基于稀疏注意力机制的超长上下文推理加速方法

黄宇翔 计11

指导教师：刘知远

研究背景：长上下文推理具有挑战性

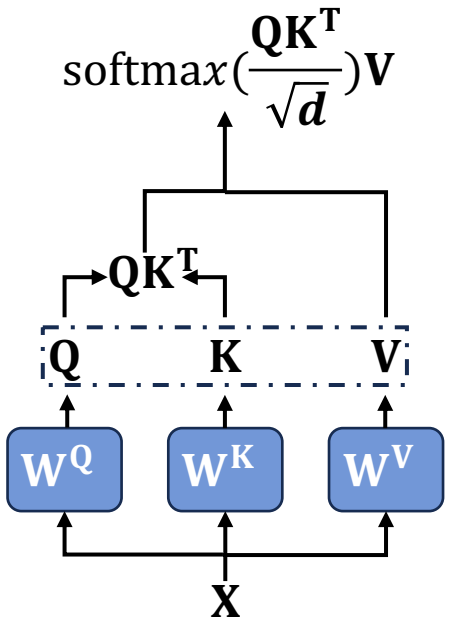
- **自注意力机制是长文本推理的瓶颈**

$$\mathrm{ATTN}(\mathrm{Q},\mathrm{K},\mathrm{V}) = softmax\left(\frac{\mathrm{QK^T}}{\sqrt{d}}\right)\mathrm{V}$$

- 当输入上下文变得极长时：

键值缓存变得极大	计算复杂度呈$O(n^2)$增长	多模态长视频推理
↕	↕	↕
开销高昂	**速度缓慢**	**双重瓶颈**

- **开销高昂：** 无法在低显存设备上推理
- **速度缓慢：** 处理长文本请求的速度慢
- **双重瓶颈：** 视频与文本编码均为计算瓶颈



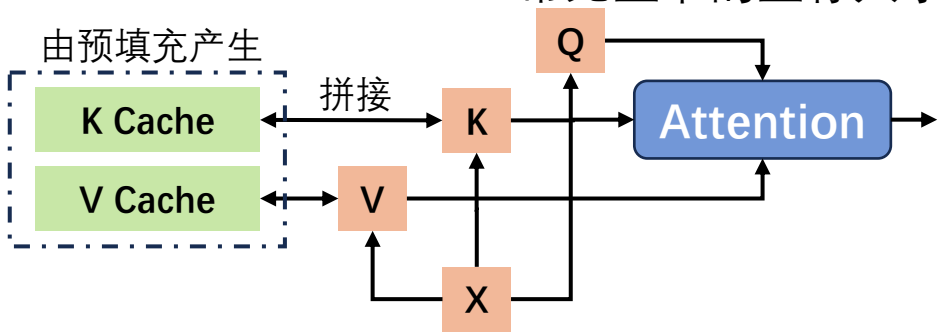
自注意力机制

模型	全量键值缓存大小
Yi-200K	100GB
MiniCPM-128K	45GB
ChatGLM3	64GB

常见模型的全量键值缓存大小

GPU 型号	显存大小
NVIDIA A100-80GB	80GB
NVIDIA A100-40GB	40GB
NVIDIA 4090	24GB

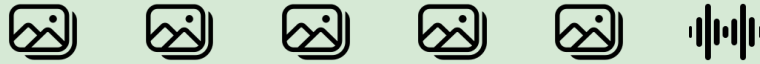
常见显卡的显存大小

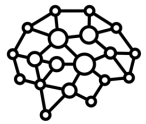


使用键值缓存解码

研究组织

长上下文任务

- 长文摘要：人工智能从上世纪五十年代开始 …… 请使用150字概括上述文本：
- 大海捞针：下面文档中有一个7位数字，找到并记住它……上述文档的7位数字是：
- 视频理解：  ……上述视频中有什么？



