Yun-Cheng (Joe) Wang

Senior Machine Learning Engineer, Apple Inc.
One Apple Park Way, Cupertino, CA 95014
Email: joewang622@gmail.com
Website: yunchengwang.github.io

Research Interests

Efficient and Scalable AI/ML: representation learning for high-dimensional data; on-device AI; time- and memory-efficient ML pipelines; privacy-preserving edge–cloud collaboration.

Knowledge-Driven Reasoning: reasoning over knowledge graphs; knowledge-grounded text generation; large language model (LLM) agents; multimodal reasoning.

Education

University of Southern California, Los Angeles, CA, USA — Jan 2021 – Dec 2023
Ph.D. in Electrical Engineering, Minor in Computer Science
Dissertation: *Green Knowledge Graph Completion and Scalable Generative Content Delivery*
Advisor: Prof. C.-C. Jay Kuo

University of Southern California, Los Angeles, CA, USA — Aug 2018 – Dec 2019
M.S. in Electrical Engineering

National Taiwan University, Taipei, Taiwan — Sep 2014 – Jun 2018
B.S. in Electrical Engineering

Employment

Senior Machine Learning Engineer — Jan 2025 – Present
Apple Inc., Cupertino, CA, USA

Leading both research and engineering efforts in foundation model agent development and evaluation. Focus areas include agent tool-use and reasoning, voice agents, and synthetic data generation.

Research Scientist — Jan 2024 – Jan 2025
Yahoo Inc., Mountain View, CA, USA

Developed models for user intent understanding and search page optimization in web search systems.

Research Intern — Jun 2023 – Aug 2023
Yahoo Inc., Mountain View, CA, USA

Researched on knowledge-enhanced entity summarization in search and recommendation pipelines.

Publications [Google Scholar]

Conference Papers

[1] Jintang Xue, **Yun-Cheng Wang**, Chengwei Wei, C.-C. Jay Kuo, "Efficient Feature Selection for Word Embedding Dimension Reduction," *Asia Pacific Signal and Information Processing Association Annual Summit and Conference (ASC)*, 2024.

[2] Zhanxuan Mei, **Yun-Cheng Wang**, C.-C. Jay Kuo, "GSBIQA: Green Saliency-guided Blind Image Quality Assessment Method," *ASC*, 2024.

[3] Tsung-Shan Yang, **Yun-Cheng Wang**, Chengwei Wei, Suya You, C.-C. Jay Kuo, "GMA: Green Multi-Modal Alignment for Image-Text Retrieval," *ASC*, 2024.

[4] Tsung-Shan Yang, **Yun-Cheng Wang**, Chengwei Wei, C.-C. Jay Kuo, "GHOI: A Green Human-Object-Interaction Detector," *IEEE MIPR*, 2024.

[5] **Yun-Cheng Wang**, Xiou Ge, Bin Wang, C.-C. Jay Kuo, "AsyncET: Asynchronous Representation Learning for Knowledge Graph Entity Typing," *ACM SIGKDD*, 2024.

[6] **Yun-Cheng Wang**, Xiou Ge, Bin Wang, C.-C. Jay Kuo, "GreenKGC: A Lightweight Knowledge Graph Completion Method," *ACL*, 2023.

[7] Xiou Ge, **Yun-Cheng Wang**, Bin Wang, C.-C. Jay Kuo, "Compounding Geometric Operations for Knowledge Graph Completion," *ACL*, 2023.

[8] Zhanxuan Mei, **Yun-Cheng Wang**, Xingze He, C.-C. Jay Kuo, "GreenBIQA: A Lightweight Blind Image Quality Assessment Method," *IEEE MMSP*, 2022. **(Top 10% Paper Award)**

[9] Bin Wang, Fenxiao Chen, **Yun-Cheng Wang**, C.-C. Jay Kuo, "Efficient Sentence Embedding via Semantic Subspace Analysis," *IEEE ICPR*, 2020.

Journal Papers

[1] Tsung-Shan Yang, **Yun-Cheng Wang**, Chengwei Wei, Suya You, C.-C. Jay Kuo, "Efficient Human-Object-Interaction (EHOI) Detection via Interaction Label Coding and Conditional Decision," *Computer Vision and Image Understanding*, 2025.

