PR #32440 完整报告

sgl-project/sglang

fix(gemma4): quantize MTP bridge projections

合并时间：2026-08-20 03:41

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/32440

执行摘要

- 一句话：修复 Gemma4 MTP 量化配置遗漏，FP8 接受率 0% 提升至 60%
- 推荐动作：值得快速精读。虽然代码改动仅 2 行，但背后是一个完整的调试故事：用 draft-token 接受率作为量化健康的观测指标，反推出 quant_config 传递遗漏。建议关注两点：一是量化配置在不同模型组件间传播时的一致性校验；二是为 pre_projection / post_projection 的量化加载补充专门的回归测试，避免同类问题复发。

功能与动机

PR body 明确指出：block-FP8 的 gemma-4-31B-it-assistant 检查点 draft-token 接受率为 0%，而 BF16 assistant 约为 65%。检查点将 pre_projection 与 post_projection 存为 block-FP8 权重并带 weight_scale 张量，但这些层以 quant_config=None 构造，导致 FP8 权重加载时缺少 block scales，破坏了 MTP 激活。作者因此将 draft model 的 quant_config 透传给这两个桥接投影层。

实现拆解

1. 定位缺陷：在 python/sglang/srt/models/gemma4_mtp.py 的 MTP 模型 __init__中，pre_projection 与 post_projection 两个桥接线性层构造时硬编码 quant_config=None，与主干 Gemma4TextModel 透传 quant_config 的行为不一致。
2. 修复方式：将两处 quant_config=None 改为 quant_config=quant_config，让桥接投影层继承 draft model 的量化配置，从而在加载 block-FP8 检查点时正确消费 weight_scale 与分块量化参数。
3. 影响分析：BF16 路径不受影响（quant_config 仍为 None）；FP8 路径的 MTP draft-token 接受率由 0% 恢复至约 60%。
4. 验证与配套：未新增单元测试；通过 /rerun-test 重跑 registered/spec/test_frozen_kv_mtp.py（1-gpu-h100）与 registered/spec/test_gemma4_mtp_31b_extra.py（2-gpu-h100）均通过。

关键文件：

- python/sglang/srt/models/gemma4_mtp.py（模块 MTP 模型；类别 source；类型 data-contract；符号 Gemma4MTP.init, pre_projection, post_projection）：唯一的变更文件，修复 Gemma4 MTP Frozen-KV 桥接投影层 pre_projection / post_projection 的量化配置传递，使 block-FP8 检查点的 weight_scale 能正确加载，将 FP8 draft-token 接受

率从 0% 提升至 60%。

关键符号：Gemma4MTP.init

关键源码片段

python/sglang/srt/models/gemma4_mtp.py

唯一的变更文件，修复 Gemma4 MTP Frozen-KV 桥接投影层 pre_projection / post_projection 的量化配置传递，使 block-FP8 检查点的 weight_scale 能正确加载，将 FP8 draft-token 接受率从 0% 提升至 60%。

```
def __init__(
    self,
    config: PretrainedConfig,
    quant_config: Optional[QuantizationConfig] = None,
    prefix: str = "",
) -> None:
    text_config = copy.deepcopy(_get_text_config(config))
    text_config.num_kv_shared_layers = 0
    PreTrainedModel.__init__(self, config=text_config)
    self.assistant_config = config
    self.config = text_config
    self.quant_config = quant_config
    self.pp_group = get_pp_group()

    self.vocab_size = text_config.vocab_size
    self.hidden_size = text_config.hidden_size
    self.backbone_hidden_size = config.backbone_hidden_size
    self.target_embed_scale = self.backbone_hidden_size ** 0.5
    self.use_ordered_embeddings = bool(
        getattr(config, "use_ordered_embeddings", False)
    )
    self.centroid_intermediate_top_k = int(
        getattr(config, "centroid_intermediate_top_k", 32)
    )

    self.target_embed_weight: Optional[torch.Tensor] = None

    # 关键修复：pre_projection / post_projection 之前硬编码 quant_config=None，
    # 导致 block-FP8 分片权重加载时丢失 weight_scale，MTP 激活被破坏，
    # draft-token 接受率从约 65%（BF16）跌到 0%（FP8）。
    # 现在透传 draft model 的 quant_config；BF16 路径不受影响（其值仍为 None）。
    self.pre_projection = ReplicatedLinear(
        2 * self.backbone_hidden_size,
        self.hidden_size,
        bias=False,
        quant_config=quant_config,
        prefix=add_prefix("pre_projection", prefix),
    )
```

```
        self.model = Gemma4TextModel(
            config=text_config,
            quant_config=quant_config,
            prefix=add_prefix("model", prefix),
        )
        self.post_projection = ReplicatedLinear(
            self.hidden_size,
            self.backbone_hidden_size,
            bias=False,
            quant_config=quant_config,
            prefix=add_prefix("post_projection", prefix),
        )

        if text_config.tie_word_embeddings:
            self.lm_head = self.model.embed_tokens
        else:
            self.lm_head = nn.Linear(self.hidden_size, self.vocab_size, bias=False)
        self.logits_processor = LogitsProcessor(text_config, skip_all_gather=True)
```

评论区精华

该 PR 几乎没有实质 review 讨论：kpham-sgl 直接 APPROVED，唯一的审查动作是通过 /rerun-test 触发两个 MTP 专项测试并全部通过；PR 作者曾请求 @Jiminator 或 @JustinTong0323 参与 review，但无人留下技术评论。gemini-code-assist[bot] 仅声明其代码审查服务已停止。值得注意的是 checklist 中"添加单元测试"未被勾选，且审查过程无人对此提出异议，说明该修复目前依赖现有 spec 测试作为回归防线。

- MTP 专项测试重跑验证 (testing): 与 Gemma4 MTP 路径相关的 spec 测试通过，为合并提供了验证支撑。
- 缺少针对量化配置传播的单元测试 (testing): 未补充单元测试，依赖现有 spec 测试间接覆盖；该缺口留待后续补强。

风险与影响

- 风险：
  1. 缺少量化配置传播的单元测试：本 PR 的 2 行修复没有附带回归测试，未来若再次出现类似的 quant_config 透传遗漏，难以被快速发现。
  2. 兼容性风险：如果某个检查点的 pre_projection / post_projection 实际未量化，但传入非 None 的 quant_config，可能导致量化加载路径误判或权重解析报错；当前仅以 gemma-4-31B-it-assistant 的实测结果作为验证。
  3. 影响面受控：修改仅发生在 Gemma4 MTP（Frozen-KV）模型构造路径，对 SRT 其他模型和普通解码路径无影响。 - 影响：对用户而言，运行 block-FP8 gemma-4-31B-it-assistant 的 speculative decoding 用户获得显著收益：draft-token 接受率从 0% 恢复至约 60%，意味着 MTP 预测从完全失效变为可用，端到端解码吞吐随之改善。对团队而言，该修复暴露了模型层量化配置传播链路的一个系统性风险点，提示后续需要更系统的量化配置一致性检查。整体影响范围限定在 Gemma4 MTP 单一路径，

影响程度中等。 - 风险标记：量化配置传播遗漏，缺少测试覆盖，核心推理路径变更

关联脉络

- 暂无明显关联 PR