When Seeing Is not Enough: Revealing the Limits of Active Reasoning in MLLMs

Hongcheng Liu*, **Pingjie Wang***, **Yuhao Wang, Siqu Ou, Yanfeng Wang, Yu Wang**†
Shanghai Jiao Tong University
{hongcheng_liu, pingjiewang, yuwangsjtu}@sjtu.edu.cn
* Equal contribution. † Corresponding author.

Abstract

Multimodal large language models (MLLMs) have shown strong capabilities across a broad range of benchmarks. However, most existing evaluations focus on passive inference, where models perform step-by-step reasoning under complete information. This setup is misaligned with real-world use, where seeing is not enough. This raises a fundamental question: ***Can MLLMs actively acquire missing evidence when faced with uncertainty?*** To bridge this gap, we require the MLLMs to actively acquire missing evidence and iteratively refine decisions under incomplete information, by selecting a target image from a candidate pool without task-specific priors. To support systematic study, we propose GUESSBENCH, a benchmark with both perception-oriented and knowledge-oriented images for evaluating active reasoning in MLLMs. We evaluate 20 superior MLLMs and find that performance on active reasoning lags far behind it on passive settings, indicating substantial room for improvement. Further analysis identifies fine-grained perception, strategic information acquisition, and timely decision-making as key challenges. Ablation studies show that perceptual enhancements benefit smaller models, whereas thinking-oriented methods provide consistent gains across model sizes. These results suggest promising directions for future research on multimodal active reasoning.

1 Introduction

Multimodal Large Language Models (MLLMs) have demonstrated impressive capabilities (Ou et al., 2025; Yin et al., 2024; Li et al., 2025), but their success is predominantly measured in passive inference settings (Wang et al., 2024; Liu et al., 2025), where they are supplied with all necessary information to make a single-pass decision (Chen et al., 2024c; Liu et al., 2026a). This benchmark-driven progress overlooks a fundamental reality: in the real world, seeing is often not enough (Zhou et al., 2025b). Many tasks, from a product recommender learning user preferences to a robot navigating an unknown space, require interaction and reasoning under incomplete information (Raza et al., 2024). This critical disparity reveals a fundamental research question: ***Can MLLMs actively acquire missing evidence when faced with uncertainty?***

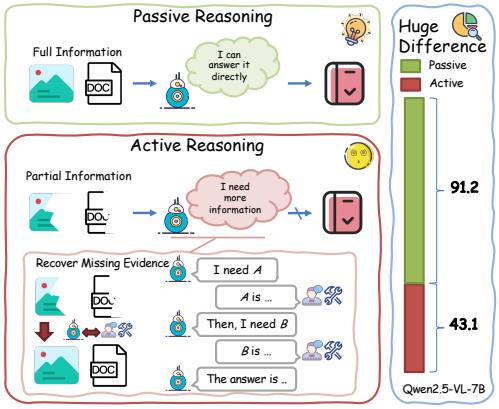


Figure 1: **Passive vs. active reasoning.** The left shows the difference between these two, and the right shows a pronounced gap between the two capabilities.

To bridge this gap, we introduce and formalize the active reasoning problem in multimodal contexts, as shown in Figure 1. Inspired by interactive games such as "Guess Who I Am" (Khasanov, 2024) and "GuessArena" (Yu et al., 2025), we define this problem as an interactive target guessing problem, where an MLLM must select a target image from a candidate set without any auxiliary information at the start. To succeed, the model must compare candidates, ask strategic questions to acquire missing evidence, and decide whether to query further or commit to an answer. Success hinges on a goal-directed cognitive cycle that requires: (1) visual abstraction (**Perceive**), to capture commonalities and subtle distinctions within the pool; and (2) knowledge-integrated reasoning (**Think**), to combine visual cues, prior world knowledge, and responses from external sources

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 27395–27422
July 2-7, 2026

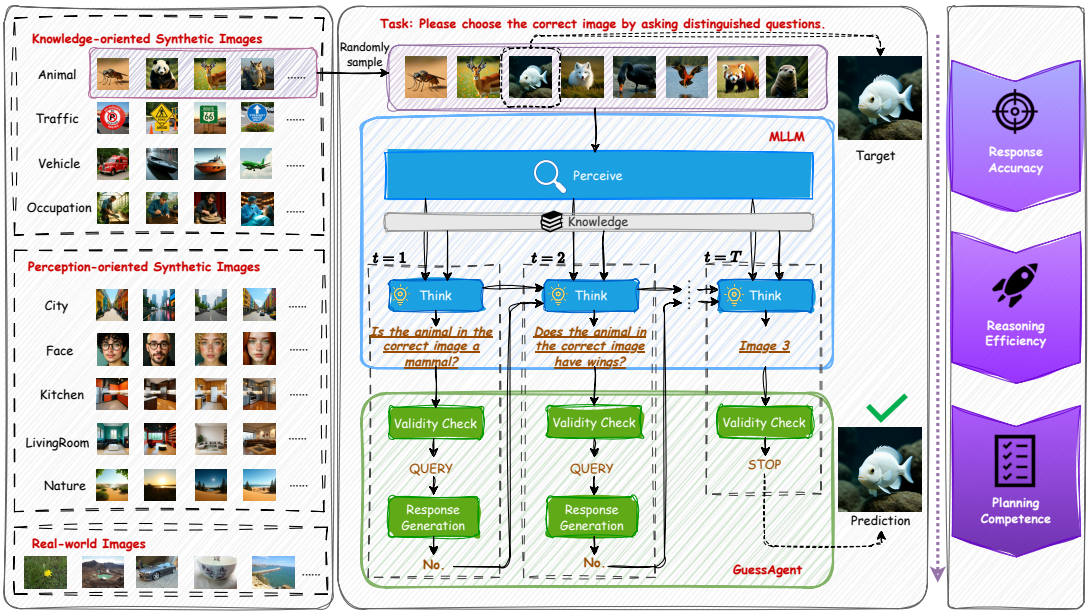


Figure 2: Overview of data distribution, interaction pipeline, and metric design.

into a strategic hypothesis.

Building on this paradigm, we propose GUESSBENCH, the first systematic framework for rigorous evaluation of active reasoning in MLLMs. To enable comprehensive and fine-grained analysis, we construct a dataset that strategically mixes real-world images with nine distinct synthetic types. These types are organized into two categories: (1) perception-oriented categories, which emphasize the ability to capture subtle visual nuances such as facial details or minor structural variations. (2) knowledge-oriented categories, which require recalling external world knowledge such as specialized occupations or complex functional roles.

Using this structured dataset, we rigorously benchmark 20 mainstream MLLMs, establishing a solid empirical foundation for this new paradigm. We reveal that state-of-the-art models that excel under passive evaluation degrade markedly on active tasks, indicating substantial headroom for improvement. We further uncover three dominant impediments to active success: limited fine-grained perception, non-strategic information acquisition, and untimely decision-making. Our analysis also identifies clear pathways toward more powerful active reasoning, highlighting the benefits of perception-oriented enhancements for smaller models and the consistent gains from thinking-oriented methods across all models.

The main contributions can be summarized as:

- **Multimodal active reasoning.** We identify and formalize the underexplored problem of multimodal active reasoning, where MLLMs must actively acquire missing evidence and iteratively refine decisions under incomplete information. This paradigm shifts the focus from passive inference to goal-directed, interactive decision cycles.
- **GUESSBENCH framework.** We propose GUESSBENCH, the first systematic framework for multimodal active reasoning. The dataset combines real-world images with nine distinct synthetic types, grouped into perception-oriented and knowledge-oriented categories, enabling comprehensive and fine-grained diagnostics of visual abstraction and knowledge-integrated reasoning.
- **Comprehensive analysis.** We empirically benchmark 20 mainstream MLLMs on GUESSBENCH and find that they fall short in active reasoning scenarios. Further analyses reveal key limiting factors and point to effective enhancement strategies.

2 GuessBench

2.1 Evaluation Framework

To push beyond passive evaluation, we introduce GUESSBENCH, a systematic framework for rigorously testing the active reasoning capabilities of MLLMs under incomplete information. We formalize this multimodal challenge as an interactive target guessing problem: given a candidate pool and no prior clues, the model must select the single target image. Success requires an iterative perceive

then think cycle. The model first (1) **Perceive** the visual candidates to extract shared and distinguishing features, then (2) **Think** by integrating these features with world knowledge to pose informative questions that acquire the missing evidence. This cycle repeats until the model commits to a confident final decision. The design of GUESSBENCH is detailed in Figure 2 and the following sections. This formulation is motivated by real-world multimodal assistants, where observations are often partial and ambiguous, and the correct decision depends on actively acquiring discriminative evidence (e.g., concise verification queries).

2.2 Data Construction

To comprehensively evaluate the active reasoning capabilities of MLLMs, we design a specialized dataset that integrates real-world and synthetic images, spanning nine domains for multi-faceted assessment. Real-world images are sampled from the IIW dataset (Garg et al., 2024), providing naturally occurring visual diversity and ambiguity. For systematic tests of the core capabilities of perceiving and thinking, we engineer synthetic images in two categories: perception-oriented (city, face, kitchen, livingroom, nature) and knowledge-oriented (animal, traffic, vehicle, occupation). The former stresses sensitivity to subtle visual nuances among similar patterns, whereas the latter requires recalling external world knowledge to resolve complex distinctions. Importantly, although synthetic, the images preserve real-world discriminative cues and knowledge demands, making performance on our synthetic suite indicative of real-world active reasoning ability. We construct these images through description generation, image synthesis, and human verification. Further details are provided in Appendix C.1.

2.3 Evaluation Procedure

Inspired by the classic game "Guess Who I Am?", GUESSBENCH formalizes the multimodal active reasoning problem as an interactive target guessing problem. Specifically, at the start of the evaluation session, B images are randomly sampled from the data pool (detailed in Section 2.2) to form an evaluation set $\mathcal{I} = \{i^1, i^2, \ldots, i^B\}$, in which a randomly selected image $i' \in \mathcal{I}$ is designed as the target image. Given the candidate pool $\mathcal{I}$, the MLLM ($\mathcal{M}$) must conduct an iterative query process by posing strategic questions (q_t) based on its own world-knowledge $\mathcal{K}_\mathcal{M}$ and obtaining facts (a_t) at each time step t. To keep queries concise and directly informative, the model is constrained to binary questions that admit Yes or No answers. We use binary QA to control interaction cost and response variance, so performance reflects informative querying and belief updating rather than open-ended answer noise. This interactive process is repeated until the model determines its final prediction $\hat{i}$, which marks the end of the session at time step T. This evaluation protocol closely mirrors real-world interactive vision assistance, where systems iteratively pose verifiable clarification queries to resolve ambiguity from partial observations. The entire decision-making process of the MLLM $\mathcal{M}$ at any step t can be formally modeled as a function of the candidate pool $\mathcal{I}$, the history of previous questions and facts $\mathcal{H}_{t-1} = \{(q_1, a_1), \ldots, (q_{t-1}, a_{t-1})\}$, and its internal knowledge $\mathcal{K}_\mathcal{M}$:

$$< \delta_t, \text{Output}_t >= \mathcal{M}(\mathcal{I}, \mathcal{H}_{t-1}, \mathcal{K}_\mathcal{M}), \tag{1}$$

where $\delta_t \in \{\text{QUERY}, \text{STOP}\}$ is the decision type, and Output_t is either the next question (if $\delta_t = \text{QUERY}$) or the final prediction $\hat{i}$ (if $\delta_t = \text{STOP}$).

To efficiently assist the MLLM in acquiring cues and to minimize incorrect or superfluous external responses, we design the GuessAgent to interpret and respond to the model's queries, which consists of a validity check and response generation components. For the validity check, we first employ Qwen3-8B (Yang et al., 2025) to determine the decision type, classifying each model output into one of <QUERY>, <STOP>, or <INVALID>. A query is labeled <INVALID> if it cannot be answered with a simple Yes or No, thereby enforcing the binary-question constraint. For response generation, GuessAgent first attempts to retrieve relevant information from the image's detailed data attributes. If the question cannot be resolved using these attributes alone, then it accesses the comprehensive image caption and image. Details for GuessAgent are provided in the Appendix F.

