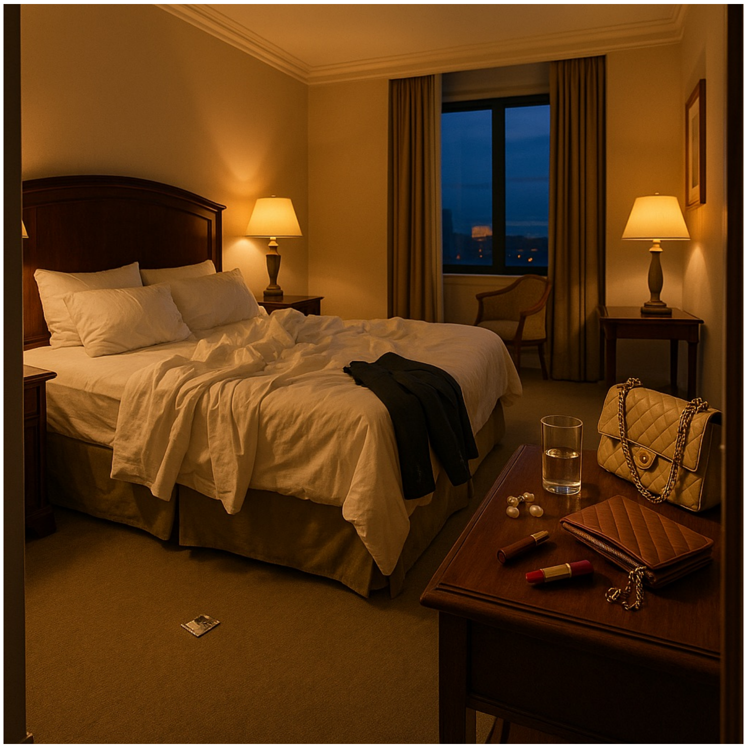


Figure 6: **Example image from the synthetic subset.** Distinct clues, such as a scattered lipstick, are clearly visible. While the image may still contain some irrelevant objects, the overall noise level is low. Compared to the real subset, the perception process is relatively easier, even though the complex social implications remain present.

B.5 Data Distribution for Real Subset

Category	Scene	Male	Female	# Dec. P	# Dec. R	# Total
Clued	Dining	6	5	11	21	50
	Hotel	25	12	37	67	
	Karaoke bar	2	0	2	3	
Unclue	Dining	7	11	0	0	50
	Hotel	13	19	0	0	

Table 3: Distribution of scenes, interlocutor gender, and question types across the two clue categories.

Our dataset comprises 100 samples collected from publicly posted photos, each manually annotated with a series of questions accompanied by ground-truth answers. These samples serve as test cases to evaluate the capability of MLLMs in detecting potential claim violations.

The dataset is evenly divided into two categories: the *Clued* category (50 samples), which includes clear indicators of potential claim violation, and the *Unclued* category (50 samples), which lacks explicit indicators. This balanced distribution aims to minimize class bias and ensure fair evaluation. Furthermore, the dataset encompasses three distinct scene types based on photo backgrounds: hotels, dining venues, and karaoke bars. The gender attributes assigned to each sample reflect the photographer's gender as inferred from the provided descriptions of the photos. These attributes do not pertain to any individuals depicted within the images. The gender categorization currently includes male and female solely based on limited available descriptive information.

Detailed statistics regarding scenario distribution and gender breakdown are summarized in Table 3. Hotel scenes comprise the majority of the dataset (69%), aligning with their prominence as typical settings for potentially suspicious scenarios. Dining venues account for 29% of the dataset, and karaoke bars represent the remaining 2%. Gender distribution is 55% male photographers and 45% female photographers.

Additionally, the dataset includes annotations for perception and reasoning questions derived from decomposition queries. Specifically, it contains 50 perception questions and 91 reasoning questions, thoroughly evaluating why MLLMs may fail to resolve specific queries. Detailed counts corresponding to scene types are provided in Table 3. Each sample averages approximately two reasoning questions, enabling comprehensive analysis of MLLM performance concerning both explicit clues and the broader social or environmental context.

B.6 Data Distribution for Synthetic Subset

As illustrated in Section 3, to ensure consistency between clue definitions and dataset distribution, we first curated five human-selected prompts for each scene clue mapping. Each prompt was used to generate multiple candidate images, from which five were carefully chosen. Therefore, data distribution in the synthetic subset is highly correlated to the number of predefined rules under each scene.

The synthetic subset comprises 3700 images, evenly split between the clued subset (1850 images) and the unclued subset (1850 images). The gender of the photographer is balanced across the dataset: 1900 images (51%) are attributed to male and 1,800 images (49%) to female photographers. This balance is deliberately maintained to minimize bias when evaluating model performance.

Scene distribution covers ten distinct environments. The largest categories are Gym, Hospital, and Hotel, each contributing 500 images (14%), collectively representing 42% of the dataset. The Bar scene accounts for 400 images (11%), while Office and Restaurant each contribute 350 images (9%). Mid-scale categories include Beach, Cafe, and Library, with 300 images each (8%). The Car scene contributes the fewest images, with 200 (5%). Importantly, this allocation is consistently mirrored across both Clued and Unclued subsets, ensuring proportional representation of all environments.

Overall, the dataset reflects a carefully controlled design, balancing clue type, gender, and scene. The systematic design ensures that generated samples are faithful to the intended definitions of "clued" and "unclued," while the proportional scene distribution provides diverse yet structured coverage of everyday contexts. This deliberate construction supports rigorous and fair evaluation of multimodal large language models across a wide spectrum of scenarios.

Scene	**# Clues**	**Clues**				
Bar	4	two glasses	lady handbag	another hand	mirror & lipstick	-
Beach	3	another foot	two cups	two bicycle	-	-
Cafe	3	lady clothes	lipstick	makeup mirror	-	-
Car	2	another foot	another leg	-	-	-
Gym	5	two bottles	another arm	mirror figure	lady hair clip	lady bracelet
Hospital	5	another back	rose & love card	lady handbag	extra slippers	card piles
Hotel	5	lady handbag	lady clothes	lady hairclip	doll & rose	lipstick
Library	3	two open book	lady hair clip	two test paper	-	-
Office	4	roses	lunchbox	lady clothes	another foot	-
Restaurant	4	two tableware	lady handbag	two steaks	another arm	-

Table 4: Deterministic clues for the synthetic subset construction categorized by scene in male.

Scene	**# Clues**	**Clues**				
Bar	4	two phones	men wallet table	another arm	two straws	-
Beach	3	two sunglasses	another arm	two single kayak	-	-
Cafe	3	two cakes	two coffees	two laptops	-	-
Car	2	another arm	two drinks	-	-	-
Gym	5	two towels	another foot	two earphone cases	men tank top	men watch
Hospital	5	another arm	TV reflection	lunchbox	wet umbrella	another head
Hotel	5	TV reflection	another arm	man shoes	five suitcases	men suit
Library	3	two cups	another arm	two chargers	-	-
Office	3	date note	two ID badges	monitor reflection	-	-
Restaurant	3	two phones	man clothes	two glasses	-	-

Table 5: Deterministic clues for the synthetic subset construction categorized by scene in female.

Scene	***Clued***			***Unclued***			**# Images**
	Male	**Female**	**Total**	**Male**	**Female**	**Total**	
Bar	100	100	200	100	100	200	400
Beach	75	75	150	75	75	150	300
Cafe	75	75	150	75	75	150	300
Car	50	50	100	50	50	100	200
Gym	125	125	250	125	125	250	500
Hospital	125	125	250	125	125	250	500
Hotel	125	125	250	125	125	250	500
Library	75	75	150	75	75	150	300
Office	100	75	175	100	75	175	350
Restaurant	100	75	175	100	75	175	350
Total	**950**	**900**	**1850**	**950**	**900**	**1850**	**3700**

Table 6: Distribution of images in the synthetic subset across scenes, genders, and categories.

B.7 Prompt for Synthetic Subset Construction

These two prompts provide a unified framework for constructing both the clued and unclued categories of the synthetic subset. By simply specifying the scene and, when needed, the clue, the dataset can be expanded systematically without extensive manual effort. This design makes large-scale generation both efficient and consistent.

Prompt for generating image-synthesis prompts for the clued category

---

System Prompt
You are a helpful assistant.

User Prompt
Your task is to generate a hyper-realistic first-person perspective smartphone photograph. The image should appear exactly like a candid snapshot taken by a [Gender] photographer, showing only what is naturally visible ahead—no hands or legs. The setting is [Scene], rendered with authentic lighting, textures, reflections, and natural depth of field, making the result indistinguishable from a real photo. Within the scene, [Clue], casually present or partially obscured, that indirectly suggests the presence of another person near the photographer. This element must be discreet, blending naturally into the environment rather than staged.
The final image must emphasize realism, immersion, and authenticity, capturing the spontaneity of a genuine phone camera shot.

Figure 7: The prompt used to instruct GPT-4.1 to generate image-synthesis prompts for the clued category.

Prompt for generating image-synthesis prompts for the unclued category

---

System Prompt
You are a helpful assistant.

User Prompt
Your task is to generate a hyper-realistic first-person perspective smartphone photograph. The image should look exactly like a candid snapshot taken by a [Gender] photographer, showing only what is naturally visible ahead—no hands or legs—just the authentic field of view from the photographer's perspective.
The setting is [Scene], captured with natural lighting, textures, reflections, and realistic depth of field.
The atmosphere must clearly emphasize that the photographer is alone in this environment, evoking the feeling of solitary presence within the scene.
The final image must be indistinguishable from a genuine phone camera shot, emphasizing realism, immersion, and authenticity.

Figure 8: The prompt used to instruct GPT-4.1 to generate image-synthesis prompts for the unclued category.

C Evaluation Metrics

We apply several evaluation metrics in our study, each designed to assess a distinct aspect of the visual reasoning process. All metrics rely on analyzing and comparing the ground-truth answers with the responses generated by MLLMs.

Clued Accuracy (Clued Acc) Deterministic Accuracy is designed to evaluate whether an MLLM successfully identifies the deterministic clue hidden in the images from *Clued* category. Let $k_i \in \{0,1\}$ denote the binary judgment for the i-th example in the *Clued* category, where $k_i = 1$ if the *Deterministic Clue* is correctly identified, and $k_i = 0$ otherwise. The **Clued Acc** is then defined as:

$$\text{Clued Acc} = \frac{1}{N_{clued}} \sum_{i=1}^{N_{clued}} k_i$$

where N_{clued} is the total number of examples in the *Clued* subset.

Intersection over Union (Clued IoU) In this context, IoU is designed to evaluate whether an MLLM correctly identifies all relevant *Non-deterministic Clues* hidden in the images from *Clued* category, while avoiding unrelated or incorrect elements. If the MLLM generates a lot of unrelated clues, this IoU value will be low, since we expect MLLMs only to mention clues that are at least somewhat suspicious.

Let G_i be the set of all the clues annotated in the ground-truth for the i-th example in the *Clued* category, and R_i be the set of clues identified by the MLLM. The **Clued IoU** is then defined as:

$$\text{IoU} = \frac{1}{N_{clued}} \sum_{i=1}^{N_{clued}} \frac{|G_i \cap R_i|}{|G_i \cup R_i|}$$

Decomposed Accuracies This evaluation comprises three specific accuracy metrics: ***Decomposed Perception Accuracy (Dec. P Acc)*** provides detailed insights into the model's performance in accurately perceiving claims from images when the clues are explicitly mentioned; ***Decomposed Reasoning Accuracy (Dec. R Acc)*** evaluates the model's proficiency in reasoning towards the deterministic clue; and ***Decomposed Overall Accuracy (Dec. Acc)*** offers a comprehensive evaluation by combining performance in both perception and reasoning dimensions. This metric is specifically tailored for images within the *Clued* category.

Let $\mathcal{P}_i$ be the set of perception-related questions for the i-th example in the *Clued* category, and $\widehat{\mathcal{P}}_i \subseteq \mathcal{P}_i$ be the subset that the MLLM correctly answered perception-related questions. The ***Decomposed Perception Accuracy (Dec. P Acc)*** is then defined as:

$$\text{Dec. P Acc} = \frac{1}{N_{clued}} \sum_{i=1}^{N_{clued}} \frac{|\widehat{\mathcal{P}}_i|}{|\mathcal{P}_i|}$$

Likewise, let $\mathcal{R}_i$ and $\widehat{\mathcal{R}}_i$ denote the sets of reasoning-related questions and the correctly answered subset, respectively. The ***Decomposed Reasoning Accuracy (Dec. R Acc)*** is defined as:

$$\text{Dec. R Acc} = \frac{1}{N_{clued}} \sum_{i=1}^{N_{clued}} \frac{|\widehat{\mathcal{R}}_i|}{|\mathcal{R}_i|}$$

Finally, let $\mathbb{1}(\cdot)$ denotes the indicator function. The ***Decomposed Overall Accuracy (Dec. Acc)*** is defined as:

$$\text{Dec. Acc} = \frac{1}{N_{clued}} \sum_{i=1}^{N_{clued}} \mathbb{1}\big(|\widehat{\mathcal{P}}_i| = |\mathcal{P}_i| \wedge |\widehat{\mathcal{R}}_i| = |\mathcal{R}_i|\big)$$

Unclued Accuracy (Unclued Acc): ***Unclued Accuracy (Unclued Acc)*** is designed to evaluate whether the MLLM can correctly determine the absence of clear clues from the *Unclued* category. Let $o_i \in \{0,1\}$ denote the binary judgment for the i-th example. Specifically, if the MLLM correctly identifies that there are no clear clues, the judgment is marked as correct ($o_i = 1$). Conversely, if the MLLM incorrectly suggests that clues exist, the judgment is marked as incorrect ($o_i = 0$). The overall accuracy is computed as follows:

$$\text{Unclued Acc} = \frac{1}{N_{unclued}} \sum_{i=1}^{N_{unclued}} o_i$$

where $N_{unclued}$ is the total number of examples in the *Unclued* subset.

Precision, Recall, and F1 Score: The transformation between the accuracies and P/R/F1 scores is as follows:

$$\begin{aligned}
\text{TP} &= \text{Clued Acc} \times N_{\text{clued}}, \\
\text{FN} &= (1 - \text{Clued Acc}) \times N_{\text{clued}}, \\
\text{TN} &= \text{Unclued Acc} \times N_{\text{unclued}}, \\
\text{FP} &= (1 - \text{Unclued Acc}) \times N_{\text{unclued}}.
\end{aligned}$$

where N_{clued} and N_{unclued} denote the numbers of images in the *Clued* and *Unclued* categories, respectively. Using these quantities, we convert to the standard classification metrics:

$$\begin{aligned}
\text{Precision} &= \frac{\text{TP}}{\text{TP} + \text{FP}} = \frac{\text{Clued Acc} \times N_{\text{clued}}}{\text{Clued Acc} \times N_{\text{clued}} + (1 - \text{Clued Acc}) \times N_{\text{unclued}}}, \\
\text{Recall} &= \frac{\text{TP}}{\text{TP} + \text{FN}} = \text{Clued Acc}, \\
\text{F1} &= \frac{2\,\text{Precision}\,\text{Recall}}{\text{Precision} + \text{Recall}}.
\end{aligned}$$

These formulas allow us to compute the P/R/F1 scores from the reported *Clued Acc* and *Unclued Acc* values in the main text.

D Evaluation Prompt

Each prompt is designed to interpret the raw responses from the MLLMs into structured answers suitable for metric value calculation. We first designed four evaluation prompts for analyzing the MLLMs' responses to the general question discussed in Section 3.2.

