UC Merced

Proceedings of the Annual Meeting of the Cognitive Science Society

Title

Metacognitive Active Perception with Memory-Guided Hypothesis Verification

Permalink

https://escholarship.org/uc/item/34f005nb

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 48(0)

Authors

Peng, Chengyi
Peng, Yifei
Wang, Zichen

Publication Date

2026

Copyright Information

This work is made available under the terms of a Creative Commons Attribution License, available at https://creativecommons.org/licenses/by/4.0/

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Metacognitive Active Perception with Memory-Guided Hypothesis Verification

Chengyi Peng[*†]**(362322020108@home.hpu.edu.cn)**[1]**, Yifei Peng**[*]**(362322020122@home.hpu.edu.cn)**[1] **& Zichen Wang(312305080407@home.hpu.edu.cn)**[1]
[1]Henan Polytechnic University, Jiaozuo, Henan, China

Abstract

Current multimodal visual models continue to improve in perceptual performance, yet still rely on passive, dense processing, resulting in high computational overhead. From a cognitive perspective, existing methods incorporate limited prior memory and resource allocation during perception, making it difficult to construct perception as goal-directed sequential decision-making. This paper proposes an active metacognitive framework inspired by biological memory priors, organizing visual understanding as a memory-driven hypothesis–verification process. The model forms initial beliefs from low-resolution global information and internal memory, and selects local observations to reduce uncertainty. As evidence accumulates, the system adaptively terminates perception based on confidence, enabling on-demand allocation of computational resources. Experiments across visual complexity settings show that this approach matches or surpasses traditional models while reducing inference time and token usage by approximately 15–30%, and produces structured, interpretable observation sequences.

Keywords: metacognitive control;active perception;memory-guided vision;sequential decision-making

Introduction

Vision is the cornerstone of human cognition. Inspired by human vision, computer vision has rapidly evolved: from convolutional neural networks (Purwono et al., 2022) to visual Transformers (Dosovitskiy, 2020), and on to multimodal large models with open-world understanding capabilities, such as GPT-4V (Hurst et al., 2024),LLaVA (Liu et al., 2023), machine perception capabilities continue to strengthen. However, state-of-the-art visual models often require energy consumption several orders of magnitude higher than humans when approaching human-level perceptual understanding Roy et al., 2019. This is typically linked to passive and dense processing patterns: regardless of task requirements or difficulty, computational resources are allocated to vast numbers of pixels with limited relevance to the current task (M. Li et al., 2023; Z. Li et al., 2017). This disparity prompts us to explore: can we draw inspiration from human visual mechanisms to reduce unnecessary computational resources while maintaining perceptual quality?

The biological visual system typically achieves efficient information acquisition through a fovea-peripheral vision mechanism: peripheral vision rapidly perceives overall contexts at lower resolution, while the fovea concentrates resources to perform high-resolution analysis of critical regions (Crayen et al., 2024). Cognitive neuroscience indicates that biological perception is not entirely passive; rather, it integrates memory-based prior knowledge with peripheral visual cues to form holistic predictive hypotheses about scenes (Rothkopf & Hayhoe, 2025). Concurrently, the oculomotor system actively controls gaze position, imparting sequential characteristics to perception (Yewbrey & Kornysheva, 2024). Research indicates that gaze shifts are typically associated with uncertainty, where new observations help reduce uncertainty in current perception (Lanillos et al., 2021). Furthermore, during perceptual tasks, biological systems often exhibit a tendency to prematurely terminate information sampling as evidence accumulates, thereby flexibly allocating perceptual resources according to varying task demands (M. Li et al., 2023; Stockart et al., 2025).

However, current multimodal models still exhibit certain limitations in incorporating biological perception mechanisms. Although models have made significant progress in logical reasoning through CoT (Zhang et al., 2023), they fail to model task-relevant uncertainty regulation and metacognitive decision-making mechanisms during perception (Liang et al., 2022). Due to the lack of prior memory utilization, models often struggle to form stable global hypotheses early on, thereby failing to effectively organize perception into a sequential observation process guided by uncertainty reduction. Consequently, under high-resolution visual inputs, mainstream paradigms still tend to adopt bottom-up dense encoding strategies, processing vast regions with limited relevance to the current task (Min et al., 2022). This approach partially constrains the model's ability to balance perceptual accuracy with computational efficiency (Rao et al., 2021; Wang et al., 2023).

