Siyuan **Liang**

Algorithm Research Engineer · Long-Context Modeling

Hangzhou, China

✉ 354560462@qq.com | 🏠 siyuanseever.github.io | siyuanseever | 🎓 Google Scholar | INFP

Summary

Research interests include long-context modeling, long-term memory, and recurrent architectures. Designed and implemented block-recurrent attention and state-continuous training, and observed consistent out-of-context-length gains (longer extrapolation leads to lower loss). Experienced in shipping algorithms at Megvii (fingerprint/face liveness, display demura, XR hand-ray stabilization), and open-sourced a minimal C inference implementation with persistent long-term memory (llama2Rnn.c).

Research

Long-term memory
Research

Block-recurrent attention + state-continuous training — *2022 – Present*

- Proposed block-recurrent attention: pass hidden states across chunks via KV cache to enable long-term memory.
- Designed state-continuous training: carry hidden states across batches with sequentially continuous data.
- **Out-of-train-length gains:** longer extrapolation yields lower prediction loss, unlike many prior Transformer extrapolation tricks.
- Drop-in attention replacement; validated length extrapolation on TinyStories (train 256, eval up to 4096).
- Structural analysis: evaluated RoPE/NTK scaling, interpolation, truncation, and local-attention baselines; identified key bottlenecks in positional extrapolation and attention-entropy dilution.
- Open-sourced a minimal C inference implementation with persistent long-term memory (llama2Rnn.c).

Length extrapolation
Research

LEDiT: length-extrapolatable Diffusion Transformer without positional encoding — *2025*

- Removed explicit positional encoding to avoid degradation under extrapolation.
- Used causal attention to implicitly encode global position, plus a local enhancement module for fine-grained details.
- Achieved up to 4× resolution extrapolation (256×256 → 512×512) on conditional and text-to-image tasks, outperforming prior extrapolation methods.

Work Experience

Megvii / JIIOV (Megvii incubation)
Beijing, China

Algorithm Researcher — *2019 – 2025*

- Researched long-context and memory architectures; drove experiments on block-recurrent attention and state-continuous training.
- Observed out-of-train-length gains (longer extrapolation ⇒ lower loss).
- Built an LLM retrieval/reranking demo; improved nDCG@10 by 20+ points.
- Open-sourced llama2Rnn.c: minimal C inference implementation with persistent long-term memory.
- Delivered and optimized fingerprint liveness across multiple products/modules; improved ModelZoo search/distillation/release pipeline.
- Improved cross-project performance by 1–10 points via randmix/blur/resize augmentation and weight averaging.
- Polarization-based face liveness demo: 2D false-negative rate 7% → 0.1%; face detection rate 86% → 92%.
- Shipped display demura pipeline; reduced runtime from 120s → 20s via compression and pipeline optimization.
- XR hand-ray stabilization: reduced jitter by 40% using temporal reference inputs.

Writing & Output

Selected Papers
NeurIPS / IEEE

(Publications) — *2019 – 2025*

- LEDiT: Your Length-Extrapolatable Diffusion Transformer without Positional Encoding, NeurIPS 2025.
- SimpleDG: Simple Domain Generalization Baseline without Bells and Whistles, ECCV Workshop 2022.
- An End-to-End Anti-jamming Target Detection Method based on CNN, IEEE Sensors Journal 2021.
- Waveform design for cognitive radar in presence of jammer using Stackelberg game, The Journal of Engineering 2019.

Selected Repositories
GitHub

(Open Source) — *2023 – Present*

- llama2Rnn.c: minimal C inference implementation of memory attention with demo and training code
- LEDiT: PyTorch implementation (NeurIPS 2025)
- SimpleDG: training and evaluation code for the ECCV 2022 Workshop NICO Challenge

Honors & Awards

2022	**2nd place in two tracks; 1st overall**, NICO Challenge 2022 (Domain Generalization)	*China*
2021	**1st place**, LivDet 2021 Fingerprint Liveness Detection Competition	*International*

Education

Xidian University *Xi’an, China*

M.S. in Electronic and Communication Engineering *2016.07 – 2019.06*

- School of Electronic Engineering (research-oriented training).

Xidian University *Xi’an, China*

B.S. in Electronic Information Science and Technology *2012.07 – 2016.06*

Skills

Research	Long-context modeling; persistent memory; structural analysis; ablation design
Models	Transformers; RNNs; attention variants; length extrapolation
Engineering	PyTorch; C/C++; inference optimization; data and tooling
Applications	Face liveness; fingerprint liveness; display demura pipeline; LLM retrieval