InfoQ 极客传媒

大模型推理系统与压缩优化：从算法到工具的工程实践

演讲人：

龚睿昊 ModelTC开源社区

AiCon

全球人工智能开发与应用大会

AiCon
全球人工智能开发与应用大会

目录

01 背景介绍 多模态发展趋势

02 概览 统一框架工具

03 LightLLM 语言和多模态理解

04 LightX2V 图片和视频生成

05 LightCompress 统一压缩工具

06

极客邦科技 2026 年会议规划

促进软件开发及相关领域知识与创新的传播

参会咨询

查看会议

4月 · 北京 · 1200人

QCon 全球软件开发大会

会议时间：4月16-18日

- Agentic Engineering
- AgentOps
- 下一代模型架构与推理优化
- AI 原生基础设施
- 知识工程实践
- AI 安全

6月 · 上海 · 1000人

AiCon 全球人工智能开发与应用大会

会议时间：6月26-27日

- AI Infra 系统工程
- 多 Agent 协作与实践
- 多模态融合
- 模型训练与推理创新
- 数据平台与特征服务

8月 · 深圳 · 1000人

AiCon 全球人工智能开发与应用大会

会议时间：8月21-22日

- Agentic AI
- 轻量化与高效推理
- 多模态应用
- AI + IoT 场景实践
- AI 工业化落地

10月 · 上海 · 1200人

QCon 全球软件开发大会

会议时间：10月22-24日

- AI Agent
- Vibe Coding
- 智能可观测
- 推理基建
- 模型攻防
- AI x 创造力

12月 · 北京 · 1000人

AiCon 全球人工智能开发与应用大会

会议时间：12月18-19日

- 大模型架构创新
- 多模态 AI 产业融合
- 具身智能
- AI for Science
- 大模型安全

InfoQ 极客传媒

模态走向融合与统一

单模态理解与生成能力持续上升，多模态呈现融合趋势

图文理解

Qwen3-VL

Gemini

视觉生成

Sora2

Seedance 1.5 pro

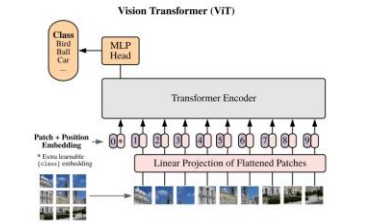
GOOGLE VEO 3

模态走向融合与统一

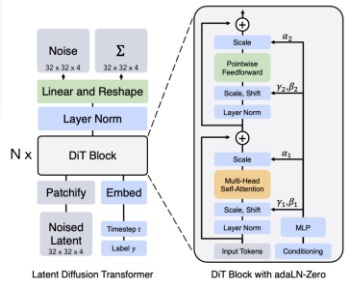
各模态对应**模型结构、存算需求、参数规模和计算模式**存在差异

结构各异

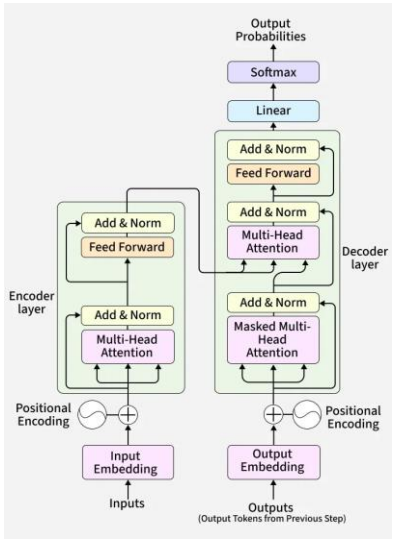
ViT



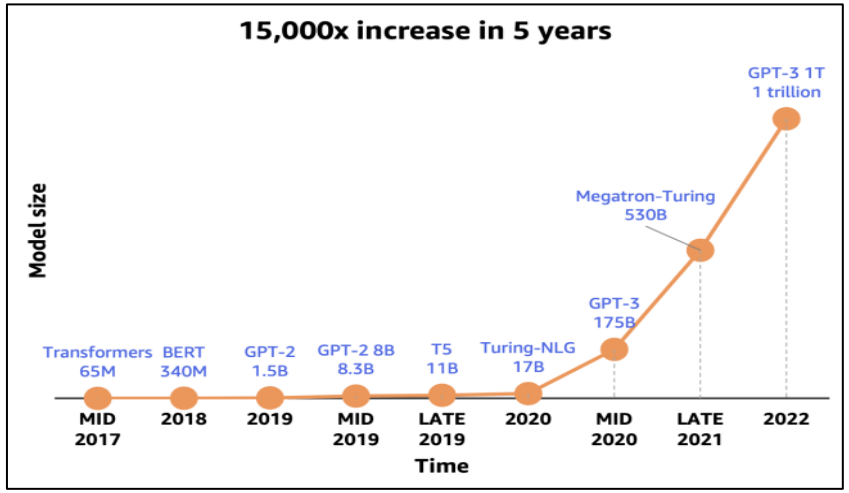
DiT



Transformer

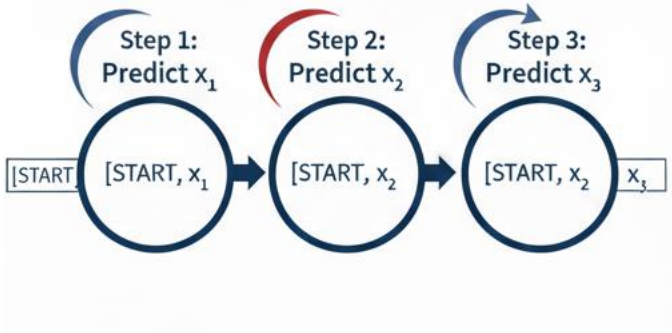


大小各异



自回归结构

顺序解码，token by token



扩散过程

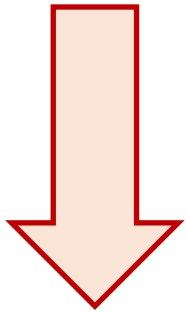
顺序去噪，step by step

硬件规格多样

延迟、吞吐、成本多重约束下，模型与硬件间需要架起技术桥梁

Qwen3　Wan　Qwen3-VL　Qwen-Image

deepseek　腾讯混元-Video　Tencent Hunyuan　腾讯混元-Image 3.0



壁仞科技 BIREN TECHNOLOGY　METAX 沐曦集成电路　Enflame 燧原科技　天数智芯 Iluvatar CoreX　平头哥　Ascend　中科曙光 Sugon　Cambricon 寒武纪　摩尔线程 MOORE THREADS　sunrise

多种模态模型高效压缩部署工具

覆盖多种模态理解生成、高效低成本的端到端压缩推理系统

理解

LightLLM

LightLLM

文/图/音频->文

LightX2V

LightX2V

Light Video Generation Inference Framework

文/图/音频->图/视频

LightTTS

Light-TTS

文本->音频

LightKernel

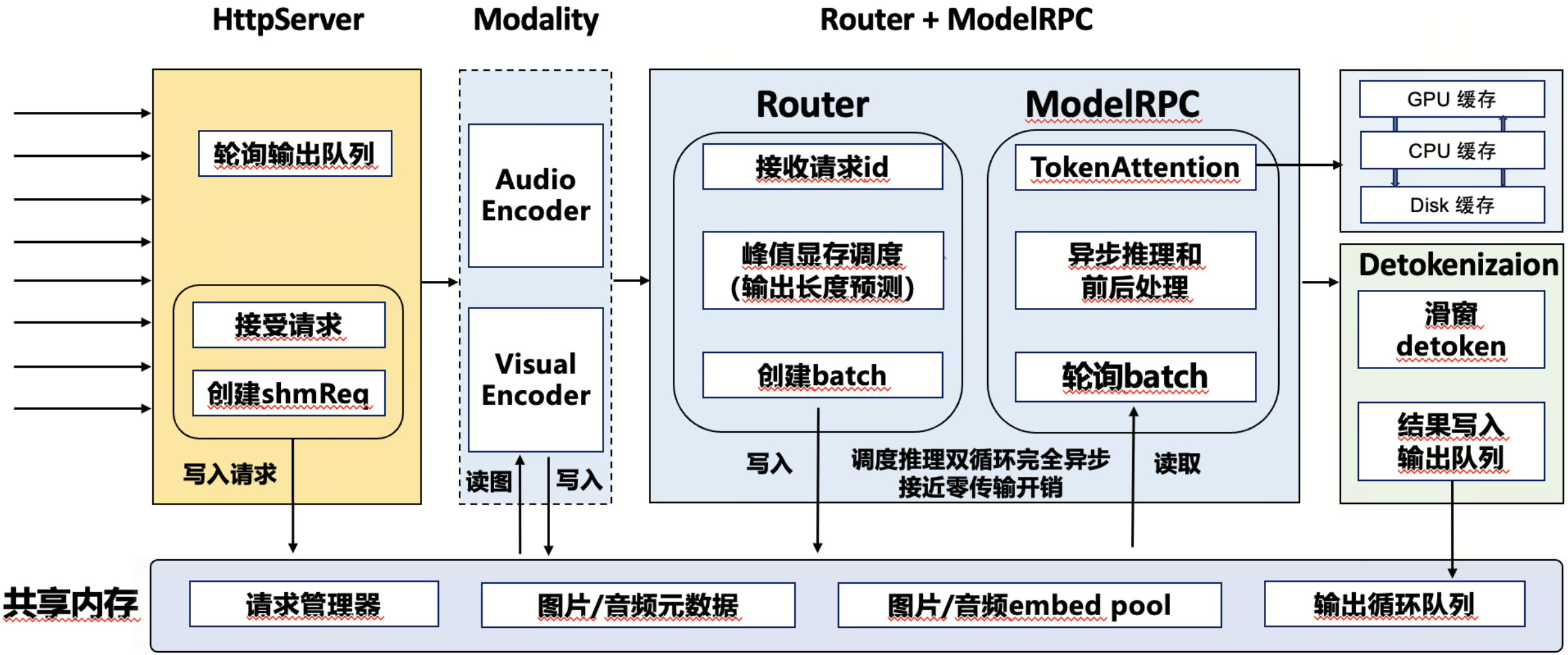
LightCompress

Quantization | Sparsification | Token Reduction

生成

视觉与语言理解：LightLLM—架构特性

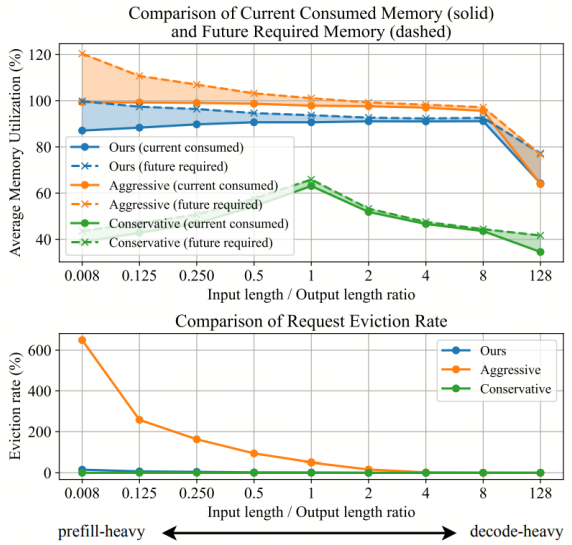
多进程异步架构 + 进程间多模态数据通信 ->高吞吐



视觉与语言理解：LightLLM—算法创新

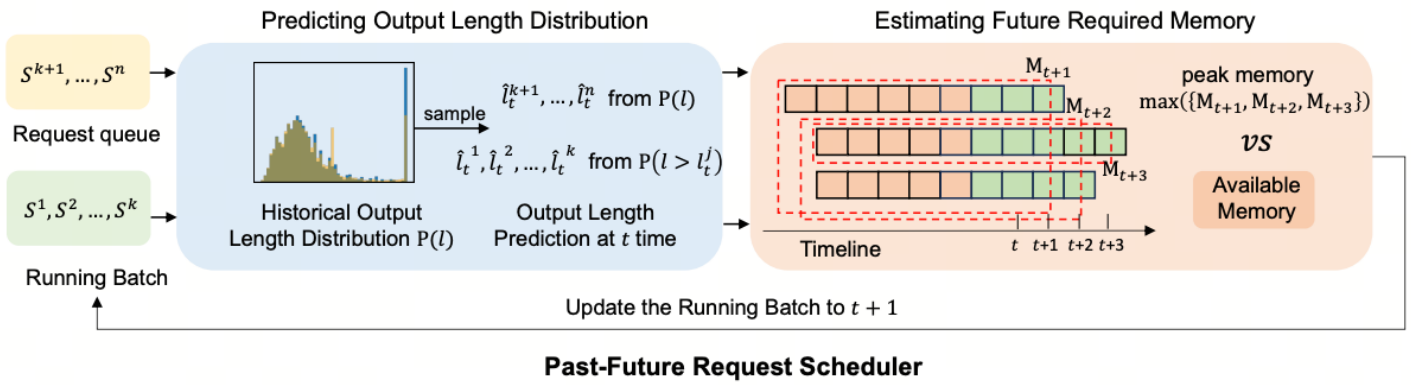
TokenAttention+输出Token量预测技术

瓶颈 **负载变化大**



长度不固定->资源难管理

创新点 **提出了基于过去未来的请求调度方法**



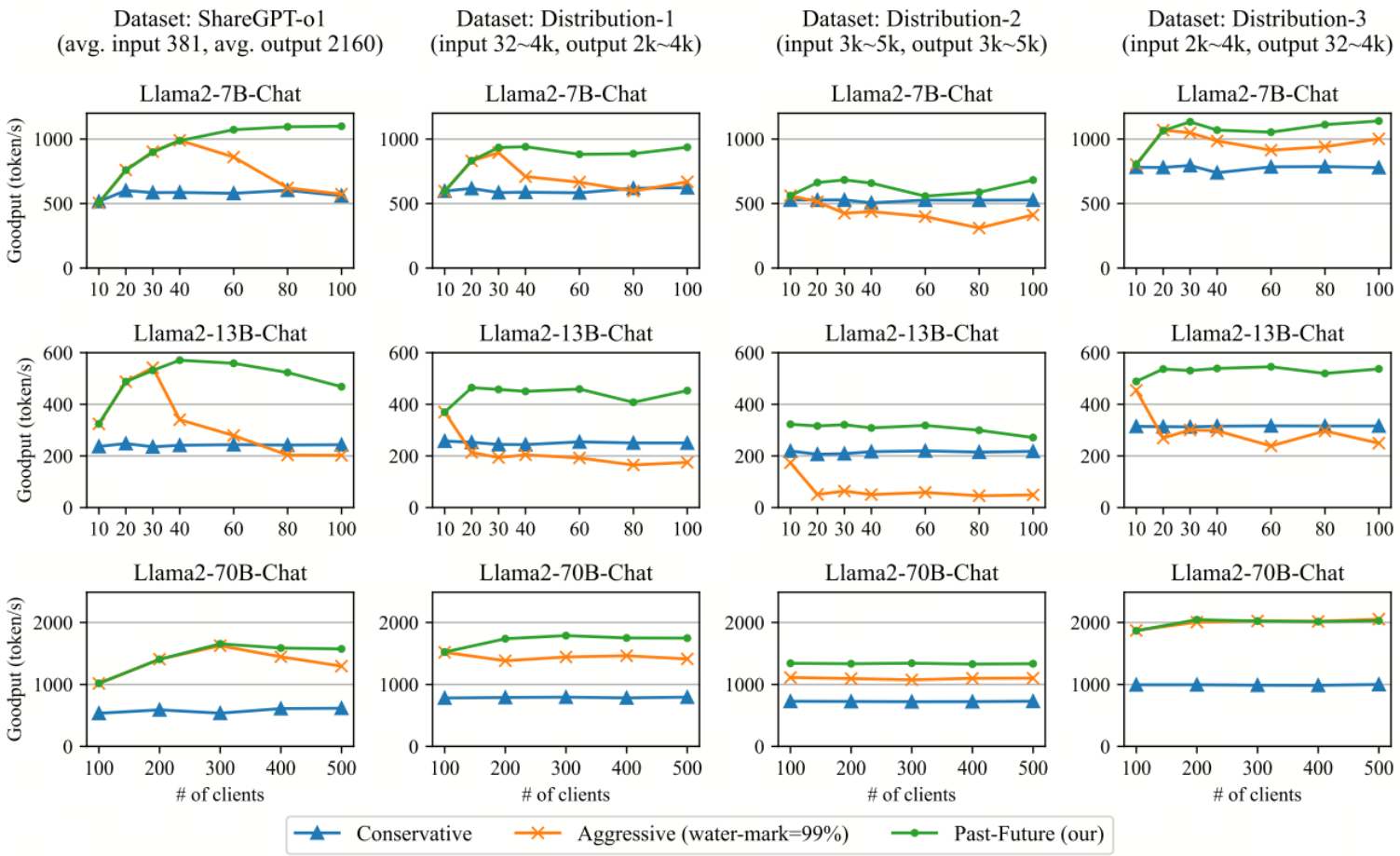
ASPLOS 2025

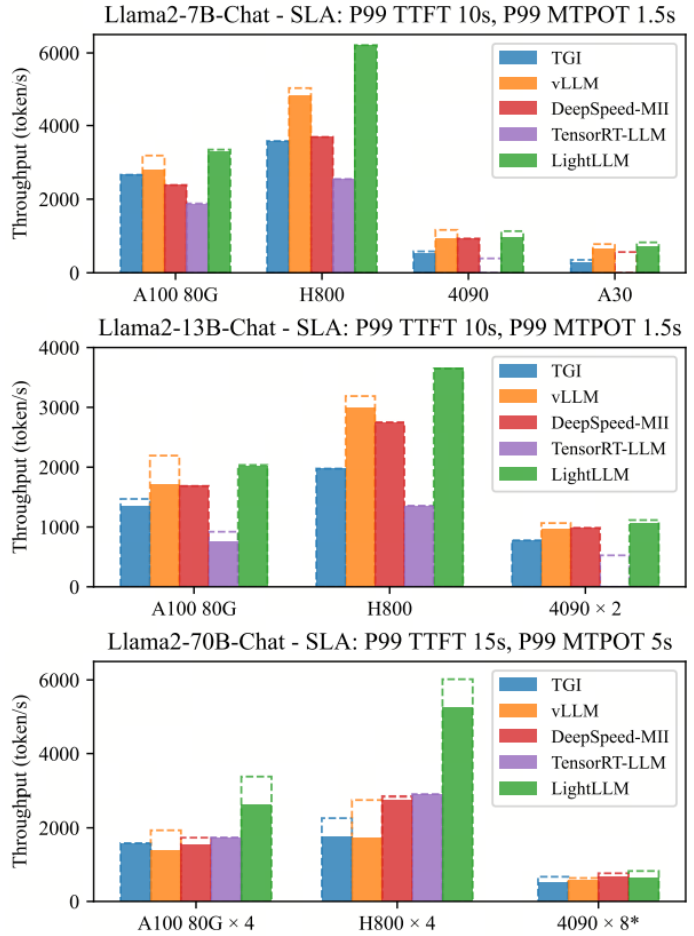
TokenAttention保证精细管理，过去未来调度保证精准调控

Gong et al., Past-Future Scheduler for LLM Serving under SLA Guarantees, ASPLOS 2025

视觉与语言理解：LightLLM—算法创新

效果：2-3倍的goodput 提升





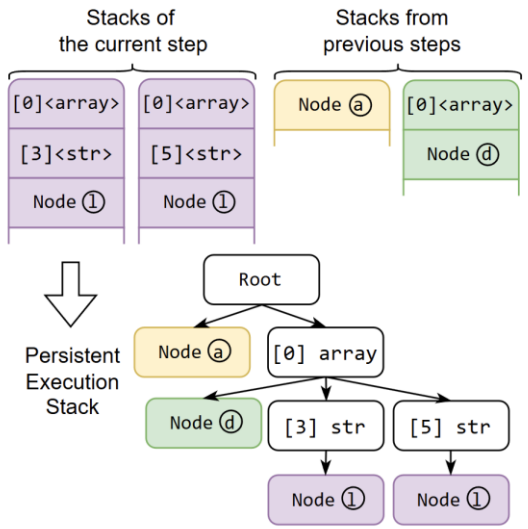
数据测试时间点：2023年12月

Gong et al., Past-Future Scheduler for LLM Serving under SLA Guarantees, ASPLOS 2025

