Tri Dao

Curriculum Vitae

Work Experience

9/24–present **Assistant Professor in Computer Science**, *Princeton University*.

7/23–present **Co-founder, Chief Scientist**, *Together AI*.

Education

9/16–06/23 **PhD in Computer Science**, *Stanford University*.
Dissertation: Hardware-aware Algorithms for Efficient Machine Learning.
Advisors: Christopher Ré, Stefano Ermon.

1/18–1/19 **MS in Statistics**, *Stanford University*.

9/14–6/16 **MS in Computer Science**, *Stanford University*.

9/12–6/16 **BS in Mathematics**, *Stanford University*.

Research Interests

Machine learning and systems, with a focus on efficient training and inference:
- Hardware-aware algorithms.
- Sequence models with long-range memory.

Research Adoption

FlashAttention.
- Used by most organizations to speed up the training and inference large language models and diffusion models (e.g., Meta, Microsoft, Nvidia, OpenAI, Google, Mistral, IBM, DeepSeek, Tencent, Alibaba).
- Integrated into virtually all ML frameworks (Pytorch, Jax, Huggingface's transformers, vLLM, SGLang, Microsoft's DeepSpeed, Nvidia's Megatron-LM and TensorRT-LLM) that benefit a large audience of researchers and practitioners.

See this page for a partial list of places where FlashAttention is being used.

Mamba.
- Used by many organizations (e.g., Microsoft, Nvidia, Mistral, IBM, TII, AI21, Tencent) to train large language models (up to 560B params) that scale better in context length than the dominant Transformer architecture.
- Integrated into many ML frameworks (Huggingface's transformers, vLLM, SGLang, Nvidia's TensorRT-LLM) that enable fast training and inference of Mamba models.
- Supported on Nvidia GPUs, AMD GPUs, and AWS Trainum 2.

See this page for a partial list of places where Mamba is being used.

Honors and Awards

2025 Schmidt Sciences AI2050 Fellowship.

2025 Google ML and Systems Junior Faculty Awards.

2025 Google Research Scholar.

2025 Conference on Machine Learning and Systems (MLSys) 2025, **Outstanding Paper Honorable Mention**.

2024 Conference on Language Modeling (COLM) 2024, **Outstanding Paper**.

2024 Inaugural **Stanford Open Source Software Prize** for Flash Attention.

2022 International Conference on Machine Learning (ICML) 2022, **Outstanding Paper runner-up**.

2022 Hardware Aware Efficient Training Workshop 2022, **Best Paper award**.

Publications

Ted Zadouri*, Jay Shah*, Markus Hohnerbach*, Timmy Liu, Vijay Thakkar, and **Tri Dao**. FlashAttention-4: Algorithm and kernel pipelining co-design for asymmetric hardware scaling. In *Machine Learning and Systems (MLSys)*, 2026.

Aakash Lahoti, Kevin Li, Berlin Chen, Caitlin Wang, Aviv Bick, Zico Kolter, , **Tri Dao**, and Albert Gu. Mamba-3: Improved sequence modeling using state space principles. In *International Conference on Learning Representations (ICLR)*, 2026.

Wentao Guo, Mayank Mishra, Xinle Cheng, Ion Stoica, and **Tri Dao**. SonicMoE: Accelerating moe with io and tile-aware optimizations. In *International Conference on Learning Representations (ICLR)*, 2026.

Tanishq Kumar, **Tri Dao**, and Avner May. Speculative speculative decoding. In *International Conference on Learning Representations (ICLR)*, 2026.

Zelei Shao, Vikranth Srivatsa, Sanjana Srivastava, Qingyang Wu, Alpay Ariyak, Xiaoxia Wu, Ameen Patel, Jue Wang, Percy Liang, **Tri Dao**, and others. Beat the long tail: Distribution-aware speculative decoding for rl training. In *Machine Learning and Systems (MLSys)*, 2026.