To address these challenges, this paper proposes a metacognitive framework inspired by prior mechanisms of biological memory, attempting to introduce the concept of active perception from biological vision into modern multimodal models. Unlike the traditional approach of passively ingesting the entire image followed by dense processing, we organize the perception process as a memory-driven hypothesis-verification loop. Specifically, the model leverages a parameterized memory prior to form initial scene hypotheses. Based on these, it evaluates the potential information value of different regions, guiding the attention mechanism to perform discrete active

[*]Equal Contribution.
[†]Corresponding author.

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society* ©2026 the author(s). This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY).

sampling. This constitutes a sequential decision-making process: as observational evidence accumulates, the inference process can terminate early via an adaptive termination mechanism when confidence reaches a preset threshold. Experimental results demonstrate that this framework maintains stable perception performance while significantly reducing computational overhead, offering a clearer behavioral perspective for analyzing the model's perceptual decision-making process.From an interdisciplinary perspective, this paper represents an exploratory endeavor: translating the concepts of memory priors and uncertainty regulation from cognitive science into computable perceptual mechanisms for analyzing and organizing complex perceptual processes. This paper contributes as follows:

- Memory-driven active perception framework. We propose a method that organizes the perception process into a hypothesis-verification closed loop driven by memory. By integrating low-resolution global inputs with prior knowledge of the internal world, this approach guides subsequent perception processes to explore more structured interpretive spaces.
- Adaptive termination of active sampling mechanisms. A heuristic active sampling strategy based on uncertainty reduction was designed, incorporating confidence levels to achieve adaptive termination of the perception process. This approach models visual perception as a controllable sequential decision-making process.
- Systematic efficiency-performance evaluation. Experimental results across varying visual complexity scenarios demonstrate that this method achieves comparable perceptual performance to high-resolution dense inference baselines while significantly reducing inference time and computational overhead, all while maintaining observation paths with clear semantic interpretation.

Related Work

Active and Sequential Vision

Existing research generally models vision as a sequential process, utilizing attention mechanisms (Shang et al., 2022) or recurrent neural networks (Lemmel & Grosu, 2025) to progressively integrate local observations. Although these approaches incorporate a temporal dimension (Lemmel & Grosu, 2025), agents typically map the next action directly from the current observation, aiming to maximize cumulative reward or task accuracy (Liu et al., 2021; Shang & Ryoo, 2023). A critical limitation lies in their disregard for the hypothesis-testing mechanism central to biological vision. Without top-down guidance from prior assumptions, the perception process degenerates into blind local exploration rather than efficient cognitive verification, making rapid convergence under conditions of extreme information scarcity challenging.

Some studies have attempted to incorporate concepts from cognitive science such as active perception and uncertainty representation into computational vision models (M. Li et al., 2021; Sharafeldin et al., 2024), treating visual understanding as a sequential decision-making process (Brohan et al., 2022). While this approach has partially strengthened the connection between models and human perceptual behavior, its primary focus remains on modeling or explaining human behavioral patterns. It rarely directly influences the perceptual workflow within modern vision models to constrain observation and resource allocation during actual reasoning processes (Bajcsy et al., 2018).

Information-Driven Sampling and Adaptive Termination

Research has also examined the relationship between attentional selection and information value (Callaway et al., 2022), positing that perceptual systems should prioritize inputs with the highest information content to reduce uncertainty and enhance decision efficiency. Such approaches often employ metrics like information gain or uncertainty reduction to guide perception or sampling strategies (Ren et al., 2021), finding widespread application in active learning and perception. Nevertheless, these studies typically do not explicitly frame attentional behavior within a hypothesis-testing framework, with information measures often serving general exploratory objectives (Lee et al., 2023; M. Li et al., 2022).

Research on decision-making processes indicates that human judgments rely on evidence accumulated over time and terminate (Yildirim, 2025) when confidence reaches a threshold. This concept is widely used to explain tradeoffs between reaction time and accuracy (Masís et al., 2023). However, relevant models often treat termination mechanisms as products of the decision phase rather than integral components of the perception process itself (Ratcliff & McKoon, 2008).

Method

Overview

This paper proposes a metacognitive framework inspired by biological memory prior mechanisms(Figure1), organizing the perception process into a sequential hypothesis generation and verification workflow. Specifically, the model first combines memory information with coarse-grained perception to form an initial hypothesis about the scene. Subsequently, during the inference process, it actively selects locally valuable regions for observation. As evidence gradually accumulates, perception is adaptively terminated via a confidence-based termination criterion. This framework aims to dynamically allocate perception resources according to task requirements, thereby ensuring perception effectiveness while avoiding unnecessary computational overhead.