视觉与语言理解：LightLLM—算法创新

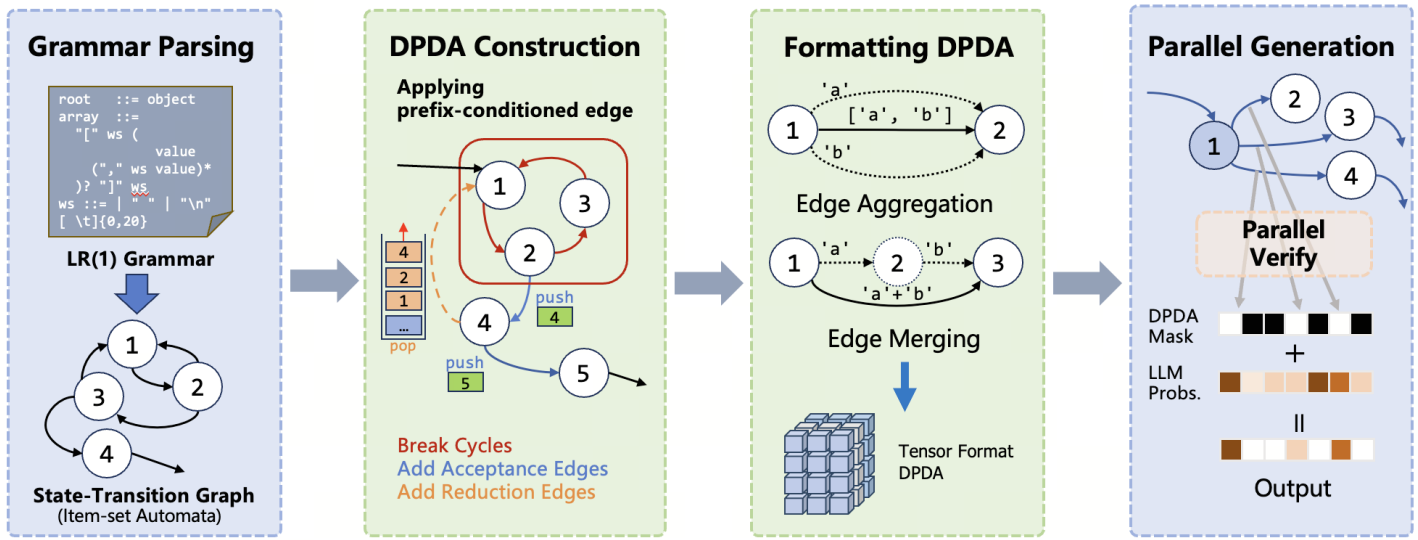
LR1文法+确定性下推自动机实现高效结构化解码

瓶颈 **结构化输出效率低**



依赖树形栈，计算开销高

创新点 **提出基于确定性下推自动机的约束解码方法**



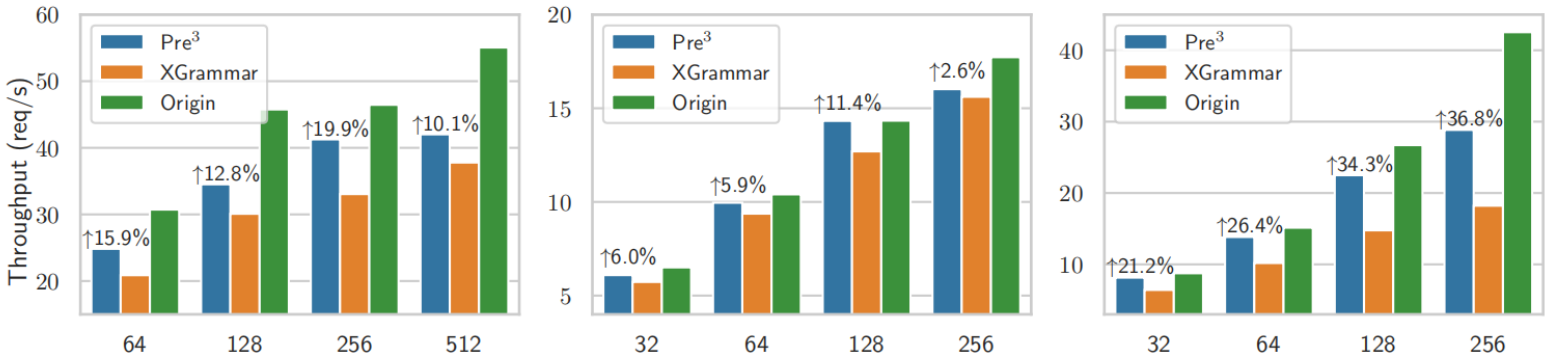
基于LR1文法+确定性下推自动机消除运行时开销

Chen et al., Pre[3]: Enabling Deterministic Pushdown Automata for Faster Structured LLM Generation, **ACL 2025 Outstanding Paper (前0.8%)**

视觉与语言理解：LightLLM—算法创新

效果：相对XGrammar取得了最大百分之40的提升

	Batch Size	16	32	64	128	256	512	1024
Llama-3-8B (Dubey et al., 2024)	XGrammar (ms)	15.19	43.69	52.07	65.21	90.98	147.64	272.77
	Pre^3 (ms)	11.77	31.12	35.88	45.32	64.42	104.46	201.16
2×H800	Reduction	**↓22.49%**	**↓28.78%**	**↓30.09%**	**↓30.50%**	**↓29.20%**	**↓29.24%**	**↓26.25%**
DeepSeek-V2-Lite-Chat (Liu et al., 2024)	XGrammar (ms)	51.76	59.45	77.74	104.06	121.46	-	-
	Pre^3 (ms)	49.91	53.71	54.41	61.63	75.47	-	-
15.7B 2×H800	Reduction	**↓3.57%**	**↓9.65%**	**↓30.01%**	**↓40.78%**	**↓37.86%**	-	-
Qwen2-14B (Yang et al., 2024a)	XGrammar (ms)	16.77	47.94	57.05	74.54	98.64	162.47	285.42
	Pre^3 (ms)	16.52	47.94	47.89	65.50	90.20	143.83	232.18
INT8 2×H800	Reduction	**↓1.52%**	**↓0.12%**	**↓2.37%**	**↓12.14%**	**↓8.55%**	**↓11.47%**	**↓18.65%**
Llama-2-70B (Touvron et al., 2023)	XGrammar (ms)	28.75	55.12	56.94	68.79	85.92	-	-
	Pre^3 (ms)	27.20	54.24	54.18	62.27	75.72	-	-
4×H800	Reduction	**↓5.39%**	**↓1.60%**	**↓4.85%**	**↓9.48%**	**↓11.87%**	-	-

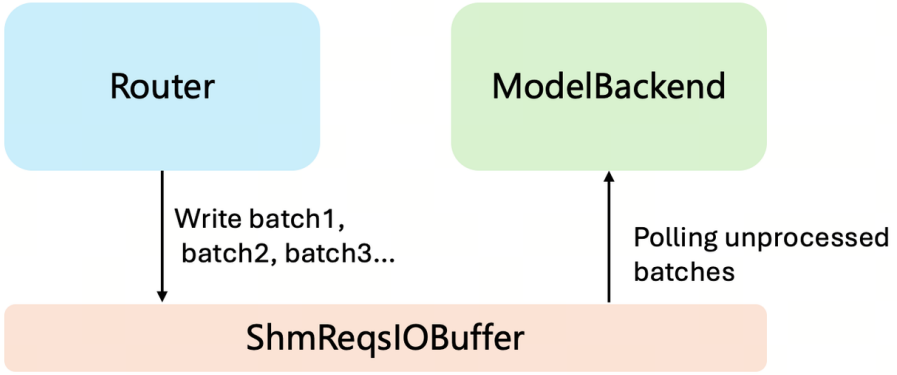


Left: Llama3-8B, Middle: Llama2- 70B, Right: DeepSeek-V2-Lite-Chat.

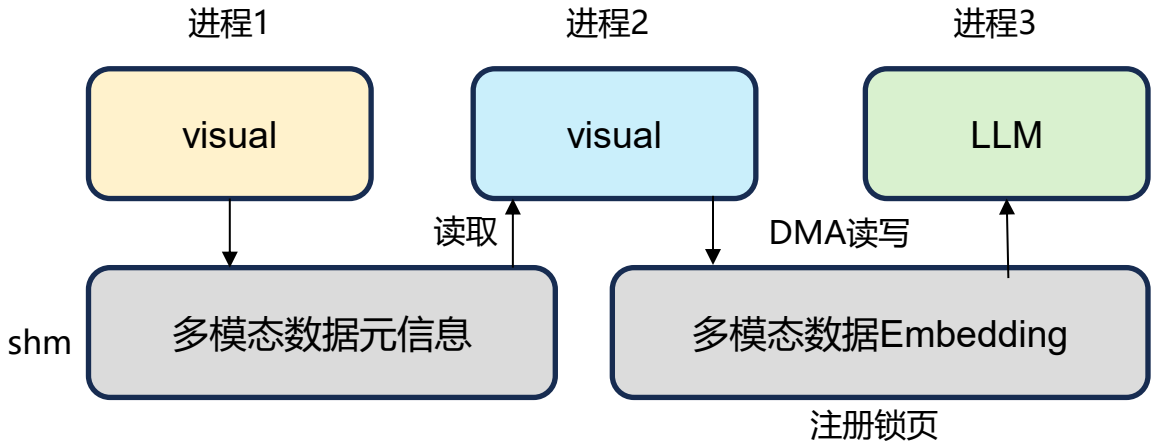
Chen et al., Pre^3: Enabling Deterministic Pushdown Automata for Faster Structured LLM Generation, **ACL 2025 Outstanding Paper (前0.8%)**

视觉与语言理解：LightLLM—工程创新

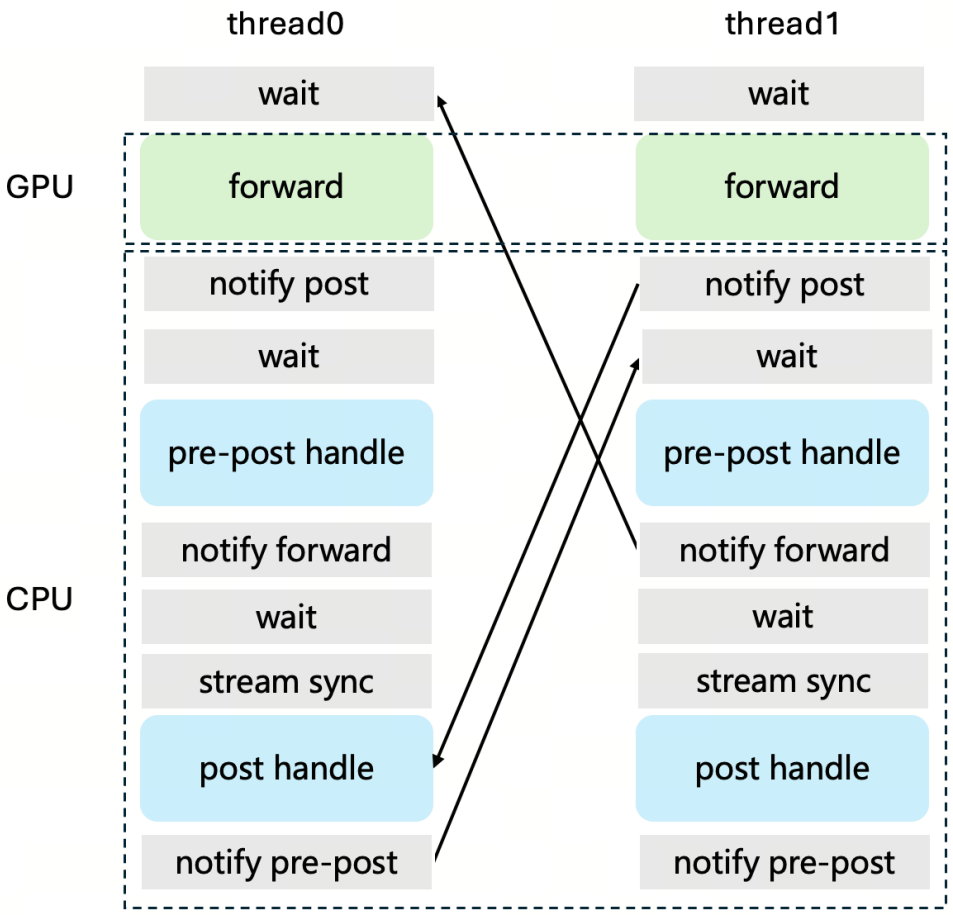
高度异步化的并行设计



调度和模型推理异步化



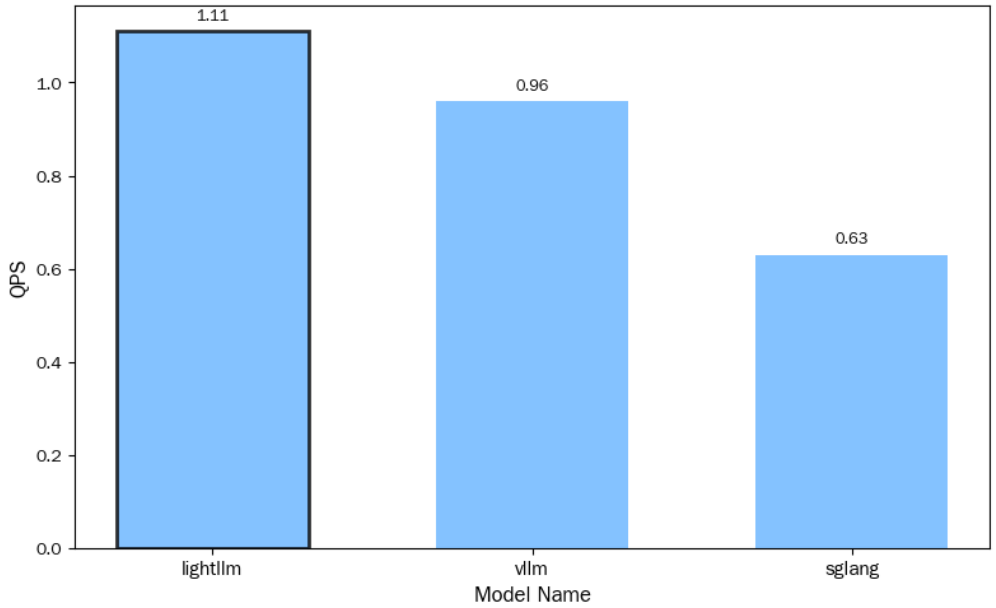
Visual模块和LLM 模块异步化和高效通信



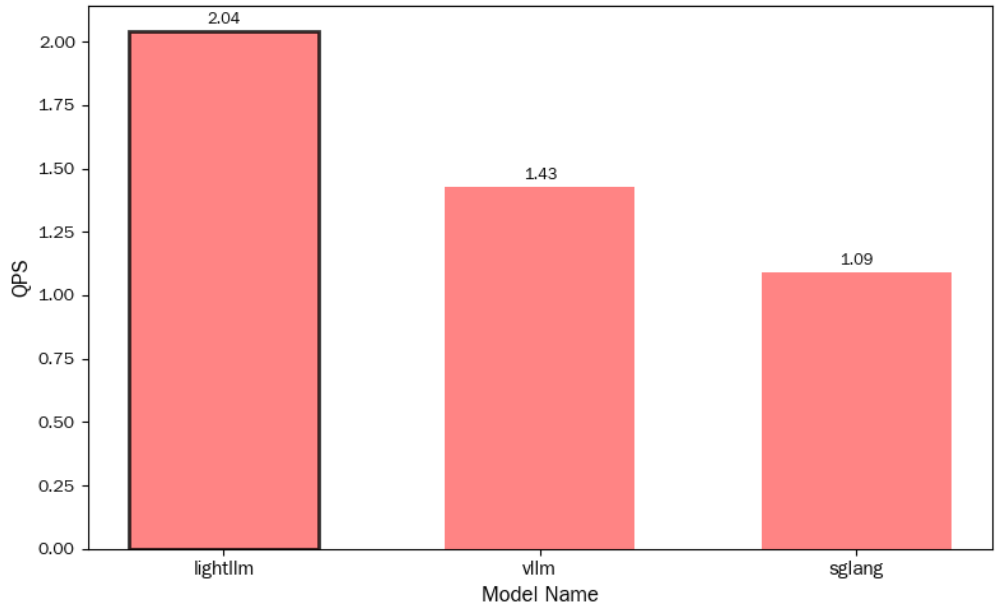
GPU推理和CPU前后处理异步化

视觉与语言理解：LightLLM—工程创新

效果：在多模态模型上优于vllm/sglang最多30%



Qwen3-VL-8B 4090单卡

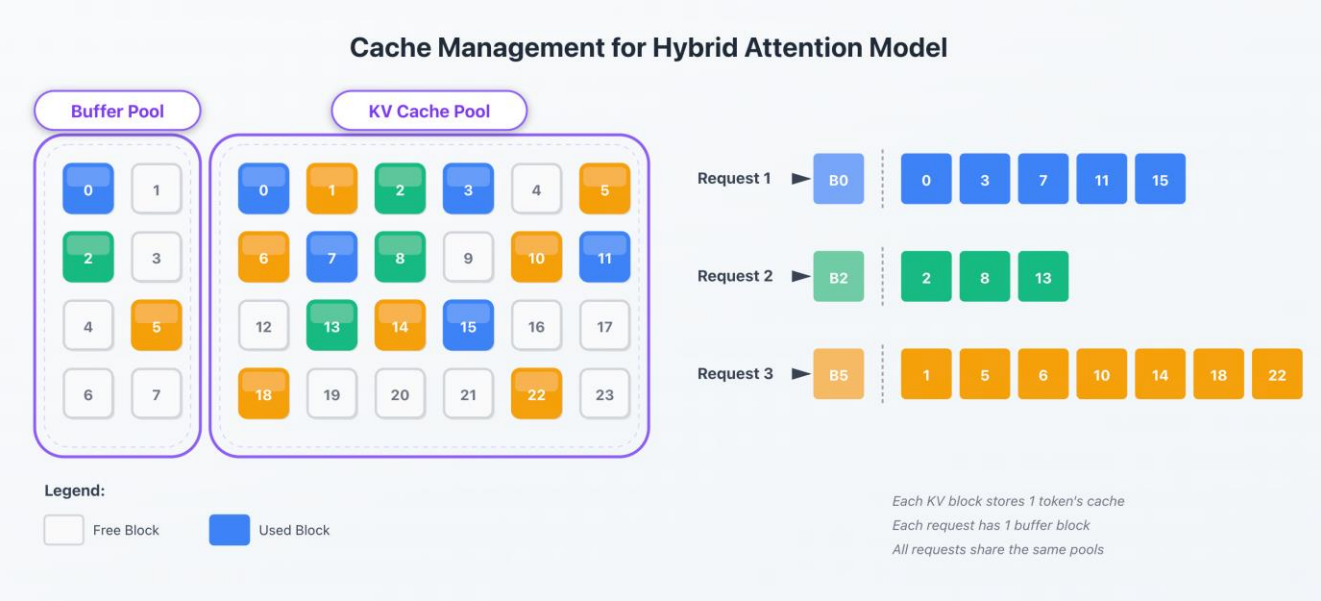


Qwen3-VL-235B H200 八卡

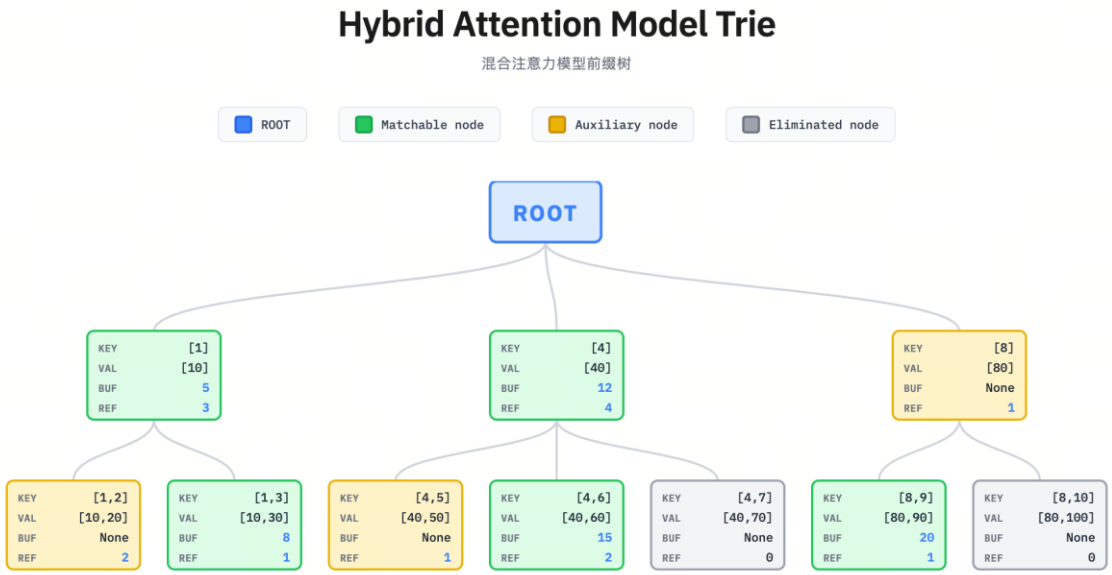
vllm/sglang版本截止 2025.12.19，测试数据集为miner-U数据集，并发32。

