Haotian Deng

12313204@mail.sustech.edu.cn Southern University of Science and Technology, Shenzhen, China

Education

Southern University of Science and Technology, B.Eng. in Biomedical Engineering — Sept 2023 – Present

- GPA: 3.59/4.00
- NCC Lab Research Advisor: Prof. Quanying Liu
- IELTS: 6.5

Core Coursework

- **Machine Learning and AI:** Natural Language Processing; Machine Learning; Computer Vision.
- **Algorithms and Optimization:** Data Structures and Algorithms; Operations Research and Optimization.
- **Systems and Signals:** Digital System Design; Signals and Systems.
- **Mathematics and Statistics:** Probability and Mathematical Statistics; Advanced Mathematics; Linear Algebra.

Research Interests

- **Multimodal LLMs and VLM Safety**
- **Human-AI Alignment**
- **Cognitive Computing**
- **Human-AI Interaction**

Research Experience

VLM Safety: Alignment under Implicit Multimodal Risks — OmniAI Lab, CUHK(SZ), 2026 – Present

- Developed Intent-Privilege On-Policy Self-Distillation (OPSD), under the supervision of Prof. Ming Li, to help vision-language models address risks arising from image-text interactions while providing safe, useful responses without unnecessary refusal.
- Built evidence-grounded intent supervision and distilled a frozen teacher's guidance into a student that uses only the original image and text, requiring no extra annotations or safety modules at inference. Submitted to ICLR 2027.

Neuro-KE: EEG-Conditioned Multimodal LLM — NCC Lab, SUSTech, Oct 2025 – Present

- Developed an EEG-conditioned multimodal LLM using signal-derived question-answer data and instruction tuning to describe EEG features, predict task labels, and generate explanations grounded in signal evidence. Accepted at NeurIPS 2026.

LLM Agents and Human Moral Cognition — NCC Lab, SUSTech, Feb 2025 – Present

- Built group- and individual-level proxy agents from human behavioral and dialogue data, and used moral-dilemma experiments to study AI uncertainty, human variability, and agent-mediated behavioral modulation.

Representation Geometry and Generalization in Neural Networks — NCC Lab, SUSTech, Jan 2025 – May 2025

- Developed a representation-centric, post-hoc generalization analysis relating the generalization gap to intrinsic embedding dimension and downstream Lipschitz sensitivity, and validated the theory across architectures and datasets.

Diffusion Models for Human Perceptual Variability — NCC Lab, SUSTech

- Designed perceptual-boundary sampling with diffusion models to generate ambiguous visual stimuli, collected large-scale behavioral data, and aligned neural networks with group- and individual-level human perception. July 2024 – Jan 2025

Publications

- **Haotian Deng***, Wenbin Xing*, Gang Xu, Tao He, Jinkai Zheng, Chun Li, Zheng Zhu, Ming Li. "Can Vision-Language Models Stay Helpful When Facing Implicit Risks? Intent-Privilege OPSD for Efficient Safety–Helpfulness Alignment." Under review at **ICLR**, 2027.
- Hanrui Chen*, Xiang Chen*, **Haotian Deng***, Kexin Lou, Shinan Wang, Chen Wei, Quanying Liu. "Neuro-KE: Knowledge-Guided Interfaces for Semantically Grounded EEG Foundation Models." **NeurIPS**, 2026.
- Junjie Yu*, Zhuoli Ouyang*, **Haotian Deng***, Chen Wei, Wenxiao Ma, Jianyu Zhang, Zihan Deng, Quanying Liu. "Understanding Generalization from Embedding Dimension and Distributional Convergence." **ICML**, 2026.
- **Haotian Deng***, Sitian Wang*, Ruxin Wang*, Chen Wei, Quanying Liu. "When Proxy Agents Disagree, Do Humans Mirror? Manipulating Human Behavior in Moral Dilemmas through Agents." **AAAI**, 2026.
- Chen Wei, Chi Zhang, Jiachen Zou, **Haotian Deng**, Dietmar Heinke, Quanying Liu. "Synthesizing Images on Perceptual Boundaries of ANNs for Uncovering and Manipulating Human Perceptual Variability." **ICML**, 2025.
- **Haotian Deng***, Chi Zhang*, Chen Wei, Quanying Liu. "Synthesizing Images on Perceptual Boundaries of ANNs for Uncovering Human Perceptual Variability on Facial Expressions." **IJCNN, oral**, 2025.
- **Haotian Deng***, Sitian Wang*, Ruxin Wang*, Xuetao Wei, Chen Wei, Quanying Liu. "When LLM Agents Disagree, Do Humans Mirror? Behavioral Comparisons on Moral Dilemmas." **MIND, Best Paper**, 2025.

* *Equal contribution.*

Skills

- **Programming:** Python, PyTorch, Java, MATLAB, C, Linux/HPC.
- **Methods:** Deep learning, diffusion models, multimodal LLMs/VLMs, behavioral data analysis, on-policy distillation.
- **Reinforcement Learning:** Monte Carlo, SARSA, Q-learning, DQN, PPO, DPO, GRPO, GDPO.
- **Leadership:** Project leader, National College Students' Innovation Training Program (2025).