This article has been accepted for publication in IEEE Intelligent Systems. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/MIS.2026.3684015

Metacognitive Artificial Intelligence in Vision Foundation Models: Research Challenges

Shahriar Rifat, *Northeastern University, Boston, MA, 02115 USA*

A.Q.M. Sazzad Sayyed, *Northeastern University, Boston, MA, 02115 USA*

Nathaniel D. Bastian, *United States Military Academy, West Point, NY, 10996 USA*

Francesco Restuccia, *Northeastern University, Boston, MA, 02115 USA*

Abstract—*The adoption of Vision Foundation Models (VFMs) in high stakes scenarios has spurred the demand for task specific, high performance models. However, the lack of explainability of VFMs makes it challenging to ensure safety, reliability, and resilience across tasks when facing data distributions different from those seen during training. Recently, approaches based on metacognition, the human ability to regulate cognitive processes, have emerged as a promising way to understand these large models. This paper surveys the interdisciplinary connection between metacognition and state of the art VFMs, and further examines its relationship with knowledge distillation (KD), a widely used technique in VFMs. The paper concludes by outlining possible avenues for future research on the topic.*

Introduction

Metacognition, or "cognition about cognition," was introduced by Flavell in 1979 and later expanded by Brown in 1987.[1, 2] In psychology, it refers to an individual's ability to monitor, regulate, and adapt their cognitive processes. Although metacognition has been extensively studied in diverse fields, including schizophrenia research, programming education, manufacturing, aerospace, and military applications, its definitions and applications vary.[10] Here, we draw upon Flavell's framework, which conceptualizes metacognition as comprising four interrelated elements: knowledge, experience, goals, and strategies. Together, these elements support self reflection, adaptive learning, and informed decision making. Recent work formalizes metacognition as a system-level capability in AI.[22, 23]

A similar paradigm applies to VFMs, which are large scale models trained on vast multi-modal data to learn general-purpose visual representations. Just as metacognition improves human cognition, it allows artificial intelligence (AI) systems to self-monitor, detect errors, adapt learning strategies, and optimize performance. Although interest in metacognitive AI has fluctuated, the rise of Artificial General Intelligence (AGI) has rekindled its significance.[3] This is particularly prevalent in agentic systems and generative AI, where adaptability and self-improvement are crucial.[4]

XXXX-XXX © IEEE
Digital Object Identifier 10.1109/XXX.0000.0000000

While systems like ChatGPT and Deepseek AI already employ metacognitive strategies to refine reasoning and outputs, their application to VFMs remains underexplored.[5] While highly cognitive agents can reflect upon and improve their own knowledge, inducing this behavior across task-specific VFMs is highly challenging in the absence of external feedback. KD provides an effective computational mechanism to emulate this process, allowing a model to refine its internal representations through guidance from a teacher model. In this sense, KD can be viewed as an operational analogue of metacognitive learning, enabling VFMs to inherit not only structured knowledge but also self-evaluative and generalization behaviors across model scales.

This work examines metacognitive AI in VFMs, particularly its role in improving explainability, uncertainty estimation, adaptive learning, and error detection. Our key contributions are: (i) unifying metacognitive approaches in AI to enhance self-awareness and interpretability in VFMs; (ii) bridging metacognition and KD, highlighting their intersection in self-regulating learning strategies, and (iii) identifying research challenges and future directions to develop more robust VFMs.

Metacognitive Framework for Vision Foundation Models

To examine VFMs from a metacognitive perspective, Figure 1 presents our adapted framework. Building on Flavell's four components, we emphasize the components that best explain key functional properties of current VFMs. In particular, we divide metacognitive strategies into monitoring and control, and omit explicit

This article has been accepted for publication in IEEE Intelligent Systems. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/MIS.2026.3684015

goals from the figure to concentrate on mechanisms that are directly reflected in existing model behavior. Although the overall aim is reliable prediction grounded in sound semantic reasoning, the framework highlights the components and processes that support this aim in practice. We then relate these metacognitive components to core properties of VFMs, showing how their interactions can improve performance, adaptability, and trustworthiness across different applications.

