SHAOYI ZHENG

+1 (201) 205-7865 |seanalpaca030818@gmail.com | Homepage |Linkedin

EDUCATION

New York University **New York, NY**
Doctor of Philosophy in Computer Science, Courant Institute of Mathematical Sciences Sep 2025 -- Present

New York University **New York, NY**
Bachelor of Science, Major in Computer Science with a Minor in Mathematics (GPA:3.9; Major GPA:4.0) Aug 2021 -- May 2025

Selected Publications

- **S. Zheng***, W. Lu*, et al. *ToMA: Token Merge with Attention for Diffusion Models.* ICML 2025, PMLR 267:40930–40951.
- **S. Zheng**, et al. *HilbertA: Hilbert Attention for Image Generation with Diffusion Models.* Under Review at ICLR 2026.
- **S. Zheng**, et al. *Submodular Context Partitioning and Compression for In-Context Learning.* Under Review at ACL 2026.

PROJECTS

ToMA: Token Merging with Attention for Diffusion Models[Paper|Github] **Sep 2023 -- May 2024**
Co-First Author *ICML 2025(Poster)*

- Proposed **ToMA**, a GPU-aligned token merging framework that reformulates merging as an attention-like linear transformation with invertible unmerge, enabling practical acceleration of diffusion models without quality degradation.
- Applied submodular optimization to select representative tokens, providing both theoretical guarantees on information coverage and replacing previous discrete token selection operation with matrix operation to improve both efficiency and generation fidelity.
- Co-designed GPU-efficient merging/unmerging using pure attention-like matrix operations to minimize overhead. Further leverage spatial locality for parallel processing across local image tiles and temporal redundancy across layer and timesteps for shared computation, significantly reducing FLOPS and real-world latency.
- Achieved notable acceleration on both Unet and DiT architectures: up to **1.3**× speedup on Flux.1 and **1.4**× speedup on SDXL, exceeding prior baseline runtime without quality degradation measured by FID, CLIP, and DINO Score.

Hilbert Attention for Image Generation with Diffusion Models[Paper] **May 2025 -- Sep 2025**
First Author *Under Review at ICLR 2026*

- Proposed **HilbertA**, a sparse attention mechanism based on the Hilbert curve that jointly preserves 2D spatial locality and enables contiguous memory access, improving both sparsity efficiency and memory throughput.
- Designed Hilbert-curve sparse attention with reordering, tiling, and sliding strategies to support local modeling and global information flow, while maintaining coalesced GPU memory access and preserving image locality structure.
- Developed custom sparse attention kernel in Triton and integrated LoRA fine-tuning to maximize information flow.
- Achieved up to **4.17**× speedup on Flux.1 while maintaining comparable image quality, demonstrating a superior speed–quality trade-off over other dense and 2D sparse attention baselines.

Efficient Long-Context LLM KV Recomputation via Small Model Guidance **May 2025 -- Present**
First Author *Manuscript in preparation for ICML 2026.*

- Proposed **Speculative-Recompute**, a method to alleviate the prefill bottleneck in long-context LLMs by leveraging a smaller sibling model to predict critical tokens for selective recomputation.
- Leveraged a hybrid guidance strategy combining token-mixing consistency of attention across model scales and token-level entropy to estimate token importance, enabling the small model to efficiently identify salient tokens for selective recomputation.
- Achieved up to **9.4**× TTFT speedup on 0.6B–8B Qwen3 models at the same recomputation ratio, with an average **8%** accuracy improvement on the Longbench dataset compared to baseline methods.

Submodular Context Partitioning and Compression for In-Context Learning [Paper] **Jan 2024 -- May 2024**
First Author *Under Review at ACL 2026 (Short Paper Track)*

- Proposed **Sub-CP**, a submodular, block-aware context selection framework that controls a diversity–coherence spectrum and supports offline precomputation for scalable ICL.
- Designed four partition strategies—Global Diverse, Global–Local Diverse, Local Diverse, Local Coherent—to balance global coverage and local structure, usable as a drop-in pre-processing step across ICL pipeline.
- Integrated Sub-CP into DENSE, ICAE, and CEPE, achieving **+29.2 %** on TREC (DENSE/PoE), average **+6.29%** on ICAE, and best overall CEPE performance with Global Diverse, showing consistent gains across SST-2/5, MR, TREC, and AG News

PROFESSIONAL EXPERIENCE

SenseTime Technology **Onsite, Shanghai, China**
Machine Learning Engineer Intern May 2023 -- Aug 2023

- Curated **8M text QA corpus** for customer service AI Chatbot through contextual augmentation and embedding-based filtering.
- Applied curated data to fine-tune chatbot model, improving F1 score by **7%** in downstream production evaluation.

Tencent Technology **Onsite, Shanghai, China**
Machine Learning Engineer Intern May 2024 -- Aug 2024

- Constructed a 1M+ synthetic facial dataset via diffusion with ControlNet and LoRA through distributed multi-node pipelines.
- Fine-tuned a 1B-parameter multimodal anti-spoofing model on curated data using **8×H100 GPUs** under a distributed data-parallel training framework, achieving **97% accuracy** on a customized spoof-detection benchmark.