Second Workshop on Metacognitive Prediction of AI Behavior

# 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 (VFM) in high-stakes scenarios has spurred the the demand for task-specific high-performance models. On the other hand, the lack of explainability of VFMs makes it challenging to ensure that these systems remain safe, reliable and resilient when encountering data that has a different distribution from the one encountered during training. Recently, approaches based on metacognition – the human ability to regulate cognitive processes – has emerged as a way to understand 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 defining 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.[16–20] Here, we adopt Flavell's framework, which categorizes metacognition into knowledge, experiences, goals, and strategies, allowingg self-reflection, adaptive learning, and improved decision-making.

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.[3–6] This is particularly prevalent in agentic systems and generativeAI, where adaptability and self-improvement are crucial.[7, 8] While systems like ChatGPT and Deepseek AI already employ metacognitive strategies to refine reasoning and outputs, their application to VFMs remains largely unexplored.[9, 10] Furthermore, KD has been explored as a means to improve self-supervised learning and adaptation in VFM by transferring structured knowledge from a teacher model to a student model. 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 observeVFMs through the lens of metacognition, Figure 1 presents our framework - integrating metacognition into VFMs to highlight its importance in enhancing model performance, adaptability, and trustworthiness across diverse applications. In this framework, we adapt the original four components of metacognition as described by Flavell and map aspects of VFMs to these components. Our mapping emphasizes the

XXXX-XXX © 2025 IEEE
Digital Object Identifier 10.1109/XXX.0000.0000000

role of these metacognitive components in improving inference reliability and decision-making with VFMs.

**Metacognitive Knowledge** refers to an awareness of one's cognitive processes. In VFMs, this translates to understanding their own learning capabilities and limitations. Given their large parameter spaces and black-box nature, acquiring this knowledge requires explainable models. Approaches such as modeling patch embedding distributions in VFMs or linearly combining concept prototypes enhance interpretability.[22, 23] Post-hoc methods, including saliency maps and attention visualization, further aid in explaining inference mechanisms.[27] Additionally, VFMs trained to detect object recognition errors without label access exemplify self-assessment capabilities.[28] Together, these techniques support explainability and error estimation—key elements for effective VFM adaptation in complex computer vision tasks.

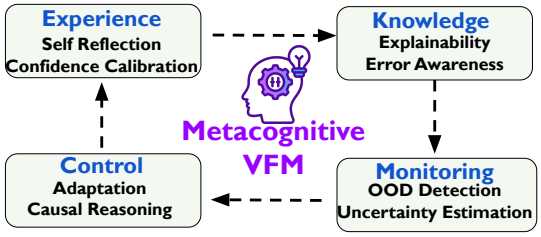


**FIGURE 1.** Metacognition in Vision Foundation Models.

**Metacognitive Monitoring** enables to assess input reliability and detect 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.[21] 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).

As VFMs undergo fine-tuning before being deployed to specific tasks, it reduces OOD detection performance of VFMs by overfitting to learned features.[24] Strategies such as parameter-efficient tuning help maintain broad feature representations crucial for OOD robustness. For example, fine-tuning vision language models like CLIP with multimodal concept matching preserves semantic richness, thus improving OOD detection.[25] Benchmark studies also show that while fine-tuning improves classification accuracy, it may affect the ability to recognize novel inputs unless explicitly addressed.[26] These findings underscore the importance of balancing task adaptation with generalizable feature retention to maintain reliable OOD detection.

Metacognitive monitoring in humans is well aligned with metacognitive monitoring for VFMs, as shown in Figure 2. For humans, metacognitive monitoring involves estimating one's confidence in their knowledge, recognizing when the provided information is incorrect, and identifying situations where they lack knowledge about a specific topic—commonly referred to as a knowledge gap. Similarly, for VFMs, uncertainty estimation ensures reliable inference, misclassification detection identifies errors, and OOD detection helps determine when the VFM lacks sufficient knowledge about a given input. The same applies to large language models (LLMs): self-consistency checks promote confident inferences, hallucination detection prevents the generation of incorrect outputs, and OOD detection facilitates the identification of knowledge gaps.

