Hongxu (Danny) Yin

dannyy@nvidia.com • https://hongxu-yin.github.io/ • Google Scholar

Research

My research focuses on multimodal intelligence with large-scale language and vision training. I am a TL for NVIDIA's herd of multimodal models (VILA and its model family series) on post-training, agent, and reasoning aspects of multimodal LLMs, with full-stack optimization for NVIDIA's hardware.

Experience

NVIDIA Research
Staff Research Scientist & Tech Lead, Learning and Perception Research (LPR) *Apr. 2024 - Now*

NVIDIA Research
Senior Research Scientist, Learning and Perception Research (LPR) *May 2022 - Mar. 2024*

NVIDIA Research
Research Scientist, Learning and Perception Research (LPR) *May 2020 - Apr. 2022*

NVIDIA Research
Research Intern, Learning and Perception Research (LPR) *May 2019 - Nov. 2019*

Alibaba U.S.
Research Intern, Machine Learning Team *May 2018 - Nov. 2018*

Education

Princeton University New Jersey, USA
Ph.D. in Electrical Computer Engineering, advised by Prof. Niraj K. Jha *2015 - 2020*
Research focus: Efficient and Secure Deep Learning

Nanyang Technological University Singapore, SG
B.Eng in Electronic & Electronics Engineering (GPA 3.9/4.0, dean's lister all four years) *2011 - 2015*
Minor in Business (GPA 4.0/4.0)

University of California, Berkeley California, USA
Undergraduate summer exchange *2012*

University of Cambridge Cambridge, UK
High school elite exchange program *2007*

Selected Awards

- 36 Kr Top 100 Global Outstanding Chinese Awards
- Forbes Top 60 Elite Chinese North America
- Princeton ECE Best Dissertation Award Finalist (Top-3 in department)
- Princeton Yan Huo *94 Fellowship (Top-3 in department)
- Princeton Natural Science and Foundation Fellowship
- Gold Medal - Defense Science and Technology
- Gold Medal - Thomas Asia Pacific Holdings
- Department Dean's Lister Award
- Nanyang Best Industrial Orientation Award
- Nanyang Presidential Scholar with Highest Distinction

Conference Publications

(*: equal contribution; †: advised intern; ◇: equal advisory;)

52. Baifeng Shi†, Boyi Li, Han Cai, Yao Lu, Sifei Liu, Marco Pavone, Jan Kautz, Song Han, Trevor Darrell, Pavlo Molchanov, **Hongxu Yin**

VILA-HD: Scaling vision pre-training to 4K resolution
Conference on Computer Vision and Pattern Recognition (CVPR), 2025
(Highlight Paper)

51. Zhijian Liu, Ligeng Zhu, Baifeng Shi, Zhuoyang Zhang, Yuming Lou, Shang Yang, Haocheng Xi, Shiyi Cao, Yuxian Gu, Dacheng Li, Xiuyu Li, Haotian Tang, Yunhao Fang, Yukang Chen, Cheng-Yu Hsieh, De-An Huang, An-Chieh Cheng, Jinyi Hu, Sifei Liu, Ranjay Krishna, Pavlo Molchanov, Jan Kautz, **Hongxu Yin**$^{\diamond}$, Song Han$^{\diamond}$, Yao Lu$^{\diamond}$
NVILA: Efficient frontier visual language models
Conference on Computer Vision and Pattern Recognition (CVPR), 2025

50. Greg Heinrich, Mike Ranzinger, **Hongxu Yin**, Yao Lu, Jan Kautz, Bryan Catanzaro, Andrew Tao, Pavlo Molchanov
RADIO Amplified: Improved baselines for agglomerative vision foundation models
Conference on Computer Vision and Pattern Recognition (CVPR), 2025

49. Vishwesh Nath, Wenqi Li, Dong Yang, Andriy Myronenko, Yao Lu, Zhijian Liu, **Hongxu Yin**, Yucheng Tang, Pengfei Guo, Ziyue Xu, Can Zhao, Yufan He, Greg Heinrich, Mingxin Zheng, Benjamin D. Simon, Stephanie Anne Harmon, Michael Zephyr, Marc Edgar, Stephen R. Aylward, Pavlo Molchanov, Yan Mee Law, Baris Turkbey, Holger R Roth, Daguang Xu
VILA-M3: Enhancing vision-language models with medical expert knowledge
Conference on Computer Vision and Pattern Recognition (CVPR), 2025
(Highlight Paper)