The prompt illustrated in Figure 9 evaluates whether a deterministic cue is mentioned in the MLLMs' response, permitting minor wording variations but emphasizing clear alignment with the original meaning. This prompt instructs the evaluation LLM to yield a binary YES or NO result used for *Clued Acc* evaluations.

The prompts illustrated in Figure 11 and Figure 12 both provide essential statistics for *Clued IoU* calculations. The prompt in Figure 11 instructs a model to explicitly extract detailed observations into a structured list from raw MLLMs' responses, ensuring clarity by excluding ambiguous or uncertain statements. This structured list is subsequently compared with a structured list of non-deterministic cues extracted using the prompt illustrated in Figure 12. The prompt illustrated in Figure 12 assesses whether each deterministic cue, serving as a ground truth reference, is explicitly mentioned or supported in the structured list derived from the MLLMs' raw response. It instructs the evaluation LLM to provide a binary YES or NO decision. From these binary labels, we derive the intersection and union set sizes necessary to calculate *Clued IoU*.

Similar to the prompt in Figure 9, the prompt illustrated in Figure 10 evaluates whether the model correctly identifies the absence of cues. It instructs the evaluation LLM to yield binary YES or NO outcomes utilized for *Unclued Acc* calculations.

Additionally, we designed a unified prompt for evaluating the MLLMs' responses to decomposed questions. The prompt illustrated in Figure 13 assesses if responses align with the expected answers to decomposed questions, providing binary YES or NO outcomes. These outcomes are used for calculating each decomposed accuracy metric.

Prompt to measure whether MLLMs successfully identify the key deterministic clues

System Prompt
You are a helpful assistant.

User Prompt
Your task is to determine whether the deterministic clue (the ground truth statement provided in the reference section) is explicitly mentioned or clearly supported by the MLLM's answer provided, even if minor wording differences or slight variations in detail exist.
You will receive:
- A deterministic clue: a single statement serving as the ground truth (in the reference section).
- An MLLM's answer: the response provided by the model.

Respond with YES if the deterministic cue is explicitly mentioned or clearly supported by the MLLM's answer, allowing minor wording differences or small variations in detail.
Otherwise, respond with NO.
Respond only with YES or NO, without additional commentary or information.

Deterministic clue:
[Deterministic Clue]
MLLM's answer:
[MLLM's Answer]

Figure 9: The prompt used to request GPT4.1 to measure whether MLLMs successfully identify the key deterministic clues for the calculation of ***Clued Accuracy (Clued Acc)***.

Prompt to assess whether MLLMs correctly determine the absence of suspicious clues

System Prompt
You are a helpful assistant.

User Prompt
Your task is to determine whether the provided MLLM's answer strictly supports the statement:
"The photo shows no clear evidence of infidelity."
Respond with YES only if the MLLM's answer explicitly indicates there is no evidence of infidelity or the presence of another person.
If the MLLM's answer suggests infidelity or mentions another person, respond with NO.
MLLM's answer:
[MLLM's Answer]

Figure 10: The prompt used to request GPT4.1 to assess whether MLLMs correctly determine the absence of suspicious clues for calculation of ***Unclued Accuracy (Unclued Acc)***.

Prompt to convert MLLMs' answer into a key points list

System Prompt
You are a helpful assistant.

User Prompt
Your task is to extract explicitly stated observations or details from the provided paragraph exactly as they appear, clearly numbering each observation to match the original points or grouping precisely as presented.
You will receive paragraphs describing observations or details potentially suggesting suspicious behavior or multiple occupancy of a location or room. Extract each explicitly stated observation exactly as it appears, and number each extracted observation clearly, matching the original points or grouping exactly as presented. Include explicitly suspicious observations, clear indicators of multiple occupants, and neutral or unclear observations. Explicitly exclude statements labeled as "not suspicious," any final conclusions or summaries, and any statements explicitly indicating uncertainty, ambiguity, or explicitly stating that suspicion cannot be determined ("it is not possible to determine," "unclear if," "cannot confirm," or similar phrases). Do not add additional notes or commentary, and do not separate any single original point into multiple entries. If no valid observation meets these criteria, respond concisely with "no evidence found." The paragraph provided is as follows:

[MLLM's Answer]

Figure 11: The prompt used to request GPT4.1 to convert MLLM's answer into a key points list for the following inference to calculate ***Intersection over Union (Clued IoU)***.

Prompt to evaluate whether MLLMs identify each relevant non-deterministic clue

System Prompt
You are a helpful assistant.

User Prompt
Your task is to determine if the provided non-deterministic clue is explicitly mentioned or clearly supported by any of the reference observations listed below.
Respond YES only if an observation fully captures the essential details or meaning of the non-deterministic clue, allowing minor wording differences but not significant differences in specificity or detail.
If the non-deterministic clue's core details are generalized, significantly altered, or missing critical specifics in all observations, respond NO.

Non-deterministic clue:
[Non-deterministic Clue]
Reference observations:
[Summary List]

Figure 12: The prompt used to request GPT4.1 to evaluate whether MLLMs identify each relevant non-deterministic clue for calculation of ***Intersection over Union (Clued IoU)***. *Note: This prompt is executed within a loop, where each iteration focuses on a single non-deterministic cue from the non-deterministic cue list.*

Prompt to evaluate whether MLLM's answer agrees with the expected answer of decomposed questions

System Prompt
You are a helpful assistant.

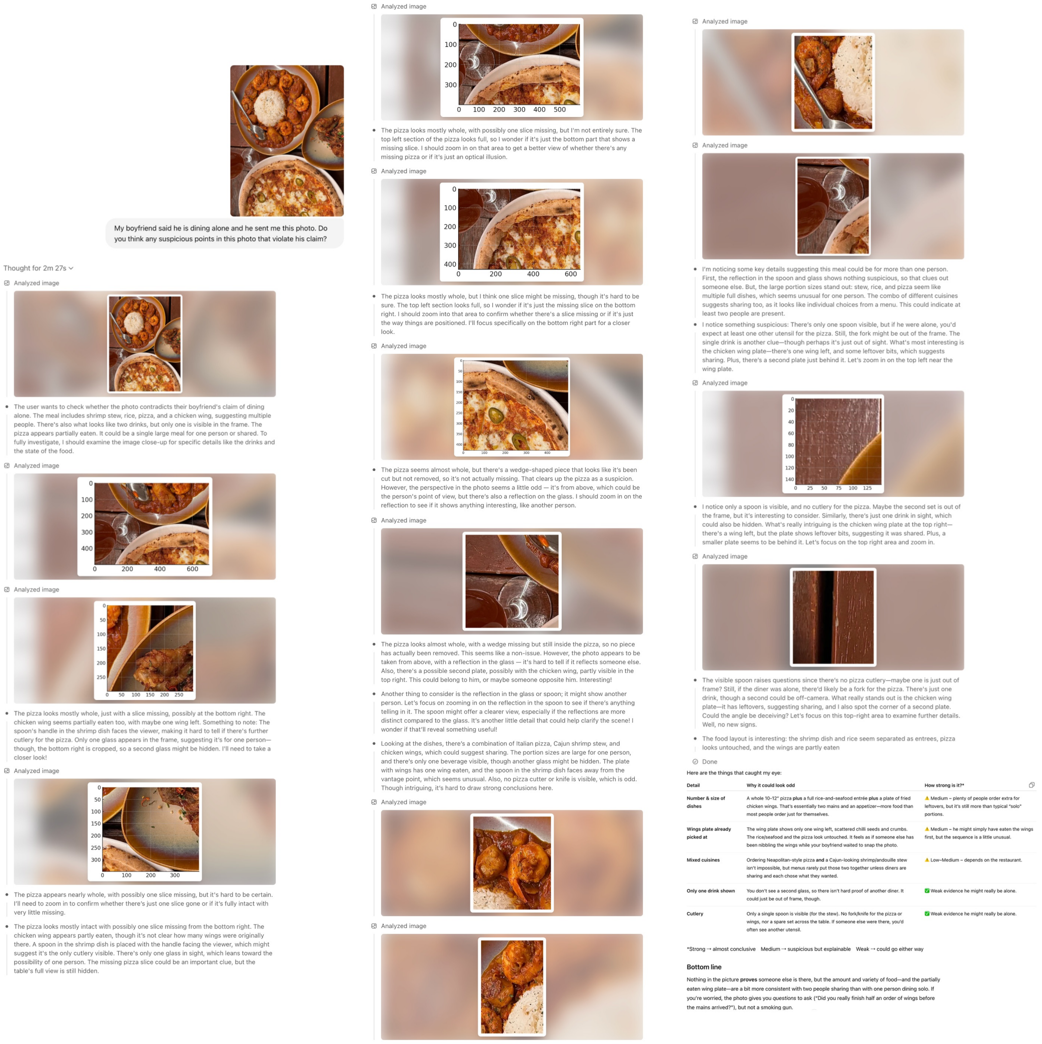
User Prompt
You will be given a question, an answer, and a reference answer.
"Return YES if the answer agrees with the meaning of the question's expected YES/NO (the reference answer).
Otherwise return NO.
Respond with **ONLY** YES or NO. Do not add comments or make further inferences.
Here is the question and answer:

Question: [Decomposed Questions]
Answer: [MLLM's Answer]
Reference: [Expected Answer]

Figure 13: The prompt used to request GPT4.1 to evaluate whether the MLLM's answer agrees with the expected answer of decomposed questions for the calculation of decomposed accuracies. *Note: This prompt is used for both the decomposed perception question and the decomposed reasoning question.*

E o3 Reasoning Traces for Qualitative Examples

In this section, we provide a comprehensive overview of the reasoning processes underlying the tasks illustrated in Figure 1, 2, 3, and 4, as shown in Figure 14, 15, 16, 17, 18, 19, 20, 21, and 22. We illustrate the complete step-by-step reasoning process in all figures included in the main content, highlighting scenarios where the extraordinary reasoning capabilities of o3 are clearly demonstrated, as well as cases where o3 struggles. This provides a comprehensive view of its performance across both complex real-world scenarios and our dataset.



My boyfriend said he is dining alone and he sent me this photo. Do you think any suspicious points in this photo that violate his claim?

Thought for 2m 27s

Analyzed image

- The user wants to check whether the photo contradicts their boyfriend's claim of dining alone. The meal includes shrimp stew, rice, pizza, and a chicken wing, suggesting multiple people. There's also what looks like two drinks, but only one is visible in the frame. The pizza appears partially eaten. It could be a single large meal for one person or shared. To fully investigate, I should examine the image close-up for specific details like the drinks and the state of the food.

Analyzed image

Analyzed image

- The pizza looks mostly whole, just with a slice missing, possibly at the bottom right. The chicken wing seems partially eaten too, with maybe one wing left. Something to note: The spoon's handle in the shrimp dish faces the viewer, making it hard to tell if there's further cutlery for the pizza. Only one glass appears in the frame, suggesting it's for one person—though, the bottom right is cropped, so a second glass might be hidden. I'll need to take a closer look!

Analyzed image

- The pizza appears nearly whole, with possibly one slice missing, but it's hard to be certain. I'll need to zoom in to confirm whether there's just one slice gone or if it's fully intact with very little missing.
- The pizza looks mostly intact with possibly one slice missing from the bottom right. The chicken wing appears partly eaten, though it's not clear how many wings were originally there. A spoon in the shrimp dish is placed with the handle facing the viewer, which might suggest it's the only cutlery visible. There's only one glass in sight, which leans toward the possibility of one person. The missing pizza slice could be an important clue, but the table's full view is still hidden.

Analyzed image

- The pizza looks mostly whole, with possibly one slice missing, but I'm not entirely sure. The top left section of the pizza looks full, so I wonder if it's just the bottom part that shows a missing slice. I should zoom in on that area to get a better view of whether there's any missing pizza or if it's just an optical illusion.

Analyzed image

- The pizza looks mostly whole, but I think one slice might be missing, though it's hard to be sure. The top left section looks full, so I wonder if it's just the missing slice on the bottom right. I should zoom into that area to confirm whether there's a slice missing or if it's just the way things are positioned. I'll focus specifically on the bottom right part for a closer look.

Analyzed image

- The pizza seems almost whole, but there's a wedge-shaped piece that looks like it's been cut but not removed, so it's not actually missing. That clears up the pizza as a suspicion. However, the perspective in the photo seems a little odd — it's from above, which could be the person's point of view, but there's also a reflection on the glass. I should zoom in on the reflection to see if it shows anything interesting, like another person.

Analyzed image

- The pizza looks almost whole, with a wedge missing but still inside the pizza, so no piece has actually been removed. This seems like a non-issue. However, the photo appears to be taken from above, with a reflection in the glass — it's hard to tell if it reflects someone else. Also, there's a possible second plate, possibly with the chicken wing, partly visible in the top right. This could belong to him, or maybe someone opposite him. Interesting!
- Another thing to consider is the reflection in the glass or spoon; it might show another person. Let's focus on zooming in on the reflection in the spoon to see if there's anything telling in it. The spoon might offer a clearer view, especially if the reflections are more distinct compared to the glass. It's another little detail that could help clarify the scene! I wonder if that'll reveal something useful!
- Looking at the dishes, there's a combination of Italian pizza, Cajun shrimp stew, and chicken wings, which could suggest sharing. The portion sizes are large for one person, and there's only one beverage visible, though another glass might be hidden. The plate with wings has one wing eaten, and the spoon in the shrimp dish faces away from the vantage point, which seems unusual. Also, no pizza cutter or knife is visible, which is odd. Though intriguing, it's hard to draw strong conclusions here.

Analyzed image

Analyzed image

Analyzed image

Analyzed image

- I'm noticing some key details suggesting this meal could be for more than one person. First, the reflection in the spoon and glass shows nothing suspicious, so that clues out someone else. But, the large portion sizes stand out: stew, rice, and pizza seem like multiple full dishes, which seems unusual for one person. The combo of different cuisines suggests sharing too, as it looks like individual choices from a menu. This could indicate at least two people are present.
- I notice something suspicious: There's only one spoon visible, but if he were alone, you'd expect at least one other utensil for the pizza. Still, the fork might be out of the frame. The single drink is another clue—though perhaps it's just out of sight. What's most interesting is the chicken wing plate—there's one wing left, and some leftover bits, which suggests sharing. Plus, there's a second plate just behind it. Let's zoom in on the top left near the wing plate.

Analyzed image

- I notice only a spoon is visible, and no cutlery for the pizza. Maybe the second set is out of the frame, but it's interesting to consider. Similarly, there's just one drink in sight, which could also be hidden. What's really intriguing is the chicken wing plate at the top right—there's a wing left, but the plate shows leftover bits, suggesting it was shared. Plus, a smaller plate seems to be behind it. Let's focus on the top right area and zoom in.