Metacognitive Knowledge refers to an awareness of one's cognitive processes. In VFMs, this corresponds to understanding their representational structure, learning capabilities, and limitations. Given their large parameter spaces and black-box nature, developing such knowledge requires interpretable modeling approaches. Techniques such as modeling patch embedding distributions in VFMs or linearly combining concept prototypes enhance interpretability by revealing structured semantic representations.[11] Post hoc methods, including saliency maps and attention visualization, further support explanation of inference mechanisms.[14] Beyond post hoc analysis, structured concept-based architectures strengthen metacognitive knowledge by introducing explicit semantic layers.For example, VLG-CBM integrates visually grounded concepts, such as objects and attributes detected by an open-vocabulary detector, as an intermediate concept layer for image classification[21]. By incorporating an interpretable concept bottleneck, the model exposes semantically meaningful reasoning pathways, improving transparency and structured decision making. Additionally, VFMs trained to detect object recognition errors without label access exemplify self-assessment capabilities.[15] Together, these approaches enhance explainability and representation awareness, which are essential for reliable adaptation in complex computer vision tasks. Figure 1 illustrates how metacognitive knowledge, along with monitoring, experience, and control, maps onto key functional properties of Vision Foundation Models, supporting explainability, adaptation, and error awareness.

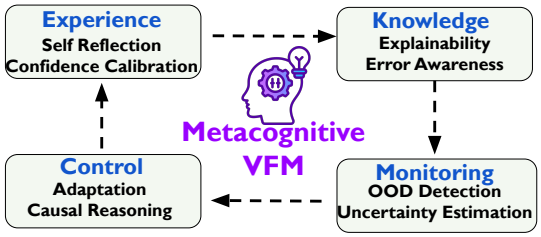

FIGURE 1. Metacognition in Vision Foundation Models.

Metacognitive Monitoring enables assessment of input reliability and detects domain shifts. This connects to VFMs in two ways. First, VFMs, due to their rich feature representations, can help verify whether an input aligns with the training distribution of an AI system, thus working as monitors themselves. Second, VFMs themselves require monitoring post-deployment, especially after fine-tuning for specific tasks. This monitoring can be used for both checking inputs for distribution shifts (out-of-distribution (OOD) detection) and misclassification detection (uncertainty estimation). Recent work on large language models further distinguishes metacognitive ability from task-level cognition, providing quantitative perspectives on monitoring.[24]

As VFMs are adapted to downstream tasks through task-specific fine-tuning, their ability to detect OOD inputs often degrades due to overfitting to task-aligned features. This loss of generality weakens the broad representations that are essential for robust OOD detection. Parameter-efficient tuning strategies help mitigate this effect by preserving shared and semantically rich features during adaptation. For example, fine-tuning Vision Language Models (VLMs) such as CLIP using multimodal concept matching maintains semantic alignment across modalities, leading to improved OOD detection performance.[12] Consistent with this observation, benchmark studies show that although fine-tuning improves task-specific accuracy, it can reduce the ability to recognize novel or shifted inputs if robustness is not explicitly considered.[13] Together, these findings highlight the need to balance task adaptation with the retention of generalizable representations in order to ensure reliable metacognitive monitoring capabilities in VFMs.

To present metacognitive monitoring in VFMs more intuitively, we draw an explicit parallel with human metacognitive monitoring, as illustrated in Figure 2. In humans, metacognitive monitoring involves estimating confidence in one's knowledge, recognizing incorrect information, and identifying situations where knowledge is insufficient, commonly referred to as knowledge gaps. Analogously, in VFMs, uncertainty estimation supports reliable inference, misclassification detection enables error awareness, and OOD detection indicates when an input lies outside the model's learned knowledge. Similar mechanisms have emerged in large language models (LLMs), where self-consistency checks improve inference reliability, hallucination detection mitigates erroneous outputs, and OOD detection facilitates the identification of knowledge gaps.

Metacognitive Control refers to the ability to regulate

This article has been accepted for publication in IEEE Intelligent Systems. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/MIS.2026.3684015

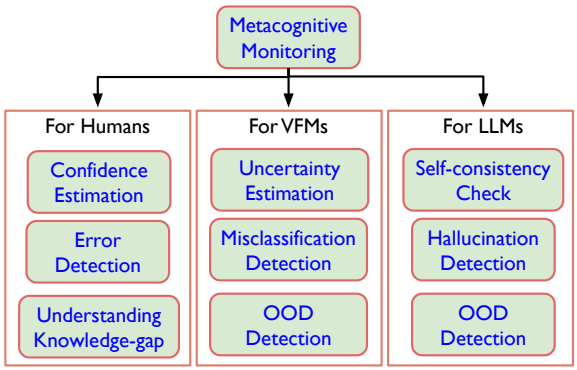

FIGURE 2. Mapping among metacognitive monitoring in humans, VFMs, and LLMs.