视觉与语言理解：LightLLM—工程创新

混合注意力模型cache管理优化



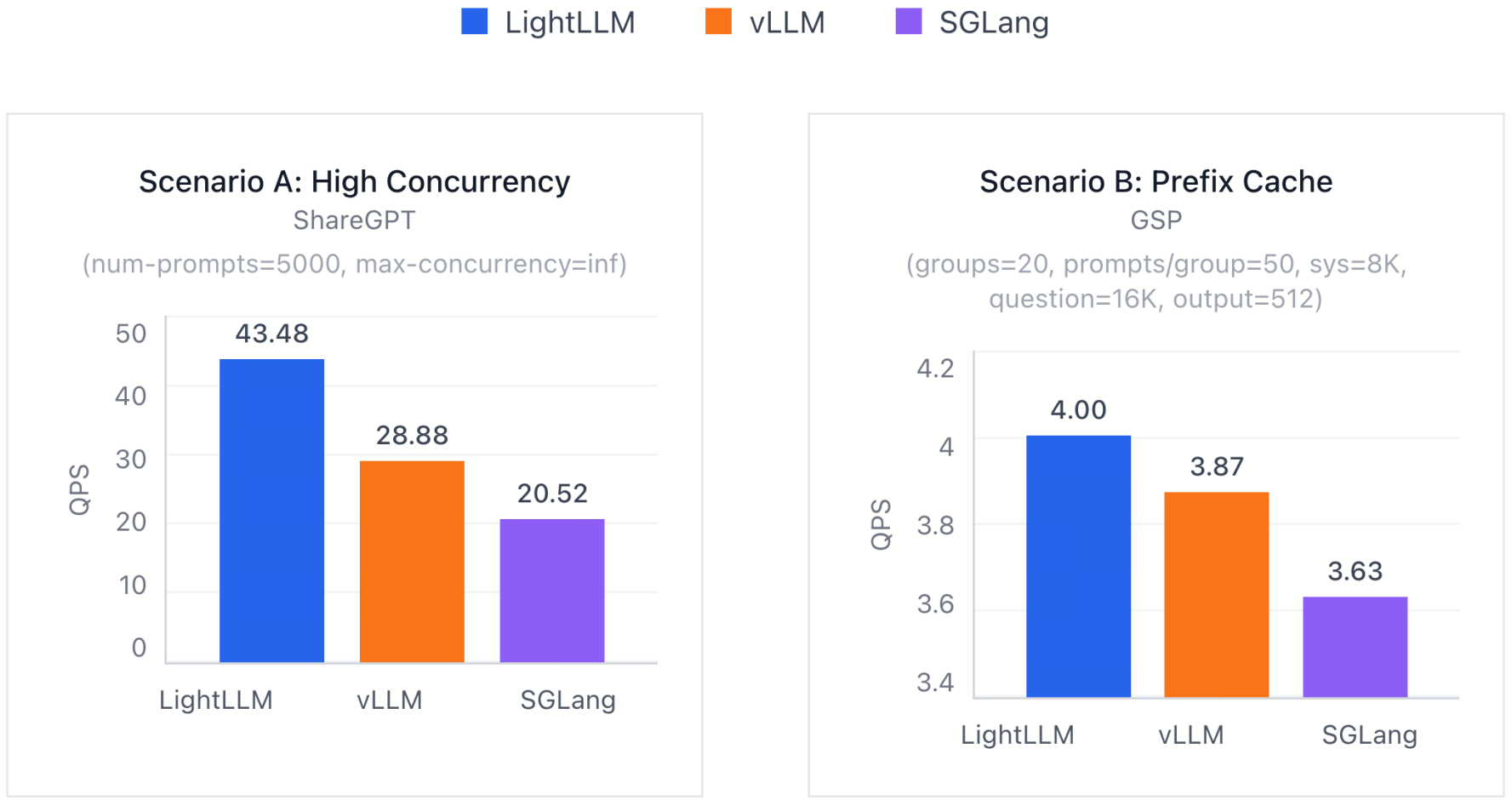
Token-level kv 和 linear buffer隔离管理，统一寻址



混合注意力前缀树

视觉与语言理解：LightLLM—工程创新

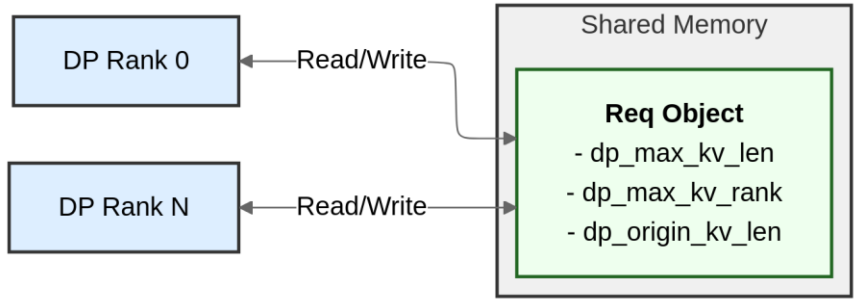
效果：领先vllm/sglang 百分之33



Qwen-Next-80B 四卡H200

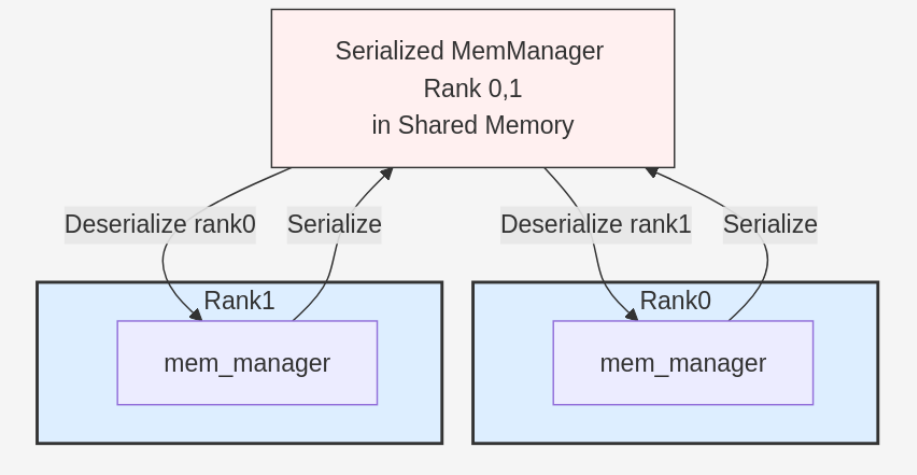
视觉与语言理解：LightLLM—工程创新

单机不同DP GPU KV cache交换

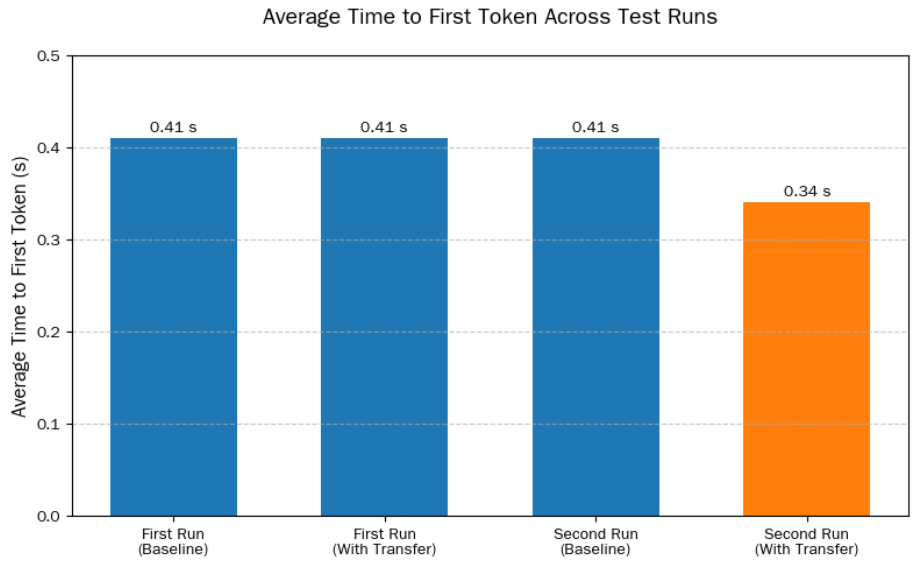


各rank都可获取请求元信息

各个dp rank高性能kv cache 交换



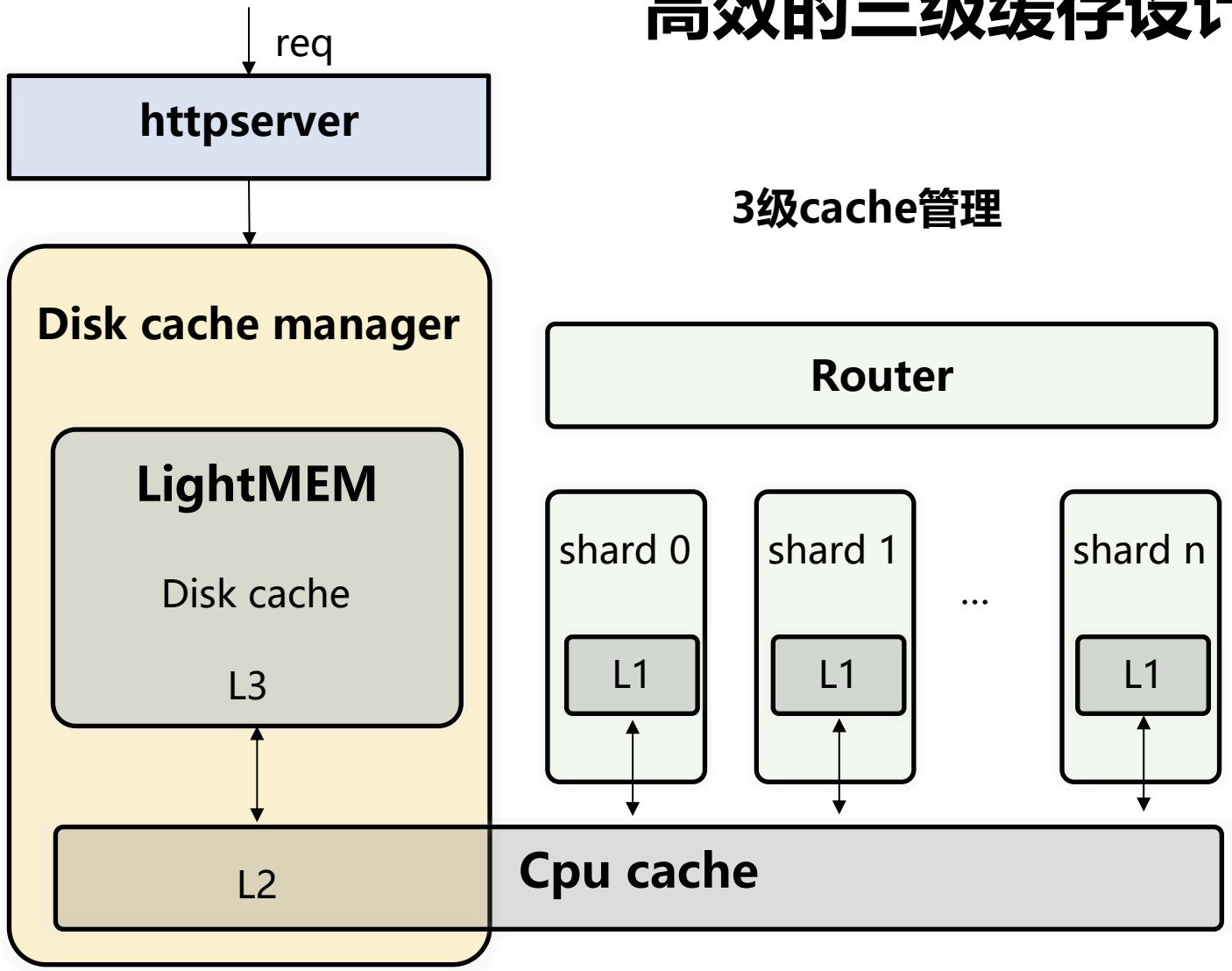
各个dp rank 交换kv cache handle



DeepSeek-V3 首字提升**17%**

视觉与语言理解：LightLLM—工程创新

高效的三级缓存设计

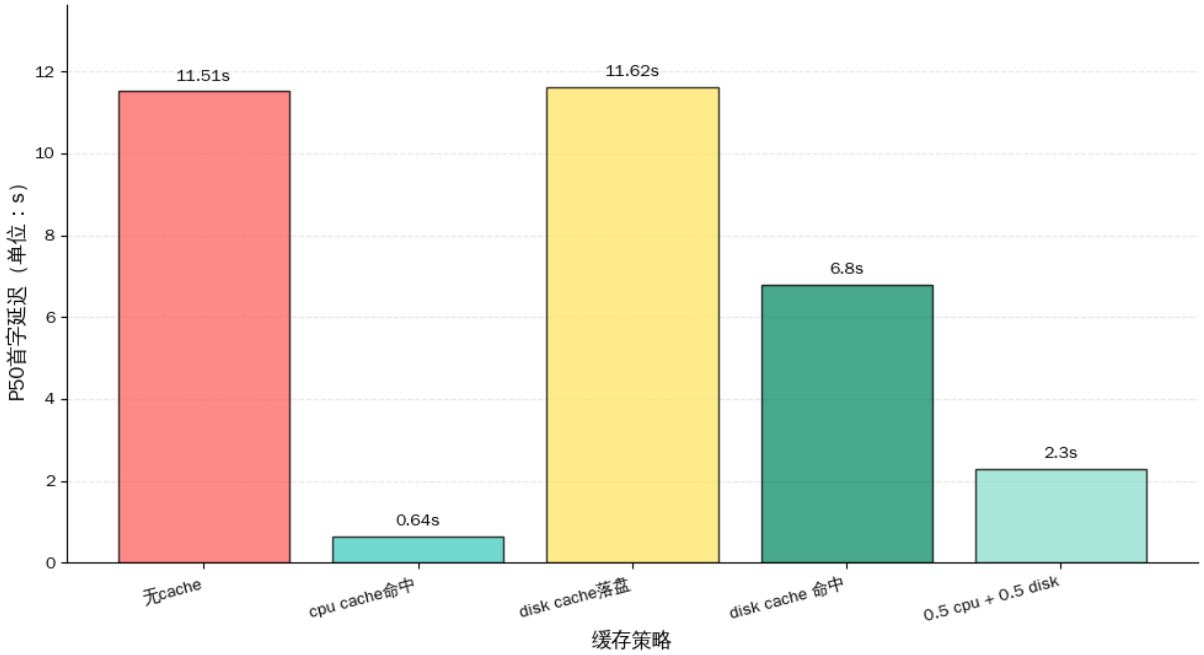


相对其他开源方案的优势

- 无大小关系限制，可灵活配比L1 L2 L3
- 储存效率高，cpu cache 储存相同数量token仅需sglang一半的空间
- 生产级可靠的稳定性
- 提供了高性能的键值缓存管理库LightMEM，不依赖于文件系统的性能

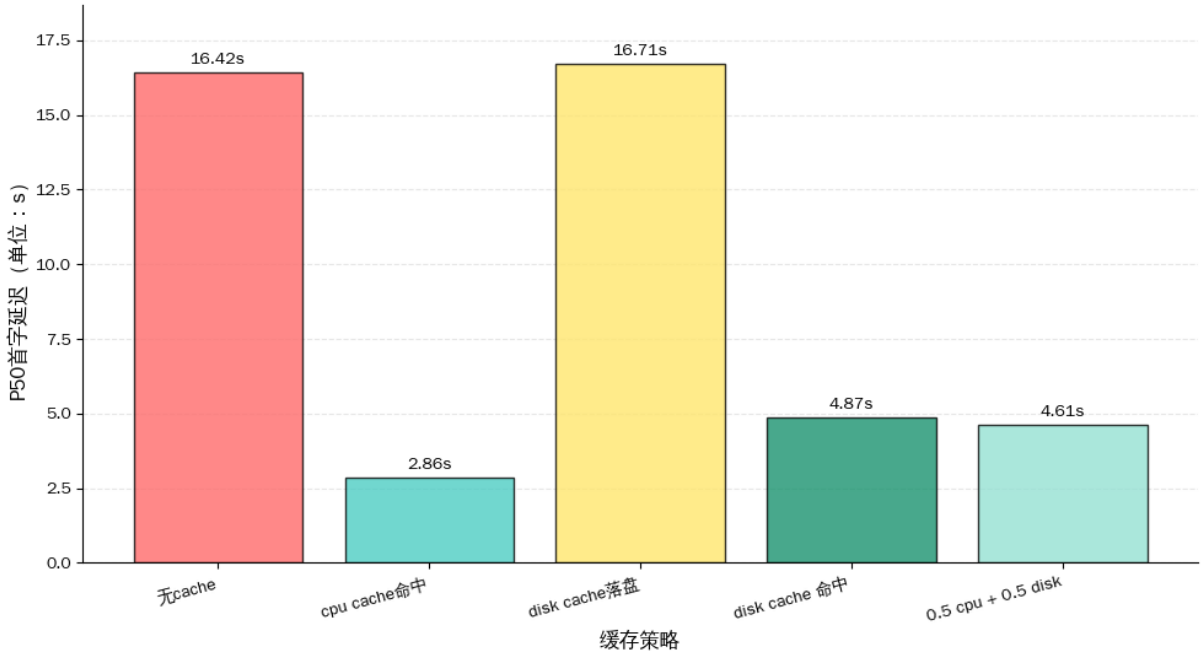
视觉与语言理解：LightLLM—工程创新

效果：DeepSeek-V3上可获得250%的首字收益



Qwen3-235B-A3, 10k上下文, 10qps, H200x8

disk write: 7.1GB/s disk read: 8.4GB/s

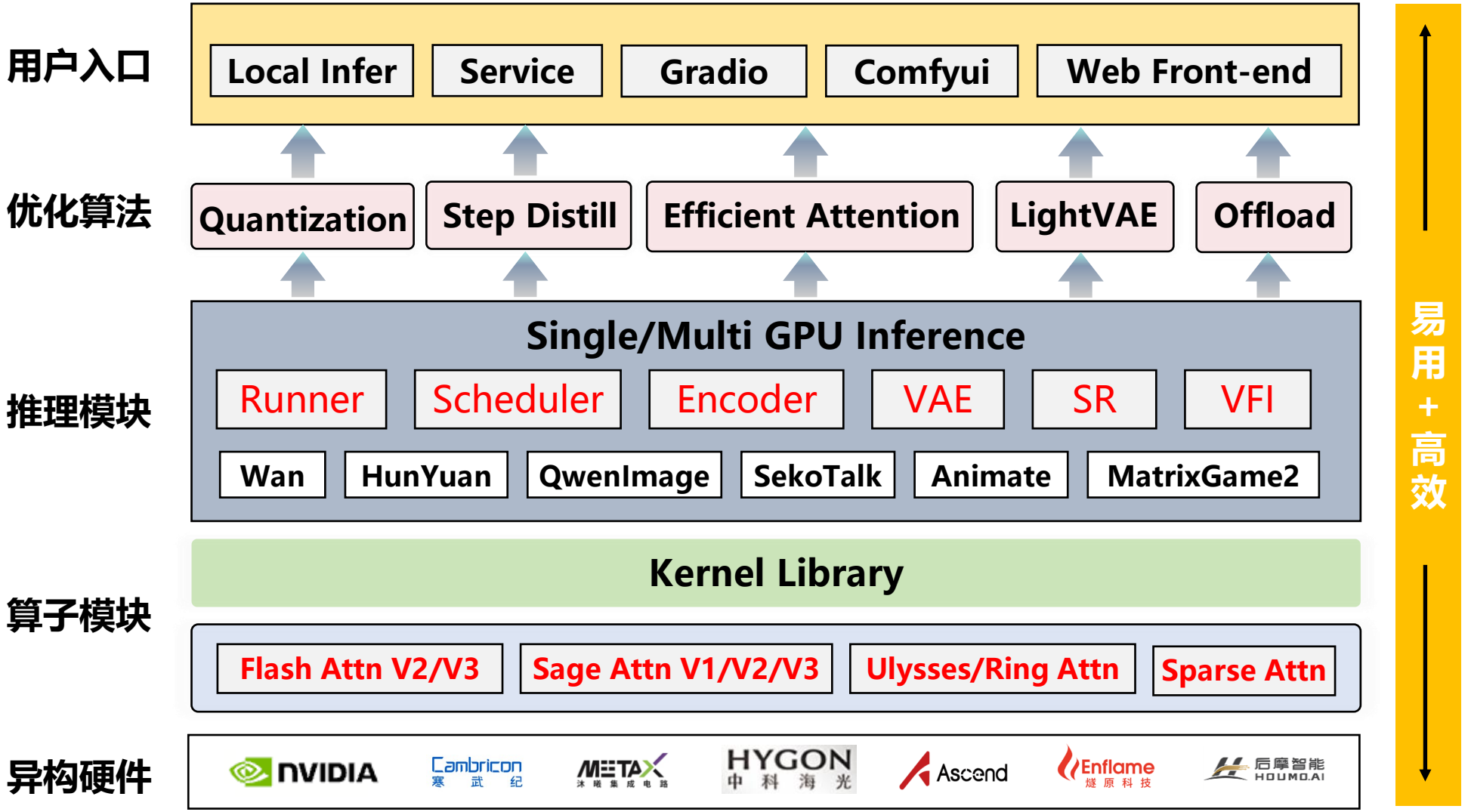


DeepSeek-v3, 10k上下文, 10qps, H200x8

disk write: 6.3GB/s disk read: 6.7GB/s

视觉与图像生成：LightX2V—架构特性

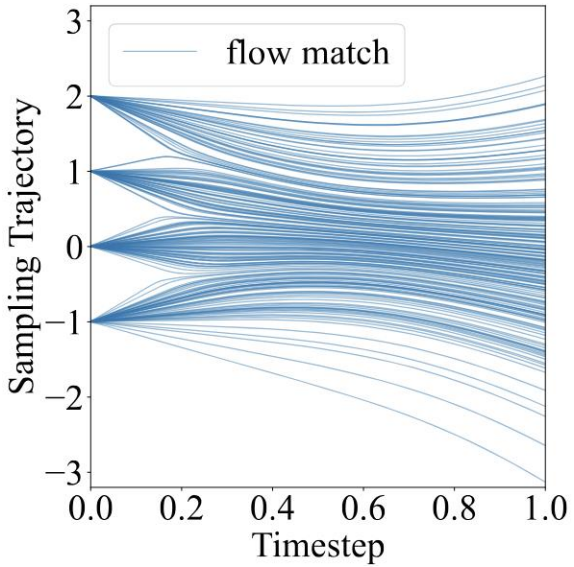
高效完备的图像视频生成推理框架



视觉与图像生成：LightX2V—算法创新

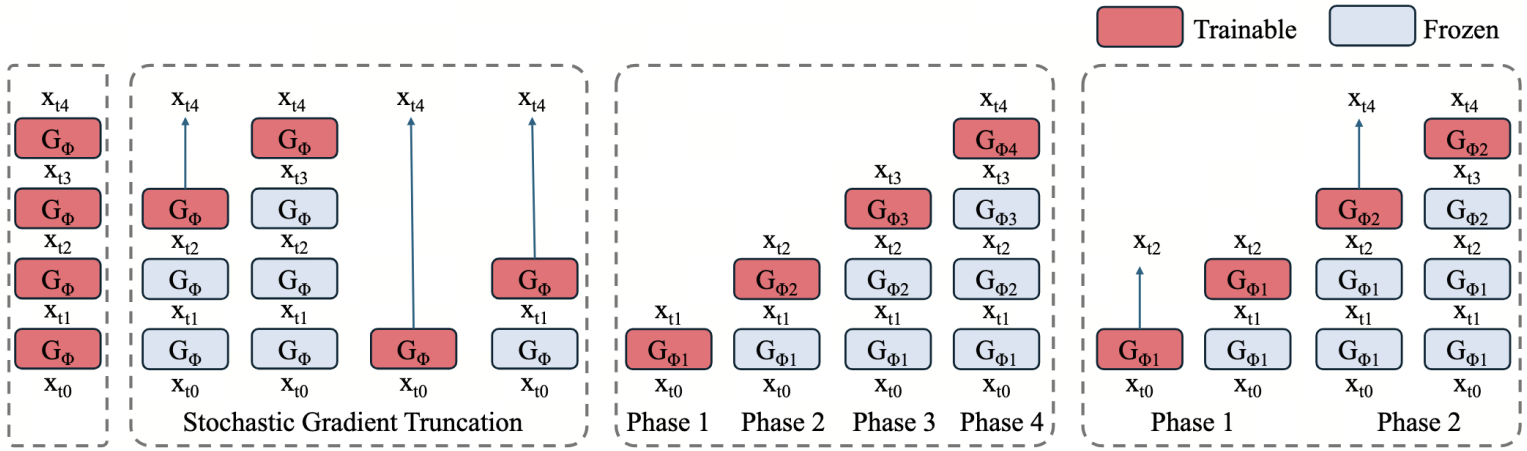
4步蒸馏:提出基于阶段性分布匹配的扩散模型步数蒸馏方法，能够对少步数扩散的时序信息进行缩减并保持生成效果