2.4 Evaluation Metrics

To comprehensively evaluate the active reasoning ability of MLLMs across diverse domains in terms of both accuracy and efficiency, we introduce a composite metric $\mathcal{S}$ that integrates response accuracy ($\mathcal{A}$), reasoning efficiency ($\mathcal{R}$), and planning competence ($\mathcal{P}$). We define $\mathcal{S}$ as

$$\mathcal{S} = f(\mathcal{A}, \mathcal{R}, \mathcal{P}), \tag{2}$$

Model	Real	Perception					Knowledge				Average		
		City	Face	Kitchen	Livingroom	Nature	Animal	Traffic	Vehicle	Occupation	Per.	Know.	Total
Inter3-VL-Series													
Intern3-VL-2B	0.0896	0.1009	0.1430	0.1224	0.1196	0.1118	0.0941	0.0390	0.0484	0.1162	0.1195	0.0744	0.0985
Intern3-VL-8B	0.2626	0.1547	0.2621	0.1759	0.0851	0.2237	0.3244	0.4004	0.2162	0.3263	0.1803	0.3168	0.2423
Intern3-VL-14B	0.4247	0.2522	0.2364	0.1915	0.2690	0.2643	0.4392	0.6395	0.4658	0.5082	0.2427	0.5131	0.3691
Intern3-VL-38B	0.5294	0.3000	0.3864	0.3057	0.2364	0.3787	0.4642	0.5191	0.4690	0.5360	0.3214	0.4970	0.4125
Intern3-VL-78B	0.4332	0.2056	0.2154	0.2923	0.2455	0.3086	0.4708	0.3884	0.3705	0.3825	0.2535	0.4031	0.3313
Qwen2.5-VL-Series													
Qwen2.5-VL-3B	0.1496	0.0958	0.1459	0.0865	0.0888	0.1151	0.2132	0.1924	0.2140	0.2214	0.1064	0.2103	0.1523
Qwen2.5-VL-7B	0.5087	0.1761	0.1509	0.1844	0.1910	0.2544	0.4863	0.4718	0.5233	0.5196	0.1914	0.5003	0.3466
Qwen2.5-VL-32B	0.7695	0.5284	0.5980	0.3717	0.5205	0.5274	0.6727	0.5237	0.6735	0.7087	0.5092	0.6446	0.5894
Qwen2.5-VL-72B	**0.7722**	0.5655	**0.6594**	**0.5700**	**0.6570**	0.6445	**0.7684**	**0.8177**	**0.7263**	**0.7968**	**0.6193**	**0.7773**	**0.6978**
Inter3.5-VL-Series													
Intern3.5-VL-2B	0.1226	0.1316	0.0994	0.0999	0.0998	0.0611	0.1136	0.0496	0.1053	0.1030	0.0983	0.0929	0.0986
Intern3.5-VL-4B	0.3977	0.1560	0.2015	0.1782	0.1462	0.1299	0.3158	0.2018	0.2274	0.1774	0.1623	0.2306	0.2132
Intern3.5-VL-8B	0.4048	0.1096	0.3464	0.2253	0.2093	0.1144	0.2071	0.2560	0.2047	0.2125	0.2010	0.2201	0.2290
Intern3.5-VL-14B	0.1675	0.2082	0.2085	0.2193	0.1501	0.2397	0.2193	0.1704	0.1683	0.2505	0.2052	0.2021	0.2002
Intern3.5-VL-38B	0.2304	0.2150	0.3203	0.1994	0.1320	0.1860	0.2878	0.3307	0.3058	0.3280	0.2105	0.3131	0.2535
Others													
Kimi-VL	0.2957	0.1296	0.2259	0.1570	0.1773	0.1773	0.3689	0.2115	0.2799	0.1778	0.1734	0.2595	0.2201
Kimi-VL-Thk.	0.3687	0.2567	0.2807	0.2703	0.2860	0.3569	0.3366	0.3532	0.3309	0.3131	0.2901	0.3334	0.3153
Qwen3-VL-Ins	0.6935	0.4524	0.5594	0.3849	0.4825	0.5425	0.6239	0.6753	0.6106	0.5605	0.4843	0.6176	0.5586
Qwen3-VL-Thk	0.7017	**0.6664**	0.6105	0.4478	0.6395	**0.7176**	0.6779	0.7071	0.5932	0.7724	0.6164	0.6877	0.6534
Mimo-VL	0.6844	0.4492	0.6202	0.4811	0.5607	0.6043	0.7221	0.6726	0.6017	0.7061	0.5431	0.6756	0.6102
GPT-4o-mini	0.5909	0.4235	0.4473	0.2972	0.4105	0.4410	0.6342	0.6070	0.6068	0.5918	0.4039	0.6100	0.5050
Average													
–	0.4299	0.2789	0.3359	0.2631	0.2853	0.3199	0.4220	0.4114	0.3871	0.4154	0.2966	0.4090	0.3549

Table 1: Model performance across different categories. The Per. and Know. are abbreviations for perception-oriented and knowledge-oriented domains. Ins and Thk are abbreviations for instruct and thinking. The best scores are in **bold**, and the secondary are underlined. Qwen3-VL refers to Qwen3-VL-30B-A3B, and the details of all evaluated models are shown in the Appendix C.2.

where $f(\cdot)$ is the aggregation function. Implementation details are provided in Appendix B.

2.5 Evaluation Data Composition

We conduct evaluations using 1500 real-world images and 500 synthetic images for each synthetic domain. For each experiment, we conduct $N = 100$ evaluation sessions with a candidate pool size B of 8.

3 Main Results

The main results on different fields of images are presented in Table 1. Several key conclusions can be drawn from these results.

Existing MLLMs struggle to gather missing evidence proactively. Performance on active visual reasoning is weak across image fields. More than half of the models yield means below 0.4, indicating poor ability to retrieve target images within the required steps. Even the strongest models, Qwen2.5-VL-72B and Qwen3-VL-Thinking, averages only 0.70 and 0.65. The gap to reliable performance remains large, and further improvements are required.

Synthetic images present greater challenges. Across categories, synthetic images yield lower scores than real-world images. The real-world images score the highest among various models, with 0.7695 for Qwen2.5-VL-32B and 0.6844 for Mimo-VL, while synthetic categories trail. GPT-4o-mini achieves 0.5909 on real-world images but only 0.2972 on kitchen. We hypothesize that synthetic domains typically exhibit repeated layouts and subtle semantic variations, creating higher demands in visual perception and world knowledge. These results indicate a need to enhance these two capabilities, particularly the recognition and utilization of fine-grained differences.

Perception-based tasks underperform knowledge-based tasks. Among all synthetic image categories, the performance on perception-based tasks is consistently lower than that on knowledge-oriented tasks. The average score for perception tasks is 0.2966, while for knowledge-oriented tasks it is 0.4090, indicating a relative difference of approximately 27.5%. The most significant disparity is observed with Intern3-VL-14B in the traffic and kitchen scenarios, where the scores are 0.6395 and 0.1915 respectively, resulting in a sub-

stantial performance gap. Our hypothesis is that perception tasks focus on fine-grained visual recognition, whereas knowledge-oriented tasks emphasize knowledge inference and assumption. These results indicate that the models exhibit stronger capabilities in knowledge reasoning compared to visual perception.

Scaling generally improves performance, but not monotonically. A key trend is the overall benefit of scaling: larger parameter sizes deliver substantial gains across model families. In the Qwen2.5-VL series, the average score rises from 0.1523 (3B) to 0.6978 (72B), representing more than 300% improvement, also happened in Intern3-VL-series. These results reinforce scaling as a reliable means of enhancing both perceiving and thinking (Chen et al., 2023), echoing two processes in active reasoning. However, the trend is not strictly monotonic: Qwen2.5-VL-32B achieves an average score of 0.5894, clearly outperforming Intern3-VL-38B at 0.4125 despite their similar scale, suggesting that beyond a certain scale, the stage and quality of training constrain both perception and thinking.

Explicit thinking significantly boosts performance. Across various models, the integration of explicit thinking mechanisms consistently leads to substantial performance improvements. For example, Kimi-VL demonstrates this clearly: its base model achieves an average score of only 0.2201, while the thinking-enhanced variant, Kimi-VL-Thinking, reaches 0.3153, an increase of 43.3%. Notably, smaller models augmented with thinking capabilities can rival or even surpass much larger models without such mechanisms. Mimo-VL (0.6102), trained with explicit thinking reinforcement learning from Qwen2.5-VL-7B, outperforms GPT-4o-mini (0.5050). These findings suggest that thinking augmentation is an effective pathway to stronger active reasoning, especially in dynamic information settings.

4 Further Discussions

We first contrast passive and active reasoning to quantify the effectiveness and limitations of interactive information acquisition in Section 4.1. We then analyze the dominant failure modes of multimodal active reasoning from the perspectives of perception, information acquisition strategy, and decision timing in Section 4.2, followed by an exploration of practical strategies for improvement in Section 4.3.

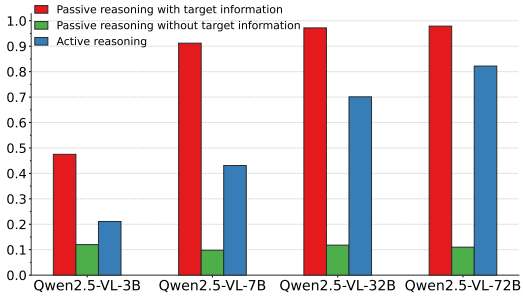


Figure 3: The accuracy compared with the passive reasoning. The no-information baseline yields near-chance accuracy, confirming the reliability of our tasks.

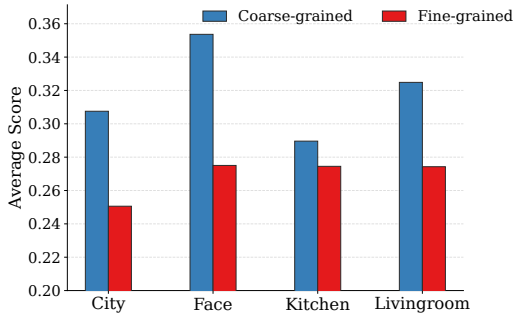
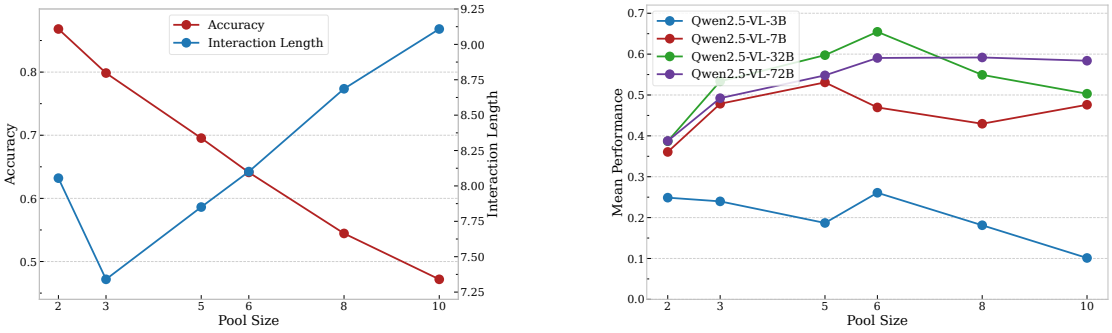


Figure 4: Performance gap between the original images (coarse-grained) and edited images (fine-grained).

4.1 Active Reasoning is Effective but Limited

To quantify how much interaction can compensate for missing evidence, we compare three conditions: (i) a no-information baseline, (ii) a passive oracle setting where the target image description is provided, and (iii) the proposed active setting. The no-information baseline yields near-chance accuracy, confirming that the task cannot be solved without informative evidence. The results are exhibited in Figure 3 and we summarize the findings below.

Oracle performance reveals a fundamental integration bottleneck. Under the passive oracle setting, performance is strongly size-dependent: large models can achieve high accuracy (> 0.9), whereas the 3B model remains below 0.5 even when the target description is fully available. This size-dependent performance on passive reasoning suggests two potential failure modes for active reasoning: small models struggle with integrating multimodal information, while larger models that perform well here may still fail in the active setting by failing to ask the right questions for information seeking incompletely.



(a) Accuracy and interaction length with different pool sizes.

(b) Average score across tasks of Qwen2.5-VL models.

Figure 5: Effect of candidate-pool size on performance.

Interaction partially recovers missing evidence, but a persistent oracle gap remains. Removing the target description leads to a drop of more than 0.72 points on average, underscoring the decisive role of task-specific evidence. With interaction enabled, models recover part of this loss with accuracy rising to about 0.54 on average and improving with scale, approaching the oracle setting. Nevertheless, a substantial gap remains between active reasoning and passive oracle performance (below 0.29 on average), suggesting that current models still fail to acquire and consolidate sufficient evidence through querying reliably.

4.2 How MLLMs fail for Active Reasoning

To pinpoint the sources of this gap, we conduct controlled analyses from three complementary perspectives: (i) *fine-grained perception* over candidate pools, (ii) *information acquisition strategy* under different interaction protocols, and (iii) *verification and decision timing* during iterative querying. Together, these experiments provide a unified diagnosis of why current MLLMs underperform in active reasoning despite strong performance under complete evidence.

4.2.1 Limited Fine-grained Perception

Fine-grained visual differences pose significant challenges. To assess whether active reasoning is sensitive to fine-grained perceptual variations, we construct edited image sets across four perception-oriented domains by modifying a single attribute of the original images (Figure 15 in Appendix E.1). As shown in Figure 4, nearly all models exhibit substantial degradation relative to the original sets, with an average drop of 15.8%. This trend indicates that subtle variations within the candidate pool can significantly affect overall performance, highlighting the importance of robust fine-grained discrimination at the start of the interaction.

Larger candidate pools exacerbate perceptual difficulty. To examine how perceptual demands scale with the candidate pool size, we evaluate Qwen2.5-VL-series models under varying numbers of candidate images (Figure 5a). As pool size increases, accuracy generally declines while the number of reasoning steps rises, with an average 45.7% relative degradation in accuracy and a 13.1% increase in steps. Larger pools amplify perceptual uncertainty, which subsequently propagates to the thinking stage and weakens question generation and evidence consolidation. These results suggest that perception is central to active reasoning, and models that can quickly identify and focus on a small, relevant subset of candidates are more likely to sustain effective interaction.

Imbalanced perception and reasoning degrades performance. To examine how perceptual difficulty and reasoning efficiency jointly shape overall performance, we analyze model scores as a function of candidate pool size (Figure 5b). We find that all models achieve peak performance at a moderate pool size, where the performance degrades at both extremes: larger pools incur an average 15.8% drop, consistent with perceptual bottlenecks, while very small pools reduce scores to roughly 70.1% of the moderate pool level, reflecting redundant interactions and limited planning. These patterns support the validity of our metric that jointly accounts for accuracy and efficiency, and indicate that strong active reasoning requires a balanced interplay between perceptual selection and reasoning.

