Yifei Cai

Postdoctoral Research Associate
Department of Computer Science
Iowa State University · Ames, Iowa, USA

yifeic@iastate.edu
caiyifei2008.github.io
Google Scholar

Research Interests: AI Security & Privacy; Efficient Secure Inference; Privacy-Preserving AI Systems.

Academic Appointments

Iowa State University — Department of Computer Science, Ames, IA Oct. 2025–Present
Postdoctoral Research Associate Supervisor: Prof. Chunsheng Xin

- Conduct research on secure and trustworthy AI systems, extending cryptography-aware optimization toward privacy-preserving foundation models, selective protection of sensitive information, and efficient inference under real-world privacy and computational constraints.
- Collaborate with faculty on research initiatives, proposal development, and privacy-preserving AI platform design.
- Co-mentor Ph.D. students in AI security, guiding research development and contributing to work published at ICML 2026.

Education

Old Dominion University, Norfolk, VA 2025
Ph.D. in Electrical and Computer Engineering
Advisors: Prof. Chunsheng Xin and Prof. Hongyi Wu
Dissertation: *Towards Efficient Privacy-Preserving Deep Learning: HE-Friendly Structures, Flexible Pruning, HE-Efficient Architectures, and Secure Transformer Token Drop*

Beijing Institute of Technology, Beijing, China
M.S. in Electrical Engineering 2018
B.S. in Electrical Engineering 2015

Peer-Reviewed Publications

[1] **Yifei Cai**, Zhuoran Li, Yizhou Feng, Qiao Zhang, Hongyi Wu, Danella Zhao, and Chunsheng Xin
"SecDTD: Dynamic Token Drop for Secure Transformers Inference."
*IEEE European Symposium on Security and Privacy (**EuroS&P**), 2026.* DOI

[2] **Yifei Cai**, Yizhou Feng, Qiao Zhang, Chunsheng Xin, and Hongyi Wu
"Towards Zero Rotation and Beyond: Architecting Neural Networks for Fast Secure Inference with Homomorphic Encryption."
*IEEE Conference on Secure and Trustworthy Machine Learning (**SaTML**), 2026.* DOI

[3] Xiaozuo Shen, **Yifei Cai**, Rui Ning, Chunsheng Xin, and Hongyi Wu
"DF-LoGiT: Data-Free Logic-Gated Backdoor Attacks in Vision Transformers."
*International Conference on Machine Learning (**ICML**), PMLR 306:110706–110728, 2026.* Paper

[4] Zhuoran Li, Hanieh Totonchi Asl, **Yifei Cai**, Ebrahim Nouri, and Danella Zhao
"TAMI-MPC: Trusted Acceleration of Minimal-Interaction MPC for Efficient Nonlinear Inference."
*Design Automation Conference (**DAC**), 2026.* arXiv

[5] Zhuoran Li, Hanieh Totonchi Asl, Ebrahim Nouri, **Yifei Cai**, and Dan Zhao
"Silentflow: Leveraging Trusted Execution for Resource-Limited MPC via Hardware-Algorithm Co-design."
*Asia and South Pacific Design Automation Conference (**ASP-DAC**), pp. 737–743, 2026.* DOI

[6] **Yifei Cai**, Qiao Zhang, Rui Ning, Chunsheng Xin, and Hongyi Wu
"MOSAIC: A Prune-and-Assemble Approach for Efficient Model Pruning in Privacy-Preserving Deep Learning."
*ACM Asia Conference on Computer and Communications Security (**AsiaCCS**), pp. 1034–1048, 2024.* DOI

[7] Yizhou Feng, Qiao Zhang, **Yifei Cai**, Hongyi Wu, and Chunsheng Xin
"TILE: Input Structure Optimization for Neural Networks to Accelerate Secure Inference."
*Annual Computer Security Applications Conference (**ACSAC**), pp. 1203–1216, 2024.* DOI