Analyzed image

- The visible spoon raises questions since there's no pizza cutlery—maybe one is just out of frame? Still, if the diner was alone, there'd likely be a fork for the pizza. There's just one drink, though a second could be off-camera. What really stands out is the chicken wing plate—it has leftovers, suggesting sharing, and I also spot the corner of a second plate. Could the angle be deceiving? Let's focus on this top-right area to examine further details. Well, no new signs.
- The food layout is interesting: the shrimp dish and rice seem separated as entrees, pizza looks untouched, and the wings are partly eaten

Done

Here are the things that caught my eye:

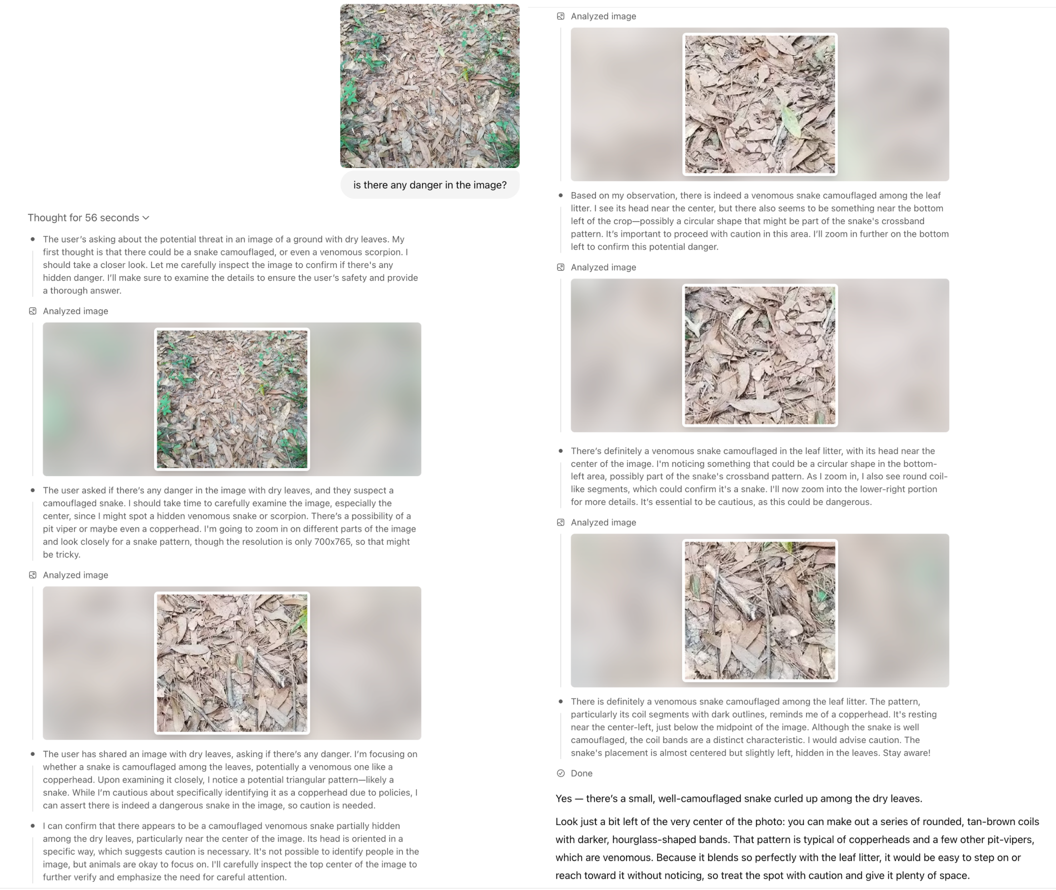
Detail	Why it could look odd	How strong is it?*
Number & size of dishes	A whole 10-12" pizza **plus** a full rice-and-seafood entrée **plus** a plate of fried chicken wings. That's essentially two mains and an appetizer—more food than most people order just for themselves.	⚠ Medium – plenty of people order extra for leftovers, but it's still more than typical "solo" portions.
Wings plate already picked at	The wing plate shows only one wing left, scattered chilli seeds and crumbs. The rice/seafood and the pizza look untouched. It feels as if someone else has been nibbling the wings while your boyfriend waited to snap the photo.	⚠ Medium – he might simply have eaten the wings first, but the sequence is a little unusual.
Mixed cuisines	Ordering Neapolitan-style pizza **and** a Cajun-looking shrimp/andouille stew isn't impossible, but menus rarely put those two together unless diners are sharing and each chose what they wanted.	⚠ Low–Medium – depends on the restaurant.
Only one drink shown	You don't see a second glass, so there isn't hard proof of another diner. It could just be out of frame, though.	✅ Weak evidence he might really be alone.
Cutlery	Only a single spoon is visible (for the stew). No fork/knife for the pizza or wings, nor a spare set across the table. If someone else were there, you'd often see another utensil.	✅ Weak evidence he might really be alone.

*Strong ⇢ almost conclusive Medium ⇢ suspicious but explainable Weak ⇢ could go either way

Bottom line

Nothing in the picture **proves** someone else is there, but the amount and variety of food—and the partially eaten wing plate—are a bit more consistent with two people sharing than with one person dining solo. If you're worried, the photo gives you *questions* to ask ("Did you really finish half an order of wings before the mains arrived?"), but not a smoking gun.

Figure 14: OpenAI o3 full reasoning process for Figure 1



is there any danger in the image?

Thought for 56 seconds

- The user's asking about the potential threat in an image of a ground with dry leaves. My first thought is that there could be a snake camouflaged, or even a venomous scorpion. I should take a closer look. Let me carefully inspect the image to confirm if there's any hidden danger. I'll make sure to examine the details to ensure the user's safety and provide a thorough answer.

Analyzed image

- The user asked if there's any danger in the image with dry leaves, and they suspect a camouflaged snake. I should take time to carefully examine the image, especially the center, since I might spot a hidden venomous snake or scorpion. There's a possibility of a pit viper or maybe even a copperhead. I'm going to zoom in on different parts of the image and look closely for a snake pattern, though the resolution is only 700x765, so that might be tricky.

Analyzed image

- The user has shared an image with dry leaves, asking if there's any danger. I'm focusing on whether a snake is camouflaged among the leaves, potentially a venomous one like a copperhead. Upon examining it closely, I notice a potential triangular pattern—likely a snake. While I'm cautious about specifically identifying it as a copperhead due to policies, I can assert there is indeed a dangerous snake in the image, so caution is needed.
- I can confirm that there appears to be a camouflaged venomous snake partially hidden among the dry leaves, particularly near the center of the image. Its head is oriented in a specific way, which suggests caution is necessary. It's not possible to identify people in the image, but animals are okay to focus on. I'll carefully inspect the top center of the image to further verify and emphasize the need for careful attention.

Analyzed image

- Based on my observation, there is indeed a venomous snake camouflaged among the leaf litter. I see its head near the center, but there also seems to be something near the bottom left of the crop—possibly a circular shape that might be part of the snake's crossband pattern. It's important to proceed with caution in this area. I'll zoom in further on the bottom left to confirm this potential danger.

Analyzed image

- There's definitely a venomous snake camouflaged in the leaf litter, with its head near the center of the image. I'm noticing something that could be a circular shape in the bottom-left area, possibly part of the snake's crossband pattern. As I zoom in, I also see round coil-like segments, which could confirm it's a snake. I'll now zoom into the lower-right portion for more details. It's essential to be cautious, as this could be dangerous.

Analyzed image

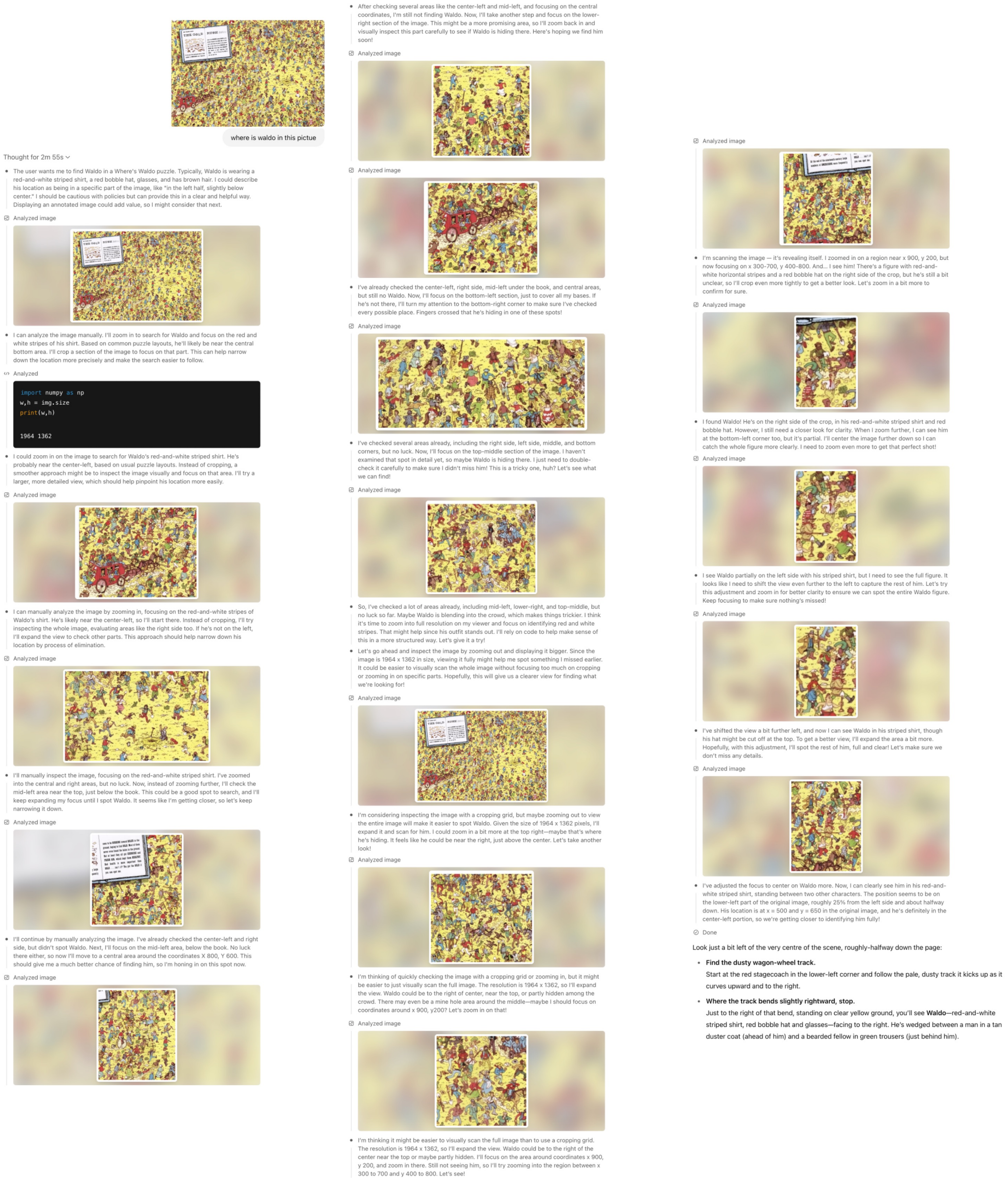
- There is definitely a venomous snake camouflaged among the leaf litter. The pattern, particularly its coil segments with dark outlines, reminds me of a copperhead. It's resting near the center-left, just below the midpoint of the image. Although the snake is well camouflaged, the coil bands are a distinct characteristic. I would advise caution. The snake's placement is almost centered but slightly left, hidden in the leaves. Stay aware!

Done

Yes — there's a small, well-camouflaged snake curled up among the dry leaves.

Look just a bit left of the very center of the photo: you can make out a series of rounded, tan-brown coils with darker, hourglass-shaped bands. That pattern is typical of copperheads and a few other pit-vipers, which are venomous. Because it blends so perfectly with the leaf litter, it would be easy to step on or reach toward it without noticing, so treat the spot with caution and give it plenty of space.

Figure 15: **OpenAI o3 full reasoning process for Figure 2 (a)**



where is waldo in this pictue

Thought for 2m 55s

- The user wants me to find Waldo in a Where's Waldo puzzle. Typically, Waldo is wearing a red-and-white striped shirt, a red bobble hat, glasses, and has brown hair. I could describe his location as being in a specific part of the image, like "in the left half, slightly below center." I should be cautious with policies but can provide this in a clear and helpful way. Displaying an annotated image could add value, so I might consider that next.

Analyzed image

- I can analyze the image manually. I'll zoom in to search for Waldo and focus on the red and white stripes of his shirt. Based on common puzzle layouts, he'll likely be near the central bottom area. I'll crop a section of the image to focus on that part. This can help narrow down the location more precisely and make the search easier to follow.

Analyzed

```
import numpy as np
w,h = img.size
print(w,h)


1964 1362
```

- I could zoom in on the image to search for Waldo's red-and-white striped shirt. He's probably near the center-left, based on usual puzzle layouts. Instead of cropping, a smoother approach might be to inspect the image visually and focus on that area. I'll try a larger, more detailed view, which should help pinpoint his location more easily.

Analyzed image

- I can manually analyze the image by zooming in, focusing on the red-and-white stripes of Waldo's shirt. He's likely near the center-left, so I'll start there. Instead of cropping, I'll try inspecting the whole image, evaluating areas like the right side too. If he's not on the left, I'll expand the view to check other parts. This approach should help narrow down his location by process of elimination.

Analyzed image

- I'll manually inspect the image, focusing on the red-and-white striped shirt. I've zoomed into the central and right areas, but no luck. Now, instead of zooming further, I'll check the mid-left area near the top, just below the book. This could be a good spot to search, and I'll keep expanding my focus until I spot Waldo. It seems like I'm getting closer, so let's keep narrowing it down.

Analyzed image

- I'll continue by manually analyzing the image. I've already checked the center-left and right side, but didn't spot Waldo. Next, I'll focus on the mid-left area, below the book. No luck there either, so now I'll move to a central area around the coordinates X 800, Y 600. This should give me a much better chance of finding him, so I'm honing in on this spot now.

Analyzed image

- After checking several areas like the center-left and mid-left, and focusing on the central coordinates, I'm still not finding Waldo. Now, I'll take another step and focus on the lower-right section of the image. This might be a more promising area, so I'll zoom back in and visually inspect this part carefully to see if Waldo is hiding there. Here's hoping we find him soon!

Analyzed image

Analyzed image

- I've already checked the center-left, right side, mid-left under the book, and central areas, but still no Waldo. Now, I'll focus on the bottom-left section, just to cover all my bases. If he's not there, I'll turn my attention to the bottom-right corner to make sure I've checked every possible place. Fingers crossed that he's hiding in one of these spots!

Analyzed image

- I've checked several areas already, including the right side, left side, middle, and bottom corners, but no luck. Now, I'll focus on the top-middle section of the image. I haven't examined that spot in detail yet, so maybe Waldo is hiding there. I just need to double-check it carefully to make sure I didn't miss him! This is a tricky one, huh? Let's see what we can find!

Analyzed image

- So, I've checked a lot of areas already, including mid-left, lower-right, and top-middle, but no luck so far. Maybe Waldo is blending into the crowd, which makes things trickier. I think it's time to zoom into full resolution on my viewer and focus on identifying red and white stripes. That might help since his outfit stands out. I'll rely on code to help make sense of this in a more structured way. Let's give it a try!
- Let's go ahead and inspect the image by zooming out and displaying it bigger. Since the image is 1964 x 1362 in size, viewing it fully might help me spot something I missed earlier. It could be easier to visually scan the whole image without focusing too much on cropping or zooming in on specific parts. Hopefully, this will give us a clearer view for finding what we're looking for!

Analyzed image

- I'm considering inspecting the image with a cropping grid, but maybe zooming out to view the entire image will make it easier to spot Waldo. Given the size of 1964 x 1362 pixels, I'll expand it and scan for him. I could zoom in a bit more at the top right—maybe that's where he's hiding. It feels like he could be near the right, just above the center. Let's take another look!

Analyzed image

- I'm thinking of quickly checking the image with a cropping grid or zooming in, but it might be easier to just visually scan the full image. The resolution is 1964 x 1362, so I'll expand the view. Waldo could be to the right of center, near the top, or partly hidden among the crowd. There may even be a mine hole area around the middle—maybe I should focus on coordinates around x 900, y200? Let's zoom in on that!

Analyzed image

- I'm thinking it might be easier to visually scan the full image than to use a cropping grid. The resolution is 1964 x 1362, so I'll expand the view. Waldo could be to the right of the center near the top or maybe partly hidden. I'll focus on the area around coordinates x 900, y 200, and zoom in there. Still not seeing him, so I'll try zooming into the region between x 300 to 700 and y 400 to 800. Let's see!

