刘天宇 Tianyu Liu

17755201083 | tianyu_liu@mail.ustc.edu.cn | 个人主页 | Scholar | GitHub

Speculative Decoding · LLM Serving · 推理优化

教育背景

中国科学技术大学（上海 AI Lab 联培博士） 信息与通信工程 博士 2022.09 — 2027.06

中央财经大学 — 计算机科学与技术 本科 专业排名 1 / 35 2018.09 — 2022.06

博士导师：孙骁、孙晓艳
荣誉奖项：华为奖学金 (2023)、国家奖学金 (2020–2021)、国家奖学金 (2019–2020)、北京市优秀毕业生 等
技能栈：Python（熟练）/ C++ / Java | PyTorch | Linux

实习经历

腾讯混元 · AngelSlim 团队 — 研究型实习（青云计划） 2026.05 — 至今

主导 **D-Cut**：面向高并发 serving 的 verification-depth 动态剪枝，已落地开源 AngelSlim。

阿里巴巴 · 通义千问 C 端事业群 — 研究型实习（人才计划） 2026.01 — 2026.04

主攻 **KVShot**：复用 KV Cache 缓解长距衰减，含 draft-model 训练与 serving 瓶颈分析。

Technical Reports

D-cut: Adaptive Verification Depth Pruning for Batched Speculative Decoding

1. 硬件 cost model 驱动的 **跨请求 verification-depth 动态剪枝**：用 prefix-product score 估计各位置边际收益，在 batch 内做全局预算分配；不改 target、免训练、全程 lossless，面向高并发 serving。
2. H20：dense 超 EAGLE-3、MoE 超 MTP（**2.83×** vs 2.00×），半预算保留约 95% 收益；作为 AngelSpec runtime 组件集成至 AngelSlim；**D-cut 被 DeepSeek DSpark 引用**，后者采用同类 load-aware 剪枝。

When Hidden States Drift: Can KV Caches Rescue Long-Range Speculative Decoding?

1. 提出 **KV-Reuse Hypothesis**：hidden states 是 query 相关的压缩表示，会抑制当前步无用、后续步有用的信息；复用 target KV Cache 能保留 token 级前缀表示，缓解长距 acceptance 衰减。
2. Qwen3-8B 上系统对比 hidden / KV / hybrid 三范式，并定位 query 估计、KV 投影梯度稀疏、gate 梯度饥饿三大训练瓶颈；重新带动复用 KV 浪潮，Gemma 4 MTP、GLM 5.2 的 KV / index share 均印证此思路。

Selected Publications

PEARL & nano-PEARL — 并行投机解码 & DT 分离

ICLR 2025 · 一作 · arXiv · GitHub ★171

1. **PEARL**（首篇 parallel speculative decoding）：识别 draft–verify 互相等待，用 pre-verify / post-verify 让二者并行、实现自适应 draft length 且无损；相比自回归最高 **4.43×**、相比 vanilla SD **1.50×**。
2. **nano-PEARL**（首提 **DT 分离**）：nano-vLLM 风格全栈引擎，打通 TP / CUDA Graph / PagedAttention 等系统优化；70B、bs=32、3×H200 达 **3546.72 tok/s**（vanilla SD **3.06×**，无损）。 · GitHub ★213

TALON — Confidence-Aware Speculative Decoding with Adaptive Token Trees

Preprint · 一作 · arXiv

1. 为 EAGLE-3 / MTP 等 AR drafter 设计的置信度门控 **动态 token tree**：固定 budget 下用 robust top-K 初始化 + confidence-gated 扩展，难时 shallow-and-wide、易时 deep-and-narrow。
2. 覆盖 5 个 backbone × 6 数据集，wall-time 稳定超 EAGLE-3，最高 **5.16×** over AR；即使 MAT 持平，也能靠减少无效 draft 计算实现提速。

LogitSpec — Retrieval-based SD via Next-Next Token Speculation

ACL 2026 Findings · 一作 · arXiv

1. 用 target 末位 logit 投机 **next-next token** 作为更强检索 anchor（比仅用上一个 token 更具区分度），training-free、**无需额外 draft model** 即可加速。
2. Spec-Bench / HumanEval 上检索成功率约 97–99%，远高于 vanilla 检索；最高 **2.61×** 加速、平均接受长度 3.28，长上下文场景收益更明显。

