

ZHENMEI SHI

Email: zhmeishi@gmail.com Tel: (+1) 608-698-2223 Homepage: <http://zhmeishi.github.io/>

PROFESSIONAL EXPERIENCE

Member of Technical Staff *at* **xAI**, Palo Alto, CA 10/2025-Now

- Technical led the webdev, terminal-use, and full-stack long-horizon agentic coding expert training. Made Grok 4.2 and Grok 4.3 to be #1 in multiple benchmarks on Design Arena.
- Created $O(100B)$ tokens of multimodal interleave pretrain data.
- Trained embedding/reranker model for agentic search.

Senior Research Scientist *at* MongoDB + **Voyage AI**, Palo Alto, CA 01/2025-10/2025

- Lead rerank-2.5 end-to-end training Blog
- Core contribute to automatic data label pipeline
- Contribute to customized finetuning Report to: Tengyu Ma

EDUCATION

University of Wisconsin-Madison, US

Ph.D. in Computer Sciences (Advised by **Yingyu Liang**) 09/2019-12/2024

Master of Science in Computer Sciences 09/2019-12/2022

Hong Kong University of Science and Technology, Hong Kong 09/2015-05/2019

Bachelor of Science in Computer Science and Pure Mathematics Advanced

ETH Zurich, Switzerland (Exchange) 02/2018-08/2018

PHD THESIS

Understanding Training and Adaptation in Feature Learning: From Two-Layer Networks to Foundation Models 2024

PUBLICATIONS

* denotes equal contribution or alphabetical order.

1. Discovering the Gems in Early Layers: Accelerating Long-Context LLMs with 1000x Input Token Reduction
Zhenmei Shi, Yifei Ming, Xuan-Phi Nguyen, Yingyu Liang, Shafiq Joty
<https://openreview.net/forum?id=9dPP4n4mfA> ACL 2026 Findings
2. On Computational Limits of FlowAR Models: Expressivity and Efficiency
Chengyue Gong*, Yekun Ke*, Xiaoyu Li*, Yingyu Liang*, Zhizhou Sha*, **Zhenmei Shi***, Zhao Song*
<https://openreview.net/forum?id=FZqLS3Y2t2> AISTATS 2026
3. T2VWorldBench: A Benchmark for Evaluating World Knowledge in Text-to-Video Generation
Yubin Chen*, Xuyang Guo*, **Zhenmei Shi***, Zhao Song*, Jiahao Zhang*
<https://openreview.net/forum?id=louUbQLq0Z> WACV 2026
4. Kernel Regression in Structured Non-IID Settings: Theory and Implications for Denoising Score Learning
Dechen Zhang, **Zhenmei Shi**, Yi Zhang, Yingyu Liang, Difan Zou
<https://openreview.net/forum?id=t0wP9z1Zde> NeurIPS 2025