瓶颈 **时序信息丢失**



少步蒸馏极易奔溃

创新点 **统一训练推理范式，学习高效历史特征缓存器**



理论+实验证明：渐进式、分段式分布匹配实现少步扩散效果保持

Fan et al., Phased DMD: Few-step Distribution Matching Distillation via Score Matching within Subintervals, Arxiv 2025

视觉与图像生成：LightX2V—算法创新

效果：扩散推理从50步压缩到2步，保持生成效果
各项指标均超过现有先进方法

Method	Wan2.1-T2V-14B		Wan2.2-T2V-A14B		Qwen-Image	
	DINOv3 ↓	LPIPS ↑	DINOv3 ↓	LPIPS ↑	DINOv3 ↓	LPIPS ↑
Base model	0.708	0.607	0.732	0.531	0.907	0.483
DMD	0.825	0.522	-	-	-	-
DMD with SGTS	0.826	0.521	0.828	0.447	**0.941**	0.309
Phased DMD (Ours)	**0.782**	**0.544**	**0.768**	**0.481**	0.958	**0.322**

可视化效果近乎无损，效果超过普通分布匹配方法

Method	Optical Flow ↑	Dynamic Degree ↑
Base model	10.66	79.35 %
DMD with SGTS	5.27	72.90 %
Phased DMD(Ours)	**7.76**	**78.71 %**

少步扩散挑战：动态性显著提升，高噪、低噪效果均有提升

Fan et al., Phased DMD: Few-step Distribution Matching Distillation via Score Matching within Subintervals, Arxiv 2025

视觉与图像生成：LightX2V—算法创新

2步蒸馏：结合分布匹配、对抗训练、平均速度蒸馏的扩散模型步数蒸馏，极大减少生成步数并保证视频的一致性

瓶颈 少步蒸馏一致性差

2步分布匹配蒸馏模型

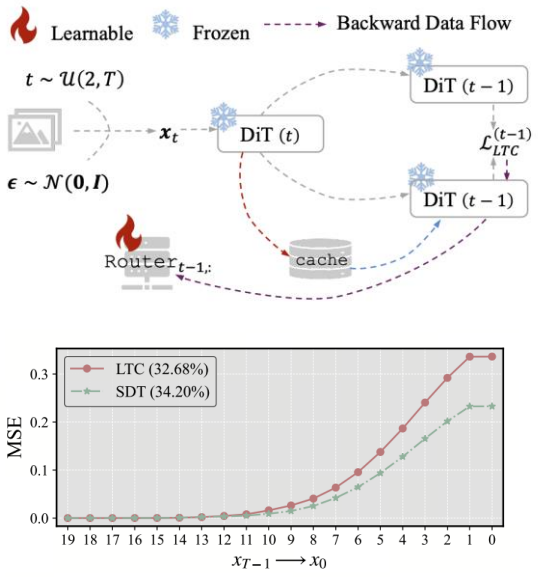
效果 结合多种训练技术保证一致性

2步分布匹配+对抗训练+平均速度蒸馏模型

视觉与图像生成：LightX2V—算法创新

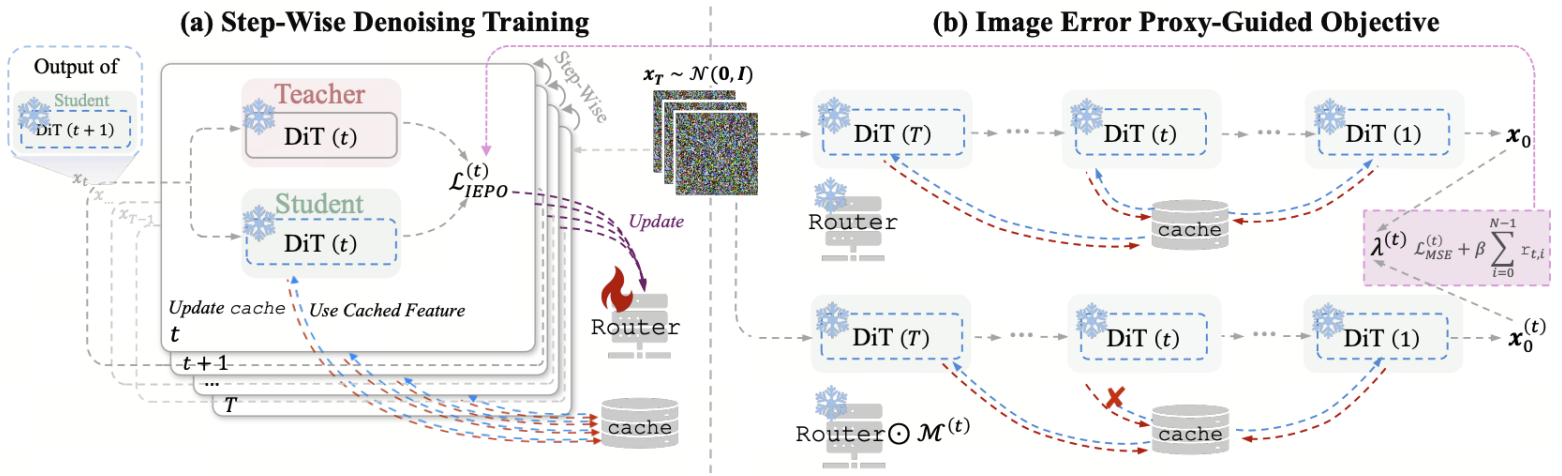
特征缓存: 提出首个基于累计误差意识的扩散模型特征缓存训练方法，以极低训练开销保持生成效果同时实现一键加速

瓶颈 **累积误差过大**



模型最终输出误差大

创新点 **辅助模块改进收敛，低秩删除消除额外开销**



分析+实验证明：引入逐步迭代式训练和最终误差引导的优化目标大幅度减小了特征缓存中累积误差

视觉与图像生成：LightX2V—算法创新

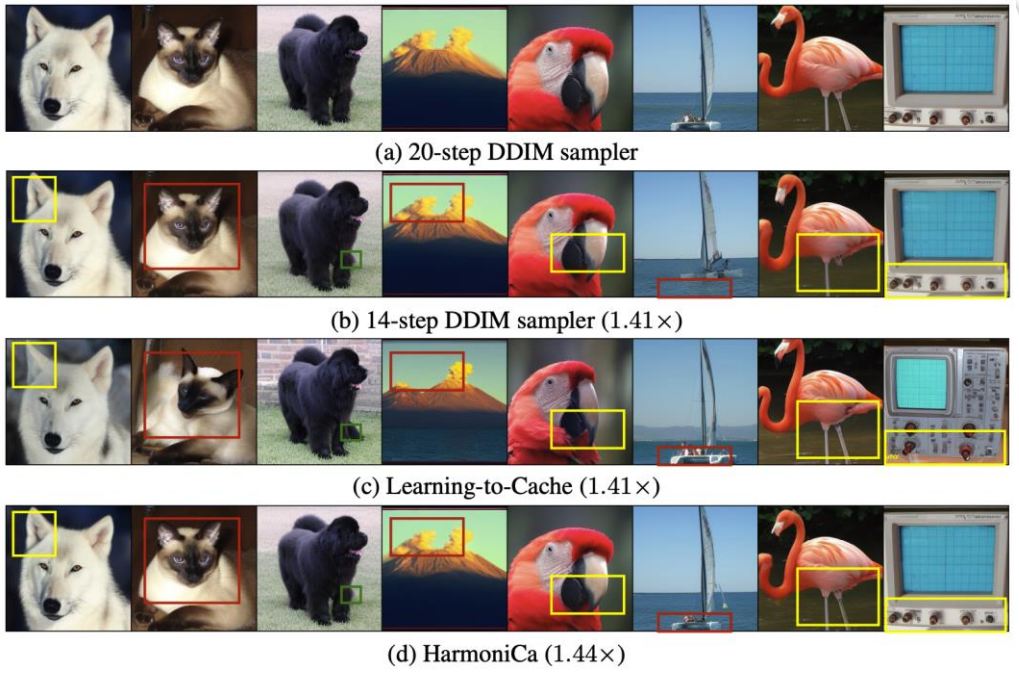
效果：大于1.5倍无损加速扩散模型，各项指标均超过现有先进方法

Method	T	DINO↑	HPSv2↑	PickScore↑	CUR(%)↑	Latency(s)↓
PixArt-α 256 × 256 (cfg = 4.5)						
DPM-Solver++ (Lu et al., 2022b)	20	0.3082	28.91	27.89	-	0.553
DPM-Solver++ (Lu et al., 2022b)	15	0.2582	27.98	23.02	-	0.418$_{(1.32\times)}$
FORA (Selvaraju et al., 2024)	20	0.2712	28.11	22.44	50.00	0.364$_{(1.52\times)}$
HarmoniCa	20	**0.3235**	**28.72**	**26.65**	**56.01**	**0.346**$_{(1.60\times)}$
PixArt-α 512 × 512 (cfg = 4.5)						
DPM-Solver++ (Lu et al., 2022b)	20	0.3339	30.53	28.52	-	1.759
DPM-Solver++ (Lu et al., 2022b)	15	0.3127	29.79	22.03	-	1.291$_{(1.36\times)}$
FORA (Selvaraju et al., 2024)	20	0.3099	29.82	21.98	50.0	1.150$_{(1.53\times)}$
HarmoniCa	20	**0.3289**	**30.28**	**27.47**	**54.64**	**1.072**$_{(1.64\times)}$

Method	T	CLIP↑	FID↓	sFID↓	CUR(%)↑	Latency(s)↓
PixArt-α 256 × 256 (cfg = 4.5)						
DPM-Solver++ (Lu et al., 2022b)	20	30.96	27.68	36.39	-	0.553
DPM-Solver++ (Lu et al., 2022b)	15	30.77	31.68	38.92	-	0.418$_{(1.32\times)}$
FORA (Selvaraju et al., 2024)	20	31.10	27.42	37.98	50.00	0.364$_{(1.52\times)}$
HarmoniCa	20	**31.13**	**26.33**	**37.85**	**56.01**	**0.346**$_{(1.60\times)}$
IDDPM (Nichol & Dhariwal, 2021)	100	31.25	24.15	33.65	-	2.572
IDDPM (Nichol & Dhariwal, 2021)	75	31.25	24.17	33.73	-	1.868$_{(1.37\times)}$
FORA (Selvaraju et al., 2024)	100	31.25	25.16	33.62	50.00	1.558$_{(1.65\times)}$
HarmoniCa	100	**31.17**	**23.73**	**32.23**	**53.24**	**1.523**$_{(1.69\times)}$
SA-Solver (Xue et al., 2024)	25	31.31	26.78	38.35	-	0.891
SA-Solver (Xue et al., 2024)	20	31.23	27.45	39.01	-	0.665$_{(1.34\times)}$
HarmoniCa	25	**31.27**	**27.07**	**38.62**	**54.19**	**0.561**$_{(1.59\times)}$
PixArt-α 512 × 512 (cfg = 4.5)						
DPM-Solver++ (Lu et al., 2022b)	20	31.30	23.96	40.34	-	1.759
DPM-Solver++ (Lu et al., 2022b)	15	**31.29**	25.12	40.37	-	1.291$_{(1.36\times)}$
HarmoniCa	20	**31.29**	**24.81**	**40.18**	**54.64**	**1.072**$_{(1.64\times)}$
SA-Solver (Xue et al., 2024)	25	31.23	25.43	39.84	-	2.263
SA-Solver (Xue et al., 2024)	20	31.19	25.85	40.08	-	1.738$_{(1.30\times)}$
HarmoniCa	25	**31.20**	**25.74**	**39.99**	**54.24**	**1.406**$_{(1.61\times)}$

(a) PixArt-Σ (b) HarmoniCa (1.73×)

可视化无损，部分指标高于加速前模型



(a) 20-step DDIM sampler

(b) 14-step DDIM sampler (1.41×)

(c) Learning-to-Cache (1.41×)

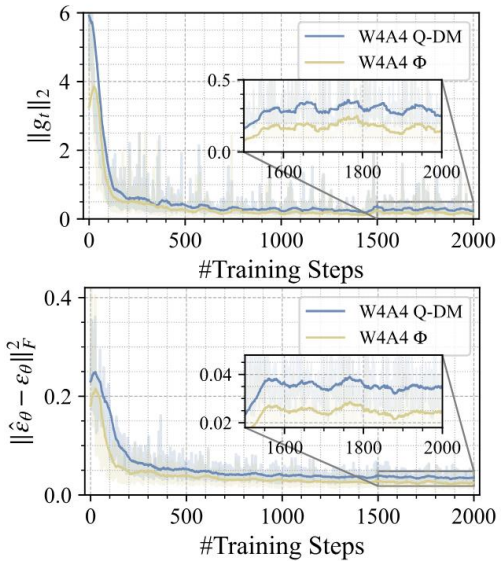
(d) HarmoniCa (1.44×)

对于细节的生成显著优于已有算法

视觉与图像生成：LightX2V—算法创新

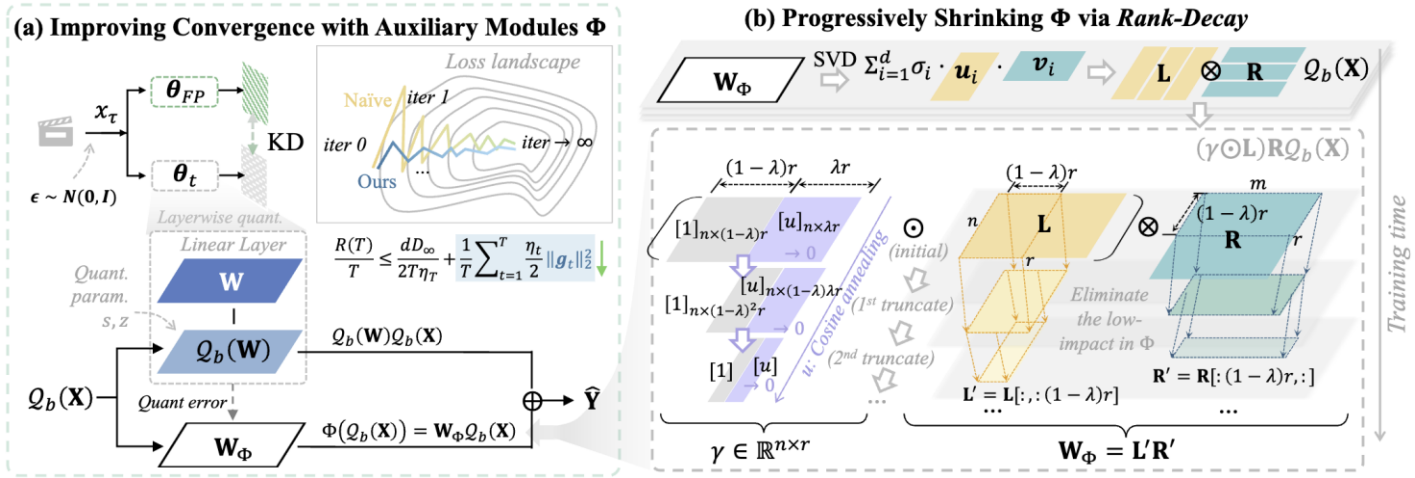
量化感知训练: 提出首个基于辅助模块增强的视频扩散模型量化感知训练方法，能够对保持生成效果同时实现极致压缩

瓶颈 收敛到次优局部



模型收敛受梯度范数影响

创新点 辅助模块改进收敛，低秩删除消除额外开销

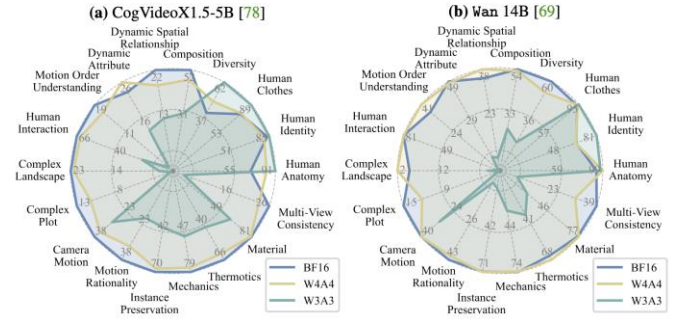


理论+实验证明：引入增强模块与渐进式低秩消除实现了压缩后的模型效果保持

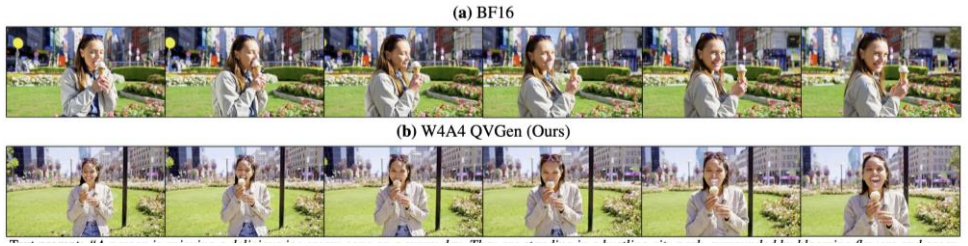
视觉与图像生成：LightX2V—算法创新

效果：4倍压缩视频扩散模型，保持生成效果
各项指标均超过现有先进方法

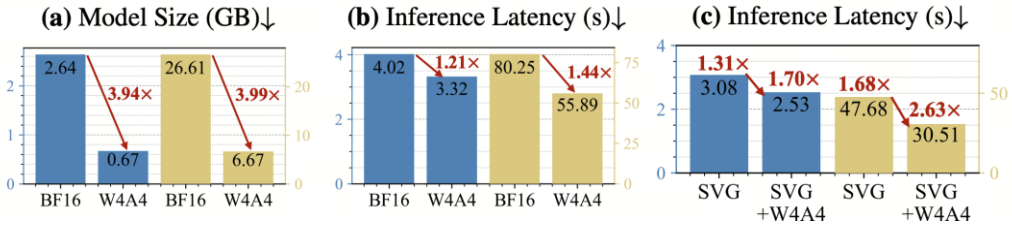
Wan 1.3B (CFG = 5.0, 480p, fps = 16)									
Full Prec.	16/16	64.30	58.21	97.37	70.28	95.94	93.84	28.05	24.67
ViDiT-Q [82]†	4/6	56.24	50.18	94.81	52.43	89.67	82.53	13.45	19.58
SVDQuant [34]†	4/6	58.16	51.27	97.05	49.44	93.74	91.71	14.18	23.26
SVDQuant [34]†♣	4/4	57.57	46.30	94.21	72.22	93.16	77.96	12.73	21.91
LSQ [10]*	4/4	59.11	49.09	**98.35**	71.11	92.66	91.67	10.38	18.83
Q-DM [37]*	4/4	60.40	52.50	97.22	76.67	93.37	89.26	13.28	21.63
EfficientDM [19]*	4/4	60.70	53.57	96.18	56.39	93.74	91.70	11.77	21.19
QVGen (Ours)*	4/4	**63.08**$_{+2.38}$	**54.67**$_{+1.10}$	98.25$_{-0.10}$	**77.78**$_{+1.11}$	**94.08**$_{+0.34}$	**92.57**$_{+0.87}$	**15.32**$_{+2.04}$	**23.01**$_{+1.38}$
LSQ [10]*	3/3	58.80	46.86	98.22	23.61	91.86	89.42	0.89	15.51
Q-DM [37]*	3/3	56.19	44.95	95.13	76.94	92.09	83.82	1.79	16.89
EfficientDM [19]*	3/3	42.32	33.52	96.50	70.28	92.10	74.79	0.04	11.38
QVGen (Ours)*	3/3	**67.35**$_{+8.55}$	**49.71**$_{+2.85}$	**98.93**$_{+0.71}$	**84.14**$_{+7.20}$	**93.62**$_{+1.52}$	**92.25**$_{+2.83}$	**5.71**$_{+3.92}$	**20.11**$_{+3.22}$