大语言模型

挑战1：开销高昂

面向单卡场景的键值缓存丢弃框架Locret

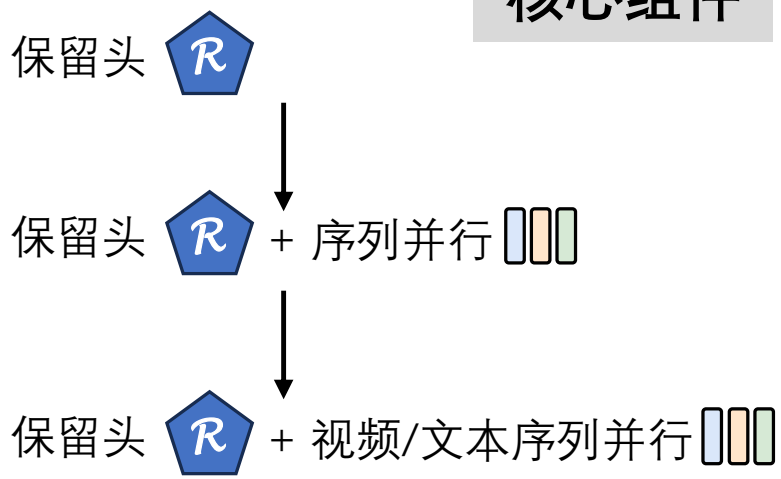
挑战2：速度缓慢

面向分布式多主机场景的长文本推理框架APB

挑战3：双重瓶颈

面向多模态模型的长视频推理框架APB-V

核心组件



1. 面向单卡场景的键值缓存丢弃框架Locret

主要挑战：单卡显存难以容纳长上下文键值缓存

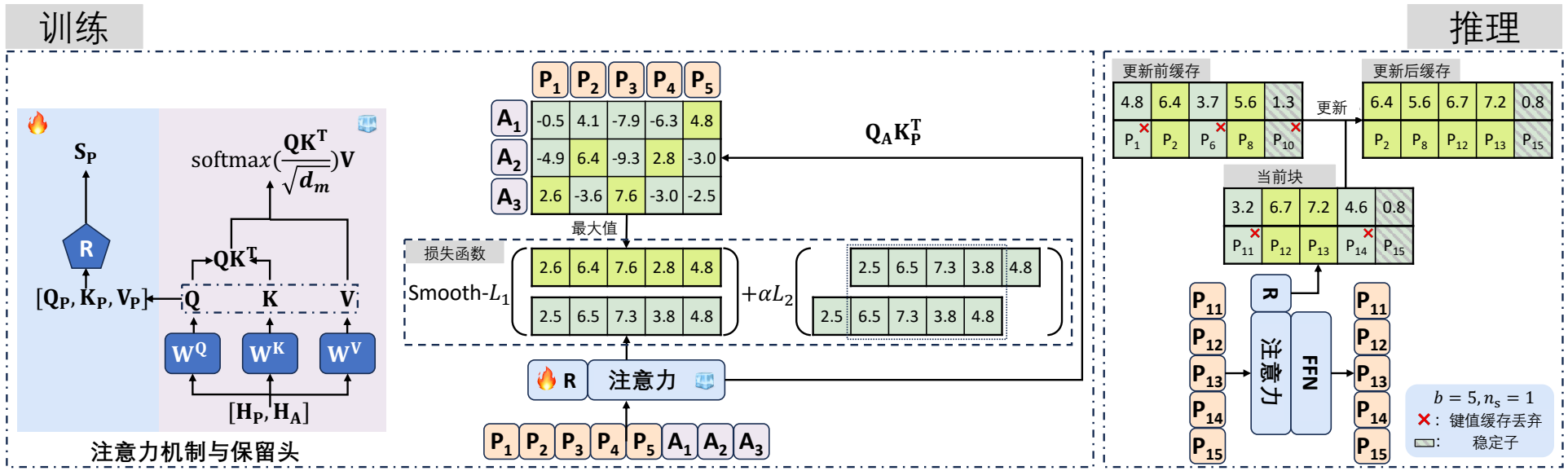
- 将分块预填充与键值缓存丢弃结合
- 在长SFT数据上训练保留头
- 保留头仅根据$[Q,K,V]$预测键值对的重要性
- 推理时在分块预填充过程中丢弃重要性较低的缓存

