蔡逸超

yichao.cai@adelaide.edu.au +61 478 927 202 yichaocai.com GitHub Google Scholar

研究兴趣

表征学习理论；基础模型训练目标；多模态学习；可靠与可解释机器学习。

研究陈述

我的研究关注现代学习目标与监督信号如何塑造模型表征。我尤其关心对比学习、掩码预测（Masked Prediction）与下一词元预测（Next-Token Prediction）等训练目标，在何种条件下能够识别潜在结构，又在什么情况下会丢失、混淆或使这些结构不可辨识。理解这些问题有助于刻画基础模型训练目标的理论边界，并区分哪些能力可以通过规模化获得，哪些限制需要通过新的目标、监督形式或数据干预来突破。

研究方法上，我主要使用可辨识性理论、潜变量建模、总体目标分析与表征几何等工具。我的长期目标是发展一套关于表征学习的理论，用以解释多模态基础模型、视觉语言模型与预测式世界模型的能力边界与结构性限制。

教育背景

计算机科学博士 2023.09 – 2027.03 (预计)

阿德莱德大学, 澳大利亚机器学习研究所

导师：Prof. Javen Qinfeng Shi

计划于 *2027* 年 *1–3* 月提交博士论文

仪器科学与技术硕士 2016.09 – 2019.06

武汉理工大学

导师：周晓教授

测控技术与仪器学士 2012.09 – 2016.06

武汉理工大学

论文发表

第一作者与共同第一作者论文

[1] **Cai, Y.**; Zhang, Z.; Liu, Y.; and Shi, J. Q. "The Geometric Mechanics of Contrastive Representation Learning: Alignment Potentials, Entropic Dispersion, and Cross-Modal Divergence." *International Conference on Machine Learning (ICML)*, 2026.

[2] **Cai, Y.**[*]; Liu, Y.[*]; Gao, E.; Jiang, T.; Zhang, Z.; van den Hengel, A.; and Shi, J. Q. "On the Value of Cross-Modal Misalignment in Multimodal Representation Learning." *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. **Spotlight**.

[*]Equal contribution.

[3] **Cai, Y.**; Liu, Y.; Zhang, Z.; and Shi, J. Q. "CLAP: Isolating Content from Style Through Contrastive Learning with Augmented Prompts." *European Conference on Computer Vision (ECCV)*, 2024.

合作论文

[4] Liu, Y.; Gong, D.; **Cai, Y.**; Gao, E.; Zhang, Z.; Huang, B.; Gong, M.; van den Hengel, A.; and Shi, J. Q. "I Predict Therefore I Am: Is Next Token Prediction Enough to Learn Human-Interpretable Concepts from Data?" *International Conference on Learning Representations (ICLR)*, 2026.

[5] Jiang, W.; Liu, Y.; **Cai, Y.**; Gao, E.; Dong, J.; Abbasnejad, E.; Yao, L.; and Shi, J. Q. "What Makes a Representation Good for Single-Cell Perturbation Prediction?" *International Conference on Machine Learning (ICML)*, 2026.

[6] Chen, J.; Yin, Z.; **Cai, Y.**; Liu, Y.; Zhang, Z.; Gong, D.; and Shi, J. Q. "Boundary Embedding Shaping with

Adaptive Contrastive Learning for Graph Structural Disentanglement." *International Conference on Machine Learning (ICML)*, 2026.

代表性研究工作

跨模态错位下的语义可辨识性 2024 – 2025

NeurIPS 2025 Spotlight，第一/共同第一作者

- 建立多模态对比学习在不完整、有偏语言监督下的潜变量模型，形式化了图文监督中的语义选择偏差与文本扰动偏差。
- 证明语义因素在何种条件下会被保留、忽略或学习为不变性的可辨识性条件，并讨论其对多模态预训练与鲁棒表征学习的启示。
- 通过合成数据、MPI3D-Complex 与 OpenCLIP probing 验证理论分析。

对比表征学习的几何机制 2025 – 2026

ICML 2026，第一作者

- 建立 InfoNCE 总体目标的几何分析框架，刻画对齐势、熵驱动的分散效应与跨模态散度。
- 解释对齐、均匀性与模态差距如何由对比学习目标产生，并讨论其对多模态表征诊断与目标设计的启示。

语言监督下的内容–风格解耦表征学习 2023 – 2024

ECCV 2024，第一作者

- 提出 CLAP，通过增强语言提示构造对比监督，在表征空间中分离内容因素与风格因素，从而提升 CLIP 类视觉语言表征的跨域泛化与鲁棒性。
- 主导方法设计、实验实现、可视化分析、论文写作与代码开源。

进行中的研究

- **亚稳态数据律下的掩码预测：**研究掩码预测目标（例如 MAE、BERT、JEPA 与 Masked Diffusion Models 中使用的目标）何时能够识别联合数据分布，以及何时会在慢混合或亚稳态结构下产生模式盲性（mode-blindness）。
- **下一词元预测器中的预测状态可辨识性：**研究有限表征容量的下一词元预测器必须保留哪些预测状态信息，以及哪些潜在结构仍然无法由目标确定。
- **目标诱导的表征等价类：**建立统一框架，刻画数据、监督信号与学习目标能够恢复哪些表征结构。

教学经历

均在阿德莱德大学任教

Guest Lecturer and Head Tutor, Statistical Machine Learning — Semester 2, 2025

Teaching Assistant, Neural Networks and Deep Learning — Semester 1, 2026

Teaching Assistant, Using Machine Learning Tools — Trimester 2, 2025

Teaching Assistant, Concepts in AI and ML — Semester 1, 2025

学术服务

会议审稿： NeurIPS 2026；ICML 2026 （*Silver Reviewer Award*）；ICLR 2026。

期刊审稿： Transactions on Machine Learning Research (TMLR)。

荣誉与奖励

The D. R. Stranks Travelling Fellowship 2026.06

NeurIPS 2025 Scholar Award, Neural Information Processing Foundation 2025.10

University of Adelaide Research Scholarship 2023.09

武汉理工大学优秀毕业研究生 2019.06

其他经历

AI 工程师 2020.06 – 2022.08

泰豪软件，人工智能研究所

- 开发面向工业巡检与安全分析的计算机视觉和自动化分析系统。

软件工程师 2019.07 – 2020.04

华为技术有限公司

- 参与通信服务软件组件的开发、集成与生产环境问题定位。

访问学生研究员 2018.05 – 2018.09

加州大学伯克利分校, California PATH

- 开展自动驾驶视觉感知算法相关研究。