申 国斌

https://floyedshen.github.io/ +86 13931425808
floyed_shen@outlook.com shenguobin2021@ia.ac.cn
Guobin Shen 北京市 海淀区 中关村东路 95 号 自动化大厦

个人简介

我的研究兴趣是生物启发的神经网络及其在认知科学和人工智能中的应用. 我特别关注将脑启发模型与先进的 AI 系统相结合, 以及探索大规模模型的安全性和可解释性.

博士期间, 以第一作者在 ***PNAS***, *Cell Patterns* 等期刊以及 *NeurIPS*, *ICLR*, *CVPR*, *ICCV* 等会议上发表论文十余篇 (除了 PNAS 等顶级综合性期刊外, 还包含 CCF-A 会议**六**篇). 此外, 还作为共同作者参与发表在 *TPAMI*, *TEVC*, *Pattern Recognition*, *IJCAI*, *AAAI* 等期刊会议上的工作. 截至目前, 总引用量已达 600 余次, H-Index 为 14.

教育经历

2021 – 2026 **Ph.D.**, 模式识别与智能系统 中国科学院自动化研究所
Bio-Inspired AI / LLM Alignment, Safety & Interpretability

2017 – 2021 **B.Eng.**, 通信工程 中山大学电子与信息工程学院
GPA: 4.05 / 5.0, **Rank**: 1 / 85 (综合)

奖励与荣誉

学术荣誉

2023.11 **Cell Press 年度论文 (学生一作)**

2022.11 **Cell Press 中国科学家最佳论文奖 (第一作者)**

奖学金

2025.06 **中科院院长奖学金** (~1%) 非应届毕业生唯一获奖

2024.11 **博士生国家奖学金** (~1%)

2020.11 **本科生国家奖学金** (~2%)

2019.11 **本科生国家奖学金** (~2%)

竞赛奖项

2019.09 **国际空中机器人大赛亚太区亚军**

2019.08 **全国大学生电子设计竞赛国家二等奖**

2018.09 **全国大学生生物医学电子创新设计竞赛二等奖**

学术服务

担任 *NeurIPS*, *ICML*, *ICLR*, *CVPR*, *ICCV*, *ECCV*, *AAAI*, *MM*, *AISTATS* 等会议以及 *IEEE Computational Intelligence Magazine*, *Neural Networks*, *Neurocomputing* 等期刊的审稿人.

出版物

LLM & AI Safety

1. **Shen, Guobin**, Zhao, Dongcheng, Tong, Haibo, Li, Jindong, Zhao, Feifei, and Zeng, Yi. "Safety Instincts: LLMs Learn to Trust Their Internal Compass for Self-Defense." *Proceedings of the 14th International Conference on Learning Representations*, 2026. `AI Safety` `Instincts Rewards` `ICLR`

2. **Shen, Guobin**, Zhao, Dongcheng, Dong, Yiting, He, Xiang, and Zeng, Yi. "Jailbreak Antidote: Runtime Safety-Utility Balance via Sparse Representation Adjustment in Large Language Models." *Proceedings of the 13th International Conference on Learning Representations*, 2025. `AI Safety` `Interpretability` `ICLR`

3. **Shen, Guobin**, Zhao, Dongcheng, He, Xiang, Feng, Linghao, Dong, Yiting, Wang, Jihang, Zhang, Qian, and Zeng, Yi. "Neuro-Vision to Language: Image Reconstruction and Interaction via Non-invasive Brain Recordings." *Proceedings of the 38th Conference on Neural Information Processing Systems (NeurIPS 2024)*, 2024. `Multimodal LLM` `fMRI` `NeurIPS`

4. **Shen, Guobin**, Zhao, Dongcheng, Bao, Aorigele, He, Xiang, Dong, Yiting, and Zeng, Yi. "StressPrompt: Does Stress Impact Large Language Models and Human Performance Similarly?" *Proceedings of the 39th AAAI Conference on Artificial Intelligence (AAAI 2025)*, 2025. `LLM Analysis` `Cognitive Science` `AAAI`

5. **Shen, Guobin**, Zhao, Dongcheng, Dong, Yiting, Zhang, Qian, and Zeng, Yi. "Convergent Evolution across Modalities, Scales and Training Trajectories: Evidence for Human Brain-AI Alignment", 2025. `Brain-AI Alignment` `Cognitive Science`

