Yixian Shen, EU permanent residence card holder

y.shen@uva.nl senianone7@gmail.com +31 649 779 351 LinkedIn
1094LA, Tweede Atjehstraat, Amsterdam, The Netherlands Google Scholar

Education

2020 – 2024 **Doctor of Philosophy** in *Computer Science*, Informatics Institute.
Universiteit van Amsterdam (UvA), Amsterdam, The Netherlands.

2017 – 2019 **Master of Science** in *Computer Science*, School of Computer Science.
Sun Yat-sen University, Guangzhou, China. GPA: 4.1/5, Rank: 2/72.

2013 – 2017 **Bachelor of Science** in *Electronic and Information Engineering*.
Communication University of Zhejiang, Hangzhou, China. GPA: 4.06/5, Rank: 1/307.

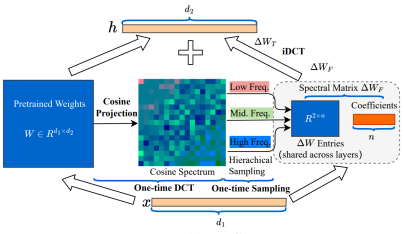
Research Interests

I am broadly interested in developing **efficient and robust AI systems** that unify **language**, **vision**, and **hardware–software co-design**. My research focuses on model efficiency, multimodal reasoning, and real-time system optimization across both algorithmic and architectural levels. Key topics include:

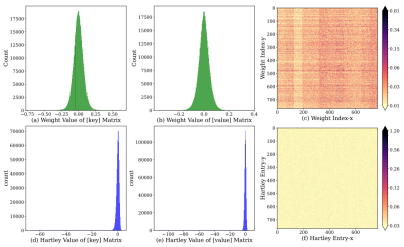
- Efficient LLM Training and Fine-Tuning
- Multimodal Vision-Language Reasoning
- Edge and Embedded Deep Learning
- Model and Token Compression for MLLMs
- Real-Time and Energy-Efficient AI Systems
- Large Language Models and Their Applications
- Multi-Core Cache and Memory Optimization
- Thermal-Aware Scheduling for 3D-Stacked Chips

Publications (ICLR, ACL, ICCV, EMNLP, NAACL, AAAI, NeurIPS, etc)

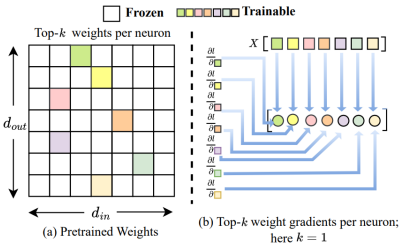
Efficient LLM Training and Fine-tuning



1 MaCP: Minimal yet Mighty Adaptation via Hierarchical Cosine Projection.
Yixian Shen, *Qi Bi, Jia-Hong Huang, Hongyi Zhu, Andy D. Pimentel, Anuj Pathania.*
In *The 63rd Annual Meeting of the Association for Computational Linguistics (**ACL 2025**).*
Best Theme Paper Paper Link Code

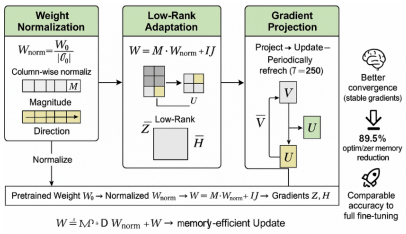


2 SSH: Sparse Spectrum Adaptation via Discrete Hartley Transformation.
Yixian Shen, *Qi Bi, Jia-Hong Huang, Hongyi Zhu, Andy D. Pimentel, Anuj Pathania.*
In *The 2025 Annual Conference of the Nations of the Americas Chapter of the ACL (**NAACL 2025**).*
Paper Link Code

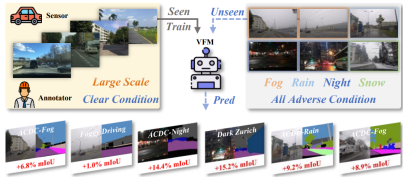


3 NeuroAda: Activating Each Neuron's Potential for Parameter-Efficient Fine-Tuning.
Zhi Zhang*, ***Yixian Shen****, Congfeng Cao, Ekaterina Shutova
In *The 2025 Conference on Empirical Methods in Natural Language Processing (**EMNLP 2025**).*
Paper Link Code * Equal contribution.