Costin-Andrei Oncescu, Qingyang Wu, Wai Tong Chung, Robert Wu, Bryan Gopal, Junxiong Wang, **Tri Dao**, and Ben Athiwaratkun. Opportunistic expert activation: Batch-aware expert routing for faster decode without retraining. *arXiv preprint arXiv:2511.02237*, 2025.

Han Guo, Songlin Yang, Tarushii Goel, Eric P Xing, **Tri Dao**, and Yoon Kim. Log-linear attention. In *International Conference on Learning Representations (ICLR)*, 2026.

Ted Zadouri, Hubert Strauss, and **Tri Dao**. Hardware-efficient attention for fast decoding. In *Conference on Language Modeling (COLM)*, 2025.

Ryan Po, Yotam Nitzan, Richard Zhang, Berlin Chen, **Tri Dao**, Eli Shechtman, Gordon Wetzstein, and Xun Huang. Long-context state-space video world models. In *International Conference on Computer Vision (ICCV)*, 2025.

Junxiong Wang, Wen-Ding Li, Daniele Paliotta, Daniel Ritter, Alexander M Rush, and **Tri Dao**. M1: Towards scalable test-time compute with mamba reasoning models. *arXiv preprint arXiv:2504.10449*, 2025.

Mingqian Ma, Guoqing Liu, Chuan Cao, Pan Deng, **Tri Dao**, Albert Gu, Peiran Jin, Zhao Yang, Yingce Xia, Renqian Luo, and others. Hybridna: A hybrid transformer-mamba2 long-range dna language model. *arXiv preprint arXiv:2502.10807*, 2025.

Daniele Paliotta, Junxiong Wang, Matteo Pagliardini, Kevin Y Li, Aviv Bick, J Zico Kolter, Albert Gu, François Fleuret, and **Tri Dao**. Thinking slow, fast: Scaling inference compute with distilled reasoners. *arXiv preprint arXiv:2502.20339*, 2025.

Muru Zhang, Mayank Mishra, Zhongzhu Zhou, William Brandon, Jue Wang, Yoon Kim, Jonathan Ragan-Kelley, Shuaiwen Leon Song, Ben Athiwaratkun, and **Tri Dao**. Ladder-residual: parallelism-aware architecture for accelerating large model inference with communication overlapping. In *International Conference on Machine Learning (ICML)*, 2025.

Rui Pan, Zhuang Wang, Zhen Jia, Can Karakus, Luca Zancato, **Tri Dao**, Ravi Netravali, and Yida Wang. Marconi: Prefix caching for the era of hybrid llms. In *Machine Learning and Systems (MLSys)*, 2025. **Outstanding Paper Honorable Mention**.

Maurice Weber, Daniel Fu, Quentin Anthony, Yonatan Oren, Shane Adams, Anton Alexandrov, Xiaozhong Lyu, Huu Nguyen, Xiaozhe Yao, Virgini Adams, Ben Athiwaratkun, Rahul Chalamala, Kezhen Chen, Max Ryabinin, **Tri Dao**, Percy Liang, Christopher Ré, Irina Rish, and Ce Zhang. Redpajama: an open dataset for training large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Junxiong Wang, Daniele Paliotta, Avner May, Alexander M Rush, and **Tri Dao**. The mamba in the llama: Distilling and accelerating hybrid models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Sukjun Hwang, Aakash Lahoti, **Tri Dao**, and Albert Gu. Hydra: Bidirectional state space models through generalized matrix mixers. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

James Liu, Guangxuan Xiao, Kai Li, Jason D Lee, Song Han, **Tri Dao**, and Tianle Cai. Bitdelta: Your fine-tune may only be worth one bit. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Jay Shah, Ganesh Bikshandi, Ying Zhang, Vijay Thakkar, Pradeep Ramani, and **Tri Dao**. FlashAttention-3: Fast and accurate attention with asynchrony and low-precision. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Roger Waleffe, Wonmin Byeon, Duncan Riach, Brandon Norick, Vijay Korthikanti, **Tri Dao**, Albert Gu, Ali Hatamizadeh, Sudhakar Singh, Deepak Narayanan, and others. An empirical study of mamba-based language models. *arXiv preprint arXiv:2406.07887*, 2024.

