Xuefei Ning

Birthday: January 11st, 1996
foxdoraame@gmail.com · https://github.com/walkerning
https://scholar.google.com/citations?user=oVslpJsAAAAJ
NICSEFC-EffAlg team I built in E.E., Tsinghua: https://nics-effalg.com/
Imagination-Research team with friends: https://github.com/imagination-research

Summary of Research Directions

- **Efficient deep learning (particularly efficient AIGC)**: I've been working on efficient deep learning (DL) since 2019. Since 2023, most of my work has been centered around efficient AIGC.

  From early 2020 to the end of 2024, in my Ph.D. advisor's laboratory that focuses on hardware design, I gradually built up a team that focuses on DL research from scratch. The information about the team I built and led for 5 years in Tsinghua University can be found at https://nics-effalg.com/.

- **Towards better reasoning and learning**: I have been thinking about the gaps between current techniques and the efficient, intelligent AI that I imagine. I am interested in improving the *fast learning ability* of an AI to generalize with very few data. As I speculate that a *strong reasoning ability* is one of the keys to the *fast learning ability*, together with collaborators, we published a work on drawing insights from human cognition and learning to improve AI reasoning performance on NeurIPS 2024. I also led a team to participate in AIMO Progress Prize 2 and won the 2nd place (announced in April 2025).

Work Experience

Tsinghua University, Research-Track Assistant Professor 2024/03 – now
Microsoft Research, Visiting Scholar 2026/05 – now
Tsinghua-Huawei Joint Postdoctoral Program, Postdoctoral Researcher 2021/12 – 2023/12
Advisor: Prof. Pinyan Lu, Prof. Yu Wang

Education

Tsinghua University, Bachelor of Electronic Engineering 2012/09 – 2016/07
Score: 92/100 *Ranking: 12/231* *Graudate with honor*

Tsinghua University, Doctor of Philosophy in Electronic Science and Technology 2016/9 – 2021/10
Advisor: Prof. Yu Wang, Prof. Huazhong Yang
Thesis: Neural Architecture Search for Efficient and Robust Convolutional Neural Networks

Mentorship Experiences

From 2020, when I began by mentoring three undergraduates, to the end of 2024, I devoted major effort to building and leading the Efficient Algorithm (EffAlg) team within the NICSEFC group.
In the past three years, one of my major focus was understanding and instructing different types of students onto the research track – help them grasp knowledge and implementation skills, improve their analytical thinking abilities, and cultivate their ideas and research sense. The frequent and in-depth reflection during this process also contributed significantly to my own growth.
Until the end of 2024, I have directly mentored 9 graduate students and various interns. I am proud that four of my mentored graduate students have earned their master's degrees, with one graduating with honors, and one of them has earned the Ph.D. degree. For more information, please visit the NICSEFC-EffAlg website.

Teaching Experiences

C/UNIX Programming 2020 Autumn, 2022 Autumn, 2024 Autumn
Lecturer Course Instructor: Prof. Huazhong Yang, Prof. Yu Wang

Computer-Aided Design of Digital Circuits and Systems 2020 Spring, 2022 Spring, 2024 Spring

Teaching Assistant Course Instructor: Prof. Yu Wang

In the 2020 Spring course, I led the course collaboration project that produced the survey paper "Machine Learning for Electronic Design Automation: A Survey", published in TODAES 2021.

Industry Internship Experiences

Douban, Beijing, China 2015-7 – 2015-9

Software Engineer Intern of the Platform Group Mentor: Guillaume Bouriez

- Participate in the development of the RPC system on the private application cloud of Douban. This system acts as a vital component of the Micro-Service Architecture in Douban.
- Specifically, my task is to: (1) optimize "circuit breaker" of the RPC system; (2) develop service supervisor that reports service status; (3) support explicit interface declaration to increase the robustness of the system.

DeePhi Tech (now part of Xilinx), Beijing, China 2016-4 – 2016-8

Software Engineer Intern, IT Manager Mentor: Hong Luo

- Build up and maintain all the networking and servers in the startup.
- Lead the development of the compression toolchain for deploying LSTM (based on Kaldi) and CNN (based on Caffe) onto FPGA, including pruning and quantization functionalities.