Memory-Driven Initial Belief Modeling

During the initial perception phase, the model must form an overall judgment of the scene within a limited budget. To achieve this, we introduce a memory-prior-driven initial hypothesis mechanism. The model constructs an initial belief distribution about the potential scene state y using only low-resolution global input x_1 and the memory representation M

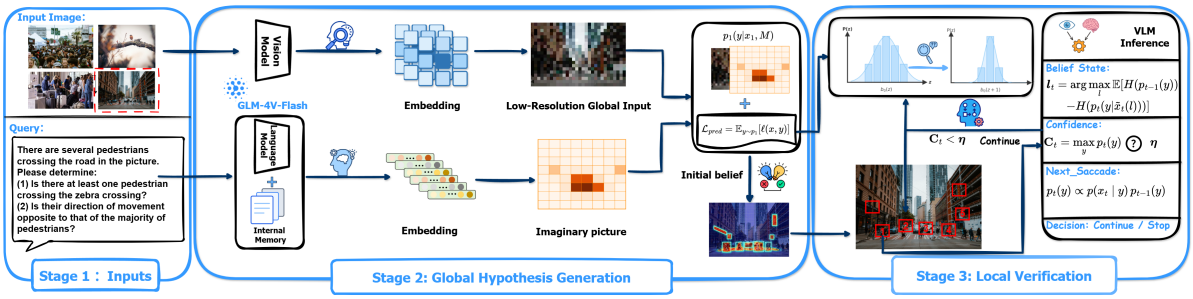

Figure 1: Overview of the proposed metacognitive perception framework. Stage 1: Receive raw image and query; Stage 2: Integrate low-resolution global visual information with internal memory representations to construct initial beliefs about the scene and form top-down prior assumptions; Stage 3: Local verification loop: under the current belief state, heuristically select observations based on maximizing uncertainty reduction, and adaptively determine whether to terminate perception via a confidence threshold mechanism.

of internal world knowledge (a fixed-dimensional vector set of general scene or task-related priors):

$$p_1(y \mid x_1, M) \tag{1}$$

where y denotes the task-relevant scene interpretation and Mrepresents internal memory, which is formalized as a set of fixed, structured representations. It draws from the pre-trained knowledge base of the underlying model. In our implementation, this is achieved through explicit contextual prompts. This design mirrors the biological mechanism whereby long-term memory guides early visual attention, thereby shaping the initial belief distribution without the need for additional parameter optimization. This belief distribution characterizes the system's relative confidence in different interpretations prior to detailed observation, serving as a predictive approximation of the full high-resolution input x. Accordingly, we constrain the initial hypothesis by minimizing the expected prediction error under this distribution:

$$\mathcal{L}_{\text{pred}} = \mathbb{E}_{y \sim p_1}\left[\ell(x, y)\right] \tag{2}$$

where $\ell(\cdot)$ denotes the deviation metric between predictions induced by the hypothesis y and the actual input. By modeling this memory-driven belief distribution during early perception, the model constrains subsequent perception within a more structured interpretive space, providing prior constraints for subsequent local verification and decision-making. It should be noted that this memory module is not used to store reasoning trajectories or linguistic context, but solely to influence early observation selection.

Uncertainty-Driven Active Observation Selection

After forming an initial hypothesis, the perception process evolves into a gradual validation and refinement of that hypothesis. We model local perception as an active sampling process whose goal is to minimize the uncertainty of the current belief distribution within a finite number of steps. At step t, the model selects the next observation location l_t from candidate regions based on the current belief distribution $p_{t-1}(y)$, guided by an uncertainty-reduction-based heuristic information gain metric:

$$l_t = \arg\max_{l} \mathbb{E}[H(p_{t-1}(y)) - H(p_t(y \mid x_t(l)))] \tag{3}$$

where $H(\cdot)$ denotes information entropy, and $x_t(l)$ represents the high-resolution local observation obtained at position l. This metric is used to compare the relative information value across different regions, rather than calculating mutual information in the strict sense. Upon acquiring new local evidence, the model updates its current belief through a normalization update rule:

$$p_t(y) \propto p(x_t \mid y)\, p_{t-1}(y) \tag{4}$$

Thus, new observations are integrated into existing beliefs. This updating process allows previously dominant hypotheses to be suppressed or modified under the influence of new evidence, enabling the perceptual process to gradually converge.