Publications (ICLR, ACL, ICCV, EMNLP, NAACL, AAAI, NeurIPS, etc 🎓) (continued)

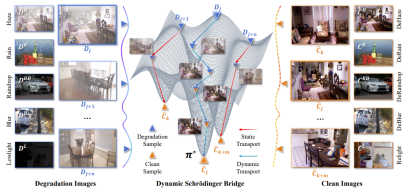


4 Gradient Weight-normalized Low-rank Projection for Efficient LLM Training.
Jia-Hong Huang*, **Yixian Shen***, Hongyi Zhu, Stevan Rudinac, Evangelos Kanoulas.
In *The 39th Annual AAAI Conference on Artificial Intelligence* (***AAAI 2025***). Paper Link Code * Equal contribution.

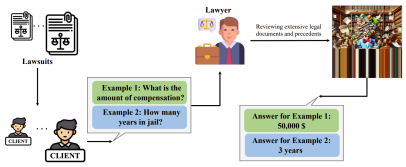


5 AdaDCP: Learning an Adapter with Discrete Cosine Prior for Clear-to-Adverse Domain Generalization.
*Qi Bi, **Yixian Shen**, Jingjun Yi, Gui-Song Xia.*
In *International Conference on Computer Vision* (***ICCV 2025***). Paper Link

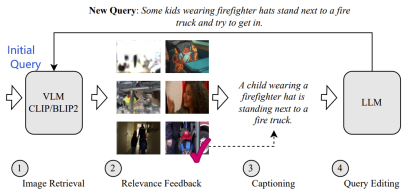
Deep Learning Applications



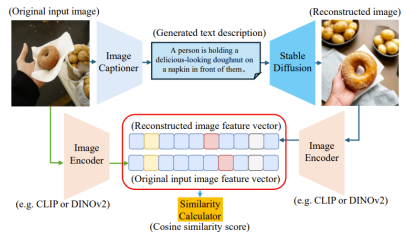
6 Degradation-aware Dynamic Schrödinger Bridge for Unpaired Image Restoration.
*Jingjun Yi, Qi Bi, Hao Zheng, Huimin Huang, **Yixian Shen**, Haolan Zhan, Wei Ji, Yawen Huang, Yuexiang Li, Xian Wu, Yefeng Zheng.*
In *The 39th Annual Conference on Neural Information Processing Systems* (***NeurIPS 2025***). Paper Link



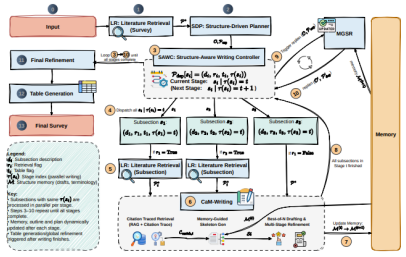
7 Optimizing numerical estimation and operational efficiency in the legal domain through large language models.
*Jia-Hong Huang, Chao-Chun Yang, **Yixian Shen**, Alessio M Pacces, Evangelos Kanoulas.*
In *The 33rd ACM International Conference on Information and Knowledge Management* (***CIKM 2024***). Paper Link Code



8 Interactive Image Retrieval Meets Query Rewriting with Large Language and Vision Language Models.
*Hongyi Zhu, Jia-Hong Huang, **Yixian Shen**, Stevan Rudinac, Evangelos Kanoulas.*
In *ACM Transactions on Multimedia Computing, Communications and Applications* (***TOMM 2025***). Paper Link



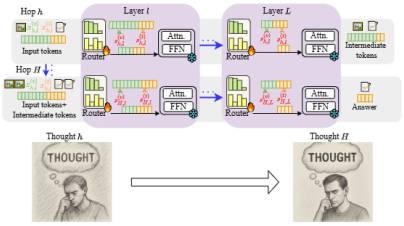
9 Image2text2image: A novel framework for label-free evaluation of image-to-text generation with text-to-image diffusion models.
*Jia-Hong Huang, Hongyi Zhu, **Yixian Shen**, Stevan Rudinac, Evangelos Kanoulas.*
In *The 31st International Conference on Multimedia Modeling* (***MMM 2024***). Paper Link