Anton Lozhkov, Raymond Li, Loubna Ben Allal, Federico Cassano, Joel Lamy-Poirier, Nouamane Tazi, Ao Tang, Dmytro Pykhtar, Jiawei Liu, Yuxiang Wei, and others. Starcoder 2 and the stack v2: The next generation. *arXiv preprint arXiv:2402.19173*, 2024.

Yair Schiff, Chia-Hsiang Kao, Aaron Gokaslan, **Tri Dao**, Albert Gu, and Volodymyr Kuleshov. Caduceus: Bi-directional equivariant long-range DNA sequence modeling. In *International Conference on Machine Learning (ICML)*, 2024.

Tri Dao* and Albert Gu*. Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. In *International Conference on Machine Learning (ICML)*, 2024.

Albert Gu* and **Tri Dao***. Mamba: Linear-time sequence modeling with selective state spaces. In *Conference on Language Modeling (COLM)*, 2024. **Outstanding Paper**.

Tianle Cai, Yuhong Li, Zhengyang Geng, Hongwu Peng, Jason D Lee, Deming Chen, and **Tri Dao**. Medusa: Simple LLM inference acceleration framework with multiple decoding heads. In *International Conference on Machine Learning (ICML)*, 2024.

Tri Dao. FlashAttention-2: Faster attention with better parallelism and work partitioning. In *International Conference on Learning Representations (ICLR)*, 2024.

Zichang Liu, Jue Wang, **Tri Dao**, Tianyi Zhou, Binhang Yuan, Zhao Song, Anshumali Shrivastava, Ce Zhang, Yuandong Tian, Christopher Ré, and Beidi Chen. Deja vu: Contextual sparsity for efficient llms at inference time. In *International Conference on Machine Learning*, 2023. **Oral**.

Michael Zhang, Khaled K Saab, Michael Poli, **Tri Dao**, Karan Goel, and Christopher Ré. Effectively modeling time series with simple discrete state spaces. In *International Conference on Learning Representations (ICLR)*, 2023.

Michael Poli*, Stefano Massaroli*, Eric Nguyen, Daniel Y Fu, **Tri Dao**, Stephen Baccus, Yoshua Bengio, Stefano Ermon, and Christopher Ré. Hyena hierarchy: Towards larger convolutional language models. In *International Conference on Machine Learning (ICML)*, 2023. **Oral**.

Daniel Y. Fu*, Elliot L Epstein*, Eric Nguyen, Armin W Thomas, Michael Zhang, **Tri Dao**, Atri Rudra, and Christopher Ré. Simple hardware-efficient long convolutions for sequence modeling. In *International Conference on Machine Learning (ICML)*, 2023.

Raymond Li, Loubna Ben Allal, Yangtian Zi, Niklas Muennighoff, Denis Kocetkov, Chenghao Mou, Marc Marone, Christopher Akiki, Jia Li, Jenny Chim, and others. Starcoder: may the source be with you! *Transactions on Machine Learning Research (TMLR)*, 2023.

Tri Dao*, Daniel Y. Fu*, Khaled K. Saab, Armin W. Thomas, Atri Rudra, and Christopher Ré. Hungry hungry hippos: Towards language modeling with state space models. In *International Conference on Learning Representations (ICLR)*, 2023. **Spotlight**.

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. FlashAttention: Fast and memory-efficient exact attention with IO-awareness. In *Advances in Neural Information Processing Systems*, 2022.

Binhang Yuan, Yongjun He, Jared Quincy Davis, Tianyi Zhang, **Tri Dao**, Beidi Chen, Percy Liang, Christopher Ré, and Ce Zhang. Decentralized training of foundation models in heterogeneous environments. In *Advances in Neural Information Processing Systems*, 2022. **Oral**.