首次在Vbench/Vbench-2指标上接近无损，部分指标高于浮点模型



可视化效果不输浮点模型



与已有稀疏化注意力方案结合，实现了超过2.5倍端到端加速

视觉与图像生成：LightX2V—算法创新

NVFP4超低比特量化结合稀疏注意力步数蒸馏

单卡5090，Wan2.1-I2V-14B-480p模型

原始模型

单步12.41s

端到端 498.92s

Nvfp4+步数蒸馏

单步3.40s

端到端 17.23s

Nvfp4+稀疏attn+步数蒸馏

单步2.45s

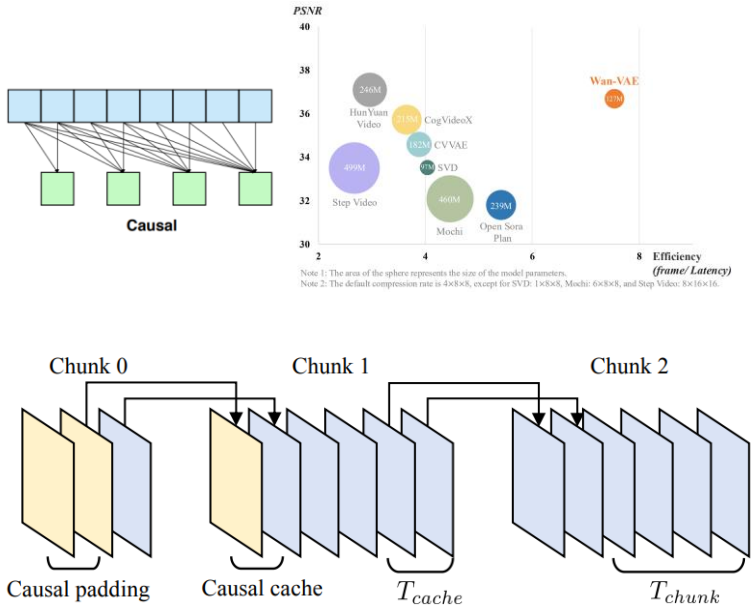
端到端 14.36s

通过在步数蒸馏中融合NVFP4线性层量化训练并结合稀疏注意力机制，实现加速35倍

视觉与图像生成：LightX2V—算法创新

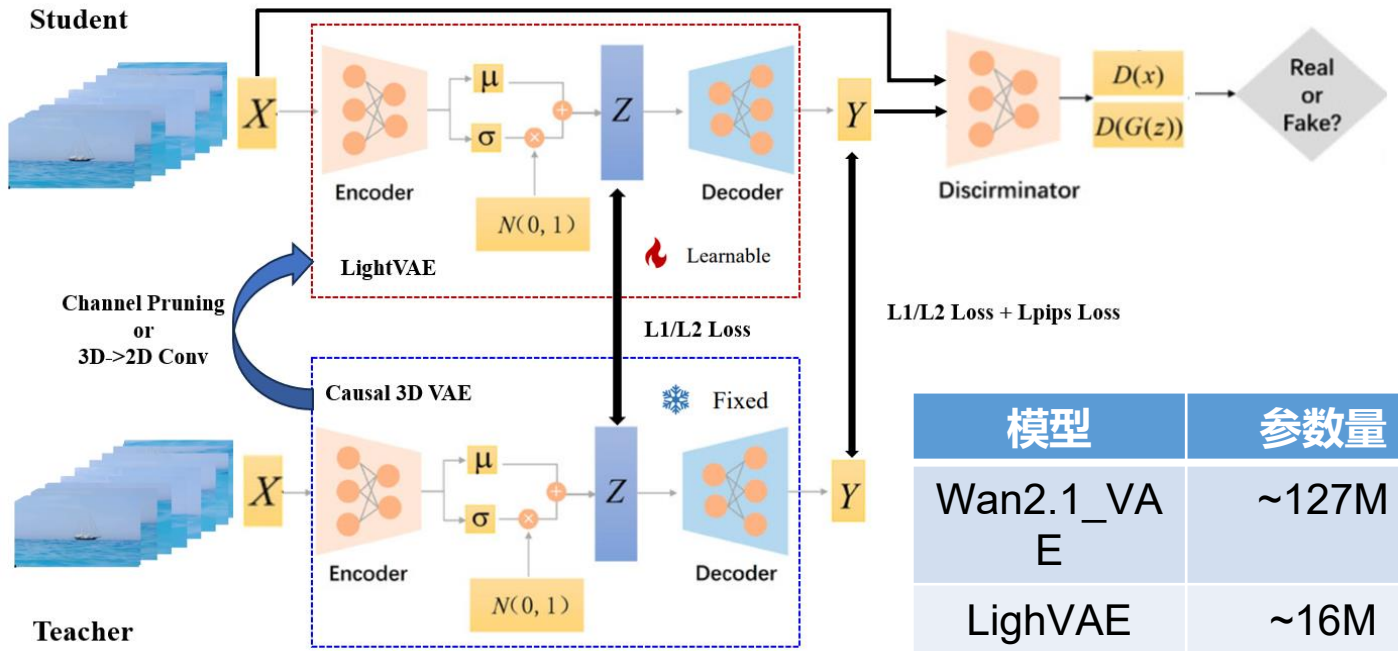
LightVAE系列:更小、更快、更省显存的高精度自编码器

瓶颈 解码时间长/显存高



解码长视频，视频帧Cache大

创新点 Causal 3D Conv裁剪/非Causal结构+蒸馏

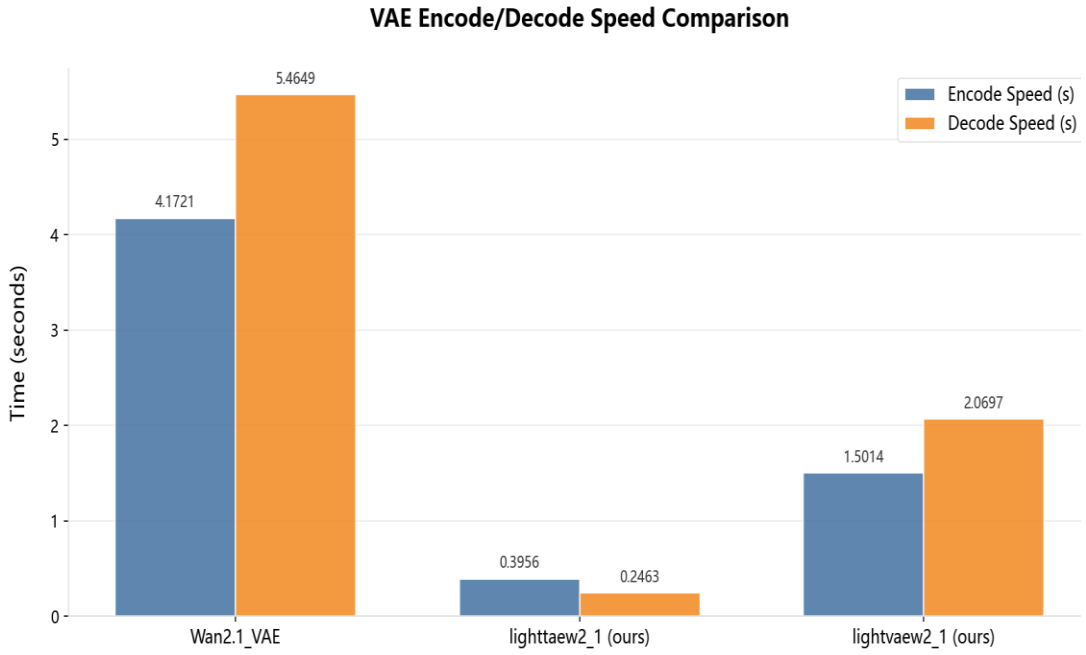


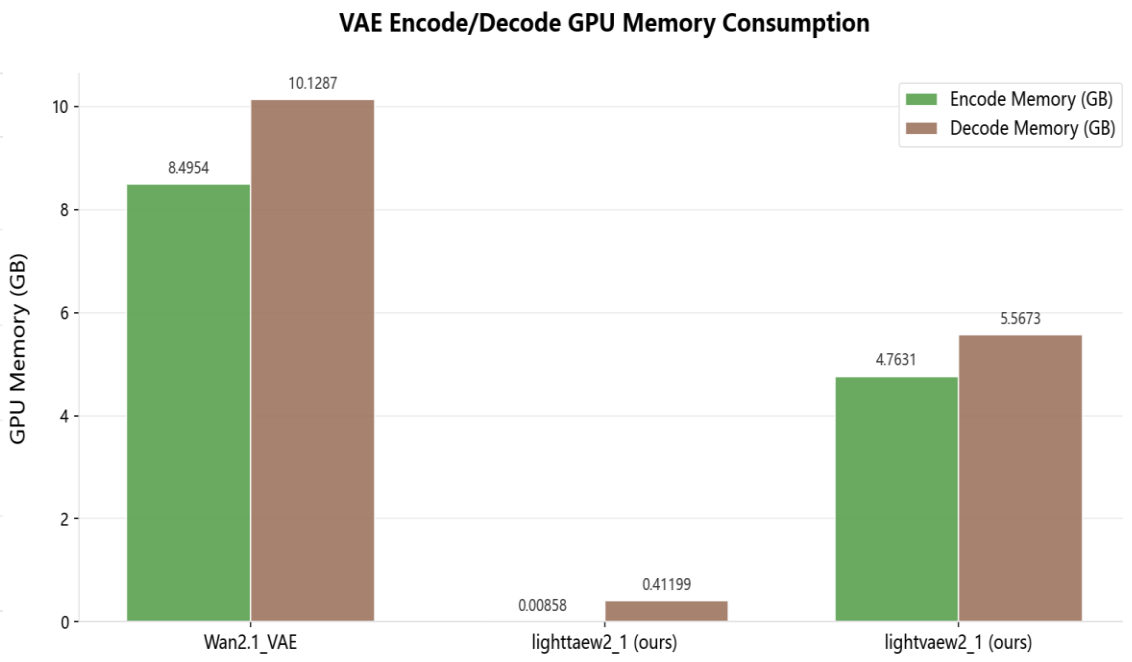
模型	参数量
Wan2.1_VAE	~127M
LighVAE	~16M
LightTAE	~11M

视觉与图像生成：LightX2V—算法创新

效果：速度显著提升，显存需求显著降低

速度与显存对比， 5s 81帧视频





解码加速20+倍，显存降低20+倍

视觉与图像生成：LightX2V—算法创新

效果：肉眼几乎无损

生成视频效果对比

Wan2.2_VAE

(Wan2.2官方VAE)

TAE

(社区开源轻量级AE, 视频背景有明显噪点)

LightTAE

(自研轻量级AE)

肉眼几乎看不到明显损失

视觉与图像生成：LightX2V—工程创新

针对消费级显卡的极致并行优化方案

瓶颈 **高通信和低互联**

单设备Ulysses通信量
4*token量*特征维度/卡数

token量巨大&低速互联

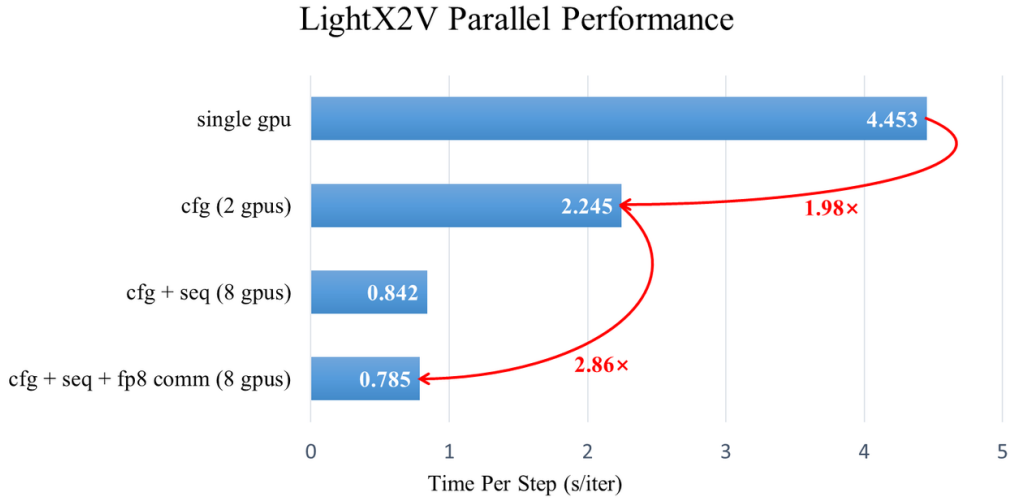
创新点 **通信量压缩+通信融合+负载均衡+计算重叠**

- FP8通信：将bf16的hidden states压缩成fp8进行传输，通信后解压回bf16，减少通信量
- qkv融合通信：大块通信，降低通信次数
- 负载均衡通信：轮询调度替代all2all，减少4090这种消费级显卡的pcie端口瞬时拥塞
- head级流水线并行：head粒度上进行计算和通信的重叠

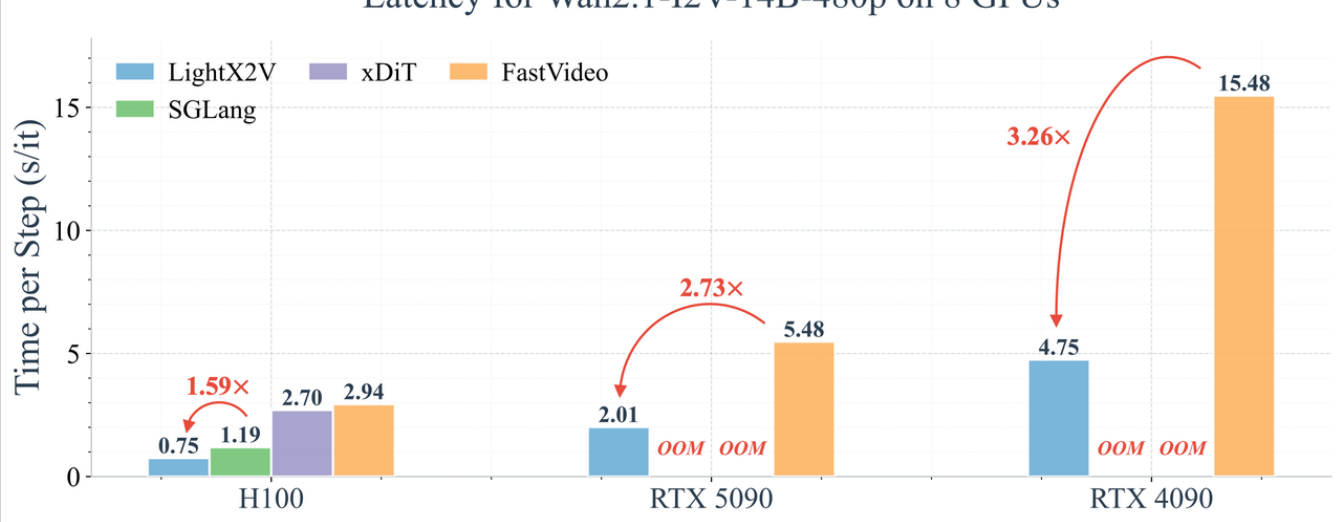
提升消费级显卡的并行效率

视觉与图像生成：LightX2V—工程创新

在低带宽的消费级显卡上实现较高通信加速比



混元1.5在5090上的并行



5090上不同框架间8卡并行对比

视觉与图像生成：LightX2V—工程创新

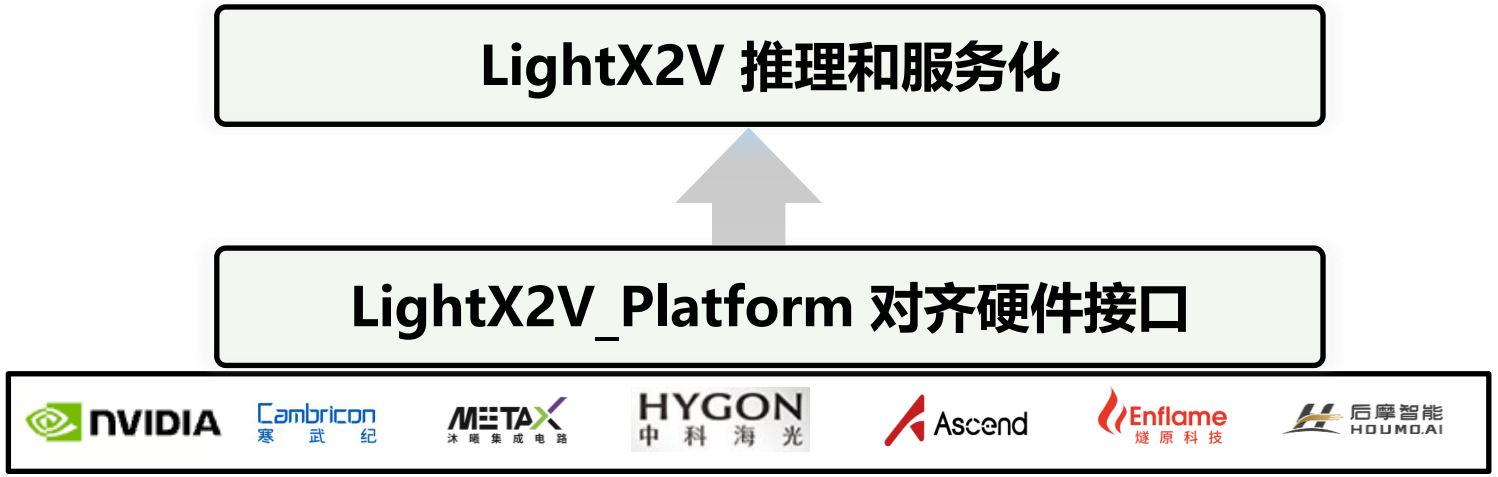
抽象设计硬件接入模式，支撑各种接口形态全国产化适配

瓶颈 **芯片接口混乱**

- torch.cuda
- torch.mlu
- torch.npu
- torch_dtu
- ...