Tencent AI Lab, Shenzhen, China 2019-3 – 2019-5

Machine Learning Research Intern Mentor: Yin Zheng

- Revise and submit the paper "Nonparametric Topic Modeling with Neural Inference", published in Neurocomputing, which designs a nonparametric and hierarchical Dirichlet Process prior for VAE for flexible topic modeling.
- Develop a Gumbel-Softmax extension of differentiable neural architecture search. *I am grateful as I started the NAS research for my Ph.D. thesis during this internship, a direction suggested by Dr. Zheng.*

Project Experiences

- I have participated in more than 10 engineering project collaboration with enterprises. I am the project PI for collaboration projects with Baidu, Oppo, and Kuaishou.
- I received the support of NSFC Young Scientists Fund (Category C) in 2025.
- I have published 10 patents as one of the first three authors.

Honors and Awards

- [Competition] *2nd prize in AI Mathematical Olympiad - Progress Prize 2 (Kaggle)* (2nd/2211) 2025/04
  (As the second code&exp contributor and project leader, with Yichen You, Zinan Lin)
- [Competition] *2nd prize in NeurIPS 2018 Adversarial Vision Challenge Competition* (2nd/339) 2018/12
  Solution: Mutual Adversarial Training with Diverse Early Stop PGD (As the main code&exp contributor and project leader, with Wenshuo Li)
- *Outstanding Graduate of Tsinghua University* (second-level, 10%) 2016/07
- *Future Scholar Scholarship of Tsinghua University* (2/103) 2016/07
- [Competition] *Grand Prize in the 5th Creativity Competition of Tsinghua University* 2016/04
  Solution: A Cross-Platform Mobile Application for Tsinghua Web Learning System (As the second code&exp contributor, with Sihan Li)
- *National Scholarship for Encouragement, Academic Excellence Award* 2015, 2014
- [Competition] *First Prize in National High Schools Physics Competition* 2011

Academic Service

- Reviewer for ICML (2022, 2023, 2024, 2025, 2026), NeurIPS (2022, 2023, 2024, 2025), ICLR (2023, 2024, 2025), CVPR (2022, 2023), ICCV / ECCV (2022, 2023, 2024, 2025, 2026), AAAI (2023, 2024), AISTATS (2025), ECML-PKDD (2022).

- Senior area chair for ACL (2025), EMNLP (2026). Area chair for CVPR (2025, 2026), ICLR (2026), COLM (2026), NeurIPS (2026).
- Reviewer for TPAMI, IJCV, CUSR, TCSVT, Pattern Recognit., TECS, TCAD, TODAES. TPC member for DAC (2025 AI track).

Selected Talks

- **Efficient Inference for Large Language Models – Algorithm, Model, and System**. Slide
  Tutorial at *EMNLP 2025*. 2025-11.
- **(Chinese) Tutorial on Efficient Deep Learning for Generative Model**.
  Invited tutorial at *CCF ADL Tutorial Series No.157*. 2025-4.
- **(Chinese) Trend Introduction on Inference Optimization**. Slide
  Invited talk at *Zhiyuan's Annual Discussion on the 10 AI Trends*. 2025-1.
- **Generative Model Compression and Acceleration**. Slide
  Invited talks at *Huawei; Apple-China; University of Chinese Academy of Sciences; University of Electronic Science and Technology of China; VIVO; AMD-China; and others*. 2024.
- **An Introduction to Quantization of Large Language Models**. Video and slide
  Invited tutorial for *an Competition organized by AWS-China*. 2023-8.
- **Model Compression Towards Efficient Deep Learning Inference**. Slide
  Invited talks at *Huawei; Inceptio.ai; Beihang University; and others*. 2023.
- **Neural Architecture Search and Architecture Encoding**. Slide
  Invited talks at *DAMO Academy of Alibaba Group (U.S.); Renmin University of China; and others*. 2022.