48. An-Chieh Cheng, Yandong Ji, Zhaojing Yang, Zaitian Gongye, Xueyan Zou, Jan Kautz, Erdem Bıyık, **Hongxu Yin**$^{\diamond}$, Sifei Liu$^{\diamond}$, Xiaolong Wang$^{\diamond}$
NaVILA: Legged Robot Vision-Language-Action Model for Navigation
Robotics: Science and Systems (RSS), 2025

47. Yecheng Wu, Zhuoyang Zhang, Junyu Chen, Haotian Tang, Dacheng Li, Yunhao Fang, Ligeng Zhu, Enze Xie, **Hongxu Yin**, Li Yi, Song Han, Yao Lu
VILA-U: A unified foundation model integrating visual understanding and generation
International Conference on Learning Representations (ICLR), 2025

46. Fuzhao Xue†*, Yukang Chen†*, Dacheng Li†*, Qinghao Hu†*, Ligeng Zhu, Xiuyu Li, Yunhao Fang, Haotian Tang, Shang Yang, Zhijian Liu, Ethan He, **Hongxu Yin**, Pavlo Molchanov, Jan Kautz, Linxi Fan, Yuke Zhu, Yao Lu, Song Han
LongVILA: Scaling long-context visual language models for long videos
International Conference on Learning Representations (ICLR), 2025

45. Min Shi, Fuxiao Liu, Shihao Wang, Shijia Liao, Subhashree Radhakrishnan, De-An Huang, **Hongxu Yin**, Karan Sapra, Yaser Yacoob, Humphrey Shi, Bryan Catanzaro, Andrew Tao, Jan Kautz, Zhiding Yu, Guilin Liu
EAGLE: Exploring the design space for multimodal LLMs with mixture of encoders
International Conference on Learning Representations (ICLR), 2025
(Spotlight Paper)

44. Ruisi Cai, Saurav Muralidharan, **Hongxu Yin**, Zhangyang Wang, Jan Kautz, Pavlo Molchanov
LLaMaFlex: Many-in-one LLMs via generalized pruning and weight sharing
International Conference on Learning Representations (ICLR), 2025

43. Gongfan Fang†, **Hongxu Yin**, Saurav Muralidharan, Greg Heinrich, Jeff Pool, Jan Kautz, Pavlo Molchanov, Xinchao Wang
MaskLLM: Learnable semi-structured sparsity for large language models
Advances in Neural Information Processing Systems (NeurIPS), 2024
(Spotlight Paper)

42. An-Chieh Cheng†, **Hongxu Yin**, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, Sifei Liu
SpatialRGPT: Grounded spatial reasoning in vision language model
Advances in Neural Information Processing Systems (NeurIPS), 2024

41. Ji Lin†*, **Hongxu Yin***, Wei Ping, Yao Lu, Pavlo Molchanov, Andrew Tao, Huizi Mao, Jan Kautz, Mohammad Shoeybi, Song Han
VILA: On pre-training for visual language models
Conference on Computer Vision and Pattern Recognition (CVPR), 2024

40. Qiushan Guo†, Shalini De Mello*, **Hongxu Yin***, Wonmin Byeon, Ka Chun Cheung, Yizhou Yu, Ping Luo, Sifei Liu

RegionGPT: Towards region understanding vision language model
Conference on Computer Vision and Pattern Recognition (CVPR), 2024

39. Shih-Yang Liu†, Chien-Yi Wang, **Hongxu Yin**, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, Min-Hung Chen
DoRA: Weight-decomposed low-rank adaptation
International Conference on Machine Learning (ICML), 2024
(Oral Presentation)

38. Jingwen Sun†, Ziyue Xu, **Hongxu Yin**, Dong Yang, Daguang Xu, Yiran Chen, Holger R. Roth
FedBPT: Efficient federated black-box prompt tuning for large language models
International Conference on Machine Learning (ICML), 2024
AAAI Symposium, 2024 **(Best Paper Award)**