国产芯片多样，接口杂乱

创新点 **LightX2V_Platform**



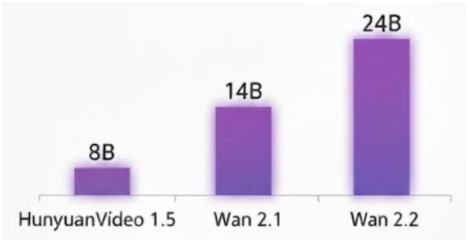
将LightX2V本身和硬件接口剥离
支持新硬件不再需要关注推理框架上层逻辑

视觉与图像生成：LightX2V—工程创新

三级offload，8G显存4060可高效运行

瓶颈 显存/内存受限

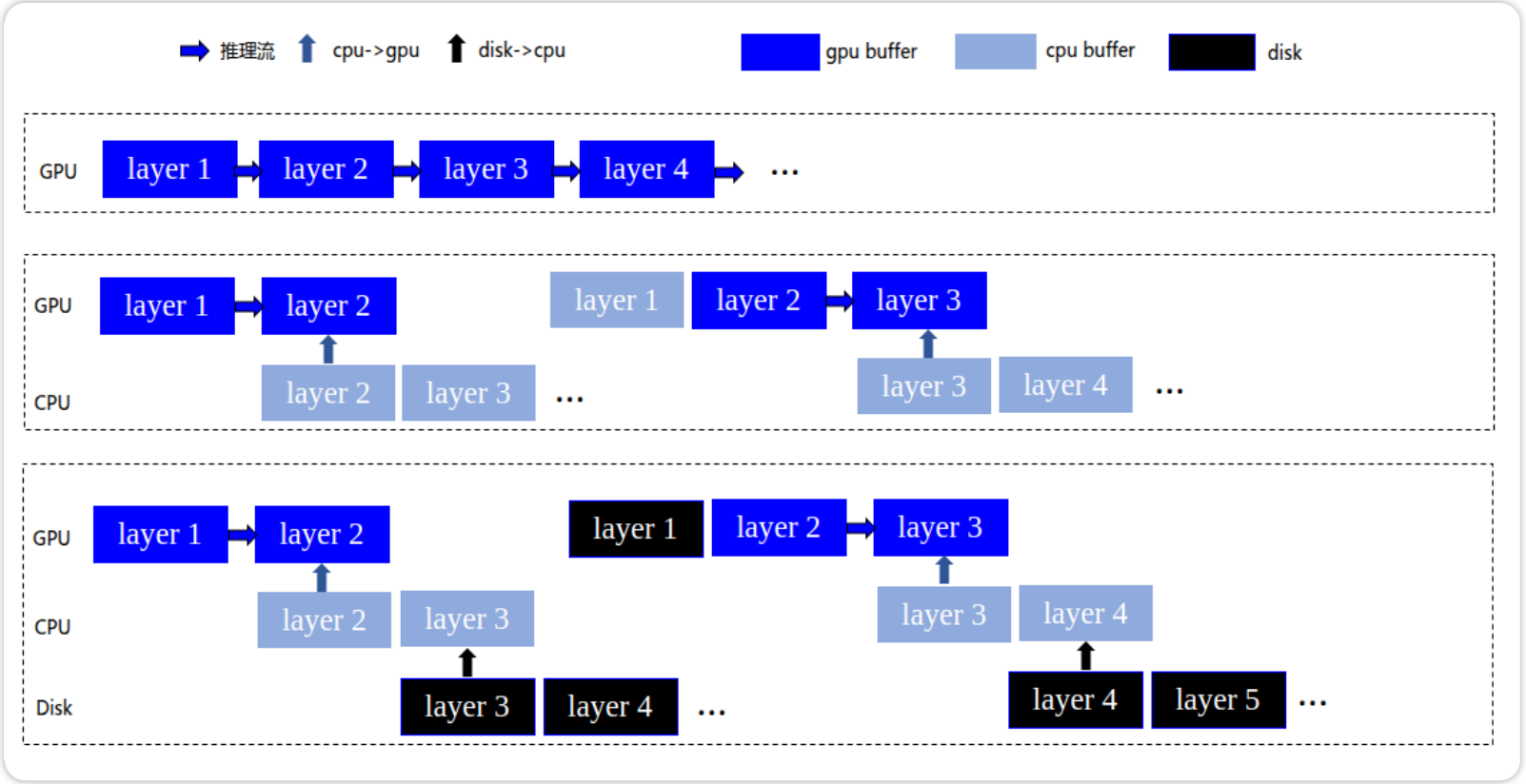
参数规模



Token量

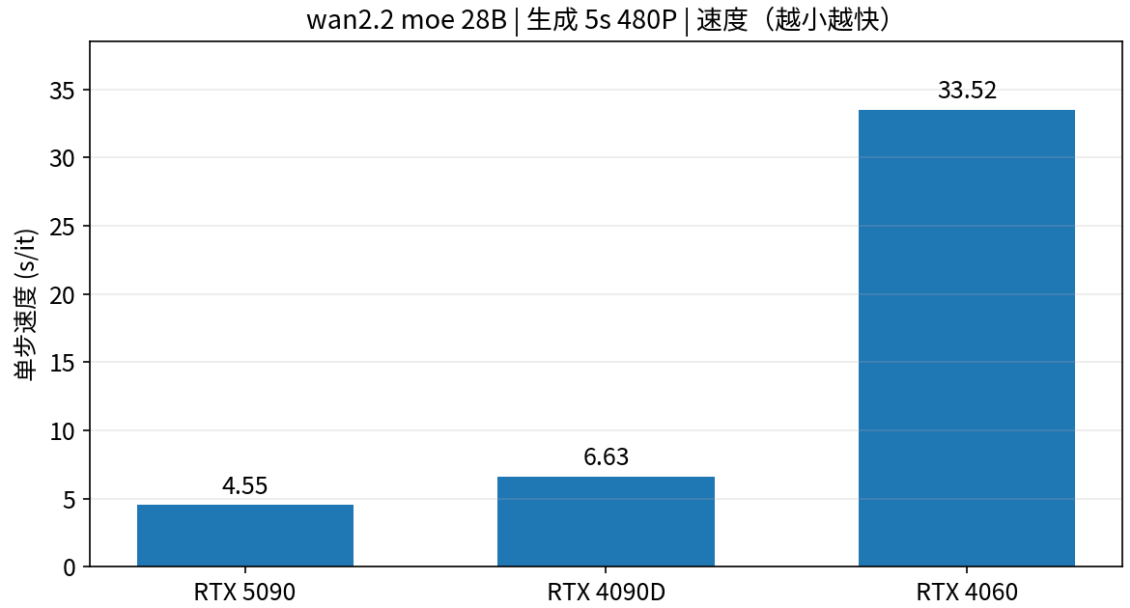
参数规模大，Token数量多

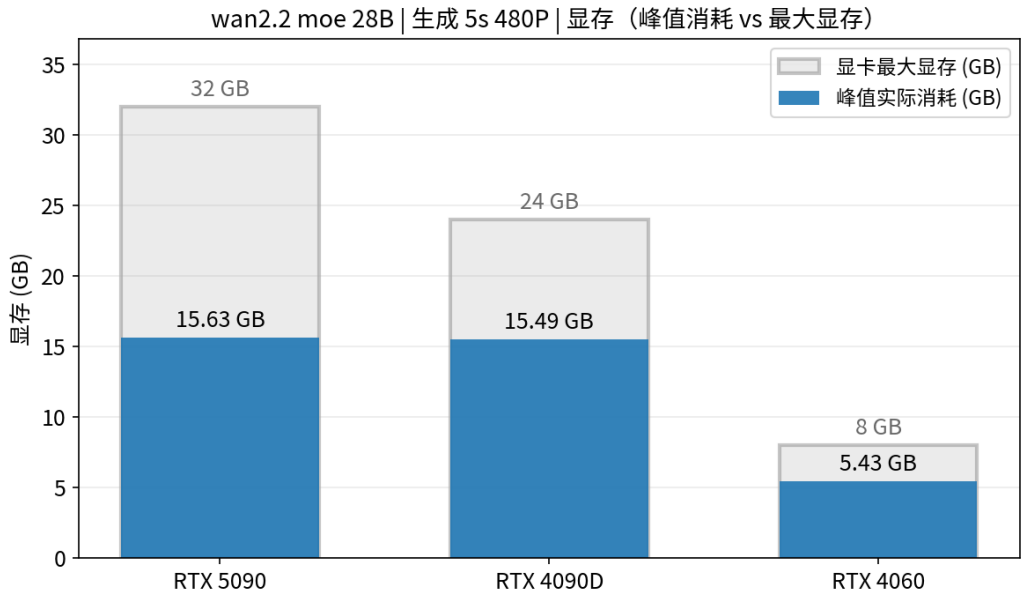
创新点 多级offload系统



视觉与图像生成：LightX2V—工程创新

三级offload，8G显存4060可高效运行





效果：最低8G显存显卡+16G系统内存，实现Wan2.2 MOE 28B模型，480P/720P视频高效生成

视觉与图像生成：LightX2V

整体效果-视频生成：Wan2.1-I2V-14B-480P

📊 Cross-Framework Performance Comparison (H100)

Framework	GPUs	Step Time	Speedup
Diffusers	1	9.77s/it	1x
xDiT	1	8.93s/it	1.1x
FastVideo	1	7.35s/it	1.3x
SGL-Diffusion	1	6.13s/it	1.6x
LightX2V	1	**5.18s/it**	**1.9x** 🚀
FastVideo	8	2.94s/it	1x
xDiT	8	2.70s/it	1.1x
SGL-Diffusion	8	1.19s/it	2.5x
LightX2V	8	**0.75s/it**	**3.9x** 🚀

📊 Cross-Framework Performance Comparison (RTX 4090D)

Framework	GPUs	Step Time	Speedup
Diffusers	1	30.50s/it	1x
FastVideo	1	22.66s/it	1.3x
xDiT	1	OOM	OOM
SGL-Diffusion	1	OOM	OOM
LightX2V	1	**20.26s/it**	**1.5x** 🚀
FastVideo	8	15.48s/it	1x
xDiT	8	OOM	OOM
SGL-Diffusion	8	OOM	OOM
LightX2V	8	**4.75s/it**	**3.3x** 🚀

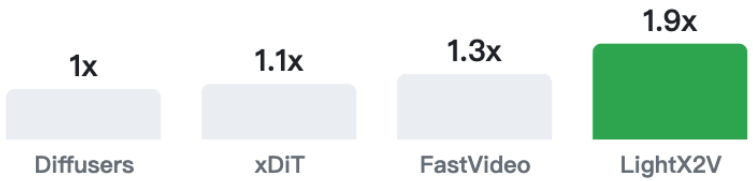
视觉与图像生成：LightX2V

整体效果-视频生成：Wan2.1-I2V-14B-480P

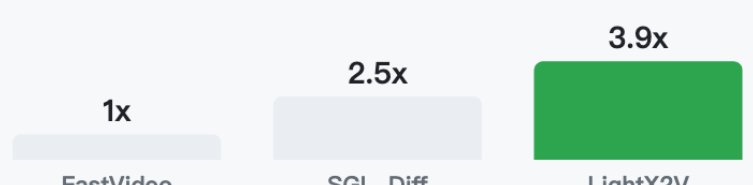
性能表现

H100 • Wan2.1–I2V

单卡性能 1.9倍 加速



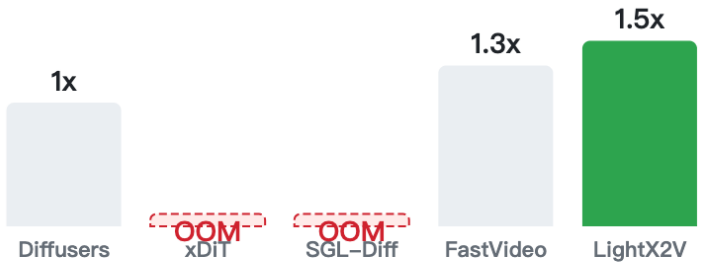
8卡分布式 0.75s / it



消费级显卡友好

测试环境：RTX 4090D (24GB)

竞品 vs LightX2V



其他框架
OOM
在 4090D 上失败

LightX2V
运行成功
流畅运行 @ 20.2s/it

无需高端集群

视觉与图像生成：LightX2V

整体效果-图像生成：Qwen-image

H100-Qwen-image-edit-单卡

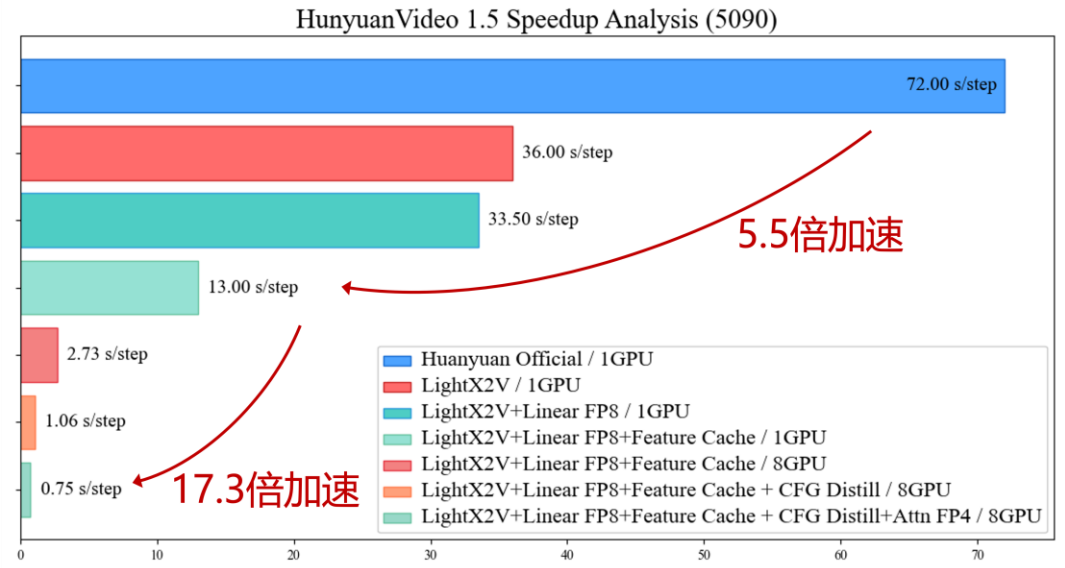
	单步时间
SGL-Diffusion(fa3)	0.85s/it (50step)
vllm-omni(fa3)	0.81s/it (50step)
LightX2V(fa3)	0.75s/it (50step)

H100-Qwen-image-edit-2509-单卡

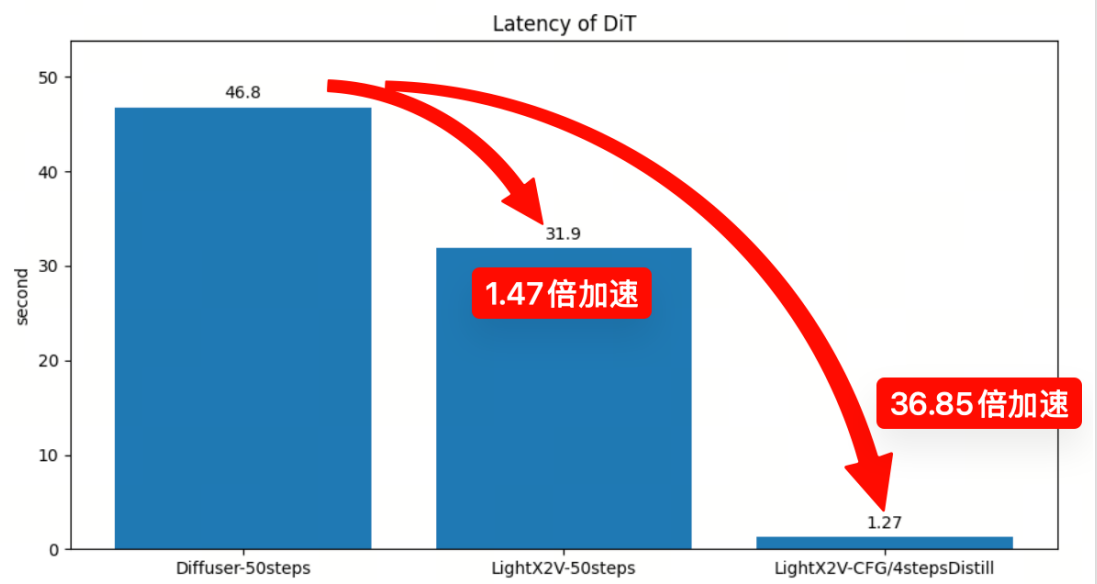
	单步时间
SGL-Diffusion(fa3)	0.73s/it (40step)
LightX2V(fa3)	0.63s/it (40step)