[8] Qiao Zhang, Tao Xiang, **Yifei Cai**, Zhichao Zhao, Ning Wang, and Hongyi Wu
"Privacy-Preserving Machine Learning as a Service: Challenges and Opportunities."
IEEE Network, 37(6):214–223, 2023. DOI

[9] **Yifei Cai**, Qiao Zhang, Rui Ning, Chunsheng Xin, and Hongyi Wu
"Hunter: HE-Friendly Structured Pruning for Efficient Privacy-Preserving Deep Learning."
*ACM Asia Conference on Computer and Communications Security (**AsiaCCS**), pp. 931–945, 2022.* DOI

Preprints and Manuscripts

[P1] **Yifei Cai**, Zhuoran Li, Xiaozuo Shen, Hongyi Wu, and Chunsheng Xin
"DegreeSpar: Structured Degree Sparsity for Efficient Secure Transformer Inference."
arXiv preprint, under review, 2026. arXiv

[P2] Xiaozuo Shen, **Yifei Cai**, Tian Tan, Rui Ning, Chunsheng Xin, and Hongyi Wu
"When Trees Are Not Enough: Learning Mixed-Topology Feature Graphs with Adaptive Graph Sparse Autoencoders."
arXiv preprint, under review, 2026. arXiv

Selected Research Contributions

Cryptography-Aware Model Optimization. Pioneered HE-aware structured pruning with *Hunter* (AsiaCCS 2022), identifying and defining HE-friendly structures that helped establish a new optimization paradigm for linear operations in HE-based privacy-preserving inference. This foundational work has influenced subsequent research and served as a reference baseline for later optimization methods. Building on this foundation, *MOSAIC* (AsiaCCS 2024) introduced a more flexible prune-and-assemble approach, while *StriaNet* (SaTML 2026) advanced beyond model pruning to architect inherently HE-efficient neural networks, achieving up to $9.78\times$ speedup.

Efficient Secure Transformer Inference. Developed *SecDTD* (EuroS&P 2026), a secure dynamic token-dropping approach achieving up to $4.47\times$ end-to-end acceleration for secure Transformer inference. Its oblivious median-selection protocol, *OMSel*, achieves up to $16.9\times$ speedup over Bitonic Sort, addressing a key latency bottleneck in secure token selection under MPC.

Selective Privacy Protection for Practical AI (Ongoing). Advancing practical and scalable privacy-preserving AI systems that enable the broader deployment of advanced AI while protecting sensitive information. Current research investigates privacy-preserving LLM inference through selective protection of sensitive information and the separation of plaintext and encrypted computation, exploring fundamental trade-offs among privacy protection, model utility, and computational efficiency across real-world AI applications.

Teaching Experience

Old Dominion University, Norfolk, VA Jan. 2023–Aug. 2025
Tutor, School of Cybersecurity

- Provided supplemental instruction to undergraduate and graduate students, reinforcing classroom teaching and helping students develop a deeper understanding of cybersecurity concepts.
- Guided students through hands-on cybersecurity labs and attack-and-defense exercises, helping them apply security concepts to realistic scenarios and develop practical cybersecurity skills.
- Supported CYSE 301 (undergraduate); CYSE 601, CYSE 602, and CYSE 695 (graduate).

Proposal Development

- **NSF SaTC Research Proposal.** Contributed to the technical development and preparation of a faculty-led NSF proposal on cryptography-aware model compression and efficient privacy-preserving inference.
- **Interdisciplinary Privacy-Preserving AI Platform.** Developing, in collaboration with faculty, the research concept and technical framework for a privacy-preserving AI platform integrating automated secure-model conversion, selective data protection, and distributed edge-cloud computing for agricultural research.

Professional Service

IEEE Transactions Manuscript Reviewing 2023–Present
Reviewed multiple manuscripts for *IEEE Transactions on Computers*, *IEEE Transactions on Cloud Computing*, *IEEE Transactions on Services Computing*, and *IEEE Transactions on Mobile Computing*.

Last updated: October 2026.