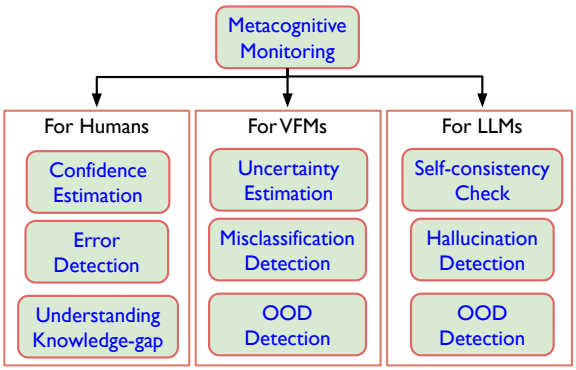


**FIGURE 2.** Mapping among metacognitive monitoring in humans, VFMs, and LLMs.

**Metacognitive Control** refers to an individual's ability to obtain a concise summary of its cognitive state as well as its ability to monitor such knowledge seamlessly. Ultimately, it helps to adapt to the optimum learning strategy dynamically by controlling the cognitive process, which we term metacognitive control. The control over this learning process becomes increasingly more necessary as such VFMs are used in specific downstream tasks through fine-tuning. A notable example of metacognitive control in VFMs is the self-distillation mechanism used in DINO.[36] In this framework, a VFM learns from soft targets generated by an exponential moving average of its own parameters, which serve as a teacher model, offering more informative feedback than hard labels. Although, it shows how metacognitive control is benificial to

learning process of VFMs, the actual process is not guided by clear and concise knowledge of the learned representation itself which can be a possible venue of improvement for the next generation of VFMs.

**Metacognitive Experience** encompasses the feelings and judgments that arise during cognitive tasks, such as confidence in an answer or awareness of difficulty. In VFMs, this translates to assessing prediction certainty, recognizing uncertainty, and adapting accordingly. Confidence calibration and self-reflection are key mechanisms for enabling this capability. Recent studies have shown that fine-tuning large transformer-based foundation models impacts confidence calibration. For example, a Bayesian parameter-efficient fine-tuning framework addresses under-confidence issues, improving reliability in few-shot settings.[29] Similarly, calibrated robust fine-tuning (CaRot) enhances both out-of-distribution accuracy and confidence calibration in vision-language models.[30] Beyond confidence calibration, self-reflection allows models to iteratively refine their outputs, thus improving decision reliability. While LLMs have leveraged self-reflection for enhanced robustness, similar techniques remain unexplored for VFMs.[31] Existing work on model introspection focus on analyzing internal states and decision processes.[32–35] However, none specifically target VFMs. Developing self-reflective VFMs could bridge this gap, enabling models to assess their own reliability and adapt dynamically – thus advancing adaptability in real-world applications.

## Knowledge Distillation in Metacognitive VFMs

We believe KD will play a critical role in enhancing the efficiency, adaptability, and generalization of VFMs by enabling a student model to learn from a teacher model through soft labels rather than hard classifications.[11] This process allows VFMs to acquire structured knowledge representations, thus leading to more efficient self-supervised learning and better feature abstraction.[12] From a metacognitive perspective, KD enhances self-monitoring and adaptation in VFMs. Self-distillation techniques, such as DINO, enable models to refine their own knowledge representations by treating their earlier predictions as a form of internal guidance.[36] This self-assessment mechanism mimics metacognitive reflection, allowing VFMs to evaluate and refine learned representations over successive training cycles. Progressive distillation, where a student model undergoes multiple iterations of knowledge refinement, further reinforces self-improving learning paradigms.[14] 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.[13] 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 hard-to-learn instances, further aligning with metacognitive self-regulation principles.[15] 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.[31] Self-consistency and rationale 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 are computationally expensive. Metacognitive control could enable *adaptive feature extraction*, optimizing energy efficiency by adjusting computational resources based on task complexity. A key research question is whether VFMs can dynamically select relevant components to improve efficiency through post-hoc dynamic neural architectures capable of metacognitive adaptation.

## Conclusion

This paper has discussed the integration of metacognitive principles into VFMs to enhance their performance, adaptability, and reliability. With the alignment of metacognitive factors – experience, monitoring, knowledge, and self-reflection – with core concepts of VFMs, we have highlighted tremendous potential in model reliability augmentation, 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." Metacogniton, 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] Schaeffer, Rylan. 2021. "An Algorithmic Theory of Metacognition in Minds and Machines." arXiv preprint arXiv:2111.03745.

[5] Wei, Hua, Paulo Shakarian, Christian Lebiere, Bruce Draper, Nikhil Krishnaswamy, and Sergei Nirenburg. 2024. "Metacognitive AI: Framework and the Case for a Neurosymbolic Approach." arXiv preprint arXiv:2406.12147.