视觉与图像生成：LightX2V

整体效果：day 0接入模型性能优于行业框架



HunyuanVideo 1.5



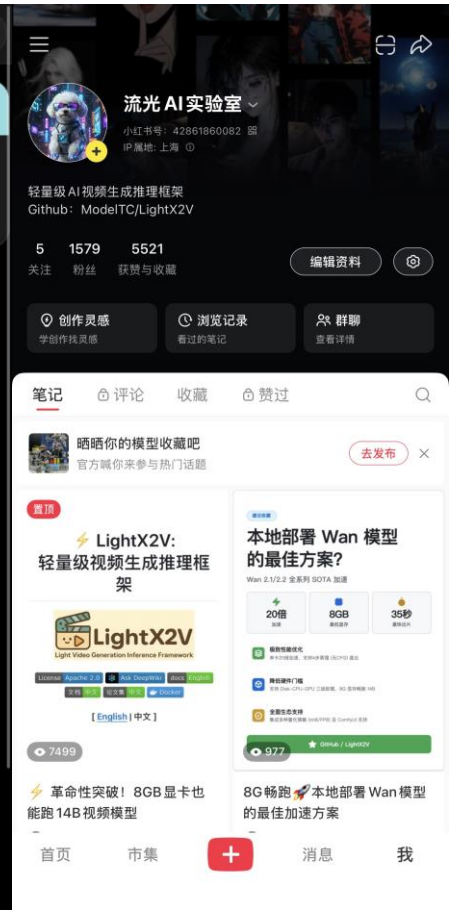
Qwen Image Edit 2511

持续不断和行业内优秀基模团队进行Day 0支持合作

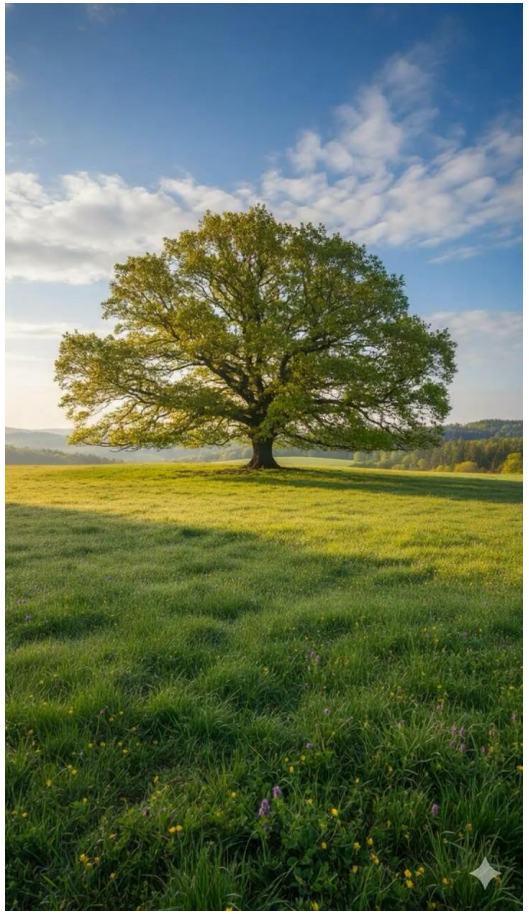
视觉与图像生成：LightX2V

体验入口：支持文生图、图生图、图片编辑、文生视频、图生视频、首尾帧生视频、角色替换、双人播客、口型驱动数字人等功能

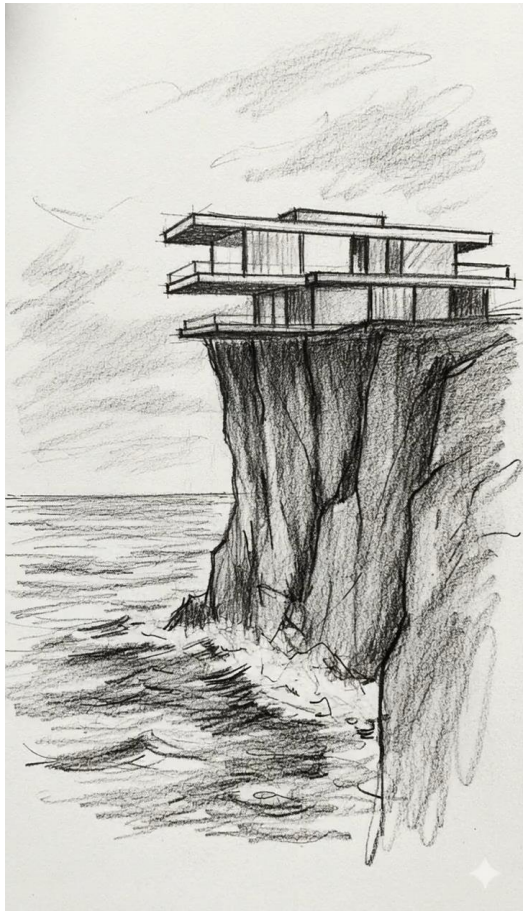
体验入口



小红书分享



示例视频

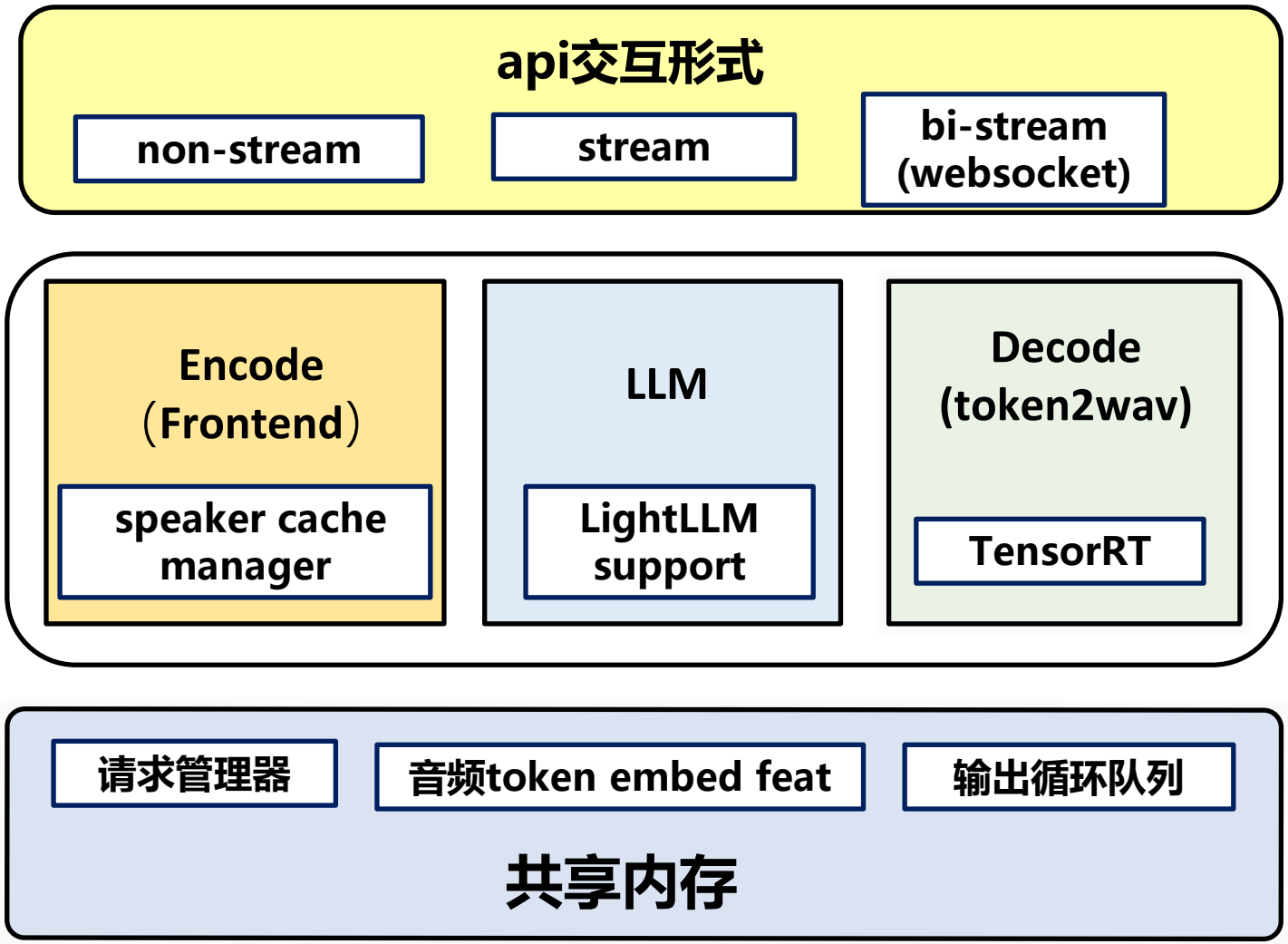


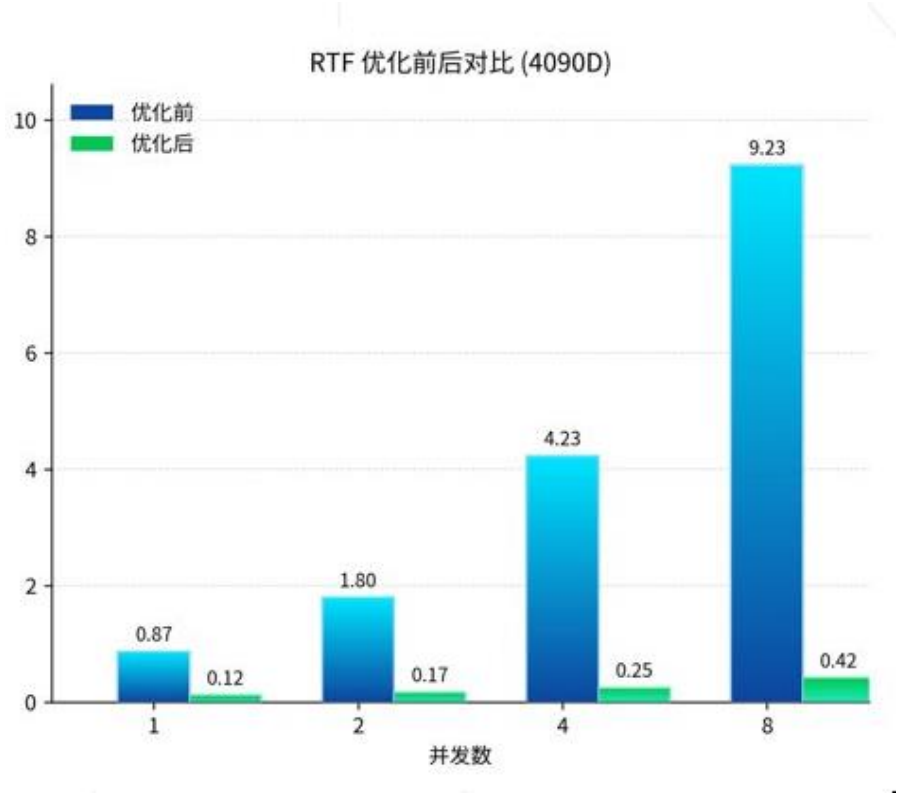
示例视频

示例视频

从文本到语音：LightTTS–架构特性

Encode、LLM、Decode多模块独立运行

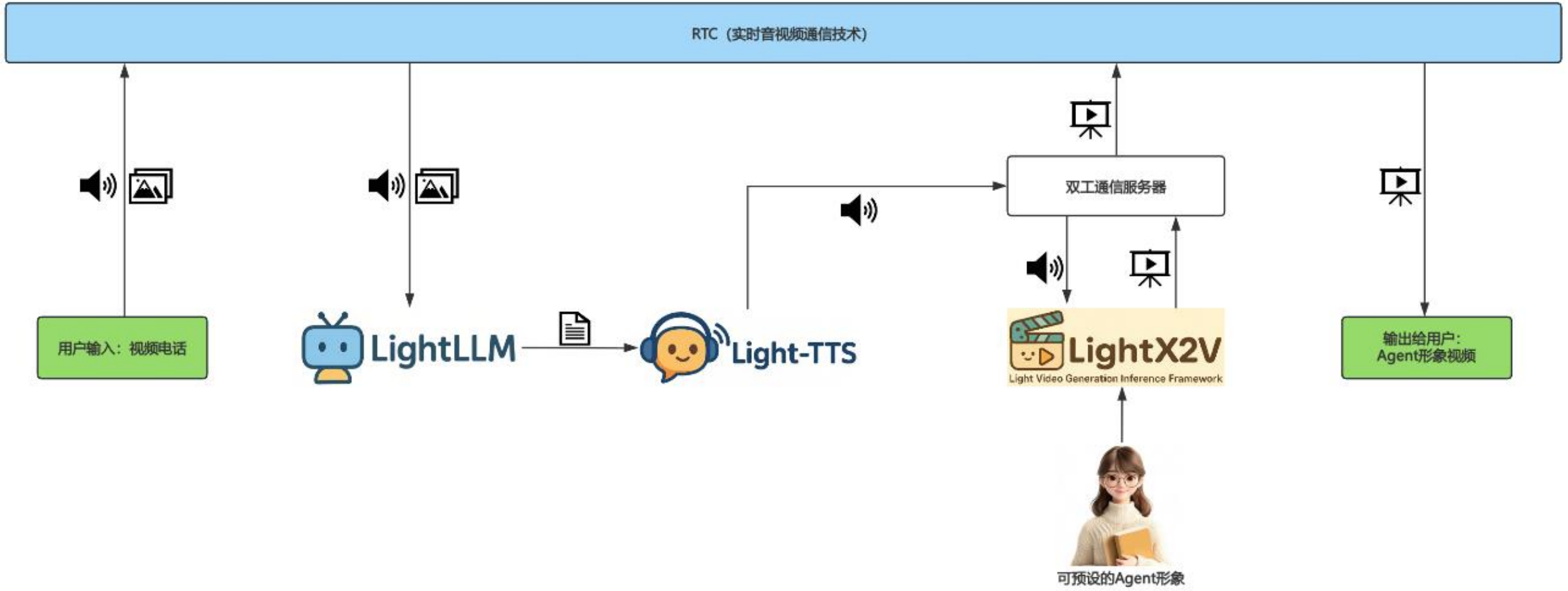




cosyvoice2模型速度测试 (vs官方库)

全模态交互：打造强实时、可交互的系统

音画驱动的实时视频对话推理系统

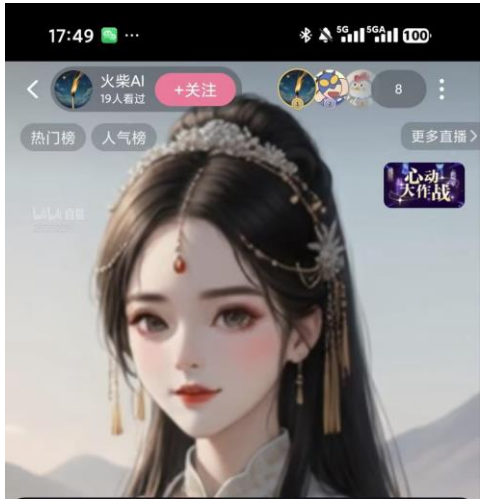


全模态交互：打造强实时、可交互的系统

应用场景：实时对话聊天室、实时交互世界模型游戏、直播

1v1聊天室

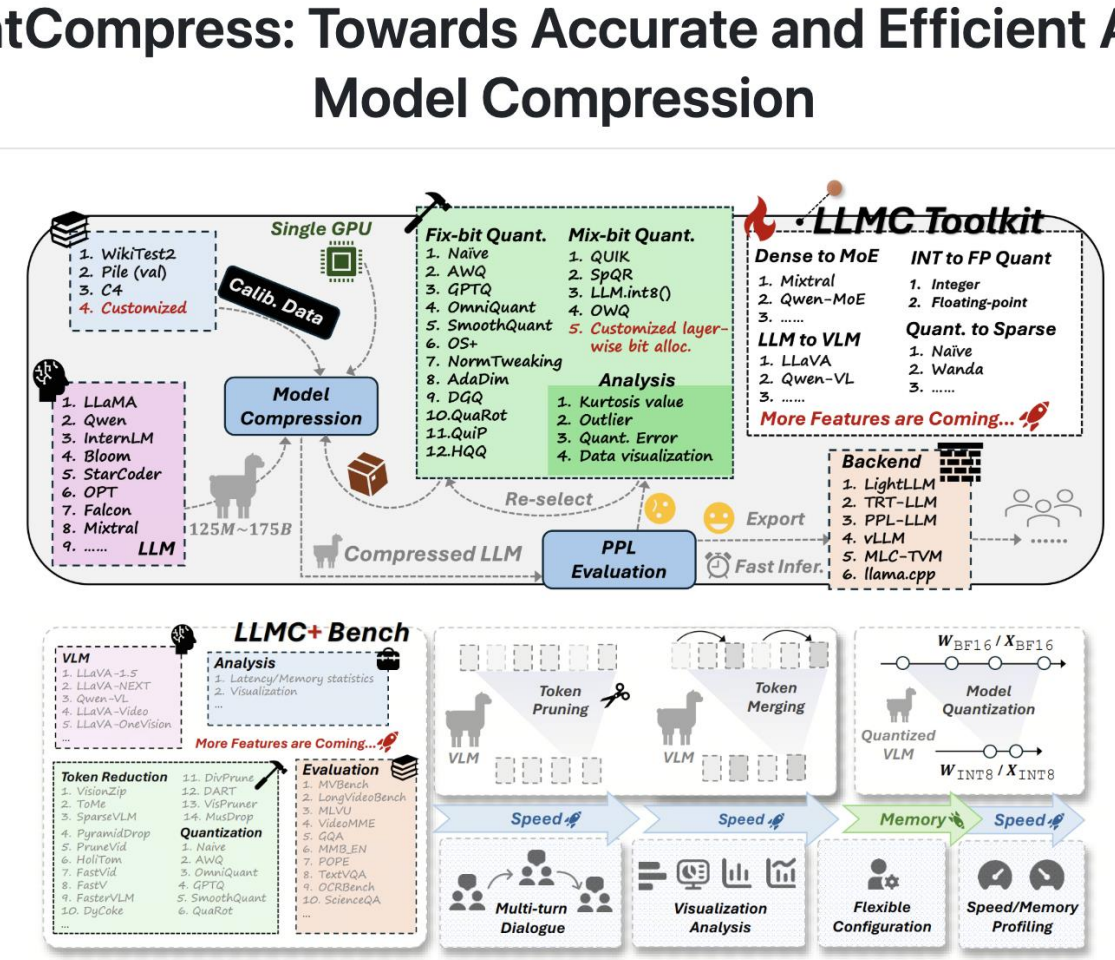
双人（猫？）播客



实时交互直播

多模态压缩：LightCompress—架构特性

支持LLM/VLM/图像和视频生成多种结构的统一压缩框架



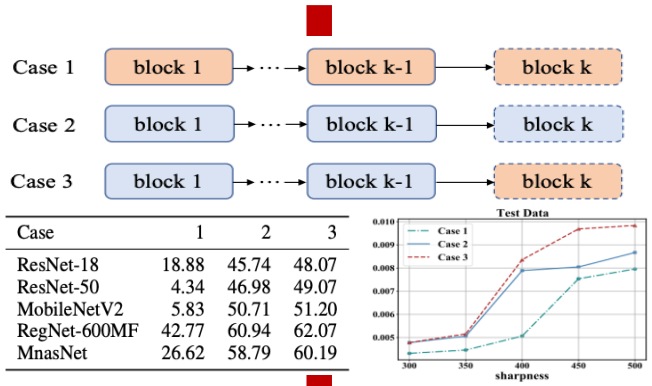
License Apache 2.0 | Ask DeepWiki | LLMC 2405.06001 | LLMC+ 2508.09981 | chat 3 online | QQ | docs English | 文档 中文

多模态压缩：LightCompress—算法创新

视觉模块压缩：基于随机失活激活的极低比特量化方法

瓶颈 **数据过拟合坍塌**

极低比特精度坍塌



Case	1	2	3
ResNet-18	18.88	45.74	48.07
ResNet-50	4.34	46.98	49.07
MobileNetV2	5.83	50.71	51.20
RegNet-600MF	42.77	60.94	62.07
MnasNet	26.62	58.79	60.19

离线重建

数据受限导致泛化性差

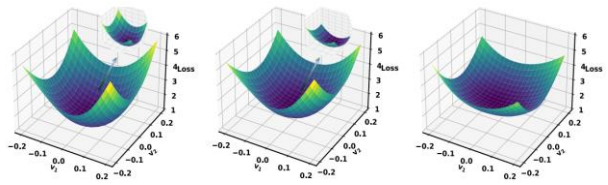
创新点 **提出平坦性分析理论，开创随机失活优化方法**

误差转移思想

$$W(a \odot \begin{bmatrix} 1+u_1(x) \\ 1+u_2(x) \\ \dots \\ 1+u_n(x) \end{bmatrix}) = (W \odot \begin{bmatrix} 1+u_1(x) & 1+u_2(x) & \dots & 1+u_n(x) \\ 1+u_1(x) & 1+u_2(x) & \dots & 1+u_n(x) \\ & \dots & & \\ 1+u_1(x) & 1+u_2(x) & \dots & 1+u_n(x) \end{bmatrix})a.$$

平坦优化曲面

$$\Delta \quad \text{QDROP}: u = \begin{cases} 0 & \text{with probability } p \\ \frac{\hat{a}}{a} - 1 & \text{with probability } 1-p \end{cases}.$$



快速准确平坦重建

Wei et al., QDrop: Randomly Dropping Quantization for Extremely Low-bit Post-Training Quantization, ICLR 2022

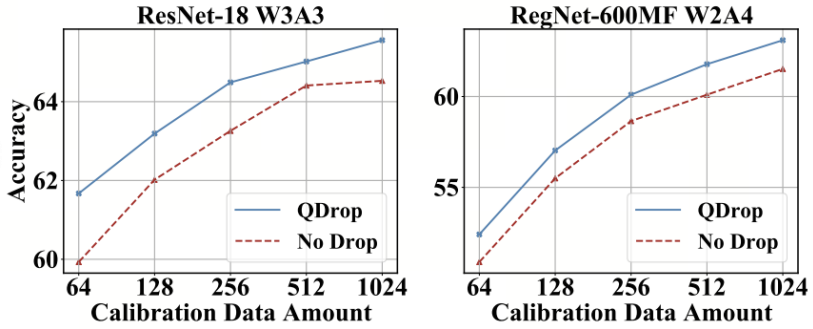
多模态压缩：LightCompress—算法创新

效果：在2/4比特场景提升达5%以上

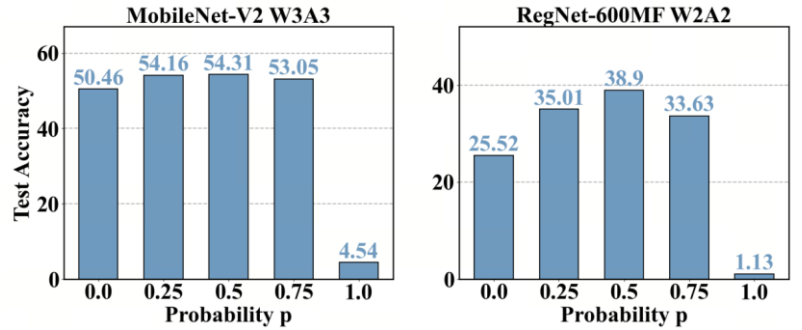
方法	比特	Res18	Res50	MNV2	Reg600M	Reg3.2G
全精度	32/32	71.06	77.00	72.49	73.71	78.36
损失感受量化[84]	4/4	60.30	70.00	49.70	57.71*	55.89*
适应性量化法[85]	4/4	**69.60**	**75.90**	47.16*	-	-
比特划分法[86]	4/4	67.56	73.71	-	-	-
上下学习取整法[11]*	4/4	67.96	73.88	61.52	68.20	73.85
随机失活激活值量化	4/4	69.10	75.03	**67.89**	**70.62**	**76.33**
上下学习取整法 †*	4/4	69.36	74.76	64.33	-	-
块重构方法 †[14]	4/4	69.60	75.05	66.57	68.33	74.21
随机失活激活值量化 †	4/4	**69.62**	75.45	**68.84**	**71.18**	**76.66**
损失感受量化 *	2/4	0.18	0.14	0.13	0.17	0.12
适应性量化法 *	2/4	0.11	0.12	0.15	-	-
上下学习取整法 *	2/4	62.12	66.11	36.31	57.00	63.89
随机失活激活值量化	2/4	**64.66**	**70.08**	**52.92**	**63.10**	**70.95**
上下学习取整法 †*	2/4	64.14	68.40	41.52	59.27	65.33
块重构法 †	2/4	64.80	70.29	53.34	59.31	67.15
随机失活激活值量化 †	2/4	**65.25**	**70.65**	**54.22**	**63.80**	**71.70**

在2比特上取得最优效果

数据



概率



对校准数据选择和随机概率大小不敏感

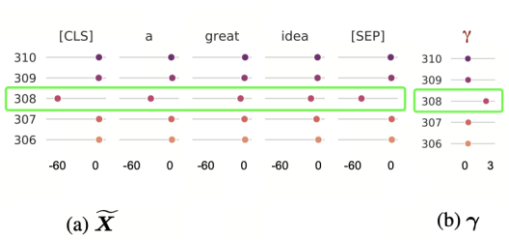
Wei et al., QDrop: Randomly Dropping Quantization for Extremely Low-bit Post-Training Quantization, ICLR 2022

多模态压缩：LightCompress—算法创新

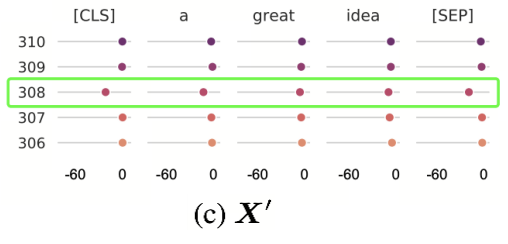
语言模块压缩：基于伽马迁移的大语言模型量化方法

瓶颈 离群值放大效应

伽马变换



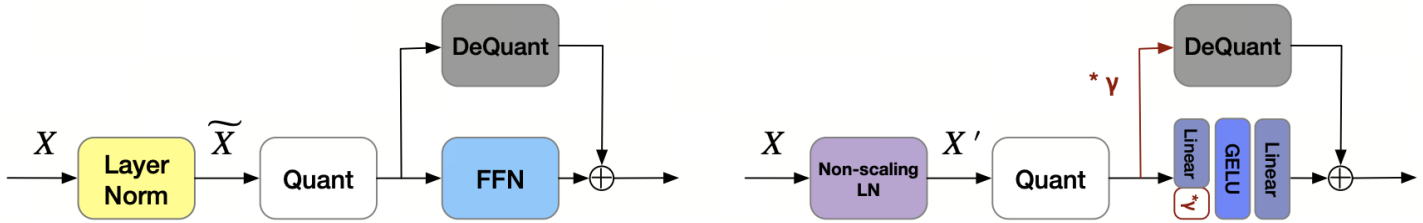
离群放大



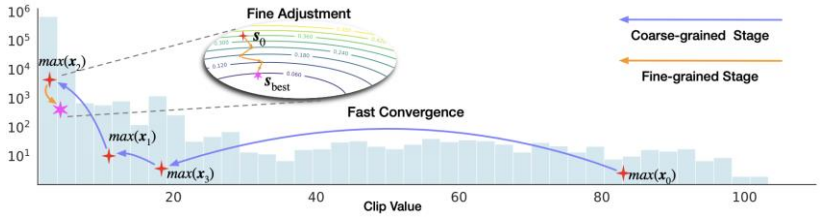
离群值放大导致精度崩溃

创新点 提出了基于伽马迁移的离群信息保持方法

Δ



$$\boldsymbol{W}(\boldsymbol{x} \odot \begin{bmatrix} \gamma_1 \\ \gamma_2 \\ \dots \\ \gamma_n \end{bmatrix}) = (\boldsymbol{W} \odot \begin{bmatrix} \gamma_1 & \gamma_2 & \dots & \gamma_n \\ \gamma_1 & \gamma_2 & \dots & \gamma_n \\ \dots & & & \\ \gamma_1 & \gamma_2 & \dots & \gamma_n \end{bmatrix})$$

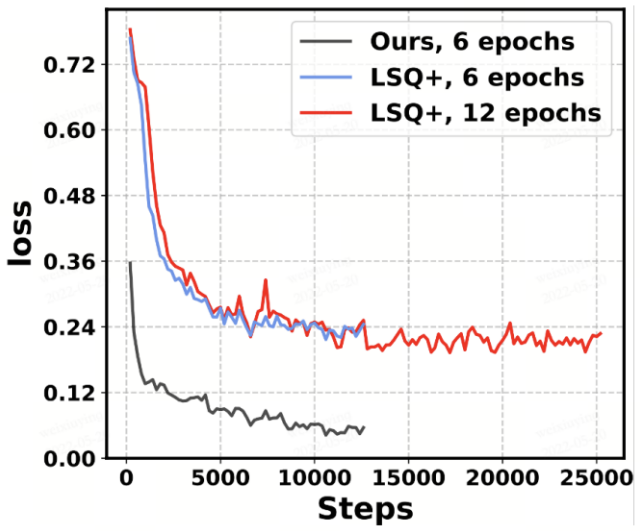


延后伽马因子缩放完成离群值抑制并实现信息保持

Wei et al., Outlier Suppression: Pushing the Limit of Low-bit Transformer Language Models, NeurIPS 2022

