



YICHEN GUO

M.Eng (by Research)
Computer Science and Engineering
Nanyang Technological University, Singapore

+65 89538476

yichen013@e.ntu.edu.sg

guoyichen021004@163.com

GitHub Profile

LinkedIn Profile

Google Scholar

EDUCATION

Nanyang Technological University (NTU), Singapore

Sept. 2025 - Jun. 2027

M.Eng (by Research) in Computer Science and Engineering

– Focusing on Large Language Model (LLM) and Large Vision-Language Model (LVLMS)

University of Electronic Science and Technology of China (UESTC)

Sept. 2021 - Jun. 2025

B.Eng. in Electronic Information Engineering

GPA: 3.77/4.00; 85.26/100

– AI Foundation Courses: Reinforcement Learning, Deep Learning, Machine Learning and AI, Introduction to Neural Networks, Probability Theory and Mathematical Statistics, Python Programming

PUBLICATIONS

STAR-Pro: Stage-Wise Token Adaptive Reduction with Progressive Refinement for Efficient Large Vision-Language Models

Yichen Guo, Q. Zhang, Y. Zhang, H. Cheng, J. Cao, S. Chen, T. Luo, C. Li, K. Tang, X. Huang, W. Wang, S. Zhang
AAAI 2027 (in preparation), transferred after ECCV 2026 rejection; meta-review: 6/4/3

MACT: An Explicit-Implicit Memory-Enhanced Multi-Agent System for Cryptocurrency Trading

Yichen Guo*, T. Luo*, R. Li*, F. Lin, J. Wu, K. Tang, M. Yu, G. Gan, Y. Lu, Y. Cai, J. You, L. Chen, H. Wang, L. W. Cong
EMNLP 2026 (under review via ACL ARR May)

TOPS: First-Principles Visual Token Pruning via Constructing Token Optimal Preservation Sets for Efficient MLLM Inference

T. Wang*, **Yichen Guo***, R. Huang, Z. Lu, Q. Zhang, C. Li, Y. Zhang, J. Cao, Z. Shen, Y. Du, G. Gan, W. Wang, L. W. Cong, S. Zhang
EMNLP 2026 (under review via ACL ARR May)

MAP: Map-Level Evidence Aggregation for Faithful Decoding in Multimodal Large Language Models

C. Li*, **Yichen Guo***, B. Qian, J. You, K. Tang, Y. Du, Z. Zhang, X. Huang
EMNLP 2026 (under review via ACL ARR May)

FADE: Mitigating Hallucinations by Reducing Language Priors Dominance in Large Vision-Language Models

Yichen Guo*, K. Tang*, J. You, F. Lin, R. Li, T. Luo, H. Cheng, F. Yang, S. Chen, G. Gan, J. Li, L. W. Cong, S. Zhang, W. Wang
AAAI 2027 (in preparation)

RegimeVGGT: Layer-Wise Spatially Preserving Redundancy Removal for Visual Geometry Grounded Transformer

S. Lyu, J. You, Z. Lyu, T. Li, Z. Zhao, J. Hu, K. Tang, **Yichen Guo**
NeurIPS 2026 (under review)

FORCE: Efficient VLA Reinforcement Fine-Tuning via Value-Calibrated Warm-up and Self-Distillation

S. Zhang, Y. Lou, H. Cheng, **Yichen Guo**, C. Fu, Y. Lyu, X. Zhang, H. Li, P. Wang, Z. Wang, S. Zhang
CoRL 2026 (under review)

SafeDojo: Safe Reinforcement Learning for VLA via Interactive World Model

K. Tang, P. Jia, Z. Chu, J. Wu, R. Ma, J. Cao, F. Zhao, S. Chen, **Yichen Guo**, X. Chi, C. Fan, K. Zhang, J. Xu, F. Yang, W. Mi, X. Ju, J. Tang, S. Zhang
CoRL 2026 (under review)

Deep Spectral Feature Fusion Network Combined with Attention Mechanism to Predict Heavy Metal Elements in Soil by XRF

Yichen Guo, S. Lyu, S. Chen, F. Li, X. Qu
Journal of Analytical Atomic Spectrometry (JAAS), 2026 (IF: 3.1, SCI Q2)

Mitigating Hallucinations in Large Language Models Via Decoder Layer Skipping

H. Li, J. You, **Yichen Guo**, K. Tang, S. Xie, X. Huang
IEEE ICASSP 2026 (CCF-B)

* Equal contribution.

