Matthew Kowal, B.A.Sc, M.Sc, Ph.D Candidate

Building Interpretable and Safe AI Systems

Toronto, Canada (open to relocation)

matt2kowal@gmail.com @MatthewKowal9

mkowal2.github.io linkedin.com/in/mkowal2

Education

2020 – Present **Ph.D. Computer Science, York University** Multi-layer and Temporal Post-Hoc Interpretability for Deep Vision Models.
Supervisor: Dr. Kosta G. Derpanis

2018 – 2020 **M.Sc. Computer Science, Ryerson University** Deep Learning, Computer Vision.
Thesis title: *An Evaluation of Modalities for Action Recognition*.
Supervisors: Dr. Kosta G. Derpanis and Dr. Neil Bruce

2013 – 2017 **B.A.Sc. Applied Mathematics and Engineering, Queens University**
Capstone title: *Region Tracking in an Image Sequence: Preventing Driver Inattention*.
Awarded Keyser Award for best capstone project in discipline.

Selected Experience

2025 – Present **Research Resident @ FAR AI** - AI Safety and mechanistic interpretability research with a focus on LLM persuasion capabilities.

2024 – 2025 **Research Intern @ Ubisoft La Forge** - Conducting research on concept-controllable diffusion models for text-to-human motion generation.

2023 – 2024 **Research Intern @ Toyota Research Institute - Machine Learning Team (Los Altos)** - Conducted research with a focus on video concept-based interpretability for robotic perception (published at CVPR 2024).

2021 – 2024 **Technical Project Lead @ Vector Institute** - Lead team of industry data-scientists in a computer vision project for multi-modal video understanding in collaboration with Intact Insurance, RBC, and Thomson Reuters.

2020 – 2022 **Lead Scientist in Residence @ NextAI** - Lead technical consultant for AI-focused startups. Provided support on the implementation of state-of-the-art deep learning algorithms for various industry applications.

2020 – 2021 **Organizing Chair @ OWCV** - Co-founder and organizing chair of the Ontario Workshop on Computer Vision, a student-focused workshop for computer vision researchers in Ontario. OWCV Website.

2015 – 2018 **Civil/Mechanical Engineer @ Morrison Hershfield** - Conducted bridge inspections in office and on site. Analysis and design of mechanical systems: controls, electrical, HVAC, hydro, fire protection.

Selected Publications

1. Fel, T., Lubana, E. S., Prince, J. S., Kowal, M., Boutin, V., Papadimitriou, I., … Konkle, T. (2025). Archetypal sae: Adaptive and stable dictionary learning for concept extraction in large vision models. *International Conference of Machine Learning*.

2. Kowal, M., Timm, J., Godbout, J.-F., Costello, T., Arechar, A. A., Pennycook, G., … Pelrine, K. (2025). It's the thought that counts: Evaluating the attempts of frontier llms to persuade on harmful topics. *arXiv preprint arXiv:2506.02873*.

3. Thasarathan, H., Forsyth, J., Fel, T., Kowal, M., & Derpanis, K. (2025). Universal sparse autoencoders: Interpretable cross-model concept alignment. *International Conference of Machine Learning*.

4. (Spotlight) Kowal, M., Dave, A., Ambrus, R., Gaidon, A., Derpanis, K. G., & Tokmakov, P. (2024). Understanding video transformers via universal concept discovery. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. Retrieved from https://arxiv.org/abs/2401.10831

5. (Spotlight) Kowal, M., Wildes, R. P., & Derpanis, K. G. (2024). Visual concept connectome (vcc): Open world concept discovery and their interlayer connections in deep models. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. Retrieved from https://arxiv.org/abs/2404.02233

6. Kowal, M., Siam, M., Islam, A., Bruce, N., Wildes, R., & Derpanis, K. (2024). Quantifying and Learning Static vs. Dynamic Information in Deep Spatiotemporal Networks. *Transactions on Pattern Analysis and Machine Intelligence*. Retrieved from https://arxiv.org/abs/2108.09929

7. Chou, S.-H., Kowal, M., Niknam, Y., Moyano, D., Mehdi, S., Pito, R., … Sigal, L. et al. (2023). Multi-modal news understanding with professionally labelled videos (reutersvilnews). In *Canadian AI Conference*.

8. Islam, A., Kowal, M., Jia, S., Derpanis, K., & Bruce, N. (2023). Position, Padding and Predictions: A Deeper Look at Position Information in CNNs. *International Journal of Computer Vision*. Retrieved from https://arxiv.org/abs/2101.12322

9. Islam, A., Kowal, M., Esser, P., Ommer, B., Derpanis, K., & Bruce, N. (2022). Maximize Mutual Shape Information. In *British Machine Vision Conference (BMVC)*.

10. Kowal, M., Siam, M., Islam, A., Bruce, N., Wildes, R., & Derpanis, K. (2022). A Deeper Dive into what Spatiotemporal Models Encode: Static vs. Dynamic Information. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. Retrieved from https://arxiv.org/abs/2206.02846

11. Islam, A., Kowal, M., Derpanis, K., & Bruce, N. (2021). SegMix: Co-occurrence Driven Mixup for Semantic Segmentation and Adversarial Robustness. *The International Journal of Computer Vision (IJCV)*. Retrieved from https://arxiv.org/abs/2108.09929

12. Islam, A., Kowal, M., Esser, P., Jia, S., Ommer, B., Derpanis, K., & Bruce, N. (2021). Shape or Texture: Understanding Discriminative Features in CNNs. In *International Conference on Learning Representations (ICLR)*. Retrieved from https://arxiv.org/abs/2101.11604

13. Islam, A., Kowal, M., Jia, S., Derpanis, K., & Bruce, N. (2021a). Global Pooling, More than Meets the Eye: Position Information is Encoded Channel-Wise in Cnns. In *International Conference on Computer Vision (ICCV)*. Retrieved from https://arxiv.org/abs/2108.07884

14. Islam, A., Kowal, M., Jia, S., Derpanis, K., & Bruce, N. (2021b). Simpler Does It: Generating Semantic Labels with Objectness Guidance. In *British Machine Vision Conference (BMVC)*. Retrieved from https://arxiv.org/abs/2110.10335

15. (Oral) Islam, A., Kowal, M., Derpanis, K., & Bruce, N. (2020). Feature Binding with Category-Dependant MixUp for Semantic Segmentation and Adversarial Robustness. In *British Machine Vision Conference (BMVC)*. Retrieved from https://arxiv.org/abs/2008.05667

Awards and Achievements

2024 **MITACs Accelerate** York University x Ubisoft La Forge ($45,000 over one year). Accepted.

2023 **NSERC CGS-D Scholarship** York University, Toronto ($105,000 over three years). Accepted.

2021 **Vector Post-Graduate Affiliate (PGA)**, Vector Institute, Toronto ($12,000). Affiliate status for two year term. Accepted.

York Graduate Scholarship (YGS), York University, Toronto ($3,000). Entrance scholarship. Accepted.

2020 **Ontario Graduate Scholarship (OGS)**, Ryerson University ($15,000). Accepted.

2017 **Keyser Award**, Queen's University ($1,000) - Best capstone project in Applied Mathematics and Engineering discipline. Accepted.

Awards and Achievements (continued)

2013 **Queen's Excellence Scholarship**, Queen's University ($8,000). Accepted.

Skills

Coding	Python, Bash, MATLAB, LaTeX.
Frameworks	PyTorch, NumPy, AWS, TensorFlow.
Communication	Skilled at conveying technical concepts in a clear and engaging way.