

# Yiming Cheng

+1 7736907675 |eaminc0328@gmail.com |Chicago,Illinois

## EDUCATION

<b>Tsinghua University</b>	Department of Electronic Engineering	<i>Sep.2019-Jul.2024</i>
Bachelor of Engineering in <b>Electronic Engineering(Major)</b>		
Minor in <b>Statistics</b> , Minor in <b>Laws</b>		
<b>University of Chicago</b>	Department of Computer Science	<i>Expected:Jan.2026</i>
Master in <b>Computer Science(Pre-doc)</b>		
• <b>GPA: 3.95/4.0</b>		

## PUBLICATIONS

- Yi Yang, **Yiming Cheng**, Hao Feng, Zhu Han “**Emotion-Aware Scene Adaptation: A Bandwidth-Efficient Approach for Generating Animated Short**”, MDPI-sensors 2024
- **X Lan, Y Cheng, L Sheng, C Gao, Y Li** “**Depression detection on social media with large language models**”,arXiv preprint arXiv:2403.10750
- Lan, x, Piao, ..., **Cheng, Y**, Gao, c, Li, Y, Niu, Y, Song, Y, Gai, K, & Li, Y(2024). Recommendation for Inclusivity of Underrepresented Producers in Usergenerated Content Platforms. In recycle
- Yi Yang, Hao Feng, **Yiming Cheng**, Yitong Ma, Zhu Han, “Minimizing Hallucinations and Communication Cost: Adversarial Debate and Voting Mechanisms in LLM-Based Multi-Agents,” MDPI Applied Science 2025
- **Yiming Cheng**, “Research on Recommendation System Technology Based on Large Language Models,” Graduation Design, Tsinghua University, 2024.
- **Patent**: Yi Yang, **Yiming Cheng**, Hao Feng, et al. “A Semantic Encoding and Decoding Framework for Converting Visual Content into Virtual Animated Visual Representations.”

## RESEARCH & PROJECT EXPERIENCES

<b>Lmcache Team</b>	<b>Sep.2024-present</b>
<b>Open Source Contributor</b>	<b>Advisor: Prof. Junchen Jiang</b>

### Working on open-source project:

- **LMCache**: The first open-source Knowledge Delivery Network (KDN) that accelerates LLM applications up to 8x faster, at 8x lower cost.
- **VLLM/production stack**: Scale from single vLLM instance to distributed vLLM deployment without changing any application code. Now an official project under vLLM.
- Maintain and contribute to Open-Source repo. Working on Router Scheduling and cloud deployment. Contributed 1262 lines of code.

<b>Future Intelligent Lab(FIBLAB), Tsinghua University</b>	<b>Jul.2022—Jun.2024</b>
<b>Research Assistant</b>	<b>Advisor: Prof. Yong Li</b>

### Recommendation for Inclusivity of Underrepresented Producers in User-generated Content Platform

- Take the pioneering step to thinking of the inclusivity issue of underrepresented producers in UGC(user-generated content) platform.
- Propose to construct a heterogeneous graph that can enrich the relations of vulnerable populations, and further propose graph neural networks to learn representations based on enriching features from multi-hop neighbors.

### City Socioeconomic Simulator based on Large Language Models

- Use UE to Build a visual model scene of Beijing (CBD district)
- Use python to write scripts for agents to interface with LLM and design the agents' memory mechanism to do POI recommendation.(POI means point of interest in the city)
- Design and plug in agent-based recommendation systems

**Signal Processing Lab, Tsinghua University**

*Mar. 2022—present*

**Research Assistant**

**Advisor: Aso Prof. Yi Yang**

**Emotion-Aware Scene Adaptation: A Bandwidth-Efficient Approach for Generating Animated Shorts**

- Use the PyTorch framework, build an image element and emotion recognition model based on the CLIP model and InceptionV3, and use PAD (Pleasure-Arousal-Dominance) for emotion scoring.
- Enhance the generated semantics using the EmoCap model trained based on PAD scores for emotion style, ultimately achieving higher emotional coherence than the baseline on the received new video frames.

**Wireless Networking, Signal Processing and Security Lab, University of Houston**

*April 2022—present*

**Research Assistant**

**Advisor: Prof. Zhu Han, Member of NAS**

**Scalable AI Generative Content for Vehicular Network Semantic Communication**

- This project aims to establish a large-model-based semantic communication channel and test its accuracy on a vehicular dataset
- Build and test a channel in PyTorch that uses CLIP to convert original images into semantics and then uses Stable Diffusion to restore semantics back into images.

## **Internship PROJECTS**

**Beijing SmartBow Information Technology Co., Ltd. (Company)**

**Software Engineer**

*June 2023—Sep. 2023*

- Refactor the Sunflower library (the main functions include JSON parsing, MQTT, B-Stack device information parsing, and data transmission encryption) for the company's Internet of Things (IoT) data platform using Go-lang
- Perform functional and performance testing on the refactored Sunflower library.
- Collaborate with hardware interns to debug and ensure successful MQTT-based data transfer of bridge deflection, vibration frequency, and temperature data from LuZhou Bridge to the company's database.

**Beijing Thunisoft Information Technology Co., Ltd.**

**Software Engineer**

*July 2022—Sep. 2022*

- Use Spring Batch to develop a batch job scheduling system supporting complex workflows and dependency management. Scheduled tasks are executed as planned using Cron expression triggers.
- Integrate Quartz scheduler for enhanced flexibility.
- Data integrity and stability are assured with Spring transaction management and JDBC operations.

## **OTHERS**

**Scholarship:** Merit-based Predoc Scholarship of \$40,000, University of Chicago

**Field:**

**Precisely as undergraduate:** Data mining (Recommendation System, Emotion Awareness, Embodied City)

**Current and future:** System for machine learning (distributed LLM deployment, distributed KV cache, efficient ml)

Machine learning for systems (machine learning for code generation)

**Programming Skills:** Python (Pytorch, CuPy), Go (Docker, K8s), Git (Github action), Linux, C, C++, Matlab, Verilog etc.

**Personal Website:** <https://eaminc.github.io/> includes github, google scholar and other detailed information