4.2.2 Non-strategic Information Acquisition

Expanding the answer space alone yields limited benefits. To test whether richer feedback encourages a better query strategy, we replace binary QA

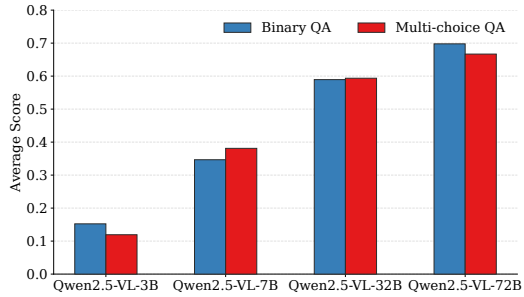


Figure 6: Performance under binary QA (default) and multiple-choice QA.

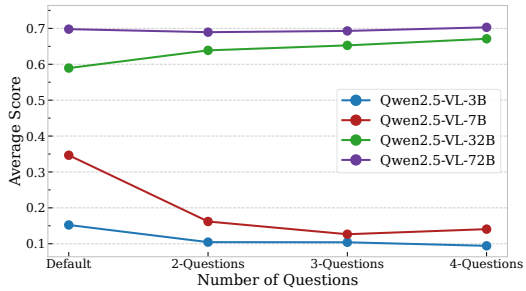


Figure 7: Performance under multiple questions.

with multiple-choice QA within a single interaction to expand the answer space. As shown in Figure 6, we find that all four models show only marginal differences from the default setting, retaining on average 98.58% of their binary-QA performance. Although multiple-choice reduces the number of rounds, it also lowers accuracy, indicating that current models do not systematically improve information acquisition capabilities and fail to exploit the enlarged answer space.

More questions do not improve performance. To further probe whether accessing more information per interaction improves performance, we adopt a n-Questions protocol where the model proposes multiple questions before receiving any answers. As demonstrated in Figure 7, we find substantial degradation for the two smaller models and only slight gains for the two larger models. This suggests that the benefit of additional signals depends critically on strategic question planning and coherent evidence synthesis. Without them, extra feedback tends to function as noise rather than improving decision quality.

4.2.3 Untimely Decision-making

Unreliable verification lengthens the reasoning process. To characterize how verification quality affects interactive progress, we track stepwise candidate-pool reduction across reasoning steps as shown in Figure 8. We find clear diminishing

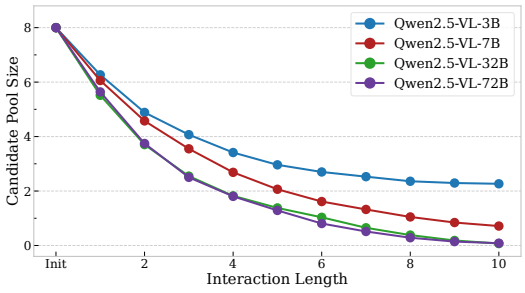


Figure 8: Candidate image–pool size versus stepwise reasoning progress. We use Qwen2.5-VL-7B as a QA-based filter to prune the pool via QA-pairs.

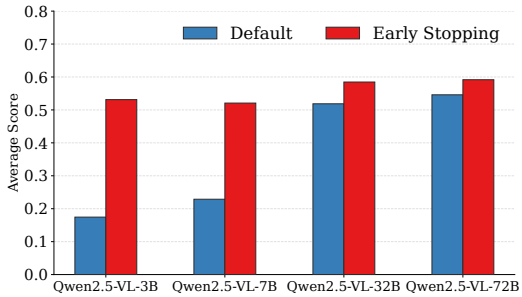


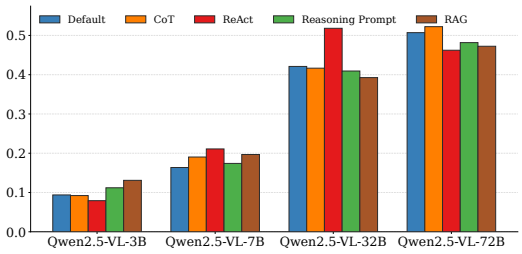
Figure 9: Performance comparison between the default setting and early stopping.

returns: across four models, the first four steps account for about 77.3% of the total reduction in candidate count, whereas the remaining steps contribute only about 22.7%. This indicates that as interactions become longer, models gain increasingly little new information, consistent with unreliable verification and limited ability to integrate evidence dynamically, thereby prolonging the reasoning process with low payoff.

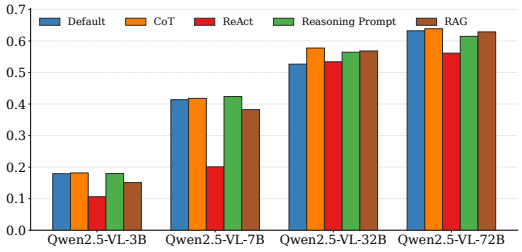
Early stopping improves efficiency and accuracy. To test whether delayed commitment is a key source of inefficiency, we recompute scores with an early-stopping rule that terminates once the candidate pool shrinks to 1 and outputs the remaining candidate. As exhibited in Figure 9, we find consistent improvements across all four models, with an average gain of 51.7%. The improvement diminishes with model size, suggesting that larger models are better at timely commitment, while smaller models benefit more from explicit stopping criteria. Overall, these results highlight verification and stopping policies as a practical strategy to improve both accuracy and interaction efficiency.

4.3 Strategies for Improving Active Reasoning

Building on the diagnosed bottlenecks, we explore several strategies to improve active reasoning from: (i) *perceptual enhancement* to strengthen visual



(a) Performance in perception-oriented images.



(b) Performance in knowledge-oriented images.

Figure 10: Performance comparison across perceptual and reasoning enhancements.

encoding and candidate selection (e.g., Reasoning Prompt and RAG), (ii) *reasoning enhancement* to improve multi-step planning and evidence integration during interaction (e.g., CoT and ReAct). Furthermore, we study *explicit thinking mechanisms* that are learned through training to better support search, verification, and timely stopping. Details are described in Appendix C.3. We summarize our observations as follows.

Perceptual enhancement yields larger gains than knowledge enhancement. We first examine the effectiveness of both strategies. As shown in Figure 10, we observe that the gains are highly task-dependent: perception-oriented images benefit consistently from targeted interventions, whereas knowledge-oriented settings show limited additional improvement from various enhancements. This pattern suggests that current MLLMs already satisfy most of the knowledge requirements in our benchmark, while perceptual encoding and candidate selection remain the primary bottlenecks for active reasoning.

Different models benefit different enhancements To examine whether the most effective strategy varies with model scale and capability, we compare the four interventions across multiple models as shown in Figure 10a. We find heterogeneous responses: Qwen2.5-VL-3B benefits most from perception-enhanced strategies, while larger models exhibit larger gains from thinking-enhanced

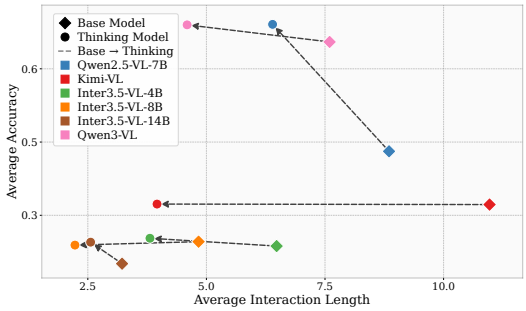


Figure 11: The accuracy and interaction length under thinking and non-thinking modes. The thinking mode of the Intern3.5-VL series is under the thinking prompt in official settings.

methods. This pattern aligns with the hypothesis in Section 4.1: smaller models are mainly limited by perceptual capacity, whereas larger models are more constrained by reasoning and verification. These results imply that effective improvements should be model-specific, targeting the dominant bottleneck rather than applying a uniform intervention across scales.

Training explicit thinking outperforms prompt-based thinking. To further assess the role of explicit thinking mechanisms, we evaluate multiple models under both non-thinking and thinking modes as shown in Figure 11. We find that enabling the thinking mode consistently improves accuracy while reducing interaction length, indicating more effective search and verification than the non-thinking mode. Moreover, the improvements from trained thinking modes exceed those from inference-time prompting strategies such as CoT and ReAct, highlighting the limitations of training-free heuristics. Notably, Mimo-VL, trained with thinking reinforcement learning from Qwen2.5-VL-7B, achieves a 76.1% improvement and substantially outperforms CoT and ReAct. These findings suggest that training explicit thinking better instills step-aware search, evidence consolidation, and verification behaviors, motivating objectives that explicitly optimize for interaction efficiency and reliable stopping.

Additional analyses of more models, larger candidate pools, noise perturbations, thinking length, and framework quality, as well as human evaluation results and detailed case studies, are provided in Appendix D and Appendix E.

5 Conclusions

To advance the evaluation of multimodal active reasoning, we introduce GUESSBENCH, a benchmark that assesses MLLMs through an interactive target-guessing task. This setting requires active evidence acquisition and dynamic information integration over both perception-oriented and knowledge-oriented images. Evaluating 20 leading MLLMs, we find that current systems perform poorly, indicating substantial headroom for improvement. Our analysis reveals three primary bottlenecks that systematically hinder performance: limited perceptual capability, non-strategic information acquisition, and delayed decision making. Further experiments show that stronger reasoning consistently improves performance. Taken together, these results highlight the need for MLLMs that better align with active multimodal tasks in real-world settings.

Limitations

Despite a two-stage procedure intended to improve responses and align scores with human judgments, the guess agent can still be inaccurate. Errors may stem from a distribution shift or propagation across stages. Moreover, although we carefully engineer prompts, a single prompt does not transfer reliably across models and can lead to mismatches at inference time. Overall performance may be marginally degraded. Furthermore, we set the maximum new tokens to 5000, which may be insufficient for some cases in thinking models and can lead to incomplete responses from MLLMs.

Ethical Considerations

All images used in this study are sourced from the publicly available intrinsic images in the IIW dataset or generated using the publicly released Omnigen2. We include only non-sensitive content and exclude material that could reasonably be considered unsafe. All evaluated models are publicly available, and we used them strictly under their respective licenses and terms of use. Human ratings were provided by college students; no personal characteristics were requested, and no personally identifiable information was collected. Accordingly, we anticipate no specific ethical concerns arising from the data or evaluation procedures in this work. Furthermore, we leveraged large language models as writing assistants for tasks such as rephrasing sentences, improving grammatical flow, and refining technical descriptions for clarity.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 62576209) and STCSM (No. 2025SHZDZX025G05).

References

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, and 8 others. 2025. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

Qiguang Chen, Libo Qin, Jin Zhang, Zhi Chen, Xiao Xu, and Wanxiang Che. 2024a. M^3cot: A novel benchmark for multi-domain multi-step multi-modal chain-of-thought. In *Proc. of ACL*.

Xi Chen, Josip Djolonga, Piotr Padlewski, Basil Mustafa, Soravit Changpinyo, Jialin Wu, Carlos Riquelme Ruiz, Sebastian Goodman, Xiao Wang, Yi Tay, Siamak Shakeri, Mostafa Dehghani, Daniel M. Salz, Mario Lucic, Michael Tschannen, Arsha Nagrani, Hexiang Hu, Mandar Joshi, Bo Pang, and 24 others. 2023. Pali-x: On scaling up a multilingual vision and language model. *ArXiv*, abs/2305.18565.

Zhe Chen, Yusheng Liao, Zhiyuan Zhu, Haolin Li, Hongcheng Liu, Yanfeng Wang, and Yu Wang. 2025. Heterorag: A heterogeneous retrieval-augmented generation framework for medical vision language tasks. *arXiv preprint arXiv:2508.12778*.

Zhe Chen, Heyang Liu, Wenyi Yu, Guangzhi Sun, Hongcheng Liu, Ji Wu, Chao Zhang, Yu Wang, and Yanfeng Wang. 2024b. M^3AV: A multimodal, multigenre, and multipurpose audio-visual academic lecture dataset. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9041–9060, Bangkok, Thailand. Association for Computational Linguistics.

Zhe Chen, Hongcheng Liu, and Yu Wang. 2024c. Dialogmcf: Multimodal context flow for audio visual scene-aware dialog. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:753–764.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. 2024. Mme: A comprehensive evaluation benchmark for multimodal large language models. *Preprint*, arXiv:2306.13394.

Roopal Garg, Andrea Burns, Burcu Karagol Ayan, Yonatan Bitton, Ceslee Montgomery, Yasumasa Onoe, Andrew Bunner, Ranjay Krishna, Jason Baldridge, and Radu Soricut. 2024. Imagein-words: Unlocking hyper-detailed image descriptions. *Preprint*, arXiv:2405.02793.

Zhang Ge, Du Xinrun, Chen Bei, Liang Yiming, Luo Tongxu, Zheng Tianyu, Zhu Kang, Cheng Yuyang, Xu Chunpu, Guo Shuyue, Zhang Haoran, Qu Xingwei, Wang Junjie, Yuan Ruibin, Li Yizhi, Wang Zekun, Liu Yudong, Tsai Yu-Hsuan, Zhang Fengji, and 3 others. 2024. Cmmmu: A chinese massive multi-discipline multimodal understanding benchmark. *arXiv preprint arXiv:2401.20847*.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. 2024. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14375–14385.