June Wang, Binhang Yuan, Luka Rimanic, Yongjun He, **Tri Dao**, Beidi Chen, Percy Liang, Christopher Ré, and Ce Zhang. Fine-tuning language models over slow networks using activation compression with guarantees. In *Advances in Neural Information Processing Systems*, 2022.

Michael Poli, Stefano Massaroli, Federico Berto, Jinkyoo Park, **Tri Dao**, Christopher Ré, and Stefano Ermon. Transform once: Efficient operator learning in frequency domain. In *Advances in Neural Information Processing Systems*, 2022.

Eric Nguyen, Karan Goel, Albert Gu, Gordon Downs, Preey Shah, **Tri Dao**, Stephen Baccus, and Christopher Ré. S4ND: Modeling images and videos as multidimensional signals with state spaces. In *Advances in Neural Information Processing Systems*, 2022.

Tri Dao, Beidi Chen, Nimit Sohoni, Arjun Desai, Michael Poli, Jessica Grogan, Alexander Liu, Aniruddh Rao, Atri Rudra, and Christopher Ré. Monarch: Expressive structured matrices for efficient and accurate training. In *International Conference on Machine Learning (ICML)*, 2022. **Outstanding Paper runner-up**.

Chenlin Meng, Linqi Zhou, Kristy Choi, **Tri Dao**, and Stefano Ermon. ButterflyFlow: Building invertible layers with butterfly matrices. In *International Conference on Machine Learning (ICML)*, 2022.

Tri Dao*, Beidi Chen*, Kaizhao Liang, Jiaming Yang, Zhao Song, Atri Rudra, and Christopher Ré. Pixelated butterfly: Simple and efficient sparse training for neural network models. In *International Conference on Learning Representations (ICLR)*, 2022. **Spotlight**.

Beidi Chen*, **Tri Dao***, Eric Winsor, Zhao Song, Atri Rudra, and Christopher Ré. Scatterbrain: Unifying sparse and low-rank attention. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Albert Gu, Isys Johnson, Karan Goel, Khaled Saab, **Tri Dao**, Atri Rudra, and Christopher Ré. Combining recurrent, convolutional, and continuous-time models with linear state space layers. *Advances in Neural Information Processing Systems*, 34, 2021.

Nicholas Roberts, Mikhail Khodak, **Tri Dao**, Liam Li, Christopher Ré, and Ameet Talwalkar. Rethinking neural operations for diverse tasks. In *Advances in Neural Information Processing Systems*, 2021.

Jared Q Davis*, Albert Gu*, Krzysztof Choromanski, **Tri Dao**, Christopher Ré, Chelsea Finn, and Percy Liang. Catformer: Designing stable transformers via sensitivity analysis. In *International Conference on Machine Learning (ICML)*, 2021.

Tri Dao, Govinda M Kamath, Vasilis Syrgkanis, and Lester Mackey. Knowledge distillation as semiparametric inference. In *International Conference on Learning Representations (ICLR)*, 2021.

Beidi Chen, Zichang Liu, Binghui Peng, Zhaozhuo Xu, Jonathan Lingjie Li, **Tri Dao**, Zhao Song, Anshumali Shrivastava, and Christopher Ré. Mongoose: A learnable LSH framework for efficient neural network training. In *International Conference on Learning Representations (ICLR)*, 2021. **Oral**.

Albert Gu*, **Tri Dao***, Stefano Ermon, Atri Rudra, and Christopher Ré. Hippo: Recurrent memory with optimal polynomial projections. In *Advances in neural information processing systems (NeurIPS)*, 2020. **Spotlight**.

Tri Dao, Nimit Sohoni, Albert Gu, Matthew Eichhorn, Amit Blonder, Megan Leszczynski, Atri Rudra, and Christopher Ré. Kaleidoscope: An efficient, learnable representation for all structured linear maps. In *The International Conference on Learning Representations (ICLR)*. 2020. **Spotlight**.

