Ashkan Shahbazi

Ph.D. Candidate in Computer Science, Vanderbilt University

Nashville, TN · +1-615-870-6448 · ashkan.shahbazi@vanderbilt.edu

ashkan-shahbazi.com · Google Scholar · LinkedIn

Education

Vanderbilt University · *Ph.D. in Computer Science; Advisor: Prof. Soheil Kolouri* — Aug 2022 – Expected 2027

Amirkabir University of Technology · *B.Sc. in Biomedical Engineering* — Sep 2016 – Jun 2020

Research Summary

My research aims to reduce the compute and memory cost of LLM inference while preserving model quality, through efficient attention, KV cache retrieval for long-context decoding, and adaptive model cascades. This work is grounded in geometric deep learning.

Research Experience

Graduate Research Assistant

Aug 2022 – Present

Vanderbilt University, Department of Computer Science — Nashville, TN

KV Cache Retrieval for Long-Context Decoding

OVAL · first author · arXiv 2026

- Designed OVAL, a training-free KV page summary weighting key directions by their effect on attention output; same format and scoring complexity as key-only spectral summaries at equal rank and page size.
- At a 2,048-token KV budget on long reasoning (DeepSeek-R1-Distill 7B–14B; MATH500, AIME24, GPQA): best of 8 compressed methods in all 9 settings (one tie), **+3.7 to +3.9 percentage points** avg@k over the strongest baseline, up to **+7.1 on AIME24**.
- At 100K+ token contexts (InfiniteBench, Llama-3.1-8B): **87.2% of full-cache accuracy** vs 83.9% for the strongest baseline, FreeKV, at 7.4% higher decode latency.

Efficient Attention via Geometry

ESPFormer · LOTFormer · LUNA

- **ESPFormer** (first author, ICML 2025): one-pass doubly stochastic attention, no Sinkhorn; improves on softmax and Sinkformer in image, point-cloud, sentiment, and translation tasks.
- **LOTFormer** (first author): linear-time doubly stochastic attention via learned pivots; on ImageNet-1K at comparable FLOPs, **+2.6 top-1 over softmax** on DeiT-Tiny (72.2 → 74.8).
- **LUNA** (co-first author): linear attention with a learned kernel; converting fine-tuned BERT and ViT-B/16 from softmax recovers most of their accuracy, well above fixed linearizations.

Industry Experience

Machine Learning Engineer Intern (Ph.D.)

Sep 2025 – Nov 2025

Instacart, Content AI — Remote (San Francisco, CA)

- Deployed a multi-LLM cascade for catalog attribute extraction (50+ attributes, ~150K products): **matches the F1 of running o3 on every item at 75–80% lower cost per item**.
- Set the escalation rule by benchmarking three confidence estimators on a 147K-sample golden set, then distilled them into a lightweight 5-level verifier at a fraction of the ensemble's compute.

Additional Research

- **Geometric machine learning** (ICLR 2025 and ICML 2024 spotlights, TMLR 2026): fast sliced transport plans and distances, including Expected Sliced Transport Plans, the theory behind ESPFormer; spherical sliced Wasserstein; and scalable Gromov–Wasserstein matching (MinGSGW).
- **Real-time surgical simulation** (co-first author, AAAI 2026; U.S. patent application): physics-informed neural models of soft-tissue deformation and cutting trained on FEM data, from multi-grasper tissue (Neural-Augmented Kelvinlet) to whole organs (SurgFormer).

Selected Publications

14 papers and preprints · Full list on Google Scholar · * equal contribution

Efficient Attention and Inference

OVAL: Output-Aware Local Page Bases for KV Cache Retrieval — arXiv 2026
A. Shahbazi, C. Thrash, S. Kolouri

ESPFormer: Doubly Stochastic Attention with Expected Sliced Transport Plans — ICML 2025
A. Shahbazi, E. Akbari, D. Salehi, X. Liu, N. Naderializadeh, S. Kolouri — arXiv · Code

LOTFormer: Doubly Stochastic Linear Attention via Low-Rank Optimal Transport — arXiv 2025
A. Shahbazi, C. Thrash, Y. Bai, K. Hamm, N. Naderializadeh, S. Kolouri — arXiv

LUNA: Linear Universal Neural Attention with Generalization Guarantees — arXiv 2025
A. Shahbazi*, P. He*, A. Abbasi, Y. Bai, X. Liu, E. Akbari, N. Naderializadeh, S. Kolouri — arXiv

Geometric Machine Learning

Expected Sliced Transport Plans — ICLR 2025
X. Liu, R. D. Martín, Y. Bai, **A. Shahbazi**, M. Thorpe, A. Aldroubi, S. Kolouri

Linear Spherical Sliced Optimal Transport: A Fast Metric for Comparing Spherical Data — ICLR 2025 SPOTLIGHT
X. Liu, Y. Bai, R. D. Martín, K. Shi, **A. Shahbazi**, B. A. Landman, C. Chang, S. Kolouri

Stereographic Spherical Sliced Wasserstein Distances — ICML 2024 SPOTLIGHT
H. Tran*, Y. Bai*, A. Kothapalli*, **A. Shahbazi**, X. Liu, R. D. Martín, S. Kolouri

A Case for Vanilla SWD: New Perspectives on Informative Slices, Sliced-Wasserstein Distances, and Learning Rates — TMLR 2026
H. Tran, Y. Bai, **A. Shahbazi**, J. R. Hershey, S. Kolouri

MinGSGW: A Scalable Path to Gromov–Wasserstein — arXiv 2026
A. Shahbazi, X. Liu, P. He, S. Kolouri — arXiv

Surgical Simulation

Neural Augmented Kelvinlet for Real-Time Soft Tissue Deformation Modeling — AAAI 2026
A. Shahbazi*, K. Pereira*, J. S. Heiselman, E. Akbari, et al., M. I. Miga, S. Kolouri — arXiv · Code

SurgFormer: Scalable Learning of Organ Deformation with Resection Support and Real-Time Inference — **arXiv 2026**
arXiv
A. Shahbazi*, E. Akbari*, K. Pereira, J. S. Heiselman, et al., M. I. Miga, S. Kolouri

System and Method for Real-Time Soft Tissue Simulation via Hybrid AI-Driven Physics-Informed Modeling — **U.S. patent app. 2025**
A. Shahbazi, K. Pereira, J. S. Heiselman, E. Akbari, et al., N. Simaan · U.S. Application No. 63/998,405

Skills

Languages	Python, C++, C, Java
ML frameworks	PyTorch, Hugging Face Transformers, Accelerate, JAX, vLLM, FlashAttention
GPU and systems	CUDA, CUTLASS, Nsight, multi-GPU inference, Linux, Docker, Git, AWS, GCP

Honors and Awards

Machine Learning Summer School (MLSS) NYC, Columbia University; admitted with program grant — 2026

Ranked 335th of 140,000+ candidates, Iran's National University Entrance Exam — 2016

Service and Teaching

Reviewer: NeurIPS, ICML, ICLR (2025, 2026); AAAI, ECCV (2026) — 2025 – Present

Teaching Assistant: Geometric Deep Learning (graduate), Vanderbilt University — Spring 2024