训练目标：

$$\underset{\mathbf{W}_1^{(i)},\mathbf{W}_2^{(i)},i=1,2\cdots,L}{\operatorname{argmin}} \mathbb{E}_{d\in\mathcal{D}}\left[\sum_{i=1}^{L}\sum_{j=1}^{h}\sum_{k=1}^{n_q(d)}\mathcal{L}\left(\tilde{\mathbf{S}}[k]_j^{(i)},\mathbf{S}[k]_j^{(i)}\right)\right]$$

损失函数：

$$\mathcal{L}\left(\tilde{\mathbf{S}}[k]_j^{(i)},\mathbf{S}[k]_j^{(i)}\right)=\text{Smooth-}\mathcal{L}_1\left(\tilde{\mathbf{S}}[k]_j^{(i)},\mathbf{S}[k]_j^{(i)}\right)+\alpha\mathcal{L}_2\left(\tilde{\mathbf{S}}[k]_j^{(i)},\tilde{\mathbf{S}}[k+1]_j^{(i)}\right)$$



Locret: Enhancing Eviction in Long-Context LLM Inference with Trained Retaining Heads on Consumer-Grade Devices (Arxiv:2410.01805)

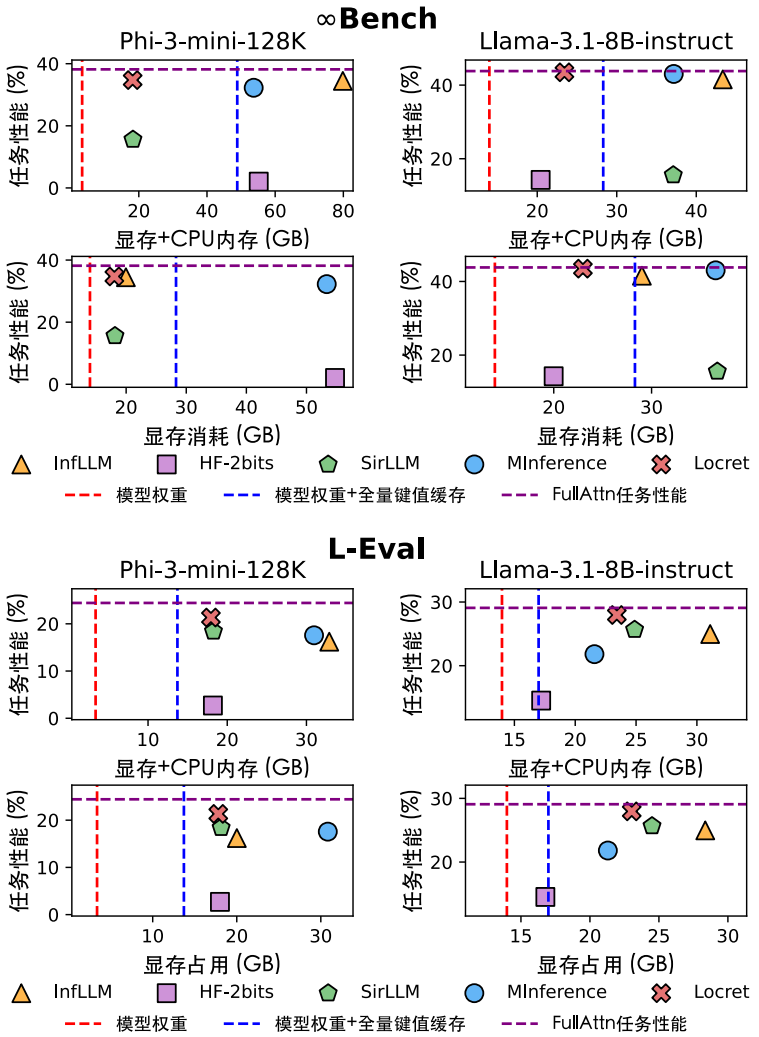
1. 面向单卡场景的键值缓存丢弃框架Locret

主要挑战：单卡显存难以容纳长上下文键值缓存

- Locret在使用更少的显存时，取得更好的端到端性能
- Locret在消费级设备上取得更高的推理速度
- Locret-Q能处理RULER等查询驱动任务