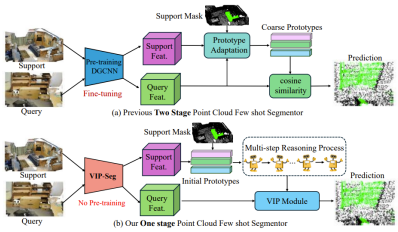
10 SurveyGen-I: Consistent Scientific Survey Generation with Evolving Plans and Memory-Guided Writing.
*Jing Chen, Zhiheng Yang,**Yixian Shen**, Jie Liu, Adam Belloum, Chrysa Papagainni, Paola Grosso.*
In *International Joint Conference on Natural Language Processing & Asia-Pacific Chapter of the Association for Computational Linguistics 2025* (***AACL 2025***). Paper Link Code

Publications (ICLR, ACL, ICCV, EMNLP, NAACL, AAAI, NeurIPS, etc 🎓) (continued)

Efficient MLLM Reasoning

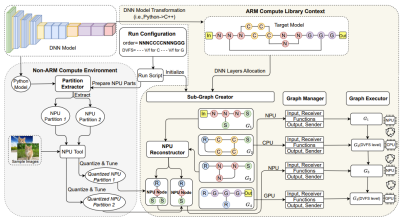


11 Efficient Multimodal Spatial Reasoning via Dynamic and Asymmetric Routing.
Yixian Shen, *Qi Bi, Zihan Wang, Zhiheng Yang, Changshuo Wang, Zhi Zhang, Prayag Tiwari, Andy D. Pimentel, Anuj Pathania.*
In *The Fourteenth International Conference on Learning Representations* (***ICLR 2026***). Paper Link

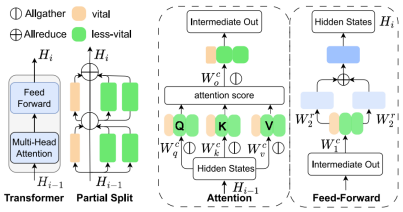


12 Reasoning Beyond Points: A Visual Introspective Approach for Few-Shot 3D Segmentation.
Changshuo Wang, Shuting He, Xiang Fang, Zhijian Hu, Jia-Hong Huang, ***Yixian Shen****, Prayag Tiwari.*
In *The 39th Annual Conference on Neural Information Processing Systems* (***NeurIPS 2025***). Paper Link

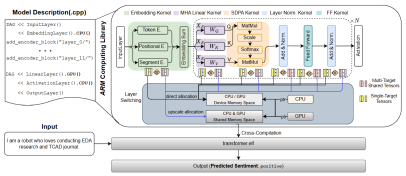
Efficient Edge DNNs



13 PiQi: Partially Quantized DNN Inference on HMPSoCs.
Ehsan Aghapour, ***Yixian Shen****, Dolly Sapra, Andy D. Pimentel and Anuj Pathania.*
In *2024 ACM/IEEE International Symposium on Low Power Electronics and Design* (***ISLPED 2024***).
Paper Link Code

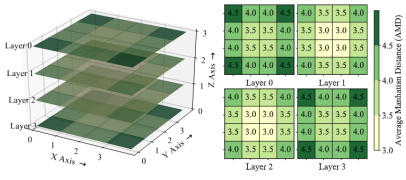


14 EASTER: Learning to Split Transformers at the Edge Robustly.
Xiaotian Guo , Quan Jiang , ***Yixian Shen****, Andy D. Pimentel and and Todor Stefanov.*
In *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems 2024* (***TCAD 2024***) .
Paper Link

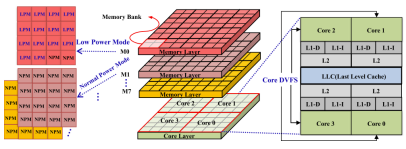


15 Low Latency Transformer Inference on HMPSoCs
Hang Xu, ***Yixian Shen****, Theo Gatea, Andy Pimentel, and Anuj Pathania.*
In *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems 2025* (***TCAD 2025***). *Under Review.*

Thermal-aware Scheduling for 3D-stacked Chips

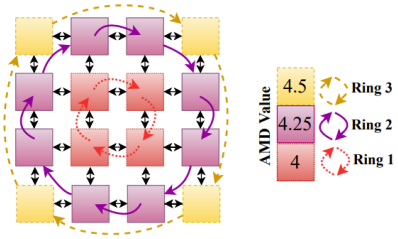


16 Taming the Heat: Thermal-Safe and Efficient LFM Inference on 3D S-NUCA Many-Cores.
Yixian Shen*,Jan Deen, George Floros, Andy Pimentel and Anuj Pathania.*
In *2026 Design Automation Conference* (***DAC 2026***), *Under Review.*