[2] Jintang Xue, **Yun-Cheng Wang**, Chengwei Wei, Xiaofeng Liu, Jonghye Woo, C.-C. Jay Kuo, "Word Embedding Dimension Reduction via Weakly-Supervised Feature Selection," *APSIPA Trans. on Signal and Information Processing*, 2024.

[3] Zhanxuan Mei, **Yun-Cheng Wang**, C.-C. Jay Kuo, "Blind Video Quality Assessment at the Edge," *IEEE Trans. on Multimedia*, 2024.

[4] Zhanxuan Mei, **Yun-Cheng Wang**, C.-C. Jay Kuo, "GreenSaliency: A Lightweight and Efficient Image Saliency Detection Method," *APSIPA Trans.*, 2024.

[5] Xiou Ge, **Yun-Cheng Wang**, Bin Wang, C.-C. Jay Kuo, "Knowledge Graph Embedding: An Overview," *APSIPA Trans.*, 2024.

[6] Jintang Xue, **Yun-Cheng Wang**, Chengwei Wei, Xiaofeng Liu, Jonghye Woo, C.-C. Jay Kuo, "Bias and Fairness in Chatbots: An Overview," *APSIPA Trans.*, 2024.

[7] Xiou Ge, **Yun-Cheng Wang**, Bin Wang, C.-C. Jay Kuo, "Knowledge Graph Embedding with 3D Compound Geometric Transformations," *APSIPA Trans.*, 2024.

[8] Zhanxuan Mei*, **Yun-Cheng Wang***, Xingze He, Yong Yan, C.-C. Jay Kuo, "Lightweight High-Performance Blind Image Quality Assessment," *APSIPA Trans.*, 2024.

[9] **Yun-Cheng Wang**, Jintang Xue, Chengwei Wei, C.-C. Jay Kuo, "An Overview on Generative AI at Scale with Edge-Cloud Computing," *IEEE Open Journal of the Communications Society*, 2023.

[10] Chengwei Wei, **Yun-Cheng Wang**, Bin Wang, C.-C. Jay Kuo, "An Overview on Language Models: Recent Developments and Outlook," *APSIPA Trans.*, 2023.

[11] Xiou Ge, **Yun-Cheng Wang**, Bin Wang, C.-C. Jay Kuo, "TypeEA: Type-Associated Embedding for Knowledge Graph Entity Alignment," *APSIPA Trans.*, 2023.

[12] **Yun-Cheng Wang**, Xiou Ge, Bin Wang, C.-C. Jay Kuo, "KGBoost: A Classification-Based Knowledge Base Completion Method with Negative Sampling," *Pattern Recognition Letters*, 2022.

[13] Xiou Ge, **Yun-Cheng Wang**, Bin Wang, C.-C. Jay Kuo, "CORE: A Knowledge Graph Entity Type Prediction Method via Complex Space Regression and Embedding," *Pattern Recognition Letters*, 2022.

[14] Fenxiao Chen, **Yun-Cheng Wang**, Bin Wang, C.-C. Jay Kuo, "Graph Representation Learning: A Survey," *APSIPA Trans.*, 2020. **(2024 Sadaoki Furui Prize Paper Award)**

[15] Bin Wang, Angela Wang, Fenxiao Chen, **Yun-Cheng Wang**, C.-C. Jay Kuo, "Evaluating Word Embedding Models: Methods and Experimental Results," *APSIPA Trans.*, 2019. **(2022 Sadaoki Furui Prize Paper Award)**

Asterisk (*) denotes equal contribution.

Talks and Presentations

Invited Talks

National Sun Yat-Sen University, EE	Mar 2024
National Taiwan University, CSIE	Dec 2023
National Tsing Hua University, CS	Dec 2023
National Yang Ming Chiao Tung University, ECE	Dec 2023
National Cheng Kung University, CS	Dec 2023
Academia Sinica, Institute of Information Science	Dec 2023
Academia Sinica, Information and Communications Technology Center	Dec 2023
National Taiwan Normal University, EE	Sep 2020

