HANCHENG YE

[Tel]: +1-(919) 672-8631 | [Email]: hancheng.ye@duke.edu | [Personal Page]: Homepage
[Google Scholar]: My Scholar Page | [Address]: 821 Louise Cir, Durham, NC 27705, USA

EDUCATION

Duke University *Durham, USA*
Ph.D. in Electrical and Computer Engineering Sep 2024 – Jun 2029 (Expected)

- GPA: 4.00/4.00 (Rank: N/A); supervised by Prof. Yiran Chen
- Research interests:
  - **Machine Learning Systems**: Cache-based Efficient LLM / Diffusion Serving & Parallelization in LLM-based Multi-agent Systems
  - **Multimodal Learning for Science**: Chart- / Geometry-related Multimodal Reasoning

Fudan University *Shanghai, China*
M.S. in Circuits and Systems Sep 2020 – Jun 2023

- GPA: 3.72/4.00 (Rank: 2/55); worked with Prof. Tao Chen and Dr. Bo Zhang
- Research interests: Efficient Multitask Inference, Model Compression, Few-shot Learning

B.E. in Biomedical Engineering Sep 2016 – Jun 2020

- GPA: 3.76/4.00 (Rank: 7/190); selected to Elite Engineering Program (top 5%)

PUBLICATIONS

Efficient Machine Learning

1. **H. Ye**, Z. Gao, M. Ma, Q. Wang, Y. Fu, M.-Y. Chung, Y. Lin, Z. Liu, J. Zhang, D. Zhuo, & Y. Chen. (2025). KVCOMM: Online cross-context KV-cache communication for efficient LLM-based multi-agent systems. In *Advances in Neural Information Processing Systems (NeurIPS 2025)*.

2. T. Jiang*, Y. Wang*, **H. Ye*** (**co-first author**), Z. Shao, J. Sun, J. Zhang, Z. Chen, J. Zhang, Y. Chen, & H. Li. (2025). SADA: Stability-guided adaptive diffusion acceleration. In *Proceedings of the 42nd International Conference on Machine Learning (ICML 2025)*.

3. **H. Ye*** (**co-first author**), J. Yuan*, R. Xia, X. Yan, T. Chen, J. Yan, B. Shi, & B. Zhang. (2024). Training-free adaptive diffusion with bounded difference approximation strategy. In *Advances in Neural Information Processing Systems (NeurIPS 2024)*.

4. **H. Ye**, C. Yu, P. Ye, R. Xia, Y. Tang, J. Lu, T. Chen, & B. Zhang. (2024). Once for both: Single stage of importance and sparsity search for vision transformer compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2024)*, 5578–5588.

5. **H. Ye**, B. Zhang, T. Chen, J. Fan, & B. Wang. (2023). Performance-aware approximation of global channel pruning for multitask CNNs. *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI 2023), 45*(8), 10267–10284.

6. Q. Wang, J. Ke, **H. Ye**, Y. Lin, Y. Fu, J. Zhang, K. Keutzer, C. Xu, & Y. Chen. (2025). Angles don't lie: Unlocking training-efficient RL through the model's own signals. In *Advances in Neural Information Processing Systems (NeurIPS 2025)*.

7. Q. Wang, **H. Ye**, M.-Y. Chung, Y. Liu, Y. Lin, M. Kuo, M. Ma, J. Zhang, & Y. Chen. (2025). CoreMatching: A co-adaptive sparse inference framework with token and neuron pruning for comprehensive acceleration of vision–language models. In *Proceedings of the 42nd International Conference on Machine Learning (ICML 2025)*.

8. C. Tu, P. Ye, W. Lin, **H. Ye**, C. Yu, T. Chen, B. Li, & W. Ouyang. (2024). Efficient architecture search via bi-level data pruning. *IEEE Transactions on Circuits and Systems for Video Technology (T-CSVT 2024)*.

9. S. Tang, W. Lin, **H. Ye**, P. Ye, C. Yu, B. Li, & T. Chen. (2024). Enhanced sparsification via stimulative training. In *European Conference on Computer Vision (ECCV 2024)*.

10. B. Zhang, **H. Ye**, G. Yu, B. Wang, Y. Wu, J. Fan, & T. Chen. (2022). Sample-centric feature generation for semi-supervised few-shot learning. *IEEE Transactions on Image Processing (T-IP 2022), 31*, 2309–2320.

Multimodal Learning

1. R. Xia*, **H. Ye*** (**co-first author**), X. Yan, Q. Liu, H. Zhou, Z. Chen, M. Dou, B. Shi, J. Yan, Y. Qiao, & B. Zhang*. (2025). ChartX & ChartVLM: A versatile benchmark and foundation model for complicated chart reasoning. *IEEE Transactions on Image Processing (T-IP 2025)*.

2. R. Xia*, M. Li*, **H. Ye**, W. Wu, H. Zhou, J. Yuan, T. Peng, X. Cai, X. Yan, B. Wang, C. He, B. Shi, T. Chen, J. Yan, & B. Zhang. (2025). GeoX: Geometric problem solving through unified formalized vision–language pre-training. In *International Conference on Learning Representations (ICLR 2025)*.

PROFESSIONAL EXPERIENCE

Duke University *Durham, USA*
Research Assistant — Supervisor: Prof. Yiran Chen Sep 2024 – Feb 2029 (Expected)

- Created KVCOMM, the first KV-cache communication framework for efficient LLM-based multi-agent systems.
- Co-developed a stability-guided adaptive diffusion generation paradigm supporting training-free acceleration of various modality generative models without quality degradation.

Shanghai Artificial Intelligence Laboratory *Shanghai, China*
Research Intern — Mentor: Dr. Bo Zhang Oct 2023 – Aug 2024

- Co-designed a structured representation to enhance chart understanding. Proposed a two-stage framework with an explicit high-dimensional chart embedding formulation.

Massachusetts Institute of Technology *Cambridge, USA*
Research Intern — Mentor: Dr. Zhijian Liu Jun 2023 – May 2024

- Developed the first prompt-adaptive acceleration strategy for efficient high-quality image/video diffusion generation.

TEACHING EXPERIENCE

Fudan University *Shanghai, China*
Teaching Assistant — Instructor: Prof. Tao Chen Sep 2021 – Jan 2022

MENTORING EXPERIENCE

Ting Jiang, Duke University B.E.	Sep 2024 – Sep 2025
Guangzhi Su, Duke Kunshan University B.E.	Dec 2024 – Present
Dongting Li, Duke University Ph.D.	Sep 2025 – Present
Baijun Zhou, Duke University Ph.D.	Sep 2025 – Present
Alejandro Paredeslato, Duke University M.S.	Sep 2025 – Present

ADDITIONAL INFORMATION

Professional Services

- Reviewer: NeurIPS, ICML, ICLR, CVPR, ECCV, AAAI, ACM MM, AISTATS, IJCV, etc.

Miscellaneous

- Programming: Python, PyTorch, MATLAB, CUDA C++, etc.
- Languages: English (proficient), Mandarin Chinese (native)

REFERENCES

- Yiran Chen, Ph.D., IEEE Fellow, NAI Fellow — John Cocke Distinguished Professor in ECE, Duke; yiran.chen@duke.edu
- Hai Li, Ph.D., IEEE Fellow, NAI Fellow — Marie Foote Reel E'46 Distinguished Professor of ECE, Duke; hai.li@duke.edu
- Danyang Zhuo, Ph.D. — Assistant Professor of CS, Duke; danyang@cs.duke.edu