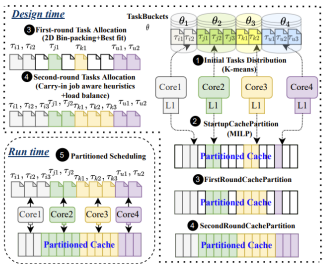
17 Thermal Management for 3D-Stacked Systems via Unified Core-Memory Power Regulation.
Yixian Shen*,Leo Schreuders, Andy D. Pimentel and Anuj Pathania.*
In *ACM Transactions on Embedded Computing Systems and ESWEEK* (***TECS and ESWEEK 2023***). Paper Link Code

Publications (ICLR, ACL, ICCV, EMNLP, NAACL, AAAI, NeurIPS, etc 🎓) (continued)

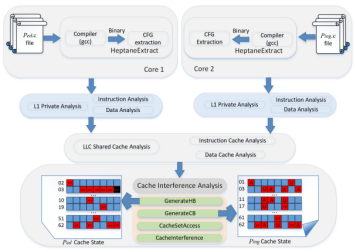


18 Thermal Management for S-NUCA Many-Cores via Synchronous Thread Rotations.
Yixian Shen,*Sobhan Niknam, Andy D. Pimentel and Anuj Pathania.*
In *2023 Design, Automation & Test in Europe Conference* (***DATE 2023***).
Paper Link Code

Multi-Core Cache Optimization



19 TCPS: A Task and Cache-Aware Partitioned Scheduler for Hard Real-Time Multi-core Systems.
Yixian Shen, *Jun Xiao and Andy D. Pimentel.*
In *The 23rd ACM SIGPLAN/SIGBED International Conference on Languages, Compilers, and Tools for Embedded Systems* (***LCTES 2022***).
Paper Link



20 Cache interference-aware task partitioning for non-preemptive real-time multi-core systems.
Jun Xiao, ***Yixian Shen*** *and Andy D. Pimentel.*
In *ACM Transactions on Embedded Computing Systems*(***TECS 2022***).
Paper Link

Practical Machine Learning Project Experiences

Efficient LLM Training and Fine-tuning

1 Low-Rank Adaptation in the Frequency Domain.
Goal: Enable parameter- and memory-efficient fine-tuning of large language models for downstream tasks by projecting adaptation into structured frequency subspaces.
Contribution: ① Designs frequency projection that encodes updates into spectral subspaces across transformer layers. ② Achieves up to 70% reduction in trainable parameters while preserving downstream accuracy. ③ Bridges frequency-domain analysis and low-rank adaptation to deliver compact, scalable LLM fine-tuning.

2 Stabilized Low-Rank Adaptation for Robust LLM Optimization.
Goal: Enhance the stability and efficiency of large language model training by mitigating gradient imbalance in low-rank adaptation mechanisms.
Contribution: ① Incorporates gradient normalization into rank-limited updates to ensure balanced parameter scaling. ② Improves convergence robustness and reduces training variance across transformer layers. ③ Achieves full fine-tuning–level performance with superior efficiency–accuracy trade-offs via adaptive gradient-controlled projection.

3 Neuron-Selective Adaptation for Efficient LLM Fine-Tuning.
Goal: Achieve fine-grained and parameter-efficient adaptation of large language models by selectively activating neuron subspaces critical for downstream tasks.
Contribution: ① Proposes neuron-level modulation that dynamically reweights activation paths for efficient task adaptation. ② Reduces redundant neuron updates while maintaining expressiveness and generalization. ③ Demonstrates superior efficiency–accuracy balance compared with existing low-rank or adapter-based fine-tuning methods.

Practical Machine Learning Project Experiences (continued)

Efficient Multimodal and Spatial Reasoning

1 Dynamic Routing for Efficient Multimodal Spatial Reasoning.
Goal: Enable efficient visual–language reasoning by dynamically selecting salient visual and textual tokens across layers and reasoning hops in multimodal large language models.
Contribution: ① Proposes a routing framework that adaptively retains essential visual–text tokens through a differentiable top-k gating mechanism. ② Introduces asymmetric retention across modalities to prioritize spatially informative regions and concise textual cues. ③ Achieves significant reductions in computation and memory while preserving reasoning accuracy on spatial–visual benchmarks such as Winoground and V-Star.

