

傅天予

学术经历

本科 Undergraduate Study	2018.8 – 2022.6
获评清华大学优秀毕业论文，清华大学优良毕业生	清华大学电子工程系
平均学分绩：3.88/4.00；学分绩排名：25/259；保研综合排名：5/87	
博士研究生 Ph.D. Student	2022.8 – 至今
直博生，清华大学 NICS-EFC 实验室；导师汪玉老师	清华大学电子工程系
主要研究方向为大模型长序列场景下的高效推理和训练算法	
海外经历 Visiting Experience	2025.10 – 2026.4
访问学者，CMU InfiniAI Lab；导师 Beidi Chen 老师	卡内基梅隆大学电子与计算机工程系
主要研究方向为大模型 Agent 在长序列场景下的行为与优化算法	

实习经历

无问芯穹 - 实习生	2023.6 – 2025.9
研究大模型高效推理和训练算法，包括稀疏注意力、多 Agent 协作等	大语言模型算法研究
Cache-to-Cache 入选无问芯穹 2025 年代表成果；MoA 在无问-联想端侧推理加速项目落地	
字节跳动 Seed - 实习生	2026.5 – 至今
研究新一代 Multi-Token Prediction 范式和模型架构	基础大语言模型 - 模型架构组
将 MTP 分支的额外计算同时用于提升推理效率和主模型能力，贯穿预训练和后训练流程	
相比站内现有实现，在预训练阶段使主模型能力进一步提升，在后训练阶段使投机解码接受长度明显增长	

代表学术研究

LLM 推理加速算法：token 级路由 Website	Roads-to-Roam (R2R), NeurIPS'25
- 训练时，构建 token 难度分类方法并训练路由网络；推理时，不同 token 路由给不同大小的模型 - 推理时混合路由 R1-1.5B 与 32B，平均激活 5.6B；其速度和精度均超过 R1-14B；较 R1-32B 加速 2.8× - 工作获 Qwen、Seed 等团队关注；NeurIPS 审稿人全正面评价；算法已全面适配 SGLang 推理系统	
LLM 新 Scaling 维度：动态深度 looped transformer Website	Think-at-Hard (TaH), ICML'26
- 对预训练 LLM 进行 SFT，构建动态 looped LLM；实现 token 级动态算力分配与 test-time scaling - 提出双因果注意力架构，支持跨迭代的信息流动并保持训练可完全并行；在 0.6B-8B 模型上训练验证 - 仅对 7% token 迭代，即可在九个基准测试上提升 5.3–6.2%；入选 ICLR'26 LIT Workshop 最佳论文提名	
Agent 协作算法：KV-Cache 通信 Website	Cache-to-Cache (C2C), ICLR'26
- 以 KV-Cache 替代文本，实现不同 LLM 间更准确、更高效的通信，支持跨模型家族、大小、代际和训练阶段 - 相比单 LLM 提点 8.5-10.5%；相比文本沟通提点 3-5%，平均加速 2.0× - 获 420+ GitHub Stars；Hugging Face 登顶 Daily Papers；Twitter 862K 浏览、马斯克评论	

主要奖项

综合优秀奖：惠妍英才奖学金	清华大学	2025.10
一二·九辅导员奖	清华大学	2024.12
综合优秀奖：清华之友一子立奖学金	清华大学	2024.10
综合优秀奖：清华之友一子立奖学金	清华大学	2023.10
优秀毕业论文	清华大学	2022.06
优良毕业生	清华大学	2022.06
综合优秀奖：清华之友一中国光谷奖学金	清华大学	2021.12
电子之星一班团之星	清华大学电子工程系	2021.12
国家奖学金	中华人民共和国教育部	2020.12
综合优秀奖	清华大学	2020.12
科技创新优秀奖	清华大学	2020.12
智能机器人设计实践比赛冠军	清华大学智能机器人课组	2020.12
社会实践优秀奖	清华大学	2019.12
文艺优秀奖	清华大学	2019.12
第三十六届全国部分地区大学生物理竞赛非物理类A组一等奖	北京物理学会	2019.12
智能无人机挑战赛三等奖	清华大学学生科协	2019.12

🏠 主要论文

For the full list of publications, please refer to [this page](#).