37. Ruisi Cai†, Saurav Muralidharan, Greg Henrich, **Hongxu Yin**, Zhangyang Wang, Jan Kaut, Pavlo Molchanov
FlexTron: Many-in-One flexible large language models
International Conference on Machine Learning (ICML), 2024
(Oral Presentation)

36. De-an Huang, Shijia Liao, Subhashree Radhakrishnan, **Hongxu Yin**, Pavlo Molchanov, Zhiding Yu, Jan Kautz
LITA: Language instructed temporal-localization assistant
European Conference on Computer Vision (ECCV), 2024

35. Anna Bair†, **Hongxu Yin**, Maying Shen, Pavlo Molchanov, Jose M. Alvarez
Adaptive sharpness-aware pruning for robust sparse networks
International Conference on Learning Representations (ICLR), 2024

34. Ali Hatamizadeh, Greg Heinrich, **Hongxu Yin**, Andrew Tao, Jose M. Alvarez, Jan Kautz, Pavlo Molchanov
FasterViT: Fast vision transformers with hierarchical attention
International Conference on Learning Representations (ICLR), 2024

33. Xinlong Sun†, Maying Shen, **Hongxu Yin**, Lei Mao, Pavlo Molchanov, Jose M Alvarez
Advancing weight and channel sparsification with enhanced saliency
Winter Conference on Applications of Computer Vision (WACV), 2024

32. Jiaming Song, Qinsheng Zhang, **Hongxu Yin**, Morteza Mardani, Ming-yu Liu, Jan Kautz, Yongxin Chen, Arash Vahdat
Loss-guided diffusion models for Plug-and-Play controllable generation
International Conference on Machine Learning (ICML), 2023

31. Ali Hatamizadeh, **Hongxu Yin**, Jan Kautz, Pavlo Molchanov
Global context vision transformer
International Conference on Machine Learning (ICML), 2023

30. Divyam Madaan†, **Hongxu Yin**, Wonmin Byeon, Jan Kautz, Pavlo Molchanov
Heterogeneous continual learning
Conference on Computer Vision and Pattern Recognition (CVPR), 2023
(Highlight Paper)

29. Huanrui Yang†, **Hongxu Yin**, Pavlo Molchanov, Hai Li, Jan Kautz
NViT: Vision transformer compression and parameter redistribution
Conference on Computer Vision and Pattern Recognition (CVPR), 2023

28. Paul Micaelli†, Pavlo Molchanov, Arash Vahdat, **Hongxu Yin**, Jan Kautz
Recurrence without recurrence: stable video landmark detection with deep equilibrium models
Conference on Computer Vision and Pattern Recognition (CVPR), 2023

27. Xin Dong†, **Hongxu Yin**, Jose Alvarez, Jan Kautz, Pavlo Molchanov
Privacy vulnerability of split computing to data-free model inversion attacks
British Machine Vision Conference (BMVC), 2022

26. Maying Shen*, **Hongxu Yin***, Pavlo Molchanov, Lei Mao, Jianna Liu, Jose Alvarez
Structural pruning via latency-saliency Knapsack
Advances in Neural Information Processing Systems (NeurIPS), 2022

25. **Hongxu Yin**, Arash Vahdat, Jose Alvarez, Arun Mallya, Jan Kautz, Pavlo Molchanov
A-ViT: Adaptive tokens for efficient vision transformer
Conference on Computer Vision and Pattern Recognition (CVPR), 2022
(Oral Presentation)

24. Ali Hatamizadeh*, **Hongxu Yin***, Holger Roth, Wenqi Li, Jan Kautz, Daguang Xu, Pavlo Molchanov
GradViT: Gradient inversion of vision transformers
Conference on Computer Vision and Pattern Recognition (CVPR), 2022

23. Maying Shen, Pavlo Molchanov, **Hongxu Yin**, Jose Alvarez
When to prune? A policy towards early structural pruning
Conference on Computer Vision and Pattern Recognition (CVPR), 2022

22. Pavlo Molchanov*, Jimmy Hall*, **Hongxu Yin***, Jan Kautz, Nicolo Fusi, Arash Vahdat
HANT: Hardware-aware network transformation
European Conference on Computer Vision (ECCV), 2022