Computer Vision Applications

1 Schrödinger-Bridge Framework for Unpaired Image Restoration.
Goal: Restore high-quality images from degraded inputs without requiring paired supervision, by learning a distributional bridge between degraded and clean image domains.
Contribution: ① Develops a degradation-aware dynamic Schrödinger Bridge that models bidirectional stochastic transport between degraded and clean image distributions. ② Incorporates physical degradation priors, such as blur, haze, and low-light noise, into the transport process for realism and robustness. ③ Achieves state-of-the-art unpaired restoration performance across diverse degradation types through consistency-constrained dynamic transport.

2 Sparse-View Underwater Reconstruction with Medium-Aware Gaussian Splatting.
Goal: Reconstruct high-fidelity 3D scenes from few underwater views by compensating absorption, scattering, and color shifts inherent to the medium.
Contribution: ① Integrates a degradation-aware appearance model into 3D Gaussian Splatting to correct view-inconsistent color and haze. ② Introduces consistency and geometry stabilizers tailored to sparse-view inputs for robust reconstruction. ③ Achieves superior quality and fidelity over vanilla 3DGS on underwater benchmarks under limited views.

3 Visual-Introspective Reasoning for Few-Shot 3D Segmentation.
Goal: Enhance 3D point-cloud segmentation under few-shot settings by enabling introspective reasoning over visual prototypes and reducing intra-class variance across support and query domains.
Contribution: ① Introduces a visual introspection framework that refines class prototypes through iterative reasoning using enhancement and difference modules. ② Designs a dynamic power convolution to better capture local geometric structures under limited supervision. ③ Achieves state-of-the-art few-shot and cross-domain 3D segmentation on S3DIS and ScanNet without requiring large-scale pretraining.

NLP Applications

1 Sentiment Analysis using LSTM Networks.
Goal: Classify text sentiment by modeling sequential dependencies through an LSTM-based architecture.
Contribution: ① Implemented a bi-directional LSTM network for sentence-level sentiment classification, capturing contextual relationships and polarity cues. ② Incorporated dropout regularization and gradient clipping to stabilize optimization and prevent overfitting. ③ The model effectively captures long-term dependencies in text, achieving robust accuracy across multiple domains.

Practical Machine Learning Project Experiences (continued)

2 Document Classification using Mamba Sequence Modeling.
Goal: Apply state-space sequence modeling (Mamba) to text classification tasks for improved long-context understanding with reduced computational overhead.
Contribution: ① Implemented a Mamba-based classifier that encodes document sequences efficiently through implicit long-range dependency modeling. ② Fine-tuned the lightweight Mamba architecture on standard benchmarks such as AG News and IMDB, achieving accuracy comparable to transformer models while requiring fewer parameters and lower latency. ③ Demonstrated that state-space models offer a practical and scalable alternative for document-level NLP tasks.

3 Speech Recognition with QuartzNet-based Acoustic Modeling.
Goal: Develop a lightweight end-to-end automatic speech recognition (ASR) system leveraging convolutional time-channel separable networks for efficient feature encoding.
Contribution: ① Implemented and fine-tuned a **QuartzNet**-style encoder for character-level ASR on the LibriSpeech dataset using **NVIDIA NeMo**. ② Optimized training with mixed-precision and data parallelism to reduce GPU memory usage by 40% without accuracy degradation. ③ Achieved over 94% word accuracy on the test-clean subset, validating the efficiency and scalability of convolutional ASR architectures for real-time speech applications.

Hardware-Aware Deep Learning Inference

1 Partially Quantized Inference on Heterogeneous Multi-Core Systems.
Goal: Optimize DNN inference efficiency on heterogeneous multi-processor SoCs by selectively quantizing layers and dynamically mapping computation across CPU, GPU, and NPU cores.
Contribution: ① Develops a partially quantized inference framework that balances precision and performance through layer-wise hardware assignment. ② Employs a multi-objective search to co-optimize energy, latency, and accuracy under system constraints. ③ Integrates a learning-based accuracy predictor to accelerate design-space exploration, achieving near-optimal performance-power trade-offs for modern HMPSoCs.