方法		FullAttn	InfLLM	HF-2bits	SirLLM	MInference	**Locret**	HF-2bits*	MInference*
Phi-3-mini-128K	tok/s↑	-	2276.38	-	2352.20	-	**5080.85**	1098.51	4099.92
	C.Len.↑	128K	128K	128K	128K	128K	**128K**	30K	14K
	Acc.↑	OOM	99.83	OOM	1.69	OOM	**100.00**	0.00	13.56
	成功	✗	✓	✗	✗	✗	✓	✗	✗
Llama-3.1-8B-Instruct	tok/s↑	-	2287.66	1365.51	1589.75	-	**3209.10**	3680.06	5135.74
	C.Len.↑	128K	128K	128K	128K	128K	**128K**	30K	25K
	Acc.↑	OOM	100.00	35.59	1.69	OOM	**100.00**	26.78	20.34
	成功	✗	✓	✗	✗	✗	✓	✗	✗

消费级设备推理速度

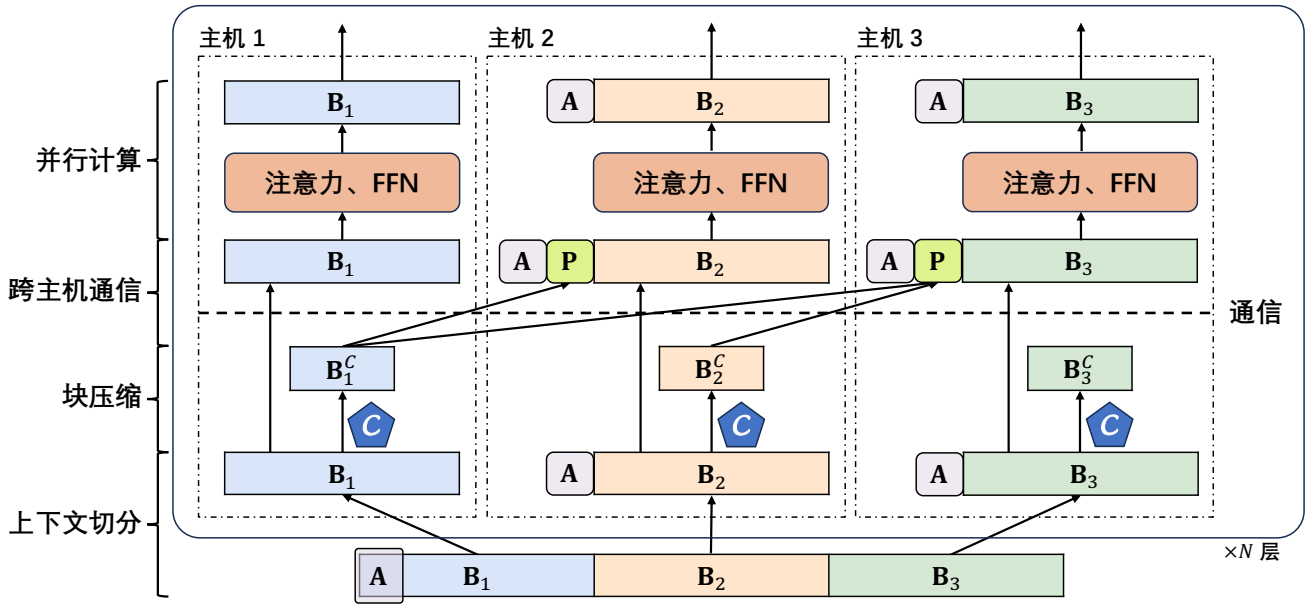


Locret: Enhancing Eviction in Long-Context LLM Inference with Trained Retaining Heads on Consumer-Grade Devices (Arxiv:2410.01805)

2. 面向分布式多主机的长文本推理框架APB

主要挑战：长文本预填充速度慢

- 现存优化方法难以结合：减少计算、提升并行度
- **同时减少计算与提升并行度：设计面向序列并行的近似注意力机制**
- 使用Locret中的保留头实现单卡内的键值缓存压缩
- 仅通信和计算压缩过的键值缓存，节省计算量