Adaptive Termination Mechanism Based on Confidence Level

To avoid unnecessary oversampling, the model evaluates the certainty level of the current belief after each perception step. Specifically, we define the confidence metric as:

$$C_t = \max_{y} p_t(y) \tag{5}$$

We compare this value with a preset threshold η. When $C_t \geq \eta$, the perception process terminates and the system directly outputs the current decision. From a utility perspective, this termination condition can be understood as determining whether the potential benefit of continuing sampling is lower than its computational cost:

$$\mathbb{E}[\Delta \mathcal{U}_{t+1}] < \lambda \cdot \text{Cost}_{t+1} \tag{6}$$

where $\Delta \mathcal{U}_{t+1}$ denotes the potential decision improvement from additional sampling, and λ represents the cost weight. This mechanism embodies a trade-off between decision confidence and computational overhead. To ensure algorithmic

Table 1: Performance comparison of methods at different visual complexity Levels. Under three distinct perception-level settings, we systematically compare our proposed method with four baseline models across three dimensions: answer quality, reasoning time, and resource consumption, with BL-1 serving as the Backbone. Results are reported under both fixed budget and adaptive budget modes.

Level	Model	Fixed Budget (K=5)			Adaptive Budget		
		Score	Time ↓	Tokens ↓	Score	Time ↓	Tokens ↓
Level 1	BL-1	**97.44**	11.74	3378	**98.52**	12.19	3489
	BL-2	82.28	**10.25** ↓12.7%	**2547** ↓24.6%	84.82	13.43 ↑10.2%	**2742** ↓21.4%
	BL-3	85.78	11.98 ↑2.0%	2689 ↓20.4%	85.84	12.66 ↑3.9%	2754 ↓21.1%
	BL-4	75.64	13.71 ↑16.8%	2764 ↓18.2%	77.23	13.28 ↑8.9%	2861 ↓18.0%
	Ours	**97.23**	**9.89** ↓15.8%	**2464** ↓27.1%	**98.24**	**8.21** ↓32.6%	**2263** ↓35.1%
Level 2	BL-1	**95.05**	16.74	3862	**95.11**	16.78	3789
	BL-2	74.53	16.52 ↓1.3%	3514 ↓9.0%	77.86	15.84 ↓5.6%	3567 ↓5.9%
	BL-3	71.79	15.79 ↓5.7%	3422 ↓11.4%	73.19	**15.23** ↓9.2%	**3542** ↓6.5%
	BL-4	66.45	**13.64** ↓18.5%	**3344** ↓13.4%	67.58	16.35 ↓2.6%	3788 ↓0.0%
	Ours	**94.73**	**11.23** ↓32.9%	**3258** ↓15.6%	**94.89**	**12.34** ↓26.5%	**3362** ↓11.3%
Level 3	BL-1	91.21	17.56	4042	92.16	18.23	4121
	BL-2	70.42	18.24 ↑3.9%	4217 ↑4.3%	71.23	20.23 ↑11.0%	4511 ↑9.5%
	BL-3	68.46	18.67 ↑6.3%	4289 ↑6.1%	69.54	19.24 ↑5.5%	4274 ↑3.7%
	BL-4	65.34	19.67 ↑12.0%	4334 ↑7.2%	65.23	20.25 ↑11.1%	4587 ↑11.3%
	Ours	**92.67**	**16.23** ↓7.6%	**3964** ↓1.9%	**94.21**	**17.54** ↓3.8%	**4024** ↓2.4%

termination, all experiments set a maximum perceptual step count T as a fallback condition.

Experiment

Settings

Visual Stimuli To investigate the hypothesis-testing behavior of models under varying levels of perceptual stress, we constructed a hierarchical perceptual stress test suite. Based on visual content density, saliency, and semantic ambiguity, we categorized test scenes into three perceptual tiers. (1) Level 1: Primarily captures basic visual information, featuring clear images with simple structures and minimal occlusions or distractions; (2) Level 2: Introduces more complex visual feature integration requirements, with images containing moderate background interference and local complexity; (3) Level 3: Scenes exhibit significant occlusions and strong interference, with highly incomplete overall information. Models must leverage top-down memory priors to guide attention allocation. For each scenario, we designed specific queries focusing on local details, relational attributes, or subtle cues, avoiding reliance on purely linguistic inference. Decision quality was jointly assessed by human experts and character agents, aiming to validate deep alignment between model behavior and human perceptual processes.

