# Dehong Xu

Phone: +1(424)440-4146, Email: xudehong1996@ucla.edu
Homepage: https://dehongxu.github.io/

ACADEMIC BACKGROUND

**University of California, Los Angeles (UCLA)**
*Ph.D. candidate in Statistics* Expected graduation: 2025
*M.S. in Statistics* Sep 2019 - Jul 2021
Co-advised by Prof. Ying Nian Wu and Prof. Song-Chun Zhu, GPA: 4.0 / 4.0

**Beijing University of Posts and Telecommunications (BUPT)**
*B.Eng. in Software Engineering* Sep 2015 - Jul 2019
GPA: 3.9 / 4.0; Ranking: 1 / 153

RESEARCH INTERESTS

LLM Alignment, Multi-modal LLM, Language Modeling, Representation Learning

SELECTED PUBLICATIONS

(* denotes equal contributions)

Deqian Kong*, Minglu Zhao*, **Dehong Xu***, Bo Pang, Shu Wang, Edouardo Honig, Zhangzhang Si, Chuan Li, Jianwen Xie, Sirui Xie, Ying Nian Wu. *"Scalable Language Models with Posterior Inference of Latent Thought Vectors."*
Preprint. In submission to ICML 2025.

Rohan Sharma, Changyou Chen, Feng-Ju Chang, Seongjun Yun, Xiaohu Xie, Rui Meng, **Dehong Xu**, Alejandro Mottini, Qingjun Cui. *"Multi-Modal Multi-Task Unified Embedding Model (M3T-UEM): A Task-Adaptive Representation Learning Framework."*
Preprint. In submission to CVPR 2025.

**Dehong Xu**, Ruiqi Gao, Wen-Hao Zhang, Xue-Xin Wei, Ying Nian Wu. *"An Investigation of Conformal Isometry Hypothesis for Grid Cells."* International Conference on Learning Representations (**ICLR**), **2025**. [Oral Presentation (1.8%)]

Deqian Kong*, **Dehong Xu***, Minglu Zhao*, Bo Pang, Jianwen Xie, Andrew Lizarraga, Yuhao Huang, Sirui Xie*, Ying Nian Wu. *"Latent Plan Transformer for Trajectory Abstraction: Planning as Latent Space Inference."* Conference on Neural Information Processing Systems (**NeurIPS**), **2024**.

**Dehong Xu**, Liang Qiu, Minseok Kim, Faisal Ladhak, Jaeyoung Do. *"Aligning Large Language Models via Fine-grained Supervision."* Annual Meeting of the Association for Computational Linguistics (**ACL**), **2024**.

Yan Xu*, Deqian Kong*, **Dehong Xu***, Ziwei Ji*, Bo Pang, Pascale Fung, Ying Nian Wu. *"Diverse and Faithful Knowledge-Grounded Dialogue Generation via Sequential Posterior Inference."* International Conference on Machine Learning (**ICML**), **2023**.

Minglu Zhao, **Dehong Xu**, Wen-Hao Zhang, Ying Nian Wu, *"A Minimalistic Representation Model for Head Direction System."* **NeurIPS** Workshop on Symmetry and Geometry in Neural Representations, 2024.

**Dehong Xu***, Ruiqi Gao*, Wen-Hao Zhang, Xue-Xin Wei, Ying Nian Wu. *"Conformal Isometry of Lie Group Representation in Recurrent Network of Grid Cells."* Proceedings of the 1st **NeurIPS** Workshop on Symmetry and Geometry in Neural Representations,

**PMLR** 197:370-387, 2023.

RESEARCH EXPERIENCE

**Amazon Inc - Search M5 Team** Jun 2024 - Sep 2024
*Applied Scientist Intern, Advisors: Xiaohu Xie and Alejandro Mottini*

**Research Topic**: Multi-modality, Representation Learning, Instruction-following VLM

- Improving Instruction-following Capability of Multi-modal Embedding Models
  - Developed a multi-modal, decoder-only framework for learning representations with instruction-following capabilities.
  - Designed and implemented a two-stage training approach: a pre-training phase for modality alignment, followed by instruction fine-tuning.
  - Our method achieved SoTA performance on multi-modal information retrieval benchmarks. (In submission to CVPR 2025).

**Amazon Inc - Alexa AGI Team & Rufus Team** Jun 2023 - Oct 2023
*Applied Scientist Intern, Advisors: Liang Qiu, Puyang Xu and Yi Xu*

**Research Topic**: LLM Post-training, Fine-grained RLHF, Reward Modeling

- Aligning Large Language Models via Fine-grained Supervision and Token-level RLHF **(Paper published in ACL 2024)**
  - Developed a fine-grained data collection method for reward training via minimal editing, which pinpoints the exact output segments that affect user choices.
  - Proposed token-level RLHF by training a token-level reward model with fine-grained supervision and incorporated it into PPO training.
  - Our method outperformed LLaMA2-chat-7B and achieved the best performance on AlpacaFarm among all 7B models.

**UCLA, Center for Vision, Cognition, Learning and Autonomy**
*Research assistant, Advisors: Prof. Ying Nian Wu and Prof. Song-chun Zhu*

- Latent-Thought Language Models
  - Proposed a novel family of language models: Latent-Thought Language Models (LTMs) – abstract tokens that *guide* the autoregressive generation of ground tokens through a Transformer decoder.
  - Dual-rate optimization framework: fast learning of local parameters for the posterior latent tokens, and slow learning of global decoder parameters.
  - Given equivalent inference budgets, LTMs demonstrate superior sample efficiency compared to conventional autoregressive models and diffusion models.
- KokoMind: A Multifaceted Evaluation Dataset of Social Interactions
  - Developed an evaluation dataset containing 150 complex multi-party social interactions with free-text questions and answers generated by GPT-4.
  - For each social interaction, we ask questions designed to assess multiple dimensions including Theory of Mind, social norms, emotion recognition, etc.

AWARDS

| | |
|---|---|
| *Doctoral Student Travel Award, UCLA* | 2019 - 2025 |
| *Graduate Summer Research Mentorship (GSRM) Award, UCLA* | Jun 2022 |
| *Cross-disciplinary Scholars in Science and Technology, UCLA* | Jun 2018 |
| *People's Daily Scholarship, BUPT* | 2017 |
| *First Prize Scholarship in BUPT (Top 1 in BUPT)* | 2016 - 2018 |

ACADEMIC SERVICES

**Peer-reviewed Conferences Reviewer**

*Conference on Neural Information Processing Systems (NeurIPS)*
*The International Conference on Learning Representations (ICLR)*
*International Conference on Machine Learning (ICML)*
*Annual Meeting of the Association for Computational Linguistics (ACL)*
*International Joint Conference on Artificial Intelligence (IJCAI)*
*International Conference on Artificial Intelligence and Statistics (AISTATS)*
*ACM Multimedia (ACM MM)*

**Journals Reviewer**

*Transactions on Machine Learning Research (TMLR)*
*IEEE Transactions on Neural Networks and Learning Systems (TNNLS)*
*IEEE Transactions on Image Processing (TIP)*
*The ISI's Journal for the Rapid Dissemination of Statistics Research (Stat)*

SKILLS

Python, PyTorch, TensorFlow, HuggingFace, Latex, C/C++
Fluent in English and Chinese