[ICML'26 & ICLR'26 LIT Workshop Best Paper Shortlist] Tianyu Fu*, Yichen You*, Zekai Chen, Guohao Dai, Huazhong Yang, Yu Wang, "Think-at-Hard: Selective Latent Iterations to Improve Reasoning Language Models," International Conference on Machine Learning, 2026. [📄 Paper](#) [🔗 Code](#)

[ICLR'26] Tianyu Fu*, Zihan Min*, Hanling Zhang*, Jichao Yan, Guohao Dai, Wanli Ouyang, Yu Wang, "Cache-to-Cache: Direct Semantic Communication Between Large Language Models," International Conference on Learning Representations, 2026. [📄 Paper](#) [🔗 Code](#)

[NeurIPS'25] Tianyu Fu*, Yi Ge*, Yichen You, Enshu Liu, Zhihang Yuan, Guohao Dai, Shengen Yan, Huazhong Yang, Yu Wang, "R2R: Efficiently Navigating Divergent Reasoning Paths with Small-Large Model Token Routing," Annual Conference on Neural Information Processing Systems, 2025. [📄 Paper](#) [🔗 Code](#)

[ICCV'25] Tianyu Fu*, Tengxuan Liu*, Qinghao Han*, Guohao Dai, Shengen Yan, Huazhong Yang, Xuefei Ning, Yu Wang, "FrameFusion: Combining Similarity and Importance for Video Token Reduction on Large Visual Language Models," IEEE/CVF International Conference on Computer Vision, 2025. [📄 Paper](#) [🔗 Code](#)

[CoLM'25] Tianyu Fu*, Haofeng Huang*, Xuefei Ning*, Genghan Zhang, Boju Chen, Tianqi Wu, Hongyi Wang, Zixiao Huang, Shiyao Li, Shengen Yan, Guohao Dai, Huazhong Yang, Yu Wang, "MoA: Mixture of Sparse Attention for Automatic Large Language Model Compression," Conference on Language Modeling, 2025. [📄 Paper](#) [🔗 Code](#)

[WSDM'24 Oral] Tianyu Fu, Chiyue Wei, Yu Wang, Rex Ying, "DeSCo: Towards Generalizable and Scalable Deep Subgraph Counting," ACM International Conference on Web Search and Data Mining, 2024. [📄 Paper](#) [🔗 Code](#)

[arXiv'24] Zixuan Zhou*, Xuefei Ning*+, Ke Hong*, Tianyu Fu, Jiaming Xu, Shiyao Li, Yuming Lou, Luning Wang, Zhihang Yuan, Xiuhong Li, Shengen Yan, Guohao Dai, Xiao-Ping Zhang, Yuhan Dong, Yu Wang+, "A Survey on Efficient Inference for Large Language Models," 2024. [📄 arXiv](#)

[DATE'23] Tianyu Fu*, Chiyue Wei*, Zhenhua Zhu, Shang Yang, Zhongming Yu, Guohao Dai, Huazhong Yang, Yu Wang, "CLAP: Locality Aware and Parallel Triangle Counting with Content Addressable Memory," IEEE Design, Automation and Test in Europe Conference, 2023. [📄 Paper](#) [🔗 Code](#)

[DAC'23] Xinhao Yang, Tianyu Fu, Guohao Dai, Shulin Zeng, Kai Zhong, Ke Hong, Yu Wang, "An Efficient Accelerator for Point-based and Voxel-based Point Cloud Neural Networks," IEEE/ACM Design Automation Conference, 2023. [📄 Paper](#)

[ISCA'22] Guohao Dai*, Zhenhua Zhu*, Tianyu Fu, Chiyue Wei, Banyan Wang, Xiangyu Li, Yuan Xie, Huazhong Yang, Yu Wang, "DIMMining: Pruning-Efficient and Parallel Graph Mining on DIMM-based Near-Memory Computing," ACM International Symposium on Computer Architecture, 2022. [📄 Paper](#)

[HPEC'20] Tianyu Fu, Ziqian Wan, Guohao Dai, Yu Wang, Huazhong Yang, LessMine: Reducing Sample Space and Data Access for Dense Pattern Mining, in 2020 IEEE High Performance Extreme Computing Conference [📄 Paper](#)

🏠 社会工作

本科-无85班长	2020.8 – 2022.7
任职期间, 班级获北京市先进班集体; 清华大学先进班集体、优良学风班	清华大学电子工程系
研究生-带班辅导员	2022.8 – 2024.7
无零年级辅导员联系人, 无07、08班带班辅导员; 获一二·九辅导员奖	清华大学电子工程系

🌐 外语水平

托福 TOEFL iBT: 102/120; 国家英语六级 CET-6: 621/710

📞 联系方式

邮箱: fty22@mails.tsinghua.edu.cn; 手机: 18256010328; 个人主页: fuvty.thesimple.ink