Baselines To ensure fair comparison among different perception and decision strategies, all methods are based on the same visual language model GLM-4V-Flash, with unified inference parameters such as temperature and decoding strategy. We constructed four baseline categories representing distinct perception approaches: (1) BL-1: Performs a single inference pass on the high-resolution full map, representing traditional perception; (2) BL-2: Utilizes only the VLM's general inference capabilities to plan observation paths based on low-resolution full-image data, representing a general-purpose agent lacking specialized internal memory modules; (3) BL-3: Selects observation points based on classical saliency maps, representing passive perception driven by underlying physics; (4) BL-4: Randomly selects observation regions, serving as a reference for evaluating the effectiveness of active sampling.

Metrics To quantify the model's performance, we use the following metrics for evaluation: (1) Score, calculates the degree of alignment between the model's output and the user's query. To ensure objectivity, scores are determined and verified jointly by human experts and an agent evaluator, and the average is taken. (2) Time, taken per query, which measures the model's decision-making efficiency. (3) Tokens, total number of tokens consumed per task, which measures the model's computational overhead.

Main Results

Overall Performance Comparison As shown in Table 1, we evaluated and compared each method across three dimensions: (1) answer quality; (2) reasoning time; and (3) resource consumption. Time and Tokens were averaged under two dis-

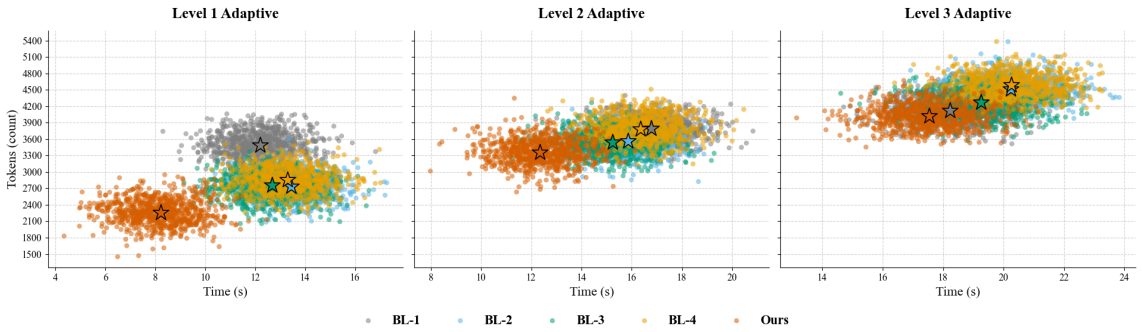

Figure 2: Comparison of inference time and resource consumption. Our approach demonstrates significant speed advantages and resource efficiency across all levels.

tinct settings: a resource-constrained configuration with a fixed high-resolution observation count (K=5), and an adaptive setting allowing models to adjust observation steps based on confidence scores. This approach captures differences in reasoning behavior under varying perception budget constraints. BL-1, representing the traditional perception model, exhibited no significant variation across these two settings. In our implementation, the observation window size is set to $[10 \times 10]$ pixels, and the adaptive termination threshold η is empirically set to $[0.85]$. Information gain is heuristically approximated by the entropy reduction of the token probability distribution generated by the VLM's language head.

Answer quality Under both inference modes, our approach achieved overall optimal or near-optimal response quality. As shown in Table 1, in Level 1 scenarios, our non-fixed mode scores approached those of BL-1 with high-resolution full-image inference, indicating that this method maintains stable perceptual performance while significantly reducing computational overhead in tasks with clear visual structures and minimal interference. As visual complexity increases in Level 2, our approach maintains response quality comparable to BL-1, demonstrating stable performance under moderate perceptual pressure. In the high-complexity Level 3 scenario, our approach outperforms BL-1 in both modes. This result indicates that relying solely on dense, high-resolution inputs does not always yield more reliable decisions in the presence of severe occlusions and semantic interference. Instead, selective observation helps mitigate the interference of irrelevant visual evidence on task judgments, thereby more effectively extracting task-relevant information.

