# 吴俊飞

junfei.wu@cripac.ia.ac.cn | (+86) 18801099385

## 教育背景

**中国科学院大学** 自动化研究所 (多模态全国重点实验室) 直博（保送） **2022.09 – 2027.06**

- **导师**：谭铁牛院士 | **研究方向**：面向可信与可解释的多模态推理研究

**北京理工大学** 计算机学院 本科 **2018.09 – 2022.06**

- **专业**：计算机科学与技术 | **学业成绩**：92.20/100，**排名**：4/359

## 科研经历

**多模态大模型 Think with Images 推理** 蚂蚁技术研究院-科研实习 **2025.01 – 至今**

- **Think with Image 赋能多模态大模型空间推理**：我们提出创新性的“Drawing to Reason in Space”范式，让 LVLMs 通过绘制辅助标注（如参考线、标记框）在视觉空间中“边画边想”，克服传统视觉转文本推理犯事的视觉信息损失问题，在 VSI-Bench 等空间推理基准上平均提升 18.4%，相关成果发表于 **NeurIPS 2025**。

**多模态大模型的幻觉检测与缓解** **2023.11 – 2025.05**

- **基于表示干预的 MLLM 幻觉缓解**：发现幻觉与非幻觉样本在模型内部表示中具有可分性，且不同幻觉诱因（如语言先验偏差、提示-视觉信息冲突）对应可区分的表示模式。因此，构建多诱因干预向量并通过线性组合联合调控内部表示，在无需修改模型结构或重新训练的前提下有效抑制幻觉，相关成果发表于 **EMNLP 2025**。
- **基于多模态大模型 (MLLM) 逻辑一致性的幻觉检测与缓解**：首次提出基于 MLLM 行为逻辑闭环性的物体幻觉检测与缓解框架，通过对大模型进行一系列逻辑拷问，判断其回答是否能够形成逻辑闭环，从而判断是否在幻觉问题。该框架从逻辑一致性的角度挖掘模型内在幻觉，以期推动 LVLM 的可信可靠应用，相关成果发表于 **ACL 2024 Findings**。

**虚假信息检测模型的性能增强与去偏** **2021.10 – 2023.10**

- **基于图神经网络和对抗对比训练的虚假信息检测模型**：由于互联网上的文本证据内容较长且存在大量冗余噪声，我们提出采用图结构建模新闻和证据文本以捕捉长距离依赖，通过图结构学习消除冗余节点信息。进一步结合对抗对比训练以增强表示学习质量。相关成果发表于 **WWW 2022** 和 **TKDE 2024**。
- **基于因果干预和双重对抗训练的虚假信息检测模型去偏**：针对偏置问题导致模型难以泛化至分布外真实场景的挑战，我们提出了一种基于因果干预的方法，以避免模型依赖新闻中的虚假关联信息。此外，我们引入双重对抗训练策略，旨在消除新闻侧和证据侧偏置的影响，从而全面提升模型的推理能力。相关成果发表于 **SIGIR 2022** 和 **TKDE 2024**。