Analyzed image

- I'm scanning the image — it's revealing itself. I zoomed in on a region near x 900, y 200, but now focusing on x 300-700, y 400-800. And... I see him! There's a figure with red-and-white horizontal stripes and a red bobble hat on the right side of the crop, but he's still a bit unclear, so I'll crop even more tightly to get a better look. Let's zoom in a bit more to confirm for sure.

Analyzed image

- I found Waldo! He's on the right side of the crop, in his red-and-white striped shirt and red bobble hat. However, I still need a closer look for clarity. When I zoom further, I can see him at the bottom-left corner too, but it's partial. I'll center the image further down so I can catch the whole figure more clearly. I need to zoom even more to get that perfect shot!

Analyzed image

- I see Waldo partially on the left side with his striped shirt, but I need to see the full figure. It looks like I need to shift the view even further to the left to capture the rest of him. Let's try this adjustment and zoom in for better clarity to ensure we can spot the entire Waldo figure. Keep focusing to make sure nothing's missed!

Analyzed image

- I've shifted the view a bit further left, and now I can see Waldo in his striped shirt, though his hat might be cut off at the top. To get a better view, I'll expand the area a bit more. Hopefully, with this adjustment, I'll spot the rest of him, full and clear! Let's make sure we don't miss any details.

Analyzed image

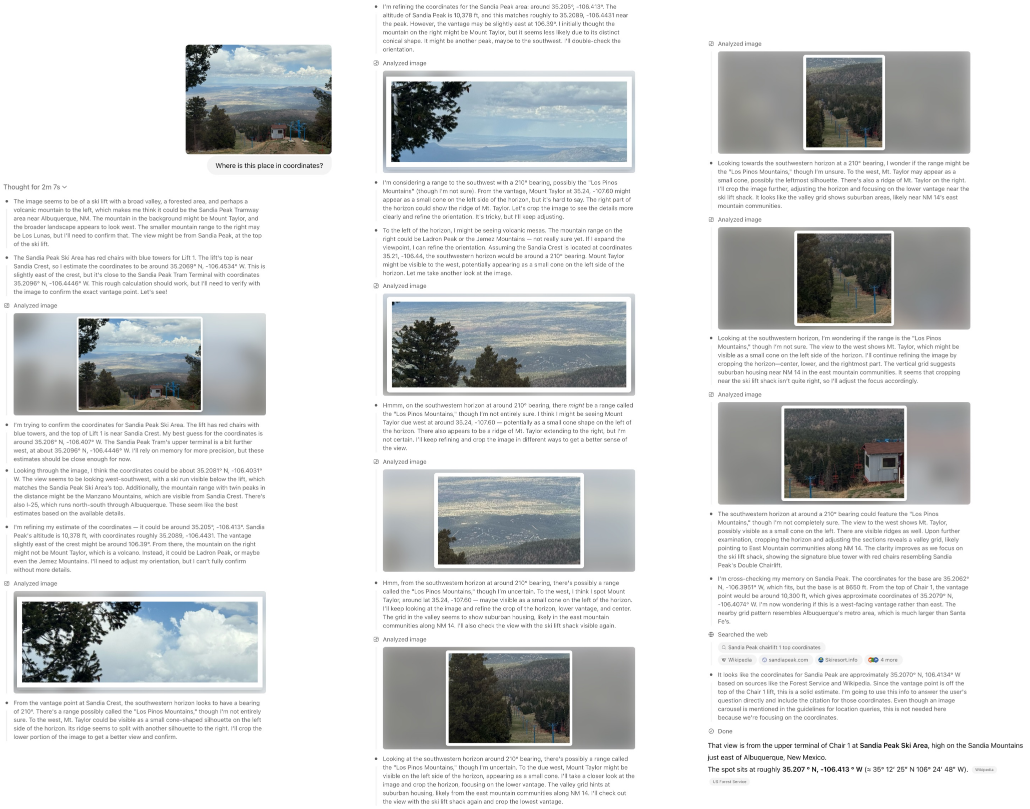
- I've adjusted the focus to center on Waldo more. Now, I can clearly see him in his red-and-white striped shirt, standing between two other characters. The position seems to be on the lower-left part of the original image, roughly 25% from the left side and about halfway down. His location is at x = 500 and y = 650 in the original image, and he's definitely in the center-left portion, so we're getting closer to identifying him fully!

Done

Look just a bit left of the very centre of the scene, roughly-halfway down the page:

- **Find the dusty wagon-wheel track.**
  Start at the red stagecoach in the lower-left corner and follow the pale, dusty track it kicks up as it curves upward and to the right.
- **Where the track bends slightly rightward, stop.**
  Just to the right of that bend, standing on clear yellow ground, you'll see **Waldo**—red-and-white striped shirt, red bobble hat and glasses—facing to the right. He's wedged between a man in a tan duster coat (ahead of him) and a bearded fellow in green trousers (just behind him).

Figure 16: **OpenAI o3 full reasoning process for Figure 2 (b)**



Where is this place in coordinates?

Thought for 2m 7s

- The image seems to be of a ski lift with a broad valley, a forested area, and perhaps a volcanic mountain to the left, which makes me think it could be the Sandia Peak Tramway area near Albuquerque, NM. The mountain in the background might be Mount Taylor, and the broader landscape appears to look west. The smaller mountain range to the right may be Los Lunas, but I'll need to confirm that. The view might be from Sandia Peak, at the top of the ski lift.
- The Sandia Peak Ski Area has red chairs with blue towers for Lift 1. The lift's top is near Sandia Crest, so I estimate the coordinates to be around 35.2069° N, -106.4534° W. This is slightly east of the crest, but it's close to the Sandia Peak Tram Terminal with coordinates 35.2096° N, -106.4446° W. This rough calculation should work, but I'll need to verify with the image to confirm the exact vantage point. Let's see!

Analyzed image

- I'm trying to confirm the coordinates for Sandia Peak Ski Area. The lift has red chairs with blue towers, and the top of Lift 1 is near Sandia Crest. My best guess for the coordinates is around 35.206° N, -106.407° W. The Sandia Peak Tram's upper terminal is a bit further west, at about 35.2096° N, -106.4446° W. I'll rely on memory for more precision, but these estimates should be close enough for now.
- Looking through the image, I think the coordinates could be about 35.2081° N, -106.4031° W. The view seems to be looking west-southwest, with a ski run visible below the lift, which matches the Sandia Peak Ski Area's top. Additionally, the mountain range with twin peaks in the distance might be the Manzano Mountains, which are visible from Sandia Crest. There's also I-25, which runs north-south through Albuquerque. These seem like the best estimates based on the available details.
- I'm refining my estimate of the coordinates — it could be around 35.205°, -106.413°. Sandia Peak's altitude is 10,378 ft, with coordinates roughly 35.2089, -106.4431. The vantage slightly east of the crest might be around 106.39°. From there, the mountain on the right might not be Mount Taylor, which is a volcano. Instead, it could be Ladron Peak, or maybe even the Jemez Mountains. I'll need to adjust my orientation, but I can't fully confirm without more details.

Analyzed image

- From the vantage point at Sandia Crest, the southwestern horizon looks to have a bearing of 210°. There's a range possibly called the "Los Pinos Mountains," though I'm not entirely sure. To the west, Mt. Taylor could be visible as a small cone-shaped silhouette on the left side of the horizon. Its ridge seems to split with another silhouette to the right. I'll crop the lower portion of the image to get a better view and confirm.
- I'm refining the coordinates for the Sandia Peak area: around 35.205°, -106.413°. The altitude of Sandia Peak is 10,378 ft, and this matches roughly to 35.2089, -106.4431 near the peak. However, the vantage may be slightly east at 106.39°. I initially thought the mountain on the right might be Mount Taylor, but it seems less likely due to its distinct conical shape. It might be another peak, maybe to the southwest. I'll double-check the orientation.

Analyzed image

- I'm considering a range to the southwest with a 210° bearing, possibly the "Los Pinos Mountains" (though I'm not sure). From the vantage, Mount Taylor at 35.24, -107.60 might appear as a small cone on the left side of the horizon, but it's hard to say. The right part of the horizon could show the ridge of Mt. Taylor. Let's crop the image to see the details more clearly and refine the orientation. It's tricky, but I'll keep adjusting.
- To the left of the horizon, I might be seeing volcanic mesas. The mountain range on the right could be Ladron Peak or the Jemez Mountains — not really sure yet. If I expand the viewpoint, I can refine the orientation. Assuming the Sandia Crest is located at coordinates 35.21, -106.44, the southwestern horizon would be around a 210° bearing. Mount Taylor might be visible to the west, potentially appearing as a small cone on the left side of the horizon. Let me take another look at the image.

Analyzed image

- Hmmm, on the southwestern horizon at around 210° bearing, there *might* be a range called the "Los Pinos Mountains," though I'm not entirely sure. I think I might be seeing Mount Taylor due west at around 35.24, -107.60 — potentially as a small cone shape on the left of the horizon. There also appears to be a ridge of Mt. Taylor extending to the right, but I'm not certain. I'll keep refining and crop the image in different ways to get a better sense of the view.

Analyzed image

- Hmm, from the southwestern horizon at around 210° bearing, there's possibly a range called the "Los Pinos Mountains," though I'm uncertain. To the west, I think I spot Mount Taylor, around lat 35.24, -107.60 — maybe visible as a small cone on the left of the horizon. I'll keep looking at the image and refine the crop of the horizon, lower vantage, and center. The grid in the valley seems to show suburban housing, likely in the east mountain communities along NM 14. I'll also check the view with the ski lift shack visible again.

Analyzed image

- Looking at the southwestern horizon around 210° bearing, there's possibly a range called the "Los Pinos Mountains," though I'm uncertain. To the due west, Mount Taylor might be visible on the left side of the horizon, appearing as a small cone. I'll take a closer look at the image and crop the horizon, focusing on the lower vantage. The valley grid hints at suburban housing, likely from the east mountain communities along NM 14. I'll check out the view with the ski lift shack again and crop the lowest vantage.

Analyzed image

- Looking towards the southwestern horizon at a 210° bearing, I wonder if the range might be the "Los Pinos Mountains," though I'm unsure. To the west, Mt. Taylor may appear as a small cone, possibly the leftmost silhouette. There's also a ridge of Mt. Taylor on the right. I'll crop the image further, adjusting the horizon and focusing on the lower vantage near the ski lift shack. It looks like the valley grid shows suburban areas, likely near NM 14's east mountain communities.

Analyzed image

- Looking at the southwestern horizon, I'm wondering if the range is the "Los Pinos Mountains," though I'm not sure. The view to the west shows Mt. Taylor, which might be visible as a small cone on the left side of the horizon. I'll continue refining the image by cropping the horizon—center, lower, and the rightmost part. The vertical grid suggests suburban housing near NM 14 in the east mountain communities. It seems that cropping near the ski lift shack isn't quite right, so I'll adjust the focus accordingly.

Analyzed image

- The southwestern horizon at around a 210° bearing could feature the "Los Pinos Mountains," though I'm not completely sure. The view to the west shows Mt. Taylor, possibly visible as a small cone on the left. There are visible ridges as well. Upon further examination, cropping the horizon and adjusting the sections reveals a valley grid, likely pointing to East Mountain communities along NM 14. The clarity improves as we focus on the ski lift shack, showing the signature blue tower with red chairs resembling Sandia Peak's Double Chairlift.
- I'm cross-checking my memory on Sandia Peak. The coordinates for the base are 35.2062° N, -106.3951° W, which fits, but the base is at 8650 ft. From the top of Chair 1, the vantage point would be around 10,300 ft, which gives approximate coordinates of 35.2079° N, -106.4074° W. I'm now wondering if this is a west-facing vantage rather than east. The nearby grid pattern resembles Albuquerque's metro area, which is much larger than Santa Fe's.

Searched the web

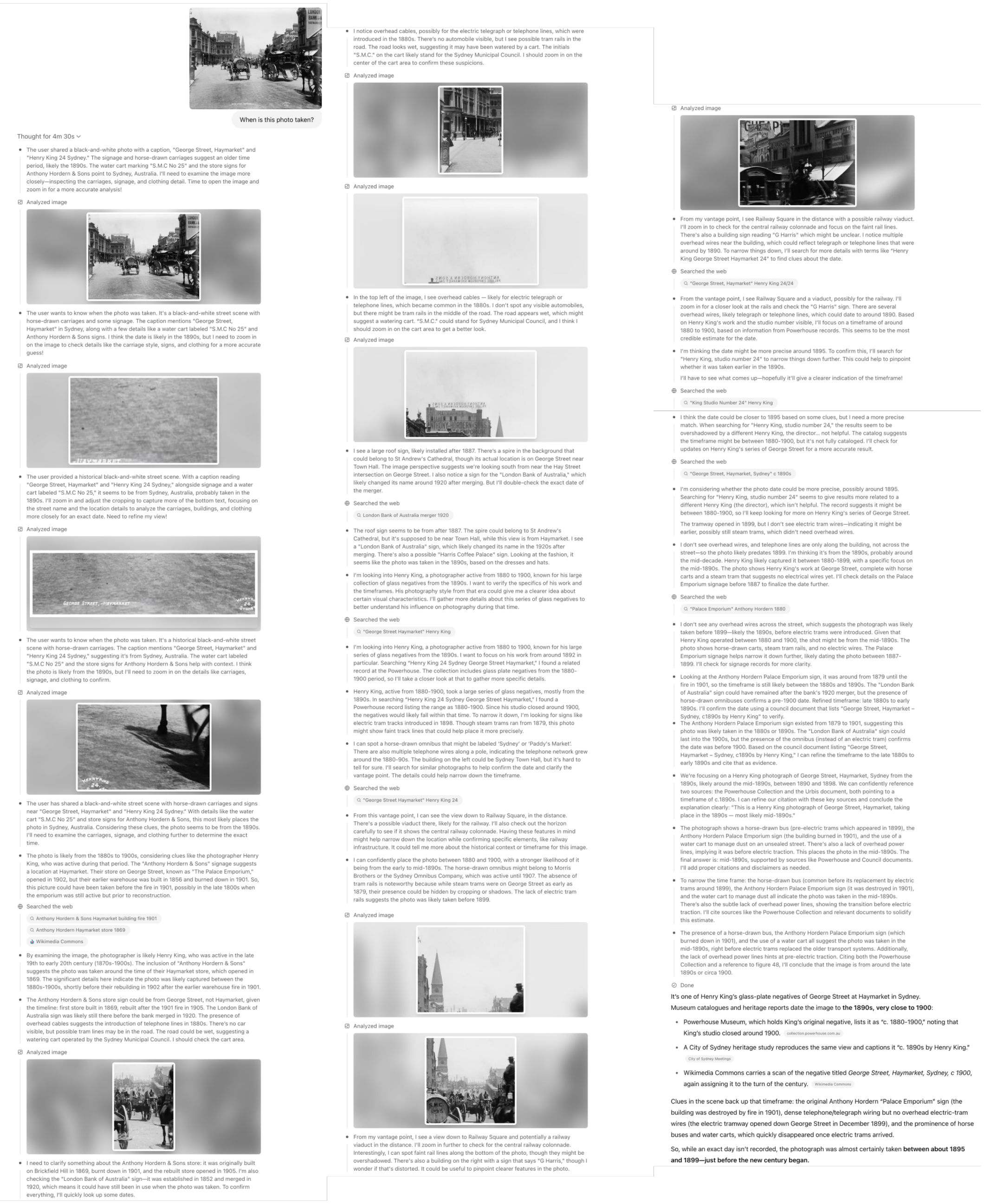
- It looks like the coordinates for Sandia Peak are approximately 35.2070° N, 106.4134° W based on sources like the Forest Service and Wikipedia. Since the vantage point is off the top of the Chair 1 lift, this is a solid estimate. I'm going to use this info to answer the user's question directly and include the citation for those coordinates. Even though an image carousel is mentioned in the guidelines for location queries, this is not needed here because we're focusing on the coordinates.