Reasoning Time Our approach demonstrates superior overall temporal efficiency across varying levels of complexity (Table 1). Under resource-constrained settings with fixed observations (K=5), our method maintains competitive inference times when compared against other approaches at identical perception steps. In non-fixed settings (Figure 2), our method achieves lower average completion times than BL-1 and other baselines across all levels, demonstrating its ability to effectively control computational overhead through dynamic adjustment of observation steps. This advantage is particularly pronounced in Level 1/2 scenarios, revealing that when task structure is clearer, the model can terminate high-resolution observations earlier, thereby avoiding unnecessary perception computations. In contrast, BL-3 and BL-4 exhibit relatively stable delays as observation steps increase due to their fixed or random observation strategies.

Resource consumption Our model demonstrates superior resource efficiency across varying complexity scenarios (Table 1). BL-1 consistently employs a dense, high-resolution observation strategy, resulting in significant token overhead in high-complexity scenarios averaging over 4,100 tokens at Level 3. In contrast, our approach maintains comparable or even higher response quality while keeping token consumption at a lower level. Particularly in non-fixed adaptive settings (Figure 2), our model dynamically adjusts the number of observation steps based on the current task difficulty. This enables effective reasoning with fewer local observations in most examples, thereby avoiding the redundant computations commonly found in baseline methods.

Explainable and Efficiency routes Although the aforementioned results demonstrate that our model outperforms baseline methods in both response quality and reasoning efficiency, it warrants further analysis whether its perceived behavior exhibits more structured exploration patterns. To this end, we visualize high-resolution observation trajectories of the model in challenging scenarios (Figure 3), comparing regional selection behaviors among BL-3, BL-4, and our model.

As shown in the figure, the observation paths of BL-4 exhibit distinct non-directional characteristics. Due to the random generation of observation locations, even after multiple iterations, the sampling points remain highly dispersed spatially. The model typically requires a large number of observations to cover task-relevant key regions, and paths vary significantly between different samples, lacking stable spatial convergence trends. After incorporating classical saliency maps, the observation paths of BL-3 become more spatially

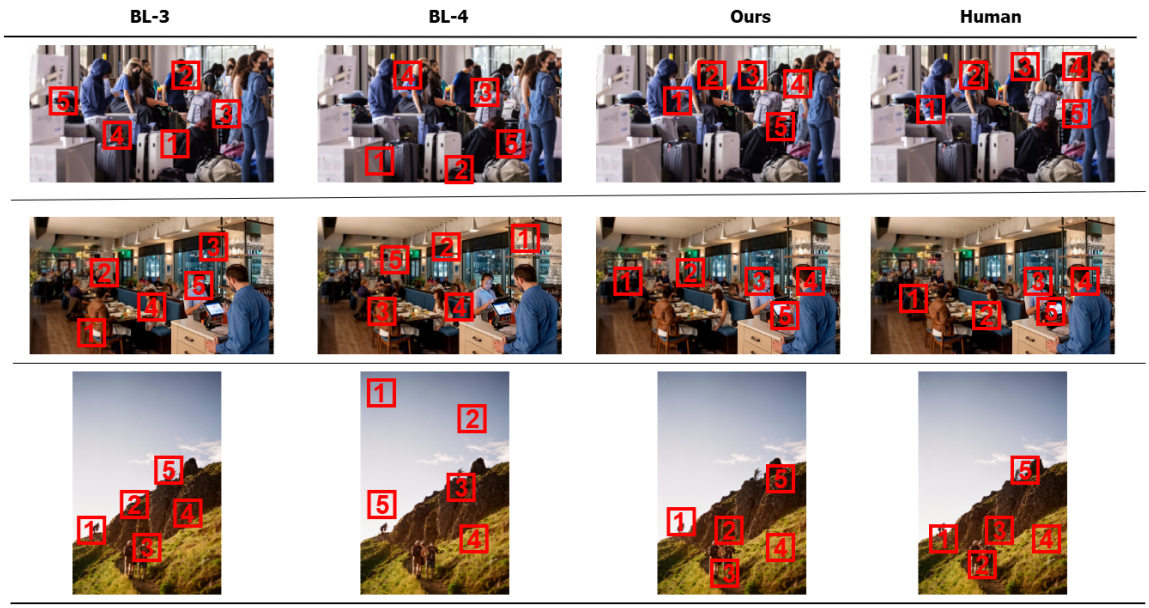

Figure 3: Visual exploration trajectory comparison.Unlike the undirected sampling of BL-4 or the saliency paths of BL-3, our approach demonstrates gradual convergence.

concentrated. The model tends to prioritize areas with more prominent structural information, thereby reducing redundant observations to some extent. This strategy primarily relies on low-level visual saliency, which does not always align with the semantic requirements of specific queries. In challenging scenarios, it may deviate from task-relevant regions

In contrast, Ours shows consistent stepwise convergence in multi-round perception. Initially, it explores regions that maximize belief uncertainty reduction. Guided by multiple observations and accumulated evidence, locations progressively reduce belief uncertainty. This process mirrors human observation trajectory structures. Ultimately, it achieves effective decisions in fewer steps and adaptively terminates based on confidence.