21. **Hongxu Yin**, Arun Mallya, Arash Vahdat, Jose Alvarez, Jan Kautz, Pavlo Molchanov
See through gradients: Image batch recovery via GradInversion
Conference on Computer Vision and Pattern Recognition (CVPR), 2021

20. Yerlan Idelbayev†, Pavlo Molchanov, Maying Shen, **Hongxu Yin**, M. C. Perpinan, Jose Alvarez
Optimal quantization using scaled codebook
Conference on Computer Vision and Pattern Recognition (CVPR), 2021

19. Akshay Chawla†, **Hongxu Yin**, Pavlo Molchanov, Jose Alvarez
Data-free knowledge distillation for object detection
Winter Conference on Applications of Computer Vision (WACV), 2021

18. **Hongxu Yin**, Arun Mallya, Arash Vahdat, Jose Alvarez, Jan Kautz, Pavlo Molchanov
Dreaming to distill: Data-free knowledge transfer via DeepInversion
Conference on Computer Vision and Pattern Recognition (CVPR), 2020
(Oral Presentation)

17. Wenhan Xia, **Hongxu Yin**, Niraj K. Jha
Efficient synthesis of compact deep neural networks
IEEE Design Automation Conference (DAC), 2020

16. Xiaoliang Dai, Peizhao Zhang, Bichen Wu, **Hongxu Yin**, Fei Sun, Yanghan Wang, Marat Dukhan, Yunqing Hu, Yiming Wu, Yangqing Jia, Peter Vajda, Matt Uyttendaele, Niraj K. Jha
ChamNet: Towards efficient network design through platform-aware model adaptation
Conference on Computer Vision and Pattern Recognition (CVPR), 2019

15. Ozge Akmandor, **Hongxu Yin**, and Niraj K. Jha
Simultaneously ensuring smartness, security, and energy efficiency in Internet-of-Things sensors
IEEE Custom Integrated Circuits Conference (CICC), 2017

14. **Hongxu Yin**, Bah Hwee Gwee, Zhiping Lin, Kumar Anil, Galul R. Sirajudeen, and Choo M. S. See
Novel real-time system design for floating-point sub-Nyquist multi-coset signal blind reconstruction
IEEE Int. Symp. on Circuits and Systems (ISCAS), 2015
(Oral Presentation)

Journal Publications

13. Ali Hatamizadeh, **Hongxu Yin**, Pavlo Molchanov, Andriy Myronenko, Wenqi Li, Prerna Dogra, Andrew Feng, Mona G Flores, Jan Kautz, Daguang Xu, Holger R. Roth
Do gradient inversion attacks make federated learning unsafe?
IEEE Transactions on Medical Imaging, 2023

12. Shayan Hassantabar, Joe Zhang, **Hongxu Yin**, Niraj K. Jha
MHDeep: Mental health disorder detection system based on body-area and deep neural networks
ACM Transactions on Embedded Computing Systems, 2022

11. **Hongxu Yin**, Guoyang Chen, Yingmin Li, Shuai Che, Weifeng Zhang, and Niraj K. Jha
Hardware-guided symbiotic training for compact, accurate, yet execution-efficient LSTMs
IEEE Trans. Emerging Topics in Computing, 2021

10. Wenhan Xia, **Hongxu Yin**, Xiaoliang Dai, Niraj K. Jha
Fully dynamic inference with deep neural networks
IEEE Trans. Emerging Topics in Computing, 2021

9. Xiaoliang Dai*, **Hongxu Yin***, and Niraj K. Jha
Grow and prune compact, fast, and accurate LSTMs
IEEE Trans. Computers, 2020

8. **Hongxu Yin**, Bilal Mukadam, Xiaoliang Dai, and Niraj K. Jha
DiabDeep: Pervasive diabetes diagnosis based on wearable medical sensors and efficient neural networks
IEEE Trans. Emerging Topics in Computing, 2020

7. Xiaoliang Dai, **Hongxu Yin**, and Niraj K. Jha
Incremental learning using a grow-and-prune paradigm with efficient neural networks
IEEE Trans. Computers, 2020

6. Xiaoliang Dai, **Hongxu Yin**, and Niraj K. Jha
NeST: A neural network synthesis tool based on a grow-and-prune paradigm
IEEE Trans. Computers, 2019