6. **Shen, Guobin**, Zhao, Dongcheng, Feng, Linghao, He, Xiang, Wang, Jihang, Shen, Sicheng, Tong, Haibo, Dong, Yiting, Li, Jindong, Zheng, Xiang, and others. "PandaGuard: Systematic Evaluation of LLM Safety in the Era of Jailbreaking Attacks." 2025. `AI Safety` `Framework`

7. **Shen, Guobin**, Zhao, Dongcheng, Dong, Yiting, Li, Yang, Li, Jindong, Sun, Kang, and Zeng, Yi. "Astrocyte-Enabled Advancements in Spiking Neural Networks for Large Language Modeling." *arXiv preprint arXiv:2312.07625*, 2023. `Astrocyte` `LLM Pre-training`

8. Wu, Ping, **Shen, Guobin**, Zhao, Dongcheng, Wang, Yuwei, Dong, Yiting, Shi, Yu, Lu, Enmeng, Zhao, Feifei, and Zeng, Yi. "CVC: A Large-Scale Chinese Value Rule Corpus for Value Alignment of Large Language Models." *arXiv preprint arXiv:2506.01495*, 2025. `Value Alignment` `Dataset`

9. Dong, Yiting, **Shen, Guobin**, Zhao, Dongcheng, He, Xiang, and Zeng, Yi. "Harnessing Task Overload for Scalable Jailbreak Attacks on Large Language Models." *arXiv preprint arXiv:2410.04190*, 2024. `Jailbreak` `LLM`

出版物 (continued)

Spiking Neural Networks & Brain-Inspired AI

Shen, Guobin, Zhao, Dongcheng, Dong, Yiting, and Zeng, Yi. "Brain-Inspired Neural Circuit Evolution for Spiking Neural Networks." *Proceedings of the National Academy of Sciences*, vol. 120, no. 39, 2023, p. e2218173120. `Neuro-Evolution` `SNN` `PNAS`

10 **Shen, Guobin**, Zhao, Dongcheng, and Zeng, Yi. "Backpropagation with Biologically Plausible Spatiotemporal Adjustment for Training Deep Spiking Neural Networks." *Cell Patterns*, vol. 3, no. 6, 2022. `SNN` `Backpropagation` `Cell Patterns`

11 **Shen, Guobin**, Zhao, Dongcheng, Li, Tenglong, Li, Jindong, and Zeng, Yi. "Are Conventional SNNs Really Efficient? A Perspective from Network Quantization." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 27538-27547. `SNN` `Efficiency` `CVPR Hightlight`

12 **Shen, Guobin**, Zhao, Dongcheng, Dong, Yiting, Li, Yang, Zhao, Feifei, and Zeng, Yi. "Learning the Plasticity: Plasticity-Driven Learning Framework in Spiking Neural Networks." *Proceedings of the 39th Conference on Neural Information Processing Systems (NeurIPS 2025)*, 2025. `Plasticity` `Learning Framework` `NeurIPS`

13 **Shen, Guobin**, Zhao, Dongcheng, and Zeng, Yi. "Exploiting High-Performance Spiking Neural Networks with Efficient Spiking Patterns." *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2025. `SNN` `Efficiency` `TETCI`

14 **Shen, Guobin**, Zhao, Dongcheng, and Zeng, Yi. "Exploiting Nonlinear Dendritic Adaptive Computation in Training Deep Spiking Neural Networks." *Neural Networks*, vol. 170, 2024, pp. 190-201. `SNN` `Dendritic Dynamic` `Neural Networks`

15 **Shen, Guobin**, Zhao, Dongcheng, Shen, Sicheng, and Zeng, Yi. "Enhancing Spiking Transformers with Binary Attention Mechanisms." *The Second Tiny Papers Track at ICLR 2024*. `Transformer` `Binary Attention` `ICLR Tiny Paper`

16 **Shen, Guobin**, Zhao, Dongcheng, Dong, Yiting, Li, Yang, and Zeng, Yi. "Dive into the Power of Neuronal Heterogeneity." *arXiv preprint arXiv:2305.11484*, 2023. `Neuronal Heterogeneity` `SNN`

17 Han, Bing, Zhao, Feifei, Zeng, Yi, and **Shen, Guobin**. "Developmental Plasticity-Inspired Adaptive Pruning for Deep Spiking and Artificial Neural Networks." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. `Pruning` `Plasticity` `TPAMI`