Selected Work

1. **Xuefei Ning**, Wenshuo Li, Yu Wang, "Second place in NeurIPS 2018 Adversarial Vision Challenge (Model Track 2nd/339)".
   NeurIPS 2018 Adversarial Vision Challenge: NNs are found to be vulnerable to adversarial examples. This work proposes two new techniques for improving NNs' robustness: 1) mutual adversarial learning that combine adversarial training and mutual learning, 2) return-early black-box adversarial example generation. Both techniques are shown to be beneficial. Combined with other techniques (mixing black-box and white-box adversarial examples, using stronger augmnetation in training, and test-time augmentation and ensemble), the resulting solution (w.o. extra data) won the 2nd place in NeurIPS 2018 Adversarial Vision Challenge, closely match the score of the 1st place solution from CMU that uses extra dataset.

2. **Xuefei Ning**, Yin Zheng, Tianchen Zhao, Yu Wang, Huazhong Yang, "A Generic Graph-based Neural Architecture Encoding Scheme for Predictor-based NAS", In ECCV 2020. [Website]
   Design a more suitable encoder for neural architecture to improve NAS efficiency: Predictor-based search method relies on a performance predictor to rapidly estimate the performance of candidates and prioritize those with high potential, thereby improving sample efficiency. The quality of the performance predictor is critical to the efficiency of the optimization process. This work targets a more efficient predictor-based optimization for neural architecture search (NAS). We propose GATES, a graph-based encoder specifically designed for neural architectures from topological search spaces. In addition, we employ ranking losses to train the performance predictor, as correct relative ordering among candidates rather than precise performance estimation is what matters for high sample efficiency. Under the same ranking loss, GATES significantly outperforms baseline encoders in correctly ranking unseen architectures (e.g., achieving a Kendall's Tau of 0.7634 vs. 0.5509 on NAS-Bench-101). Moreover, with the same encoder, training the performance predictor using ranking losses consistently outperforms regression-based training.

3. **Xuefei Ning**, Changcheng Tang, Wenshuo Li, Zixuan Zhou, and Others, "Evaluating Efficient Performance Estimators of Neural Architectures", In NeurIPS 2021. [Website]
   Assess one-shot and zero-shot evaluators of architecture performance: To reduce the architecture training costs in NAS, one-shot evaluators (OSEs) amortize the architecture training costs by sharing the parameters of one supernet between all architectures. Zero-shot evaluators (ZSEs) involve no training are proposed to further reduce the architecture evaluation cost. This work assesses OSEs and ZSEs on five NAS benchmarks. Specifically, we employ a set of NAS-oriented criteria to study the behavior of OSEs and ZSEs and

reveal that they have certain biases and variances. After analyze how and why the estimations are unsatisfying, we explore how to mitigate the correlation gap of OSEs from several perspectives.

4. **Xuefei Ning***, Zinan Lin*, Zixuan Zhou*, Zifu Wang, and Others, "Skeleton-of-Thought: Prompting Large Language Models for Efficient Parallel Generation", In ICLR 2024. [Website]
Let LLMs organize their own output structure and leverage parallelism for efficient generation: This work proposes Skeleton-of-Thought (SoT) that prompts off-the-shelf LLMs to first generate a high-level skeleton of an answer and then complete the corresponding contents via batched decoding in parallel. This approach increases hardware utilization and reduces end-to-end generation latency. Not all questions admit a fully parallelizable skeleton; therefore, we introduce a question-level router to determine whether a given query is suitable for SoT-based generation. Although SoT is not fully practical, it is conceptually novel in that it explores a new data- or agent-level pathway for efficiency improvement by allowing the LLM to explicitly organize its own output structure. The idea can be extended by replacing the parallel skeleton with a more general graph structure, or by enabling the model to trigger parallel generation dynamically during decoding (e.g., in a manner similar to tool invocation), rather than relying on in-advance skeleton planning.