5. Circuit Complexity Bounds for RoPE-based Transformer Architecture
Bo Chen*, Xiaoyu Li*, Yingyu Liang*, Jiangxuan Long*, **Zhenmei Shi***, Zhao Song*, Jiahao Zhang*
<https://openreview.net/forum?id=dsrWTxVeos> EMNLP 2025
6. Toward Infinite-Long Prefix in Transformer
Yingyu Liang*, **Zhenmei Shi***, Zhao Song*, Chiwun Yang*
<https://openreview.net/forum?id=ihEy7xRq1M> EMNLP 2025
7. Conv-Basis: A New Paradigm for Efficient Attention Inference and Gradient Computation in Transformers
Yingyu Liang*, Heshan Liu*, **Zhenmei Shi***, Zhao Song*, Zhuoyan Xu*, Jiale Zhao*, Zhen Zhuang*
<https://openreview.net/forum?id=029Fi0m2cB> EMNLP Findings 2025
8. Force Matching with Relativistic Constraints: A Physics-Inspired Approach to Stable and Efficient Generative Modeling
Yang Cao*, Bo Chen*, Xiaoyu Li*, Yingyu Liang*, Zhizhou Sha*, **Zhenmei Shi***, Zhao Song*, Mingda Wan*
<https://arxiv.org/abs/2502.08150> CIKM 2025
9. Unraveling the Smoothness Properties of Diffusion Models: A Gaussian Mixture Perspective
Yingyu Liang*, Zhizhou Sha*, **Zhenmei Shi***, Zhao Song*, Mingda Wan*, Yufa Zhou*
<https://openreview.net/forum?id=E9MPcFpU1A> ICCV 2025
10. NRFlow: Towards Noise-Robust Generative Modeling via High-Order Mechanism
Bo Chen*, Chengyue Gong*, Xiaoyu Li*, Yingyu Liang*, Zhizhou Sha*, **Zhenmei Shi***, Zhao Song*, Mingda Wan*, Xugang Ye*
<https://openreview.net/forum?id=hCONRCHDEL> UAI 2025
11. Dissecting Submission Limit in Desk-Rejections: A Mathematical Analysis of Fairness in AI Conference Policies
Yuefan Cao*, Xiaoyu Li*, Yingyu Liang*, Zhizhou Sha*, **Zhenmei Shi***, Zhao Song*, Jiahao Zhang*
<https://openreview.net/forum?id=wDKlybjm7T> ICML 2025
12. Fundamental Limits of Visual Autoregressive Transformers: Universal Approximation Abilities
Yifang Chen*, Xiaoyu Li*, Yingyu Liang*, **Zhenmei Shi***, Zhao Song*
<https://openreview.net/forum?id=mag0SIm8UT> ICML 2025
13. Beyond Linear Approximations: A Novel Pruning Approach for Attention Matrix
Yingyu Liang*, Jiangxuan Long*, **Zhenmei Shi***, Zhao Song*, Yufa Zhou*
<https://openreview.net/forum?id=sgbI8Pxwie> ICLR 2025
14. Fast Gradient Computation for RoPE Attention in Almost Linear Time
Yifang Chen*, Jiayan Huo*, Xiaoyu Li*, Yingyu Liang*, **Zhenmei Shi***, Zhao Song*
<https://openreview.net/forum?id=nVXcDUTkcf> Workshop ICLR 2025
15. RichSpace: Enriching Text-to-Video Prompt Space via Text Embedding Interpolation
Yuefan Cao*, Chengyue Gong*, Xiaoyu Li*, Yingyu Liang*, Zhizhou Sha*, **Zhenmei Shi***, Zhao Song*
<https://openreview.net/forum?id=fk9svZT2KR> Workshop ICLR 2025
16. High-Order Matching for One-Step Shortcut Diffusion Models
Bo Chen*, Chengyue Gong*, Xiaoyu Li*, Yingyu Liang*, Zhizhou Sha*, **Zhenmei Shi***, Zhao Song*, Mingda Wan*
<https://openreview.net/forum?id=be8vS1Dcix> Workshop ICLR 2025
17. When Can We Solve the Weighted Low Rank Approximation Problem in Truly Subquadratic Time?
Chenyang Li*, Yingyu Liang*, **Zhenmei Shi***, Zhao Song*
<https://openreview.net/forum?id=SQkur6MNOR> AISTats 2025