5. **Hongxu Yin**, Zeyu Wang, and Niraj K. Jha
A hierarchical inference model for Internet-of-Things
IEEE Trans. Multi-scale Computing Systems, 2018

4. **Hongxu Yin** and Niraj K. Jha
A health decision support system for disease diagnosis based on wearable medical sensors and machine learning ensembles
IEEE Trans. Multi-scale Computing Systems, 2017

3. Ozge Akmandor, **Hongxu Yin** and Niraj K. Jha
Smart, secure, yet energy-efficient, Internet-of-Things sensors
IEEE Trans. Multi-scale Computing Systems, 2017

Technical Report

2. NVIDIA Corporation
Nemotron-H: A family of accurate and efficient hybrid mamba-transformer models
Technical Report, 2025

Book Chapter

1. **Hongxu Yin**, Ozge Akmandor, Arsalan Mosenia, and Niraj K. Jha
Smart healthcare
Foundations and Trends, 2017

Workshop & Tutorial Organizer

- *Efficient Computer Vision Workshop*
CVPR 2025 *2025*
- *Full-Stack, GPU-based Acceleration of Deep Learning Tutorial*
CVPR 2025 *2025*
- *Efficient Deep Learning for Foundation Models Workshop*
ECCV 2024 *2024*
- *Efficient Computer Vision Workshop*
CVPR 2024 *2024*
- *Full-Stack, GPU-based Acceleration of Deep Learning Tutorial*
CVPR 2024 *2024*
- *Data-efficient Learning for Large Model Tutorial*
ICCV 2023 *2023*
- *Full-Stack, GPU-based Acceleration of Deep Learning Tutorial*
CVPR 2023 *2023*
- *Transformers for Vision Workshop*
CVPR 2022 *2022*

Invited Keynote & Talk (till Dec. 2022)

- *Efficient Deep Learning*
Invited Panelist, Open Compute Project (OCP) Global Summit *Oct. 2022*

- *Towards Efficient and Secure Deep Learning*
  Invited Keynote, Design & Automation Conference (DAC'60) *Jul. 2022*
- *Towards Efficient and Secure Deep Nets*
  University of British Columbia ECE Department *May 2022*
- *Inverting Deep Nets*
  Princeton University, Department of Computer Science research groups *Aug. 2021*
- *See through Gradients*
  Europe ML meeting *Apr. 2021*
- *Dreaming to Distill*
  Synced AI (largest AI media in Asia) *Jul. 2020*
- *Dreaming to Distill*
  Facebook AR/VR *Jun. 2020*
- *Making Neural Networks Efficient*
  Alibaba Cloud / Platform AI group *Feb. 2020*
- *Efficient Neural Networks*
  NVIDIA Research, Facebook Research *Dec. 2019*
- *Efficient Neural Networks*
  Baidu Research, ByteDance A.I. Lab US *Dec. 2019*
- *Efficient Neural Networks*
  Alibaba A.I. Research, Kwai Lab *Nov. 2019*
- *Applied Machine Learning: From Theory to Practice*
  Invited Keynote, IEEE Circuits and Systems Society (Singapore Chapter) *Feb. 2018*
- *A Health Decision Support System for Disease Diagnosis*
  New Jersey Tech Council *Jun. 2016*

Patents (till Jun. 2022)

25. *Pruning Neural Networks*
    NVIDIA *2022*
24. *Neural Network Training Technique*
    NVIDIA *2022*
23. *Techniques to Identify Data used to Train One or More Neural Networks*
    NVIDIA *2022*
22. *Pruning Vision Transformers under Latency Budget and a Method to Distribute Parameters across Layers*
    NVIDIA *2022*
21. *GradViT: Gradient Inversion of Vision Transformers*
    NVIDIA *2022*
20. *Adaptive Token Depth Adjustment Algorithm for Networks with Transformer Blocks*
    NVIDIA *2022*
19. *Global Context Model for Transformer Neural Networks*
    NVIDIA *2022*
18. *Towards Understanding the Risks of Gradient Inversion in Federated Learning*
    NVIDIA *2022*
17. *When to Prune? A Policy for Early Structural Pruning*
    NVIDIA *2021*
16. *See Through Gradients: Image Batch Recovery via GradInversion*
    NVIDIA *2021*
15. *Network similarity metric as a Pruning Indicator*
    NVIDIA *2021*
14. *Zero-shot Model Inversion for Data-free Distillation*
    NVIDIA *2021*