18 Pan, Wenxuan, Zhao, Feifei, **Shen, Guobin**, Han, Bing, and Zeng, Yi. "Brain-Inspired Multi-Scale Evolutionary Neural Architecture Search for Deep Spiking Neural Networks." *IEEE Transactions on Evolutionary Computation*, 2024. `NAS` `Evolution` `TEVC`

19 Zhao, Dongcheng, **Shen, Guobin**, Dong, Yiting, Li, Yang, and Zeng, Yi. "Improving Stability and Performance of Spiking Neural Networks through Enhancing Temporal Consistency." *Pattern Recognition*, vol. 159, 2025, p. 111094. `SNN` `Stability` `Pattern Recognition`

出版物 (continued)

20 Zeng, Yi, Zhao, Dongcheng, Zhao, Feifei, **Shen, Guobin**, Dong, Yiting, Lu, Enmeng, Zhang, Qian, Sun, Yinqian, Liang, Qian, Zhao, Yuxuan, and others. "BrainCog: A Spiking Neural Network Based, Brain-Inspired Cognitive Intelligence Engine for Brain-Inspired AI and Brain Simulation." *Patterns*, 2023, p. 100789. `Framework` `Brain-inspired` `Patterns`

21 Han, Bing, Zhao, Feifei, Zeng, Yi, Pan, Wenxuan, and **Shen, Guobin**. "Enhancing Efficient Continual Learning with Dynamic Structure Development of Spiking Neural Networks." *Proceedings of the 32nd International Joint Conference on Artificial Intelligence (IJCAI 2023)*, 2023. `Continual Learning` `SNN` `IJCAI`

22 Yu, Yonghao, Zhao, Dongcheng, **Shen, Guobin**, Dong, Yiting, and Zeng, Yi. "Brain-Inspired Stepwise Patch Merging for Vision Transformers." *IJCAI*, 2025. `Vision Transformer` `Brain-inspired` `IJCAI`

23 Shen, Sicheng, Zhao, Dongcheng, **Shen, Guobin**, and Zeng, Yi. "TIM: An Efficient Temporal Interaction Module for Spiking Transformer." *Proceedings of the 33rd International Joint Conference on Artificial Intelligence (IJCAI 2024)*, 2024. `Transformer` `Temporal` `IJCAI`

24 He, Xiang, Liu, Xiangxi, Li, Yang, Zhao, Dongcheng, **Shen, Guobin**, Kong, Qingqun, Yang, Xin, and Zeng, Yi. "CACE-Net: Co-guidance Attention and Contrastive Enhancement for Effective Audio-Visual Event Localization." *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 985-993. `Multimodal` `Event Localization` `MM`

25 He, Xiang, Zhao, Dongcheng, Li, Yang, **Shen, Guobin**, Kong, Qingqun, and Zeng, Yi. "An Efficient Knowledge Transfer Strategy for Spiking Neural Networks from Static to Event Domain." *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 1, 2024, pp. 512-520. `Knowledge Transfer` `Event Domain` `AAAI`

26 Feng, Linghao, Zhao, Dongcheng, Shen, Sicheng, Dong, Yiting, **Shen, Guobin**, and Zeng, Yi. "Time Cell Inspired Temporal Codebook in Spiking Neural Networks for Enhanced Image Generation." *arXiv preprint arXiv:2405.14474*, 2024. `Image Generation` `Time Cell`

27 Shen, Sicheng, Zhao, Dongcheng, Feng, Linghao, Yue, Zeyang, Li, Jindong, Li, Tenglong, **Shen, Guobin**, and Zeng, Yi. "STEP: A Unified Spiking Transformer Evaluation Platform for Fair and Reproducible Benchmarking." *Proceedings of the 39th Conference on Neural Information Processing Systems (NeurIPS 2025) Dataset and Benchmark Track*, 2025. `Spiking Transformer` `Benchmark Platform` `NeurIPS`

Hardware Acceleration & System Optimization

28 **Shen, Guobin**, Li, Jindong, Li, Tenglong, Zhao, Dongcheng, and Zeng, Yi. "SpikePack: Enhanced Information Flow in Spiking Neural Networks with High Hardware Compatibility." *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. `SNN` `Hardware` `ICCV`