18. Fourier Circuits in Neural Networks and Transformers: A Case Study of Modular Arithmetic with Multiple Inputs
Chenyang Li*, Yingyu Liang*, **Zhenmei Shi***, Zhao Song*, Tianyi Zhou*
<https://openreview.net/forum?id=RFpIosT9WI> AISTats 2025
19. Looped ReLU MLPs May Be All You Need as Practical Programmable Computers
Yingyu Liang*, Zhizhou Sha*, **Zhenmei Shi***, Zhao Song*, Yufa Zhou*
<https://openreview.net/forum?id=XjNFbBqrEi> AISTats 2025
20. Bypassing the Exponential Dependency: Looped Transformers Efficiently Learn In-context by Multi-step Gradient Descent
Bo Chen*, Xiaoyu Li*, Yingyu Liang*, **Zhenmei Shi***, Zhao Song*
<https://openreview.net/forum?id=4C1aRz2gRq> AISTats 2025
21. Differential Privacy Mechanisms in Neural Tangent Kernel Regression
Jiuxiang Gu*, Yingyu Liang*, Zhizhou Sha*, **Zhenmei Shi***, Zhao Song*
<https://arxiv.org/abs/2407.13621> WACV 2025
22. The Computational Limits of State-Space Models and Mamba via the Lens of Circuit Complexity
Yifang Chen*, Xiaoyu Li*, Yingyu Liang*, **Zhenmei Shi***, Zhao Song*
<https://openreview.net/forum?id=bImLLT3r62> Oral CPAL 2025
23. Fast John Ellipsoid Computation with Differential Privacy Optimization
Xiaoyu Li*, Yingyu Liang*, **Zhenmei Shi***, Zhao Song*, Junwei Yu*
<https://openreview.net/forum?id=yfmFSc5ZPG> Oral CPAL 2025
24. Curse of Attention: A Kernel-Based Perspective for Why Transformers Fail to Generalize on Time Series Forecasting and Beyond
Yekun Ke*, Yingyu Liang*, **Zhenmei Shi***, Zhao Song*, Chiwun Yang*
<https://openreview.net/forum?id=bd0mItHgU5> CPAL 2025
25. HSR-Enhanced Sparse Attention Acceleration
Bo Chen*, Yingyu Liang*, Zhizhou Sha*, **Zhenmei Shi***, Zhao Song*
<https://openreview.net/forum?id=wso1gABiPZ> CPAL 2025
26. Is A Picture Worth A Thousand Words? Delving Into Spatial Reasoning for Vision Language Models
Jiayu Wang, Yifei Ming, **Zhenmei Shi**, Vibhav Vineet, Xin Wang, Yixuan Li, Neel Joshi
<https://openreview.net/forum?id=cvaSru8Le0> NeurIPS 2024
27. Multi-Layer Transformers Gradient Can be Approximated in Almost Linear Time
Yingyu Liang*, Zhizhou Sha*, **Zhenmei Shi***, Zhao Song*, Yufa Zhou*
<https://openreview.net/forum?id=1LJIPZ4SvS> Workshop NeurIPS 2024
28. Tensor Attention Training: Provably Efficient Learning of Higher-order Transformers
Yingyu Liang*, **Zhenmei Shi***, Zhao Song*, Yufa Zhou*
<https://openreview.net/forum?id=1vrVar7FDC> Workshop NeurIPS 2024
29. A Tighter Complexity Analysis of SparseGPT
Xiaoyu Li*, Yingyu Liang*, **Zhenmei Shi***, Zhao Song*
<https://openreview.net/forum?id=443oNjXhsT> Workshop NeurIPS 2024
30. Differential Privacy of Cross-Attention with Provable Guarantee
Yingyu Liang*, **Zhenmei Shi***, Zhao Song*, Yufa Zhou*
<https://openreview.net/forum?id=GttuYQVARs> Workshop NeurIPS 2024
31. Do Large Language Models Have Compositional Ability? An Investigation into Limitations and Scalability

- Zhuoyan Xu*, **Zhenmei Shi***, Yingyu Liang
<https://openreview.net/forum?id=iI1CzEhEMU> COLM 2024
32. Why Larger Language Models Do In-context Learning Differently?
Zhenmei Shi, Junyi Wei, Zhuoyan Xu, Yingyu Liang
<https://openreview.net/forum?id=W0a96EG26M> ICML 2024
33. Towards Few-Shot Adaptation of Foundation Models via Multitask Finetuning
Zhuoyan Xu, **Zhenmei Shi**, Junyi Wei, Fangzhou Mu, Yin Li, Yingyu Liang
<https://openreview.net/forum?id=1jbh2e0b2K> ICLR 2024
34. Domain Generalization via Nuclear Norm Regularization
Zhenmei Shi*, Yifei Ming*, Ying Fan*, Frederic Sala, Yingyu Liang
<https://openreview.net/forum?id=hJd66ZzXEZ> **Oral** CPAL 2024
35. Provable Guarantees for Neural Networks via Gradient Feature Learning
Zhenmei Shi*, Junyi Wei*, Yingyu Liang
<https://openreview.net/forum?id=5F04bU79eK> NeurIPS 2023
36. A Graph-Theoretic Framework for Understanding Open-World Semi-Supervised Learning
Yiyu Sun, **Zhenmei Shi**, Yixuan Li
<https://openreview.net/forum?id=ZIT0HWeAy7> **Spotlight** NeurIPS 2023
37. When and How Does Known Class Help Discover Unknown Ones? Provable Understandings Through Spectral Analysis
Yiyu Sun, **Zhenmei Shi**, Yingyu Liang, Yixuan Li
<https://openreview.net/forum?id=JHodnaW5WZ> ICML 2023
38. The Trade-off between Universality and Label Efficiency of Representations from Contrastive Learning
Zhenmei Shi*, Jiefeng Chen*, Kunyang Li, Jayaram Raghuram, Xi Wu, Yingyu Liang, Somesh Jha
https://openreview.net/forum?id=rvsbw2YthH_ (Accept Rate: 7.95%) **Spotlight** ICLR 2023
39. A Theoretical Analysis on Feature Learning in Neural Networks: Emergence from Inputs and Advantage over Fixed Features
Zhenmei Shi*, Junyi Wei*, Yingyu Liang
https://openreview.net/forum?id=wMpS-Z_AI_E ICLR 2022
40. Attentive Walk-Aggregating Graph Neural Networks
Mehmet F. Demirel, Shengchao Liu, Siddhant Garg, **Zhenmei Shi**, Yingyu Liang
<https://openreview.net/forum?id=TWSTyYd2Rl> TMLR 2022
41. Deep Online Fused Video Stabilization
Zhenmei Shi, Fuhao Shi, Wei-Sheng Lai, Chia-Kai Liang, Yingyu Liang
<https://zhmeishi.github.io/dvs/> WACV 2022
42. Structured Feature Learning for End-to-End Continuous Sign Language Recognition
Zhaoyang Yang*, **Zhenmei Shi***, Xiaoyong Shen, Yu-Wing Tai
<https://arxiv.org/abs/1908.01341> News, 2019