Discussion

Resource-Rational Perceptual Loop

While traditional vision models rely on passive, dense processing, our framework demonstrates that integrating memory priors and uncertainty-driven active sampling can yield a highly efficient, resource-rational perceptual loop. This approach shifts machine perception toward a goal-directed decision-making process, endowing intermediate visual explorations with transparent functional semantics.

Limitations in Cognitive Alignment

However, we must clarify the scope of our current study's contributions. At this stage, our framework primarily serves as a biologically inspired computational model rather than a rigorously validated cognitive theory. Although the exploratory trajectories of our model resemble human center-periphery visual scanning patterns, the lack of quantitative validation of these results remains a key limitation of the present work. Future work should involve quantitative evaluation of these behaviors using metrics such as ScanMatch or KL divergence. We also plan to investigate whether our adaptive termination mechanism correlates with human reaction times under varying cognitive loads. Ultimately, we view this work as an exploratory bridge: integrating cognitive science into practical models of perceptual processes to provide new insights for research on resource-constrained perception.

Conclusion

This paper introduces a memory-driven active perception framework that shifts visual understanding from passive, dense processing to a sequential hypothesis-verification loop. By integrating internal memory priors with uncertainty-driven active sampling, our model adaptively allocates computational resources on demand. Experiments demonstrate that this approach maintains perceptual performance comparable to dense models while significantly reducing inference time and token consumption, particularly in high-complexity scenarios. Ultimately, this framework provides a computable bridge between cognitive theories of active perception and the development of resource-rational visual agents.

Acknowledgement

We would like to express our gratitude to MetaSci (26255817624) and Chengyue Zhao for their assistance in the topic selection, scheme design and experimental analysis.

References

Bajcsy, R., Aloimonos, Y., & Tsotsos, J. K. (2018). Revisiting active perception. *Autonomous Robots*, *42*(2), 177–196.

Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Dabis, J., Finn, C., Gopalakrishnan, K., Hausman, K., Herzog, A., Hsu, J., et al. (2022). Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*.

Callaway, F., van Opheusden, B., Gul, S., Das, P., Krueger, P. M., Griffiths, T. L., & Lieder, F. (2022). Rational use of cognitive resources in human planning. *Nature human behaviour*, *6*(8), 1112–1125.

Crayen, M. A., Treue, S., & Esghaei, M. (2024). Interactions between the fovea and the periphery shape misbinding of visual features in a continuous report paradigm. *Scientific Reports*, *14*(1), 28381.

Dosovitskiy, A. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Hurst, A., Lerer, A., Goucher, A. P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al. (2024). Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Lanillos, P., Meo, C., Pezzato, C., Meera, A. A., Baioumy, M., Ohata, W., Tschantz, A., Millidge, B., Wisse, M., Buckley, C. L., et al. (2021). Active inference in robotics and artificial agents: Survey and challenges. *arXiv preprint arXiv:2112.01871*.

Lee, D. G., Daunizeau, J., & Pezzulo, G. (2023). Evidence or confidence: What is really monitored during a decision? *Psychonomic Bulletin & Review*, *30*(4), 1360–1379.

Lemmel, J., & Grosu, R. (2025). Real-time recurrent reinforcement learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, *39*(17), 18189–18197.

Li, M., Liu, H., Wang, Y., & Zheng, H.-T. (2021). Bilingual self-attention network: Generating headlines for online linguistic questions. *2021 International Joint Conference on Neural Networks (IJCNN)*, 1–8. https://doi.org/10.1109/IJCNN52387.2021.9533288

Li, M., Shi, X., Leng, H., Zhou, W., Zheng, H.-T., & Zhang, K. (2023). Learning semantic alignment with global modality reconstruction for video-language pre-training towards retrieval. *Proceedings of the AAAI Conference on Artificial Intelligence*, *37*(1), 1377–1385. https://doi.org/10.1609/aaai.v37i1.25222

Li, M., Wang, L., Liu, H., Wang, W., & Zheng, H.-T. (2022). Context reasoning attention network: Generating plausible distractors for multi-choice questions. In E. Pimenidis, P. Angelov, C. Jayne, A. Papaleonidas, & M. Aydin (Eds.), *Artificial neural networks and machine learning – icann 2022* (pp. 593–604). Springer Nature Switzerland.

