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 — 2022/09 - Present

Advisor: Professor Kai-Wei Chang

- *Research Interests: Building trustworthy multimodal understanding & generative models*

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

RESEARCH INTEREST

My research interest is building high-performance controllable multimodal generative and understanding models, including generative LLMs, VLM/MLLMs, image generation and image editing models, as well as agent systems. I am experienced with large-scale real and synthetic data curation, mid- and post-training methods (SFT and RL), as well as inference-time methods (RAG, prompt enrichment, etc.).

WORKING EXPERIENCE

Research Scientist Intern, **Meta Superintelligence Labs (FAIR - TBD Friends Team)** — 2026/06 – Now

- **Research focus:** Developing data and mid-training recipes for MLLMs to natively emit multi-task representations.
- Design task-specific generative heads for MLLMs to natively emit structured representations for downstream tasks (e.g. perception).
- Conduct large scale experiments on multi-task joint mid-training recipes for internal MLLMs (from 1B dense model to >200B MoE model).
- Establish perception tool-augmented agentic pipeline to construct 200k+ challenging grounding-related data for training and evaluation.

Research Intern, **Tencent Hunyuan** — 2025/06 – 2025/12

- **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 text prompts.
- Collects and constructs *MotionEdit*, a novel dataset and benchmark for motion image editing task. Performed large-scale data pre- and post-processing on over 30K real and synthetic videos, cleaned data from over 70K raw image pairs.
- Proposes *Motion-NFT*, a novel DiffusionNFT-based reinforcement learning 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 - 2024/11

- **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.
- Designed and delivered research project on correlation between sequence-level certainty and 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 Knowledge Distillation (KD) framework for Deep Learning-based Noise Suppression (DNS) task.

PRE-PRINTS

*1. Pruksachatkun, Y., **Wan, Y.,** Chen, X., Chang, K.W., Wu, C. S. (2026) CustomerSim: Benchmarking and Aligning Multimodal Language Models as Retail User Simulators. arXiv preprint.*

*2. **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**., Chen, X., and Chang, K.W.. Which Cultural Lens Do Models Adopt? Unmasking Cultural Positioning Bias in Large Language Model-Generated Interview Scripts. ACL 2026 Main.*

2. Wu, D., **Wan, Y.***, and Chang, K. W.. Visualized Text-to-Image Retrieval. ACL 2026 Main. (Co-first author)*

*3. Liu, Z., Kim, D., **Wan, Y.,** … & Jiang, M. (2026). MTMCS-Bench: Evaluating Contextual Safety of Multimodal Large Language Models in Multi-Turn Dialogues. ACL 2026 Findings.*

*4. **Wan, Y.,** Ke, L., Yu, W., Chang, K.W., and Yu, D. MotionEdit: Benchmarking and Learning Motion-Centric Image Editing. CVPR 2026.*

*5. **Wan, Y.,** Ramakrishna, 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*

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

*7. 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*

8. 🏆(Best Short Paper Award)* **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.*

*9. **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*

*10. **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*

*11. 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*

*12. **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*

*13. **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*

*14. **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*

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

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 & AWARD

- *Reviewer:* ACL 2023, EMNLP 2023, ICASSP 2024, NeurIPS 2025, ICLR 2025, ICLR 2026, CVPR 2026, ECCV 2026, COLM 2026, ACL Rolling Review
- *Program / Organizing Committee:* TrustNLP Workshop 2024 – 2026, SEMEval 2025 Challenge
- *Best Short Paper Award, NAACL 2024 TrustNLP Workshop (non-archival track)*
- *Amazon AI PhD Fellowship Recipient,* 2025