2 Thermal- and Cache-Aware Scheduling for Large Model Inference on 3D Many-Cores.
Goal: Enhance inference efficiency of large foundation models on 3D-stacked S-NUCA many-core processors by jointly optimizing cache locality, thermal safety, and execution latency.
Contribution: ① Proposes an Active Imitation Learning framework (AILFM) that dynamically schedules transformer-block snippets across thermally heterogeneous cores. ② Integrates cache- and temperature-aware cost modeling to balance performance and reliability. ③ Achieves significant latency reduction and thermal stability compared to heuristic schedulers, enabling scalable and safe LFM inference on stacked architectures.

Work Experience

2020.07–2023.10 **University of Amsterdam, Parallel Computing Systems (PCS) Group**
Ph.D. Researcher, Supervised by Anuj Pathania and Prof. Andy D. Pimentel
① Conducting research on efficient and scalable AI systems, focusing on large foundation models (LFMs) and multimodal large language models (MLLMs).
② Proposed several frameworks for memory- and parameter-efficient adaptation, including **MaCP**, **SSH**, and **NeuroAda**, achieving substantial training efficiency with minimal accuracy loss.
③ Developed the **DARE** framework for dynamic and asymmetric routing in multimodal spatial reasoning.
④ Introduced **AILFM**, a thermal- and cache-aware scheduling framework for LFM inference on 3D-stacked S-NUCA systems, addressing energy–latency trade-offs.

Work Experience (continued)

2023.10–2024.04 **Imec R&D, Nano Electronics and Digital Technologies**
Research Intern, Supervised by Dr. Vinay B. Y. Kumar and Prof. Francky Catthoor
① Investigated chiplet-based co-design of heterogeneous systems with thermal-aware placement and 2.5D/3D integration.
② Applied system-level modeling to balance inter-chip communication, thermal coupling, and architectural scalability in advanced packaging systems.

2024.05–2024.09 **Junior Lecturer, University of Amsterdam**
Research Collaboration Project
① Supervised a group of eight master's students on projects focused on enhancing the accessibility and semantic enrichment of EPUB files.
② Contributed to the development of AI-driven methods for automated content description and multimodal accessibility improvement.

2024.09 – present **Postdoctoral Researcher, University of Amsterdam**
Computer Science, Informatics Institute
① Designed lightweight adaptation frameworks for large multimodal models, enabling efficient reasoning and perception under strict compute and memory constraints.
② Investigated agentic inference mechanisms, including dynamic token routing and selective activation, to improve multi-step multimodal reasoning efficiency.
③ Developed hardware-aware optimization strategies for scalable deployment, integrating model compression, quantization, and heterogeneous execution.

Skills

Languages — Fluent in English; Dutch (A2); native in Mandarin Chinese.

Programming — Python, C++, SQL, Bash; Git, Docker, Kubernetes, Conda; systems-oriented Linux development.

Machine Learning — Large language models, agentic reasoning and planning, tool-augmented inference, diffusion models, multimodal learning.

Scientific Writing — Strong publication record (NeurIPS, ICLR, ACL, ICCV, DATE, NAACL, EMNLP, ICASSP, CODES+ISSS, etc.); expert in LaTeX, TikZ, and reproducible research.

Hardware-Aware AI — Efficient agent execution and LLM inference; CPU–GPU–NPU co-design, quantization, CUDA optimization, Slurm/DAS-6 clusters, 3D many-core modeling.

Professional Service

- **ICCAD 2023:** External Reviewer, *The International Conference on Computer-Aided Design.*
- **CGO 2024, 2025:** Artifact Evaluation Committee Member, *IEEE/ACM International Symposium on Code Generation and Optimization.*
- **CIKM 2024, 2025:** Reviewer, *International Conference on Information and Knowledge Management.*
- **ICLR 2025,2026:** Reviewer, *International Conference on Learning Representations.*
- **ACM MM 2025:** Reviewer, *ACM International Conference on Multimedia.*
- **AAAI 2026:** Reviewer, *The 40th Annual AAAI Conference on Artificial Intelligence.*
- **IEEE Transactions on Computers (TC):** Reviewer, *A leading IEEE journal covering computer architecture and system-level design.*
- **Scientific Reports (Nature Portfolio) and Springer Nature journals.** *Areas: artificial intelligence, computer systems, and computational modeling.*