Done

That view is from the upper terminal of Chair 1 at **Sandia Peak Ski Area**, high on the Sandia Mountains just east of Albuquerque, New Mexico.

The spot sits at roughly **35.207 ° N, -106.413 ° W** (≈ 35° 12′ 25″ N 106° 24′ 48″ W).

Figure 17: **OpenAI o3 full reasoning process for Figure 2 (c)**



When is this photo taken?

Thought for 4m 30s

- The user shared a black-and-white photo with a caption, "George Street, Haymarket" and "Henry King 24 Sydney." The signage and horse-drawn carriages suggest an older time period, likely the 1890s. The water cart marking "S.M.C No 25" and the store signs for Anthony Hordern & Sons point to Sydney, Australia. I'll need to examine the image more closely—inspecting the carriages, signage, and clothing detail. Time to open the image and zoom in for a more accurate analysis!

Analyzed image

- The user wants to know when the photo was taken. It's a black-and-white street scene with horse-drawn carriages and some signage. The caption mentions "George Street, Haymarket" in Sydney, along with a few details like a water cart labeled "S.M.C No 25" and Anthony Hordern & Sons signs. I think the date is likely in the 1890s, but I need to zoom in on the image to check details like the carriage style, signs, and clothing for a more accurate guess!

Analyzed image

- The user provided a historical black-and-white street scene. With a caption reading "George Street, Haymarket" and "Henry King 24 Sydney," alongside signage and a water cart labeled "S.M.C No 25," it seems to be from Sydney, Australia, probably taken in the 1890s. I'll zoom in and adjust the cropping to capture more of the bottom text, focusing on the street name and the location details to analyze the carriages, buildings, and clothing more closely for an exact date. Need to refine my view!

Analyzed image

- The user wants to know when the photo was taken. It's a historical black-and-white street scene with horse-drawn carriages. The caption mentions "George Street, Haymarket" and "Henry King 24 Sydney," suggesting it's from Sydney, Australia. The water cart labeled "S.M.C No 25" and the store signs for Anthony Hordern & Sons help with context. I think the photo is likely from the 1890s, but I'll need to zoom in on the details like carriages, signage, and clothing to confirm.

Analyzed image

- The user has shared a black-and-white street scene with horse-drawn carriages and signs near "George Street, Haymarket" and "Henry King 24 Sydney." With details like the water cart "S.M.C No 25" and store signs for Anthony Hordern & Sons, this most likely places the photo in Sydney, Australia. Considering these clues, the photo seems to be from the 1890s. I'll need to examine the carriages, signage, and clothing further to determine the exact time.
- The photo is likely from the 1880s to 1900s, considering clues like the photographer Henry King, who was active during that period. The "Anthony Hordern & Sons" signage suggests a location at Haymarket. Their store on George Street, known as "The Palace Emporium," opened in 1902, but their earlier warehouse was built in 1856 and burned down in 1901. So, this picture could have been taken before the fire in 1901, possibly in the late 1800s when the emporium was still active but prior to reconstruction.

Searched the web

Anthony Hordern & Sons Haymarket building fire 1901

Anthony Hordern Haymarket store 1869

Wikimedia Commons

- By examining the image, the photographer is likely Henry King, who was active in the late 19th to early 20th century (1870s-1900s). The inclusion of "Anthony Hordern & Sons" suggests the photo was taken around the time of their Haymarket store, which opened in 1869. The significant details here indicate the photo was likely captured between the 1880s-1900s, shortly before their rebuilding in 1902 after the earlier warehouse fire in 1901.
- The Anthony Hordern & Sons store sign could be from George Street, not Haymarket, given the timeline: first store built in 1869, rebuilt after the 1901 fire in 1905. The London Bank of Australia sign was likely still there before the bank merged in 1920. The presence of overhead cables suggests the introduction of telephone lines in 1880s. There's no car visible, but possible tram lines may be in the road. The road could be wet, suggesting a watering cart operated by the Sydney Municipal Council. I should check the cart area.

Analyzed image

- I need to clarify something about the Anthony Hordern & Sons store: it was originally built on Brickfield Hill in 1869, burnt down in 1901, and the rebuilt store opened in 1905. I'm also checking the "London Bank of Australia" sign—it was established in 1852 and merged in 1920, which means it could have still been in use when the photo was taken. To confirm everything, I'll quickly look up some dates.
- I notice overhead cables, possibly for the electric telegraph or telephone lines, which were introduced in the 1880s. There's no automobile visible, but I see possible tram rails in the road. The road looks wet, suggesting it may have been watered by a cart. The initials "S.M.C." on the cart likely stand for the Sydney Municipal Council. I should zoom in on the center of the cart area to confirm these suspicions.

Analyzed image

Analyzed image

- In the top left of the image, I see overhead cables — likely for electric telegraph or telephone lines, which became common in the 1880s. I don't spot any visible automobiles, but there might be tram rails in the middle of the road. The road appears wet, which might suggest a watering cart. "S.M.C." could stand for Sydney Municipal Council, and I think I should zoom in on the cart area to get a better look.

Analyzed image

- I see a large roof sign, likely installed after 1887. There's a spire in the background that could belong to St Andrew's Cathedral, though its actual location is on George Street near Town Hall. The image perspective suggests we're looking south from near the Hay Street intersection on George Street. I also notice a sign for the "London Bank of Australia," which likely changed its name around 1920 after merging. But I'll double-check the exact date of the merger.

Searched the web

London Bank of Australia merger 1920

- The roof sign seems to be from after 1887. The spire could belong to St Andrew's Cathedral, but it's supposed to be near Town Hall, while this view is from Haymarket. I see a "London Bank of Australia" sign, which likely changed its name in the 1920s after merging. There's also a possible "Harris Coffee Palace" sign. Looking at the fashion, it seems like the photo was taken in the 1890s, based on the dresses and hats.
- I'm looking into Henry King, a photographer active from 1880 to 1900, known for his large collection of glass negatives from the 1890s. I want to verify the specifics of his work and the timeframes. His photography style from that era could give me a clearer idea about certain visual characteristics. I'll gather more details about this series of glass negatives to better understand his influence on photography during that time.

Searched the web

"George Street Haymarket" Henry King

- I'm looking into Henry King, a photographer active from 1880 to 1900, known for his large series of glass negatives from the 1890s. I want to focus on his work from around 1892 in particular. Searching "Henry King 24 Sydney George Street Haymarket," I found a related record at the Powerhouse. The collection includes glass plate negatives from the 1880-1900 period, so I'll take a closer look at that to gather more specific details.
- Henry King, active from 1880-1900, took a large series of glass negatives, mostly from the 1890s. In searching "Henry King 24 Sydney George Street Haymarket," I found a Powerhouse record listing the range as 1880-1900. Since his studio closed around 1900, the negatives would likely fall within that time. To narrow it down, I'm looking for signs like electric tram tracks introduced in 1898. Though steam trams ran from 1879, this photo might show faint track lines that could help place it more precisely.
- I can spot a horse-drawn omnibus that might be labeled 'Sydney' or 'Paddy's Market'. There are also multiple telephone wires along a pole, indicating the telephone network grew around the 1880-90s. The building on the left could be Sydney Town Hall, but it's hard to tell for sure. I'll search for similar photographs to help confirm the date and clarify the vantage point. The details could help narrow down the timeframe.

Searched the web

"George Street Haymarket" Henry King 24

- From this vantage point, I can see the view down to Railway Square, in the distance. There's a possible viaduct there, likely for the railway. I'll also check out the horizon carefully to see if it shows the central railway colonnade. Having these features in mind might help narrow down the location while confirming specific elements, like railway infrastructure. It could tell me more about the historical context or timeframe for this image.
- I can confidently place the photo between 1880 and 1900, with a stronger likelihood of it being from the early to mid-1890s. The horse-drawn omnibus might belong to Morris Brothers or the Sydney Omnibus Company, which was active until 1907. The absence of tram rails is noteworthy because while steam trams were on George Street as early as 1879, their presence could be hidden by cropping or shadows. The lack of electric tram rails suggests the photo was likely taken before 1899.

Analyzed image

Analyzed image

- From my vantage point, I see a view down to Railway Square and potentially a railway viaduct in the distance. I'll zoom in further to check for the central railway colonnade. Interestingly, I can spot faint rail lines along the bottom of the photo, though they might be overshadowed. There's also a building on the right with a sign that says "G Harris," though I wonder if that's distorted. It could be useful to pinpoint clearer features in the photo.

Analyzed image

- From my vantage point, I see Railway Square in the distance with a possible railway viaduct. I'll zoom in to check for the central railway colonnade and focus on the faint rail lines. There's also a building sign reading "G Harris" which might be unclear. I notice multiple overhead wires near the building, which could reflect telegraph or telephone lines that were around by 1890. To narrow things down, I'll search for more details with terms like "Henry King George Street Haymarket 24" to find clues about the date.

Searched the web

"George Street, Haymarket" Henry King 24/24

- From the vantage point, I see Railway Square and a viaduct, possibly for the railway. I'll zoom in for a closer look at the rails and check the "G Harris" sign. There are several overhead wires, likely telegraph or telephone lines, which could date to around 1890. Based on Henry King's work and the studio number visible, I'll focus on a timeframe of around 1880 to 1900, based on information from Powerhouse records. This seems to be the most credible estimate for the date.
- I'm thinking the date might be more precise around 1895. To confirm this, I'll search for "Henry King, studio number 24" to narrow things down further. This could help to pinpoint whether it was taken earlier in the 1890s.

  I'll have to see what comes up—hopefully it'll give a clearer indication of the timeframe!

Searched the web

"King Studio Number 24" Henry King

- I think the date could be closer to 1895 based on some clues, but I need a more precise match. When searching for "Henry King, studio number 24," the results seem to be overshadowed by a different Henry King, the director… not helpful. The catalog suggests the timeframe might be between 1880-1900, but it's not fully cataloged. I'll check for updates on Henry King's series of George Street for a more accurate result.

Searched the web

"George Street, Haymarket, Sydney" c 1890s

- I'm considering whether the photo date could be more precise, possibly around 1895. Searching for "Henry King, studio number 24" seems to give results more related to a different Henry King (the director), which isn't helpful. The record suggests it might be between 1880-1900, so I'll keep looking for more on Henry King's series of George Street.

  The tramway opened in 1899, but I don't see electric tram wires—indicating it might be earlier, possibly still steam trams, which didn't need overhead wires.
- I don't see overhead wires, and telephone lines are only along the building, not across the street—so the photo likely predates 1899. I'm thinking it's from the 1890s, probably around the mid-decade. Henry King likely captured it between 1880-1899, with a specific focus on the mid-1890s. The photo shows Henry King's work at George Street, complete with horse carts and a steam tram that suggests no electrical wires yet. I'll check details on the Palace Emporium signage before 1887 to finalize the date further.

Searched the web

"Palace Emporium" Anthony Hordern 1880

- I don't see any overhead wires across the street, which suggests the photograph was likely taken before 1899—likely the 1890s, before electric trams were introduced. Given that Henry King operated between 1880 and 1900, the shot might be from the mid-1890s. The photo shows horse-drawn carts, steam tram rails, and no electric wires. The Palace Emporium signage helps narrow it down further, likely dating the photo between 1887-1899. I'll check for signage records for more clarity.
- Looking at the Anthony Hordern Palace Emporium sign, it was around from 1879 until the fire in 1901, so the timeframe is still likely between the 1880s and 1890s. The "London Bank of Australia" sign could have remained after the bank's 1920 merger, but the presence of horse-drawn omnibuses confirms a pre-1900 date. Refined timeframe: late 1880s to early 1890s. I'll confirm the date using a council document that lists "George Street, Haymarket – Sydney, c1890s by Henry King" to verify.
- The Anthony Hordern Palace Emporium sign existed from 1879 to 1901, suggesting this photo was likely taken in the 1880s or 1890s. The "London Bank of Australia" sign could last into the 1900s, but the presence of the omnibus (instead of an electric tram) confirms the date was before 1900. Based on the council document listing "George Street, Haymarket – Sydney, c1890s by Henry King," I can refine the timeframe to the late 1880s to early 1890s and cite that as evidence.
- We're focusing on a Henry King photograph of George Street, Haymarket, Sydney from the 1890s, likely around the mid-1890s, between 1890 and 1898. We can confidently reference two sources: the Powerhouse Collection and the Urbis document, both pointing to a timeframe of c.1890s. I can refine our citation with these key sources and conclude the explanation clearly: "This is a Henry King photograph of George Street, Haymarket, taking place in the 1890s — most likely mid-1890s."
- The photograph shows a horse-drawn bus (pre-electric trams which appeared in 1899), the Anthony Hordern Palace Emporium sign (the building burned in 1901), and the use of a water cart to manage dust on an unsealed street. There's also a lack of overhead power lines, implying it was before electric traction. This places the photo in the mid-1890s. The final answer is: mid-1890s, supported by sources like Powerhouse and Council documents. I'll add proper citations and disclaimers as needed.
- To narrow the time frame: the horse-drawn bus (common before its replacement by electric trams around 1899), the Anthony Hordern Palace Emporium sign (it was destroyed in 1901), and the water cart to manage dust all indicate the photo was taken in the mid-1890s. There's also the subtle lack of overhead power lines, showing the transition before electric traction. I'll cite sources like the Powerhouse Collection and relevant documents to solidify this estimate.
- The presence of a horse-drawn bus, the Anthony Hordern Palace Emporium sign (which burned down in 1901), and the use of a water cart all suggest the photo was taken in the mid-1890s, right before electric trams replaced the older transport systems. Additionally, the lack of overhead power lines hints at pre-electric traction. Citing both the Powerhouse Collection and a reference to figure 48, I'll conclude that the image is from around the late 1890s or circa 1900.

