

Hossein Entezari Zarch

(213)709-9486 ◇ entezari@usc.edu ◇ <https://hoenza.github.io/>

EDUCATION

University of Southern California, Los Angeles, California

2023 – Present

Ph.D. in Computer Science

Advisor: Prof. Murali Annavaram

Thesis: “**Efficient Large-Scale Machine Learning Systems;**

Application in Efficient Large Language Models Inference & Training”

University of Southern California, Los Angeles, California

2023 – 2024

M.Sc. in Computer Science

GPA: 3.95/4.0

University of Tehran, Tehran, Iran

2017 – 2022

B.Sc. in Computer Engineering[Software]

GPA: 18.59/20.0

Advisors: Prof. Hamed Kebriaei & Prof. Pooya Shariatpanahi

Thesis: “**Incentive Mechanism for Reliable Coded Federated Learning;**

Application in Distributed Edge Computation”

RESEARCH DIRECTIONS

- ◇ Efficient LLM Inference & Training
- ◇ Large-Scale ML Systems
- ◇ Memory–Compute Trade-offs in Transformers
- ◇ Dynamic/Sparse Attention Mechanisms

PUBLICATIONS (* indicates equal contribution.)

- ◇ **Hossein Entezari Zarch**, Lei Gao, Chaoyi Jiang, Murali Annavaram. “*DELTA: Dynamic Layer-Aware Token Attention for Efficient Long-Context Reasoning.*” arXiv preprint, 2025. [PDF]
- ◇ **Hossein Entezari Zarch***, Lei Gao*, Chaoyi Jiang, Murali Annavaram. “*DEL: Context-Aware Dynamic Exit Layer for Efficient Self-Speculative Decoding.*” **COLM** 2025. [PDF]
- ◇ Chaoyi Jiang*, Lei Gao*, **Hossein Entezari Zarch**, Murali Annavaram. “*KVPR: Efficient LLM Inference with I/O-Aware KV Cache Partial Recomputation.*” **ACL Findings** 2025. [PDF]
- ◇ Chaoyi Jiang*, Sungwoo Kim*, Lei Gao, **Hossein Entezari Zarch**, Won Woo Ro, Murali Annavaram. “*MARché: Fast Masked Autoregressive Image Generation with Cache-Aware Attention.*” arXiv preprint, 2025. [PDF]
- ◇ Arun Ramachandran, R. Govindarajan, Prakash Raghavendra, Murali Annavaram, **Hossein Entezari Zarch**, Chaoyi Jiang, Lei Gao. “*Balancing Memory and Compute (BMC) of Attention Blocks: An Effective Technique for Speculative LLM Inferencing.*” (under review)
- ◇ **Hossein Entezari Zarch**, Abdulla Alshabanah, Chaoyi Jiang, Murali Annavaram. “*CADC: Encoding User-Item Interactions for Compressing Recommendation Model Training Data.*” **RecSys** Workshop, 2024. [PDF]
- ◇ Chaoyi Jiang*, Abdulla Alshabanah*, **Hossein Entezari Zarch**, Keshav Balasubramanian, Murali Annavaram. “*HuffmanEmbed: Using Huffman Coding for Embedding Table Compression in Deep Learning Recommendation Models.*” **EuroSys** Poster, 2025. [PDF]
- ◇ **Hossein Entezari Zarch***, Milad Soltany*, Hesam Mojtahedi*, Amirhossein Kazerouni*, Alireza Morsali, Azra Abtahi, Farokh Marvasti. “*Ensemble Neural Representation Networks.*” arXiv preprint, 2022. [PDF]
- ◇ Seyed Masoud Rezaei, **Hossein Entezari Zarch**, Hesam Mojtahedi, Nahid Chegeni, Amir Danyaei. “*Feasibility Study of Synthetic DW-MR Images Using GANs.*” **AMR**, 2022. [PDF]
- ◇ Seyed Masoud Rezaei, Mohammadreza Ghorvei, Razzagh Abedi-Firouzjah, Hesam Mojtahedi, **Hossein Entezari Zarch**. “*Detecting COVID-19 in Chest Images via Transfer Learning.*” **EJRN**, 2021. [PDF]

RESEARCH & INDUSTRY EXPERIENCE

Graduate Research Assistant, SCIP Lab, USC

Jan. 2023 – Present

Advisor: Prof. Murali Annamalai

Research on efficient LLM inference and recommendation systems. Contributed to multiple projects published or under review at top-tier venues.

- ◊ **DELTA:** Built a dynamic sparse attention module combining layer-aware token selection, page-based KV caching, and adaptive query refresh for efficient long-context LLM inference.
- ◊ **DEL:** Designed a dynamic exit framework that adapts layer depth and speculation length during self-speculative decoding using token-per-layer metrics and confidence-based control.
- ◊ **MARChé:** Developed a training-free cache-aware attention framework with selective KV refresh for efficient masked autoregressive image generation.
- ◊ **KVPR:** Developed an I/O-aware LLM inference framework using partial KV-cache recomputation and asynchronous CPU–GPU overlap to minimize PCIe bottlenecks and maximize throughput.
- ◊ **CADC:** Designed matrix-factorized compression for efficient large-scale recommender training.
- ◊ **HuffmanEmbed:** Built frequency-aware embedding compression with Huffman coding for DLRMs.

Software Engineer Intern, Divar, Tehran, Iran

Sept. 2022 – Dec. 2022

Team: Search & Submit

Contributed to large-scale backend search systems while gaining experience in microservice design and integration.

Undergraduate Research Assistant, University of Tehran

Sept. 2020 – Jul. 2022

Advisor: Prof. Behnam Bahrak

- ◊ **Efficient INR:** Developed an ensemble neural representation model with parallel lightweight sub-networks and FLOP-constrained optimization for efficient signal reconstruction.
- ◊ **Real-Time Object Detection:** Optimized YOLO and Fast-RCNN pipelines for robotic sorting, achieving real-time inference with balanced accuracy and speed.

Undergraduate Research Assistant, MSL Lab, Sharif University of Technology

Mar. 2019 – Oct. 2021

Advisor: Prof. Farokh Marvasti

- ◊ **Neural Machine Translation:** Explored RNN and Transformer architectures (LSTM, GPT, BERT) for bilingual translation, analyzing accuracy–efficiency trade-offs.

SKILLS

LLM Inference & Systems Optimization:

- ◊ Efficient LLM Serving, KV Page Management, Request Scheduling, Prefix Caching
- ◊ Sparse Attention, Memory-Aware Inference, Retrieval-Augmented Generation
- ◊ Speculative Decoding, Early-Exit and Layer-Skipping Strategies

Machine Learning & Modeling:

- ◊ Transformers (GPT, BERT), Signal Reconstruction, Federated Learning, GANs
- ◊ Recommender Systems (DLRMs, Embedding Compression, Matrix Factorization)
- ◊ Object Detection (YOLOv3/v5, Fast-RCNN, MobileNet)

Frameworks & Infrastructure:

- ◊ PyTorch, Hugging Face, vLLM, SGLang
- ◊ C++, Python, CUDA, Bash
- ◊ Docker, Kubernetes, gRPC

VOLUNTEER SERVICE

Mentorship: USC CURVE (Fall 2024, Spring & Fall 2025), USC VSI (Summer 2025)

Talks: DEL for Efficient Speculative Decoding LLM Inference (AMD 2025)

TEACHING EXPERIENCE

- ◊ CS 102: Fundamentals of Computation
- ◊ CS 585: Database Systems
- ◊ CS 100: Explorations in Computing

Spring 2023 - 2025
Summer 2023, Fall 2025
Fall 2023