雷宇 (Yu Lei)

+86-135-6890-2648 leiyu2648@gmail.com torpedo2003.github.io github.com/torpedo2003

研究兴趣

当前研究方向：个性化与安全的 AI 生成 • 生成模型的偏好优化 • 世界模型与交互式生成
过往研究方向 (2022-2024)：医学图像分割 • 弱监督学习

教育背景

计算机科学与技术博士研究生 2025 年 9 月 – 至今
中国科学院计算技术研究所 北京
导师：李亮研究员
研究方向：人类对齐的生成式 AI、偏好学习、世界模型

人工智能工学学士 2021 年 9 月 – 2025 年 6 月
四川大学计算机学院 成都
导师：王利团教授、张蕾教授
荣誉：国家奖学金、宝钢奖学金（本科生前 6 名）、优秀毕业生

科研经历

研究实习生 2025 年 2 月 – 2025 年 8 月
中国电信人工智能研究院 *(TeleAI)* 北京
导师：白晋斌博士、喻凯东博士

- **研究方向：**文生图扩散模型的个性化安全对齐
- **主要贡献：**开发 PSAlign 框架，通过交叉注意力适配器实现用户特定的安全控制；构建包含 1,000 个虚拟用户的 Sage 数据集，涵盖多样化安全偏好
- **论文成果：**第一作者论文投稿至 TMLR 2025；UPO 论文（CVPR 2026 投稿）共同作者

研究助理 2022 年 9 月 – 2024 年 8 月
四川大学机器智能实验室 成都
导师：王利团教授、张蕾教授

- **研究方向：**基于涂鸦标注的弱监督医学图像分割
- **主要工作：**开发 PCLMix 方法，结合不确定性引导的像素级对比学习
- **论文成果：**第一作者论文发表于 ICIC 2024，已被引用 6 次

论文发表

审稿中

[1] **雷宇**, 白晋斌, 史清宇, 冯傲松, 喻凯东. “Personalized Safety Alignment for Text-to-Image Diffusion Models（文生图扩散模型的个性化安全对齐）.” 投稿至 *TMLR*, 2025. [arXiv] [代码] [项目]

[2] 白晋斌, 冯傲松, **雷宇**, 赵卓然, 喻凯东. “Unpaired Preference Optimization: Aligning Visual Generative Models with Scalar Feedback（非配对偏好优化：使用标量反馈对齐视觉生成模型）.” 投稿至 *CVPR*, 2026. [论文]

已发表

[3] **雷宇**, 罗浩伦, 王利团, 张振威, 张蕾. “PCLMix: Weakly Supervised Medical Image Segmentation via Pixel-Level Contrastive Learning and Dynamic Mix Augmentation（PCLMix：通过像素级对比学习和动态混合增强的弱监督医学图像分割）.” 国际智能计算会议 *(ICIC)*, 2024. [arXiv] [代码] [Springer]

预印本与综述论文

[4] 白晋斌, **雷宇**, 吴鹤聪, 朱宇辰, 李淑凡, 辛毅, 李湘泰, 陶墨磊, Aditya Grover, 杨明轩. “From Masks to Worlds: A Hitchhiker’s Guide to World Models（从掩码到世界：世界模型漫游指南）.” *arXiv* 预印本, 2025. [arXiv] [GitHub]

精选开源项目

PSAlign • M-E-AGI-Lab/PSAlign • Python, 8 stars, 1 fork
文生图扩散模型的个性化安全对齐框架，配有完整文档。

PCLMix • torpedo2003/PCLMix • Python, 8 stars
基于不确定性引导对比学习的弱监督医学图像分割框架。

Awesome World Models • M-E-AGI-Lab/Awesome-World-Models • Markdown, 58 stars
世界模型研究精选集，配套综述论文。

荣誉奖项

优秀毕业生 • 四川大学 2024–2025
国家奖学金 • 中华人民共和国教育部 2023–2024
宝钢奖学金（全校仅本科生 6 名） • 四川大学 2023–2024
全球总决赛一等奖 • 华为 ICT 大赛 2023–2024

学术服务

会议评审： International Joint Conference on Neural Networks (IJCNN) 2024）
开源贡献： 活跃维护研究代码库，项目总星标数 70+

技能专长

编程语言： Python (PyTorch, TensorFlow, Hugging Face), C++, MATLAB, JavaScript
研究工具： Git, Docker, Linux, LaTeX, Weights & Biases, Gradio
专业方向： 扩散模型、对比学习、偏好优化、医学图像处理