and modify cognitive processes based on internally monitored signals. Unlike metacognitive monitoring, which estimates the current cognitive state, metacognitive control involves taking deliberate actions to adjust reasoning, learning strategies, or resource allocation in order to improve performance and reliability. In the context of VFMs, this may include dynamically adapting feature usage, modifying learning dynamics, or reallocating computational effort based on uncertainty, task complexity, or performance feedback. Such control mechanisms become particularly important when VFMs are adapted to downstream tasks through fine-tuning, where maintaining generalization and reliability is challenging. A representative example is the self-distillation mechanism used in DINO,[20] where a model learns from soft targets generated by an exponential moving average of its own parameters. This process enables the model to refine its representations using informative internal guidance rather than relying solely on hard labels. However, existing approaches do not explicitly ground this adaptive behavior in structured knowledge about the learned representations themselves. Incorporating more explicit representational awareness into such control mechanisms remains an important direction for the next generation of VFMs.

Metacognitive Experience refers to the internal evaluative signals that arise from a model's inference process and shape how it interprets its own cognitive state. Unlike metacognitive monitoring, which estimates properties such as uncertainty or distribution shift, metacognitive experience concerns how the model internally evaluates those signals and whether they influence subsequent reasoning or adaptation. In humans, such experiences include feelings of confidence, perceived difficulty, or a sense of knowing.

In VFMs, metacognitive experience can be interpreted as calibrated confidence, perceived prediction difficulty, or reflective reassessment of outputs. Confidence calibration plays a central role in this process, as it determines whether predicted probabilities meaningfully reflect reliability. Recent work shows that fine-tuning transformer-based foundation models can significantly alter calibration behavior. For example, Bayesian parameter-efficient fine-tuning mitigates underconfidence and improves reliability in few-shot settings.[16] Similarly, calibrated robust fine-tuning (CaRot) enhances both OOD generalization and confidence alignment in VLMs.[17]

Beyond calibration, self-reflection represents a stronger form of metacognitive experience, where a model revisits and refines its own outputs based on internal evaluative cues. While LLMs have demonstrated improvements through reflective reasoning strategies,[18] comparable mechanisms remain largely unexplored for VFMs. Existing work on model introspection primarily analyzes internal representations or decision pathways,[19] but does not explicitly frame these processes as evaluative signals that guide subsequent inference. Developing self-reflective VFMs that leverage internal confidence and difficulty signals to dynamically refine predictions could enable a clearer separation between monitoring and experience, ultimately improving reliability and adaptability in real-world deployments.

Knowledge Distillation in Metacognitive VFMs

Building on the metacognitive framework described earlier, we reinterpret KD as a computational mechanism that aligns with metacognitive control and self-reflection in VFMs. We believe KD will play a critical role in enhancing the efficiency, adaptability, and generalization of VFMs by distilling structured knowledge into more task-oriented and informative representations.[6] This process allows VFMs to acquire structured knowledge representations, thus leading to more efficient self-supervised learning and better feature abstraction.[7] From a metacognitive perspective, KD can act as self-reflection on current knowledge and sort out irrelevant information thus enhancing self-monitoring and adaptation in VFMs. Self-distillation methods like DINO allow VFMs to refine internal representations using earlier predictions as guidance, serving as an operational analogue of metacognitive reflection, enabling VFMs to evaluate and refine learned representations over successive training cycles.[20] Progressive distillation, where a student model undergoes multiple iterations of knowledge refinement, further reinforces self-improving learning paradigms.[9] However, challenges

remain in integrating explicit self-awareness mechanisms into KD for VFMs. Current KD approaches often inherit biases from the teacher model, lack mechanisms for interpretable feature selection, and struggle with adapting to novel, OOD scenarios.[8] Future research should explore self-reflective distillation where VFMs actively assess their learning trajectory and uncertainty levels to optimize knowledge transfer dynamically. Additionally, uncertainty-aware distillation techniques could allow VFMs to focus on challenging instances, further aligning with metacognitive self-regulation principles. By embedding metacognitive control and monitoring into KD, VFMs can evolve into more self-aware, efficient, and generalizable AI systems that are capable of adaptive learning, robust decision making, and improved inference reliability.

Future Research Directions

While metacognition has gained traction in language models through chain of thought reasoning, its integration into VFMs remains substantially unexplored. Although VFMs enhance general AI performance, their ability to assess prediction reliability (e.g., confidence calibration) and refine outputs (e.g., self-reflection) is still in its infancy. In short, we highlight the following key research opportunities:

Confidence Calibration: Fine tuning often degrades VFMs' confidence calibration. Future research could explore *meta learning strategies* to preserve calibration post fine-tuning or *self-assessment modules* that analyze internal features to detect anomalies.