Ibragim Khasanov. 2024. Guess, who i am? Apple App Store. IOS app, version 1.3.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji rong Wen. 2023. Evaluating object hallucination in large vision-language models. In *Conference on Empirical Methods in Natural Language Processing*.

Yunxin Li, Zhenyu Liu, Zitao Li, Xuanyu Zhang, Zhenran Xu, Xinyu Chen, Haoyuan Shi, Shenyuan Jiang, Xintong Wang, Jifang Wang, Shouzheng Huang, Xinping Zhao, Borui Jiang, Lanqing Hong, Longyue Wang, Zhuotao Tian, Baoxing Huai, Wenhan Luo, Weihua Luo, and 3 others. 2025. Perception, reason, think, and plan: A survey on large multimodal reasoning models. *arXiv preprint arXiv:2505.04921*.

Fuxiao Liu, Xiaoyang Wang, Wenlin Yao, Jianshu Chen, Kaiqiang Song, Sangwoo Cho, Yaser Yacoob, and Dong Yu. 2023. Mmc: Advancing multimodal chart understanding with large-scale instruction tuning. *arXiv preprint arXiv:2311.10774*.

Hongcheng Liu, Yixuan Hou, Heyang Liu, Yuhao Wang, Yanfeng Wang, and Yu Wang. 2025. Vocalbench-df: A benchmark for evaluating speech llm robustness to disfluency. *arXiv preprint arXiv:2510.15406*.

Hongcheng Liu, Yusheng Liao, Siqv Ou, Yuhao Wang, Heyang Liu, Yanfeng Wang, and Yu Wang. 2024a. Med-pmc: Medical personalized multi-modal consultation with a proactive ask-first-observe-next paradigm. *arXiv preprint arXiv:2408.08693*.

Hongcheng Liu, Pingjie Wang, Heyang Liu, Zhiyuan Zhu, Yusheng Liao, Yanfeng Wang, and Yu Wang. 2026a. Anchornet: Adaptive anchor token enhancement in video-grounded dialogue generation. *IEEE Journal of Selected Topics in Signal Processing*, pages 1–11.

Hongcheng Liu, Yuhao Wang, Zhe Chen, Pingjie Wang, Zhiyuan Zhu, Yixuan Hou, Yanfeng Wang, and Yu Wang. 2026b. Cross-modal coreference alignment: Enabling reliable information transfer in omni-llms. *arXiv preprint arXiv:2604.05522*.

Yexin Liu, Zhengyang Liang, Yueze Wang, Muyang He, Jian Li, and Bo Zhao. 2024b. Seeing clearly, answering incorrectly: A multimodal robustness benchmark for evaluating mllms on leading questions. *Preprint*, arXiv:2406.10638.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations (ICLR)*.

OpenAI. 2024. Gpt-4o mini: Advancing cost-efficient intelligence. OpenAI blog post.

Siqu Ou, Hongcheng Liu, Pingjie Wang, Yusheng Liao, Chuan Xuan, Yanfeng Wang, and Yu Wang. 2025. Bridging the dynamic perception gap: Training-free draft chain-of-thought for dynamic multimodal spatial reasoning. *Preprint*, arXiv:2505.16579.

Shaina Raza, Mizanur Rahman, Safiullah Kamawal, Armin Toroghi, Ananya Raval, Farshad Navah, and Amirmohammad Kazemeini. 2024. A comprehensive review of recommender systems: Transitioning from theory to practice. *ArXiv*, abs/2407.13699.

Kimi Team. 2025a. Kimi-VL technical report. *Preprint*, arXiv:2504.07491.

Qwen Team. 2025b. Qwen3-vl: Sharper vision, deeper thought, broader action. *Qwen Blog*. Accessed: 2025-10-04.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, and 1 others. 2025. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*.

Yuhao Wang, Yusheng Liao, Heyang Liu, Hongcheng Liu, Yu Wang, and Yanfeng Wang. 2024. Mm-sap: A comprehensive benchmark for assessing self-awareness of multimodal large language models in perception. *Preprint*, arXiv:2401.07529.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, Ze Liu, Ziyi Xia, Chaofan Li, Haoge Deng, Jiahao Wang, Kun Luo, Bo Zhang, Defu Lian, Xinlong Wang, and 3 others. 2025. Omnigen2: Exploration to advanced multimodal generation. *arXiv preprint arXiv:2506.18871*.

LLM-Core-Team Xiaomi. 2025. Mimo-vl technical report. *Preprint*, arXiv:2506.03569.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Yue Yang, Shuibai Zhang, Wenqi Shao, Kaipeng Zhang, Yi Bin, Yu Wang, and Ping Luo. 2024. Dynamic multimodal evaluation with flexible complexity by vision-language bootstrapping. *ArXiv*, abs/2410.08695.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. *ArXiv*, abs/2210.03629.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2024. A survey on multimodal large language models. *National Science Review*, 11(12).

Fangyi Yu, Lee Quartey, and Frank Schilder. 2023. Exploring the effectiveness of prompt engineering for legal reasoning tasks. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13582–13596, Toronto, Canada. Association for Computational Linguistics.

Qingchen Yu, Zifan Zheng, Ding Chen, Simin Niu, Bo Tang, Feiyu Xiong, and Zhiyu Li. 2025. Guessarena: Guess who i am? a self-adaptive framework for evaluating llms in domain-specific knowledge and reasoning. *arXiv preprint arXiv:2505.22661*.

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, and 3 others. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of CVPR*.

Xiongtao Zhou, Jie He, Lanyu Chen, Jingyu Li, Haojing Chen, Víctor Gutiérrez-Basulto, Jeff Z. Pan, and Hanjie Chen. 2025a. Miceval: Unveiling multimodal chain of thought's quality via image description and reasoning steps. *Preprint*, arXiv:2410.14668.

Zhanke Zhou, Xiao Feng, Zhaocheng Zhu, Jiangchao Yao, Sanmi Koyejo, and Bo Han. 2025b. From passive to active reasoning: Can large language models ask the right questions under incomplete information? *Preprint*, arXiv:2506.08295.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, and 1 others. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.

A Related Works

There have been numerous multimodal benchmarks for evaluating MLLMs across real-world scenarios (Liu et al., 2024b; Zhou et al., 2025a; Chen et al., 2024b). For example, POPE (Li et al., 2023) targets object perception, HallusionBench (Guan et al., 2024) examines image–text hallucination, and dynamicME (Yang et al., 2024) further explores the performance under generative images. To move beyond single-capability probes and toward broader utility, subsequent work has emphasized complementary axes (Liu et al., 2026b). MME (Fu et al., 2024) assesses knowledge utilization, and MathVista (Lu et al., 2024) evaluates visual mathematics. Reasoning-oriented evaluation has become a central focus in recent work, including M^{3}CoT (Chen et al., 2024a), MMMU (Yue et al., 2024), and so on (Ge et al., 2024; Liu et al., 2023). However, most benchmarks still present models with full information at inference time and thus primarily test passive reasoning (Liu et al., 2024a). This setting diverges from real-world use, where models must reason actively under incomplete or ambiguous evidence. To bridge this gap, we propose GuessBench, the first benchmark that systematically evaluates the active reasoning capability of multimodal large language models, complementing and extending existing evaluations.

B Evluation Metric

To comprehensively evaluate the active reasoning ability of MLLMs across diverse domains while accounting for both accuracy and efficiency, we propose a composite evaluation metric $\mathcal{S}$ that integrates response accuracy ($\mathcal{A}$), reasoning efficiency ($\mathcal{R}$), and planning competence ($\mathcal{P}$). Inspired by the design of GuessArena, we define $\mathcal{S}$ as follows:

$$\mathcal{S} = \mathcal{A} \cdot \frac{\omega + \mathcal{R} + \mathcal{P}}{\omega + 2}, \tag{3}$$

where $\omega \geq 1$ is a constant that preserves the baseline contribution of accuracy, which is set as 1 by default. In this way, the metric can primarily assess identification accuracy while quantitatively rewarding efficiency and strategic knowledge utilization.

Response Accuracy evaluates whether the MLLM correctly identifies the target image. It is formally defined as the average accuracy across N evaluation sessions, adjusted by a correction factor α as:

$$\mathcal{A} = \alpha \cdot \frac{1}{N} \sum_{n=1}^{N} \mathbf{I}(\hat{i}_n = i'_n). \tag{4}$$

The indicator function $\mathbf{I}(\cdot)$ equals 1 if the condition is satisfied and 0 otherwise. The correction factor $\alpha = 1/R_{\text{agent}}$ represents human-verified adjustments based on human labels. The term R_{agent} (GuessAgent reliability) measures the rate at which the GuessAgent provides a correct and valid response; a lower R_{agent} results in a larger α, ensuring the accuracy score is less influenced by the external mistakes from the GuessAgent. We provide an accuracy analysis of R_{agent} in Appendix D.3.

Reasoning Efficiency quantifies how quickly the MLLM completes the reasoning process, rewarding models that reach the solution in fewer steps. It is modeled as:

$$\mathcal{R} = \frac{1}{N} \sum_{n=1}^{N} \exp(-\beta \frac{T_n - T_{\text{min}}}{T_{\text{max}} - T_{\text{min}}}), \tag{5}$$

where β is a reward coefficient, set to 1 in our experiments, and T_n denotes the actual number of interaction length taken in the n-th evaluation session. The minimal reasoning length, $T_{\text{min}} = \lceil \log_2 B \rceil$, is derived from information-theoretic principles (ideal binary search), where B is the candidate set size and $\lceil \cdot \rceil$ denotes the ceiling operator. We set the practical maximum length T_{max} to 10. Consequently, models that complete the task with a query length T closer to T_{min} receive higher scores, directly reflecting greater efficiency in active reasoning progress.

Planning Competence measures the model's ability to formulate effective, adaptive strategies based on acquired evidence. It is defined using an exponential penalty function:

$$\mathcal{P} = \frac{1}{N} \sum_{n=1}^{N} \exp\left(-\max\left(0, \frac{T_n - T_B}{T_B}\right)\right), \tag{6}$$

where T_B is the upper bound implied by a simple one-by-one querying strategy, instantiated as the candidate-pool size B (i.e., $T_B = B$). If the actual reasoning step T exceeds T_B, the term inside the $\max(\cdot)$ function becomes positive, resulting in an exponential penalty. This formulation specifically penalizes models whose active reasoning steps exceed the limit of a non-strategic, exhaustive search, indicating a failure to re-plan effectively using the newly acquired evidence.

C Experimental Settings

C.1 Data Construcction

For real-world images, we sample from the IIW dataset (Garg et al., 2024). For synthetic data, we

first use GPT-4o-mini (OpenAI, 2024) to generate domain-specific, discriminative attributes, which are randomly composed into full descriptions (see Table 2 for details). These descriptions are then fed to OmniGen2 (Wu et al., 2025) to create images, using 50 inference steps at a resolution of 1024×1024. Next, we employ Qwen2.5-VL-7B (Bai et al., 2025) with the prompt "Describe the image in detail." to produce detailed captions that extend the initial attributes and serve as references for external sources when answering the model's strategic queries. Finally, we perform human verification to rigorously validate the accuracy and quality of all synthesized images and captions.

Group	Subtitle
City	Time of day, Weather, Street type, Buildings, Vehicles, People activity, Street elements
Face	Hair, Eyes, Eyebrows, Nose, Lips, Beard, Skin, Face shape, Other features
Kitchen	Overall kitchen style, Cabinets, Countertop material, Appliances, Flooring, Backsplash, Island/bar setup, Lighting, Decor
Livingroom	Overall living room style, Sofa, Coffee table, Wall decor, Lighting, Color scheme, Flooring, Window view, Notable accessory
Nature	Terrain type, Time of day, Weather condition, Vegetation, Water feature, Wildlife element, Light quality, Overall atmosphere
Animal	Animal types, Color and patterns, General actions, Habitats, Moods and characteristics
Traffic	Generated by GPT4o
Vehicle	Vehicle types, Vehicle colors, Materials and details, Conditions, Eras and design styles, Vehicle environments, Functional descriptions
Occupation	Profession list, Profession-specific actions, Work environments, Styles and moods

Table 2: Overview of subtitles among categories. More details can be found in the official code.

C.2 Evaluation Models

Model Name	Thinking Mode	Parameters
Intern3-VL (2025)	Non-think	2B, 8B, 14B, 38B, 78B
Intern3.5-VL (2025)	Both	2B, 4B, 8B, 14B, 38B
Qwen-2.5-VL (2025)	Non-think	3B, 7B, 32B, 72B
Qwen-3-VL-Instrust (2025b)	Non-think	30B-A3B
Qwen-3-VL-Thinking (2025b)	think	30B-A3B
Mimo-VL (2025)	Think	7B
Kimi-VL (2025a)	Non-Think	16.4B-A3B
Kimi-VL-Thinking (2025a)	Think	16.4B-A3B
GPT4o-mini (2024)	Non-think	-

Table 3: Details of evaluation MLLMs.

To comprehensively investigate the factors that influence active reasoning performance, we evaluate 20 mainstream MLLMs of varying sizes, architectures, and training strategies. Since the reasoning mode is closely tied to active reasoning performance, we report it separately. We evaluated these models with official code and set the temperature to 0.1 for consistent performance. The details are summarized in Table 3.

