

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 Machine Learning

M.S. in Machine Learning

Advised by Prof. Ying Nian Wu, GPA: 4.0 / 4.0

Expected graduation: 2025

Sep 2019 - Jul 2021

Beijing University of Posts and Telecommunications (BUPT)

B.Eng. in Computer Science

GPA: 3.9 / 4.0; Ranking: 1 / 153

Sep 2015 - Jul 2019

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

RESEARCH EXPERIENCE **Amazon Inc - Alexa AGI Team**

Applied Scientist Intern

Jun 2023 - Oct 2023

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

- Aligning LLMs 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.

Amazon Inc - Search Team

Applied Scientist Intern

Jun 2024 - Sep 2024

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

- Improving Instruction-following Capability of Multi-modal Embedding Models
(**In submission to ICCV 2025**)
 - 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.

University of California, Los Angeles

Research assistant, Advisor: Prof. Ying Nian Wu

Research Topic: **Language Modeling, Decision-making**

- 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.
- Latent Plan Transformer
 - Planning as latent variable inference: Developed *Latent Plan Transformer*, an unsupervised solution to decision-making via sequence modeling by inferring a latent variable from a target return to guide policy execution as a plan.
- 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.

SELECTED (* denotes equal contributions)
PUBLICATIONS

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 ICCV 2025.

Deqian Kong*, Minglu Zhao*, **Dehong Xu***, Bo Pang, Shu Wang, Edouardo Honig, Zhangzhang Si, Chuan Li, Jianwen Xie, Sirui Xie, Ying Nian Wu. “*Latent Thought Models with Variational Bayes Inference Time Computation.*” International Conference on Machine Learning (ICML), 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 Lizarraaga, 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.

AWARDS	<i>Doctoral Student Travel Award, UCLA</i>	2019 - 2025
	<i>Graduate Summer Research Mentorship (GSRM) Award, UCLA</i>	Jun 2022
	<i>Cross-disciplinary Scholars in Science and Technology, UCLA</i>	Jun 2018
	<i>People’s Daily Scholarship, BUPT</i>	2017
	<i>First Prize Scholarship in BUPT (Top 1 in BUPT)</i>	2016 - 2018
ACADEMIC SERVICES	Peer-reviewed Conferences Reviewer	
	<i>Conference on Neural Information Processing Systems (NeurIPS)</i>	
	<i>The International Conference on Learning Representations (ICLR)</i>	
	<i>International Conference on Machine Learning (ICML)</i>	
	<i>Annual Meeting of the Association for Computational Linguistics (ACL)</i>	
	<i>International Conference on Computer Vision (ICCV)</i>	
	<i>International Joint Conference on Artificial Intelligence (IJCAI)</i>	
	<i>International Conference on Artificial Intelligence and Statistics (AISTATS)</i>	
	<i>ACM Multimedia (ACM MM)</i>	
	Journals Reviewer	
	<i>Transactions on Machine Learning Research (TMLR)</i>	
	<i>IEEE Transactions on Neural Networks and Learning Systems (TNNLS)</i>	
	<i>IEEE Transactions on Image Processing (TIP)</i>	
	<i>The ISI’s Journal for the Rapid Dissemination of Statistics Research (Stat)</i>	
SKILLS	Python, PyTorch, TensorFlow, HuggingFace, Latex, C/C++	
	Fluent in English and Chinese	