锚块:

查询 B_1的前缀

传递块:

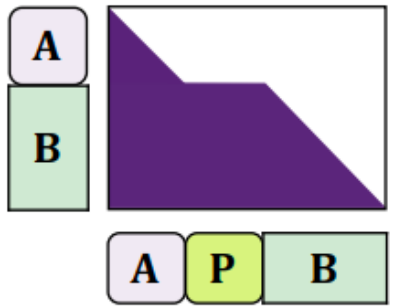
主机 2:

$P = B_1^C$

主机 3:

$P = B_1^C B_2^C$

注意力掩码:

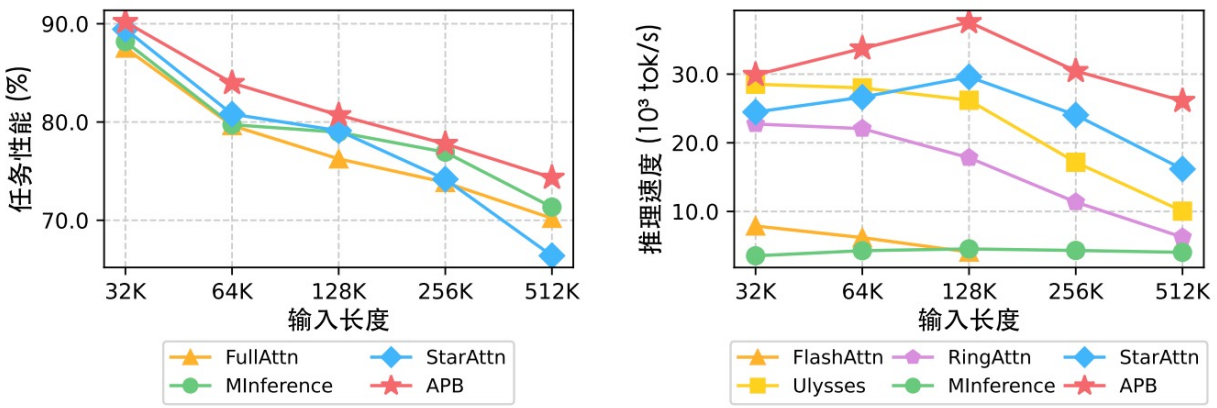


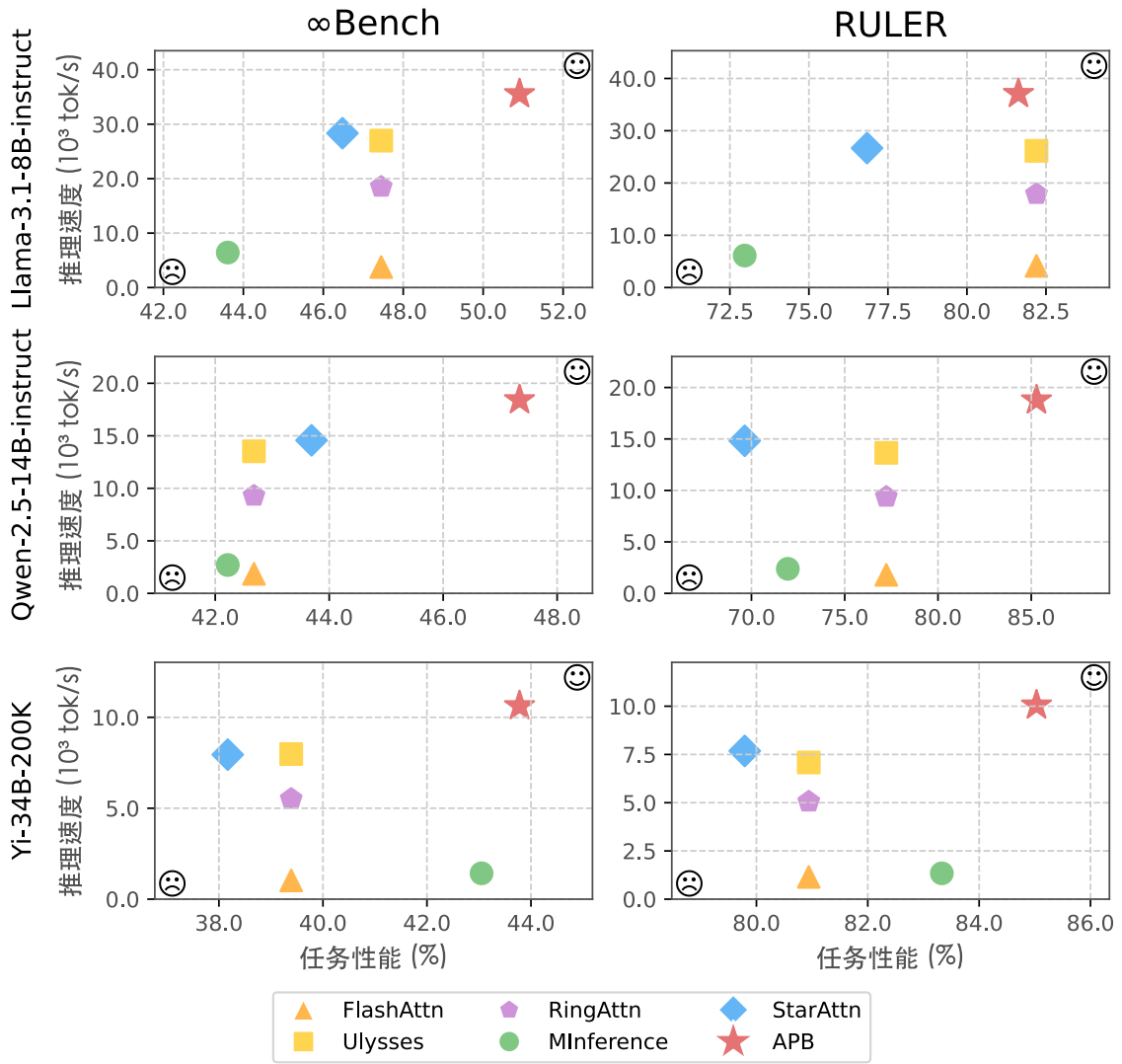
APB: Accelerating Distributed Long-Context Inference by Passing Compressed Context Blocks across GPUs (ACL 2025 main)

2. 面向分布式多主机的长文本推理框架APB

主要挑战：长文本预填充速度慢

- APB能在无性能损失时，显著提升推理速度
- APB适用于不同长度上下文的推理加速
- APB通过同时并行与加速Attn与FFN模块





APB: Accelerating Distributed Long-Context Inference by Passing Compressed Context Blocks across GPUs (ACL 2025 main)