Metacognitive Feedback Loops: Similar to human learning, VFMs could benefit from a *human in the loop framework*, incorporating real time feedback for more reliable adaptation. Unlike standard lifelong learning, this approach actively integrates external validation into the learning process.

Limited Self Reflection: While self reflection has improved reasoning in LLMs, its application to VFMs remains unclear.[18] Self-consistency and rational reflection could refine vision-based predictions, especially in high-stakes tasks like medical imaging or autonomous navigation. A lack of these mechanisms leads to persistent errors in complex scenarios.

Task Specific Adaptation: VFMs require substantial computational resources due to large-scale pretraining and high inference cost. Metacognitive control could enable *adaptive feature extraction*, allocating computational resources based on task complexity to improve efficiency. A key research question is whether VFMs can dynamically select relevant components through post hoc dynamic architectures that support metacognitive adaptation.

Conclusion

This paper has discussed the integration of metacognitive principles into VFMs to enhance their performance, adaptability, and reliability. By aligning metacognitive factors such as experience, monitoring, knowledge, and self-reflection with the core concepts of VFMs, we highlight the potential for improving model reliability, decision making, and overall generalization. While metacognitive approaches have been well researched in language models, their generalization to computer vision is still underdeveloped. Future research in areas such as confidence calibration, self-monitoring, and adaptive learning algorithms can pave the way for more powerful and efficient VFMs to solve real-world problems across different applications.

REFERENCES

[1] Flavell, John H. "Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry." American psychologist 34, no. 10 (1979): 906. (Journal)

[2] Brown, A. "Metacognition, executive control, self-regulation, and other more mysterious mechanisms." Metacognition, motivation, and understanding/Lawrence Erlbaum Associates (1987). (Journal)

[3] Schmill, M., Tim Oates, Michael L. Anderson, Darsana Josyula, Don Perlis, Shomir Wilson, and Scott Fults. "The role of metacognition in robust AI systems." In Workshop on Metareasoning at the Twenty-Third AAAI Conference on Artificial Intelligence. 2008.

[4] "The Metacognitive Demands and Opportunities of Generative AI." 2024. Proceedings of the ACM on Human-Computer Interaction 8 (CSCW2): Article 290.

[5] Wu, Siwei, Zhongyuan Peng, Xinrun Du, Tuney Zheng, Minghao Liu, Jialong Wu, Jiachen Ma et al. "A Comparative Study on Reasoning Patterns of OpenAI's o1 Model." CoRR (2024).

[6] Hinton, Geoffrey, Oriol Vinyals, and Jeff Dean. "Distilling the Knowledge in a Neural Network." *Proceedings of the NIPS Deep Learning and Representation Learning Workshop*, 2015.

[7] Yuan, L., Tay, F. E. H., Li, G., Wang, T., & Feng, J. (2020). Revisiting Knowledge Distillation via Label Smoothing Regularization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 3903–3911).

[8] Yang, C., Xie, L., & Liu, X. (2019). Snapshot Distillation: Teacher-Student Optimization in One Generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 2859–2868).

[9] Mirzadeh, S. I., Farajtabar, M., Li, A., & Ghasemzadeh, H. (2020). Improved Knowledge Distillation via Teacher Assistant. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 34, No. 04, pp. 5191–5198).

[10] Moritz, Steffen, and Paul H. Lysaker. "Metacognition–what did James H. Flavell really say and the implications for the conceptualization and design of metacognitive interventions." Schizophrenia Research 201 (2018): 20-26.

[11] Wang, Hengyi, Shiwei Tan, and Hao Wang. "Probabilistic Conceptual Explainers: Trustworthy Conceptual Explanations for Vision Foundation Models." Forty-first International Conference on Machine Learning.

[12] Borlino, Francesco Cappio, Lorenzo Lu, and Tatiana Tommasi. "Foundation Models and Fine-Tuning: A Benchmark for Out Of Distribution Detection." IEEE Access (2024).

[13] Ming, Yifei, and Yixuan Li. "How Does Fine-Tuning Impact Out-of-Distribution Detection for Vision-Language Models?." International Journal of Computer Vision 132, no. 2 (2024): 596-609.

[14] Kazmierczak, Rémi, et al. "Explainability for Vision Foundation Models: A Survey." arXiv preprint arXiv:2501.12203 (2025).