PREPRINT

* denotes equal contribution or alphabetical order.

1. Can Language Models Compose Skills In-Context?
Zidong Liu, Zhuoyan Xu, **Zhenmei Shi**, Yingyu Liang
<https://arxiv.org/abs/2510.22993>

2. Neural Algorithmic Reasoning for Hypergraphs with Looped Transformers
Xiaoyu Li*, Yingyu Liang*, Jiangxuan Long*, **Zhenmei Shi***, Zhao Song*, Zhen Zhuang*
<https://arxiv.org/abs/2501.10688> 2025
3. On the Computational Capability of Graph Neural Networks: A Circuit Complexity Bound Perspective
Xiaoyu Li*, Yingyu Liang*, **Zhenmei Shi***, Zhao Song*, Wei Wang*, Jiahao Zhang*
<https://arxiv.org/abs/2501.06444> 2025
4. Theoretical Constraints on the Expressive Power of RoPE-based Tensor Attention Transformers
Xiaoyu Li*, Yingyu Liang*, **Zhenmei Shi***, Zhao Song*, Mingda Wan*
<https://arxiv.org/abs/2412.18040> 2024
5. On the Expressive Power of Modern Hopfield Networks
Xiaoyu Li*, Yuanpeng Li*, Yingyu Liang*, **Zhenmei Shi***, Zhao Song*
<https://arxiv.org/abs/2412.05562> 2024
6. Advancing the Understanding of Fixed Point Iterations in Deep Neural Networks: A Detailed Analytical Study
Yekun Ke*, Xiaoyu Li*, Yingyu Liang*, **Zhenmei Shi***, Zhao Song*
<https://arxiv.org/abs/2410.11279> 2024
7. Fine-grained Attention I/O Complexity: Comprehensive Analysis for Backward Passes
Xiaoyu Li*, Yingyu Liang*, **Zhenmei Shi***, Zhao Song*, Yufa Zhou*
<https://arxiv.org/abs/2410.09397> 2024
8. Exploring the Frontiers of Softmax: Provable Optimization, Applications in Diffusion Model, and Beyond
Jiuxiang Gu*, Chenyang Li*, Yingyu Liang*, **Zhenmei Shi***, Zhao Song*
<https://arxiv.org/abs/2405.03251> 2024
9. Dual Augmented Memory Network for Unsupervised Video Object Tracking
Zhenmei Shi*, Haoyang Fang*, Chi-Keung Tang, Yu-Wing Tai
<https://arxiv.org/abs/1908.00777> <https://zhmeishi.github.io/DAWN/>, 2019

INTERNSHIP EXPERIENCE

- Research Intern at **Google Cloud AI**, Sunnyvale, CA 10/2024-12/2024
- LLM Post-training. Supervised by: Sercan Arik
- AI Research Scientist Intern at **Salesforce**, Palo Alto, CA 06/2024-08/2024
- Long Context LLM Understanding, In-context Learning, and Inference Acceleration.
Supervised by: Shafiq Joty
- Software Engineering Intern at Google YouTube Ads ML, Mountain View, CA 06/2021-09/2021
- Unsupervised Clustering in Recommendation System. Supervised by: Myra Nam
- Software Engineering Intern at Google Pixel Camera, Mountain View, CA 05/2020-08/2020
- Deep Video Stabilization. Supervised by: Fuhao Shi
- Research Intern at Megvii (Face++) Foundation Model, Beijing 06/2019-08/2019
- Neural Architecture Search. Supervised by: Xiangyu Zhang
- Machine Learning Research Intern at Tencent YouTu, Shenzhen 12/2018-02/2019
- Deep Sign Language Recognition. Supervised by: Yu-Wing Tai News