C.3 Enhanced Strategies

The enhanced strategies are divided into two categories: perception-enhanced and thinking-enhanced approaches. The former includes RAG (Chen et al., 2025) and the reasoning prompt (Yu et al., 2023), while the latter comprises CoT (Wei et al., 2023) and ReAct (Yao et al., 2022). Specifically, RAG provides retrieval results based on external information, and the reasoning prompt encourages the model to focus on specific aspects of the experimental setting. In contrast, CoT guides the model to reason step by step, whereas ReAct enables the model to plan subsequent actions as new information is obtained. The detailed prompts are shown in Appendix F.

D Additional Discussion

D.1 Further Results on Qwen3-VL

We further evaluate additional Qwen3-VL models, with results shown in Table 4. Compared with the models in Table 1, these variants achieve generally better performance, especially on perception-based tasks. Despite this improvement, the findings are consistent with Section 3: synthetic images are more challenging, perception-based tasks underperform knowledge-based tasks, and explicit thinking significantly boosts performance. These results further corroborate our analysis in Section 3.

D.2 Extended Analysis of Active Reasoning

Larger candidate pools can degrade performance. To assess how these bottlenecks scale in more realistic settings, we further evaluate GuessBench with larger candidate pool sizes, $N \in \{16, 32, 64\}$. The results in Table 5 and Table 6 reveal a clear scaling trend. As the candidate pool increases, accuracy drops substantially while the number of interaction rounds rises, suggesting that active reasoning becomes markedly more difficult as the search space expands. This degradation is particularly pronounced for smaller models, whereas larger models retain partial capability even under larger candidate pools. These results high-

Model	Real	Perception					Knowledge				Average		
		City	Face	Kitchen	Livingroom	Nature	Animal	Traffic	Vehicle	Occupation	Per	Know	Total
Qwen3-VL-Series													
Qwen3-VL-2B-Ins	0.1072	0.1168	0.1327	0.1366	0.1614	0.1341	0.1917	0.1470	0.1084	0.1291	0.1363	0.1441	0.1365
Qwen3-VL-2B-Thk	0.2506	0.1645	0.2023	0.1719	0.1425	0.2392	0.2127	0.2131	0.1458	0.2379	0.1841	0.2024	0.1981
Qwen3-VL-4B-Ins	0.6882	0.4763	0.4130	0.2980	0.4546	0.4474	0.7152	0.7646	0.6444	0.6727	0.4179	0.6992	0.5574
Qwen3-VL-4B-Thk	0.4226	0.4143	0.4071	0.3060	0.3628	0.4261	0.4493	0.4750	0.4059	0.6138	0.3833	0.4860	0.4283
Qwen3-VL-8B-Ins	0.7119	0.4935	0.5934	0.4714	0.4797	0.5730	0.7286	0.7338	0.6582	0.7479	0.5222	0.7171	0.6191
Qwen3-VL-8B-Thk	0.6782	**0.7048**	0.4943	**0.5936**	0.6018	0.6828	0.7482	**0.7703**	0.7160	0.7051	0.6155	0.7349	0.6695
Qwen3-VL-32B-Ins	0.7363	0.5183	0.5954	0.5364	0.6061	0.6165	0.7652	0.7680	0.7221	**0.7976**	0.5745	0.7632	0.6662
Qwen3-VL-32B-Thk	**0.8188**	0.6822	**0.7368**	0.5715	**0.7392**	**0.7562**	**0.8074**	0.7314	**0.8191**	0.7554	**0.6972**	**0.7783**	**0.7418**
Average													
–	0.5517	0.4463	0.4469	0.3857	0.4435	0.4844	0.5773	0.5754	0.5275	0.5824	0.4414	0.5657	0.5021

Table 4: Qwen3-VL performance across different categories without alpha correction. The Per and Know are abbreviations for perception and knowledge domains. Ins and Thk are the abbreviations for instruction and thinking. The best scores are in **bold**, and the secondary are underlined.

Size	QV-3B	QV-7B	QV-32B	QV-72B	Avg
2	0.62	0.88	0.99	0.97	0.86
3	0.42	0.81	0.94	0.95	0.78
5	0.36	0.68	0.88	0.91	0.71
6	0.33	0.64	0.85	0.89	0.68
8	0.24	0.47	0.80	0.84	0.59
10	0.16	0.37	0.68	0.85	0.52
16	0.08	0.33	0.45	0.71	0.39
32	0.02	0.10	0.14	0.31	0.14
64	0.00	0.04	0.07	0.15	0.07

Table 5: Accuracy under different candidate sizes, where "QV-" is an abbreviation for Qwen2.5-VL.

Size	QV-3B	QV-7B	QV-32B	QV-72B	Avg
2	10.59	7.89	7.09	6.59	8.04
3	9.53	7.53	6.14	6.27	7.37
5	10.09	7.95	6.77	7.17	8.00
6	10.11	8.30	7.15	7.46	8.26
8	10.44	8.90	7.50	7.88	8.68
10	10.39	9.68	8.02	8.47	9.14
16	18.86	14.75	9.35	11.57	13.63
32	17.39	15.95	12.82	15.60	15.44
64	20.61	17.43	15.69	17.14	17.72

Table 6: Interaction length under different candidate sizes, where "QV-" is an abbreviation for Qwen2.5-VL.

light a substantial robustness gap between current MLLMs and the demands of realistic interactive settings.

Longer reasoning does not guarantee higher accuracy. To examine whether thinking more reliably translates into better active-reasoning outcomes, we bucket Mimo-VL sessions by the length of their generated analysis and report accuracy within each range, using Qwen2.5-VL-7B as the baseline since Mimo-VL is trained from it. As shown in Table 7, overall, reasoning length is slightly negatively correlated with accuracy ($r = -0.15$), indicating a non-monotonic relationship. Moderate-length analyses (100∼400 tokens) substantially outperform the baseline, suggesting that a reasonable amount of structured reasoning helps active reasoning, whereas very long reasoning is more often associated with upstream perceptual or knowledge errors that trigger prolonged self-correction yet still end in incorrect decisions. Taken together, these results imply that improving fine-grained perception and multimodal evidence integration is more consequential than merely increasing reasoning length.

CoT length range	Accuracy	Count (n)
0 (baseline)	0.4310	1000
(0, 50]	0.0556	18
(50, 100]	0.8298	47
(100, 200]	0.7736	349
(200, 400]	0.7176	386
(400, 800]	0.6241	133
(800, 1600]	0.4615	39
(1600, ∞)	0.4286	28

Table 7: Average accuracy grouped by generated CoT length ranges.

Active reasoning is robust to pixel-level visual distortions. To further examine robustness to different types of noise in the image condition, we apply four common image distortions: brightness shift, color masking, contrast change, and Gaussian noise. Specifically, brightness shift randomly scales brightness, color masking blends the image with a random-color overlay, contrast change randomly scales contrast, and Gaussian noise adds random pixel noise. We apply each at two severity levels to the original images, and evaluate models under identical protocols. As reported in Table 8, both accuracy and the number of interaction rounds change only marginally across distortion types and severities, suggesting that conventional low-level corruptions have a limited impact on active reason-

Type	Level	Acc.	Rounds
Qwen2.5-VL-3B			
Default	–	**0.211**	**10.164**
Brightness	mild	0.200	10.231
	severe	0.223	10.300
Color Mask	mild	0.207	10.301
	severe	0.196	10.464
Contrast	mild	0.204	10.159
	severe	0.220	10.235
Gaussian Noise	mild	0.215	10.232
	severe	0.211	10.421
Avg. (w/ Noise)	–	**0.210**	**10.293**
Qwen2.5-VL-7B			
Default	–	**0.431**	**8.850**
Brightness	mild	0.455	8.885
	severe	0.464	8.908
Color Mask	mild	0.451	8.999
	severe	0.418	9.073
Contrast	mild	0.449	9.008
	severe	0.455	8.881
Gaussian Noise	mild	0.441	8.981
	severe	0.436	9.242
Avg. (w/ Noise)	–	**0.446**	**8.997**
Qwen2.5-VL-32B			
Default	–	**0.701**	**7.823**
Brightness	mild	0.716	7.893
	severe	0.718	7.911
Color Mask	mild	0.714	8.001
	severe	0.647	8.104
Contrast	mild	0.718	7.961
	severe	0.694	7.908
Gaussian Noise	mild	0.718	7.926
	severe	0.707	7.938
Avg. (w/ Noise)	–	**0.704**	**7.955**
Qwen2.5-VL-72B			
Default	–	**0.822**	**7.820**
Brightness	mild	0.823	7.886
	severe	0.803	8.022
Color Mask	mild	0.782	8.135
	severe	0.725	8.481
Contrast	mild	0.811	7.942
	severe	0.817	8.075
Gaussian Noise	mild	0.832	8.040
	severe	0.777	8.051
Avg. (w/ Noise)	–	**0.796**	**8.079**

Table 8: Performance under visual noise perturbations, including brightness shifts, color masking, contrast changes, and Gaussian noise.

ing in our setting. Among these distortions, color masking induces the largest drop, likely because it alters global color characteristics that can be informative for recognition, yet the decrease remains below 10%. In contrast, the pronounced degradation under semantic-level perturbations in Figure 4 indicates that the primary challenge lies in identifying and leveraging fine-grained, discriminative cues, rather than in handling generic images.

D.3 Extended Analysis of Evaluation Pipeline

GuessAgent provides reliable and consistent responses. To validate that our evaluation pipeline does not confound model performance with unreliable external feedback, we assess whether GuessAgent provides appropriate answers to model queries. Specifically, we randomly sample 100 sessions per model and ask human annotators to label each GuessAgent response as *True* or *False* with respect to its reasonableness and consistency with the target evidence. As summarized in Table 9, all models achieve a correctness rate above 95%, indicating that GuessAgent can reliably return evidence aligned with model queries. This supports that GuessBench evaluates the model's active reasoning ability rather than artifacts from the response mechanism.

Model	**Reliability Rate**
InternVL3-2B	99.01%
InternVL3-8B	98.71%
InternVL3-14B	94.12%
InternVL3-38B	98.34%
InternVL3-78B	97.82%
Qwen2.5-VL-7B	98.74%
Qwen2.5-VL-32B	98.58%
Qwen2.5-VL-3B	98.33%
Qwen2.5-VL-72B	98.01%
InternVL3.5-2B	100.00%
InternVL3.5-4B	96.14%
InternVL3.5-8B	98.25%
InternVL3.5-14B	99.39%
InternVL3.5-38B	98.15%
Kimi-VL-A3B	98.07%
Kimi-VL-A3B-Thinking	95.98%
Qwen3-VL-Instruct	98.28%
Qwen3-VL-Thinking	98.41%
MiMo-VL	99.04%
GPT-4o-mini	98.16%

Table 9: Reliability rate of GuessAgent by human verification.

GuessAgent Achieves the Best Results with Hybrid Evidence. To test whether the evaluation is overly influenced by caption-centric judging, we restrict GuessAgent to different evidence sources and compare both downstream performance and oracle correctness. Table 10 shows that the hybrid mechanism achieves the best overall results, while removing either the textual or visual component leads to lower accuracy, more interaction rounds, and reduced oracle correctness. These findings suggest that the visual component mainly acts as a complementary corrective signal when textual descriptions or attribute summaries are incomplete, rather than simply amplifying caption bias.

The evaluation pipeline is robust to GuessAgent noise. To further quantify the sensitivity of the evaluation pipeline, we simulate GuessAgent noise by randomly forcing GuessAgent to output "I do

Type	Accuracy	Rounds	Correctness
Hybrid	0.59	8.68	98.42%
LLM-only	0.57	8.99	96.23%
MLLM-only	0.51	9.58	90.79%

Table 10: Performance and oracle correctness under different evidence sources for GuessAgent.

Rate	QV-3B	QV-7B	QV-32B	QV-72B	Avg
0 (Original)	0.24	0.47	0.80	0.84	0.59
0.05	0.26	0.47	0.72	0.81	0.57
0.10	0.25	0.44	0.69	0.78	0.54
0.15	0.25	0.42	0.65	0.72	0.51
0.20	0.21	0.42	0.61	0.70	0.48

Table 11: Accuracy under different perturbation rates, where "QV-" is an abbreviation for Qwen2.5-VL.

not know" at rates of $5\%, 10\%, 15\%, 20\%$. The results on accuracy and interaction rounds are reported in Table 11 and Table 12. As the noise rate increases, accuracy consistently declines while the number of interaction rounds rises, confirming that GuessAgent reliability has a measurable effect on absolute performance. Importantly, however, the relative ranking and performance gaps across models remain broadly stable, suggesting that our main comparative conclusions are driven primarily by differences in model capability rather than artifacts of the benchmark pipeline.

The pipeline generalizes beyond image selection to other domains. To test whether our framework captures active reasoning in modalities beyond image identification, we instantiate the same interactive protocol on two additional tasks: text selection and medical consultation. In text selection, the model identifies a target historical figure from a candidate pool via active questioning. In medical consultation, it infers a patient's illness through iterative inquiry. As shown in Table 13, both tasks exhibit the same qualitative trend as image selection: larger models achieve higher performance with fewer interaction rounds. These results suggest that the proposed pipeline is transferable and can serve as a general evaluation template for active reasoning across domains.

Composite scores are stable under reasonable hyperparameter choices. To ensure that our conclusions are not artifacts of a particular scoring configuration, we perform a sensitivity analysis of the composite score S, which combines accuracy and reasoning efficiency via hyperparameters ω and β. We vary $\omega, \beta \in \{0.5, 1.0, 1.5\}$, and report results for the four Qwen2.5-VL models in Table 14 and category-wise results for Qwen2.5-VL-7B in Table 15. From these results, we find that changing (ω, β) mainly rescales the composite scores while largely preserving model rankings, especially among larger models. In addition, larger models are consistently less sensitive to (ω, β), reflecting their stronger accuracy and higher interaction efficiency. Across categories, perception-oriented domains (e.g., city, face, kitchen) show relatively larger score variation than real-world or knowledge-centric domains, consistent with our main finding that fine-grained perception remains the dominant bottleneck. Overall, these trends indicate that our core conclusions and evaluation design are robust under reasonable hyperparameter choices.