13. *MHDeep: Mental Health Disorder Detection System based on Body-Area and Deep Neural Networks*
Princeton University *2019*

12. *Optimal MSE Quantization with Fixed Codebook and Rescaling*
NVIDIA *2020*

11. *Dreaming Data for Continual Learning*
NVIDIA *2020*

10. *Data-Free Knowledge Distillation for Object Detection*
NVIDIA *2020*

9. *Hardware-aware Latency Neural Network Pruning*
NVIDIA *2020*

8. *Image Generation for Data Free Pruning*
NVIDIA *2019*

7. *Hardware-guided Symbiotic Training for Compact, Accurate, yet Execution-efficient LSTMs*
Alibaba *2019*

6. *Incremental Learning using a Grow-and-prune Paradigm with Efficient Neural Networks*
Princeton University *2019*

5. *DiabDeep: Pervasive Diabetes Diagnosis based on Wearable Medical Sensors and Efficient Neural Networks*
Princeton University *2019*

4. *Smart, Secure, yet Energy-efficient Internet-of-Things Sensors*
Princeton University *2019*

3. *NeST: A Neural Network Synthesis Tool based on a Grow-and-prune Paradigm*
Princeton University *2018*

2. *Grow and Prune Compact, Fast, yet Accurate LSTMs*
Princeton University *2018*

1. *A Hierarchical Health Decision support System based on Wearable Medical Sensors and Machine Learning Ensembles*
Princeton University *2017*

Academic Services

Teaching Assistant - Princeton University

ELE 364, Machine Learning for Predictive Data Analytics *Fall, 17-18*
ELE464, Embedded Computing *Spring, 16-17*

Conference Reviewer & Committee

Computer Vision and Pattern Recognition (CVPR)
Conference on Neural Information Processing Systems (NeurIPS)
International Conference on Learning Representations (ICLR)
International Conference on Machine Learning (ICML)
International Conference on Computer Vision (ICCV)
European Conference on Computer Vision (ECCV)
British Machine Vision Conference (BMVC)
Winter Conference on Applications of Computer Vision (WACV)
AAAI Conference on Artificial Intelligence (AAAI)
Design Automation Conference (DAC)
High-Performance Computer Architecture (HPCA)

Journal Reviewer & Committee

IEEE Transactions on Pattern Analysis and Machine Intelligence
IEEE Transactions on Neural Networks and Learning Systems
International Journal of Computer Vision
IEEE Journal of Biomedical and Health Informatics
IEEE Journal of Selected Topics in Signal Processing
IEEE Sensors Journal
IEEE Consumer Electronics Magazine
International Journal on Artificial Intelligence Tools
International Journal of Systems Architecture
International Journal of Healthcare Technology and Management
International Journal of Electronic Imaging

Mentorship

NVIDIA Research Mentees

Baifeng Shi, University of California, Berkeley	*2023-2024*
Hanrong Ye, Hong Kong University of Science and Technology	*2023-2024*
Ji Lin, Massachusetts Institute of Technology	*2022-2023*
Huanrui Yang, Duke University	*2021-2022*
Zhen Dong, University of California, Berkeley	*2021-2022*
Xin Dong, Harvard University	*2021-2022*
Annamarie Bair, Carnegie Mellon University	*2022-2023*
Divyam Madaan, New York University	*2022-2023*
Paul Micaelli, University of Edingbugh	*2021-2022*
Yerlan Idelbayev, University of California, Merced	*2020-2021*
Vu Nguyen, Stony Brooks University	*2020-2021*
Akshay Chawla, Carnegie Mellon University	*2020-2021*

Princeton Senior Thesis Mentees

Joe Zhang, now Ph.D. at Stanford	*2019-2020*
Hari Santhanam, now Ph.D. at University of Pennsylvania	*2019-2020*
Frederick Hertan, now at SIG Trading	*2018-2019*
Kyle Johnson, now at Princeton University	*2018-2019*
Bilal Mukadam, now at Microsoft	*2018-2019*
Chloe Song, now at Astra Inc.	*2017-2018*