5. **Xuefei Ning***, Zifu Wang*, Shiyao Li*, Zinan Lin*, Peiran Yao*, and Others, "Can LLMs Learn by Teaching for Better Reasoning? A Preliminary Study", In NeurIPS 2024. [Website]
Explore whether current LLMs can learn by teaching weaker models for better reasoning: In human learning, teaching enhances not only the students but also the teachers by fostering more rigorous and clearer reasoning, as well as deeper knowledge building. This work explores "can LLMs also learn by teaching (LbT) for better reasoning". We implement the LbT idea into existing LLM training/prompting pipelines (search-based output generation, generation-scoring-finetuning, input prompt optimization), and observe some improvements.

Systematic survey and tutorial on efficient deep learning:

- Zixuan Zhou*, **Xuefei Ning***+, Ke Hong*, Tianyu Fu, and Others, "A Survey on Efficient Inference for Large Language Models", In arXiv 2024.
A survey on the algorithm-, model-, and system-level techniques for efficient LLM inference
- **Xuefei Ning**, Guohao Dai, Haoli Bai, Lu Hou, Yu Wang, Qun Liu, "Efficient Inference for Large Language Models – Algorithm, Model, and System", EMNLP 2025 Tutorial. Website: [Website]
A tutorial on the algorithm-, model-, and system-level techniques for efficient LLM inference
- Yu Wang, **Xuefei Ning**, "Efficient Deep Learning: Model Compression and Design", Published by the Publishing House of Electronics Industry, 2024/07.
A Chinese book introducing model compression and efficient model design techniques systematically

PUBLICATIONS * INDICATES EQUAL CONTRIBUTION, + INDICATES CORRESPONDING AUTHOR & MAIN ADVISOR

Note: Representative publications are highlighted in the brown color.

Towards Better Reasoning and Learning

- **Xuefei Ning***, Zifu Wang*, Shiyao Li*, Zinan Lin*, Peiran Yao*, and Others, "Can LLMs Learn by Teaching for Better Reasoning? A Preliminary Study", In NeurIPS 2024.

Efficient NN Inference

- Zixuan Zhou*, **Xuefei Ning***+, Ke Hong*, Tianyu Fu, and Others, "A Survey on Efficient Inference for Large Language Models", In arXiv 2024.

1. **Algorithm-level Techniques**:
   - **[Vision Generation]** Enshu Liu*, **Xuefei Ning***+, Zinan Lin*, Huazhong Yang, Yu Wang+, "OMS-DPM: Deciding The Optimal Model Schedule for Diffusion Probabilistic Model", In ICML 2023.
   - **[Language Generation] Xuefei Ning***, Zinan Lin*, Zixuan Zhou*, Zifu Wang, and Others, "Skeleton-of-Thought: Prompting Large Language Models for Efficient Parallel Generation", In ICLR 2024.
   - **[Vision Generation]** Enshu Liu, **Xuefei Ning**+, Huazhong Yang, Yu Wang+, "A Unified Sampling Framework for Solver Searching of Diffusion Probabilistic Models", In ICLR 2024.

- **[Vision Generation]** Enshu Liu, **Xuefei Ning**, Yu Wang, Zinan Lin+, "Distilling Auto-regressive Models into Few Steps 1: Image Generation", In ICLR 2025.
- **[Vision Generation]** Yao Teng, Han Shi, Xian Liu, **Xuefei Ning**, and Others, "Accelerating Auto-regressive Text-to-Image Generation with Training-free Speculative Jacobi Decoding", In ICLR 2025.