Conference Presentations

AsyncET: Asynchronous Representation Learning for Knowledge Graph Entity Typing	**KDD**, 2024
GreenKGC: A Lightweight Knowledge Graph Completion Method	**ACL**, 2023
Improving Knowledge Graph Embeddings for Entity Typing	**USC EE Research Festival**, 2023
A Multi-stage Classification Framework for Knowledge Graph Completion	**USC EE Research Festival**, 2022
Rule-Guided Knowledge Graph Completion	**USC EE Research Festival**, 2021

Awards

APSIPA Sadaoki Furui Prize Paper Award (Pioneering in Speech Processing)	2022, 2024
IEEE MMSP Top 10% Paper Award	2022
USC Viterbi Graduate Fellowship	2021
USC EE MS Honors Program	2019

Teaching Experience

Guest Lecture

EE669 Multimedia Compression (Graduate Level)	Fall 2022, Fall 2023

Teaching Assistant

EE512 Stochastic Process for Financial Engineering (Graduate Level)	Spring 2023, Fall 2023
EE503 Probability for Electrical and Computer Engineers (Graduate Level)	Spring 2019

Academic Service

Conference PC/Reviewer

ACL Rolling Review (ARR)	2024, 2025
Int'l Conf. on Knowledge Discovery and Data Mining (KDD)	2024, 2025
Int'l Conf. on Computational Linguistics (COLING)	2024, 2025
Asian Conference on Machine Learning (ACML)	2024, 2025
Conference on Language Modeling (CoLM)	2025
IEEE Int'l Conf. on Acoustics, Speech and Signal Processing (ICASSP)	2025
Empirical Methods in Natural Language Processing (EMNLP)	2023
European Conference on Machine Learning (ECML)	2022

Journal Invited Reviewer

ACM Transactions on Multimedia Computing, Communications and Applications (TOMM)
APSIPA Transactions on Signal and Information Processing
IEEE/ACM Transactions on Audio, Speech and Language Processing (T–ASLP)
IEEE Consumer Electronics Magazine (CEM)
IEEE Internet of Things Magazine (IoTM)
IEEE Journal of Biomedical and Health Informatics (JBHI)
IEEE Journal on Emerging and Selected Topics in Circuits and Systems (JETCAS)
IEEE Transactions on Artificial Intelligence (TAI)

Professional Affiliations

ACM SIGKDD, Member — Since 2024
Association for Computational Linguistics (ACL), Student Member/Member — Since 2023
IEEE Signal Processing Society, Student Member/Member — Since 2019

Research Projects

Efficient and Interpretable Language Models — Aug 2023 – Present

- Integrated structured knowledge into lightweight language models for controllable text generation.
- Developed efficient and robust generative models for high-stakes domains, including biomedical applications.
- Explored local n-gram and constituency-based mechanisms to simplify transformer self-attention.

Multimodal Retrieval and Alignment, *Sponsored by Army Research Lab (ARL)* — Aug 2023 – Present

- Unified cross-modal information via joint embedding spaces to support multimodal reasoning.
- Extracted human–object interactions using hierarchical classifiers over spatial and latent features.
- Designed an alignment module for bidirectional text–image retrieval tasks.

Efficient Reasoning on Knowledge Graphs with Lightweight Models — Jan 2021 – Dec 2023

- Predicted missing entity types and relations using compact, efficient knowledge graph models.
- Applied feature pruning to reduce model size while achieving state-of-the-art accuracy.
- Proposed new modeling of entity types to improve generalization with minimal overhead.
- Developed an asynchronous training framework that achieved top-tier results in link prediction and type classification.
- Reduced inference FLOPs by over $100\times$ and parameters by over $15\times$.

Scalable Generative AI Services under Edge–Cloud Computing — Jan 2023 – Oct 2023

- Analyzed resource constraints for GenAI deployment across edge–cloud infrastructure.
- Estimated latency under various communication and computation frameworks.
- Identified key trade-offs for building efficient, privacy-preserving, and scalable GenAI systems.

Blind Image and Video Quality Assessment, *Sponsored by Meta Inc.* — Aug 2021 – Dec 2022

- Designed no-reference ML models to estimate perceptual quality of media in real-time.
- Built a lightweight inference pipeline for efficient deployment on edge devices.
- Achieved state-of-the-art results while reducing model size by $54\times$, enabling scalable application.