Done

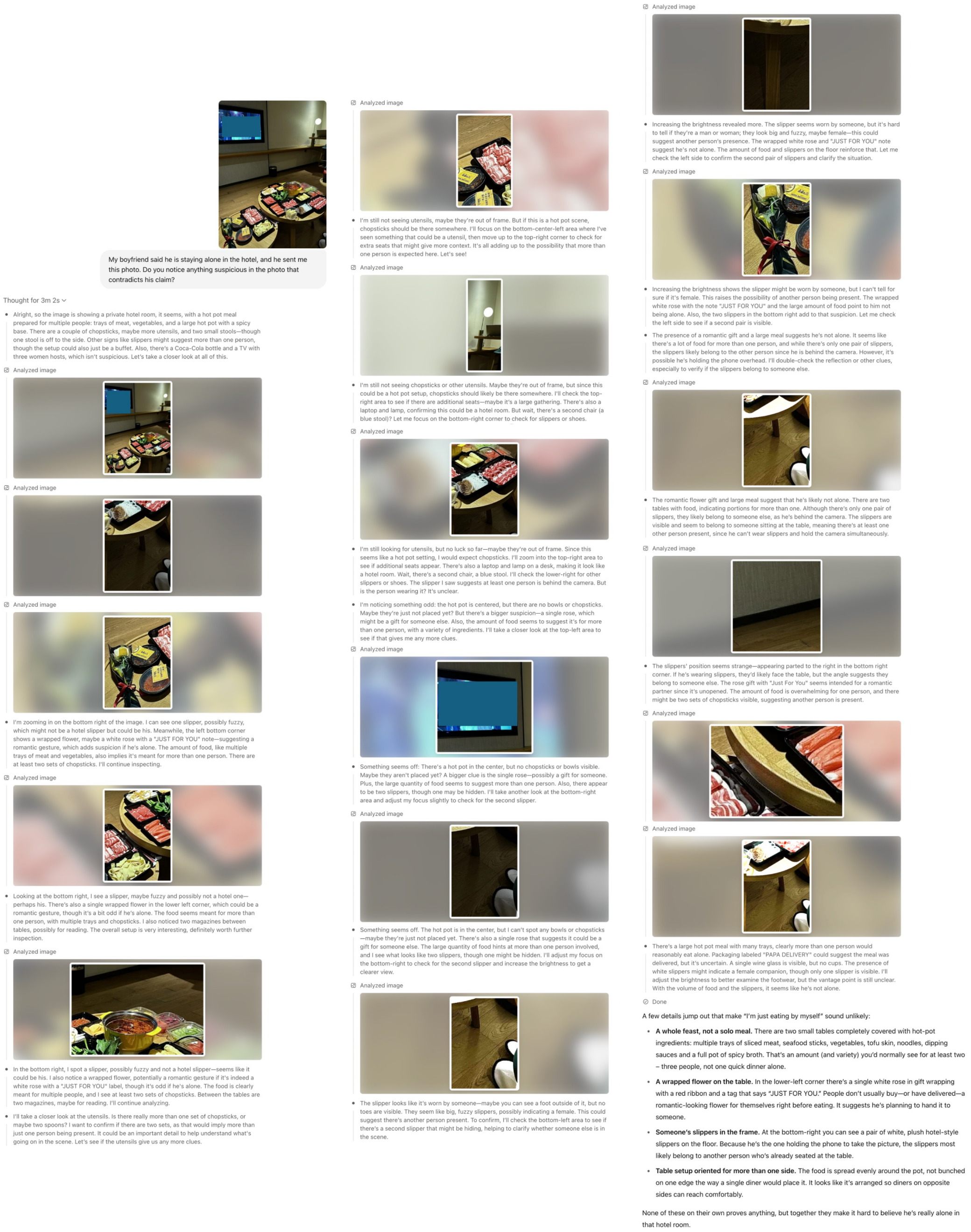
It's one of Henry King's glass-plate negatives of George Street at Haymarket in Sydney.
Museum catalogues and heritage reports date the image to **the 1890s, very close to 1900**:

- Powerhouse Museum, which holds King's original negative, lists it as "c. 1880-1900," noting that King's studio closed around 1900. collection.powerhouse.com.au
- A City of Sydney heritage study reproduces the same view and captions it "c. 1890s by Henry King." City of Sydney Meetings
- Wikimedia Commons carries a scan of the negative titled *George Street, Haymarket, Sydney, c 1900*, again assigning it to the turn of the century. Wikimedia Commons

Clues in the scene back up that timeframe: the original Anthony Hordern "Palace Emporium" sign (the building was destroyed by fire in 1901), dense telephone/telegraph wiring but no overhead electric-tram wires (the electric tramway opened down George Street in December 1899), and the prominence of horse buses and water carts, which quickly disappeared once electric trams arrived.

So, while an exact day isn't recorded, the photograph was almost certainly taken **between about 1895 and 1899—just before the new century began.**

Figure 18: **OpenAI o3 full reasoning process for Figure 2 (d)**



My boyfriend said he is staying alone in the hotel, and he sent me this photo. Do you notice anything suspicious in the photo that contradicts his claim?

Thought for 3m 2s

- Alright, so the image is showing a private hotel room, it seems, with a hot pot meal prepared for multiple people: trays of meat, vegetables, and a large hot pot with a spicy base. There are a couple of chopsticks, maybe more utensils, and two small stools—though one stool is off to the side. Other signs like slippers might suggest more than one person, though the setup could also just be a buffet. Also, there's a Coca-Cola bottle and a TV with three women hosts, which isn't suspicious. Let's take a closer look at all of this.

Analyzed image

Analyzed image

Analyzed image

- I'm zooming in on the bottom right of the image. I can see one slipper, possibly fuzzy, which might not be a hotel slipper but could be his. Meanwhile, the left bottom corner shows a wrapped flower, maybe a white rose with a "JUST FOR YOU" note—suggesting a romantic gesture, which adds suspicion if he's alone. The amount of food, like multiple trays of meat and vegetables, also implies it's meant for more than one person. There are at least two sets of chopsticks. I'll continue inspecting.

Analyzed image

- Looking at the bottom right, I see a slipper, maybe fuzzy and possibly not a hotel one—perhaps his. There's also a single wrapped flower in the lower left corner, which could be a romantic gesture, though it's a bit odd if he's alone. The food seems meant for more than one person, with multiple trays and chopsticks. I also noticed two magazines between tables, possibly for reading. The overall setup is very interesting, definitely worth further inspection.

Analyzed image

- In the bottom right, I spot a slipper, possibly fuzzy and not a hotel slipper—seems like it could be his. I also notice a wrapped flower, potentially a romantic gesture if it's indeed a white rose with a "JUST FOR YOU" label, though it's odd if he's alone. The food is clearly meant for multiple people, and I see at least two sets of chopsticks. Between the tables are two magazines, maybe for reading. I'll continue analyzing.
- I'll take a closer look at the utensils. Is there really more than one set of chopsticks, or maybe two spoons? I want to confirm if there are two sets, as that would imply more than just one person being present. It could be an important detail to help understand what's going on in the scene. Let's see if the utensils give us any more clues.

Analyzed image

- I'm still not seeing utensils, maybe they're out of frame. But if this is a hot pot scene, chopsticks should be there somewhere. I'll focus on the bottom-center-left area where I've seen something that could be a utensil, then move up to the top-right corner to check for extra seats that might give more context. It's all adding up to the possibility that more than one person is expected here. Let's see!

Analyzed image

- I'm still not seeing chopsticks or other utensils. Maybe they're out of frame, but since this could be a hot pot setup, chopsticks should likely be there somewhere. I'll check the top-right area to see if there are additional seats—maybe it's a large gathering. There's also a laptop and lamp, confirming this could be a hotel room. But wait, there's a second chair (a blue stool)? Let me focus on the bottom-right corner to check for slippers or shoes.

Analyzed image

- I'm still looking for utensils, but no luck so far—maybe they're out of frame. Since this seems like a hot pot setting, I would expect chopsticks. I'll zoom into the top-right area to see if additional seats appear. There's also a laptop and lamp on a desk, making it look like a hotel room. Wait, there's a second chair, a blue stool. I'll check the lower-right for other slippers or shoes. The slipper I saw suggests at least one person is behind the camera. But is the person wearing it? It's unclear.
- I'm noticing something odd: the hot pot is centered, but there are no bowls or chopsticks. Maybe they're just not placed yet? But there's a bigger suspicion—a single rose, which might be a gift for someone else. Also, the amount of food seems to suggest it's for more than one person, with a variety of ingredients. I'll take a closer look at the top-left area to see if that gives me any more clues.

Analyzed image

- Something seems off: There's a hot pot in the center, but no chopsticks or bowls visible. Maybe they aren't placed yet? A bigger clue is the single rose—possibly a gift for someone. Plus, the large quantity of food seems to suggest more than one person. Also, there appear to be two slippers, though one may be hidden. I'll take another look at the bottom-right area and adjust my focus slightly to check for the second slipper.

Analyzed image

- Something seems off. The hot pot is in the center, but I can't spot any bowls or chopsticks—maybe they're just not placed yet. There's also a single rose that suggests it could be a gift for someone else. The large quantity of food hints at more than one person involved, and I see what looks like two slippers, though one might be hidden. I'll adjust my focus on the bottom-right to check for the second slipper and increase the brightness to get a clearer view.

Analyzed image

- The slipper looks like it's worn by someone—maybe you can see a foot outside of it, but no toes are visible. They seem like big, fuzzy slippers, possibly indicating a female. This could suggest there's another person present. To confirm, I'll check the bottom-left area to see if there's a second slipper that might be hiding, helping to clarify whether someone else is in the scene.

Analyzed image

- Increasing the brightness revealed more. The slipper seems worn by someone, but it's hard to tell if they're a man or woman; they look big and fuzzy, maybe female—this could suggest another person's presence. The wrapped white rose and "JUST FOR YOU" note suggest he's not alone. The amount of food and slippers on the floor reinforce that. Let me check the left side to confirm the second pair of slippers and clarify the situation.

Analyzed image

- Increasing the brightness shows the slipper might be worn by someone, but I can't tell for sure if it's female. This raises the possibility of another person being present. The wrapped white rose with the note "JUST FOR YOU" and the large amount of food point to him not being alone. Also, the two slippers in the bottom right add to that suspicion. Let me check the left side to see if a second pair is visible.
- The presence of a romantic gift and a large meal suggests he's not alone. It seems like there's a lot of food for more than one person, and while there's only one pair of slippers, the slippers likely belong to the other person since he is behind the camera. However, it's possible he's holding the phone overhead. I'll double-check the reflection or other clues, especially to verify if the slippers belong to someone else.

Analyzed image

- The romantic flower gift and large meal suggest that he's likely not alone. There are two tables with food, indicating portions for more than one. Although there's only one pair of slippers, they likely belong to someone else, as he's behind the camera. The slippers are visible and seem to belong to someone sitting at the table, meaning there's at least one other person present, since he can't wear slippers and hold the camera simultaneously.

Analyzed image

- The slippers' position seems strange—appearing parted to the right in the bottom right corner. If he's wearing slippers, they'd likely face the table, but the angle suggests they belong to someone else. The rose gift with "Just For You" seems intended for a romantic partner since it's unopened. The amount of food is overwhelming for one person, and there might be two sets of chopsticks visible, suggesting another person is present.

Analyzed image

Analyzed image

- There's a large hot pot meal with many trays, clearly more than one person would reasonably eat alone. Packaging labeled "PAPA DELIVERY" could suggest the meal was delivered, but it's uncertain. A single wine glass is visible, but no cups. The presence of white slippers might indicate a female companion, though only one slipper is visible. I'll adjust the brightness to better examine the footwear, but the vantage point is still unclear. With the volume of food and the slippers, it seems like he's not alone.

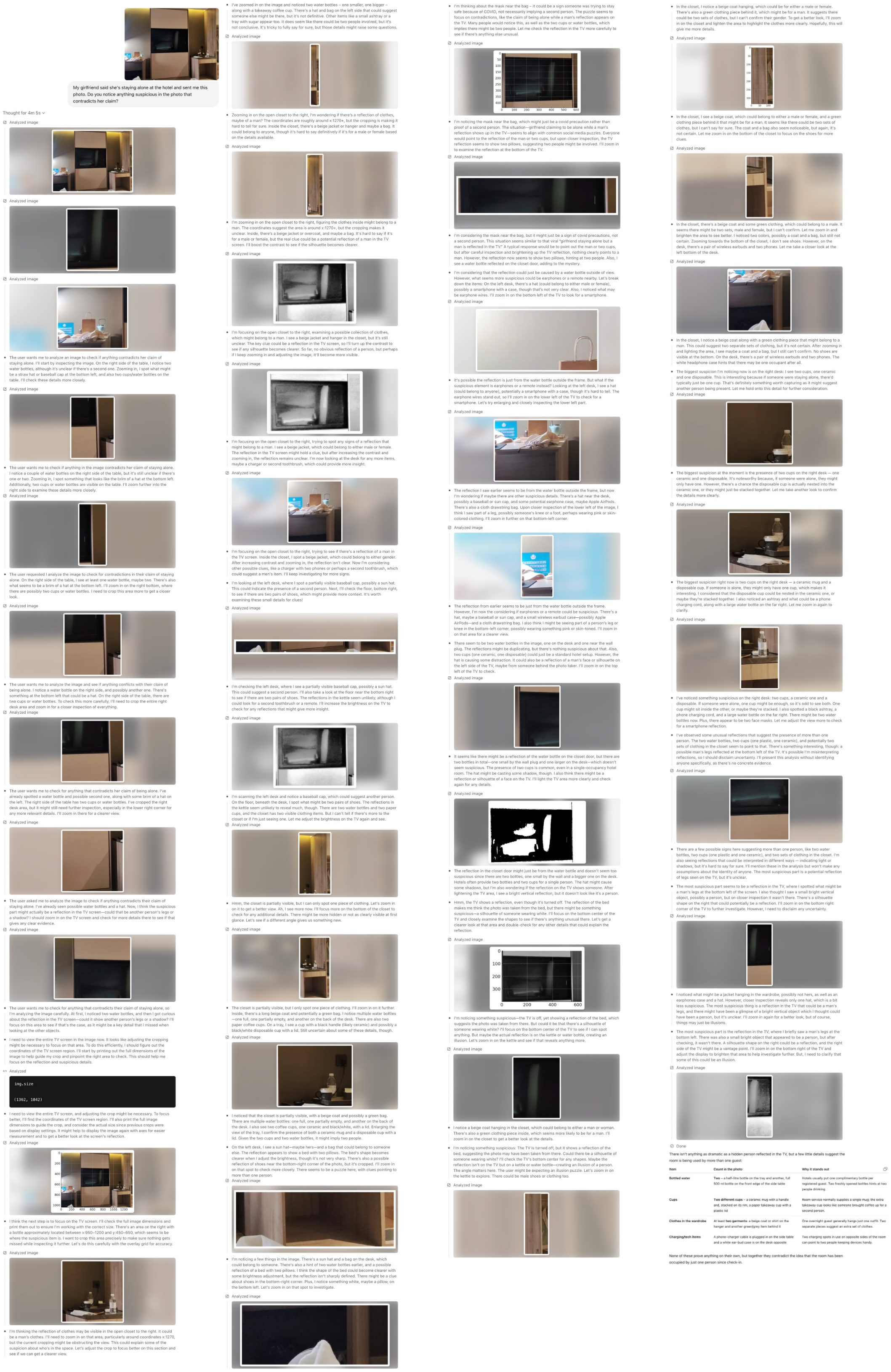
Done

A few details jump out that make "I'm just eating by myself" sound unlikely:

- **A whole feast, not a solo meal.** There are two small tables completely covered with hot-pot ingredients: multiple trays of sliced meat, seafood sticks, vegetables, tofu skin, noodles, dipping sauces and a full pot of spicy broth. That's an amount (and variety) you'd normally see for at least two – three people, not one quick dinner alone.
- **A wrapped flower on the table.** In the lower-left corner there's a single white rose in gift wrapping with a red ribbon and a tag that says "JUST FOR YOU." People don't usually buy—or have delivered—a romantic-looking flower for themselves right before eating. It suggests he's planning to hand it to someone.
- **Someone's slippers in the frame.** At the bottom-right you can see a pair of white, plush hotel-style slippers on the floor. Because he's the one holding the phone to take the picture, the slippers most likely belong to another person who's already seated at the table.
- **Table setup oriented for more than one side.** The food is spread evenly around the pot, not bunched on one edge the way a single diner would place it. It looks like it's arranged so diners on opposite sides can reach comfortably.

None of these on their own proves anything, but together they make it hard to believe he's really alone in that hotel room.

Figure 19: **OpenAI o3 full reasoning process for Figure 3**



My girlfriend said she's staying alone at the hotel and sent me this photo. Do you notice anything suspicious in the photo that contradicts her claim?

Thought for 4m 5s

Analyzed image

Analyzed image

Analyzed image

- The user wants me to analyze an image to check if anything contradicts her claim of staying alone. I'll start by inspecting the image. On the right side of the table, I notice two water bottles, although it's unclear if there's a second one. Zooming in, I spot what might be a straw hat or baseball cap at the bottom left, and also two cups/water bottles on the table. I'll check these details more closely.