29 Li, Jindong, **Shen, Guobin**, Zhao, Dongcheng, Zhang, Qian, and Zeng, Yi. "Firefly v2: Advancing Hardware Support for High-Performance Spiking Neural Network with a Spatiotemporal FPGA Accelerator." *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 2024. `Hardware` `FPGA` `TCAD`

出版物 (continued)

30 Li, Jindong, **Shen, Guobin**, Zhao, Dongcheng, Zhang, Qian, and Zeng, Yi. "Firefly: A High-Throughput Hardware Accelerator for Spiking Neural Networks with Efficient DSP and Memory Optimization." *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 31, no. 8, 2023, pp. 1178-1191. `Hardware` `Accelerator` `TVLSI`

31 Li, Jindong, Li, Tenglong, **Shen, Guobin**, Zhao, Dongcheng, Zhang, Qian, and Zeng, Yi. "Revealing Untapped DSP Optimization Potentials for FPGA-Based Systolic Matrix Engines." *2024 34th International Conference on Field-Programmable Logic and Applications (FPL)*, IEEE, 2024, pp. 197-203. `FPGA` `Optimization` `FPL`

32 Li, Tenglong, Li, Jindong, **Shen, Guobin**, Zhao, Dongcheng, Zhang, Qian, and Zeng, Yi. "FireFly-S: Exploiting Dual-Side Sparsity for Spiking Neural Networks Acceleration with Reconfigurable Spatial Architecture." *IEEE Transactions on Circuits and Systems I: Regular Papers*, 2024. `Acceleration` `Sparsity` `TCAS-I`

33 Li, Jindong, Li, Tenglong, **Shen, Guobin**, Zhao, Dongcheng, Zhang, Qian, and Zeng, Yi. "Pushing Up to the Limit of Memory Bandwidth and Capacity Utilization for Efficient LLM Decoding on Embedded FPGA." *Proceedings of the 2025 Design, Automation & Test in Europe Conference (DATE)*, 2025. `LLM Acceleration` `FPGA` `DATE`

34 Li, Jindong, Li, Tenglong, Chen, Ruiqi, **Shen, Guobin**, Zhao, Dongcheng, Zhang, Qian, and Zeng, Yi. "Hummingbird: A Smaller and Faster Large Language Model Accelerator on Embedded FPGA." *Proceedings of the 2025 International Conference on Computer-Aided Design (ICCAD)*, 2025. `LLM Acceleration` `FPGA` `ICCAD`

Datasets & Data Augmentation

35 **Shen, Guobin**, Zhao, Dongcheng, and Zeng, Yi. "EventMix: An Efficient Data Augmentation Strategy for Event-Based Learning." *Information Sciences*, vol. 644, 2023, p. 119170. `Event-based` `Augmentation` `Information Sciences`

36 Dong, Yiting, He, Xiang, **Shen, Guobin**, Zhao, Dongcheng, Li, Yang, and Zeng, Yi. "EventZoom: A Progressive Approach to Event-Based Data Augmentation for Enhanced Neuromorphic Vision." *Proceedings of the 39th AAAI Conference on Artificial Intelligence (AAAI 2025)*, 2025. `Event-based` `Neuromorphic` `AAAI`

37 Dong, Yiting, Li, Yang, Zhao, Dongcheng, **Shen, Guobin**, and Zeng, Yi. "Bullying10K: A Large-Scale Neuromorphic Dataset Towards Privacy-Preserving Bullying Recognition." *Advances in Neural Information Processing Systems*, vol. 36, 2024. `Dataset` `Neuromorphic` `NeurIPS`

项目经历

PandaGuard: 大语言模型安全评估框架 Stars 42 **项目负责人**

- 设计并实现系统性 LLM 越狱攻击安全评估框架, 集成多种攻击和防御算法
- 构建大规模基准数据集 PandaBench, 提供多维度安全评估指标体系

BrainCog: 脑启发认知智能引擎 Stars 548 **项目负责人**

- 领导开发综合性脉冲神经网络框架, 支持脑启发 AI 和脑仿真研究
- 实现 50+ 功能性 SNN 算法, 涵盖感知学习、决策制定、知识表示等认知功能

项目经历 (continued)

aw_nas: 模块化神经架构搜索框架  **算法贡献者**

- 为开源 NAS 框架贡献 Once-for-All 等算法的复现, 以及相关数据集的构建.