Li, Z., Yang, Y., Liu, X., Zhou, F., Wen, S., & Xu, W. (2017). Dynamic computational time for visual attention. *Proceedings of the IEEE international conference on computer vision workshops*, 1199–1209.

Liang, Y., Ge, C., Tong, Z., Song, Y., Wang, J., & Xie, P. (2022). Not all patches are what you need: Expediting vision transformers via token reorganizations. *arXiv preprint arXiv:2202.07800*.

Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2023). Visual instruction tuning. *Advances in neural information processing systems*, *36*, 34892–34916.

Liu, H., Li, M., Wang, Y., Chen, W., & Zheng, H.-T. (2021). Spherecf: Sphere embedding for collaborative filtering. In T. Mantoro, M. Lee, M. A. Ayu, K. W. Wong, & A. N. Hidayanto (Eds.), *Neural information processing* (pp. 570–583). Springer International Publishing.

Masís, J. A., Melnikoff, D., Barrett, L. F., & Cohen, J. (2023). When to choose: Information seeking in the speed-accuracy tradeoff. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *45*(45).

Min, J., Zhao, Y., Luo, C., & Cho, M. (2022). Peripheral vision transformer. *Advances in Neural Information Processing Systems*, *35*, 32097–32111.

Purwono, P., Ma'arif, A., Rahmaniar, W., Fathurrahman, H. I. K., Frisky, A. Z. K., & ul Haq, Q. M. (2022). Understanding of convolutional neural network (cnn): A review. *International Journal of Robotics and Control Systems*, *2*(4), 739–748.

Rao, Y., Zhao, W., Liu, B., Lu, J., Zhou, J., & Hsieh, C.-J. (2021). Dynamicvit: Efficient vision transformers with dynamic token sparsification. *Advances in neural information processing systems*, *34*, 13937–13949.

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural computation*, *20*(4), 873–922.

Ren, P., Xiao, Y., Chang, X., Huang, P.-Y., Li, Z., Gupta, B. B., Chen, X., & Wang, X. (2021). A survey of deep active learning. *ACM computing surveys (CSUR)*, *54*(9), 1–40.

Rothkopf, C. A., & Hayhoe, M. M. (2025). Computational elements of natural vision. *Journal of Vision*, *25*(12), 4–4.

Roy, K., Jaiswal, A., & Panda, P. (2019). Towards spike-based machine intelligence with neuromorphic computing. *Nature*, *575*(7784), 607–617.

Shang, J., Kahatapitiya, K., Li, X., & Ryoo, M. S. (2022). Starformer: Transformer with state-action-reward representations for visual reinforcement learn-

ing. *European conference on computer vision*, 462–479.

Shang, J., & Ryoo, M. S. (2023). Active vision reinforcement learning under limited visual observability. *Advances in Neural Information Processing Systems*, *36*, 10316–10338.

Sharafeldin, A., Imam, N., & Choi, H. (2024). Active sensing with predictive coding and uncertainty minimization. *Patterns*, *5*(6).

Stockart, F., Msheik, R., Robin, A., Jurkovičová, L., Goueytes, D., Rouy, M., Mareček, R., Hoffmann, D., Mudrik, L., Roman, R., et al. (2025). Cortical evidence accumulation for visual perception occurs irrespective of reports. *Nature Communications*, *16*(1), 8458.

Wang, L., Li, M., & Zheng, H.-T. (2023). Rethinking rule-based approaches in session-based recommendation. *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. https://doi.org/10.1109/ICASSP49357.2023.10094851

Yewbrey, R., & Kornysheva, K. (2024). The hippocampus preorders movements for skilled action sequences. *Journal of Neuroscience*, *44*(45).

Yildirim, I. (2025). New perspectives in computational modeling of human attention. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *47*.

Zhang, Z., Zhang, A., Li, M., Zhao, H., Karypis, G., & Smola, A. (2023). Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*.