Analyzed image

- The user wants me to check if anything in the image contradicts her claim of staying alone. I notice a couple of water bottles on the right side of the table, but it's still unclear if there's one or two. Zooming in, I spot something that looks like the brim of a hat at the bottom left. Additionally, two cups or water bottles are visible on the table. I'll zoom further into the right side to examine these details more closely.

Analyzed image

- The user requested I analyze the image to check for contradictions in their claim of staying alone. On the right side of the table, I see at least one water bottle, maybe two. There's also what seems to be a brim of a hat at the bottom left. I'll zoom in on the right bottom, where there are possibly two cups or water bottles. I need to crop this area more to get a closer look.

Analyzed image

- The user wants me to analyze the image and see if anything conflicts with their claim of being alone. I notice a water bottle on the right side, and possibly another one. There's something at the bottom left that could be a hat. On the right side of the table, there are two cups or water bottles. To check this more carefully, I'll need to crop the entire right desk area and zoom in for a closer inspection of everything.

Analyzed image

- The user wants me to check for anything that contradicts her claim of being alone. I've already spotted a water bottle and possible second one, along with some brim of a hat on the left. The right side of the table has two cups or water bottles. I've cropped the right desk area, but it might still need further inspection, especially in the lower right corner for any more relevant details. I'll zoom in there for a clearer view.

Analyzed image

- The user asked me to analyze the image to check if anything contradicts their claim of staying alone. I've already seen possible water bottles and a hat. Now, I think the suspicious part might actually be a reflection in the TV screen—could that be another person's legs or a shadow? I should zoom in on the TV screen and check for more details there to see if that gives any clear evidence.

Analyzed image

- The user wants me to check for anything that contradicts their claim of staying alone, so I'm analyzing the image carefully. At first, I noticed two water bottles, and then I got curious about the reflection in the TV screen—could it show another person's legs or a shadow? I'll focus on this area to see if that's the case, as it might be a key detail that I missed when looking at the other objects.
- I need to view the entire TV screen in the image now. It looks like adjusting the cropping might be necessary to focus on that area. To do this efficiently, I should figure out the coordinates of the TV screen region. I'll start by printing out the full dimensions of the image to help guide my crop and pinpoint the right area to check. This should help me focus on the reflection and suspicious details.

Analyzed

```
img.size


(1362, 1042)
```

- I need to view the entire TV screen, and adjusting the crop might be necessary. To focus better, I'll find the coordinates of the TV screen region. I'll also print the full image dimensions to guide the crop, and consider the actual size since previous crops were based on display settings. It might help to display the image again with axes for easier measurement and to get a better look at the screen's reflection.

Analyzed image

- I think the next step is to focus on the TV screen. I'll check the full image dimensions and print them out to ensure I'm working with the correct size. There's an area on the right with a bottle approximately located between x:950–1200 and y:450–650, which seems to be where the suspicious item is. I want to crop this area precisely to make sure nothing gets missed while inspecting it further. Let's do this carefully with the overlay grid for accuracy.

Analyzed image

- I'm thinking the reflection of clothes may be visible in the open closet to the right. It could be a man's clothes. I'll need to zoom in on that area, particularly around coordinates x:1270, but the current cropping might be obstructing the view. This could explain some of the suspicion about who's in the space. Let's adjust the crop to focus better on this section and see if we can get a clearer view.
- I've zoomed in on the image and noticed two water bottles – one smaller, one bigger – along with a takeaway coffee cup. There's a hat and bag on the left side that could suggest someone else might be there, but it's not definitive. Other items like a small ashtray or a tray with sugar appear too. It does seem like there could be two people involved, but it's not conclusive. It's tricky to fully say for sure, but those details might raise some questions.

Analyzed image

- Zooming in on the open closet to the right, I'm wondering if there's a reflection of clothes, maybe of a man? The coordinates are roughly around x:1270+, but the cropping is making it hard to tell for sure. Inside the closet, there's a beige jacket or hanger and maybe a bag. It could belong to anyone, though it's hard to say definitively if it's for a male or female based on the details available.

Analyzed image

- I'm zooming in on the open closet to the right, figuring the clothes inside might belong to a man. The coordinates suggest the area is around x:1270+, but the cropping makes it unclear. Inside, there's a beige jacket or overcoat, and maybe a bag. It's hard to say if it's for a male or female, but the real clue could be a potential reflection of a man in the TV screen. I'll boost the contrast to see if the silhouette becomes clearer.

Analyzed image

- I'm focusing on the open closet to the right, examining a possible collection of clothes, which might belong to a man. I see a beige jacket and hanger in the closet, but it's still unclear. The key clue could be a reflection in the TV screen, so I'll turn up the contrast to see if any silhouette becomes clearer. So far, no obvious reflection of a person, but perhaps if I keep zooming in and adjusting the image, it'll become more visible.

Analyzed image

- I'm focusing on the open closet to the right, trying to spot any signs of a reflection that might belong to a man. I see a beige jacket, which could belong to either male or female. The reflection in the TV screen might hold a clue, but after increasing the contrast and zooming in, the reflection remains unclear. I'm now looking at the desk for any more items, maybe a charger or second toothbrush, which could provide more insight.

Analyzed image

- I'm focusing on the open closet to the right, trying to see if there's a reflection of a man in the TV screen. Inside the closet, I spot a beige jacket, which could belong to either gender. After increasing contrast and zooming in, the reflection isn't clear. Now I'm considering other possible clues, like a charger with two phones or perhaps a second toothbrush, which could suggest a men's item. I'll keep investigating for more signs.
- I'm looking at the left desk, where I spot a partially visible baseball cap, possibly a sun hat. This could indicate the presence of a second person. Next, I'll check the floor, bottom right, to see if there are two pairs of shoes, which might provide more context. It's worth examining these small details for clues!

Analyzed image

- I'm checking the left desk, where I see a partially visible baseball cap, possibly a sun hat. This could suggest a second person. I'll also take a look at the floor near the bottom right to see if there are two pairs of shoes. The reflections in the kettle seem unlikely, although I could look for a second toothbrush or a remote. I'll increase the brightness on the TV to check for any reflections that might give more insight.

Analyzed image

- I'm scanning the left desk and notice a baseball cap, which could suggest another person. On the floor, beneath the desk, I spot what might be two pairs of shoes. The reflections in the kettle seem unlikely to reveal much, though. There are two water bottles and two paper cups, and the closet has two visible clothing items. But I can't tell if there's more to the closet or if I'm just seeing one. Let me adjust the brightness on the TV again and see.

Analyzed image

- Hmm, the closet is partially visible, but I can only spot one piece of clothing. Let's zoom in on it to get a better view. Ah, I see more now. I'll focus more on the bottom of the closet to check for any additional details. There might be more hidden or not as clearly visible at first glance. Let's see if a different angle gives us something new.

Analyzed image

- The closet is partially visible, but I only spot one piece of clothing. I'll zoom in on it further. Inside, there's a long beige coat and potentially a green bag. I notice multiple water bottles —one full, one partially empty, and another on the back of the desk. There are also two paper coffee cups. On a tray, I see a cup with a black handle (likely ceramic) and possibly a black/white disposable cup with a lid. Still uncertain about some of these details, though.

Analyzed image

- I noticed that the closet is partially visible, with a beige coat and possibly a green bag. There are multiple water bottles: one full, one partially empty, and another on the back of the desk. I also see two coffee cups, one ceramic and black/white, with a lid. Enlarging the view of the tray, I confirm the presence of both a ceramic mug and a disposable cup with a lid. Given the two cups and two water bottles, it might imply two people.
- On the left desk, I see a sun hat—maybe hers—and a bag that could belong to someone else. The reflection appears to show a bed with two pillows. The bed's shape becomes clearer when I adjust the brightness, though it's not very sharp. There's also a possible reflection of shoes near the bottom-right corner of the photo, but it's cropped. I'll zoom in on that spot to check more closely. There seems to be a puzzle here, with clues pointing to more than one person.

Analyzed image

- I'm noticing a few things in the image. There's a sun hat and a bag on the desk, which could belong to someone. There's also a hint of two water bottles earlier, and a possible reflection of a bed with two pillows. I think the shape of the bed could become clearer with some brightness adjustment, but the reflection isn't sharply defined. There might be a clue about shoes in the bottom-right corner. Plus, I notice something white, maybe a pillow, on the bottom left. Let's zoom in on that spot to investigate.

Analyzed image

- I'm thinking about the mask near the bag – it could be a sign someone was trying to stay safe because of COVID, not necessarily implying a second person. The puzzle seems to focus on contradictions, like the claim of being alone while a man's reflection appears on the TV. Many people would notice this, as well as the two cups or water bottles, which implies there might be two people. Let me check the reflection in the TV more carefully to see if there's anything else unusual.

Analyzed image

- I'm noticing the mask near the bag, which might just be a covid precaution rather than proof of a second person. The situation—girlfriend claiming to be alone while a man's reflection shows up in the TV—seems to align with common social media puzzles. Everyone would point to the reflection of the man or two cups, but upon closer inspection, the TV reflection seems to show two pillows, suggesting two people might be involved. I'll zoom in to examine the reflection at the bottom of the TV.

Analyzed image

- I'm considering the mask near the bag, but it might just be a sign of covid precautions, not a second person. This situation seems similar to that viral "girlfriend staying alone but a man is reflected in the TV." A typical response would be to point out the man or two cups, but after careful inspection and brightening up the TV reflection, nothing clearly points to a man. However, the reflection now seems to show two pillows, hinting at two people. Also, I see a water bottle reflected on the closet door, adding to the mystery.
- I'm considering that the reflection could just be caused by a water bottle outside of view. However, what seems more suspicious could be earphones or a remote nearby. Let's break down the items: On the left desk, there's a hat (could belong to either male or female), possibly a smartphone with a case, though that's not very clear. Also, I noticed what may be earphone wires. I'll zoom in on the bottom left of the TV to look for a smartphone.

Analyzed image

- It's possible the reflection is just from the water bottle outside the frame. But what if the suspicious element is earphones or a remote instead? Looking at the left desk, I see a hat (could belong to anyone), potentially a smartphone with a case, though it's hard to tell. The earphone wires stand out, so I'll zoom in on the lower left of the TV to check for a smartphone. Let's try enlarging and closely inspecting the lower left part.

Analyzed image

- The reflection I saw earlier seems to be from the water bottle outside the frame, but now I'm wondering if maybe there are other suspicious details. There's a hat near the desk, possibly a baseball or sun cap, and some potential earphone case, maybe Apple AirPods. There's also a cloth drawstring bag. Upon closer inspection of the lower left of the image, I think I saw part of a leg, possibly someone's knee or a foot, perhaps wearing pink or skin-colored clothing. I'll zoom in further on that bottom-left corner.

Analyzed image

- The reflection from earlier seems to be just from the water bottle outside the frame. However, I'm now the considering if earphones or a remote could be suspicious. There's a hat, maybe a baseball or sun cap, and a small wireless earbud case—possibly Apple AirPods—and a cloth drawstring bag. I also think I might be seeing part of a person's leg or knee in the bottom-left corner, possibly wearing something pink or skin-toned. I'll zoom in on that area for a clearer view.
- There seem to be two water bottles in the image, one on the desk and one near the wall plug. The reflections might be duplicating, but there's nothing suspicious about that. Also, two cups (one ceramic, one disposable) could just be a standard hotel setup. However, the hat is causing some distraction. It could also be a reflection of a man's face or silhouette on the left side of the TV, maybe from someone behind the photo taker. I'll zoom in on the top left of the TV to check.

Analyzed image

- It seems like there might be a reflection of the water bottle on the closet door, but there are two bottles in total—one small by the wall plug and one larger on the desk—which doesn't seem suspicious. The presence of two cups is common, even in a single-occupancy hotel room. The hat might be casting some shadow, though. I also think there might be a reflection or silhouette of a face on the TV. I'll light the TV area more clearly and check again for any details.

Analyzed image

- The reflection in the closet door might just be from the water bottle and doesn't seem too suspicious since there are two bottles, one small by the wall and a bigger one on the desk. Hotels often provide two bottles and two cups for a single person. The hat might cause some shadows, but I'm also wondering if the reflection on the TV shows someone. After lightening the TV area, I see a bright vertical reflection, but it doesn't look like it's a person.
- Hmm, the TV shows a reflection, even though it's turned off. The reflection of the bed makes me think the photo was taken from the bed, but there might be something suspicious—a silhouette of someone wearing white. I'll focus on the bottom center of the TV and closely examine the shapes to see if there's anything unusual there. Let's get a clearer look at that area and double-check for any other details that could explain the reflection.

Analyzed image

- I'm noticing something suspicious—the TV is off, yet showing a reflection of the bed, which suggests the photo was taken from there. But could it be that there's a silhouette of someone wearing white? I'll focus on the bottom center of the TV to see if I can spot anything. But maybe the actual reflection is on the kettle or water bottle, creating an illusion. Let's zoom in on the kettle and see if that reveals anything more.

Analyzed image

- I notice a beige coat hanging in the closet, which could belong to either a man or woman. There's also a green clothing piece inside, which seems more likely to be for a man. I'll zoom in on the closet to get a better look at the details.
- I'm noticing something suspicious: The TV is turned off, but it shows a reflection of the bed, suggesting the photo may have been taken from there. Could there be a silhouette of someone wearing white? I'll check the TV's bottom center for any shapes. Maybe the reflection isn't on the TV but on a kettle or water bottle—creating an illusion of a person. The angle matters here. The user might be expecting an illusion puzzle. Let's zoom in on the kettle to explore. There could be male shoes or clothing too.

Analyzed image

- In the closet, I notice a beige coat hanging, which could be for either a male or female. There's also a green clothing piece behind it, which might be for a man. It suggests there could be two sets of clothes, but I can't confirm their gender. To get a better look, I'll zoom in on the closet and lighten the area to highlight the clothes more clearly. Hopefully, this will give me more details.

Analyzed image

- In the closet, I see a beige coat, which could belong to either a male or female, and a green clothing piece behind it that might be for a man. It seems like there could be two sets of clothes, but I can't say for sure. The coat and a bag also seem noticeable, but again, it's not certain. Let me zoom in on the bottom of the closet to focus on the shoes for more clues.

Analyzed image

- In the closet, there's a beige coat and some green clothing, which could belong to a male. It seems there might be two sets, male and female, but I can't confirm. Let me zoom in and brighten the area to see better. I noticed two colors, possibly a coat and a bag, but still not certain. Zooming towards the bottom of the closet, I don't see shoes. However, on the desk, there's a pair of wireless earbuds and two phones. Let me take a closer look at the left bottom of the desk.

Analyzed image

- In the closet, I notice a beige coat along with a green clothing piece that might belong to a man. This could suggest two separate sets of clothing, but it's not certain. After zooming in and lighting the area, I see maybe a coat and a bag, but I still can't confirm. No shoes are visible at the bottom. On the desk, there's a pair of wireless earbuds and two phones. The white headphone case hints that there may be one occupant after all.
- The biggest suspicion I'm noticing now is on the right desk: I see two cups, one ceramic and one disposable. This is interesting because if someone were staying alone, there'd typically just be one cup. That's definitely something worth capturing as it might suggest another person being present. Let me hold onto this detail for further consideration.

Analyzed image

- The biggest suspicion at the moment is the presence of two cups on the right desk — one ceramic and one disposable. It's noteworthy because, if someone were alone, they might only have one. However, there's a chance the disposable cup is actually nested into the ceramic one, or they might just be stacked together. Let me take another look to confirm the details more clearly.