3. 面向多模态模型的长视频推理框架APB-V

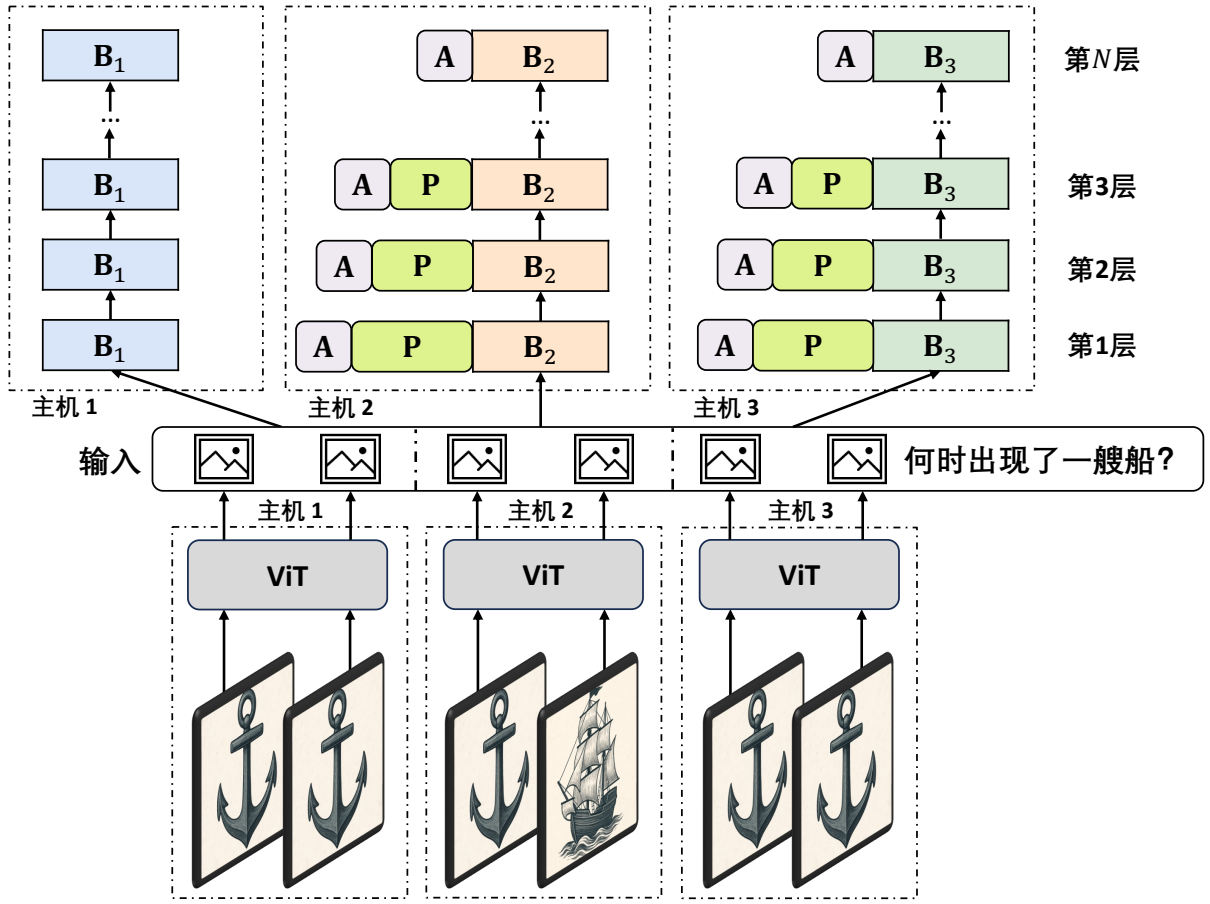
主要挑战：长视频在视频编码与文本基座上均存在计算瓶颈

- 统一使用序列并行作为加速技术
- 视频编码（ViT）：按帧分配到各GPU上并行计算
- 文本基座（LLM）：运行优化过的APB算法
- 多模态模型上的APB改进：
  - 注意力分数较为接近：使用排序学习训练保留头

$$\mathcal{L}_{\text{ListMLE}}\left(g_\theta(\mathbf{x}), y\right) = -\log P\left(\mathbf{y}|\mathbf{x}, g_\theta(\mathbf{x})\right)$$

$$P\left(\mathbf{y}|\mathbf{x}, g_\theta(\mathbf{x})\right) = \prod_{i=1}^{n} \frac{\exp\left(g(x_{y(i)})\right)}{\sum_{j=i}^{n} \exp\left(g(x_{y(j)})\right)}$$