Rate	QV-3B	QV-7B	QV-32B	QV-72B	Avg
0 (Original)	10.44	8.90	7.50	7.88	8.68
0.05	10.45	9.11	7.88	8.02	8.87
0.10	10.49	9.21	8.13	8.15	9.00
0.15	10.46	9.37	8.42	8.33	9.15
0.20	10.56	9.35	8.65	8.39	9.24

Table 12: Interaction length under different perturbation rates, where "QV-" is an abbreviation for Qwen2.5-VL.

Text Selection		
Model	Accuracy	Rounds
Qwen2.5-VL-3B	0.69	10.6
Qwen2.5-VL-7B	0.77	8.14
Qwen2.5-VL-32B	0.77	6.42
Qwen2.5-VL-72B	0.95	5.72
Medical Consultation		
Model	ROUGE-1 Recall	Rounds
Qwen2.5-VL-3B	0.37	9.43
Qwen2.5-VL-7B	0.34	8.13
Qwen2.5-VL-32B	0.58	6.83
Qwen2.5-VL-72B	0.58	6.33

Table 13: Results on additional active reasoning tasks under the same framework. The text selection task contains 100 samples, and the medical consultation task contains 30 samples.

D.4 Failure Mode Analysis

Analysis on Normal Cases To identify the dominant sources of failure in interactive target guessing, we manually inspect 200 unsuccessful sessions and categorize the failure modes as summarized in Table 16. We find that planning-related issues (e.g., planning failure and repetitive questioning, often accompanied by exceeding the maximum number of turns) occur more frequently than isolated per-

ω	β	Qwen2.5-VL-3B	Qwen2.5-VL-7B	Qwen2.5-VL-32B	Qwen2.5-VL-72B
0.5	0.5	0.1772	0.3835	0.6403	0.7564
0.5	1.0	0.1694	0.3784	0.6377	0.7551
0.5	1.5	0.1626	0.3736	0.6353	0.7539
1.0	0.5	0.1600	0.3518	0.5919	0.6990
1.0	1.0	0.1523	0.3466	0.5894	0.6978
1.0	1.5	0.1455	0.3419	0.5870	0.6965
1.5	0.5	0.1497	0.3305	0.5576	0.6584
1.5	1.0	0.1420	0.3253	0.5551	0.6571
1.5	1.5	0.1352	0.3206	0.5527	0.6559

Table 14: Composite scores S of four Qwen2.5-VL models under different (ω, β).

ω	β	Real	City	Face	Kitchen	Livingroom	Nature	Animal	Traffic	Vehicle	Profession
0.5	0.5	0.5536	0.2031	0.1756	0.2084	0.2201	0.2870	0.5363	0.5112	0.5633	0.5767
0.5	1.0	0.5524	0.1953	0.1679	0.2030	0.2117	0.2799	0.5306	0.5112	0.5633	0.5683
0.5	1.5	0.5511	0.1884	0.1612	0.1981	0.2042	0.2732	0.5251	0.5112	0.5633	0.5603
1.0	0.5	0.5100	0.1839	0.1586	0.1898	0.1993	0.2615	0.4920	0.4718	0.5233	0.5280
1.0	1.0	0.5087	0.1761	0.1509	0.1844	0.1910	0.2544	0.4863	0.4718	0.5233	0.5196
1.0	1.5	0.5074	0.1692	0.1442	0.1795	0.1835	0.2477	0.4807	0.4718	0.5233	0.5116
1.5	0.5	0.4796	0.1721	0.1484	0.1778	0.1866	0.2450	0.4620	0.4441	0.4938	0.4955
1.5	1.0	0.4783	0.1643	0.1407	0.1725	0.1782	0.2379	0.4563	0.4441	0.4938	0.4871
1.5	1.5	0.4771	0.1574	0.1340	0.1675	0.1707	0.2313	0.4507	0.4441	0.4938	0.4791

Table 15: Composite score S of Qwen2.5-VL-7B across categories under different (ω, β).

Error type	**Count**
Misunderstanding instructions	34
Planning failure	58
Repetitive questions	23
Perception errors	46
Missing knowledge	26
Other	13

Table 16: Failure mode analysis over 200 erroneous cases.

Error type	**Count**
Failed to notice subtle differences	33
Misperceived subtle differences	18
Planning failure	12
Misunderstanding instructions	18
Repetition questions	10
Other	9

Table 17: Failure mode analysis of 100 erroneous cases on edited-image examples.

ception or knowledge errors. This suggests that many failures stem from ineffective information aggregation and suboptimal query planning, rather than from a lack of visual recognition or factual knowledge alone. This diagnosis is consistent with our analysis of verification and stopping behavior in Section 4.2.3. Qualitative examples for each error category are provided in Figure 12 and Figure 13.

Analysis on Edited-Images Our current results identify fine-grained perception as a key bottleneck: edited-image settings cause a substantial drop in accuracy, and our failure taxonomy shows that perception-related errors are among the dominant failure modes. To investigate this issue more closely, we further categorize failures on edited-image examples, with the results reported in Table 17. We find that most errors stem from missing subtle differences, rather than from incorrect reasoning after the difference has already been detected. In particular, the largest category is failure to notice subtle differences, followed by misperception of subtle differences. These findings suggest that fine-grained perceptual limitations, rather than downstream reasoning alone, are a primary source

of difficulty in active reasoning.

E Case Study

E.1 Dataset Cases

We provide examples of images in Figure 14 and Figure 15, including images in the main procedure and edited images in further experiments.

E.2 Active Reasoning Cases

We provide three cases for illustration in Figure 16 and Figure 17, including the normal case and an explicit thinking case.

F The details of the evaluation procedure.

We provide the detailed prompts on the whole evaluation procedure and enhanced methods for reproduction, which are shown in Table 18-Table 24.

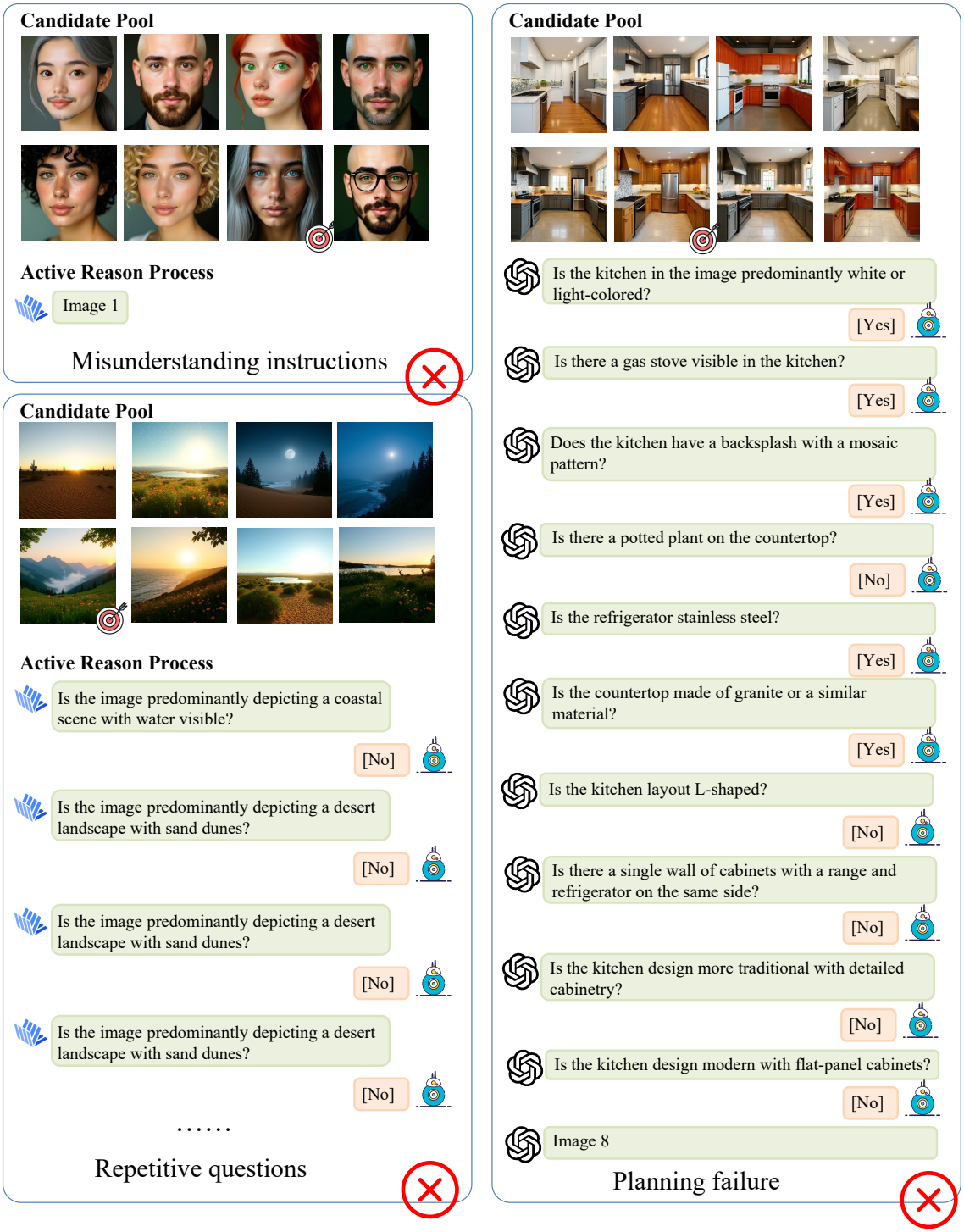


Figure 12: Representative error cases, including instruction misunderstanding, repetitive questioning, and planning failures.

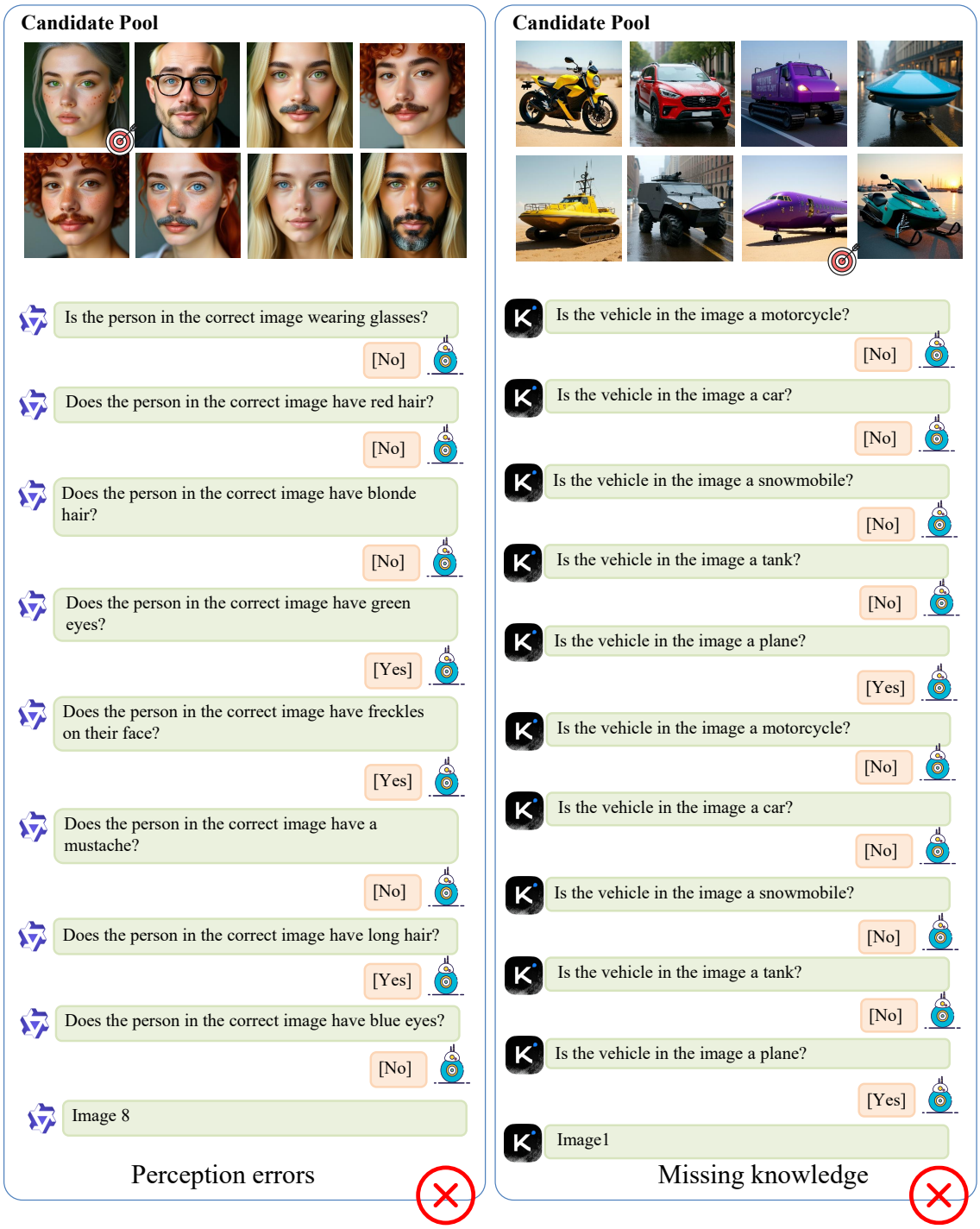


Figure 13: Representative error cases, including perception errors and missing knowledge.