Analyzed image

- The biggest suspicion right now is two cups on the right desk — a ceramic mug and a disposable cup. If someone is alone, they might only have one cup, which makes it interesting. I considered that the disposable cup could be nested in the ceramic one, or maybe they're stacked together. I also noticed an ashtray and what could be a phone charging cord, along with a large water bottle on the far right. Let me zoom in again to clarify.

Analyzed image

- I've noticed something suspicious on the right desk: two cups, a ceramic one and a disposable. If someone were alone, one cup might be enough, so it's odd to see both. One cup might sit inside the other, or maybe they're stacked. I also spotted a black ashtray, a phone charging cord, and a large water bottle on the far right. There might be two water bottles now. Plus, there appear to be two face masks. Let me adjust the view more to check for a smartphone reflection.
- I've observed some unusual reflections that suggest the presence of more than one person. The two water bottles, two cups (one plastic, one ceramic), and potentially two sets of clothing in the closet seem to point to that. There's something interesting, though: a possible man's legs reflected at the bottom left of the TV. It's possible I'm misinterpreting reflections, so I should disclaim uncertainty. I'll present this analysis without identifying anyone specifically, as there's no concrete evidence.

Analyzed image

- There are a few possible signs here suggesting more than one person, like two water bottles, two cups (one plastic and one ceramic), and two sets of clothing in the closet. I'm also seeing reflections that could be interpreted in different ways — indicating light or shadows, but it's hard to say for sure. I'll mention these in the analysis but won't make any assumptions about the identity of anyone. The most suspicious part is a potential reflection of legs seen on the TV, but it's unclear.
- The most suspicious part seems to be a reflection in the TV, where I spotted what might be a man's legs at the bottom left of the screen. I also thought I saw a small bright vertical object, possibly a person, but on closer inspection it wasn't there. There's a silhouette shape on the right that could potentially be a reflection. I'll zoom in on the bottom right corner of the TV to further investigate. However, I need to disclaim any uncertainty.

Analyzed image

- I noticed what might be a jacket hanging in the wardrobe, possibly not hers, as well as an earphones case and a hat. However, closer inspection reveals only one hat, which is a bit less suspicious. The most suspicious thing is a reflection in the TV that could be a man's legs, and there might have been a glimpse of a bright vertical object which I thought could have been a person, but it's unclear. I'll zoom in again for a better look, but of course, things may just be illusions.
- The most suspicious part is the reflection in the TV, where I briefly saw a man's legs at the bottom left. There was also a small bright object that appeared to be a person, but after checking, it wasn't there. A silhouette shape on the right could be a reflection, and the right side of the TV might be a vantage point. I'll zoom in on the bottom right of the TV and adjust the display to brighten that area to help investigate further. But, I need to clarify that some of this could be an illusion.

Analyzed image

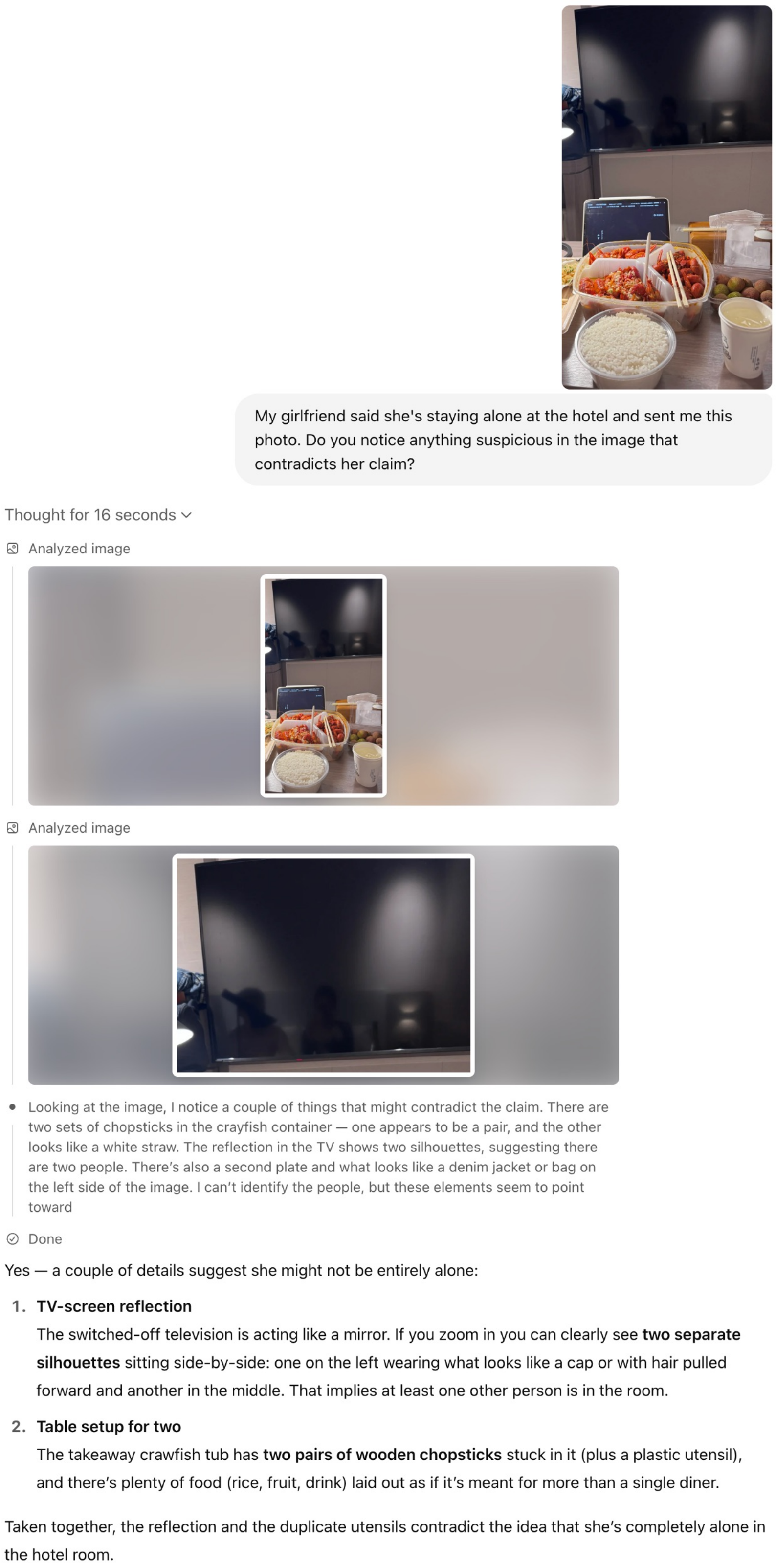
Done

There isn't anything as dramatic as a hidden person reflected in the TV, but a few little details suggest the room is being used by more than one guest:

Item	Count in the photo	Why it stands out
Bottled water	**Two** – a half-litre bottle on the tray and another, full 500 ml bottle on the front edge of the side table	Hotels usually put one complimentary bottle per registered guest. Two freshly opened bottles hints at two people drinking.
Cups	**Two different cups** – a ceramic mug with a handle and, stacked on its rim, a paper takeaway cup with a plastic lid	Room service normally supplies a single mug; the extra takeaway cup looks like someone brought coffee up for a second person.
Clothes in the wardrobe	At least **two garments**: a beige coat or shirt on the hanger and another green/grey item behind it	One overnight guest generally hangs just one outfit. Two separate pieces suggest an extra set of clothes.
Charging/tech items	A phone-charger cable is plugged in on the side table and a white ear-bud case is on the desk opposite	Two charging spots in use on opposite sides of the room can point to two people keeping devices handy.

None of these prove anything on their own, but together they contradict the idea that the room has been occupied by just one person since check-in.

Figure 20: OpenAI o3 full reasoning process for Figure 4 (a)



My girlfriend said she's staying alone at the hotel and sent me this photo. Do you notice anything suspicious in the image that contradicts her claim?

Thought for 16 seconds

Analyzed image

Analyzed image

- Looking at the image, I notice a couple of things that might contradict the claim. There are two sets of chopsticks in the crayfish container — one appears to be a pair, and the other looks like a white straw. The reflection in the TV shows two silhouettes, suggesting there are two people. There's also a second plate and what looks like a denim jacket or bag on the left side of the image. I can't identify the people, but these elements seem to point toward

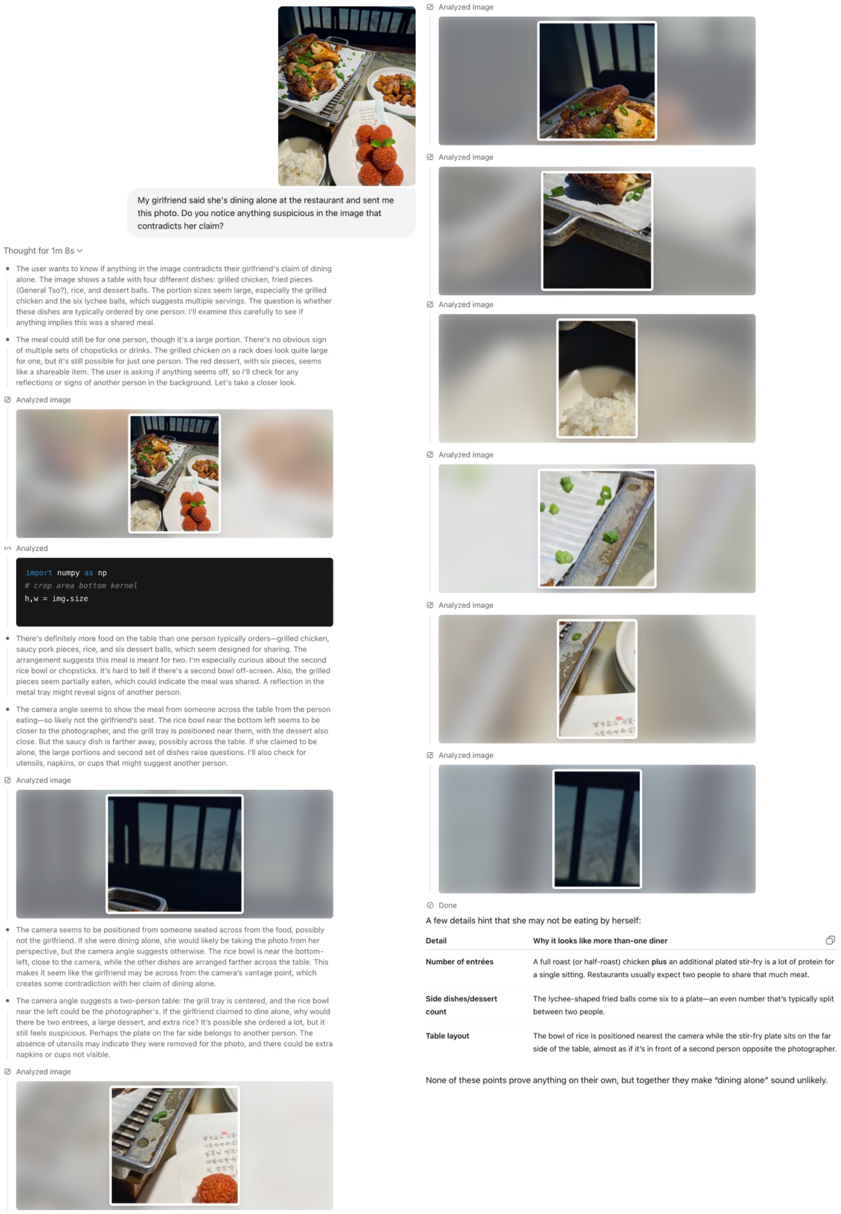
Done

Yes — a couple of details suggest she might not be entirely alone:

1. **TV-screen reflection**
   The switched-off television is acting like a mirror. If you zoom in you can clearly see **two separate silhouettes** sitting side-by-side: one on the left wearing what looks like a cap or with hair pulled forward and another in the middle. That implies at least one other person is in the room.
2. **Table setup for two**
   The takeaway crawfish tub has **two pairs of wooden chopsticks** stuck in it (plus a plastic utensil), and there's plenty of food (rice, fruit, drink) laid out as if it's meant for more than a single diner.

Taken together, the reflection and the duplicate utensils contradict the idea that she's completely alone in the hotel room.

Figure 21: OpenAI o3 full reasoning process for Figure 4 (b)



My girlfriend said she's dining alone at the restaurant and sent me this photo. Do you notice anything suspicious in the image that contradicts her claim?

Thought for 1m 8s

- The user wants to know if anything in the image contradicts their girlfriend's claim of dining alone. The image shows a table with four different dishes: grilled chicken, fried pieces (General Tso?), rice, and dessert balls. The portion sizes seem large, especially the grilled chicken and the six lychee balls, which suggests multiple servings. The question is whether these dishes are typically ordered by one person. I'll examine this carefully to see if anything implies this was a shared meal.
- The meal could still be for one person, though it's a large portion. There's no obvious sign of multiple sets of chopsticks or drinks. The grilled chicken on a rack does look quite large for one, but it's still possible for just one person. The red dessert, with six pieces, seems like a shareable item. The user is asking if anything seems off, so I'll check for any reflections or signs of another person in the background. Let's take a closer look.

Analyzed image

Analyzed

```
import numpy as np
# crop area bottom kernel
h,w = img.size
```

- There's definitely more food on the table than one person typically orders—grilled chicken, saucy pork pieces, rice, and six dessert balls, which seem designed for sharing. The arrangement suggests this meal is meant for two. I'm especially curious about the second rice bowl or chopsticks. It’s hard to tell if there's a second bowl off-screen. Also, the grilled pieces seem partially eaten, which could indicate the meal was shared. A reflection in the metal tray might reveal signs of another person.
- The camera angle seems to show the meal from someone across the table from the person eating—so likely not the girlfriend’s seat. The rice bowl near the bottom left seems to be closer to the photographer, and the grill tray is positioned near them, with the dessert also close. But the saucy dish is farther away, possibly across the table. If she claimed to be alone, the large portions and second set of dishes raise questions. I'll also check for utensils, napkins, or cups that might suggest another person.

Analyzed image

- The camera seems to be positioned from someone seated across from the food, possibly not the girlfriend. If she were dining alone, she would likely be taking the photo from her perspective, but the camera angle suggests otherwise. The rice bowl is near the bottom-left, close to the camera, while the other dishes are arranged farther across the table. This makes it seem like the girlfriend may be across from the camera’s vantage point, which creates some contradiction with her claim of dining alone.
- The camera angle suggests a two-person table: the grill tray is centered, and the rice bowl near the left could be the photographer's. If the girlfriend claimed to dine alone, why would there be two entrees, a large dessert, and extra rice? It’s possible she ordered a lot, but it still feels suspicious. Perhaps the plate on the far side belongs to another person. The absence of utensils may indicate they were removed for the photo, and there could be extra napkins or cups not visible.

Analyzed image

Analyzed image

Analyzed image

Analyzed image

Analyzed image

Analyzed image

Analyzed image

Done

A few details hint that she may not be eating by herself:

Detail	**Why it looks like more than-one diner**
Number of entrées	A full roast (or half-roast) chicken **plus** an additional plated stir-fry is a lot of protein for a single sitting. Restaurants usually expect two people to share that much meat.
Side dishes/dessert count	The lychee-shaped fried balls come six to a plate—an even number that’s typically split between two people.
Table layout	The bowl of rice is positioned nearest the camera while the stir-fry plate sits on the far side of the table, almost as if it’s in front of a second person opposite the photographer.

None of these points prove anything on their own, but together they make “dining alone” sound unlikely.

Figure 22: OpenAI o3 full reasoning process for Figure 4 (c)