竞赛获奖

天池大赛 · CCKS 2023 — 归纳式知识图谱关系推理，冠军（1 / 401） 2023.04 — 2023.07

天池大赛 · CAAI-BDSC 2023 — 社交图谱小样本场景链接预测，亚军（2 / 739） 2023.04 — 2023.06

中国科学技术大学第二届量化交易研究大赛 — 第二名 2023.07 — 2023.08

美国大学生数据建模竞赛 (MCM/ICM) — Meritorious Winner 2021

全国大学生数学建模竞赛 — 北京市一等奖 2021

Publication List

一作 / 共一 9 篇；* 共同一作，† 通讯作者；完整列表与链接见 Google Scholar。

[1] **Tianyu Liu**, Yuhao Shen, Rui Cen, Junhan Shi, Jiebin Zhang, Guangshuo Qin, Hong Liu, Song Liu, Guanghua Yu†, Jianchen Zhu. "D-cut: Adaptive Verification Depth Pruning for Batched Speculative Decoding" · Technical Report · arXiv · docs · vLLM PR

[2] Hong Liu, Rui Cen, Junhan Shi, Guangshuo Qin, Jiebin Zhang, **Tianyu Liu**, Runzhi Fan, Guoliang Zhao, Ruobing Xie, Kai Zhang, Song Liu, Guanghua Yu†, Jianchen Zhu. "AngelSpec: Towards Real-World High Performance Inference with Speculative Decoding" · Technical Report · arXiv · code

[3] **Tianyu Liu**, Yun Li, Qitan Lv, Kai Liu, Jianchen Zhu, Winston Hu, Xiao Sun. "Parallel Speculative Decoding with Adaptive Draft Length" · ICLR 2025 · arXiv · code

[4] **Tianyu Liu**, Yuhao Shen, Xinyi Hu, Baolin Zhang, Hengxin Zhang, Jun Dai, Jun Zhang, Shuang Ge, Lei Chen, Yue Li, Mingcheng Wan. "When Hidden States Drift: Can KV Caches Rescue Long-Range Speculative Decoding?" · Preprint · arXiv

[5] **Tianyu Liu***, Qitan Lv*, Hao Li, Xing Gao, Xiao Sun, Xiaoyan Sun. "LogitSpec: Accelerating Retrieval-based Speculative Decoding via Next-Next Token Speculation" · ACL 2026 Findings · arXiv · code

[6] **Tianyu Liu***, Qitan Lv*, Yuhao Shen, Xiao Sun, Xiaoyan Sun. "TALON: Confidence-Aware Speculative Decoding with Adaptive Token Trees" · Preprint · arXiv

[7] **Tianyu Liu**, Qitan Lv, Jie Wang, Shuling Yang, Hanzhu Chen. "Learning Rule-Induced Subgraph Representations for Inductive Relation Prediction" · NeurIPS 2023 · arXiv · code

[8] Qitan Lv*, **Tianyu Liu***, Wen Wu, Xuenan Xu, Bowen Zhou, Feng Wu, Chao Zhang. "HIPPO: Accelerating Video LLM Inference via Holistic-aware Parallel Speculative Decoding" · Preprint · arXiv

[9] Qitan Lv*, **Tianyu Liu***, Hong Wang. "Exploiting Edited Large Language Models as General Scientific Optimizers" · NAACL 2025 · arXiv

[10] Qitan Lv*, **Tianyu Liu***, Qiaosheng Zhang, Xingcheng Xu, Chaochao Lu. "KALE: Enhancing Knowledge Manipulation in LLMs via Knowledge-aware Learning" · Preprint · arXiv

[11] Yuhao Shen, **Tianyu Liu**, Junyi Shen, Jinyang Wu, Quan Kong, Li Huan, Cong Wang. "Double: Breaking the Acceleration Limit via Double Retrieval Speculative Parallelism" · ACL 2026, Oral · arXiv

[12] Yuhao Shen, Junyi Shen, Quan Kong, **Tianyu Liu**, Yao Lu, Cong Wang. "Speculative Decoding via Hybrid Drafting and Rollback-Aware Branch Parallelism" · ICLR 2026 · arXiv

[13] Yuhao Shen, **Tianyu Liu**, Xinyi Hu, Quan Kong, Baolin Zhang, Jun Dai, Jun Zhang†, Shuang Ge, Lei Chen, Yue Li, Mingcheng Wan, Cong Wang†. "Draft Less, Retrieve More: Hybrid Tree Construction for Speculative Decoding" · Preprint · arXiv