多模态压缩：LightCompress—算法创新

效果：实现INT6/INT8方案的重建速度提升和稳定精度



离群信息保持效果提升，
重建速度提升>2x

Method	Bits (W-E-A)	CoLA (Matt.)	MNLI (acc m/mm)	MRPC (f1/acc)	QNLI (acc)	QQP (f1/acc)	RTE (acc)	SST-2 (acc)	STS-B (Pear./Spear.)	Avg.
BERT	32-32-32	59.60	84.94/84.76	91.35/87.75	91.84	87.82/90.91	72.56	93.35	89.70/89.28	83.83
MinMax	8-8-8	57.08	82.77/83.47	89.90/85.78	90.76	87.84/90.74	69.68	92.78	86.83/88.56	82.28
OMSE [28]	8-8-8	57.15	84.04/84.29	90.10/85.78	91.12	87.64/90.54	72.20	93.23	87.90/88.65	82.90
Ours	8-8-8	**61.64**	**84.38/84.53**	**91.44/87.75**	**91.49**	**87.92/90.77**	**72.20**	**93.81**	**89.23/89.01**	**83.96**
OMSE	6-6-6	35.44	74.00/73.30	81.54/76.47	84.66	76.07/82.12	64.26	86.27	85.57/86.05	73.52
Percentile [29]	6-6-6	37.32	72.40/71.69	85.09/79.90	79.37	72.58/80.19	61.73	87.27	86.38/87.29	72.93
EasyQuant [40]	6-6-6	38.16	75.82/75.66	82.51/77.45	84.94	75.31/81.81	65.34	87.27	85.50/86.33	74.49
Ours	6-6-6	**54.40**	**82.02/81.69**	**87.45/83.33**	**89.82**	**84.69/88.94**	**70.76**	**91.86**	**88.65/88.55**	**81.19**
PEG [26] ♣	8-8-8	59.43	81.25	88.53	**91.07**	**89.42**	69.31	92.66	87.92	82.45
Ours ♣	8-8-8	**59.83**	**82.93/82.59**	**91.33/87.99**	90.02	87.45/90.34	**70.04**	**92.66**	**88.42/88.81**	**82.81**
PEG ♣	6-6-6	9.46	32.44/32.77	83.64/78.43	49.46	29.93/62.97	**70.76**	90.14	52.79/53.22	54.11
Ours ♣	6-6-6	**42.27**	**78.54/78.32**	**85.33/81.13**	**85.36**	**78.47/84.66**	68.59	**91.74**	**87.33/87.19**	**77.31**

有效抑制异常值范围，首次实现INT6精度可用

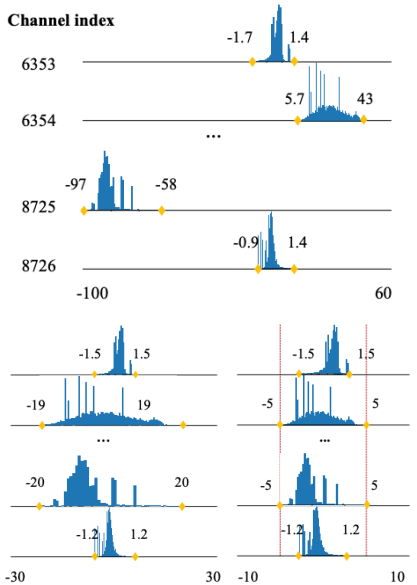
Wei et al., Outlier Suppression: Pushing the Limit of Low-bit Transformer Language Models, NeurIPS 2022

多模态压缩：LightCompress—算法创新

语言模块压缩：基于等价变换的大语言模型量化方法

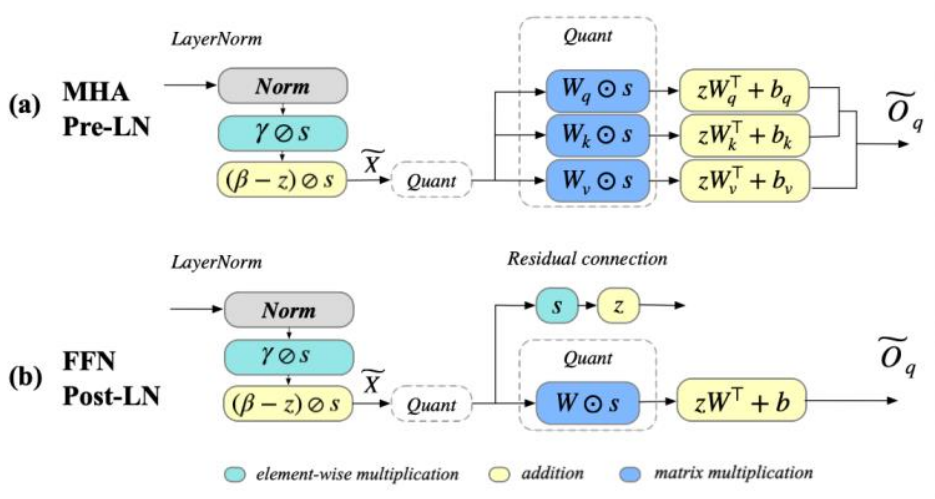
瓶颈 量化不友好分布

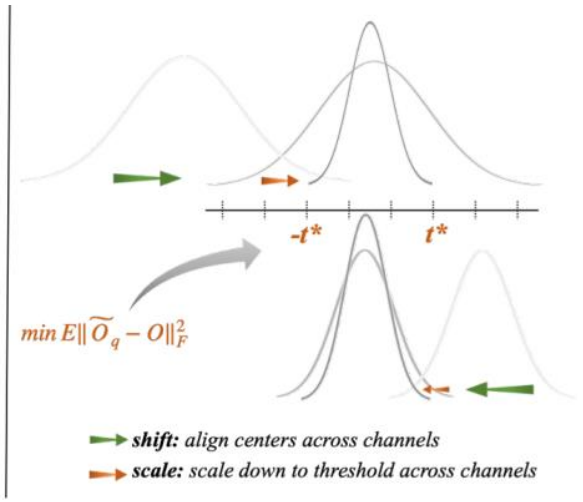
通道偏移 尺度不均



分布不友好导致精度崩溃

创新点 提出了基于等价变换的离群信息保持方法



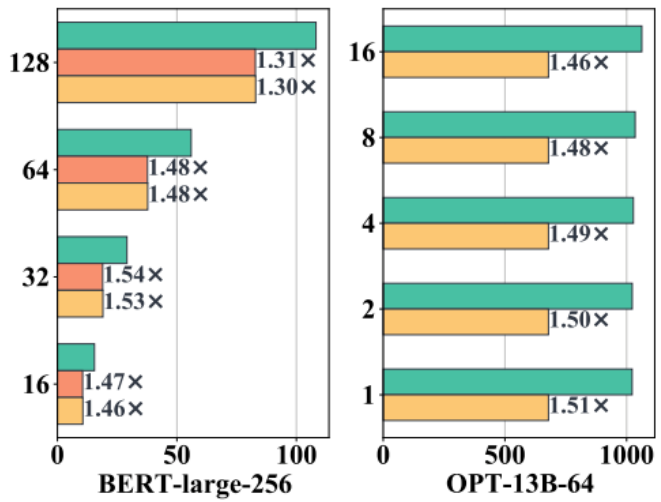


平移+缩放完成离群值抑制并实现信息保持

Wei et al., Outlier Suppression+: Accurate quantization of large language models by equivalent and effective shifting and scaling, EMNLP 2023

多模态压缩：LightCompress—算法创新

效果：实现INT4方案的速度提升和稳定精度



设计低比特优化算子，实现1.5倍速度提升

Model	Method	PIQA (↑)			Winogrande (↑)			HellaSwag (↑)			WikiText2 (↓)		
		FP16	INT6	INT4	FP16	INT6	INT4	FP16	INT6	INT4	FP16	INT6	INT4
LLaMA-1-7B	MinMax	77.37	77.26	55.98	66.93	66.54	49.64	72.99	71.78	32.28	5.68	6.00	473.97
	SmoothQuant		77.18	70.08		65.51	52.96		72.10	58.13		5.85	16.87
	OS+		**77.48**	**72.31**		**67.01**	**56.67**		**72.32**	**61.24**		**5.76**	**14.17**
LLaMA-1-13B	MinMax	79.05	78.56	50.65	70.09	69.53	50.28	76.22	75.26	26.34	5.09	5.58	3410.45
	SmoothQuant		78.45	66.49		**69.69**	51.78		75.20	58.95		5.25	56.75
	OS+		**78.73**	**75.03**		69.53	**61.17**		**75.74**	**67.21**		**5.22**	**18.95**
LLaMA-1-30B	MinMax	80.09	78.40	50.00	72.77	72.45	50.12	79.21	77.25	27.09	4.10	5.09	2959.15
	SmoothQuant		78.78	71.55		73.01	54.54		78.13	60.97		4.40	51.47
	OS+		**79.98**	**73.01**		**73.64**	**60.38**		**78.77**	**68.03**		**4.30**	**22.61**
LLaMA-1-65B	MinMax	80.85	77.58	50.27	77.11	69.46	49.33	80.73	78.72	24.59	3.56	5.25	14584.66
	SmoothQuant		78.40	65.02		74.30	51.14		78.57	59.78		3.77	19.37
	OS+		**80.47**	**74.43**		**75.14**	**61.72**		**79.76**	**67.65**		**3.65**	**9.33**

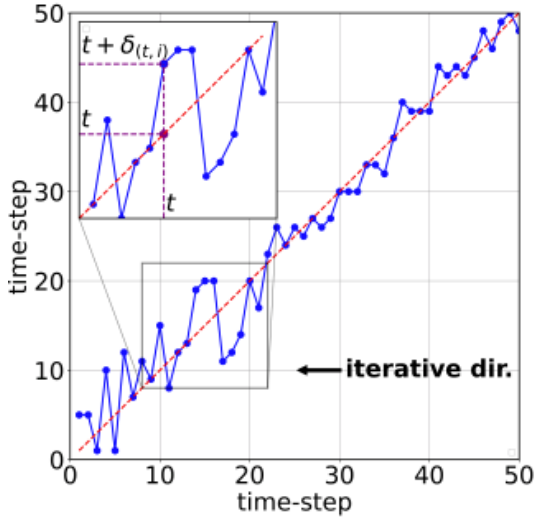
有效抑制异常值范围，首次实现INT4精度可用

Wei et al., Outlier Suppression+: Accurate quantization of large language models by equivalent and effective shifting and scaling, EMNLP 2023

多模态压缩：LightCompress—算法创新

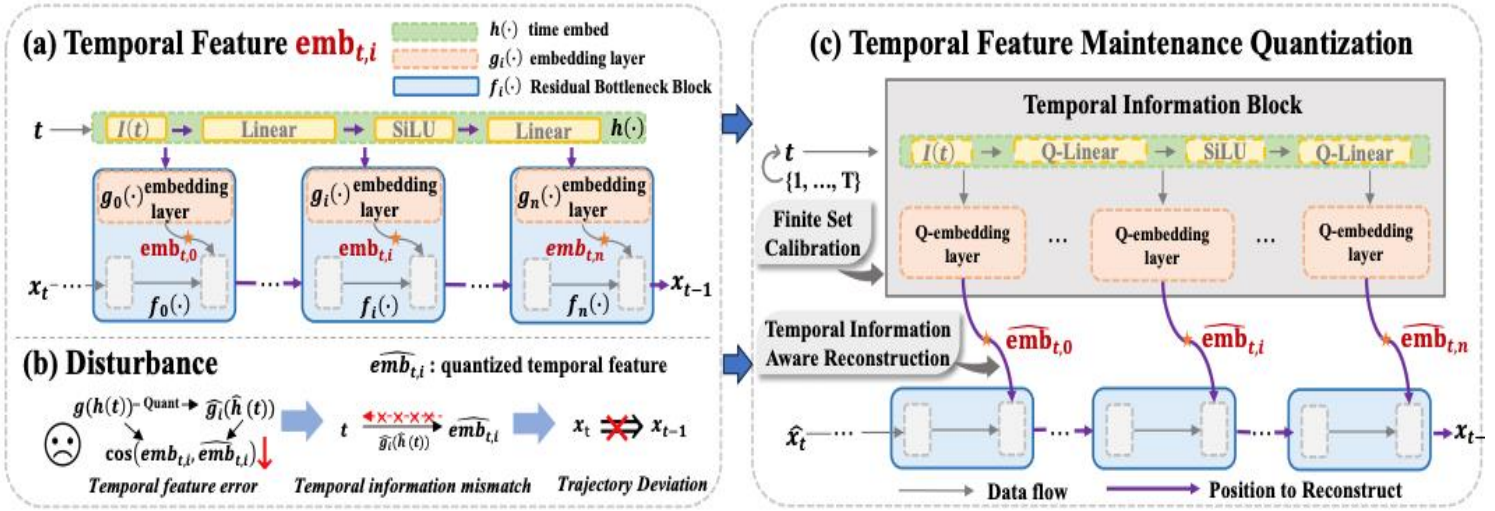
图像生成模块压缩：基于时序特征保持的扩散模型量化

瓶颈 时序信息扰动失配



多个时间步引入
累积扰动误差

创新点 首次提出时序特征保持思想



使用少量校准数据集收集时序特征，最小化时间步之间的扰动

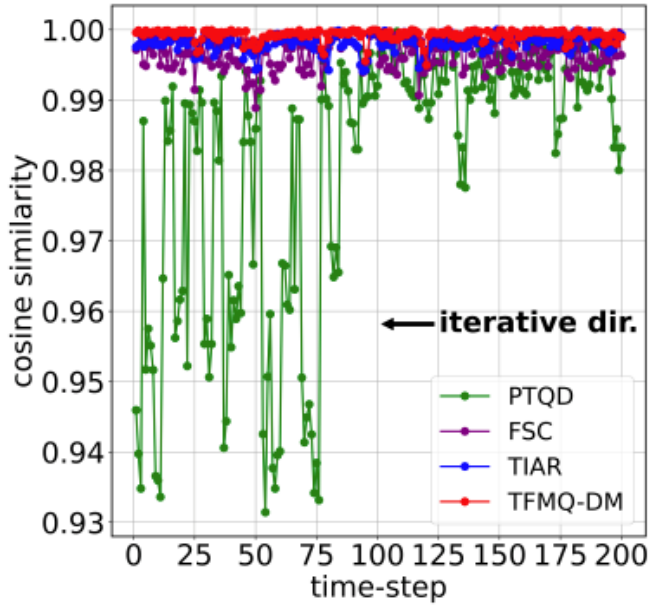
Huang et al., TFMQ-DM: Temporal Feature Maintenance Quantization for Diffusion Models, CVPR 2024 (highlight)
Huang et al., Temporal Feature Matters: A Framework for Diffusion Model Quantization, TPAMI 2025

多模态压缩：LightCompress—算法创新

效果：显著超越所有扩散模型量化方法，最多提升超6%，时序失配现象显著降低

Methods	Bits (W/A)	LSUN-Bedrooms 256 × 256		LSUN-Churches 256 × 256		CelebA-HQ 256 × 256		FFHQ 256 × 256	
		FID↓	sFID↓	FID↓	sFID↓	FID↓	sFID↓	FID↓	sFID↓
Full Prec.	32/32	2.98	7.09	4.12	10.89	8.74	10.16	9.36	8.67
PTQ4DM* [42]	4/32	4.83	7.94	4.92	13.94	13.67	14.72	11.74	12.18
Q-Diffusion† [23]	4/32	4.20	7.66	4.55	11.90	11.09	12.00	11.60	10.30
PTQD* [10]	4/32	4.42	7.88	4.67	13.68	11.06	12.21	12.01	11.12
TFMQ-DM (Ours)	4/32	**3.60** (-0.60)	**7.61** (-0.05)	**4.07** (-0.48)	**11.41** (-0.49)	**8.74** (-2.32)	**10.18** (-1.82)	**9.89** (-1.71)	**9.06** (-1.24)
PTQ4DM* [42]	8/8	4.75	9.59	4.80	13.48	14.42	15.06	10.73	11.65
Q-Diffusion† [23]	8/8	4.51	8.17	4.41	12.23	12.85	14.16	10.87	10.01
PTQD [10]	8/8	3.75	9.89	4.89*	14.89*	12.76*	13.54*	10.69*	10.97*
TFMQ-DM (Ours)	8/8	**3.14** (-0.61)	**7.26** (-0.91)	**4.01** (-0.40)	**10.98** (-1.25)	**8.71** (-4.05)	**10.20** (-3.34)	**9.46** (-1.23)	**8.73** (-1.28)

精度近乎无损，超越所有baseline



明显降低时序累积误差

Huang et al., TFMQ-DM: Temporal Feature Maintenance Quantization for Diffusion Models, CVPR 2024 (highlight)
Huang et al., Temporal Feature Matters: A Framework for Diffusion Model Quantization, TPAMI 2025

开源生态

主流框架集成

- Qdrop　　扩散模型量化方法的基础模块
- OutlierSuppression　大模型量化的主流技术
- DSQ　　主流基准，谷歌学术引用排名第4
- BRECQ　集成至百度等商用模型部署平台

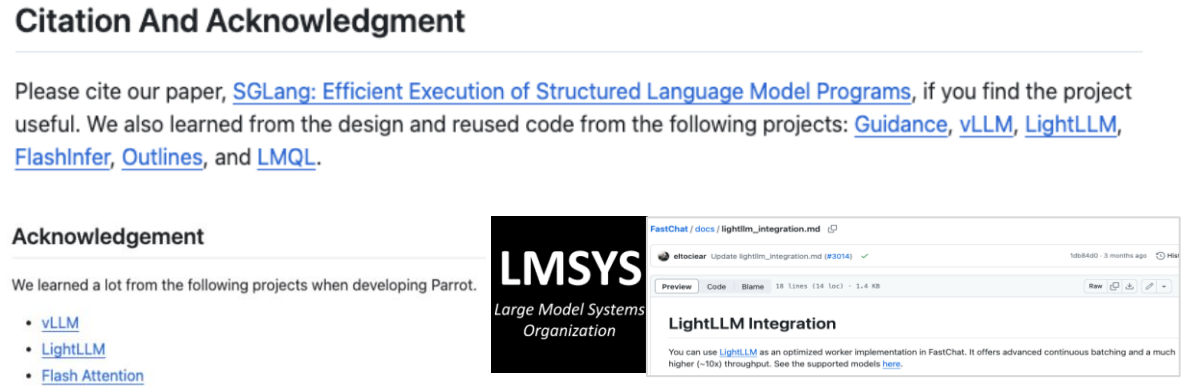
…

6+
深度学习框架

10+
国际著名企业

开源社区认可

- 7千+星标，被机器人公司、云厂商、硬件公司使用

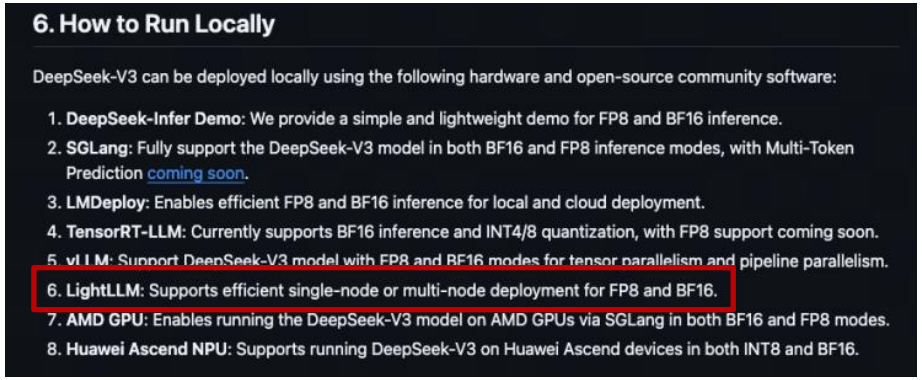


- 斯坦福、UC Berkeley、微软等多个项目借鉴，致谢表示受到启发

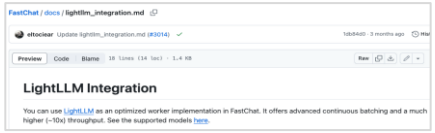
开源生态

LightLLM

- DeepSeek官方推荐
- 被20余机构的17个框架使用/致谢



LMSYS
Large Model Systems Organization

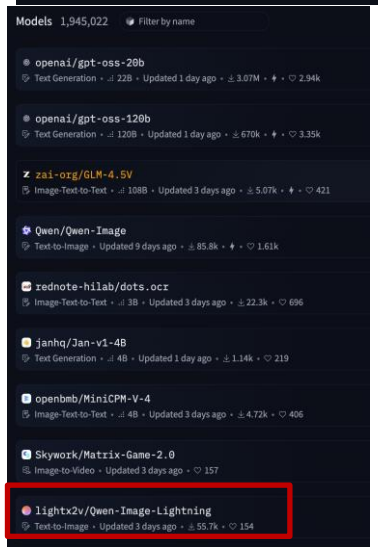


Stanford University

Microsoft

LightX2V步数蒸馏模型

- 上传首周登HuggingFace下载榜前十
- 累计下载量超530万
- WAN/HuanYuan Video官方推荐



代表模型	下载量
Wan2.2-Distill-Loras	~211M
QwenImageLightining	~304M

极客邦科技 2026 年会议规划

促进软件开发及相关领域知识与创新的传播

参会咨询

查看会议

4月 · 北京 · 1200人

QCon 全球软件开发大会

会议时间：4月16-18日

- Agentic Engineering
- AgentOps
- 下一代模型架构与推理优化
- AI 原生基础设施
- 知识工程实践
- AI 安全

6月 · 上海 · 1000人

AiCon 全球人工智能开发与应用大会

会议时间：6月26-27日

- AI Infra 系统工程
- 多 Agent 协作与实践
- 多模态融合
- 模型训练与推理创新
- 数据平台与特征服务

8月 · 深圳 · 1000人

AiCon 全球人工智能开发与应用大会

会议时间：8月21-22日

- Agentic AI
- 轻量化与高效推理
- 多模态应用
- AI + IoT 场景实践
- AI 工业化落地

10月 · 上海 · 1200人

QCon 全球软件开发大会

会议时间：10月22-24日

- AI Agent
- Vibe Coding
- 智能可观测
- 推理基建
- 模型攻防
- AI x 创造力

12月 · 北京 · 1000人

AiCon 全球人工智能开发与应用大会

会议时间：12月18-19日

- 大模型架构创新
- 多模态 AI 产业融合
- 具身智能
- AI for Science
- 大模型安全

InfoQ 极客传媒

InfoQ 极客传媒

THANKS

探索 AI 应用边界

Explore the limits of AI applications

AiCon
全球人工智能开发与应用大会

群聊：LightX2V交流群

该二维码7天内(12月27日前)有效，重新进入将更新