  - 多模态输入稀疏度更高：传递长度随模型层数加深而线性递减

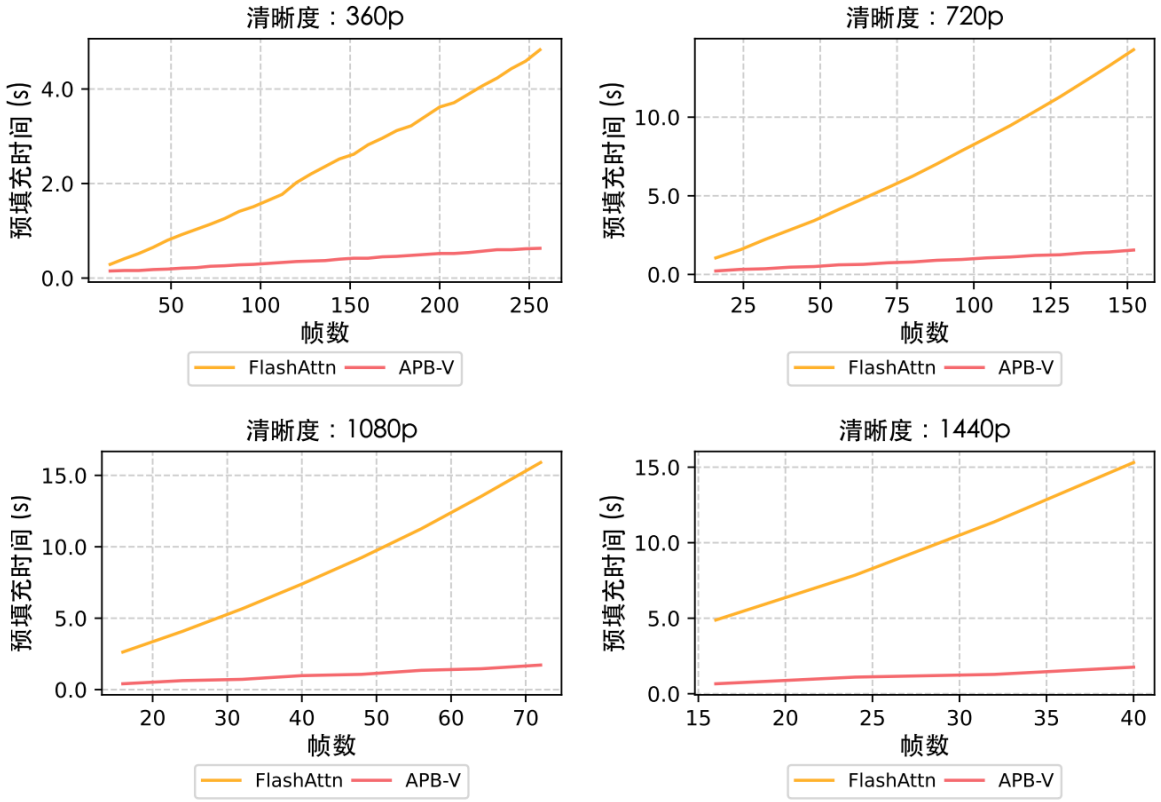


3. 面向多模态模型的长视频推理框架APB-V

主要挑战：长视频在视频编码与文本基座上均存在计算瓶颈

- APB-V不会影响长视频理解的任务性能
- APB-V加速效果明显，且在高清超长视频上加速效果更为显著

方法	8s-15s (%)	15s-60s (%)	180s-600s (%)	900s-3600s (%)	全部数据 (%)
Qwen-2.5-VL-3B-Instruct					
FullAttn	**73.54**	69.77	54.85	48.23	56.62
APB-V	70.37	**70.93**	**56.55**	**48.40**	**56.92**
Qwen-2.5-VL-7B-Instruct					
FullAttn	73.02	**72.09**	57.28	**48.94**	57.89
APB-V	**77.25**	70.35	**59.22**	48.58	**58.71**



总结与展望

挑战1：长上下文推理开销高昂

Locret框架：使用后训练方法结合分块预填充与键值缓存丢弃，单卡推128K长文本

挑战2：长上下文推理速度缓慢

APB框架：面向序列并行场景设计近似注意力机制，实现15s内预填充512K上下文

挑战3：长视频推理在视频编码与文本基座上均存在计算瓶颈

APB-V框架：通过序列并行+文本基座近似注意力机制，支持超长视频秒级问答

感谢北自然启研计划对本课题的支持

APB框架：已经被ACL 2025 主会接收

Locret框架：投稿至COLM会议，在审

未来工作

1. 集成至vLLM等开源推理引擎
2. 在更多端侧硬件上测试算法
3. 验证更多模型架构
4. 实现推理感知的预训练
5. 拓展本工作的理论框架

基于稀疏注意力机制的超长上下文推理加速方法

敬请各位老师批评指正！