[6] Cox, Michael, Zahiduddin Mohammad, Sravya Kondrakunta, Ventaksamapth Raja Gogineni, Dustin Dannenhauer, and Othalia Larue. 2022. "Computational Metacognition." arXiv preprint arXiv:2201.12885.

[7] Toy, Jason, Josh MacAdam, and Phil Tabor. 2024. "Metacognition is All You Need? Using Introspection in Generative Agents to Improve Goal-Directed Behavior." arXiv preprint arXiv:2401.10910.

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

[9] 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).

[10] Guo, Daya, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu et al. "Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning." arXiv preprint arXiv:2501.12948 (2025).

[11] 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.

[12] 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).

[13] 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).

[14] 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).

[15] Finn, C., Abbeel, P., & Levine, S. (2017). Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. Proceedings of the 34th International Conference on Machine Learning (ICML), 1126–1135.

[16] Izzo, D., M. Märtens, and B. Pan. "A survey on artificial intelligence trends in spacecraft guidance dynamics and control. Astrodynamics 3 (4): 287–299." 2019

[17] Li, Bo-hu, et al. "Applications of artificial intelligence in intelligent manufacturing: a review." Frontiers of Information Technology and Electronic Engineering 18.1 (2017): 86-96

[18] Svenmarck, P., Luotsinen, L., Nilsson, M., Schubert, J.: Possibilities and challenges for artificial intelligence in military applications. In: Proceedings of the NATO Big Data and Artificial Intelligence for Military Decision Making Specialists' Meeting, pp. 1–16 (2018)

[19] Prather, James, et al. "What do we think we think we are doing? Metacognition and self-regulation in programming." Proceedings of the 2020 ACM conference on international computing education research. 2020.

[20] 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.

[21] Keser, Nert, Halil Ibrahim Orhan, Niki Amini-Naieni, Gesina Schwalbe, Alois Knoll, and Matthias Rottmann. "Benchmarking Vision Foundation Models for Input

Monitoring in Autonomous Driving." arXiv preprint arXiv:2501.08083 (2025).
[22] Turbé, Hugues, et al. "ProtoS-ViT: Visual foundation models for sparse self-explainable classifications." arXiv preprint arXiv:2406.10025 (2024).
[23] Wang, Hengyi, Shiwei Tan, and Hao Wang. "Probabilistic Conceptual Explainers: Trustworthy Conceptual Explanations for Vision Foundation Models." Forty-first International Conference on Machine Learning.
[24] Vaze, Sagar, Kai Han, Andrea Vedaldi, and Andrew Zisserman. "Open-set recognition: A good closed-set classifier is all you need?." (2021).
[25] Borlino, Francesco Cappio, Lorenzo Lu, and Tatiana Tommasi. "Foundation Models and Fine-Tuning: A Benchmark for Out Of Distribution Detection." IEEE Access (2024).
[26] 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.
[27] Kazmierczak, Rémi, et al. "Explainability for Vision Foundation Models: A Survey." arXiv preprint arXiv:2501.12203 (2025).
[28] Berke, Marlene, et al. "MetaCOG: A Heirarchical Probabilistic Model for Learning Meta-Cognitive Visual Representations." The 40th Conference on Uncertainty in Artificial Intelligence
[29] 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.
[30] 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.
[31] 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.
[32] Daftry, Shreyansh, Sam Zeng, J. Andrew Bagnell, and Martial Hebert. 2016. "Introspective Perception: Learning to Predict Failures in Vision Systems." arXiv preprint arXiv:1607.08665.
[33] Rosenfeld, Amir, and Shimon Ullman. 2016. "Visual Concept Recognition and Localization via Iterative Introspection." arXiv preprint arXiv:1603.04186
[34] Pardo, Arturo, José A. Gutiérrez-Gutiérrez, José Miguel López-Higuera, Brian W. Pogue, and Olga M. Conde. 2019. "Coloring the Black Box: Visualizing Neural Network Behavior with a Self-Introspective Model." arXiv preprint arXiv:1910.04903.
[35] 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.
[36] Caron, Mathilde, et al. "Emerging properties in self-supervised vision transformers." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.

## 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 Mathematical Sciences at the United States Military Academy at West Point, as well as chief scientist and director, Office of Science & Engineering at the Army Cyber Institute. He received his Ph.D. from the Pennsylvania State University. His primary research interests are artificial intelligence security, assurance and robustness with cybersecurity and command & control military applications. Contact him at nathaniel.bastian@westpoint.edu.

**Francesco Restuccia** is currently an Assistant Pro-

fessor 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.