

KV Cache 公式纸

对应文章《从逐词生成理解 KV Cache》· 一页，方便对照数量级

显存

KV 显存与层数 L、头维 H、序列长度 T、批量 batch 成正比。

$$\text{KV memory} \propto 2 \times L \times H \times T \times \text{batch}$$

数量级

模型	batch × 序列	现象
70B	长上下文	KV 会先于权重满一张卡
7B	batch 8 × 32K	消费卡也会先被 KV 撑满

和计算的关系

- 朴素 decode：6 个 token 共 21 次投影，15 次是浪费
- Prefill 一次写满 K/V；Decode 每步只追加一行
- 缓存换的是时间，付的是显存