

杨童耘

Tony

+86-15692192736 | +31-0645770356

tonyyunyang@outlook.com

阿姆斯特丹(现居 时差晚-6h) | 上海 | 深圳

tonyyunyang.github.io

研究领域

我的研究兴趣围绕人工智能系统如何理解世界、参与世界。具体方向包括世界模型、大语言模型、计算机视觉、以及医疗 AI 等。我关注模型能力如何被放入真实环境中，与人的需求、社会结构和具体任务发生连接。我尤其关注两个问题。**第一**，AI 如何从语言和文本中的智能，进一步走向对现实世界的感知、建模和行动。**第二**，AI 如何成为扩展人类能力的工具，并使这种能力以更公平、可及的方式影响社会。

工作经历

2026年3月 — 至今

独立研究与产学研合作

独立研究员

- 与工业界和学术界团队合作，包括 Tencent（腾讯）、Gradient Network、MeetaVista（米塔视界）等公司，以及 McGill（加拿大麦吉尔大学）、清华大学等学术机构。研究内容涵盖低成本高效 LLM 及其面向复杂任务的路由系统、面向组合优化的 LLM 推理框架、智能体评测与运行框架、世界模型等方向。

2025年7月 — 2026年5月

IMDEA Networks | 马德里高等研究院

玛丽居里博士学者

- 研究隐私保护型无线感知系统，探索如何将无线信号转化为可用于人和障碍物检测的有效信息。研究重点包括无线信号的表征学习，以及面向无线信号的世界模型，以支持特征提取和动态预测推理。

2025年1月 — 2025年7月

TU Delft | 荷兰代尔夫特理工大学

医疗图像AI研究员

- 研究 nnU-Net 中的权重冗余问题，先后采用非结构化剪枝和结构化剪枝方法提升模型效率。实验表明，nnU-Net 可以在结构化剪枝后达到 99% 稀疏度，同时仅产生很小的性能损失。该方法带来了约 6 倍训练效率提升，并提高了推理速度，相关成果已被 MIDL 2025 接收发表。

2023年6月 — 2024年10月

TU Delft | 荷兰代尔夫特理工大学

情感计算AI研究助理

- 构建了首个仅基于沉浸式环境中眼动数据的大规模公开情绪识别数据集，覆盖七类离散情绪。同时开发了高效、可扩展的识别方法。相关成果已被 IMWUT / UbiComp 2025 接收发表。

2020年12月 — 2021年7月

NXP | 恩智浦半导体

芯片技术支持工程师

- 开发用于 CAN 芯片的自动化测试平台，集成完整的诊断流程，用于提升缺陷分析效率，并支持芯片量产后的稳定性验证。

论文发表

- Yang, T.**, Regmi, B., Du, L., Bulling, A., Zhang, X., & Lan, G. (2025). Through the Eyes of Emotion: A Multi-faceted Eye Tracking Dataset for Emotion Recognition in Virtual Reality. In Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (UbiComp/ISWC, IMWUT), 9(3), 1-41. ACM, New York, NY, USA.
- Yang, T.**, Zhao, Y., & Tao, Q. (2025). Pruning nnU-Net with Minimal Performance Loss. In Medical Imaging with Deep Learning (MIDL).
- Zhao, Y., Kellman, P., Xue, H., **Yang, T.**, Zhang, Y., Han, Y., Simonetti, O., & Tao, Q. (2025). Reverse Imaging for Wide-spectrum Generalization of Cardiac MRI Segmentation. In International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), 15962, 555-565. Springer Nature Switzerland.

项目经历

2026年1月

Diffusion Weights | 扩散权重预测

- 探索将神经网络微调过程重新表述为生成式参数预测问题，而非传统的逐步 SGD 优化过程。使用 U-Net 扩散模型，根据早期训练检查点预测模型未来的参数状态，并将短期、中期和长期预测结果融合生成最终模型。在 MMLU、OpenBookQA 和 WMT16 等基准任务上的实验表明，该方法可实现约 3.2 倍的端到端训练时间加速，同时保持与完整微调相近的下游性能。该项目成果已投稿至 ACL 2026。

2026年2月

LLM Router | 复杂任务的低成本大语言模型路由系统

- 与 Gradient Networks 合作构建 LLM 路由评测基准，用于评估不同 LLM 路由策略的有效性。项目重点考察复杂任务中的简单步骤是否可以交由成本更低的模型完成，同时不牺牲整体准确率。实验覆盖了 SWE-bench、BFCL 等近期常用基准。结果显示，与在所有步骤中都使用单一 SOTA 模型相比，合理的路由策略可以在保持相同性能的同时，将成本降低超过 90%。该项目成果已投稿至 NeurIPS 2026，目前正在审稿阶段。
- 项目链接：<https://github.com/CommonstackAI/CommonRouterBench>

2026年3月

Cost-Adaptive LLM Routing | 成本自适应的大语言模型路由

- 在 LLM Router 项目的基础上，进一步探索如何通过专用小模型降低推理成本。项目主要面向重复出现的日常工作流：这些任务在结构上相似，但具体上下文不同。我们收集强模型生成的轨迹，并用这些数据微调较小模型，使其适应特定场景。项目的核心思路是：随着系统使用次数增加，可用于训练的数据不断积累，小模型能力逐步提升，系统整体使用成本也随之下降。该项目成果正在准备投稿至 EMNLP 2026。

2026年3月

Human Intent World Model | 意图理解的视觉语言世界模型

- 与 MeetaVista 合作，探索如何提升 AI 销售场景中的客户体验。我们基于经典销售学文献中的知识，构建了一个用于建模客户意图的合成数据集。数据包括视觉与行为线索、意图标签，以及与销售知识相连接的推理过程。随后，我们基于该数据微调视觉语言模型，使其能够从视觉线索中推断客户意图，并结合相关销售逻辑做出回应。模型部署后，在真实交互任务中取得了良好表现。该项目成果正在准备投稿至顶级 AI 会议。

2026年4月

LLM for Optimization | 组合优化的大语言模型推理框架

- 与 Tencent 合作开发一个将 LLM 应用于优化问题的框架，例如旅行商问题等。该框架通过 optimization harness 引导模型进行结构化搜索，并结合反馈与记忆检索，使模型能够更高效地接近最优解，而不是仅依赖直接生成答案。项目目前仍在进行中，后续计划投稿至 AAAI 2026。

教育经历

2022年9月-2024年10月

TU Delft | 荷兰代尔夫特理工大学

计算机工程 | 硕士

2017年9月-2021年7月

上海海事大学 (1.5年荷兰海外交换)

电气工程及其自动化 | 学士

教学经历

2023年9月 — 2025年1月

TU Delft | 荷兰代尔夫特理工大学

助教与学生导师

- ET 4310 Supercomputing for Big Data, 助教, 2024/2025 Q1
- CESE 4030 Embedded Systems Lab, 助教, 2023/2024 Q3
- CESE 4000 Software Fundamentals, 助教, 2023/2024 Q1
- CESE 4010 Advanced Computing Systems, 助教, 2023/2024 Q1
- CESE 硕士项目学生导师, 2023/2024

专业技能

AI-Native 工具: Claude Code, Codex, 及其他开源框架如 Openclaw, pi, Hermer-Agent 等

编程语言: Python, C/C++, Rust, Bash

机器学习框架: PyTorch, HuggingFace, JAX

开发工具: Git, Docker, Linux, LaTeX

语言水平

中文: 母语

英文: 专业学术/工作流利 | 海外6年 | 雅思8.0

兴趣爱好

跑步, 半程马拉松 PB 1:43:53

健身 | 做饭 | 徒步 | 网球