## 论文成果

### 以第一作者、共一身份发表论文 8 篇

- **Junfei Wu**, Jian Guan, Kaituo Feng, Qiang Liu, Shu Wu, Liang Wang, Wei Wu, Tieniu Tan. Reinforcing spatial reasoning in vision-language models with interwoven thinking and visual drawing[C]. Conference on Neural Information Processing Systems, 2025. (**CCF-A 会议**)
- **Junfei Wu**, Ding Yue, Guofan Liu, Tianze Xia, Ziyue Huang, Dianbo Sui,Qiang Liu, Shu Wu, Liang Wang, Tieniu Tan. SHARP: Steering Hallucination in LVLMs via Representation Engineering[C]. Conference on Empirical Methods in Natural Language Processing, 2025.(**CCF-B 会议**)
- **Junfei Wu**, Qiang Liu, Ding Wang, Jinghao Zhang, Shu Wu, Liang Wang, Tieniu Tan. Logical Closed Loop: Uncovering Object Hallucinations in Large Vision-Language Models[C]. Annual Meeting of the Association for Computational Linguistics Findings, 2024. (**CCF-A 会议**)
- **Junfei Wu**, Weizhi Xu, Qiang Liu, Shu Wu, Liang Wang. Adversarial contrastive learning for evidence-aware fake news detection with graph neural networks[J]. IEEE Transactions on Knowledge and Data Engineering, 2024. (**CCF-A 期刊**)
- Qiang Liu, **Junfei Wu**, Shu Wu, Liang Wang. Out-of-distribution Evidence-aware Fake News Detection via Dual Adversarial Debiasing[J]. IEEE Transactions on Knowledge and Data Engineering, 2024. (**CCF-A 期刊，学生一作**)
- **Junfei Wu**, Qiang Liu, Weizhi Xu, Shu Wu. Bias mitigation for evidence-aware fake news detection by causal intervention[C]. Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2022. (**CCF-A 会议**)
- Jian Guan*, **Junfei Wu***, Jia-Nan Li*, Chuanqi Cheng*, Wei Wu. A Survey on Personalized Alignment–The Missing Piece for Large Language Models in Real-World Applications[C]. Annual Meeting of the Association for Computational Linguistics Findings, 2025. (**CCF-A 会议**)
- Weizhi Xu*, **Junfei Wu***, Qiang Liu, Shu Wu, Liang Wang. Evidence-aware fake news detection with graph neural networks[C]. Proceedings of the ACM Web Conference 2022. (**CCF-A 会议**)

### 以共同作者身份参与发表论文 5 篇

- Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, **Junfei Wu**, Xiaoying Zhang, Benyou Wang, Xiangyu Yue. Video-r1: Reinforcing video reasoning in mllms[C]. Conference on Neural Information Processing Systems, 2025. (**CCF-A 会议**)

- Yi-Fan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, **Junfei Wu**, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, Liang Wang, Rong Jin, Tieniu Tan. Mme-realworld: Could your multimodal llm challenge high-resolution real-world scenarios that are difficult for humans?[C]. The Thirteenth International Conference on Learning Representations, 2025.
- Haisong Gong, Jing Li, **Junfei Wu**, Qiang Liu, Shu Wu, Liang Wang. STRIVE: Structured Reasoning for Self-Improvement in Claim Verification[J]. Machine Intelligence Research, 2025.
- Guofan Liu, Jinghao Zhang, Qiang Liu, **Junfei Wu**, Shu Wu, Liang Wang. Uni-Modal Event-Agnostic Knowledge Distillation for Multimodal Fake News Detection[J]. IEEE Transactions on Knowledge and Data Engineering, 2024. (**CCF-A 期刊**)
- Qiang Liu, Xiang Tao, **Junfei Wu**, Shu Wu, Liang Wang. Can Large Language Models Detect Rumors on Social Media?[J]. arXiv preprint arXiv:2402.03916, 2024.

## 获奖情况

- 博士国家奖学金 2025.10
- 中国科学院大学三好学生 2024.06
- 中国科学院大学研究生学业奖学金 2022, 2023，2024, 2025
- 第 5 届 IKCEST 国际大数据竞赛（多模态虚假信息检测），二等奖 2023.11
- 2020 ICPC 国际大学生程序设计竞赛沈阳赛区，铜奖 2021.07
- 中国高校计算机大赛团队设计天梯赛全国总决赛，银奖 2021.04
- 第十一届蓝桥杯全国总决赛，三等奖 2020.11
- 北京市优秀毕业生 2022.06
- 北京理工大学优秀学生（标兵） 2019，2020，2021
- 解放领航奖学金 2021.11
- 华瑞世纪奖学金 2020.11
- 本科生国家奖学金 2019.11

## 自我评价

- **编程能力**：掌握 Python, C/C++ 等编程语言，熟悉 PyTorch 深度学习框架，以及 LLaMA-Factory、VERL 等大模型训练框架。
- **专业能力**：熟悉大语言模型、多模态大模型、不确定性估计等相关算法，当前学习大模型可解释性相关工作。
- **英语能力**：通过 CET-4(620) 和 CET-6(577)，日常阅读英文专业文献，具备良好的英语读写能力。
- **综合能力**：学习能力强，具备较强的动手能力和实践能力，具备较强的自我管理能力，高效完成学习和工作任务。