Machine Learning Research Intern <i>at</i> Tencent YouTu, Shenzhen	12/2017-02/2018
<ul style="list-style-type: none"> • Deep Colorization. Supervised by: Yu-Wing Tai 	News
Research Assistant <i>at</i> UW-Madison CS School	09/2019 - 12/2024
Research Assistant <i>at</i> HKUST CS Department	02/2019-06/2019
Research Assistant <i>at</i> Oak Ridge National Lab, US	05/2017-08/2017
Part-Time Programmer <i>at</i> CUHK, HKBU, HKUST	06/2016-08/2016, 04/2017, 12/2018

ACADEMIC SERVICES

Conference Reviewer *at* ICLR 2022-2026, NeurIPS 2022-2025, ICML 2022 and 2024-2026, COLM 2025, AISTATS 2025-2026, AAAI 2026, IJCAI 2025, CVPR 2021-2022 and 2025-2026, ICCV 2021-2025, ECCV 2020-2022, WACV 2022 and 2025-2026

Journal Reviewer *at* JVCI, IEEE Transactions on Information Theory

TEACHING EXPERIENCE

Teaching Assistant of CS220 (Data Programming I) <i>at</i> UW-Madison	Spring 2020
Teaching Assistant of CS301 (Intro to Data Programming) <i>at</i> UW-Madison	Fall 2019

EXTRA-CURRICULAR

Visiting Student <i>at</i> Microsoft Research Asia (MSRA)	08/2017
Mentee <i>at</i> Hong Kong X-Tech	2017-2018
Vice Chairman <i>at</i> Microsoft Student Club @ HKUST	2017-2018
S.S. Chern Class Member <i>at</i> HKUST Mathematics Department	2016-2019
Volunteer Teacher <i>at</i> Ociva, Maldives	12/2016-01/2017
Software Team Member <i>at</i> HKUST Robotics Team	2015-2016

AWARDS & SCHOLARSHIPS

ICLR 2025 Notable Reviewers	05/2025
NeurIPS 2023 Scholar Award	10/2023
Student Research Grants Competition <i>from</i> UW-Madison	04/2023
ICLR 2023 Financial Assistance	03/2023
CS Departmental Scholarship <i>from</i> UW-Madison	09/2019
Academic Achievement Medal <i>from</i> HKUST	11/2019
S.S. Chern Class Achievement Scholarship <i>from</i> HKUST Math Department	07/2019
Undergraduate Research Opportunity Program Award <i>from</i> HKUST	04/2019
Best Student Team <i>from</i> HC2 (Switzerland's Biggest Programming Contest)	03/2018
Champion of Micro Innovation Award <i>from</i> Tencent	02/2018
1st Runner-up <i>from</i> HKUST x Radica Big Datathon	12/2017
Technology Star <i>from</i> Beauty of Programming Competition held by MSRA	08/2017
1st Prize of University's Scholarship for Undergraduate Student <i>from</i> HKUST	2016-2019
Reaching Out Award <i>from</i> HKSAR Government Scholarship Fund	2017-2018
Exchange Award, Lee Hysan Foundation Exchange Scholarship <i>from</i> HKUST	2017-2018
The Cheng Foundation Scholarship for Mainland Students <i>from</i> HKUST	2016-2017
Dean's List <i>from</i> HKUST	2015-2019

TECHNICAL SKILLS

Python: vLLM, SGLang, Megatron-LM, DeepSpeed, JAX, TensorFlow, PyTorch, Numpy, sklearn, Pandas, BeautifulSoup, re, Spark, and many more

C++: Caffe, OpenCV, OpenMP, MPI

Slurm, Java, MATLAB, C, R, SQL, CUDA, Wandb, L^AT_EX, Bash Script

INVITED TALKS

Understanding Training and Adaptation in Feature Learning: From Two-Layer Networks to Foundation Models

- MML seminar, April 2025, Remote

Provable Guarantees for Neural Networks via Gradient Feature Learning

- AI Time Idea Seminar, October 2023, Remote

The Trade-off between Universality and Label Efficiency of Representations from Contrastive Learning

- AI Time Idea Seminar, June 2023, Remote
- MLOPT Idea Seminar, March 2023, Madison, USA