Yixin Wan (Elaine)

https://elainew728.github.io

elaine1wan@g.ucla.edu | https://scholar.google.com/citations?hl=en&user=hZPIICQAAAAJ

EDUCATION

University of California, Los Angeles   Los Angeles, CA, United States

PhD in Computer Science   2022/06 - Present

Advisor: Professor Kai-Wei Chang

- *Research Interests: Building trustworthy multimodal generative models*

Bachelor of Science in Applied Mathematics, Double Major in Economics   2018/09 - 2022/03

WORKING EXPERIENCE

Research Intern, **Tencent AI Lab**   2025/06 - Present

- **Research focus:** Benchmarking and improving motion image editing in state-of-the-art image editing models.
- Proposes the task of motion image editing to target posture, action, and interaction changes in image subject(s) based on natural language prompts.
- Collects and constructs *MotionEdit*, a novel dataset and benchmark for motion image editing task.
- Proposes Motion-NFT, a novel reward training framework for improving the motion image editing ability of state-of-the-art models through a motion-specific reward for fine-grained guidance.

Applied Scientist Intern, **Amazon AGI**   2024/06 - 2024/09

- **Research focus:** Building better machine unlearning algorithms, which removes unsafe/copyright content from models without retraining.
- Synthesized and curated forget and retain datasets, conducted model fine-tuning on the constructed dataset, and benchmarked state-of-the-art unlearning methods for the SEMEval 2025 Challenge: *'Unlearning sensitive content from Large Language Models'*.
- Proposed and delivered a selective unlearning method that remarkable improves model performance on retain data post-unlearning.

Applied Scientist Intern, **Amazon**   2023/06 - 2023/09

- **Research focus:** Explore the correlation between hallucination and certainty in LLMs, using insights to reduce nonfactual generations.
- Independently designed, owned and delivered a research project on the correlation between sequence-level certainty and model hallucinations in Natural Language Generation.

Research Intern, **Microsoft Research Asia (MSRA)**   2022/05 - 2022/09

- **Research focus:** Distilling the ability to remove noise in audio signals from larger and stronger models to more efficient smaller models.
- Developed a general Knowledge Distillation (KD) framework for Deep Learning-based Noise Suppression (DNS) task and contributed >5,000 lines of project code to research group's repository.

PRE-PRINTS & SUBMISSIONS

*1. **Wan, Y.**, Chen, X., and Chang, K.W., Which Cultural Lens Do Models Adopt? Unmasking Cultural Positioning Bias in Large Language Model-Generated Interview Scripts. arXiv preprint.*

*2. Wu, D., **Wan, Y.**, and Chang, K. W., Visualized Text-to-Image Retrieval. Submission to EACL 2025*

*3. **Wan, Y.**, & Chang, K. W. Compalign: Improving compositional text-to-image generation with a complex benchmark and fine-grained feedback. arXiv preprint.*

*4. **Wan, Y.**, Subramonian, A., Ovalle, A., Lin, Z., Suvarna, A., Chance, C., ... & Chang, K. W. (2024). Survey of Bias In Text-to-Image Generation: Definition, Evaluation, and Mitigation. arXiv preprint.*

SELECTED PUBLICATIONS

*1. **Wan, Y.**, Ramakrishn, A., Chang, K. W., Cevher, V., Gupta, R. (2025). Not Every Token Needs Forgetting: Selective Unlearning Balancing Forgetting and Utility in Large Language Models. EMNLP 2025 Findings*

*2. Ramakrishn, A., **Wan, Y.**, Jin, X., Chang, K. W., Bu, Z., Vinzamuri, B., ... & Gupta, R. (2025). Lume: Llm unlearning with multitask evaluations. EMNLP 2025 Findings*

*3. Huang, J. T., Yan, Y., Liu, L., **Wan, Y.**, Wang, W., Chang, K. W., & Lyu, M. R. (2025). Fact-or-fair: A checklist for behavioral testing of ai models on fairness-related queries. EMNLP 2025 Findings*

🏆*(Best Short Paper Award)* *4. **Wan, Y.**, & Chang, K. W. (2024). White Men Lead, Black Women Help: Uncovering Gender, Racial, and Intersectional Bias in Language Agency. NAACL 2024 TrustNLP Workshop (non-archival track), ACL 2025 Main.*

*5. **Wan, Y.**, & Chang, K. W. (2024). The Male CEO and the Female Assistant: Probing Gender Biases in Text-To-Image Models Through Paired Stereotype Test. ACL 2025 Main*

6. **Wan, Y.,** Wu, D., Wang, H., & Chang, K. W. (2024). *The Factuality Tax of Diversity-Intervened Text-to-Image Generation: Benchmark and Fact-Augmented Intervention. EMNLP 2024 Main*

7. Lin, Z., Xu, Z., **Wan, Y.,** Yao, S. X., Song, X., Lin, T. H., … & Sun, Y. (2024). *VISUAL-ALPHASOCIAL: Benchmark and Self-Reflective Chain-of-Thought Generation for Visual Social Commonsense Reasoning. ACL 2025 Findings*

10. **Wan, Y.,** Pu, G., Sun, J., Garimella, A., Chang, K. W., & Peng, N. (2023, December). *"Kelly is a Warm Person, Joseph is a Role Model": Gender Biases in LLM-Generated Reference Letters. EMNLP 2023 Findings*

11. **Wan, Y.,** Zhao, J., Chadha, A., Peng, N., & Chang, K. W. (2023, December). *Are Personalized Stochastic Parrots More Dangerous? Evaluating Persona Biases in Dialogue Systems. EMNLP 2023 Findings*

12. **Wan, Y.,** Wu, F., Xu, W., & Sengamedu, S. H. (2023). *Sequence-level certainty reduces hallucination in knowledge-grounded dialogue generation. ICLR 2024 SeT-LLM Workshop*

13. **Wan, Y.,** Huang, K. H., & Chang, K. W. (2023). *PIP: Parse-instructed prefix for syntactically controlled paraphrase generation. ACL 2023 Findings*

SCHOLARSHIP

- Amazon AI PhD Fellowship Receiver, 2025.

TEACHING

Teaching Assistant

- UCLA CS 263, Natural Language Processing, Spring 2023, with Professor Kai-Wei Chang.
- UCLA CS 263, Natural Language Processing, Spring 2024, with Professor Nanyun Peng.
- UCLA CS 31, Introduction to Computer Science, Winter 2025.
- UCLA CS 35L, Software Construction, Spring 2025.

SERVICES

- *Reviewer:* ACL 2023, EMNLP 2023, ICASSP 2024, NeurIPS 2025, ICLR 2025, ICLR 2026, ACL Rolling Review
- *Program / Organizing Committee:* TrustNLP Workshop 2024 – 2025, SEMEval 2025 Challenge