ZISHAN SHAO

+1 (336) 391-4963 Durham, NC
zishan.shao@duke.edu linkedin.com/in/zishan-shao zishan-shao.github.io

RESEARCH INTERESTS

Machine Learning System (MLSys); Efficient LLM/Diffusion Inference; High-Performance Computing (HPC)

EDUCATION

Duke University, Durham, NC — Aug 2024 – May 2026
M.S. in Statistical Science — GPA: 3.8

Wake Forest University, Winston-Salem, NC — Aug 2020 – May 2024
B.S. in Computer Science & B.S. in Statistical Science, *summa cum laude* — GPA: 3.97

SELECTED PUBLICATIONS

* denotes equal contribution

- **Zishan Shao**, Yixiao Wang, Qinsi Wang, Ting Jiang, Zhixu Du, Hancheng Ye, Danyang Zhuo, Yiran Chen, Hai Li. *"FlashSVD: Memory Efficient Approach for SVD-Based Low Rank Model Inference."* Annual AAAI Conference on Artificial Intelligence (***AAAI***), 2026. [PDF]
- Yixiao Wang*, **Zishan Shao***, Ting Jiang, Aditya Devarakonda. *"Enhanced Cyclic Coordinate Descent Methods for Elastic Net Penalized Linear Models."* Neural Information Processing Systems (***NeurIPS***), 2025. [PDF]
- Ting Jiang*, Hancheng Ye*, Yixiao Wang*, **Zishan Shao**, Jingwei Sun, Jingyang Zhang, Jianyi Zhang, Zekai Chen, Yiran Chen, Hai Li. *"SADA: Stability-guided Adaptive Diffusion Acceleration."* International Conference on Machine Learning (***ICML***), 2025. [PDF]
- **Zishan Shao**, Aditya Devarakonda. *"Scalable Dual Coordinate Descent for Kernel Methods."* International Conference on High Performance Computing in Asia-Pacific Region (***HPCAsia, Outstanding Paper Award***, 2025). [PDF]

SELECTED MANUSCRIPTS

- Yixiao Wang*, Ting Jiang*, **Zishan Shao***, Hancheng Ye, Jingwei Sun, Mingyuan Ma, Qinsi Wang, Jianyi Zhang, Yiran Chen, Hai Li. *"Accelerating Denoising Generative Models is as Easy as Predicting Second-Order Difference."* In submission to International Conference on Learning Representations (***ICLR)***, 2026.
- Ting Jiang*, **Zishan Shao***, Kangning Cui, Yueqian Lin, Yiran Chen, Julian McAuley, Taylor Berg-Kirkpatrick, Zachary Novack. *"$\mathbf{A^2}$CT: Adaptive Anchored Consistency Tuning for Domain Adaptation of Timestep-Distilled Generative Models."* In submission to IEEE International Conference on Acoustics, Speech, and Signal Processing (***ICASSP***), 2026.
- Qinsi Wang, Jing Shi, Kun Wan, Handong Zhao, Hancheng Ye, **Zishan Shao**, Jinghan Ke, Yudong Liu, Daniel Miranda, Purvak Lapsiya, Yiran Chen, Wentian Zhao *"Seeing is Solving: Unlocking Efficient Vision-Language Models RLFT via Adaptive Curriculum and Continuous Rewards."* In submission to Conference on Computer Vision and Pattern Recognition (***CVPR***), 2026.

RESEARCH EXPERIENCE

FlashSVD: Memory-Efficient Inference & Training for Low-Rank Transformers GitHub
Center of Computational Evolutionary Intelligence (CEI), Duke University

- Built rank-aware streaming kernels (*FlashSVDAttention*, *FlashSVDFFN v1/v2*) in CUDA/Triton to push SVD factors through attention/FFN without full activation buffers; sustained high GPU occupancy.
- Reduced **peak activations by 70.2%** and **transients by 75%** on BERT/GPT-class models with **no loss in accuracy or latency**.

ZEUS: Zero-shot Efficient Unified Sparsity for Generative Models GitHub
Center of Computational Evolutionary Intelligence (CEI), Duke University

- Training-free, plug-and-play skipping (2nd-order extrapolation + parity-aware reuse) with one-line Diffusers integration.
- Delivers 1.9×–3.6× **speedups** with near-linear efficiency–fidelity scaling; lower LPIPS than prior training-free baselines at matched speed across SD/SDXL/Flux, Wan2, CogVideo.

ECCD: Enhanced Cyclic Coordinate Descent for GLMs (Elastic Net) GitHub
Sparstitute, Wake Forest University

- Introduced Hessian-approximate updates enabling batched coordinate ops, removing nonlinear gradient hotspots.
- Achieved up to 13× **faster** training vs. `glmnet`, `BigLasso`, `ncvreg`, `ABESS`, `skglm` on real and synthetic data with negligible error increase.

SADA: Stability-Guided Adaptive Diffusion Acceleration GitHub
Center of Computational Evolutionary Intelligence (CEI), Duke University

- Introduced a training-free paradigm exploiting step and token-wise sparsity to accelerate diffusion sampling-achieving $\geq 1.8\times$ speedups on SD-2, SDXL, and Flux (EDM & DPM++ solvers) with LPIPS ≤ 0.10 and FID ≤ 4.5-and proposed a unified skipping method that outperforms existing training-free approaches.
- Demonstrated cross-modal generalization with approximate $1.81\times$ acceleration on MusicLDM and approximate $1.41\times$ on ControlNet-no fine-tuning required.

Scalable Dual Coordinate Descent for Kernel Methods GitHub
Sparstitute, Wake Forest University

- Developed scalable s-step variants of Dual Coordinate Descent (DCD) and Block DCD for kernel SVM and ridge regression, reducing communication on distributed-memory systems while matching standard methods' accuracy; derived theoretical bounds for computation and communication costs.
- Built high-performance C/MPI implementations and validated strong scaling speedups up to 9.8× on 512 cores (Cray EX cluster); demonstrated numerical stability and practical tuning of s-step methods in large-scale experiments.

HONORS & MEMBERSHIPS

- Phi Beta Kappa; Upsilon Pi Epsilon
- COMAP ICM 2022 — Meritorious Winner; Dean's List (all semesters)
- Wake Forest Research Fellowship (2023); George Washington Greene Scholarship (2023); UPE Scholarship (2023)

SKILLS

Programming: Python, PyTorch, C/C++, Triton, NumPy, CuPy, JAX, R, MATLAB
ML/DS: Deep Learning, LLMs/Transformers, Diffusion Models, Computer Vision, Graphs/Networks, Time Series, GLMs/Regularization, Bayesian/Hierarchical Models
Systems: Kernel fusion, memory tiling, mixed precision, profiling (nsys, nvprof), HPC, Linux
Data/Tools: Git, LaTeX, Pandas, Scikit-learn
Languages: English (advanced), Mandarin (native)