WORK EXPERIENCE

HMI Lab, School of Computer Science, Peking University

Research Assistant; Supervisor: Prof. Shanghang Zhang

Jul. 2025 - Present

China

- Researching multimodal model efficiency and inference acceleration, including visual token pruning and adaptive token reduction for LVLMS
- Working on multimodal hallucination mitigation, faithful decoding, and language-prior reduction for Large Vision-Language Models
- Exploring embodied intelligence through Unified World Model, Action Following, and Self-Evolving Unified World-Action Model

Binance

Data Science Intern; Supervisor: Harvin Wang

Jun. 2026 - Present

Remote

- Assisting the development and public-beta rollout of Binance Ai Pro, an OpenClaw-based AI trading agent product
- Working on multi-agent cryptocurrency trading workflows, including market analysis, strategy reasoning, and trading-signal generation
- Supporting end-to-end model training, inference, and deployment for agent development, including supervised fine-tuning and reinforcement learning

Tencent LightSpeed Studios, Tencent

Data Scientist Intern; Supervisor: Don Li

Aug. 2025 - Dec. 2025

Singapore

- Conducted LLM post-training for game-specific language understanding, generation, and interactive decision-making scenarios
- Built multi-agent economic systems for gaming environments, including agent coordination, resource dynamics, and simulation workflows

Harvard AI and Robotics Lab, Harvard University

Visiting Student; Supervisor: Prof. Mengyu Wang

Sept. 2025 - Nov. 2025

United States (Remote)

- Developing multi-objective reinforcement learning agents for financial decision-making and LLM-powered trading scenarios
- Exploring multimodal visual-memory reasoning for medical models, including memory-enhanced post-training and multi-image inference
- Applying LLM/VLM post-training methods such as supervised fine-tuning and GRPO across financial and medical agent tasks

REPRESENTATIVE RESEARCH EXPERIENCE

HMI Lab, School of Computer Science, Peking University

Research Assistant; Supervisor: Prof. Shanghang Zhang

Jul. 2025 - Present

- **Topic:** Multimodal Large Language Model: Model Efficiency, Inference Acceleration
- **Purpose:** Proposed STAR-Pro, an adaptive-to-progressive visual token pruning framework for efficient LVLMM inference
- **Contributions:**
 - * Designed a training-free framework with adaptive token selection after vision encoder and progressive refinement inside LLM decoder
 - * Proposed balancing instruction relevance and representation diversity in adaptive stage, followed by progressive refinement guided by multi-token text raters
 - * Demonstrated that STAR-Pro achieves up to 90% token reduction and $9.9\times$ FLOPs reduction while maintaining 96.8% performance on LLaVA-NeXT-7B
 - * Validated strong generalizability across LLaVA series, InternVL3-8B, and LLaVA-Video for image understanding, high-resolution, and video tasks
 - * Received ECCV 2026 meta-review scores of 6/4/3 and is iterating on reviewer feedback

TECHNICAL SKILLS

LLM/VLM Training: LLaMA-Factory, verl, SLIME, PEFT/LoRA, SFT, PPO, GRPO, RLHF-style post-training

LLM/VLM Inference & Deployment: SGLang, vLLM, Hugging Face Transformers, efficient LVLMM inference, model serving, inference optimization

Evaluation: lmms-eval, VLMEvalKit, LLM/VLM benchmark evaluation, evaluation harnesses, error analysis

Trading Agents: OpenClaw, trading skills, strategy workflows, market analysis pipelines, multi-agent crypto trading, LLM-based trading assistants

Engineering & Writing: Python, PyTorch, Docker, Git, LaTeX, AI-assisted development, agentic coding, agent harness engineering, tool orchestration