PR #26049 完整报告

sgl-project/sglang

Fix GLM NextN draft value head dim

合并时间：2026-06-10 07:13

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/26049

执行摘要

- 一句话：修复 GLM NextN draft 的 v_head_dim 推导
- 推荐动作：值得精读。虽然改动了 11 行，但其背后的 root cause 揭示了 MHA/MLA 混合架构下的 v_head_dim 继承陷阱。核心设计要点是：当模型配置源自 MLA 架构时，v_head_dim 可能远大于 MHA draft 的真实 head_dim，必须在架构重写时显式修正。此模式可作为未来类似 MHA-over-MLA draft adapter 的参考。

功能与动机

GLM Lite 模型使用 MLA 配置，v_head_dim 可能远大于 head_dim（如前者 256 后者 64）。当 NextN draft worker 重写架构为 MHA 后，KV pool 因从 model_config.v_head_dim 读取了过大的值，分配了比 draft 模型实际写入更宽的缓存，导致存储 draft KV 时 shape 不匹配崩溃。

实现拆解

1. 在 model_config.py 的 _derive_model_shapes 方法中增加分支：当架构名包含 Glm4MoeForCausalLMNextN 时，进入专有处理。
2. 回退 head_dim 来源：若 head_dim 为 None，则从 hidden_size // num_attention_heads 计算标准 MHA 的 head_dim（等同于 draft 模型实际使用的值）。
3. 修正 v_head_dim 与 swa_v_head_dim：将这两个维度显式设为与 head_dim 一致（v_head_dim = self.head_dim; swa_v_head_dim = self.swa_head_dim），覆盖来自 MLA target 配置的过大值。
4. 标记注意力架构：设置 self.attention_arch = AttentionArch.MHA 确保后续逻辑按 MHA 处理。

此变更仅影响 Glm4MoeForCausalLMNextN 架构的 KV cache 形状推导，无其他模块影响。

关键文件：

- python/sglang/srt/configs/model_config.py（模块 配置层；类别 source；类型 data-contract）：唯一修改的文件，新增 Glm4MoeForCausalLMNextN 架构分支，修正 v_head_dim 和 swa_v_head_dim，标记 MHA 类型。

关键符号：_derive_model_shapes

关键源码片段

python/sglang/srt/configs/model_config.py

唯一修改的文件，新增 Glm4MoeForCausalLMNextN 架构分支，修正 v_head_dim 和 swa_v_head_dim，标记 MHA 类型。

```
# python/sglang/srt/configs/model_config.py (modified)
# 在 _derive_model_shapes 中，DeepseekV4 分支之后插入
elif "Glm4MoeForCausalLMNextN" in self.hf_config.architectures:
    # 如果 head_dim 未显式指定，根据标准的 MHA 公式计算
    if self.head_dim is None:
        self.head_dim = (
            self.hf_text_config.hidden_size
            // self.hf_text_config.num_attention_heads
        )
    # swa_head_dim 同样回退到 head_dim
    if self.swa_head_dim is None:
        self.swa_head_dim = self.head_dim
    # 关键修正：将 v_head_dim 强制设回 head_dim，而非沿用 target 模型
    # 的 MLA 大 v_head_dim（如 256 vs 64），避免 KV cache 分配过宽
    self.v_head_dim = self.head_dim
    self.swa_v_head_dim = self.swa_head_dim
    # 明确标记为 MHA 类型，不再走后续的 MLA 分支
    self.attention_arch = AttentionArch.MHA
```

评论区精华

无 review 评论。仅 reviewer Qiaolin-Yu 直接批准。

- 暂无高价值评论线程

风险与影响

- 风险：低风险。
- 变更仅在特定架构 Glm4MoeForCausalLMNextN 分支内新增逻辑，不影响其他分支。
- 若 head_dim 提前被设为正确值（如从 HF config 直接读取），追加的 if self.head_dim is None 回退不会覆盖；但若其他代码路径在 _derive_model_shapes 前已依赖 v_head_dim 的原始 MLA 值，则按新逻辑 v_head_dim = head_dim 可能反而缩小为 draft 真实宽度，此修正正是预期行为。
- 无新增依赖或配置开关。
- 自带 CI 测试（test_dsa_glm5_dp_mtp.py 等）已通过。
- 影响：影响范围：仅 Glm4MoeForCausalLMNextN 架构的 draft worker 初始化阶段。 用户影响：修复了 GLM Lite 模型在使用 NextN 推测解码时因 KV cache shape mismatch 导致的崩溃；使用普通 GLM 或其他架构的用户不受影响。 影响程度：修复性变更，无功能新增或退化。
- 风险标记：核心路径变更，跨架构兼容

关联脉络

- PR #27607 Support spec v2 for Frozen-KV MTP; remove v1 worker: 同为 speculative decoding 架构适配，涉及 draft worker 的架构重写和 KV cache 配置。