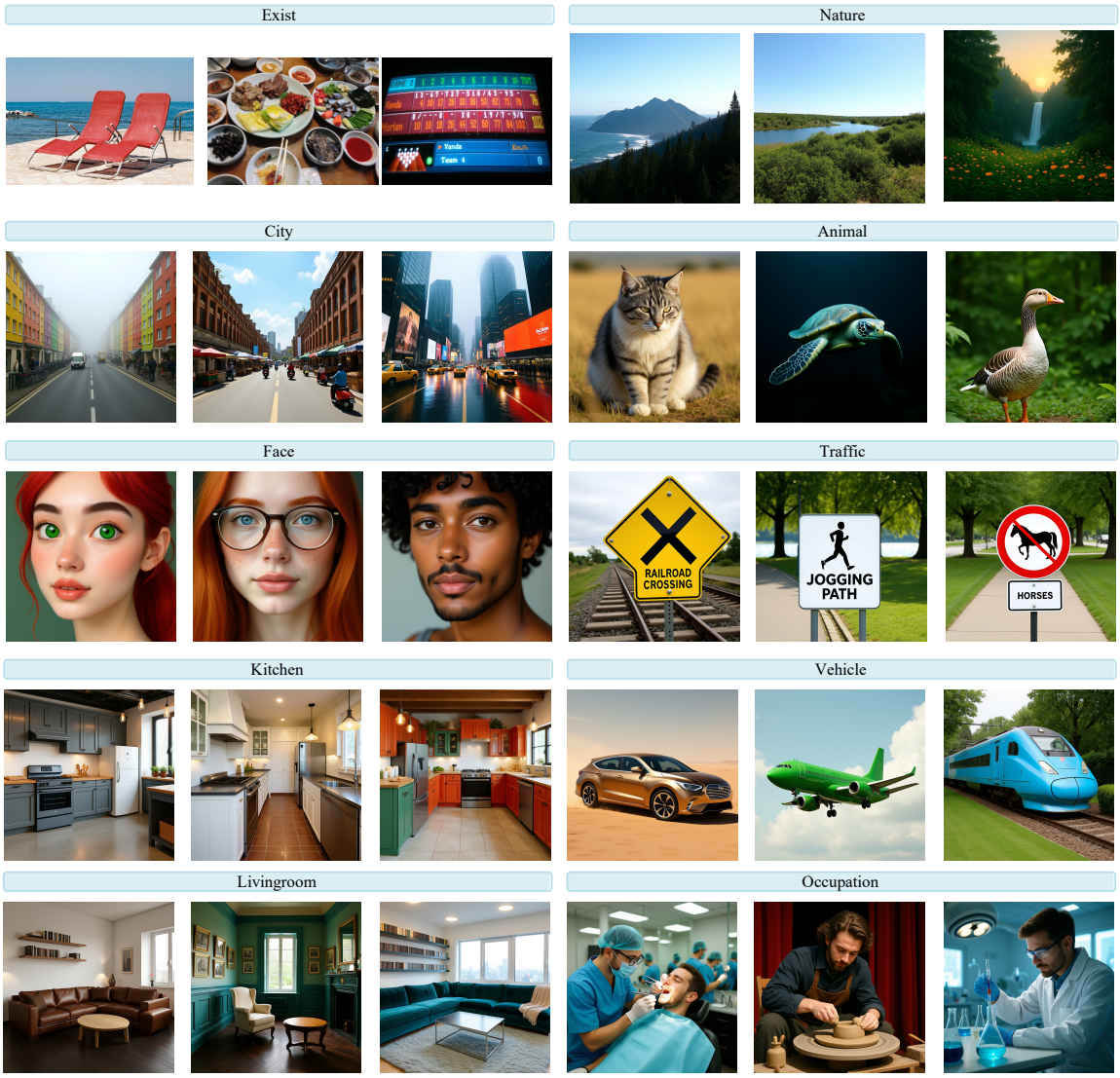


Figure 14: The cases of ten image fields.

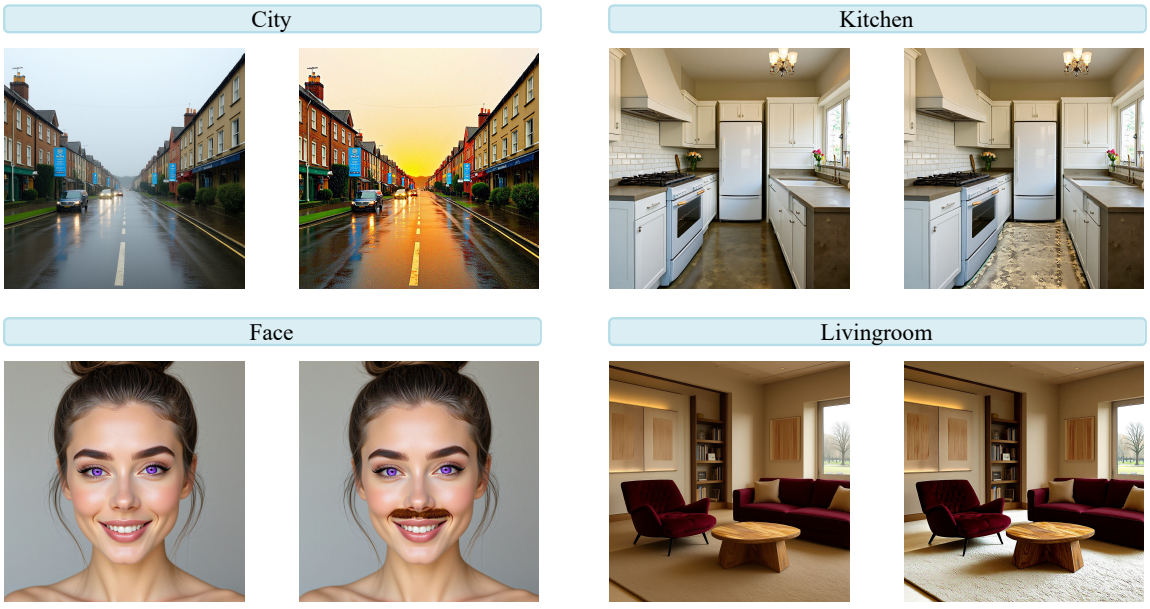


Figure 15: The cases of original images and edited images.

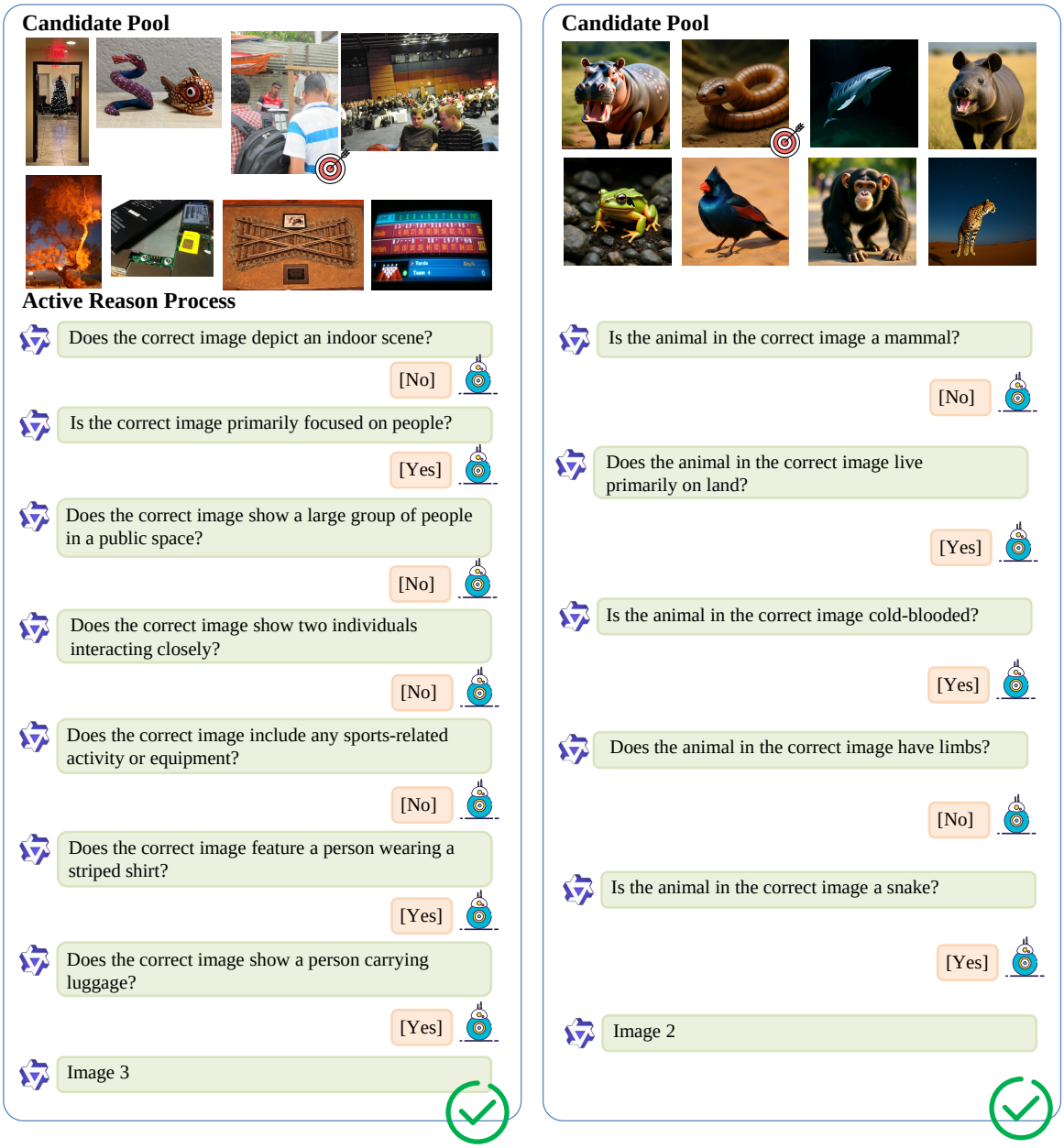


Figure 16: The cases of active reasoning.

Candidate Pool

<think>\nSo, let's start by analyzing the images. First, I need to find a distinguishing feature. Let's check if there's a bookshelf. Image 5 has a red bookshelf, image 6 also has a red bookshelf. Wait, no, let's list features. Let's think about the presence of a bookshelf. Let's ask if the image contains a bookshelf. That can eliminate some. Let's go with that.\n</think>Does the image contain a bookshelf?

[No]

<think>\nGot it, since the answer is No, we can eliminate images with bookshelves. Let's check which images have bookshelves: Image 5 and Image 6. So now we focus on the remaining images (1-4,7-8). Next, let's think of another feature. Maybe the type of seating: is there a single armchair? Let's check. Image 1 has a single brown armchair. Let's ask about that.\n</think>Does the image contain a single armchair?

[No]

<think>\nGot it, answer is No, so eliminate images with single armchairs. Image 1 has a single armchair, so now remaining are 2-4,7-8. Next, check for a large red sectional sofa. Image 2 has a large red sectional, Image 7 has a large red sectional. Let's ask about that.\n</think>Does the image contain a large red sectional sofa?

[No]

<think>\nGot it, answer is No, so eliminate images 2 and 7. Now remaining: 1,3,4,6,8. Wait, no: Image 2 and 7 have large red sectionals, so they're out. Now check other features. Let's look at the ceiling type. Image 3 has a smooth ceiling with recessed lights, Image 4 has an exposed ceiling with fluorescent lights, Image 1 has a dark ceiling with track lights, Image 6 has a dark ceiling with track lights, Image 8 has a smooth ceiling with recessed lights. Let's ask about the ceiling having exposed ducts or pipes. Image 4 has exposed ducts, Image 1 and 6 have exposed pipes/ducts. Let's ask: Does the image have exposed ducts or pipes on the ceiling?

[Yes]

<think>\nGot it, answer is Yes, so eliminate images without exposed ducts/pipes. Images 1,4,6 have exposed ducts/pipes. Now remaining: 1,4,6. Next, check the type of seating. Image 1 has a single armchair, Image 4 has a large sectional, Image 6 has a single armchair. Wait, earlier we eliminated single armchairs? Wait no, earlier question was \"Does the image contain a single armchair?\" and answer was No, so Image 1 and 6 have single armchairs, so they should be eliminated. Wait, maybe I made a mistake. Let's recheck. Wait, the second question was \"Does the image contain a single armchair?\" and answer was No, so Image 1 (has single armchair) and Image 6 (has single armchair) are eliminated. So remaining after second question: 2,3,4,7,8. Then third question was \"Does the image contain a large red sectional sofa?\" answer No, so eliminate 2 and 7. Now remaining: 3,4,8. Then fourth question: \"Does the image have exposed ducts or pipes on the ceiling?\" answer Yes, so eliminate 3 and 8 (since 3 has smooth ceiling, 8 has smooth ceiling). So remaining is 4. Wait, let's check Image 4: it has exposed ducts on the ceiling, yes. So now, let's confirm. Let's list steps:\n\n1. Q1: Does the image contain a bookshelf? → No → eliminate 5,6.\n\n2. Q2: Does the image contain a single armchair? → No → eliminate 1,6.\n\n3. Q3: Does the image contain a large red sectional sofa? → No → eliminate 2,7.\n\n4. Q4: Does the image have exposed ducts or pipes on the ceiling? → Yes → eliminate 3,8 (since 3 has smooth ceiling, 8 has smooth ceiling). So remaining is 4.\n\nNow, let's check Image 4: it has exposed ducts on the ceiling, a large sectional (but wait, earlier Q3 was No, so Image 4's sofa is dark, not red. So yes, Image 4 is the one. Now, next question to confirm. Let's check the type of coffee table. Image 4 has a round glass coffee table. Let's ask: Is the coffee table round? → Yes. That would confirm.\n</think>Is the coffee table round?