[15] Berke, Marlene, et al. "MetaCOG: A Hierarchical Probabilistic Model for Learning Meta-Cognitive Visual Representations." The 40th Conference on Uncertainty in Artificial Intelligence

[16] Pandey, Deep Shankar, Spandan Pyakurel, and Qi Yu. "Be Confident in What You Know: Bayesian Parameter Efficient Fine-Tuning of Vision Foundation Models." In The Thirty-eighth Annual Conference on Neural Information Processing Systems. 2024.

[17] Oh, Changdae, Mijoo Kim, Hyesu Lim, Junhyeok Park, Euiseog Jeong, Zhi-Qi Cheng, and Kyungwoo Song. "Towards Calibrated Robust Fine-Tuning of Vision-Language Models." In NeurIPS 2023 Workshop on Distribution Shifts: New Frontiers with Foundation Models.

[18] Wang, Xuezhi, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. "Self-Consistency Improves Chain of Thought Reasoning in Language Models." In The Eleventh International Conference on Learning Representations.

[19] Prabhushankar, Mohit, and Ghassan AlRegib. "Introspective Learning: A Two-Stage Approach for Inference in Neural Networks." In Advances in Neural Information Processing Systems (NeurIPS), New Orleans, LA, November 29 – December 1, 2022.

[20] Caron, Mathilde, et al. "Emerging properties in self-supervised vision transformers." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.

[21] Srivastava, Divyansh, Ge Yan, and Lily Weng. "Vlg-cbm: Training concept bottleneck models with vision-language guidance." Advances in Neural Information Processing Systems 37 (2024): 79057-79094.

[22] Wei, Hua, et al. "Metacognitive AI: Framework and the case for a neurosymbolic approach." International Conference on Neural-Symbolic Learning and Reasoning. Cham: Springer Nature Switzerland, 2024.

[23] Shakarian, Paulo. "Toward Artificial Metacognition." AAAI Conference on Artificial Intelligence, 2026.

[24] Wang, Guoqing, et al. "Decoupling metacognition from cognition: a framework for quantifying metacognitive ability in LLMs." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 39. No. 24. 2025.

ACKNOWLEDGMENTS

The views expressed in this paper are those of the authors and do not reflect the official policy or position of the United States Military Academy, Department of the Army, Department of Defense, or U.S. Government.

Shahriar Rifat is a Ph.D. fellow at the Institute for the Wireless Internet of Things (WIoT), Northeastern University, where he is pursuing his Ph.D. in Electrical and Computer Engineering (ECE). His research focuses on real-time, secure, and efficient dynamic adaptation of deep neural networks for resource-constrained applications. He earned his Bachelor's degree in Electrical and Electronic Engineering from the Bangladesh University of Engineering and Technology (BUET). For inquiries, he can be reached at rifat.s@northeastern.edu.

A. Q. M. Sazzad Sayyed is a Ph.D. candidate at the Institute for the Wireless Internet of Things (WIoT), Northeastern University, pursuing a doctorate in Electrical and Computer Engineering. His research focuses on designing secure and robust deep learning algorithms for resource-constrained applications, emphasizing interpretability. He completed his Bachelor's study in Electrical and Electronic Engineering from the Bangladesh University of Engineering and Technology (BUET). He can be reached at sayyed.a@northeastern.edu.

Nathaniel D. Bastian is currently an Assistant Professor in the Department of Electrical Engineering & Computer Science at the United States Military Academy at West Point, as well as Deputy Director of the Robotics Research Center. Dr. Bastian's main research focus is designing and developing intelligent autonomous systems for cybersecurity, networking, and other C5ISRT applications in resource-constrained settings, as well as neuro-symbolic artificial intelligence for representa-

tion, learning, inferencing, reasoning, and metacognition. He is a Senior Member of IEEE and INFORMS. Contact him at nathaniel.bastian@westpoint.edu.

Francesco Restuccia is currently an Assistant Professor in the Department of Electrical and Computer Engineering at Northeastern University. Dr. Restuccia's main research focus is addressing the fundamental challenges related to edge-assisted data-driven resilient mobile systems. Dr. Restuccia has received the ONR Young Investigator Award, the AFOSR Young Investigator Award, the ACM SIGMOBILE Research Highlights Award, the Mario Gerla Award in Computer Science, as well as best paper awards at IEEE INFOCOM and IEEE WOWMOM. Dr. Restuccia is in the editorial board of Computer Networks, IEEE Transactions on Cognitive Communications and Networking and IEEE Transactions on Mobile Computing. He is a Senior Member of IEEE and ACM. Contact him at f.restuccia@northeastern.edu.