Avner May, Jian Zhang, **Tri Dao**, and Christopher Ré. On the downstream performance of compressed word embeddings. In *Advances in Neural Information Processing Systems (NeurIPS) 32*, 2019. **Spotlight**.

Jonathan Kuck, **Tri Dao**, Hamid Rezatofighi, Ashish Sabharwal, and Stefano Ermon. Approximating the permanent by sampling from adaptive partitions. In *Advances in Neural Information Processing Systems (NeurIPS) 32*, 2019.

Jonathan Kuck, **Tri Dao**, Shengjia Zhao, Burak Bartan, Ashish Sabharwal, and Stefano Ermon. Adaptive hashing for model counting. In *Proceedings of the 35th Conference on Uncertainty in Artificial Intelligence (UAI)*. 2019.

Tri Dao, Albert Gu, Matthew Eichhorn, Atri Rudra, and Christopher Ré. Learning fast algorithms for linear transforms using butterfly factorizations. In *The International Conference on Machine Learning (ICML) 36*. 2019. **Oral**.

Tri Dao, Albert Gu, Alexander J Ratner, Virginia Smith, Christopher De Sa, and Christopher Ré. A kernel theory of modern data augmentation. In *The International Conference on Machine Learning (ICML) 36*. 2019.

Jian Zhang, Avner May, **Tri Dao**, and Christopher Ré. Low-precision random Fourier features for memory-constrained kernel approximation. In *The International Conference on Artificial Intelligence and Statistics (AISTATS) 22*. 2019.

Anna T Thomas, Albert Gu, **Tri Dao**, Atri Rudra, and Christopher Ré. Learning compressed transforms with low displacement rank. In *Advances in Neural Information Processing Systems (NeurIPS) 31*. 2018.

Tri Dao, Christopher M De Sa, and Christopher Ré. Gaussian quadrature for kernel features. In *Advances in Neural Information Processing Systems (NeurIPS) 30*. 2017. **Spotlight**.

Industry Experience

10/22 – 06/23 **Adept AI**, PhD Fellow (part-time), San Francisco, CA.
- Develop multi-modal Transformers to model users' interaction with browser tools.
- Speed up large-scale Transformer distributed training and inference.
- Train large language models on long context (16K).

6/20 – 9/20 **Microsoft Research**, Research Intern, Cambridge, MA.
- Developed a novel loss function for knowledge distillation that improves the performance of the student model.

6/16 – 9/16 **Citadel Securities**, Quantitative Researcher, Chicago, IL.
- Developed a novel feature generation method to analyze large quantitative trading datasets.
- Built a black-box optimization system for state-of-the-art quantitative trading strategies.

6/14 – 9/14 **Google**, Software Engineering Intern, Mountain View, CA.
- Designed machine learning algorithms to find best advertisements for each Ad Group.

Teaching

1/26 – 5/26 COS 484: Natural Language Processing, Instructor, Princeton University.

9/25 – 12/25 COS 597A: Efficient Systems for Foundation Models, Instructor, Princeton University.

1/25 – 5/25 COS 484: Natural Language Processing, Instructor, Princeton University.

1/20 – 3/20 CS 228: Probabilistic Graphical Models, Teaching Assistant, Stanford University.

4/19 – 6/19 CS 229: Machine Learning, Teaching Assistant, Stanford University.

1/16 – 3/16 EE 364A: Convex Optimization I, Teaching Assistant, Stanford University.

9/15 – 12/15 EE 103: Intro to Matrix Methods, Teaching Assistant, Stanford University.

Service

Organizer: Efficient Systems for Foundation Models workshop at ICML 2023, ICML 2024, ICML 2025.

Area Chair: COLM 2024, 2025, 2026, ICLR 2026, ICML 2026.

Reviewer: NeurIPS, ICML, ICLR, AISTATS, JMLR, ICCV, NeurIPS 2019 **best reviewers**, ICML 2019 **best reviewers**.