2. **Model-level Design or Compression**:
   - **(Chinese Book)** Yu Wang, **Xuefei Ning**, "Efficient Deep Learning: Model Compression and Design", Published by the Publishing House of Electronics Industry, 2024/07. (Introduce the model compression and efficient model design techniques systematically. The project manager and also the main writer, writing ≥80% proportion of the final contents.)
   - **Xuefei Ning***, Tianchen Zhao*, Wenshuo Li, and Others, "DSA: More Efficient Budgeted Pruning via Differentiable Sparsity Allocation", In ECCV 2020 (**Spotlight**).
   - Xiangsheng Shi*, **Xuefei Ning***+, Lidong Guo*, Tianchen Zhao, and Others, "Memory-Oriented Structural Pruning for Efficient Image Restoration", In AAAI 2023.
   - Tianchen Zhao, **Xuefei Ning**+, Ke Hong, Zhongyuan Qiu, and Others, "Ada3D: Exploiting the Spatial Redundancy with Adaptive Inference for Efficient 3D Object Detection", In ICCV 2023.
   - **[Language Generation]** Shiyao Li, **Xuefei Ning**+, Luning Wang, Tengxuan Liu, and Others, "Evaluating Quantized Large Language Models", In ICML 2024.
   - **[Vision Generation]** Tianchen Zhao*, **Xuefei Ning***+, Tongcheng Fang*, and Others, "MixDQ: Memory-Efficient Few-Step Text-to-Image Diffusion Models with Metric-Decoupled Mixed Precision Quantization", In ECCV 2024.
   - **[Vision Generation]** Zhihang Yuan*, Hanling Zhang*, Pu Lu*, **Xuefei Ning**+, and Others, "DiTFastAttn: Attention Compression for Diffusion Transformer Models", In NeurIPS 2024.
   - **[Vision Generation]** Tianchen Zhao, Tongcheng Fang, ..., **Xuefei Ning**+, Yu Wang+, "ViDiT-Q: Efficient and Accurate Quantization of Diffusion Transformers for Image and Video Generation", In ICLR 2025.
   - **[Language Generation]** Tianyu Fu*, Haofeng Huang*, **Xuefei Ning***+, Genghan Zhang, and Others, "MoA: Mixture of Sparse Attention for Automatic Large Language Model Compression", In CoLM 2025.
   - **[Language Generation]** Enshu Liu*, Junyi Zhu*, Zinan Lin+, **Xuefei Ning**+, and Others, "Efficient Expert Pruning for Sparse Mixture-of-Experts Language Models: Enhancing Performance and Reducing Inference Costs", In arXiv 2024.
   - **[Multi-Modality]** Shiyao Li*, Yingchun Hu*, **Xuefei Ning**+, Xihui Liu, Ke Hong, Xiaotao Jia+, Xiuhong Li, Yaqi Yan, Pei Ran, Guohao Dai, Shengen Yan, Huazhong Yang, Yu Wang+, "MBQ: Modality-Balanced Quantization for Large Vision-Language Models", In CVPR 2025.

Efficient NN Training

- Kai Zhong, **Xuefei Ning**, Guohao Dai, Zhenhua Zhu, and Others, "Exploring the Potential of Low-bit Training of Convolutional Neural Networks", In TCAD 2022.
- Minxue Tang, **Xuefei Ning**, Yitu Wang, Jingwei Sun, and Others, "FedCor: Correlation-Based Active Client Selection Strategy for Heterogeneous Federated Learning", In CVPR 2022.
- **[Vision Generation]** Enshu Liu*, Junyi Zhu*, Zinan Lin+, **Xuefei Ning**+, and Others, "Linear Combination of Saved Checkpoints Makes Consistency and Diffusion Models Better", In ICLR 2025.

Efficient and Automatic Optimization Process

1. **Efficient Neural Architecture Search:**
   Summary Website: https://sites.google.com/view/nas-nicsefc
   Unified NAS Framework (containing following researches): https://github.com/walkerning/aw_nas
   - **Xuefei Ning**, Yin Zheng, Tianchen Zhao, Yu Wang, Huazhong Yang, "A Generic Graph-based Neural Architecture Encoding Scheme for Predictor-based NAS", In ECCV 2020.
   - Wenshuo Li*, **Xuefei Ning***, Guangjun Ge, Xiaoming Chen, Yu Wang, Huazhong Yang, "FTT-NAS: Discovering Fault-Tolerant Neural Architecture", In ASP-DAC 2020.
   - Shulin Zeng*, Hanbo Sun*, Yu Xing, **Xuefei Ning**, Yi Shan, Xiaoming Chen, Yu Wang, Huazhong Yang, "Black Box Search Space Profiling for Accelerator-Aware Neural Architecture Search", In ASP-DAC 2020.
   - **Xuefei Ning**, Changcheng Tang, Wenshuo Li, Zixuan Zhou, and Others, "Evaluating Efficient Performance Estimators of Neural Architectures", In NeurIPS 2021.