[Yes]

<think>\nGot it, answer is Yes. Now, check which image has a round coffee table. Image 4 has a round glass coffee table. Let's confirm other features. Image 4 has exposed ducts on the ceiling (from Q4), no bookshelf (Q1), no single armchair (Q2), no large red sectional (Q3), and round coffee table (Q5). So this confirms Image 4.\n</think>Image 4

Figure 17: The cases of active reasoning with explicit thinking.

Prompt

Real images and perception-oriented images

You are presented with a series of images, from which one is the correct answer. Your task is to identify this correct image by asking a series of questions within 10 turns. You must use all available images to formulate your questions, and ultimately output the number corresponding to the correct image.

Instructions

1. **Objective:** Identify the correct image among the provided options by asking yes/no questions.
2. **Question Format:** Each question must be a single, clear, and concise **yes/no question**. Do not request additional hints.
3. **Iterative Process:** After each question, you will receive a "[Yes]" or "[No]" answer. Use this information, along with previous answers, to formulate your next question.
4. **Final Answer:** When you are confident you have identified the correct image, state its corresponding number as your final answer (e.g., "Image 3").
5. **Efficiency:** The goal is to identify the correct image with the fewest possible questions.
6. **Repeating the same question is strictly prohibited.** Each question must aim to gather new information and actively eliminate a significant number of remaining incorrect images. Questions should progress logically from broader characteristics to more specific details.

Start

Let's begin. Here are the images you need to consider:

Knowledge-oriented images

You are presented with a series of images, from which one is the correct answer. Your task is to identify this correct image by asking a series of questions within 10 turns. You must use all available images to formulate your questions, and ultimately output the number corresponding to the correct image.

Instructions

1. **Objective:** Identify the correct image among the provided options by asking yes/no questions.
2. **Question Format:** Each question must be a single, clear, and concise **yes/no question**. Do not request additional hints.
3. **Iterative Process:** After each question, you will receive a "[Yes]" or "[No]" answer. Use this information, along with previous answers, to formulate your next question.
4. **Final Answer:** When you are confident you have identified the correct image, state its corresponding number as your final answer (e.g., "Image 3").
5. **Efficiency:** The goal is to identify the correct image with the fewest possible questions.
6. **Repeating the same question is strictly prohibited.** Each question must aim to gather new information and actively eliminate a significant number of remaining incorrect images.
7. **Questions should progress logically from broader characteristics to more specific details**, emphasizing the underlying knowledge, function, history, cultural significance, scientific principles, or relationships to other objects that differentiate them.

Start

Let's begin. Here are the images you need to consider:

Table 18: Prompt for multimodal large language models.

Prompt for state detector

You are a text classifier. Your task is to analyze the user's input and determine its type.

Classification Rules:

1. **Output <QUERY>**: If the input is clearly a **question**. A question typically ends with a question mark (?) or is phrased in an interrogative manner.
2. **Output <STOP>**: If the input is in the format of **"Image [number]"** (e.g., "Image 3", "Image 10").
3. **Output <INVALID>**: if the input cannot be answer by yes/no.

Here is the text:

Table 19: Prompt for state detector.

Prompt for evidence provided by LLM

Prompt for evidence provided by LLM

You are the judge in an image selection game. Your role is to evaluate player questions against a set of given images, one of which has been secretly designated as the **correct image**. You must answer player questions truthfully and objectively based on the description of this chosen image.

Judging Rules:

1. **Response Format:** You may only use four standard responses: **"[Yes]"**, **"[No]"** and **"[Invalid]"**. Provide no additional explanations or hints. Maintain absolute objectivity and do not reveal any specific information about the correct image beyond the yes/no answer.

2. **Question Evaluation:**

Respond with **"[Yes]"** or **"[No]"** if the question can be answered based on the description of the **correct image**.

Respond with **"[Invalid]"** if the question can not be answered based on the description of the **correct image**.

Judge's Responsibilities:

Maintain absolute neutrality; provide no additional hints.

Strictly follow the above rules to ensure the fairness of the game process.

The description of the correct image is:

The question is:

Prompt for evidence provided by MLLM

You are the judge in an image selection game. Your role is to evaluate player questions against a set of given images, one of which has been secretly designated as the **correct image**. You must answer player questions truthfully and objectively based on the description of this chosen image.

Judging Rules:

1. **Response Format:** You may only use four standard responses: **"[Yes]"**, **"[No]"** and **"[Invalid]"**. Provide no additional explanations or hints. Maintain absolute objectivity and do not reveal any specific information about the correct image beyond the yes/no answer.

2. **Question Evaluation:**

Respond with **"[Yes]"** or **"[No]"** if the question can be answered based on the description of the **correct image**.

Respond with **"[Invalid]"** if the question can not be answered based on the description of the **correct image**.

Judge's Responsibilities:

Maintain absolute neutrality; provide no additional hints.

Strictly follow the above rules to ensure the fairness of the game process.

The correct image is:

The question is:

Table 20: Prompt for evidence provided.

Prompt

Real images and perception-oriented images

You are presented with a series of images, from which one is the correct answer. Your task is to identify this correct image by asking a series of questions within 10 turns. You must use all available images to formulate your questions, and ultimately output the number corresponding to the correct image.

Instructions

1. **Objective:** Identify the correct image among the provided options by asking yes/no questions.

2. **Question Format:** Each question must be a single, clear, and concise **yes/no question**. Do not request additional hints.

3. **Iterative Process:** After each question, you will receive a "[Yes]" or "[No]" answer. Use this information, along with previous answers, to formulate your next question.

4. **Final Answer:** When you are confident you have identified the correct image, state its corresponding number as your final answer (e.g., "Image 3").

5. **Efficiency:** The goal is to identify the correct image with the fewest possible questions.

6. **Repeating the same question is strictly prohibited.** Each question must aim to gather new information and actively eliminate a significant number of remaining incorrect images. Questions should progress logically from broader characteristics to more specific details.

Intelligent Guessing Strategy

1. **Chained Reasoning, Step by Step**: Before each question, construct a rigorous chain of reasoning based on your current clues. Clearly define all current possibilities and precisely pinpoint the "decisive question" that will most effectively eliminate distractions and rapidly lead you closer to the truth.

2. **Feature Filtering, Efficient Focus**: Prioritize asking about key features that possess strong differentiating power. Your goal is to swiftly eliminate a large number of ineligible cards, efficiently narrowing your focus to a few high-potential options.

3. **Know When to Stop, Avoid Redundancy**: Once you're confident about the target card, make your judgment decisively. Avoid unnecessary additional questions; these not only waste opportunities but could also negatively impact your score. Let's begin.

Start

Let's begin. Here are the images you need to consider:

Knowledge-oriented images

You are presented with a series of images, from which one is the correct answer. Your task is to identify this correct image by asking a series of questions within 10 turns. You must use all available images to formulate your questions, and ultimately output the number corresponding to the correct image.

Instructions

1. **Objective:** Identify the correct image among the provided options by asking yes/no questions.

2. **Question Format:** Each question must be a single, clear, and concise **yes/no question**. Do not request additional hints.

3. **Iterative Process:** After each question, you will receive a "[Yes]" or "[No]" answer. Use this information, along with previous answers, to formulate your next question.

4. **Final Answer:** When you are confident you have identified the correct image, state its corresponding number as your final answer (e.g., "Image 3").

5. **Efficiency:** The goal is to identify the correct image with the fewest possible questions.

6. **Repeating the same question is strictly prohibited.** Each question must aim to gather new information and actively eliminate a significant number of remaining incorrect images.

7. **Questions should progress logically from broader characteristics to more specific details**, emphasizing the underlying knowledge, function, history, cultural significance, scientific principles, or relationships to other objects that differentiate them.

Intelligent Guessing Strategy

1. **Chained Reasoning, Step by Step**: Before each question, construct a rigorous chain of reasoning based on your current clues. Clearly define all current possibilities and precisely pinpoint the "decisive question" that will most effectively eliminate distractions and rapidly lead you closer to the truth.

2. **Feature Filtering, Efficient Focus**: Prioritize asking about key features that possess strong differentiating power. Your goal is to swiftly eliminate a large number of ineligible cards, efficiently narrowing your focus to a few high-potential options.

3. **Know When to Stop, Avoid Redundancy**: Once you're confident about the target card, make your judgment decisively. Avoid unnecessary additional questions; these not only waste opportunities but could also negatively impact your score. Let's begin.

Start

Let's begin. Here are the images you need to consider:

Table 21: Prompt for chain-of-thought.

Prompt

Knowledge background prompt

real world:Subject Matter, Camera Angle, Shot Type and Composition, Lighting and Atmosphere, Color and Specific Details.

face: hair style and colore, eyes colore and shape, eyebrows shape and color, nose shape, lips thickness and color, beard type and color, skin tone and marking, face shape, glasses.

animal: Mammals, Birds, Reptiles, Fish, Insects and Arachnids, Farm Animals, Extinct.

city: Time of Day, Weather, Street Type, Buildings, Vehicles, People Activity, Street Elements.

kitchen: Style, Cabinets color and style, Countertop, Appliances stove and fridge, Flooring, Backsplash, Island Presence and feature, Lighting, Decor.

livingroom: Style, sofa type and material, coffee table material and shape, wall decor, lighting, color scheme, flooring, window view, notable accessory.

nature: Terrain, Time of Day, Weather Condition, Vegetation, Water Feature, Wildlife Element, Light Quality, Overall Atmosphere.

profession: professions, actions, environments, styles moods.

traffic: traffic sigh.

vehicle: vehicle types, colors, materials details, conditions, eras styles, environments, function.

Real images and perception-oriented images

You are presented with a series of images, from which one is the correct answer. Your task is to identify this correct image by asking a series of questions within 10 turns. You must use all available images to formulate your questions, and ultimately output the number corresponding to the correct image.

Instructions

1. **Objective:** Identify the correct image among the provided options by asking yes/no questions.
2. **Question Format:** Each question must be a single, clear, and concise **yes/no question**. Do not request additional hints.
3. **Iterative Process:** After each question, you will receive a "[Yes]" or "[No]" answer. Use this information, along with previous answers, to formulate your next question.
4. **Final Answer:** When you are confident you have identified the correct image, state its corresponding number as your final answer (e.g., "Image 3").
5. **Efficiency:** The goal is to identify the correct image with the fewest possible questions.
6. **Repeating the same question is strictly prohibited.** Each question must aim to gather new information and actively eliminate a significant number of remaining incorrect images. Questions should progress logically from broader characteristics to more specific details.

Domain Background

Here's the domain-specific knowledge you can use to guide your guesses as a player: knowledge background and knowledge background prompt

Start

Let's begin. Here are the images you need to consider:

Knowledge-oriented images

You are presented with a series of images, from which one is the correct answer. Your task is to identify this correct image by asking a series of questions within 10 turns. You must use all available images to formulate your questions, and ultimately output the number corresponding to the correct image.

Instructions

1. **Objective:** Identify the correct image among the provided options by asking yes/no questions.
2. **Question Format:** Each question must be a single, clear, and concise **yes/no question**. Do not request additional hints.
3. **Iterative Process:** After each question, you will receive a "[Yes]" or "[No]" answer. Use this information, along with previous answers, to formulate your next question.
4. **Final Answer:** When you are confident you have identified the correct image, state its corresponding number as your final answer (e.g., "Image 3").
5. **Efficiency:** The goal is to identify the correct image with the fewest possible questions.
6. **Repeating the same question is strictly prohibited.** Each question must aim to gather new information and actively eliminate a significant number of remaining incorrect images.
7. **Questions should progress logically from broader characteristics to more specific details**, emphasizing the underlying knowledge, function, history, cultural significance, scientific principles, or relationships to other objects that differentiate them.

Domain Background

Here's the domain-specific knowledge you can use to guide your guesses as a player: knowledge background and knowledge background prompt

Start

Let's begin. Here are the images you need to consider:

Table 22: Prompt for reason prompt. The text denotes denotes the input information.

Prompt for RAG model

Task:
Answer this question based on the image.
Note:
- Your output must be exclusively and only [Yes] or [No].

Table 23: Prompt for RAG models.

Prompt for ReAct

Based on the previous questions and the [Yes/No] answer you received, consider which images have been eliminated and what information you still need to narrow down the possibilities. Specifically, detail:
1. **Which images have been eliminated** based on all previous questions and their corresponding [Yes/No] answers.
2. **Why these images were eliminated** (i.e., which specific characteristics or features contradicted the answers).
3. **What characteristics or features are you now focusing on** for the remaining possible images.
4. **What is your strategic approach for the next question?** Are you trying to broadly eliminate more images, or narrow down specific details of a few remaining options?
The history is:

Table 24: Prompt for ReAct.