- **Xuefei Ning**, Guangjun Ge, Wenshuo Li, Zhenhua Zhu, and Others, "FTT-NAS: Discovering Fault-Tolerant Convolutional Neural Architecture, In TODAES 2021.
- Zixuan Zhou*, **Xuefei Ning***+, Yi Cai, Jiashu Han, and Others, "CLOSE: Curriculum Learning On the Sharing Extent Towards Better One-shot NAS", In ECCV 2022.
- **Xuefei Ning***, Zixuan Zhou*, Junbo Zhao, Tianchen Zhao, and Others, "TA-GATES: An Encoding Scheme for Neural Network Architectures", In NeurIPS 2022 (**Spotlight**).
- Junbo Zhao*, **Xuefei Ning***+, Enshu Liu, Binxin Ru, and Others, "Dynamic Ensemble of Low-fidelity Experts: Mitigating NAS Cold-Start", In AAAI 2023 (**Oral**).
- **Xuefei Ning**, Yin Zheng, Zixuan Zhou, Tianchen Zhao, Huazhong Yang, Yu Wang, "A Generic Graph-based Neural Architecture Encoding Scheme with Multifaceted Information", In TPAMI 2023.
- Hanbo Sun, Zhenhua Zhu, Chenyu Wang, **Xuefei Ning+**, and Others, "Gibbon: Efficient Co-Exploration of NN Model and Processing-In-Memory Architecture", In DATE 2022 & TCAD 2023.

2. **Efficient Performance Evaluation:**
   - **[Vision Generation]** Lin Zhao*, Tianchen Zhao*, Zinan Lin, **Xuefei Ning+**, and Others, "FlashEval: Towards Fast and Accurate Evaluation of Text-to-image Diffusion Generative Models", In CVPR 2024.

Other Researches

- **Xuefei Ning**, Yin Zheng, Zhuxi Jiang, Yu Wang, and Others, "Nonparametric Topic Modeling with Neural Inference", In Neurocomputing 2020.
- Tong Wu, **Xuefei Ning**, Wenshuo Li, Ranan Huang, Huazhong Yang, Yu Wang, "Physical Adversarial Attack on Vehicle Detector in the Carla Simulator", Technical report in arXiv 2020.
- Ye Mu*, Weilin Liu*, Chao Yu, **Xuefei Ning+**, and Others, "Multi-Agent Vulnerability Discovery for Autonomous Driving with Hazard Arbitration Reward", In CASE 2024.
- Lidong Guo*, **Xuefei Ning***+, Yonggan Fu, Tianchen Zhao, and Others, "Rad-NeRF: Ray-decoupled Training of Neural Radiance Field", In NeurIPS 2024.
- Tao Yuan, **Xuefei Ning+**, Dong Zhou, Zhijie Yang, and Others, "LV-Eval: A Balanced Long-Context Benchmark with 5 Length Levels Up to 256K", In CoLM 2025.
- Kaiyi Huang, Yukun Huang, **Xuefei Ning**, Zinan Lin, Yu Wang, Xihui Liu, "GenMAC: Compositional Text-to-Video Generation with Multi-Agent Collaboration", In AAAI 2026.
- Federico Marcuzzi, **Xuefei Ning**, Roy Schwartz, Iryna Gurevych, "How Quantization Shapes Bias in Large Language Models", In EACL 2026.
- Anyuan Zhuo, **Xuefei Ning+**, Ningyuan Li, Yu Wang, Pinyan Lu+, "Understanding the Ability of LLMs to Handle Character-Level Perturbation", ICML 2026.

Last update: 2026/05